Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

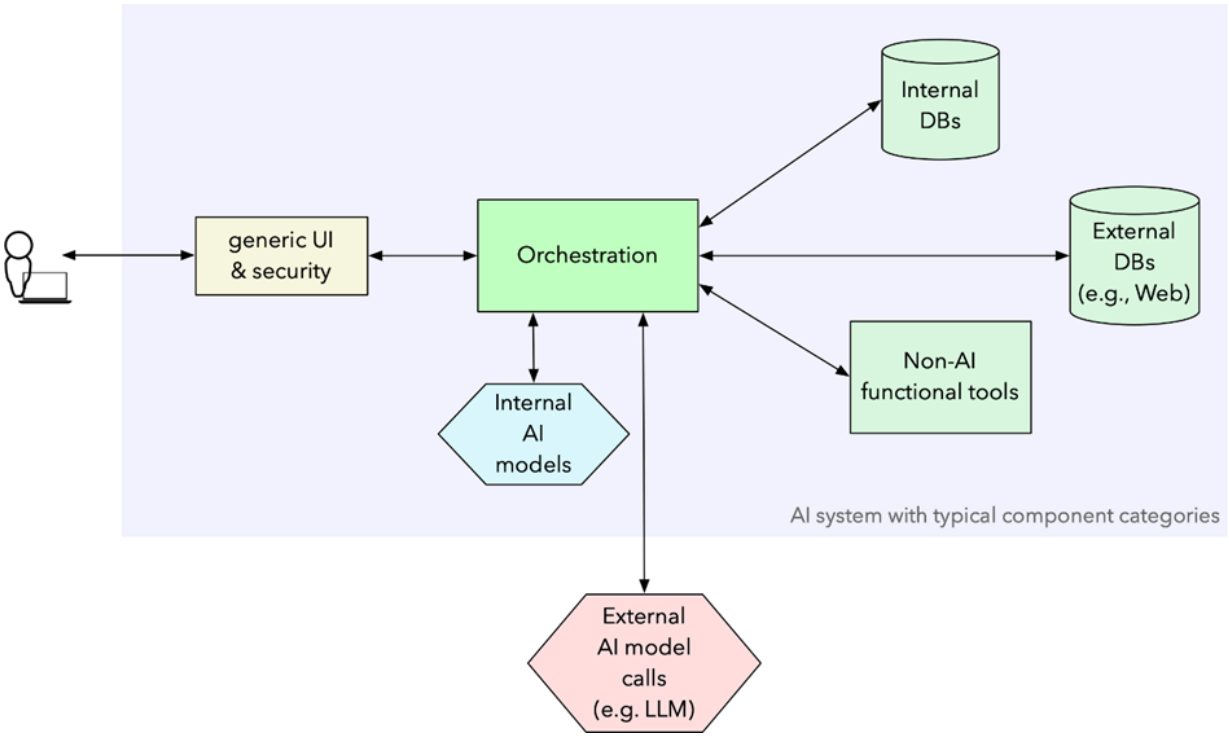
Federal Department of the Environment,
Transport, Energy and Communications DETEC

Swiss Federal Office of Energy
Energy Research and Cleantech

Final report from 12 May 2026

SAFE-AI

Sustainability **A**ssessment **F**ramework for the **E**nvironmental Impacts of **A**rtificial **I**ntelligence



Source: Vlad Coroamă, 2026

Publisher:
Swiss Federal Office of Energy SFOE
Energy Research and Cleantech
CH-3003 Berne
www.energy-research.ch

Subsidy recipients:
Roegen Centre for Sustainability GmbH
8002 Zürich
www.roegen.ch

Autors:
Vlad C. Coroamă, Roegen Centre for Sustainability GmbH, vlad.coroama@roegen.ch
Daniel Schien, University of Bristol, daniel.schien@bristol.ac.uk

SFOE project coordinators:
Dr. Michael Moser, michael.moser@bfe.admin.ch
Roland Brüniger, roland.brueniger@brueniger.swiss

SFOE contract number: SI/502781-01

The authors bear the entire responsibility for the content of this report and for the conclusions drawn therefrom.

Summary

The rapid proliferation of artificial intelligence (AI) brings not only major opportunities, but also growing environmental challenges. In particular, energy consumption and the demand for materials are increasing significantly, which in turn leads to higher greenhouse gas (GHG) emissions and further environmental burdens. However, accurately assessing these impacts is difficult. This is partly because AI systems are highly complex and diverse, and are constantly evolving. In addition, reliable data from developers is often lacking, and there are still no standardised assessment methods. Different areas of AI application also require vastly different amounts of energy, making comparisons even more difficult.

The SAFE-AI framework ("Sustainability Assessment Framework for the Environmental Impacts of Artificial Intelligence") was developed to address these problems. SAFE-AI provides a structured basis for better understanding and assessing the environmental impacts of AI, while focusing on four key aspects: identifying the main drivers of energy consumption, proposing suitable assessment methods, defining meaningful units of comparison, and providing a clear process for assessing AI systems.

SAFE-AI distinguishes between three main assessment levels: the *AI model*, the *AI system*, and individual *AI usage*. For the widely used large language models, the assessment of the *AI model* level is reserved for the providers of foundational models, as only they have access to relevant data such as energy consumption during training and use. Companies that use such models via interfaces must therefore rely on published data and apply it to their own applications.

At the level of the AI system, all further components are typically considered, such as internally deployed AI models and other elements of the AI ecosystem. These can often be measured directly and are therefore particularly relevant for companies. This report is thus aimed primarily at practitioners who deploy *AI systems* and wish to better understand their energy consumption.

An important question is how AI usage can be measured in a meaningful way. Average values can be misleading, as individual queries vary considerably. The report shows that the number of output and input tokens is particularly suitable for estimating energy consumption.

On this basis, the report proposes a directly applicable operational workflow for assessments on *AI system* level. The workflow is designed to be generic and can be applied to a wide range of AI systems. For external AI models that cannot be measured directly, the process relies on a stochastic analysis of the token distribution per request. This enables more robust aggregation at system level when only the total number of requests and tokens is known.

Using various system architectures, the report shows how different AI systems can be structured. One concrete example is a sustainability chatbot developed by the city of Zurich. The analysis shows that, in this case, the majority of energy consumption is not caused by the external AI models, but by the local system itself, mainly due to its low utilisation.

This finding highlights that a holistic perspective is crucial: anyone wishing to understand the environmental impacts of AI must not look only at individual models, but must take the entire system into account.

Beyond their direct footprint, AI systems also induce indirect environmental consequences (which can be both beneficial and detrimental) in their respective deployment sectors. These are typically varied, complex, and deeply intertwined. As such, they are difficult to assess generically. This report performs two such assessments: for the same sustainability chatbot of the city of Zurich and for a clinical AI conversational agent. They show how AI may bring about environmental benefits in two different sectors (circular economy and healthcare, respectively), but also how contextual such assessments are.

Zusammenfassung

Die rasante Verbreitung künstlicher Intelligenz (KI) bringt nicht nur grosse Chancen, sondern auch wachsende ökologische Herausforderungen mit sich. Insbesondere steigen der Energieverbrauch und der Bedarf an Materialien deutlich an, was wiederum zu höheren Treibhausgasemissionen und weiteren Umweltbelastungen führt. Die genaue Bewertung dieser Auswirkungen ist jedoch schwierig. Das liegt unter anderem daran, dass KI-Systeme sehr komplex und vielfältig sind und sich ständig weiterentwickeln. Zudem fehlen oft verlässliche Daten von den Entwicklern, und es gibt noch keine einheitlichen Methoden zur Bewertung. Unterschiedliche Einsatzbereiche von KI benötigen ausserdem sehr unterschiedliche Energiemengen, was Vergleiche zusätzlich erschwert.

Um diese Probleme anzugehen, wurde das SAFE-AI-Framework entwickelt. Es bietet eine strukturierte Grundlage, um die Umweltwirkungen von KI besser zu verstehen und zu bewerten. Dabei konzentriert es sich auf vier zentrale Aspekte: die wichtigsten Ursachen für Energieverbrauch zu identifizieren, geeignete Bewertungsmethoden vorzuschlagen, sinnvolle Vergleichseinheiten zu definieren und einen klaren Ablauf für die Bewertung von KI-Systemen bereitzustellen.

Ein zentrales Element des Frameworks ist die Unterscheidung zwischen drei Ebenen: dem *KI-Modell*, dem *KI-System* und der individuellen *KI-Nutzung*. Besonders bei den weit verbreiteten grossen Sprachmodellen (large language models, LLMs) liegt die Verantwortung für die Bewertung auf Seiten der Anbieter, da nur sie Zugriff auf relevante Daten wie den Energieverbrauch beim Training und bei der Nutzung haben. Unternehmen, die solche Modelle über Schnittstellen nutzen, müssen sich daher auf veröffentlichte Daten stützen und diese auf ihre eigenen Anwendungen übertragen.

Auf Ebene des *KI-Systems* werden hingegen alle weiteren Komponenten betrachtet, etwa eigene Modelle, sowie weitere Bestandteile des KI-Ökosystems. Diese lassen sich oft direkt messen und sind daher für Unternehmen besonders relevant. Der Bericht richtet sich deshalb vor allem an Praktikerinnen und Praktiker, die *KI-Systeme* einsetzen und deren Energieverbrauch besser verstehen möchten.

Ein wichtiger Punkt ist die Frage, wie man Nutzung sinnvoll misst. Durchschnittswerte können täuschen, da einzelne Anfragen sehr unterschiedlich sind. Der Bericht zeigt, dass sich insbesondere die Anzahl der verarbeiteten Ein- und Ausgabeeinheiten (Tokens) gut eignet, um den Energieverbrauch abzuschätzen.

Auf dieser Basis schlägt der Bericht einen unmittelbar anwendbaren Ablauf für Bewertungen auf Ebene des KI-Systems vor. Dieser ist generisch angelegt und kann auf eine Vielzahl von KI-Systemen angewendet werden. Für nicht unmittelbare messbare externe KI-Modelle stützt sich der Ablauf auf eine stochastische Analyse der Tokenverteilung pro Anfrage. Dies ermöglicht robustere Aggregationen auf Systemebene, wenn lediglich die Gesamtzahl der Anfragen und Tokens bekannt ist.

Anhand verschiedener Systemarchitekturen wird gezeigt, wie unterschiedliche KI-Systeme aufgebaut sein können. Ein konkretes Beispiel ist ein Nachhaltigkeits-Chatbot der Stadt Zürich. Die Analyse zeigt, dass in diesem Fall der Grossteil des Energieverbrauchs nicht durch die externen KI-Modelle entsteht, sondern durch das lokale System selbst – vor allem wegen dessen geringer Auslastung.

Diese Erkenntnis macht deutlich, dass eine ganzheitliche Betrachtung entscheidend ist: Wer die Umweltwirkungen von KI verstehen will, darf nicht nur auf einzelne Modelle schauen, sondern muss das gesamte System berücksichtigen.

Über ihrem direkten Fussabdruck hinaus, verursachen KI-Systeme jedoch auch indirekte Umweltfolgen in ihren jeweiligen Anwendungsgebieten, welche sowohl positiv wie auch negativ sein können. Diese sind in der Regel vielfältig, komplex und eng miteinander verwoben; daher lassen sie sich nur schwer allgemein bewerten. Der Bericht führt zwei solche Bewertungen durch: für denselben Nachhaltigkeits-Chatbot der Stadt Zürich sowie für einen klinischen KI-Konversationsagenten. Sie zeigen, wie KI in zwei unterschiedlichen Sektoren – Kreislaufwirtschaft bzw. Gesundheitswesen – Umweltvorteile bewirken kann, aber auch, wie stark solche Bewertungen vom jeweiligen Kontext abhängen.

Résumé

La prolifération rapide de l'intelligence artificielle (IA) offre non seulement de grandes opportunités, mais pose aussi des défis environnementaux croissants. En particulier, la consommation d'énergie et la demande en matériaux augmentent fortement, entraînant des émissions accrues de gaz à effet de serre (GES) et d'autres impacts environnementaux. Il reste toutefois difficile d'évaluer précisément ces effets, notamment parce que les systèmes d'IA sont très complexes, variés et en constante évolution. En outre, les données fiables des développeurs font souvent défaut, et il n'existe pas encore de méthodes d'évaluation standardisées. Les différents domaines d'application de l'IA nécessitent aussi des quantités d'énergie très variables, ce qui complique encore les comparaisons.

Le cadre SAFE-AI (« Sustainability Assessment Framework for the Environmental Impacts of Artificial Intelligence ») a été développé pour répondre à ces problèmes. Il fournit une base structurée pour mieux comprendre et évaluer les impacts environnementaux de l'IA, en se concentrant sur quatre aspects clés : identifier les principaux moteurs de consommation d'énergie, proposer des méthodes d'évaluation adaptées, définir des unités de comparaison pertinentes et fournir un processus clair d'évaluation des systèmes d'IA.

SAFE-AI distingue trois niveaux d'évaluation principaux : le *modèle d'IA*, le *système d'IA* et *l'usage individuel de l'IA*. Pour les grands modèles de langage largement utilisés, l'évaluation au niveau du *modèle d'IA* revient aux fournisseurs de modèles fondamentaux, car eux seuls ont accès aux données pertinentes, telles que la consommation d'énergie lors de l'entraînement et de l'utilisation. Les entreprises qui utilisent ces modèles via des interfaces doivent donc s'appuyer sur les données publiées et les appliquer à leurs propres applications.

Au niveau du *système d'IA*, tous les autres composants sont généralement pris en compte, comme les modèles d'IA déployés en interne et d'autres éléments de l'écosystème IA. Ceux-ci peuvent souvent être mesurés directement et sont donc particulièrement pertinents pour les entreprises. Ce rapport s'adresse ainsi principalement aux praticiennes et praticiens qui déploient des systèmes d'IA et souhaitent mieux comprendre leur consommation d'énergie.

Une question importante est de savoir comment mesurer l'usage de l'IA de manière pertinente. Les valeurs moyennes peuvent être trompeuses, car les requêtes individuelles varient fortement. Le rapport montre que le nombre de jetons d'entrée et de sortie est particulièrement adapté pour estimer la consommation d'énergie.

Sur cette base, le rapport propose un flux opérationnel directement applicable aux évaluations au niveau du *système d'IA*. Ce flux est conçu de manière générique et peut s'appliquer à une large gamme de systèmes d'IA. Pour les modèles d'IA externes non mesurables directement, le processus s'appuie sur une analyse stochastique de la distribution des jetons par requête. Cela permet une agrégation plus robuste au niveau du système lorsque seuls le nombre total de requêtes et de jetons est connu.

À partir de diverses architectures, le rapport montre que les systèmes d'IA peuvent être structurés très différemment. Un exemple concret est un chatbot de durabilité développé par la Ville de Zurich. L'analyse montre que, dans ce cas, l'essentiel de la consommation d'énergie ne provient pas des modèles d'IA externes, mais du système local lui-même, principalement en raison de sa faible utilisation.

Ce constat souligne l'importance d'une perspective globale : pour comprendre les impacts environnementaux de l'IA, il ne suffit pas d'examiner des modèles individuels, il faut prendre en compte l'ensemble du système.

Au-delà de leur empreinte directe, les systèmes d'IA entraînent aussi des conséquences environnementales indirectes — positives comme négatives — dans leurs secteurs de déploiement. Celles-ci sont généralement variées, complexes et étroitement liées. Elles sont donc difficiles à évaluer de manière générique. Ce rapport réalise deux évaluations de ce type : pour le même chatbot de durabilité de la Ville de Zurich et pour un agent conversationnel d'IA clinique. Elles montrent comment l'IA peut apporter des bénéfices environnementaux dans deux secteurs différents — l'économie circulaire et la santé —, mais aussi combien ces évaluations dépendent du contexte.

Main findings («Take-Home Messages»)

- AI can be assessed on *model*, *system*, or *usage* level. Both meaningful level and available data sources depend on the aim of the assessment and who is performing it.
- For *AI models*, impacts arise across 3 orthogonal dimensions: the *why* (along the AI model's lifecycle: research & design, development incl. training, and deployment incl. inference), *when* (along the environmental lifecycle: production, usage, end-of-life), and *where* (ICT device category: data centres, networks, end devices). For energy, GHGs, and water, impacts occur mainly in DCs due to the electricity consumption in research, training, and inference; water additionally due to DC cooling.
- *AI systems* can employ internal and/or external AI models, which has an influence on where the assessment needs to start (on system or usage level, respectively) and which primary data can be measured. Equally important, however, is the assessment of the rest of the AI ecosystem, which in some cases can outweigh the impact of the AI models by orders of magnitude.
- For *AI usage*, different functional units can be used. In Transformer-based models, the number of input and output tokens is most characteristic. For the typical API-based usage of external models, simply using the average number of tokens per query would skew the result, as the impact does not scale linearly with the number of tokens. A stochastic analysis of token distribution across queries is thus preferred.

Contents

List of figures

List of tables

List of abbreviations

AI	artificial intelligence
ANN	artificial neural network
API	application programming interface
CoT	chain-of-thought
DB	database
DC	data centre
DL	deep learning
EE-MRIO	environmentally-enhanced multi-regional input-output
EoL	end-of-life (the last phase of an LCA)
FU	functional unit
GPU	graphics processing unit (often deployed in AI computations)
HBM	high-bandwidth memory
LCA	lifecycle assessment
LLM	large language model
LRM	large reasoning models
ML	machine learning (the prevalent type of AI nowadays)
MoE	mixture-of-experts
PCF	product carbon footprint
PUE	power usage effectiveness
RAG	retrieval-augmented generation
SFOE	Swiss Federal Office of Energy
SAM	segment anything model
SSM	state space model
TPU	tensor processing unit (a type of AI accelerator)
UI	user interface
UML	unified modelling language (visual language for diagrams of software systems)
VM	virtual machine

1 Introduction

Artificial Intelligence (AI) has emerged as a transformative technology with the potential to address critical sustainability challenges across various domains (Rolnick et al. 2022). At the same time, however, its rapid proliferation is accompanied by a growing energy and carbon footprint and other impacts, including water consumption, air and noise pollution, and rising power prices. Consequently, AI has been both hailed as a crucial tool for emission reductions and criticised for its environmental and societal risks.

However, assessing the environmental impacts of AI is extremely challenging due to a lack of data and established methodologies. This has resulted in wide-ranging estimates and projections, which causes confusion for decision-makers. For example, recent assessments of the near-term (i.e., 2030) energy consumption and greenhouse gas (GHG) emissions of AI diverge by an order of magnitude (Kamiya and Coroamă 2025), using a variety of methodologies, assumptions, and system boundaries. Existing assessments typically focus on operational impacts (use phase) and do not deploy full lifecycle assessments (LCAs) that would assess impacts from production and end-of-life (EoL) treatment as well.

More importantly, environmental assessments of AI have so far mainly focused either on estimating the energy consumption of training and/or inferencing individual AI models (such as a ChatGPT query or image generated), or on the global energy consumption of AI data centres (both today and in the future). Such assessments can be useful for AI model developers to measure and disclose their impacts, for global energy modelling, or for understanding the current and projected share of the AI sector among all economic sectors and human activities.

However, for companies integrating AI in their workflow, such assessments are not so useful. Instead, they require an assessment of integrated AI systems, which are important for understanding the overall impact of the entire AI ecosystem, comparability among systems, assessments of individual AI usage instances, and ultimately accountability and informed choices as well as policymaking.

1.1 Research objectives

Addressing and mitigating the direct environmental footprint of AI systems requires an accurate understanding of the energy consumption and environmental impacts of complex, real-world AI systems. In this context, the current study proposes a conceptual framework to enhance the consistency and comprehensiveness of environmental impact assessments of AI systems. The framework is named **Sustainability Assessment Framework for the Environmental Impacts of Artificial Intelligence (SAFE-AI)**.

The **key research questions** the SAFE-AI framework seeks to address are fourfold:

RQ1. *Identify most important sources of impact*: Both along the environmental lifecycle and across the different AI system components, which are the most important sources of energy consumption and greenhouse gas emissions?

RQ2. *Propose robust systemic assessment approaches*: What are the system dynamics shaping the environmental impact of AI, including the main drivers for energy consumption of AI models (and in particular large language models) and the entire AI ecosystem alike? Who has access to which data and what are the assessment implications of data visibility?

RQ3. *Examine possible functional unit definitions*: Which are the most promising approaches for defining suitable functional units (FUs) in heterogeneous AI systems and thus break down the overall system effect to individual usages?

RQ4. *Provide an AI system assessment workflow*: Can practitioners from companies that deploy in-house AI systems be provided with an energy consumption and GHG emissions assessment workflow? Can this workflow be generic enough to also cover external AI model calls?

Correspondingly, the core **target audiences** of SAFE-AI are:

- Companies that deploy AI systems and want to better understand – both conceptually and practically – the energy consumption and GHG impact of these systems.
- Further target audiences are policymakers who require the same understanding for policy decisions.
- As the framework also gives guidance on how to assess individual AI queries, another possible audience are individual users who aim to understand their own AI footprint.
- Finally, the research community is also addressed by parts of the framework, which proposes a stochastic method to compute average per-query energy consumption of transformer models when only the average number of input and output tokens are known.

1.2 Scope and focus

1.2.1. Focus on direct effects (footprint)

As with the digital sector more generally (Bremer et al. 2023), the relationship between AI and environmental impacts is complex and multifaceted. There are various ways to conceptualise this relation, which differ in several details. All of the better-known taxonomies such as (Hilty 2008; Williams 2011; Börjesson Rivera et al. 2014; Horner et al. 2016; Pohl et al. 2019; Coroamă et al. 2020), however, agree on the fundamental difference between:

- the **direct effects**, which occur within AI itself throughout the lifecycle of ICT components, i.e., during raw material extraction, device production, device operation (or usage), and end-of-life, and
- the **indirect effects**, which are the effects that occur in other domains (i.e., outside AI), being triggered by AI.

Direct effects (i.e., the footprint) are inherently detrimental to the environment. By contrast, the indirect effects – a part of which is sometimes referred to as “systemic effects” or “higher-order effects” – may be either beneficial or detrimental to the environment, as depicted in Figure 1.

Indirect beneficial effects include AI-triggered efficiency gains or dematerialisation such as building management systems that conserve energy or the support for circular economy processes. Indirect detrimental effects, meanwhile, often occur because of new technological possibilities, which lead to behavioural and systemic changes and ultimately additional consumption. These might be, for example, AI-powered autonomous vehicles, which are expected to substitute public transportation (Coroamă and Pargman 2020), or the deployment of AI for more efficient (but environmentally damaging) oil and gas drilling. As also shown in Figure 1, the net impact of AI (whether of a particular system or AI in general) is defined as the sum of all direct and indirect effects.

Both types of effects are important, and indirect effects are potentially larger than the direct effects, as qualitatively indicated in Figure 1. They are, however, also more uncertain and the methodologies to quantify them are much less developed than the already quite uncertain direct effects. Additionally, direct effects are more immediate while indirect effects are likely to take more time to take effect.

The indirect effects of AI are thus not at the core of this study. Given the rapid recent growth of AI, its wide potential environmental implications and uncertain future developments, the study focuses on the *direct environmental effects of AI*. The indirect effects, however, are present in two case studies, which will not only demonstrate the application of the framework, but also compare direct and indirect effects and thus address AI’s net impact.

As indicated in Figure 1, both energy consumption and GHG emissions are important to the analysis. They also correlate quite well, as for AI (and digitalisation more generally) electricity consumption is typically the dominant source of GHGs. The principles discussed throughout this study apply equally to both indicators, and more widely to any type of environmental impact and indicator used to measure it.

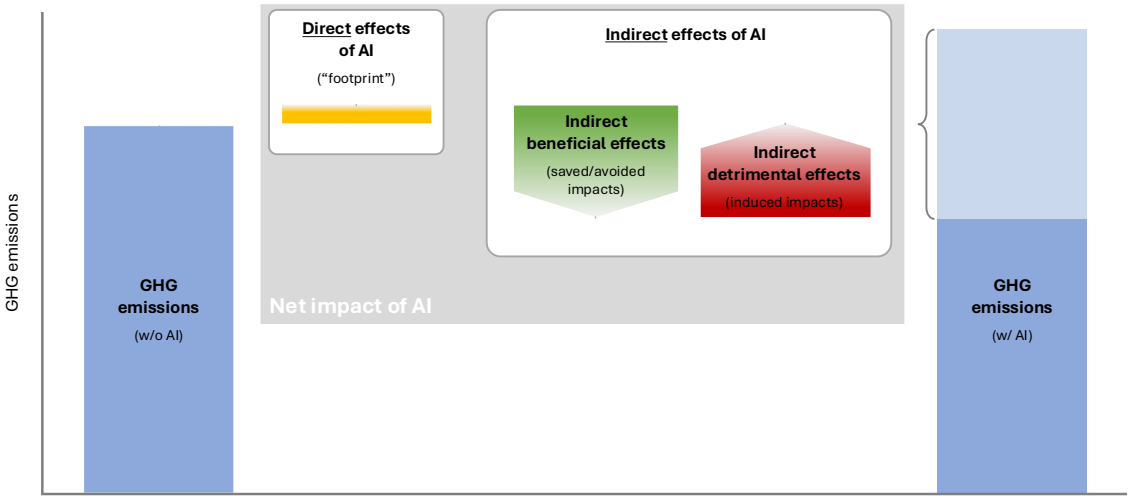


Figure 1: The three main types of environmental effects of AI: direct effects (i.e., footprint) which are by definition detrimental to the environment, indirect beneficial effects, and indirect detrimental effects. The figure refers to energy and GHGs specifically, but the principle is applicable to any type of environmental (and indeed societal) impact. Adapted from (Bremer et al. 2023); reprinted with permission.

There is nevertheless an important qualitative distinction between the two. Electricity-related GHG emissions depend not only on the amount of electricity consumed, but also on the *carbon intensity of electricity*. This can vary by more than one order of magnitude between different grid mixes or between location-based and market-based assessments as defined by the GHG protocol (Sotos 2015). These differences can easily lead to confusions and misunderstandings when comparing different estimates. The assumed grid mix can also significantly skew the GHG results. This study will thus use energy as main indicator; when necessary, the results can be easily transformed into GHGs based on different carbon intensity scenarios (Kamiya and Coroamă 2025).

Beyond energy and GHGs, the study also briefly addresses the water consumption of AI. The principles it lays out, however, are applicable to any other type of environmental impact such as resource depletion or toxic pollution. Finally, while its main focus lies on the direct impact of AI, the study also explores the relation between environmental costs and benefits in two specific case studies.

1.2.2. Focus on generative AI

Artificial intelligence (AI) is the computer science discipline focused on creating systems that emulate human intelligence. *Machine learning* (ML) is a prominent branch of AI that enables system to learn from data, discern underlying patterns, generate predictive models, and make decisions without being specifically programmed to do so. ML represents the dominant type of AI nowadays, and the one leading to a rapidly growing energy consumption and environmental impact. Other types of AI such as symbolic AI, which have been the focus of research for decades, are now marginal.

Within ML, *artificial neural networks* (ANNs) represent a class of algorithms whose layered architecture is inspired by the human brain. ANNs excel at approximating complex functions, making them highly effective for tasks such as computer vision, speech processing, and time-series forecasting. Finally, deep learning (DL) are complex ANNs, with a substantial amount of hidden layers (Schmidhuber 2015).

On the orthogonal dimension of an ML model's purpose, recent years have witnessed a rapid expansion of generative AI ("GenAI") models. GenAI models – including large language models (LLMs) such as GPT-5, Gemini, or Claude – create new content, such as text, images or video. They produce coherent, human-like outputs, being highly effective at summarising information and generating insights in natural language. These capabilities made GenAI gain rapid popularity since the first public release of the GPT model in late 2022, helping AI in general to become both a widely used technology and a household name around the world.

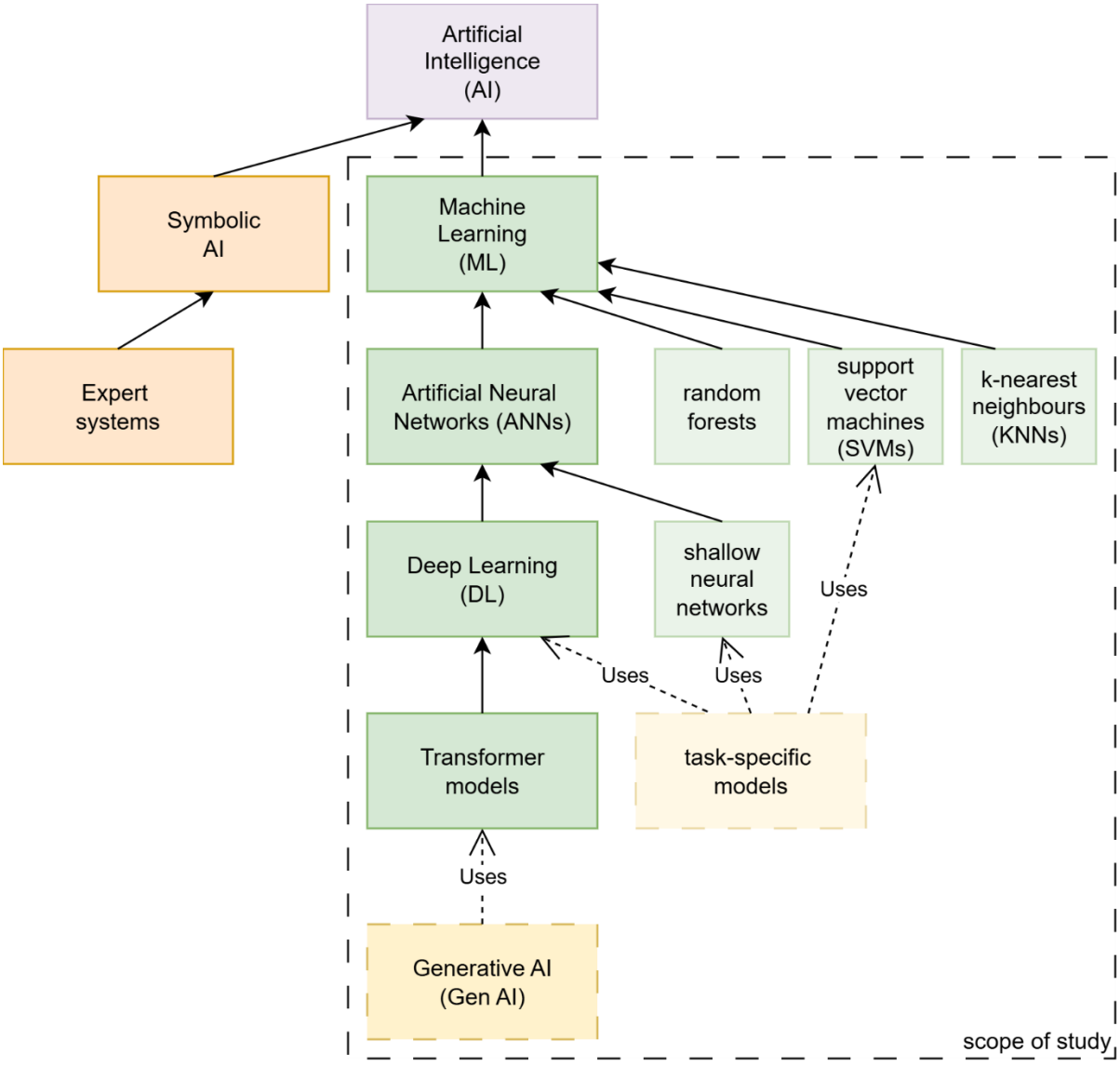


Figure 2: An overview of several types of AI, together with their relations. Filled arrows have the same semantic as in UML diagrams: they denote conceptual specialisation via a “is a” relation. ML, for example, is a type of AI, and ANNs are a type of ML. Dashed arrows show a different relation, “uses”. Everything in the large, dashed rectangle is within scope, both the types of AI (green) and their usage (yellow). Darker colours indicate the main scope, lighter ones categories that are also within scope, but which are not thoroughly addressed in the study.

The concept of GenAI predates deep learning, and simpler models with limited capabilities existed before. The complexity and realism of the content that modern GenAI can produce, however, are almost entirely attributable to the power of deep learning – specifically, a type of neural network architecture called the “Transformer”, presented in 2017 by Google researchers (Vaswani et al. 2017). Although thus not strictly a subset of DL and even residing in a different dimension (of model purpose, not of underlying technology), all current state-of-the-art GenAI models rely on deep learning as foundational technology – and in particular on Transformer models, as will be discussed in Section 4.2.1. GenAI is thus often described as a subset of DL – formally wrong; factually, currently true.

The focus of SAFE-AI is consequently limited to ML. The terms AI and ML are thus also used interchangeably and synonymously, meaning machine learning. While the currently prevailing ML technology are neural networks, others are also in use, and the methodological considerations developed throughout the study are agnostic towards the exact flavour of ML algorithms. Within ANNs, GenAI (and

consequently the DL models supporting it) results in by far the largest share of energy consumption and GHG emissions – and at least for the next decade this is not expected to change (Paccou and Wijnhoven 2024). While they might thus be deployed mainly for GenAI, the principles developed in this work can also be applied to much smaller neural network models – and even to any ML model more generally.

These terms and their relations are summarised in Figure 2, which also presents the scope of the study. All of ML is within scope, and the SAFE-AI framework is applicable to all ML, but the main focus lies on Gen AI.

1.3 Structure of the report

The remainder of this report is structured as follows: Chapter 2 presents the core principles of the SAFE-AI framework, including an archetypal view of an AI system and the three possible assessment levels: AI model, AI system, and AI usage.

Chapters 3 to 5 then present these three levels in detail: Chapter 3 discusses the impacts of AI models, presenting the four relevant assessment dimensions and their relationships: Environmental lifecycle, ML model lifecycle, device categories, and impact types.

Chapter 4 introduces several types of AI systems together with their specific individual ecosystems. It then shows the influence of various LLM architectures on the assessment, and discusses data visibility for different actors.

Chapter 5 discusses the allocation of system-level impacts to individual usages and conversely, the aggregation of usage-level impacts to AI systems. It also discusses possible functional units and their suitability for these perspective changes among assessment levels, and performs a stochastic analysis of per-query token distribution that is essential for aggregation.

Based on these insights, Chapter 6 suggests an assessment workflow for AI service providers. Rooted on AI system level, the workflow integrates the assessment of both internal ML models and externally employed models such as LLMs. Chapter 7 demonstrates the application of SAFE-AI in two case studies; the first one in particular follows the workflow from Chapter 6. Finally, Chapter 8 summarises the work, showing its limitations and possible venues for further research.

Table 1 presents the match between this report structure and the research questions it addresses.

Table 1: Correspondence between research questions and parts of the report addressing them.

RQ#	Research question	Addressed in
RQ1	*Identify most important sources of impact*: Which are the most important sources of energy consumption and greenhouse gas emissions?	Chapter 3
RQ2	*Propose robust systemic assessment approaches*: What are the main drivers for energy consumption of • AI models (and in particular large language models)? • the entire AI ecosystem? • Who has access to which data and what are the assessment implications of data visibility?	 Section 4.2 Section 4.1 Section 4.3
RQ3	*Examine possible functional unit definitions*: Which are the most suitable functional units (FUs) to break down the overall system effect to individual usages?	Sections 5.1, 5.2
RQ4	*Provide an AI system assessment workflow*: How can companies deploying AI systems assess their energy consumption and GHG emissions in a structured manner?	Sections 5.4, 5.5 Chapter 6

2 SAFE-AI framework: Core principles

This chapter presents the core principles of the SAFE-AI assessment framework. It starts by introducing in Section 2.1 a high level, archetypal AI system together with its possible generic components. These components can be either trained AI models (which can be both system-internal or system-external) or various types of non-AI tools and components.

Section 2.2 then addresses three possible levels of energy and environmental assessment: **Model level**, **System level**, and individual **Usage level**. Circling back to the generic system components introduced in Section 2.1, it then discusses on which of these three levels the assessment of each component usually resides, and draws a first sketch of the overall assessments.

2.1 Components of an AI system

The central concept in – and simultaneously the core product of – machine learning algorithms is the **ML model**. A model is first *trained* on what is called "training data" to recognise specific patterns and relationships within this data. It is then deployed to *infer* predictions or classifications on new, previously unseen data. Its tasks can range from very specific (such as recognising specific patterns e.g. in radiology) to extremely broad ones, such as producing texts or images on a wide range of topics as in generative AI.

Some machine learning models are deployed in a tightly confined environment and directly used by one or a few users. They are typically highly specific and rather small models, often not connected to the Internet, and deployed either on a user's device (such as a health analysis model embedded in a smartwatch) or in a restricted, internal environment such as a hospital's in-house data centre and used by the few internal radiologists.

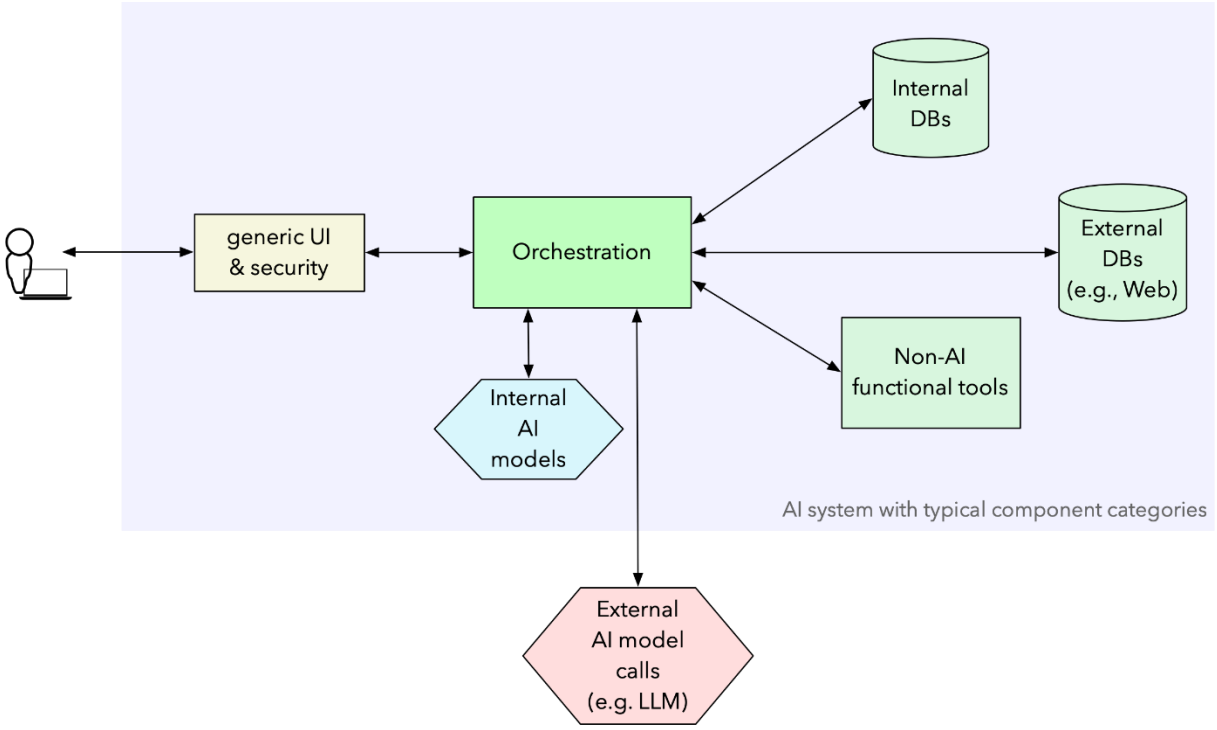


Figure 3: Archetypal view of an entire AI system. Next to the core orchestration component and at least one ML model, all other components are optional.

In general, however, ML models – and LLMs in particular – do not operate in isolation. They are embedded in a surrounding **AI system** that mediates how they interact with data, tools, users, and environments. This ecosystem orchestrates the information flow and the usage of the various AI and non-AI tools and algorithms to produce the desired service. It can be as simple as a single user issuing a prompt, or as elaborate as a production pipeline combining multiple models, databases, retrieval systems, and agentic frameworks. These components will be addressed in Section 4.1.

Figure 3 provides an archetypal overview of such an entire AI ecosystem. At least one AI model is required to qualify the entire system as an "AI system". This model can be either internal to the system (i.e., operated by the same service provider) or it can consist of application programming interface (API) calls to one or several external AI models. This latter paradigm is often deployed in conjunction with LLMs.

Other than this, all other types of components are optional: further (internal and/or external) ML models, internal and external databases (such as internal documents and web searches), non-AI functional tools and possibly user UI and security tools. Some of these components might also be shared with other systems (AI or non-AI) developed by the same service provider. These are often generic encryption/decryption services (such as SSL) or generic security such as firewalls, and are depicted in yellow in Figure 3 and only partly belonging to the AI system under investigation.

A core component of each system performs the orchestration of the entire functionality, coordinating all the information and computation flows. This core component is thus often also addressed as "orchestration layer" or "control layer". It will be more thoroughly addressed in Section 4.1.1.

2.2 Levels and principles of assessment, and their intricate relations

A first approach to assess the energy and environmental impacts of AI systems – and subsequent AI usages – is the assessment pipeline depicted in Figure 4. As argued in Section 2.1 above, the ML model is the core product of any AI effort. It thus seems natural to start there with the assessment.

One or several individual ML models contribute to a larger AI system, which typically also encompasses non-AI components. This AI system then provides AI services in terms of individual inferences. A logical assessment pipeline would thus follow these steps, by first assessing all contributing ML models, then the analysing the contribution of both ML and non-AI ecosystem components to the AI system, and the distributing the impact to individual AI services.

Figure 4: First approach to a high-level assessment pipeline of the SAFE-AI framework. One or several individual ML models contribute to a larger AI system, which typically encompasses further non-AI components. The impact of the entire system can then be allocated to individual AI usage instances. This seemingly straightforward pipeline, however, only works for AI models that have been both trained and are deployed internally – for most organisations and their AI systems, this is not the case.

While appealing, this approach, however, only works for a specific subset of AI systems: the ones that only rely on internal models that have ideally also been trained internally. For external use of AI models via API calls, there is not sufficient information to follow this pipeline.

As shown in the previous Figure 3, external models are often used by AI systems. In fact, as of late 2025, a large majority of AI systems deployed by organisations use external LLMs via APIs rather than self-hosted models (Typedef 2025). Among the general public, Web-based interaction with proprietary LLMs is even more the dominating use case. Section 4.1 will address these two usage paradigms, including the retrieval-augmented generation often deployed by companies and other organisations to leverage the usage of general-purpose LLMs to their own needs and internal documents.

As AI foundation model developers (such as OpenAI, Google, or Anthropic) rarely disclose data on the overall energy consumption of training and running their models, any external user of these models encounters epistemic uncertainty of the overall impact of these models. Another well-guarded business secret, inducing additional uncertainty, is the usage intensity of these models. Even if the energy used in training a model was known, the number and types of usages of the model would be needed to allow a meaningful amortisation to the various systems using it, and subsequently to individual usages.

It is consequently not feasible to follow the pipeline suggested in Figure 4, which would need to start with the assessment of the AI model(s). Section 4.2 explores in more detail the issues of data visibility and their influence of the types of assessment, depending also on who is performing the assessment.

An external organisation developing an AI service that performs API calls to external AI models such as LLMs thus needs to take a different route. Typically missing access to primary, model-specific data, it needs to resort to a different approach: Sometimes, the AI model developers disclose data on the energy and water consumption of individual usages; at least mean or median usages, that is. For assessing the impact of API calls to external AI models, the natural assessment based on 3rd-party data, is often the exact opposite from the one in Figure 4: It needs to start from individual usages, which can be aggregated for a system-wide impact.

Figure 5 thus provides a more thorough view of the overall SAFE-AI assessment framework. Its main feature is that it allows for various types of AI systems. As argued in the previous discussion, by reflecting this variety of systems which differ in their respective assessments, the framework cannot suggest a single assessment pipeline. It does, however, contribute with the following assessment principles:

- It depicts the three main assessment levels: *model*, *system*, and *usage*. The subsequent three chapters discuss each of these three levels of assessment in detail.
- The framework further shows the main dependencies among these levels, sketching how impacts are allocated from *system* to the *usage* level and vice versa aggregated from *usage* to *system* level.
- Depicting all types of components (as introduced in the archetypal view in Section 2.1), the framework also shows where the estimation of each such component typically starts, when the assessment is performed by an AI service developer / provider or a third-party researcher acting on the provider's behalf.

As can be seen in Figure 5, most assessments start either on *system* or *usage* level. One exception is the external training of models, which resides on the *model* level. The other exception are the shared (i.e., not AI-system-specific) tools and components, such as firewalls or encryption-decryption modules for Web-based services. These also reside on a different, system-wide level (albeit not the *AI system* level), which is not depicted in the figure. As the contribution of these components is usually negligible, this latter aspect is less relevant.

While Figure 5 sketches allocations and aggregations between *system* and *usage* level as simple divisions and multiplications with the number usages, respectively, this process is in fact much more complex and challenging. It also relates to the heterogeneity of AI usages: The more homogeneous a model’s inferences are (typically correlating with a specialised purpose of the model) and the narrower its users’ basis, the more straightforward allocation and aggregation of impacts will be. For complex AI systems, however, using repeated inferences from several ML models and with tasks of varying complexity, the allocation of impacts is all but trivial. For such complex systems, even defining the functional unit (FU), which is the basic, “atomic” unit to allocate impacts to, becomes a challenging task, as Section 5.1 discusses.

Section 4.3 will address in more detail data visibility for different actors, the usage of primary versus secondary data, as well as the concrete consequences for the assessment, depending on which actor performs it. Subsequently, Section 5.5 presents a stochastic analysis of token distribution among individual queries, which is essential for the aggregation from usage to system level, as depicted in Figure 5. Based on this analysis, Chapter 6 provides a generic assessment workflow on AI system level that includes said aggregation.

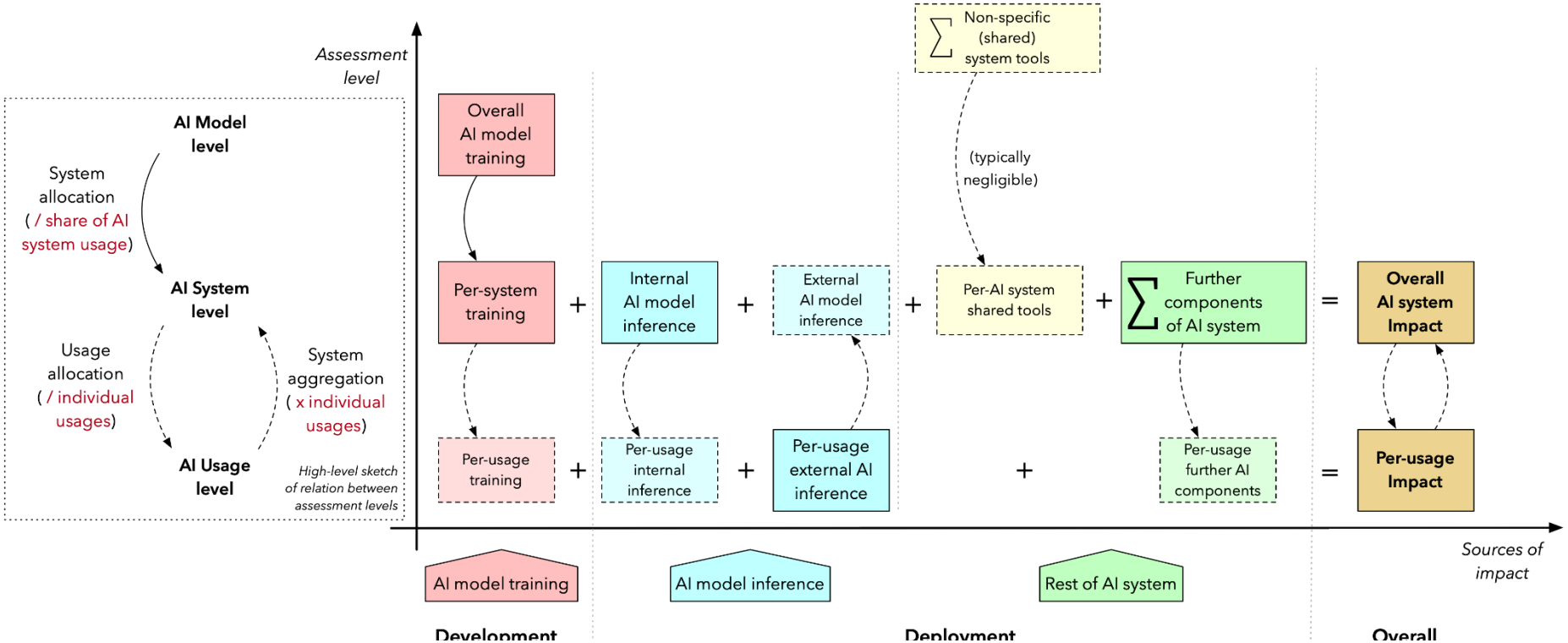


Figure 5: The three possible assessment levels of SAFE-AI, including assessment principles and a sketch of the main dependencies among the layers. Filled and more vividly coloured boxes show where usually an assessment starts; dashed boxes with less pronounced colour and smaller text are usually derived. Other than model training (which naturally starts on the *model* level), all other assessments start either on *system* or *usage* level. Allocating impacts from *system* to the *usage* level, and vice versa aggregating from the *usage* to the *system* level, is sketched as dividing and multiplying with the number usages, respectively. In reality, such allocations and aggregations are anything but trivial and depend on the heterogeneity of AI usages and the chosen functional units, as discussed in Chapter 5. Shared system tools are also not AI-system-specific, and their assessments starts on a higher system (but not *AI-system*) level; as their contribution is usually negligible, however, their assessment can typically be skipped. Allocation and aggregation can be performed both for each impact source as well as for their respective sums.

Although the framework does to some degree build on the assessment of individual AI models, this topic is only peripherally relevant. Instead, the focus of SAFE-AI lies on the assessment of the overall AI system as well as the individual AI usage level. The framework also presents principles of switching between these two layers, along with the related uncertainties and functional unit options and resulting variability.

Nonetheless, the AI model level is also important, both because the training of the model often represents a substantial share of the overall impact of an AI system, and to be able to better assess the quality of third-party assessments and developer-published data, when primary data are unavailable.

For these reasons, the next three chapters are dedicated to each of the assessment levels in turn: model, system, and usage.

3 AI model lifecycle assessment

The environmental impacts of AI are multifaceted: From the resource-intensive hardware manufacture to the rapidly growing energy consumption of data centres hosting most of the computation, the mounting problem of electronic waste (e-waste), and water consumption issues particularly in water-scarce regions, the impacts extend across several dimensions. This complexity does not only imply far-reaching consequences, but also inherent assessment challenges.

For a comprehensive assessment of the energy and environmental impacts of an ML model, this complexity needs to be addressed. Several dimensions thus need to be taken into account:

- the ML model lifecycle phases (the *why*?),
- the environmental lifecycle stages (the *when*?),
- the various ICT subsector categories that are sources of impact (the *where*?), and
- the types of environmental impact that are generated (the *what*?).

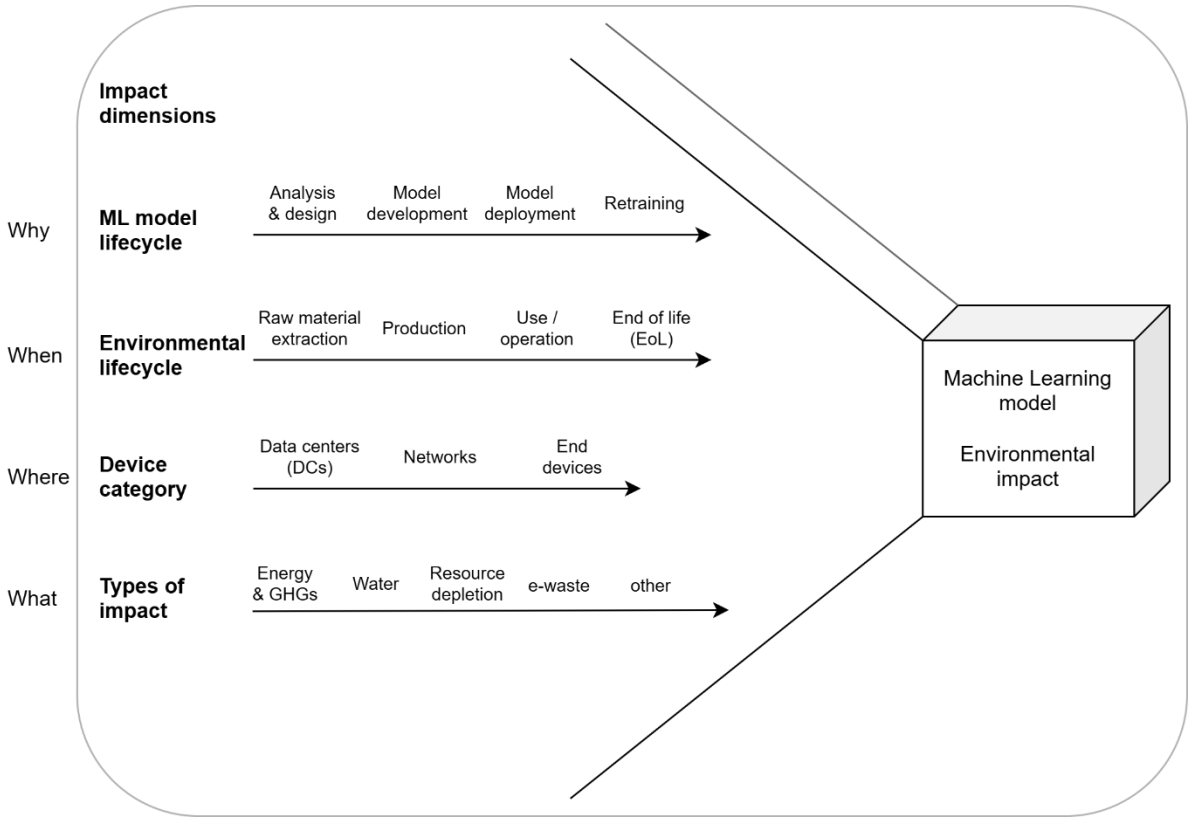


Figure 6: The four main dimensions relevant for a comprehensive analysis of the environmental impact of a machine learning model: i) the ML model's own lifecycle, ii) the environmental lifecycle as defined by ISO 14040 (ISO 2006a), iii) the category of devices which contribute to the environmental impact, and iv) the various types of environmental impacts. The four dimensions are orthogonal.

Figure 6 summarises these four dimensions. They are addressed throughout this chapter as follows: Sections 3.1 – 3.3 address the *why*, *when*, and *where*, respectively (i.e., the individual ML model lifecycle phases, the environmental lifecycle stages and in particular the relation between the environmental and the ML lifecycles as well as the three main device categories). Sections 3.4 and 3.5 then jointly address

the last of the four dimensions above, discussing energy and GHGs (Section 3.4) as well as water consumption (Section 3.5).

3.1 The *why*: Machine learning model lifecycle

There are various ways to describe the individual stages of a machine learning model pipeline, and no widely accepted taxonomy exists. While various proposals use different terminology, highlight different aspects and group stages and sub-stages differently, most of the literature agrees on a few main stages, which typically comprise:

0) *Idea*: Problem identification, inception.

1) *Analysis and design*: Data
    a. *collection* and
    b. *preparation* (for later model training) as well as
    c. *model selection*.

2) *Model development*, which consists mainly of *model training* but also comprises the c) *evaluation* of the trained result. For large models such as those deployed in generative AI, model training consists itself of two main stages: pre-training and fine-tuning. Correspondingly, the main development phases are:
    a. *Pre-training*, which typically deploys a very large neural network trained on an equally large dataset, resulting in a general-purpose model (Thomas and Avery 2023). The typically unsupervised pre-training (Devlin et al. 2019) is timewise, computationally, and energetically complex.
    b. *Fine-tuning*, which adapts a pre-trained model – by tuning all or a subset of its parameters (Howard and Ruder 2018) – to a specific task or domain, often with a smaller dataset (Thomas and Avery 2023). Fine-tuning has traditionally been accomplished via supervised learning (Devlin et al. 2019), but is increasingly shifting towards reinforcement learning (Roberts 2025).
    c. *Evaluation*, which quantifies the model's performance using unseen, typically labelled data, to ensure it generalises well from its training data. For comparability, evaluation often employs specific metrics such as accuracy, precision, or mean squared error as well as techniques such as cross-validation (Rainio et al. 2024).

3) *Model deployment*, which mainly consists of
    a. *inferencing* (or "*serving*") the model base on queries, but also comprises
    b. *model monitoring* and
    c. (*continuous*) *re-evaluation*, which might ultimately lead to

4) *Model retraining*, which becomes necessary, for example, in case of model drift.

Retirement, which occurs at the useful lifetime of the ML model.

3.1.1. Terminology used in the literature

Using heterogeneous terminology, highlighting the individual phases to various extents, and grouping them differently, several sources from industry and international bodies as well as standardisation institutions distinguish main phases of the ML lifecycle similar to those discussed above. This body of related work is presented below. For an overview, Table 2 compares the terminology used in these sources for the various stages of the ML model lifecycle. As can be seen from Table 2, the correspondence between the individual sources is far from one-to-one; often, several (sub)stages of one

Table 2: Comparison of terminologies used in the literature for the stages of the machine learning model lifecycle.

Stage	Substages	Google	ISO/IEC	ITU
0) Idea	-	-	Inception	Problem identification
1) Analysis and design	a. Data collection	Exploratory data analysis	Design and development	Data [acquisition]
	b. Data preparation	Data preparation and feature engineering		[Data] pre-processing
	c. Model selection			Model building
2) Model development	a. Pre-training	Model training and tuning		Model training
	b. Fine-tuning			
	c. Evaluation	Model review and governance	Verification and validation	Model evaluation
3) Model deployment	a. Inferencing	Model inference and serving	Deployment	Model deployment (inference)
	b. Model monitoring	Model monitoring	Operation and monitoring	Model monitoring & management
	c. Re-evaluation	-	Continuous validation	
4) Model retraining	-	Automated model retraining	-	-
5) Retirement	-	-	Retirement	-

Google
Google (Google Cloud 2025), for example, calls the individual phases as follows (in parentheses, the correspondence to numbers 0 – 5 from the enumeration above):

A. *exploratory data analysis* (1a)
B. *data prep and feature engineering* (1b, 1c)
C. *model training and tuning* (2a, 2b)
D. *model review and governance* (2c)
E. *model inference and serving* (3a)
F. *model monitoring* (3b)
G. *automated model retraining* (4)

These stages, which reflect earlier work from Google (Baylor et al. 2017), largely coincide with 4 of the 6 main stages listed above. As can be seen in the enumeration, (C) and (D) correspond to sub-phases of (2), (E) and (F) likewise to sub-phases of (3), and there is no correspondent to (6).

ISO/IEC
The 5338 AI standard jointly developed by the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) distinguishes the following main phases (ISO/IEC 2023):

A. *inception* (0)
B. *design and development* (1, 2a, 2b)
C. *verification and validation* (2c)
D. *deployment* (3a)

E. *operation and monitoring* (3b)
F. *continuous validation* (3c)
G. *re-evaluation* (3c)
H. *retirement* (5)

There is no explicit emphasis on data handling in the (ISO/IEC 2023) standard, nor is model training explicitly mentioned; however, both these main phases seem included in the *design and development* phase from (ISO/IEC 2023). This is confirmed by the literature; when discussing the (ISO/IEC 2023) standard, (Farzan and Kallio 2024) argue that "the design and development stage can include substages such as acquiring training data, data preparation, algorithm selection, and model training". These (sub-phases) are in line with (ITU 2024), discussed below.

The standard further has two phases – (F) *continuous validation* and (G) *re-evaluation* – which together correspond to phase 3c in the enumeration above. There is no explicit *retraining* phase; however, there is a loop from *re-evaluation* back to both *inception* and *design and development*. This loop, which would trigger a new *training*, thus implicitly refers to *retraining*.

ITU
By contrast, the ITU terminology places a strong emphasis on data acquisition and handling. Its main phases include (ITU 2024):

A. *problem identification* (0)
B. *data [acquisition]* (1a)
C. *[data] pre-processing* (1b)
D. *model building* (1c)
E. *model training* (2a, 2b)
F. *model evaluation* (2c)
G. *model deployment (inference)* (3a)
H. *model monitoring & management* (3b, 3c)

There is a loop from *model evaluation* back to *data [acquisition].*, which conceptually corresponds to stage (4), *model retraining*. The name used for stage (G) *model deployment (inference)*, shows that the two concepts are considered synonymous, as opposed to other literature, which treats inference as a sub-category of deployment (albeit indeed the most important one).

3.1.2. Suggested taxonomy, with relevance for the environmental impact

From the 6 stages introduced in the beginning of Section 3.1, (0) *idea* – called *problem identification* (ITU 2024) or *inception* (ISO/IEC 2023) – describes the initial idea and is thus rather conceptual in nature. From the perspective of an environmental impact analysis, it thus plays a negligible role. One might exist, of course, for example in the form of office energy or preliminary computations, but these are likely marginal and – to the extent they exist – can be attributed to phase (1) *research, analysis, and design*. Industry-based frameworks such as (Baylor et al. 2017; Google Cloud 2025), which are perhaps more operational and less theoretical, also tend to ignore this stage.

The sometimes distinguished last stage (5) *retirement* (ISO/IEC 2023) phase, meanwhile, can be understood either conceptually as well (i.e., to mark the moment that an ML model has been retired), and thus similarly to (0) *idea* without any notable environmental impact. It can also, however, describe the *end-of-life* (EoL) phase of the environmental lifecycle.

This second interpretation would rely on the (explicit or implicit) assumption that all the other impacts are specific to the operation (and perhaps production) phases. A white paper from Nokia research (Farzan and Kallio 2024), which discusses the (ISO/IEC 2023) ML model lifecycle and how it relates to the lifecycle assessment (LCA) methodology as described by the ISO standard 14040 (ISO 2006a), could imply this interpretation. Section 3.2 below discusses the relation between ML model and environmental lifecycle, and argues that the two are orthogonal. As a consequence, such interpretation would be wrong

and mixing the EoL phase of the environmental lifecycle into the ML model lifecycle semantically not meaningful.

Stage (5) *retirement* is thus either a conceptual abstraction with little to no environmental impact attached to it, or denotes the environmental EoL phase and does not belong to the ML model lifecycle. In either case, together with stage (0) *idea*, it can be ignored from a taxonomy that focuses on the ML model lifecycle stages relevant for the model's environmental impact.

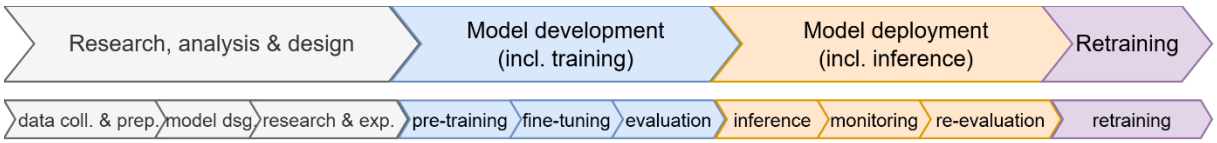


Figure 7: Suggested taxonomy of the ML model pipeline, which harmonises several classifications from the literature. The taxonomy focuses on the stages relevant to the environmental impact. Lengths of arrows do not correlate with the relative importance of their environmental impact.

The resulting model, as shown in Figure 7, has 4 main lifecycle stages, most of which with sub-stages:

1. *Research, analysis, and design*, which includes the data preparation and choice of model, comprising
    a. *data collection and preparation*
    b. *model design*
    c. *research and experimentation*
2. *Model development*, which consists mainly of the two training sub-stages as well as evaluating the model
    a. *model pre-training*
    b. *model fine-tuning*
    c. *model evaluation*
3. *Model deployment*, which mainly consists of inference, but also continuous monitoring and re-evaluation of the model's performance
    a. *model inference*
    b. *model monitoring*
    c. *model re-evaluation*
4. *Retraining*, which typically takes place either when new data becomes available or when the model is no longer precise enough (i.e., when "model drift" occurs).

While initial *training* builds a functional model from scratch, *retraining* is about updating an existing model based on real-world performance (and possibly occurring performance drift) or when new data becomes available.[1] Newer paradigms such as online learning thus retrain on the fly during model deployment.[2] This phase could thus be considered a sub-phase of (3) *model deployment*. However, as the *monitoring*, *re-evaluation* and *retraining* are often represented as a feedback loop that goes back to the (2) *model development (incl. training)* phase, it can also be seen as part of it.[3] Or it can be conceived as an entirely separate phase of the ML model lifecycle.[4] Our suggested taxonomy chooses this last option, both for clarity but also because it is more generic and can be understood as either of the other options.

[1] See https://www.evidentlyai.com/blog/retrain-or-not-retrain.
[2] See https://www.striim.com/blog/machine-learning-streaming-data/.
[3] See https://randomtrees.medium.com/mastering-model-retraining-in-mlops-4bb961ee7070.
[4] See https://www.seldon.io/machine-learning-model-inference-vs-machine-learning-training.

3.2 *Why* vs. *when*: Relation between the environmental and ML model lifecycles

A white paper from Nokia research (Farzan and Kallio 2024) discusses the ML model lifecycle as put forward by (ISO/IEC 2023) – see description in Section 3.1.1 above – and how it relates to the environmental lifecycle as defined in the life cycle assessment (LCA) methodology described by the ISO standard 14040 (ISO 2006a). The two distinguish 8 and 4 main stages, respectively. Stages A – H of (ISO/IEC 2023) were presented in Section 3.1.1. Meanwhile, the environmental lifecycle as standardized by (ISO 2006a) has the four main stages. The environmental lifecycle assessment is presented in more detail in Appendix A:

1. raw material extraction
2. production
3. use / operation
4. end-of-life (EoL)

Putting these two lifecycles side by side, (Farzan and Kallio 2024) represent them in parallel, mapping ML lifecycle stages to the environmental lifecycle stages as shown in Table 3.

Table 3: Mapping of lifecycle stages between the environmental lifecycle as standardized by ISO 14040 (ISO 2006a) and the ML lifecycle as defined in another ISO standard, 5338 (ISO/IEC 2023), as performed by (Farzan and Kallio 2024).

ISO 14040 lifecycle stage	**ISO 5338 ML model lifecycle stage**
Raw material acquisition (1)	– [none]
Production (2)	Inception (A) Design and development (B) Verification and validation (C)
Use (3)	Deployment (D) Operation and monitoring (E) Continuous validation (F) Re-evaluation (G)
End of life (4)	Retirement (H)

This mapping stands to reason. After all, designing and training an ML model obviously equates with producing it, while the deployment and inference phase mark its usage. The mapping seems so obvious, in fact, that it almost bears motivation: “Inference can be easily mapped to the LCA use stage, as well as re-training during use,” as (Farzan and Kallio 2024) argue.

Obviously, however, this does not mean that for the production phase of the ML model only the production of devices such as graphics processing units (GPUs) is relevant – or that similarly, for the use phase of an ML model only the operational electricity used by such devices matters.

On the contrary, the impacts of ML model training, for example, are not only – and perhaps not mainly – production impacts of the devices used for training. The training also requires substantial amounts of (operational) electricity. And the devices deployed for training will at some future point require EoL treatment, so a part of that impact needs to be allocated to any model trained on them as well.

Likewise, model inference belongs to the use phase of an ML model. And while the hardware devices hosting the inference do require operational electricity, those chips, servers, cooling systems, and encompassing data centres were also produced and will at some point be scrapped. Part of the production and EoL processes need to be allocated to the inference.

These considerations will come to no surprise to an LCA practitioner. In any LCA of a product or service, each source of impact in any of the lifecycle phases is a subprocess, which typically has its own material

extraction, production, use, and EoL phases. And all lifecycle phases of these subprocesses contribute to the overall impact of the higher-level product or service.

For all practical purposes then, **the two lifecycles – of the ML model and the environmental LCA – are orthogonal**. To compute the impact of each ML model lifecycle phase, the raw material extraction, production, operational, and EoL impacts of all devices involved in that phase (such as model training or inference) need to be computed and allocated.

This is why a linear mapping, such as the one in (Farzan and Kallio 2024), might be misleading for non-LCA specialists. Seeing similar names and linear mapping, one might wrongly assume that the operational impact of an ML model only consists of the use-phase electricity involved in inference, and ignore the other environmental lifecycle stages of the devices involved. An explicit independent / orthogonal representation of these two dimensions – as done in Figure 6 – can be helpful in avoiding such misinterpretation.

3.3 *Where*: The ICT subsector categories data centres, networks, and end devices

The literature on the environmental impact of digital services (Baliga et al. 2009, 2011; Coroamă and Hilty 2014; Coroama et al. 2015; Schien et al. 2015; Kamiya 2020; Coroamă 2021) typically distinguishes three main categories of devices that contribute to a service's environmental impact: data center devices, network devices, and end devices. Figure 8 presents a corresponding high-level topology of devices involved in a digital service in general, and AI service in particular.

Meanwhile, existing assessments of the environmental impact of AI tend to focus exclusively on the impact that occurs in data centers. While DCs are indeed the main source of impact for most categories of AI impacts, and in many cases networks and end devices can be ignored, sometimes the contribution of the other categories is not negligible. Section 3.4 on energy and GHGs addresses the other device types as well.

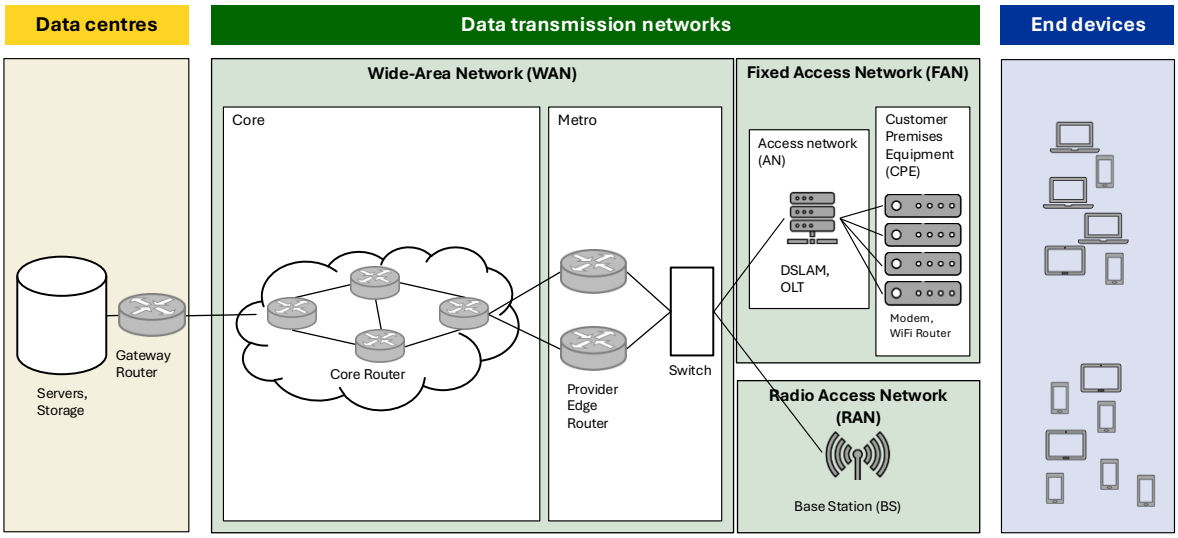


Figure 8: High-level topology of devices involved in an AI service, showing the three main categories of devices: data centers devices, networking devices, and end devices. Modified from (Coroamă 2021); reprinted with permission.

3.4 Energy consumption and GHG emissions of AI models

After theoretically addressing the impact sources (the *why*, *when*, and *where*), the current section discusses one of the most important *what*'s: the energy consumption and related GHG emissions of AI models across the three dimensions above. Similarly, Section 3.5 will address the water consumption.

3.4.1. Operational energy in data centres

Early studies on the environmental footprint of AI focused on the energy and carbon impacts associated with *training* the AI models (Lacoste et al. 2019; Strubell et al. 2019, 2020; Schwartz et al. 2020). A more recent review (Verdecchia et al. 2023) also shows that the *training* phase has been the focus of early research on the energy and GHG impact of AI. This focus stood to reason while large deep neural networks were intensively trained to then be deployed relatively seldomly, as is the case of a Go-playing ML model such as AlphaGo.

With the advent of Gen AI, however, this balance changed. LLMs still require large amounts of energy to be trained. They are, however, also inferred billions of times daily (Lee 2025), so the *inference* phase requires substantial amounts of energy itself; over the entire lifetime of a model, perhaps considerably more than training.

Three years ago, for example, it was estimated that the inference of ChatGPT required 564 MWh of energy per day (de Vries 2023); a little more than two days of inference thus already outweighing the estimated training costs of 1,287 MWh of the same model (Luccioni et al. 2022). Data from Meta and Google confirm this changed balance, indicating that in the meanwhile the training phase only accounts for 20-40% ML-related energy consumption in their DCs, while inference accounts for 60-80% (Wu et al. 2022; Patterson et al. 2022).

Building on these and further sources, (EPRI 2024) also estimates a share of 10% for a model's pre-training development, 30% for its training, and 60% for its inference. However, using a top-down quantitative system dynamics method to estimate the yearly energy footprint of all AI worldwide, (Paccou and Wijnhoven 2024) arrive at the opposite result: for Gen AI, they estimate a 2025 energy consumption of 47 TWh for training and only 15 TWh for inference, i.e., a 76% – 24% split in favour of training. For traditional, industrial AI, the authors see a balance between the two: 20 TWh/year for training and 18 TWh/year for inference.

Both training and inference – and, more generally, the development and deployment phases – are important and must be accounted for.

3.4.2. Further environmental lifecycle phases

As most of the literature, the considerations above focus exclusively on AI's impact in DCs during the operational phase. Relatively few studies look beyond this point in either of the two dimensions (i.e., either further environmental lifecycle phases or other categories of devices). This is likely due to poor data availability (Masanet et al. 2024); there are, for example, no publicly available LCA data of GPUs (Luccioni et al. 2022), the main devices deployed for AI training and inference.

Nevertheless, the few studies that do consider the production of microelectronics show they are responsible for a relatively small amount of GHGs as compared to their later operation in DCs:

For the open-source LLM BLOOM, (Luccioni et al. 2022) compared the emissions embodied in both GPUs and servers to the operational impact of model training and inference. No production data was available for the servers deployed in BLOOM training and inference, which were approximated via the similar HPE ProLiant DL345 Gen10 Plus, which is responsible for about 2.5 t CO_2eq. (HPE 2021).

For the Nvidia A100 GPUs, however, no data was available at the time of publication, not even for similar GPUs. The paper thus used an estimate of 150 kg CO_2eq. embodied GHG impact per GPU from the literature (Davy 2021). This number correlates surprisingly well with data published three years later by Nvidia on the server platform HGX H100, which yielded 1,312 kg CO_2eq. embodied emissions (Nvidia 2025). Given that HGX H100 encompasses eight H100 GPUs (the GPU generation following the A100)

plus some additional hardware, the 150 kg CO_2eq. embodied GHG impact per GPU seem remarkably accurate.

Deploying further assumptions, such as a refreshment cycle of 6 years and an average utilisation rate of 85%, both of which are in line with other studies (Coroamă et al. 2025; Schmid et al. 2025), (Luccioni et al. 2022) estimate 18% embodied emissions versus 82% operational emissions as follows:

- 11.2 tonnes of embodied CO_2eq. emissions (of which 7.57 embodied in the servers and 3.64 in the GPUs), as compared to
- 45.30 tonnes CO_2eq. emissions due to training operational electricity (of which 24.69 tons CO2eq. for useful work (i.e., the training itself), 14.6 tonnes due to the idle power consumption, and the remaining 6.01 tonnes due to DC infrastructure), and
- an average 19 kg CO_2eq. per day for inference. Assuming one year of inference, we compute 6.94 tonnes CO_2eq. for inference.

In another study, (Falk et al. 2025) recently presented a detailed LCA of Nvidia's A100 GPU. This third-party estimate is based on disassembling the hardware and performing elemental analysis. Assuming a three-year lifespan and activities similar to the training of the GPT-4 LLM along this lifespan yields 96.8% GHG contribution for the use phase and only 3.2% for production

Although the operational phase dominates the overall footprint in these studies, they are both nevertheless likely to understate it. The open-source BLOOM model analysed by (Luccioni et al. 2022) is used much less than its commercial peers; the share of inference energy (and thus of the entire operational energy) is atypically low. Additionally, both studies used for operation the low-carbon French electricity emitting 57 g CO_2eq / kWh. For more carbon-intensive electricity (such as in the US and China, where most data centres are located), the use phase is thus bound to dominate even more.

On the other hand, the results do not include the production of the DC infrastructure such as the cooling systems or the uninterruptible power supply, which would by contrast raise to some extent the share of the production phase. Overall, though, it seems reasonable to assume that for a mature commercial foundation model, the production phase does not account for more than 3-5% of lifecycle GHGs at most.

The operation phase dominates the environmental lifecycle GHG emissions of AI computation, while the production of microelectronics amounts to a maximum of a few percentage points.

3.4.3. GHG impacts and the lack of robustness of market-based accounting

When comparing production and operation as in Section 3.4.2 above, GHG accounting is necessary since emissions are usually a more meaningful metric for the production phase than energy consumption. When an assessment (whether individual or comparative) only considers AI's use phase, however, energy is the much more relevant and informative indicator, since it can lead to fewer misunderstandings and is harder to manipulate.

Since 2015, the GHG Protocol (Sotos 2015) introduced the "market-based accounting method" for Scope 2 reporting. By contrast to the traditional, "location-based method" (which uses the grid average GHG intensity), market-based accounting acknowledges the usage of renewable electricity by companies, either produced on-site, or purchased via power purchase agreements (PPAs) or renewable energy certificates.

Using market-based accounting for the use-phase electricity can yield substantially lower operational GHG results, since many of the largest data centre operators are also the largest buyers of renewable energy. If a comparative assessment deploys different assumptions for the respective carbon intensities, it can easily lead to misconceptions. This can also happen when comparing lifecycle phases. A 2022 assessment by Meta using market-based accounting, for example, concluded that the embodied emissions (which are much harder to decarbonise) were dominating the lifecycle emissions of Meta's AI (Wu et al. 2022).

This conclusion, however, stands in stark contrast to the analysis in Section 3.4.2 above. Even with very low-carbon (location-based) French electricity, the operations phase was dominating lifecycle emissions

in both (Luccioni et al. 2022) and (Falk et al. 2025). (Falk et al. 2025). However, using market-based accounting resulted in the operation phase being dwarfed by embodied emissions to the brink of disappearance.

Consequently, market-based accounting is under increasing scrutiny from both the media and the public for the greenwashing it enables (O'Brien 2024). Additionally, it has also been criticised for the double-counting of low-carbon electricity. For example, certificates issued for renewable electricity used by a buyer for market-based accounting might also be counted towards the location-based grid mix of its origin (Holzapfel et al. 2023).

We concur with the essence of this criticism, and do not take market-based accounting into consideration here.

3.4.4. Further device categories

In a life cycle assessment of GenAI, (Berthelot et al. 2024, 2025) compare not only all environmental lifecycle stages of devices, but also all three devices categories introduced in Section 3.3: data centres, networks, and end devices. The case study analysed the impacts of "Stable Diffusion", an open-source text-to-image Gen AI model.

For both energy and GHGs, data centres are clearly the main impact sources: for energy, DCs account for 37.6% (training) and 48.5% (inference) of the total impact, and for GHGs, for 46.3% (training) and 38.2% (inference). The contribution of end terminals, however, is not negligible, representing 13.5% of the total impact for energy (9.1% manufacture + 4.4% use) and 15.2% of the impact for GHGs (14.6% manufacture + 0.6% use). With only 0.4% of the energy and 0.3% of the GHGs, the contribution of networks, on the other hand, is negligible. This comes to little surprise as AI is computationally intensive, but the data transmitted is relatively modest (with the possible exception of federated learning, which is still a niche application though).

Similar results, but more skewed towards model training and inference, were presented by Mistral AI, a French company which recently published an LCA of its AI model (Mistral AI 2025). This LCA defines in two dimensions even larger system boundaries:

- along the ML model lifecycle, it also considers the "model conception" stage, which corresponds to the phase (1) *research, analysis and design* in Figure 7.
- in the environmental lifecycle dimension, it also takes into account the data centre construction.

Both additions, however, are not material, as they contribute with less than 1% each to the total impact. Similarly, the contribution of network traffic is also negligible. Model training and inference (presented jointly) dominate with 96.5% of the overall GHG impact (85.5% during the use phase, 11% embodied), while the end-user equipment only accounts for 3% of the GHG impact (production and operation jointly). Unfortunately, the full LCA has not been released yet, but only its summary (Mistral AI 2025), so the assumptions and calculations cannot be double-checked at the moment.

The energy consumption of transmission networks is entirely negligible, while end devices can account for a few percentages of the overall service impact.

3.4.5. The relative contribution of individual lifecycle phases (ML and environmental) and device categories

Section 3.4.1 discussed several estimates of the DC operational energy consumption of ML, and in particular of LLMs. Some of them saw the inference being more substantial, others the training phase. A possible explanation for these seeming discrepancies might be found in a recent analysis by the Epoch AI research institute.

Analysing OpenAI's 2024 compute expenses, the analysis estimates the ratio between the Chat GPT-4.5 training and its inference was indeed about 1 to 5 – i.e., 400 million USD for training versus 2 billion USD for inference (You 2025b). It appears, however, that the much larger share of costs was spent on R&D compute which did not make it into any final product. According to the Epoch AI analysis, no less

than 4.5 billion USD were spent by OpenAI on “basic research, experimental and derisking runs for final training runs as well as unreleased models” (You 2025b).

This implies a share of about 65% for research (both basic research unrelated to a specific later production model and the model itself, but not its final training run), only 6% for its final training, and 29% for the subsequent inferences. There are not many direct data (i.e., from the model providers themselves) discussing this overhead; for Meta Llama 3, however, 419 failures over 54 pre-training runs have been reported (Grattafiori et al. 2024).

If this estimate is essentially correct, this could explain the discrepancies. Several of the components categorised here under “research” could be attributed to “development/training” as well, the border between the two being somehow blurry. Beyond explaining the discrepancies, however, this estimate raises a further and very important point: **There is an overhead to ML development (such as basic research or the development of intermediate, never released models) that can be easily overlooked, particularly by bottom-up energy consumption models**.

In related fields, it has already been established that bottom-up models tend to have an overly narrow scope, easily overseeing components and thus resulting in understatements, such as when assessing the energy consumption of networks (Coroamă 2021). Given the potentially dominating contribution of this overhead (You 2025b), it is thus crucial to not miss it and meaningfully allocated it to ML models.

Table 4 presents an overview of the literature as discussed since Section 3.4.1. Its upper five rows show the estimates for the operational consumption in DCs, and the discussed dilemma of the relative contributions of the individual ML model lifecycle phases: research, training and inference. The lower four rows expand the discussion in the other two dimensions by considering also other environmental lifecycle phases (i.e., production) and/or device categories (i.e., end devices or networks).

Table 4: Overview of analyses quantifying the energy or GHG impact of ML along device categories (first distinction criterion), environmental lifecycle (second criterion), and ML lifecycle (third one): (Wu et al. 2022), (Paccou and Wijnhoven 2024), (EPRI 2024), (You 2025b), (Luccioni et al. 2022), (Berthelot et al. 2024, 2025), (Falk et al. 2025), (Mistral AI 2025).

Source	Data centres					Networks	End devices	
	Production		Operation			Operation	Production	Operation
	Servers	Processors	Research	Training	Inference			
(Wu et al. 2022)				35%	65%			
(Paccou and Wijnhoven 2024)				76%	24%			
(Paccou and Wijnhoven 2024)				53%	47%			
(EPRI 2024)				40%	60%			
(You 2025)			65%	6%	29%			
(Luccioni et al. 2022)	12%	6%		71%	11%			
(Berthelot et al. 2024, 2025)				37.6%	48.5%	0.4%	9.1%	4.4%
(Falk et al. 2025)		3.2%		96.8%				
(Mistral AI 2025)		11%		85.5%			3%	

Although the estimates (and thus also the data points) are relatively few, a couple of insights can nevertheless be gained:

- The production phase of data centre processors and servers is not large, but not always negligible. For the three cases assessed, it accounted for about 18% (Luccioni et al. 2022), 11% (Mistral AI 2025), and 3% (Falk et al. 2025), respectively. These studies, however, were performed on ML models trained and run in France, with its low-carbon electricity, thus resulting in a higher relative impact from production than with a global average grid. For a dirtier grid and/or large-scale LLMs with their heavy usage, the operation phase is probably even more important and the share of device production correspondingly diminished. **The 97%-3% split from** (Falk

et al. 2025) **is probably closer to reality for large-scale LLMs, in which case the production can be neglected after all**.

- The impact from networks is negligible and can be ignored. Although only one data point in Table 4 shows this – the 0.4% from (Berthelot et al. 2024, 2025) – we are very confident affirming this. The network consumption data in (Berthelot et al. 2024, 2025) relies on a 2021 assessment (Bordage et al. 2021), which puts forward the following energy intensities of the Internet: 30.7 Wh/GB for fixed networks and 96 Wh/GB for mobile networks. Other research shows, however, that these numbers were already outdated and overstated by about a factor of 5 in 2021; (Coroamă 2021) computes 7 Wh/GB and 20 Wh/GB for the two, respectively. **In short, AI is a computing-intensive process, but not a data transfer-intensive one**.
- End devices are probably not negligible. The two sources considering them, evaluate their contribution to 3% (Mistral AI 2025) and to 13.5% (Berthelot et al. 2024, 2025).

In conclusion, most of the energy and GHG impacts take place in data centres during operation (typically 80-95%), with the remainder from the production of data centre servers and processors as well as both production and operation of end devices). Along the ML lifecycle, training and inference are both substantial, while the research and experimental phase could also be highly relevant and need to be attributed to the finished AI models. Their contribution, however, is poorly understood.

3.5 Water impacts of AI models

Assessments of the water impact of AI are more recent than assessments of its energy consumption, with fewer studies and less developed and consistent methodologies. There is also a lack of homogeneous system boundaries and, consequently, a degree of terminological confusion surrounding some of the assessments. Beyond discussing the water impact of AI, this section thus also aims to bring some conceptual and terminological clarity.

3.5.1.The *why* and the *where*: Water along the ML model lifecycle and ICT subsectors

Among the orthogonal dimensions determining the environmental impact of AI (as introduced in the beginning of the chapter and discussed in Figure 6), the most important for water consumption is the environmental lifecycle dimension (the *when*), which will be discussed in Section 3.5.3 below.

As for the other two dimensions, the device categories are less relevant, because during the use phase, substantial water consumption can occur in data centres, but none in networks and end devices. During device production and EoL, water consumption occurs for all device categories, but DCs have nevertheless an outstanding importance as sources of water consumption.

The ML model lifecycle is relevant to the causality of water consumption. As model training and inference typically take place in the same data centres, however, the distinction is not operationally important. Additionally, in any DC, the water consumption is linearly proportional to the energy consumption of individual tasks. Hence, if the shares of individual ML model lifecycle stages in the energy consumption are known, the responsibilities for water consumption are also clear.

3.5.2. Measuring the *what*: Water consumption and water withdrawal as main indicators

Two metrics for water impact are often distinguished: *water consumption* and *water withdrawal*. Consumption is thereby a subset of withdrawal. The latter represents the total amount of freshwater withdrawn from surface and underground sources, while consumption is the share of withdrawal that is not returned to the original source.

Consumed water includes, for example, the water evaporated, transpired, incorporated into products, or otherwise lost from the system: “Withdrawals are the total amount of water withdrawn from sources including surface water and groundwater. Consumption represents the portion of withdrawals not returned to the original water source after use but lost, e.g. through evaporation” (IEA 2025).

Making the same distinction, (Li et al. 2025) indicate the main aims of the two indicators:

- "As water is a finite shared resource, water withdrawal indicates the level of competition as well as dependence on water resources among different sectors."
- "Water consumption reflects the impact on downstream water availability and is crucial for assessing watershed-level scarcity."

These concepts stem from the general water footprint literature (i.e., not AI-specific). In this established field, "water footprint", when not otherwise specified, refers to the narrower concept of water consumption, and not to withdrawal (Li et al. 2025). In the more recent literature on AI water impact – and more so in media reports – the same rigour is not always applied, and the two are often thrown together.

3.5.3. The *when*: Water impact along the environmental lifecycle

For the water impact of DCs, several sources (Lei and Masanet 2022; Shehabi et al. 2024; IEA 2025; Li et al. 2025) differentiate between *direct water impact* and *indirect water impact*:

- Direct water impact, also called *on-site* impact, is the water directly used in the DC – mainly for cooling, but also humidification, blowdown in cooling towers, and other minor sources (Coroamă and Dumbravă 2026).
- The indirect water impact occurs *offsite* (or *upstream*) for electricity generation (Lei and Masanet 2022; Shehabi et al. 2024) or for both electricity generation and manufacturing of hardware such as servers (Li et al. 2025) or semiconductors and microchips (IEA 2025).

The direct impact is thus a use-phase impact, equivalent in nature to the "scope 1" GHG impact as defined in the Greenhouse Gas Protocol of the World Resources Institute (WRI) and World Business Council for Sustainable Development (WBCSD) (Sotos 2015). Due to this equivalence, this has also been named "scope 1 water usage" (Li et al. 2025).

Indirect impacts could be distinguished between "scope 2" (i.e., electricity generation) and "scope 3" (i.e., for hardware manufacturing), as (Li et al. 2025) do. From the environmental lifecycle perspective, however, the important feature is that they occur upstream, during the *production* phase of either electricity or hardware components deployed in AI. For hardware devices specifically, some water consumption likely takes place during *raw material extraction* as well.

3.5.4. Direct and indirect, consumption and withdrawal

Based on the considerations above, Figure 9 brings these two dimensions (*when* and indicators for *what*) together. One dimension distinguishes between the two indicators: water consumption (i.e., permanent withdrawal) and total water withdrawal (permanent and temporary together). The other dimension distinguishes the individual environmental lifecycle stages where the water impact occurs. While the proportions in Figure 9 are not exact, they aim to qualitatively convey a sense for their relative sizes, as discussed below.

Direct consumption and withdrawal

The use-phase water impacts (both consumption and withdrawal) occur exclusively in data centres. Water consumption occurs due to the cooling of DCs. By contrast, end devices obviously do not require any water during usage. And while the large servers in Internet exchange points are likely to induce some water consumption, it is negligible compared to that of DCs.

Various DC cooling technologies induce different amounts of water consumption. The most "thirsty" technologies are *cooling towers* (both closed and open circuit ones) as well as the *adiabatic* (or *evaporative*) support, as described in detail by (Coroamă and Dumbravă 2026). While some cooling techniques – in particular dry coolers, air-cooled chillers, and airside economisers – do not induce any water consumption by themselves, they often require adiabatic support, thus consuming water as well (Coroamă and Dumbravă 2026).

Temporary withdrawal for DCs occurs when water is removed from a surface or underground source, used briefly as a medium for heat transfer, and then returned to the same body of water. Irrespective of the source, the process is similar: river/lake/sea water or groundwater is extracted, passes a heat exchanger in which it receives heat from the (closed-loop) cooling circuit of the data centre, and is discharged back to the same stream, body of water or aquifer. Aside from minimal losses, the volume is returned entirely, but warmer.

While water withdrawal is important in power generation (for the once-through cooling in electricity generation, see below), for DC cooling, it is a fairly marginal phenomenon. The International Energy Agency, for example, estimates that in 2023, the direct DC water consumption due to AI was 150 billion litres, and the entire withdrawal (including consumption) 200 billion litres (IEA 2025). The numbers estimated for 2030 are 370 billion litres (consumption) and 500 billion litres (all withdrawal). As qualitatively indicated in Figure 9, direct withdrawal is only marginally higher than consumption.

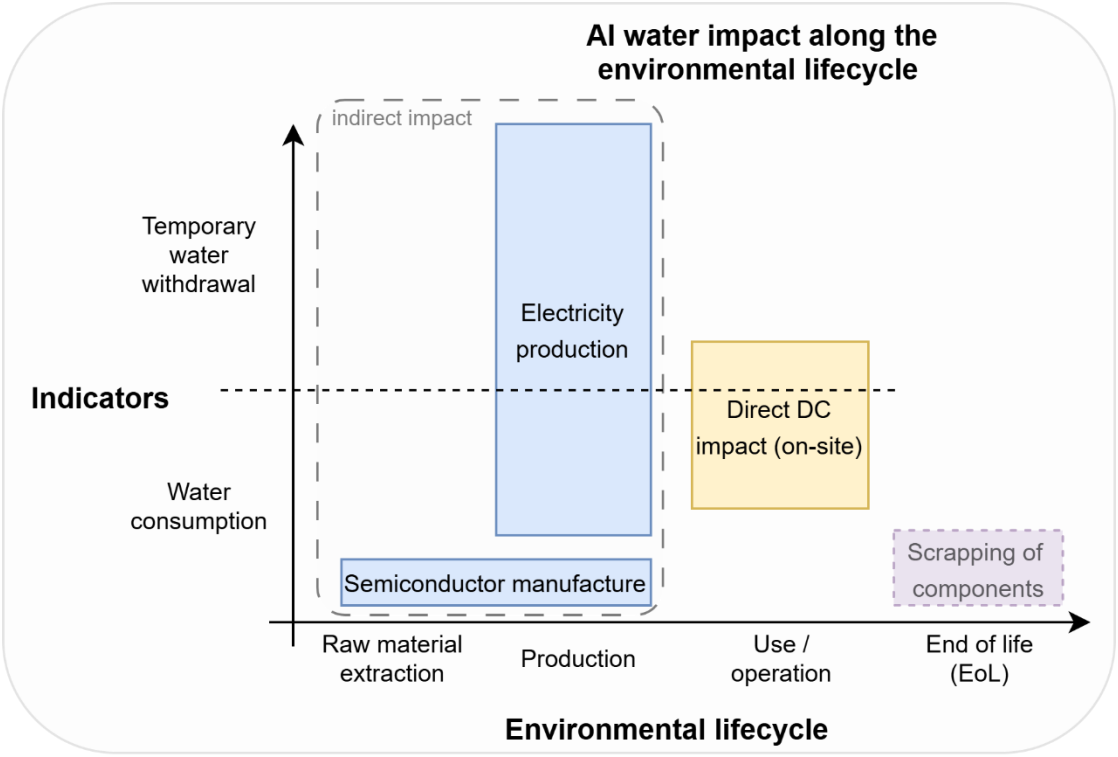


Figure 9: An indicative taxonomy of AI water impact types distinguishing two indicators (water consumption and water withdrawal) along the environmental lifecycle of AI. Direct water consumption occurs mainly for data center cooling (due to cooling towers or adiabatic support); temporary withdrawal is small. For electricity production, the consumption occurs mainly due to evaporation in the reservoirs of hydroelectric power plants, and temporary withdrawal mainly due to once-through cooling of thermal power plants. Semiconductor manufacturing induces some water consumption during both raw material extraction and the production process, but likely smaller than both on-site and for electricity production. The water impact during the EoL phase is not directly discussed in the AI water footprint literature, but likely to exist, so it is included for completeness.

Indirect consumption and withdrawal

This is quite different for electricity production, as will be discussed shortly. Depending on the generating technology, electricity production can induce both water consumption and/or temporary withdrawal.

Temporary withdrawal for electricity generation occurs mainly due to:

- The once-through cooling of thermal power plants. Whether coal, oil, gas, or nuclear steam plants, they all use a condenser to convert the steam used to drive turbines back to water. These condensers are typically cooled with water from nearby rivers or seas.
- Run-of-river hydro power plants work on temporary water withdrawal, by diverting a substantial part of the river's water and return it downstream.
- Modern geothermal power plants return most geofluid to the underground reservoir after having extracted heat from it.

Electricity generation, however, can also induce water consumption, for example by:

- Reservoir evaporation for large hydroelectric power plants, a consumption which can be substantial.
- Panel washing for photovoltaic power production or concentrating solar power (CSP).
- Older-generation geothermal power plants have steam losses.
- In all the other technologies, there are small losses due to leakage, maintenance or cleaning; these, however, are usually negligible.

As for hardware production and in particular semiconductor manufacturing, they also induce some water consumption. According to the IEA, however, this is fairly small compared to the other sources, as also optically indicated in Figure 9: For 2023, an estimated 30 billion litres were used in semiconductor manufacturing, as compared to 150 billion litres direct water consumption and 250 billion litres for electricity generation (IEA 2025). The water impact during the EoL phase is not discussed in the literature; hence, it is only hinted at in Figure 9.

The ratio of consumption and temporary withdrawal is very different for electricity generation than for on-site water: For 2023, the IEA estimated 250 billion litres of water were consumed electricity generation, but at the same time, the water withdrawal is estimated to have been more than 20-fold higher at 4,850 billion litres (IEA 2025). For the US, (Shehabi et al. 2024) estimate a smaller but still large ratio of about 12: 66 billion litres consumption and 800 billion litres withdrawal.

3.5.5. Water consumption: Relation between direct and indirect consumption

Unlike, for example, GHGs, the environmental impact of water is mainly local and relates to its local scarcity. And as argued by (Li et al. 2025), consumption is the water indicator that is relevant to relate to the watershed-level scarcity. As also argued in (Coroamă and Dumbravă 2026), we thus consider the *water consumption* as indicator for the water impact and ignore withdrawal. This corresponds to what is shown below the dashed horizontal line in Figure 9.

Three sources of water consumption have been evaluated in the literature. The IEA estimates the water consumption for electricity generation to be a factor of 1.6 times higher than the on-site water consumption, 250 billion litres as compared to 150 billion litres (IEA 2025). As stated above, the upstream consumption due to semiconductor manufacturing is substantially lower than both of the above, being estimated at 30 billion litres.

Using as case study the training of one AI model (GPT-3) and considering the US average water intensity of electricity production, (Li et al. 2025) estimate that 0.7 million litres were used on-site for cooling and 4.7 million litres upstream in electricity production. This is a substantially higher factor of about 6.7 between upstream and on-site water consumption, as compared to IEA's global estimate of 1.6.

The authors, however, might have used an overstated number of 4.6 litres of water consumed on average for the electricity generation in the US; the speciality literature in the field computes a number that is 2.5 times lower: 1.8 litres of water per kWh in the US (Peer et al. 2019). Correcting the upstream water consumption by this factor yields a ratio of only about 2.7 between upstream and on-site water consumption; much closer to the IEA's global estimate of 1.6.

3.5.6. Water usage effectiveness: Relating on-site water to electricity consumption

A metric to relate the water consumption of a DC to its electricity consumption was proposed in 2011 by the non-profit organisation "The Green Grid" (Azevedo et al. 2011). The "water usage effectiveness" (WUE) assesses how much water a facility uses relative to the energy consumption of its IT equipment and is defined as $WUE = \frac{Water\ usage}{IT\ equipment\ energy}\ \left[\frac{litres}{kWh}\right]$.

Depending on the cooling technology deployed by individual DCs (see Section 3.5.4), the WUE can take values that span several orders of magnitude, from as low as 10^{-3} (which equates to only 1ml / kWh) for airside economisers and air-cooled chillers to as high as 10^{0}-10^{1} (1-10 litres / kWh) for water-cooled chillers deploying cooling towers (Lei et al. 2025).

Users of AI models (whether individual users or companies) are usually unaware of the exact data centre where a query is executed, and much less of where the models have been trained. Given this uncertainty, from a user's perspective, industry average WUE values are more relevant than individual DC values.

Across all of its data centres, for example, the large colocation DC operator Equinix estimates a WUE of 1.07 (Higgins 2024). In a recent paper, Google does not directly estimate the WUE, but both energy and water consumption per median query; these are 0.24 Wh / prompt and 0.26 ml / prompt, respectively (Elsworth et al. 2025). Relating these two yields a WUE value of 1.08, virtually identical to the 1.07 put forward by Equinix. An average across all the DC industry puts forward a slightly higher value of 1.8 litres / kWh (Tozzi 2025), while Amazon claims a much lower value of 0.15 litres / kWh across its DCs (Amazon Sustainability 2026).

3.5.7. Reasonable default values

Given this discussion in Section 3.5.6, in absence of precise, site-specific data, it thus seems reasonable to work with an **average WUE value of 1 litre / kWh for the direct / on-site water consumption**.

Likewise, given the considerations in Section 3.5.5 and the fact that most AI compute takes place in the US, it also reasonable to assume a default average water intensity of electricity of **1.8 litres / kWh for the indirect / upstream water consumption**.

Given more precise water consumption data for either on-site or upstream water, these values should be used, of course. Otherwise, considering an **overall water consumption of 2.8 litres / kWh of DC electricity consumption**, is thus a reasonable default value.

3.5.8. Energy and water trade-offs

Given the cooling methods deployed in a DC, there can be a trade-off between energy and water consumption in a DC (Higgins 2024): by using more on-site water, DCs can reduce their cooling energy. Actually, there are two (related) trade-offs: between on-site and upstream water consumption and between overall water and electricity consumption (Coroamă and Dumbravă 2026).

These trade-offs are stronger, the "drier" the electricity is. For very "wet" electricity, saving cooling energy by spending more on-site water also saves a relative large amount of water upstream. It is thus worth both from a total energy and a total water perspective to spend more on-site water, as this saves both energy and water upstream (Coroamă and Dumbravă 2026). For dry electricity, however, the additional water spent on-site is not compensated by water saved during electricity production – there is this a strong trade-off between total energy consumption and total water consumption, which occurs mainly on-site (Coroamă and Dumbravă 2026).

4 AI system assessment

The current chapter discusses how the taxonomy of AI systems and its implications for assessments. Section 4.1 first presents the wider AI software ecosystem, which can comprise several ML models as well as various non-AI components. Section 4.2 discusses the tightly connected LLM architectures; although strictly speaking they could have been discussed in the previous chapter on AI models, the topic is more meaningful after having discussed architectures of AI systems. Finally, Section 4.3 addresses data visibility on different levels and its implications for the assessment.

4.1 The wider AI software ecosystem

As briefly addressed in Section 2.1, ML models – and LLMs in particular – typically do not operate in isolation, but are embedded in a larger ecosystem. While there is no universally accepted term for this ecosystem, researchers and practitioners have converged on several overlapping notions that attempt to describe it: the *LLM application stack*, the broader *generative AI ecosystem*, the emerging *agentic AI stack*, or – in more theoretical circles – the *cognitive architecture* perspective.

The expression *LLM application stack* (or "LLM tech stack") is widely used among developers and applied-AI engineers to describe the modular layers that make up modern LLM systems: the base model, retrieval components, orchestration logic, and monitoring infrastructure (Hada 2025). A good (i.e., quite comprehensive, yet easily understandable) overview schematic is presented by (Bornstein and Radovanovic 2023). (Cheung 2024) introduces terminology to describe the various types of ecosystem components, distinguishing between *model layer* (which hosts the LLM models themselves), *data layer* (for the databases), *orchestration layer* (bringing everything together) and possibly an *operational layer* (for monitoring and validation). (Khattab et al. 2023) show how retrieval, prompting, and reasoning modules can be compiled into declarative pipelines.

A closely related but broader framing, the *generative AI ecosystem* appears in both industry analyses, but also foundational academic work such as (Bommasani et al. 2022). The latter places LLMs within a multi-modal landscape of generative systems spanning text, code, image, and audio, emphasising shared infrastructures and social implications rather than specific engineering layers.

As LLMs progressively gain reasoning and tool usage capabilities, they started to be increasingly independent, while at the same time interacting within more complex ecosystems. This paradigm is often referred to *agentic AI* (see Section 4.1.3 below), while its ecosystem is correspondingly sometimes referred to as "agentic AI ecosystem" (Lemarchal and Laifa 2025) or "agentic AI stack" (Kaur 2025). "Agentic AI" describes systems which can reason, plan, and act through tools and APIs, whereby reasoning and action and interleaved in iterative loops (Shinn et al. 2023; Yao et al. 2022).

Finally, the term *cognitive architecture* is sometimes used in academic and advanced AI development contexts to describe the high-level, structural design of an intelligent system where an LLM is a core component. (Park et al. 2023), for example, use the term while demonstrating a working cognitive architecture with perception, memory, and reflection. The term, however, stems from cognitive sciences and computational cognitive science (Chase 2024), being historically used in AI research to describe efforts towards non-ML cognitive architectures. To avoid confusions, this term will thus not be used further.

The following subsections briefly present some of the most important components of an AI ecosystem.

4.1.1. Web searches, database search, and further services complementing ML models

Already for what is arguably the most familiar AI usage case for most end users – i.e., Web-based interaction with a large language model – the ML model is embedded on provider side into an own ecosystem. As shown (in a simplified way) in Figure 10, users do not interact with the LLM directly. Instead, their prompt reaches a web user interface (Web UI) first. For large LLMs, there are substantial numbers of virtual machines (VMs) taking care of this user interaction. The Web UI typically further interacts with some security and encryption/decryption services such as firewalls or an SSL proxy, then

sends the information further to the *orchestration layer* (Cheung 2024), sometimes also called *control layer*. As already addressed in the overview in Section 2.1 as well as in Figure 3, the orchestration layer is the main application on provider side, which coordinates the entire logic of handling the user's prompt.

For an LLM provider – and only for an LLM provider – the LLM is internal to the system. In Figure 10, it is shown accordingly within the AI system internal boundaries and is coloured in blue.

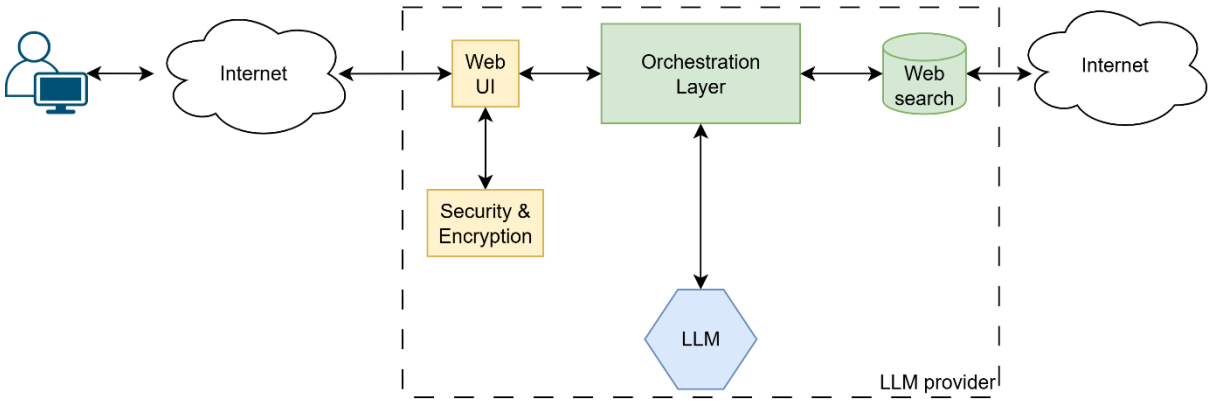


Figure 10: A simplified architecture of the provider-side ecosystem around an LLM, consisting of a Web UI receiving the prompts and handling security and encryption, and a control layer deciding which prompt will be sent unmodified to the LLM, and how it might be enhanced beforehand, e.g. by providing current knowledge from web searches.

Additional services can be, for example, web searches. While any LLM encapsulates an impressive amount of knowledge, this knowledge is frozen in time. It is limited to the time of – and the sources used while – training the model. This phenomenon is known as "knowledge cutoff date" and it used to characterise early, publicly usable LLMs (Peham 2024). News and current events, for example, can inherently not be known to the model. Similarly, the user's wish for real-time verification (e.g., of a reference) can also not be performed by the LLM itself.

Early publicly available LLM systems (around 2022 – 2023) were indeed unable to perform tasks such as informing about events after their knowledge cutoff date or double-checking their statements against sources which they themselves had indicated (indeed, sometimes not only the information provided, but even the stated source could be a hallucination). As these represent a popular class of prompts, however, this has to a large extent changed today: As also shown in Figure 10, orchestration layers have the ability to perform web searches, navigate the sites resulting from the search, embedding relevant parts of these answers into the user's prompt, and sending this new, enhanced prompt to the LLM. This process is outlined in detail by (De Koninck 2024).

In a similar way, the communication between users and other types of ML models is also orchestrated by a control layer. Figure 11 shows a simple example for a company-internal ML model, which may be a smaller LLM instance or anything else. As before, the user interacts with the orchestration layer, which mediates access to the ML model, while also being able to access company-internal information and resources such as vector SQL and NoSQL databases, customer-relationship management software, and so on.

These two simplified and archetypal examples aim to show that whether the ML model is internal or external, whether the additional services are also internal and/or external, there is usually an ecosystem that needs to be taken into account when addressing the energy consumption or environmental impact of an ML-base service.

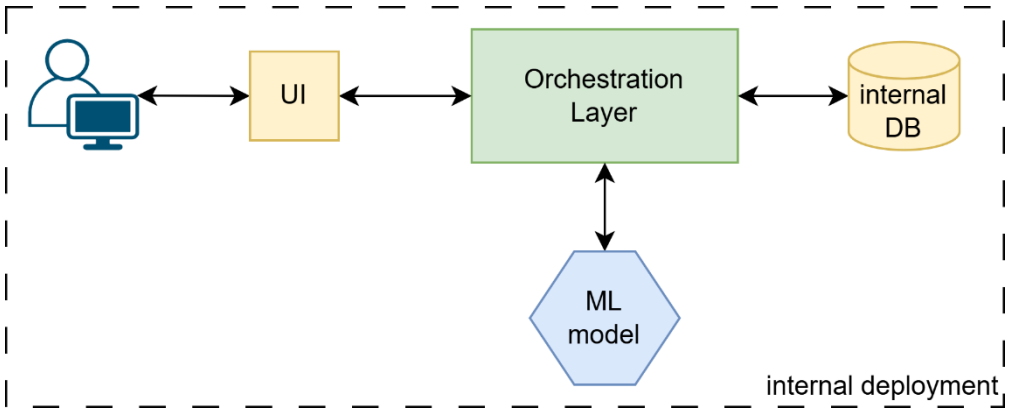


Figure 11: Example for a simple ecosystem of an internal ML deployment. Here as well, the user does not directly interact with the model. The control layer mediates interaction with the model, having access to additional internal resources such as different types of databases.

Some of the services of this ecosystem are dedicated to the ML model and thus need to be allocated to 100% to it. Others – which, in line with Figure 3, are also depicted yellow in Figure 10 and Figure 11, may be shared among several usages, as is often the case for generic firewalls or encryption services – their impact needs to be partially allocated to the ML model.

4.1.2. Retrieval-augmented generation (RAG)

The two examples above highlighted a core feature of current LLMs: The contextualisation of a user's query with additional data sources. Section 4.1.1 did not discuss how exactly the orchestration layer performs this task. While there are several ways to achieve it, the current dominant paradigm is that of retrieval-augmented generation (RAG).

RAG is a fairly new concept that was first proposed in 2020 (Lewis et al. 2020). As in the simple examples from Figure 10 and Figure 11, a RAG system also uses various information sources, often called "retrieval corpus" or "knowledge base". These can be company-internal sources, such as a collection of PDF documents or knowledge management systems such as SharePoint, as well as external sources such as web searches or live databases.

For efficient retrieval of context that is semantically related to the user's prompt from a large collection of documents, most RAG deployments deploy *vector databases*. To populate this database, the documents from the knowledge base are first cut into chunks. The text chunks are then converted into numeric vector representations within a high-dimensional space and stored in the vector DB. They will later allow fast retrieval based on semantic similarity (Google 2024).

The high-dimensional vector representations are often called "embeddings" (LangChain 2023). The transformation of text into embeddings is correspondingly named "embedding". The process is performed by an "embedding model," which is typically a specialised and thus relatively nimble ML model itself.

To search within the vector DBs and use the sematic context, the RAG pipeline is shown in Figure 12: The user query (1), after going through security checks and decryption (2) arrives at the control layer (3) and is then typically enriched with the relevant chat history (4). Original query (which might have undergone further transformations such as correction of typos; not show in the image) plus chat history are sent to the embedding model (5). There, they are treated as a chunk of text themselves and transformed into a numeric vector representation ("embedding", 6). This embedding representing the prompt is then used to retrieve semantically similar vectors from the vector DB (7). The text chunks these semantical similar vectors represent are then retrieved (8), appended to the user's prompt, and finally sent to the LLM (9) to generate its response (10) (Martineau 2023).

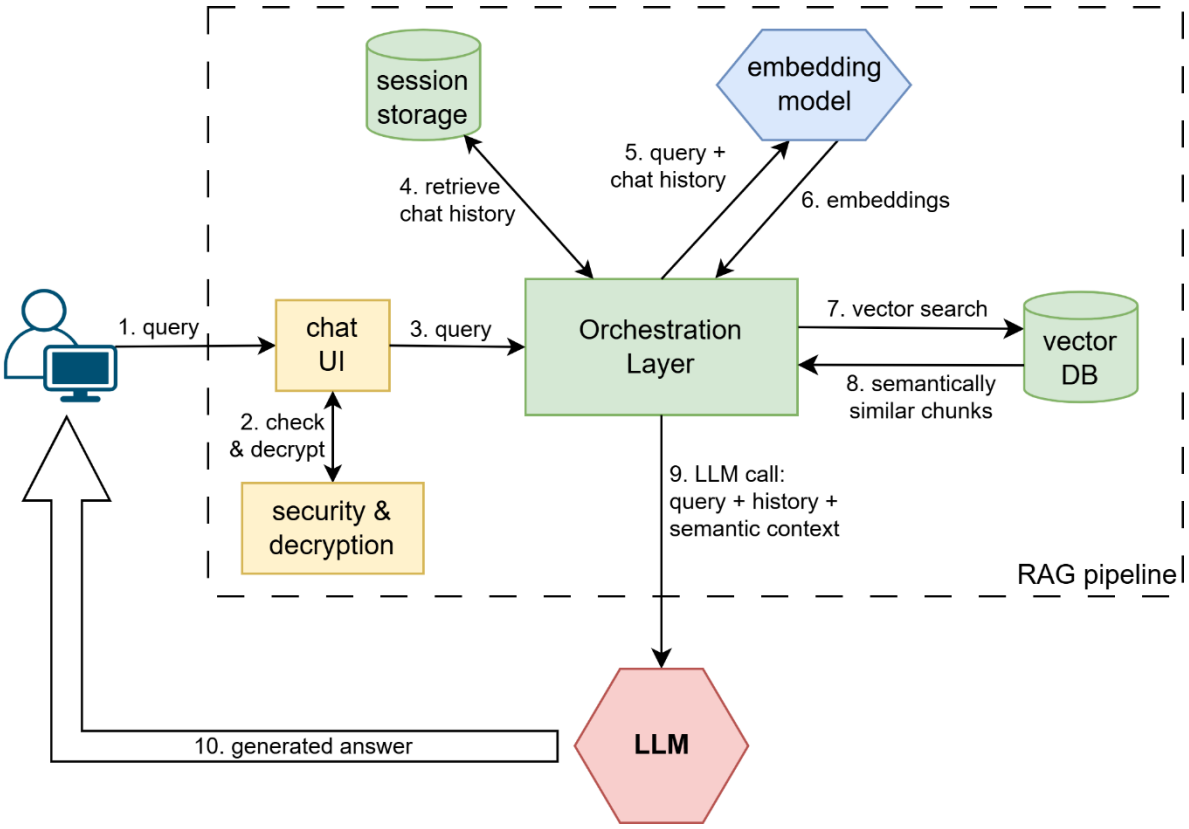


Figure 12: Simplified, archetypal RAG method, consisting of the RAG pipeline for retrieval and context assembly as well as the LLM for subsequent retrieval-augmented generation. Cylinders represent databases, hexagons ML models, rectangles further parts of the LLM ecosystem.

The pipeline aims at the *retrieval* of semantically similar information from the knowledge base via embeddings and the vector DB. The very last step is the *generation* of the answer by the LLM – hence the "retrieval-augmented generation" name. Although the two are tightly connected and together constitute the *RAG method* (i.e., software pattern), only the retrieval and context assembly stage is considered as part of the *RAG pipeline*, while the LLM-based generation is considered an external component.

While most RAGs are indeed based on embedding and vector search, this is not the only way to retrieve relevant information. Retrieval-augmented setups can also use keywords of full-text search (e.g., via Elasticsearch), in which retrieval is based on word overlap, not vector distance, or even (bur rarer) structured queries on SQL databases. Finally, some advanced search engines employ *hybrid search*, which combines both semantic search and keyword search (Google 2024).

RAG systems have been so successful because they are relatively easy to implement and bring several advantages (Google 2024; Merritt 2025):

- provide *factual grounding*, mitigating GenAI hallucinations – i.e., plausible but incorrect LLM answers,
- provide up-to-date information, *circumventing knowledge cutoff*,
- enable models to *provide sources* and hence users to check their claims,
- are a faster and less expensive method than retraining to keep *the model up-to-date*, and
- allow users to *swap sources* on the fly, and thus deploy a generic model to a variety of contexts.

4.1.3. Agentic AI

Agentic AI takes the core idea of RAG – *augmenting a model's answers with retrieved, up-to-date, domain-specific context* – and turns it into just one capability inside a broader *goal-directed loop*. In a RAG setup, the shape of the solution is typically fixed: retrieve relevant passages, then generate an answer grounded in them. In an agentic setup, retrieval becomes a tool the system chooses to use (or not) while it plans and executes steps towards an objective: It can iteratively search, ask follow-up questions, compare sources, write intermediate artifacts, call APIs, and revise its approach based on what it finds.

In other words, agentic AI expands RAG from *context injection* into *context-driven action*: While RAG supplies information, the agent decides how to sequence information gathering, reasoning, and which tool to use to actually complete a task. According to IBM, "agentic AI" are thus "systems that can accomplish a specific goal with limited supervision […] Unlike traditional AI models, which operate within predefined constraints and require human intervention, agentic AI exhibits autonomy, goal-driven behaviour and adaptability. The term "agentic" refers to these models' agency, or, their capacity to act independently and purposefully" (Stryker 2025).

Fundamentally new is thus the dynamic decision-making. An agent is not following predefined steps, but *selects* actions under uncertainty – planning, branching, retrying, and adapting based on feedback from the environment. This is what separates it from mere automation. Automation follows a predefined workflow ("if X then do Y" in a fixed sequence), whereas an agent performs on-the-fly reasoning and planning.

For reasoning, the agent typically uses an LLM (or rather, a subcategory of LLMs that have undergone reasoning-focused fine tuning and are called "large reasoning models", LRMs). Like RAGs, it also deploys memory to carry forward relevant context across sessions as well as tools to interact with the outside world (e.g., information retrieval, actions such as updating systems or sending messages, and orchestration such as the delegation of sub-tasks to other agents). The workflow, however, emerges at runtime: While the goal is stable, the path is negotiated step-by-step rather than hard-coded.

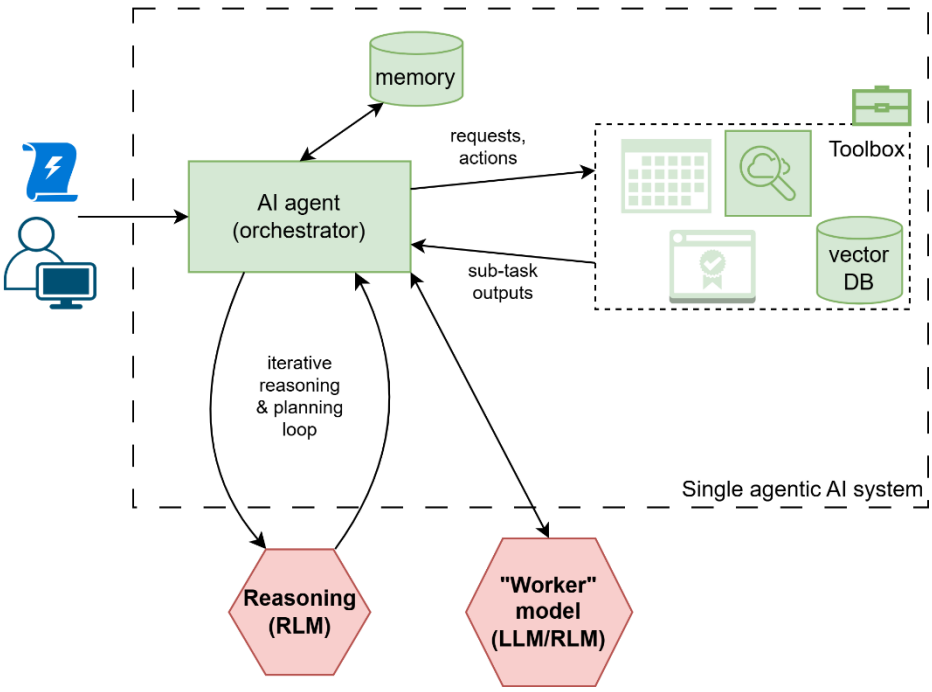


Figure 13: Simple archetypal example for a single-agent agentic AI system. The reasoning model and worker model can be both either internal or (as in this example) external. Additionally, while architecturally these two roles are clearly distinguishable, in practice a single external model will be often used for both roles.

Figure 13 shows an archetypal agentic AI system. It is the simplest possible such system, which does not employ sub-agents; it is thus a “single-agent” system. Its central piece is an *AI agent*, which performs the controlling / orchestration role, very similar to the orchestration layer in RAGs. As opposed to a RAG orchestrator, however, the agent decides contextually how to follow its goals. After perceiving a relevant piece of context (which may come from either memory or the numerous tools it can interact with), it often sends this information to a reasoning model, which decides upon the next steps, chooses the tools to be deployed, manages the state, etc, sending this info back to the controlling agent.

Following the reasoning steps thus indicated, the AI agent will then again interact with tools, which can be both internal – the “toolbox” represented in Figure 13, which can contain internal and external databases, web searches, a calendar, and more complex tools such as code compilers – and external to the system. Usually, there is also a complex ML model involved, which addresses the complex tasks such as text generation, summarising text, code writing, etc. This is represented as the “worker” model in the figure, and is often an LLM (or rather, an RLM) itself.

Architecturally, an agentic AI system thus deploys two different AI models: A *reasoning model*, which supports the agent in planning future tasks, and the *worker model*, which performs some of these tasks (usually the more complex ones). In many real systems, both of these roles are implemented by a single LLM using different prompts (also called “modes” or “policies”). This is fairly common because it is simpler for the agentic AI system developers, and modern frontier models are powerful enough to both plan and execute.

However, two-model (or multi-model) setups are also common for cost, latency, reliability, or safety reasons. One such architecture used for relatively simple agents, for example, consists of a smaller (and thus also cheaper and often system-internal) reasoning model, and a bigger model for the more difficult generation. Two external models from the same family, but different variants (such as “normal” and “mini” or “fast” and “thinking”), could be deployed**:** a fast model for reasoning and tool selection, and the stronger one for the final response. By contrast to Figure 13, one or both of the models could also be internal, in particular the reasoning model (if the two are in fact different).

A final remark regarding this archetypal agentic AI system refers to its initialisation: As opposed to the non-agentic architectures discussed in Section 4.1, the initial trigger can come from either a human user or it can be triggered by various machine processes. A very simple such trigger might be a timer, which for example could start the agent every morning at 6am. But it could also be triggered by other AI agents, which rely on the current agent to perform a sub-task. This leads to *multi-agent agentic AI systems*, briefly discussed below.

AI agents can not only employ tools, memory, reasoning and worker models to accomplish their tasks, but also other AI agents in turn. A simple example of such a multi-agent setup is sketched in Figure 14: A first agent employs a second one for the execution of a sub-task. This second agent could in turn employ further agents for some of its sub-tasks, and so on.

In multi-agent setups, external delegation is thus a qualitative break: As an external agent may itself deploy various models, tools, and further agents in turn, this yields a potentially unbounded compute tree of uncertain energy consumption.

While any external ML model used by an AI system cannot be directly assessed but needs to be approximated, this can be done with some level of confidence, as Chapter 5 below will discuss. The energy consumption of sub-agents, however, is so intricate that it cannot be defined on a generic level. Multi-agent systems are thus out of scope for the SAFE-AI framework, which only analyses non-agentic AI systems and single-agent systems.

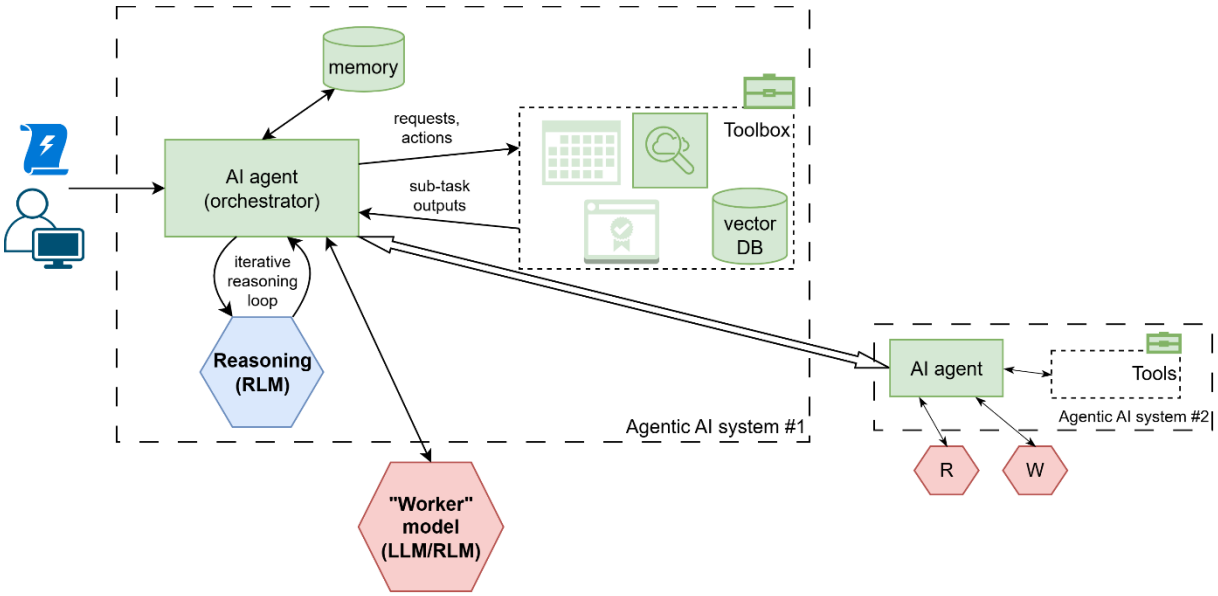


Figure 14: Example of a multi-agent setup, in which an agentic AI system delegates sub-tasks to a second agent, merely sketched in the figure. This second agent could in turn employ further sub-agents (not shown). For variety, the figure shows the reasoning model of the main agent as internal model, and thus also physically different from the external worker model.

4.2 LLM architectures

4.2.1. Transformer architectures

Transformer models are fundamental to many recent advances in artificial intelligence. The term refers to a neural network architecture based on stacked (i.e. repeated) self-attention and feed-forward layers, implemented using large-scale parallel matrix operations. In each layer, learned parameter matrices are applied to vector representations of the input sequence to model diverse patterns of dependency (e.g. syntactic or semantic relationships in text). Notably, the learning during training and the generation of text during inference are both served by the same Transformer architecture.

Transformers can be trained and operated **autoregressively**; in this mode, models generate output sequences, most commonly text, by iteratively predicting statistically likely tokens conditioned on preceding context. While Transformers are fundamental to LLMs, they are also used for image recognition (Dosovitskiy et al. 2020), protein generation (Jumper et al. 2021) or robot control (Brohan et al. 2023).

While the transformer architecture has been applied for a few years, current LLMs are predominantly built as **decoder-only**. The earlier generation were encoder–decoder variants (Vaswani et al. 2017). These were designed for sequence-to-sequence processing; in other words, consuming a complete input sequence before producing a complete output sequence, for example for translation tasks – but also used in the state-of-the-art “segment anything model” (SAM) segmentation models (Carion et al. 2025). Decoder-only architectures (e.g., GPT-4, Llama), on the other hand, are autoregressive; where each new token is predicted based on the elements that came before it and appended at the end of a sequence. In this sense, decoders are iteratively working on growing fragments of the output: they process an initial input sequence and iteratively predict subsequent tokens, appending each output to the context window to drive the next generation step.

4.2.2. Compute characteristics of inference: Prefill vs. decoding

The energy profile of decoder-only LLMs is defined by two distinct phases, each exhibiting fundamentally different hardware efficiency characteristics.

Prefill: Upon receiving a prompt, the model processes all input tokens in parallel. This phase is characterized by dense matrix–matrix multiplications, which allow for high arithmetic intensity and the saturation of parallel compute units (e.g., GPU tensor cores). Consequently, prefill is relatively energy-efficient per-token, as the primary energy cost is allocated to active computation rather than data movement. The prefill phase produces vector representations for each prompt token as input to the subsequent generation phase (called "decoding"), along with cached key‑value encodings of the prompt context for each layer.

Decoding: During the generation phase, the model produces tokens sequentially. Each step requires repeated transferring the model weight matrix from high-bandwidth memory (HBM) and the expanding key–value cache from memory to perform a relatively small vector calculation. This creates a memory bandwidth bottleneck, forcing the accelerator into a memory-bound regime where compute cores sit idle awaiting data transfer from HBM. As a result, the energy cost per token during decoding is significantly higher than during prefill. To compensate for this bottleneck, decoding from several concurrent requests are 'batched' together. Empirical analysis confirms that higher batch sizes can increase utilisation of the GPU (Fernandez et al. 2025).

One central determinant of inference energy consumption is the context window length, i.e. the sum of prompt and output length of a request. As the context length increases, the size of the key-value cache grows linearly, increasing memory traffic during decoding, while prefill cost grows super-linearly due to attention computations. Thus, the context length (as the sum of system prompt, session context, input prompt and output length) increases computational complexity and drives demand for infrastructure capacity (number of GPUs, size of memory per GPU, and indirect dependencies on cooling and power supply).

Two important trends are further driving workloads to become more **decode-heavy**, and thus dependent on the generated token number, as opposed to the prompt length. Both of these trends were already introduced in Section 4.1.3:

- Firstly, the recent LLM evolution of LRMs (i.e., reasoning models**)** also use this decoder-only transformer structure but are trained (via reinforcement learning) to generate a long, internal chain-of-thought (CoT) (Wei et al. 2023). These output tokens (often hidden from the user) further shift the ratio between the complexity of prefill and decode phases, keeping the GPU in memory-bound states for longer and imposing bottlenecks on the scalability of the hardware.
- Secondly, the emergence of agentic workflows (models invoking other models or APIs as part of a request, also discussed in Section 4.1.3) has further altered the inference workload profile as tool execution can block sequential token generation for extended periods. This results in more complex job scheduling, including off-loading of tasks and keeps hardware in the inefficient, memory-bound regime for longer durations, amplifying bandwidth bottlenecks and significantly raising the average energy consumption per task.

4.2.3. Mixture-of-experts (MoE) models

While the performance of neural language models often follows predictable scaling laws, i.e. larger models trained on *proportionally larger* datasets tend to achieve higher task performance (Hoffmann et al. 2022), this increase in model size also worsens the decode-bottleneck. To decouple model capacity from inference cost, recent architectures increasingly adopt mixture-of-experts (MoE) designs (Fedus et al. 2022). In these models, dense feed-forward layers are replaced by a set of specialized "expert" subnetworks that are invoked dynamically across separate GPUs. This architecture allows models to scale to massive parameter counts while maintaining the active computations of a much smaller model. However, MoE introduces trade-offs: while arithmetic operations per token are reduced, it increases interchip communication overhead, and its efficiency gains can diminish during decoding if expert utilisation becomes unbalanced, i.e. in cases where one expert is being used much more than others (Lepikhin et al. 2020). The effect of MoE on energy efficiency is thus highly workload- and implementation-dependent.

CoT and MoE together reinforce memory as the limiting factor, mainly driven by movement of the key value cache from HBM to the GPU cores. Scheduling in order to reduce idle wait time of GPU cores thus becomes more important for operational efficiency, driving in turn innovation in schedulers, such as vLLM (Kwon 2025).

4.2.4.Beyond transformers: Emerging architectures

While transformers remain the industry standard, they are facing scaling barriers for massive contexts. Emerging alternatives, such as State Space Models (SSMs) (Gu and Dao 2024), aim to address the memory-bottlenecks from growing context caches by reducing memory access. Despite their potential for energy efficiency, these architectures currently face challenges with exact information retrieval over longer interactions and lack the mature hardware optimization ecosystem of the transformer architecture. Consequently, the decoder-only transformer is expected to remain the dominant architecture for the near-term.

4.2.5. Architecture of image and video generation models

State-of-the-art image generation does not use transformer architectures but diffusion models (Ho et al. 2020). Such models are trained from images to which noise was added. The model learns to predict how much noise was added. During generation, the process is reversed, by sequentially removing noise from an initially entirely randoms set of pixels.

In image generation models, the term “token” typically refers to a spatial unit of representation in image or latent space rather than a linguistic unit. Unlike decoder-only language models, most modern image generators (e.g. diffusion models) update all latent representations in parallel over a fixed number of denoising steps, rather than generating tokens autoregressively. As a result, token-based functional units that are meaningful for text generation do not translate directly to image generation.

Instead, energy consumption of this process is determined by configuration choices: the number of denoising steps and the number of pixels, i.e. the resolution of the image. Correspondingly, energy consumption does not depend on the content of the image.

These principles extend to video generation. Here the energy consumption is determined by number of images that make a video sequence, their resolution and the number of denoising steps.

4.3 Hierarchies of information access

As already mentioned in Section 2.2 whilst introducing the three levels of assessment from Figure 5, the combination of system architecture and who is performing an assessment has a large influence on data visibility. This, in turn, shapes the types of assessment that can be performed as well as the primary and secondary data that can be deployed for the assessment. After having discussed possible system architectures in Section 4.1 above, the current section now addresses data visibility.

The feasibility and reliability of assessing the environmental impact of both entire systems and – as will be discussed in the next chapter – a single AI service instance depend not only on technical measurability, but also on *who* conducts the assessment and what information they are able or willing to disclose. In practice, access to detailed operational data is shaped by supply chain relationships, institutional roles, commercial incentives, and power asymmetries between model providers, service deployers, and external assessors.

As a result, the most causally accurate indicators of impact—such as fine-grained compute utilisation, token-level efficiency, or hardware-level energy use—are often available only to actors with direct control over the infrastructure, and are unlikely to be reported publicly in the near term. This section therefore examines assessment perspectives not merely as technical viewpoints, but as positions within an information landscape that fundamentally constrains how impacts can be allocated and interpreted.

To adhere to the PCF guiding principle of *accuracy (see [cross-ref 2.2.3]*, exchanges should preferably be quantified with *primary* data that was collected (ideally measured) for the unit processes in the

assessment. Existing data, from external sources or outside of the assessment context is called *secondary*. For ICT services, it is usually possible to collect primary data of the use-phase energy consumption for at least part of the digital value chain of the end-to-end service (Preist et al. 2014); see also Table 5. However, no single party in *isolation* has access to primary data across the entire value chain.

Access to comprehensive primary data is unfeasible for life cycle phases other than the use-phase, i.e. embodied impacts; due to the deep supply chain of within the ICT sector. Furthermore, there has always been a lack of reliable secondary LCA data for both energy use as well as material inputs along the digital supply chain, among others due to technical constraints such as rapid innovation, size of the supply chains (Williams 2011), that are easy for companies to hide behind (Friday et al. 2024). The lack of transparency has been called out, for example, for the ICT sector-wide carbon footprint (Freitag et al. 2021), cloud (Mytton 2020a), video streaming (Makonin et al. 2022), and now AI services (Luccioni et al. 2024).

For `connected' cloud services there is also a tension between the richness of real-time monitoring data already collected during the operation of the digital infrastructure (but held back), on the one hand, and the lack of robust data on use-phase energy consumption (Mytton 2020b). This lack of transparency has extended to use-phase energy consumption by AI services running in the cloud (Luccioni et al. 2025).

The need to use estimated secondary data thus affects assessments differently, depending on who carries out or commissions an assessment. Table 5 summarises access to primary data to different system layers relative on the perspective of the assessment. Notably, there is no party (or perspective) that can source primary data throughout the distributed system.

Table 5: Overview of typical boundaries to access primary (measured) data of energy consumption as they vary with perspective of the party carrying out an assessment. a) if the AI model is hosted by a Frontier model provider or b) self-hosted by a dedicated service provider who also host the business logic.

	Hardware (embod.)	**Use Phase (Energy)**			
		Data Centre		**Network**	**User Devices**
Perspective		AI Model	Business Logic		
Frontier-model Provider	No	**Yes** (Training and Inference)	n/a	No	No
AI-enabled Service Provider	No	"AI as-a-Service" Training and Inference: **No** "Self-hosted open weights" Inference: **Yes** Training: **No** "Self-trained" Training: **Yes** Inference**: Yes**	**Yes (all cases)**	Modelled with measured DC in/egress and billed CDN in/egress	Modelled with analytics data to infer power consumption
3rd-Party (e.g. Public Research)	No	**No**	**No**	Modelled with network coefficients and typical traffic measurable on the user device	*Modelled w*ith data for "typical" devices

In more condensed form, this is shown in Figure 15, showing how access to measured data on processes in data centres drops away from the model providers who control visibility of energy consumption

data. The columns in Figure 15 are to be interpreted as follows: the first column (“model provider”) is for a deployment such as that in Figure 10, in which the LLM is accessed via an Web API. In this case, the model provider is also service provider and – as the entity originally training the model – can directly measure training, inference, and orchestration. The second represents the case when a different service provider uses an LLM that is deployed either internally or externally. Finally, the third column shows the (lack of) visibility for a third party in either of these cases.

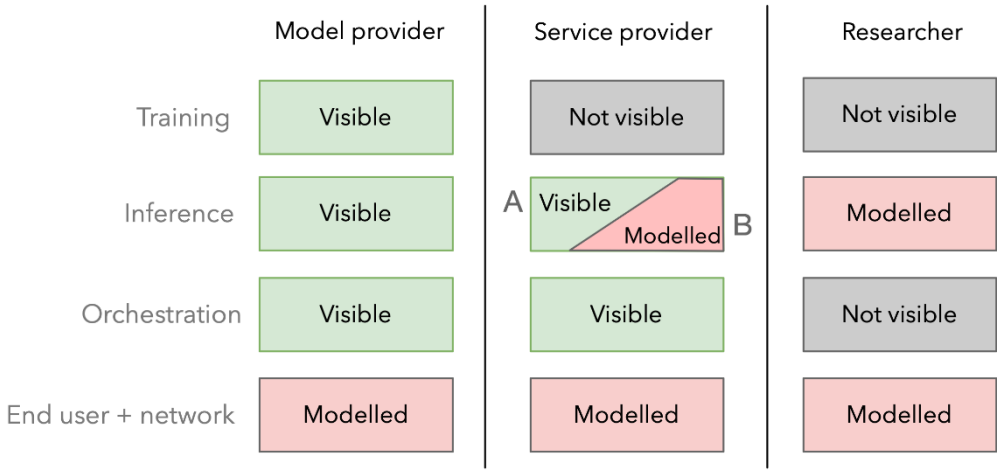


Figure 15: Typical visibility of measured energy consumption data for elementary processes part of an AI system, either for ‘A’ a self-deployed model or a ‘B’ cloud-based model.

5 AI usage assessment

For many assessments goals, the setting of boundaries around an *AI system* or around only the *AI model* itself (e.g., for frontier model providers), is sufficient. As discussed in the two previous chapters, these can be system boundaries to support goals relevant to both academics and practitioners, such as:

- On the micro-level, quantifying the energy, GHG, or water of individual ML models or the larger AI systems they are enclosed in,
- On the meso-level, building on assessments of individual AI systems and aggregating them towards an organisational assessment, or
- On the macro-level, estimating AI's yearly global consumption such as (IEA 2025; Kamiya and Coroamă 2025).

Other questions, however, require breaking down the overall impact of an AI system to *individual usage instances*. This can be required, for example, when comparing the per-usage impact of an ML model to its predecessor, or of two AI systems from different providers. It is also necessary for consumer-oriented eco-labelling, the assessment of a user's personal (usage-based) AI footprint, and further similar questions.

As the measurement of energy consumption and other resources is not fine-grained enough to measure individual usage, an attribution step called *allocation* is necessary. This introduces uncertainty and thus needs to be carefully chosen. The relationship between usage of the AI system and its energy consumption also depends on the way these systems work and how that drives their energy consumption. Hence, the allocation presented in Section 5.1 draws on the LLM architectures discussed in Section 4.2 above.

Robust methods also include a deliberate choice of a *functional unit* that is precisely and unambiguously defined to assure a fair and accurate comparison between different systems – including a measure for the quality of the AI service delivered – while finding a balance between comparability and generality that is appropriate for a given decision context. This topic is addressed in Section 5.2.

There is an inherent tension between the practical limitations of data access as described in Section 4.3, the need for accuracy in the choice of allocation key, and the desire for comparability of functional units. We call this an "assessment trilemma", within which every assessment is positioned. The assessment trilemma is presented in Section 5.3.

Finally, Section 5.4 suggests an assessment metric and presents default values for its parameters. Importantly, however, as Figure 5 has also shown, system-level and usage-level assessments are tightly interconnected. Even when the aim of the assessment lies on system level, if the AI system employs API calls to external models, their assessment needs to start at the AI usage level, to then aggregate these external calls. Consequently, AI usage assessment is also required for companies that want to assess their Scope 3 emissions while using external AI models, or for corresponding policymaking. The chapter thus finishes by presenting in Section 5.5 a stochastic analysis for the aggregation from AI usage to AI system level.

5.1 Allocation

Allocation is required whenever the environmental impacts of a system cannot be directly attributed to a single service instance or output. For example, devices such CPUs, GPU accelerators, cooling systems, and networks are shared across many users, services, and tasks at the same time. As a result, determining the energy consumption of a single request requires somehow splitting up the observed energy consumption; for example, (and grossly simplified) by dividing the measured energy by a GPU over a month by the number of requests served during this time. This is called **allocation**. It provides a systematic way to apportion these shared impacts to individual services or service instances, enabling results to be reported per functional unit (e.g. per request, per generated output, or per user session).

The mathematical formula to split up the total energy consumption across requests is called an *allocation key*; such as request count, token count, or GPU time. The choice of allocation key directly influences the estimated impact per usage and implicitly encodes assumptions about how resource use scales across heterogeneous workloads. Allocation is therefore not merely a technical step, but a necessary modelling choice that determines how shared infrastructure impacts are interpreted at the level of individual AI services.

The choice of allocation also depends on whether the assessment frame is attributional or consequential. Attributional framing assigns a share of existing emissions to a user, while consequential framing estimates the change in emissions caused by a user's actions (see Appendix A for more details). Both assessment frames face limitations. Consequential assessments are currently not possible given, besides a lack of data, the fast rate of change of the technology, as infrastructure is not currently built as a response to demand but speculatively. In the absence of data on steady-state market behaviour, attributional assessments (or, perhaps, short-term marginal assessments) are the only ones feasible (Schien et al. 2024a).

Importantly, however, attributional assessments do not provide a robust account of the environmental consequences of decisions. This means that it is currently not possible to estimate the longer-term environmental impacts from additional demand (this applies to both increasing and reducing demand). Instead, impacts are currently assessed in an attributional way. This is particularly relevant for on embodied impacts (from making devices) and training, which are 'sunk' and invariable in the short-term.While attributional methods assign some share of the impact *per-usage*, it must not be interpreted as the proportion of carbon saved if the AI service was not consumed – this latter number can be orders of magnitude different (Schien et al. 2025). The actual consequential impact from usage on these short-term invariable aspects is highly uncertain.

Reducing uncertainty is a key factor for the choice of allocation principles for a process (Appendix A). Uncertainty enters LCA through a combination of variability and epistemic uncertainty (lack of knowledge); both are discussed below.

5.1.1. Variability

As discussed in Section 4.2, variations in query length, generated output, reasoning depth, tool use, and batching can lead to orders-of-magnitude differences in energy consumption between otherwise similar service instances. If not made transparent, such variability in the input parameters increases uncertainty of results (Schien et al. 2013). Variability enters LCA in almost all input parameters. While energy consumption per average or median inference may be used as pragmatic functional units, as discussed in Section 5.2 below, it is important to note that LLM inference costs are neither uniform nor stationary (i.e., they are changing quickly over time with efficiency gains).

Similarly, user behaviour is a source of variability. Users seek help from GenAI for wide range of use cases, from niche applications such as games and role play to major categories such as tutoring and teaching (Chatterji et al. 2025). Within each type of usage, individual user behaviour and styles to communicate vary (Brandtzaeg and Følstad 2017). As communication style affects the number of requests made to a system and token count, this variability affects energy consumption (Wilkins 2025).

Section 5.2 describes how heterogeneity in prompt length, output token count, CoT or agentic tool invocations affect energy consumption. The architectural and system-level characteristics discussed in Sections 5.2 and 5.3 imply that the environmental impact of an AI usage instance is highly variable and context-dependent.

One way to manage variability is to report descriptive statistics. However, user behaviour is frequently power-law distributed (Comscore 2016). Hence, the choice of the median as descriptive statistics can conceal potential bias from heavy users – that are nonetheless real (Coroamă 2025a).

5.1.2. Varying levels of uncertainty for different impact sources

Among the end-to-end system parts that require allocation, training stands out as the one that is most uncertain – not only for third-party assessments, but also for operators themselves. This is because

amortisation must be made ‘in advance’ based on the likely model amortisation rate. This is however, at best, an experience-based judgement that can be disrupted by market behaviour. This includes the derisking runs preceding the main training as well as model-unrelated research and the development of experimental models, which all first need to be allocated to the models that get into production, as discussed in Section 3.4.5. Amortisation of model training based on the number of inference invocations is further not feasible for open weight models which can be freely downloaded and used without tracking their use.

Allocation of inference-phase use-phase impacts from idle power and other overheads is relatively less uncertain compared to sunk costs. These are visible to model operators and can be measured in principle with sufficient accuracy.

Finally, the short-term marginal impacts from induced additional power consumption (per-token or per-request) are the only impacts for which impacts can be associated to use with tolerable confidence. This is the category that is most widely studied at the moment by energy-efficiency benchmarks, e.g. (Hugging Face 2025). In this way, uncertainty from allocation decreases the nearer the change of an impact is relative to the delivery of the service (Figure 16).

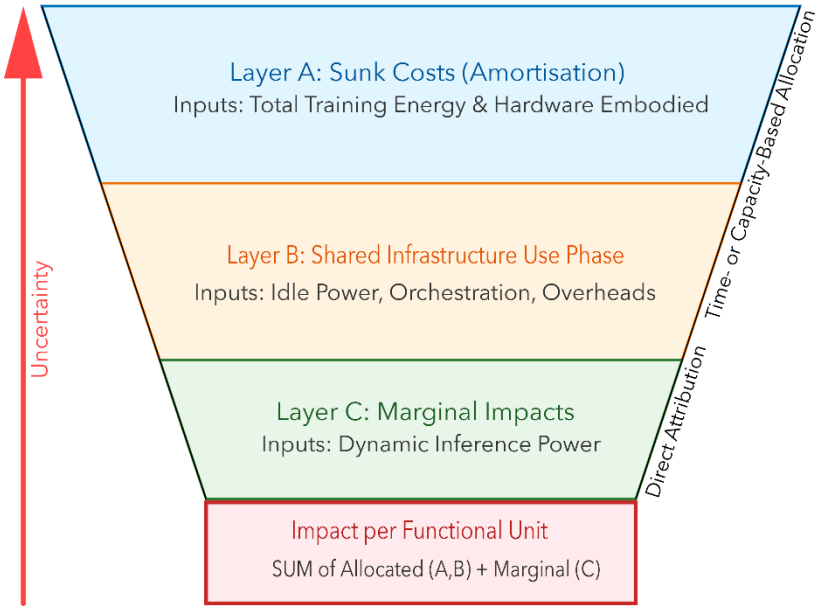


Figure 16: Allocation of increasingly specific and less uncertain impacts.

5.2 Functional unit

As mentioned in Section 2.2 and Figure 5, the choice of functional unit is considered central to LCA. What is often left unacknowledged is that it is the choice of allocation key that defines and shapes an assessment as much as the choice of functional unit. The two are analytically distinct, but in AI service assessments they are often coupled through the same proxy variable. As a result, uncertainty attributed to the choice of functional unit frequently originates in the choice of allocation approach.

The current section thus starts in Section 5.2.1 with an introduction to functional units, which addresses principles, limitations, and trade-offs, both in general and with specific relevance to AI. Possible functional units for AI usage are then discussed in Section 5.2.2.

5.2.1. Characteristics of a functional unit, and their AI-specific relevance

In this section, we consider the overarching principles that apply to the choice of functional unit in general, and illustrate them in the context of AI services.

In the context of environmental lifecycle assessment (see Appendix A), the functional unit represents a "quantified performance of a product system for use as a reference unit" (ISO 2018). The choice of functional unit needs to be relevant as a *description of the actual use* of the product in the market. For example, when assessing the lifecycle impacts of a passenger car, the typical functional units are vehicle-km or passenger-km. This directly suggests functional units related to user value: per completed user task or per session (in order of relevance). A completed user task could for example be the generation of one minute of video, or of a 500-word page of content (Ren et al. 2024), the translation of one 300-word document, etc.

The functional unit should be *quantitatively* described, including not only h*ow much* (e.g., 500 words), but also at *what quality* the service is provided (e.g., a video at 1080p resolution). Quality can vary between alternative services. These variances could potentially be resolved by calculating quality-weighted scores. However, for many functional units it is challenging to determine which qualities are most relevant, let alone defining a quantitative scale. A possibly simpler approach can be to define a *minimum threshold for acceptable quality*, e.g. using Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores for natural-language generation tasks (C.-Y. Lin 2004).

The functional unit plays an important role as the *basis for comparison* of environmental performance with alternative products or services. A functional unit can support comparability better if it describes *what the user receives* rather than what the system consumes. This can refer for example to "one pair of dried hands" when comparing paper hand towels with electric driers. In the digital domain, this could be e.g. "delivery of 1 hour of HD video content to one end-user device at 1080p resolution" instead of "1 GB of data streamed". This example shows that the quality of the service is relevant when comparing two product systems.

The specificity of a functional unit definition depends on the purpose of a study: If the goal is a comparison of a service with itself (e.g., hotspot analysis and design support of a single product) a less precise, yet consistently interpreted functional unit is sufficient (for example, a functional unit of "1,000 user requests processed"). However, if the goal is to compare different services (e.g., for environmental performance comparison or eco-labelling), then specificity of the functional unit definition is more important. An example for such an FU could be a benchmark test such as "The generation of 1,000 images at a resolution of 1024x1024 pixels, based on a standardised set of text prompts, achieving a minimum CLIP score of 0.3" (where the acronym CLIP stands for "Contrastive Language-Image Pre-training", a measure of how well an image matches a text description)

Further, the definition of a functional unit for comparison must trade-off between *accuracy* (by specifying attributes that are most deterministic on the environmental performance) and *genericity* (as accuracy in a definition narrows the set of use cases that are being covered). For a general-purpose technology such as GenAI that can be used for a great variety of tasks, each task could motivate the definition of a different functional unit. This would increase per-usage assessment accuracy but make generalisations and thus also system-wide assessments extremely challenging. Figure 17 illustrates this trade-off.

For example, when comparing coding agents, a highly comparable functional unit (with narrow scope) could be "The completion of a Python function to sort a list of integers using the Bubble Sort algorithm." And a highly general FU (with broad scope) could be "One hour of active coding assistance." While the former provides precise comparison of energy for this specific task, it ignores how the models handle complex system design or debugging. More importantly, as it is so specific and cannot be generalised to other tasks, its significance beyond very specific contexts is limited. The latter functional unit, on the other hand, has all coding tasks in scope, but is highly uncertain because one user might be writing simple HTML while another could be debugging complex programs, leading to vastly different compute loads for the same functional unit. While generalisable, this FU might not be precise enough to allow comparison between services.

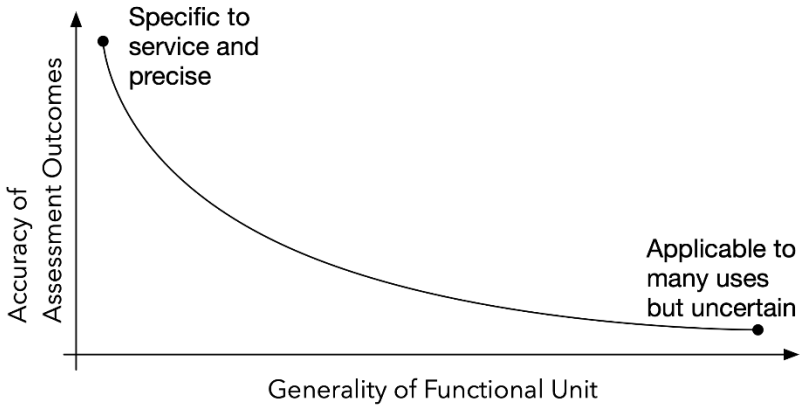


Figure 17: Inherent trade-off between the accuracy of functional units for a specific task and their generalisability.

An additional challenge with specific task-based FUs is that some GenAI sessions do not have a clearly defined measure of completion. By giving up specificity and aggregating impacts of several requests to a *per-session functional unit*, some of those challenges can be overcome. This requires, however, the definition of a typical session, which requires, in turn, operator data that is usually not available for independent third-party assessments (see Section 4.3). This idea of aggregation can be further extended to a full top-down assessment. Then, it is no longer necessary to model the energy consumption of individual requests. Such a fallback on metered energy consumption data is likely simpler and cheaper to undertake but comes at the expense of specificity.

Between two otherwise identical functional units, the one with the more parsimonious set of attributes is preferable. However, the knowledge of what properties are most deterministic of the environmental impact might change as the environmental assessment progresses and should thus be allowed to evolve during an assessment. It is thus useful to approach the choice of FU iteratively and refine it as the domain is better understood.

There is also a practical side to this, as it is often *difficult to quantify a quality attribute*. For example, even if statistically consistent, the user's judgement of the accuracy of a random GenAI response is subjective, and two users might prefer different responses that might appear to have the same factual content. Such quantification of quality is thus exceptionally challenging and subjective and not currently applied in the literature.

Notably, when analysing the benefits of substituting an AI service for another, non-AI service (such as a GenAI service substituting human text generation), the challenges identified above are compounded (Berthelot et al. 2025).

5.2.2. Possible functional units for AI usage

A choice of functional unit is also a pragmatic expression of access to data (see Section 4.3). Here, functional units reflect a trade-off between causal fidelity – how closely the unit aligns with the technical drivers of energy use identified in Section 4.2 – and practical observability for the actor carrying out the assessment. Functional units that are most closely tied to internal execution paths (e.g. directly measured electricity consumption, GPU utilisation, or per-token energy intensity) offer greater accuracy but are typically unavailable to external assessors, whereas service-level or task-based units are more readily observable but mask substantial internal variability. We thus do not propose a single 'best' functional unit but distinguish alternative functional units within this accuracy–disclosure trade-off. To be clear: there is no functional unit whose energy intensity can be robustly estimated without reliance on provider-controlled information.

Table 6 summarises the subsection following below by distinguishing between categories of functional units, with examples and explaining trade-offs.

Table 6: Overview of functional units. Both, bottom-up and top-down metrics can be compatible with LCA standards. Model-centric functional units are frequently used, yet do not relate directly to the use of a service from the user's perspective.

Category	Functional unit	Example	Trade-off
Bottom-up	Per completed task with quality threshold	Per chat response/image/video generated at given minimum quality	Pro: LCA standard Con: Task definitions possibly ambiguous
	Per user session	Per typical user session	Pro: LCA standard Con: Requires operator data to define 'typical'
Top-Down	Per user per time frame	Per active user per month	Pro: Bypass some allocation challenges Con: Not related to intensity of individual use
Model-centric	Per token (output or input)	Per 1000 tokens	Pro: aligned to scaling of infrastructure Con: Not easily relatable for user, tokens are not equal to words
	Per pixel and image/frame	Per 1024x1024 image, per 8 seconds 720p video	Pro: aligned to scaling of infrastructure
	Per quality score point	per point of improvement in BLEU translation quality metric;	Pro: useful for engineering Con: Not straightforwardly relatable for user

While the bottom-up and top-down task- and user-related functional units above suit the requirements of LCA standards best, they are not commonly found in the literature on energy and carbon footprints of GenAI services. Much more prevalent here are model-centric approaches stemming from engineering communities that normalise energy consumption or embodied impacts relative to operational KPIs, such as carbon or energy footprints per-floating point operations (FLOPs) (Schneider et al. 2025), per-inference (Hugging Face 2025) or per-token (or 1000 tokens) for text-based models (Samsi et al. 2023). The latter metric is most straightforward to operationalise. Given the different characteristics between prefill and decode, estimates of per-token energy intensity should be decoded tokens, or better, separately account for energy intensity of input and output tokens. This is further supported by the notable price difference (between 4x to 8x) between input and output tokens (OpenAI 2025).

As we outline in the description of model architectures (Section 4.2), the token output rate is a limiting factor to the performance of LLMs. And energy per-token (and by extension carbon per token) is thus an attractive model benchmark (Ramos 2025). This explains some of its appeal: it is relatively easy to empirically determine in the lab with state-of-the-art hardware. Additionally, modelling energy consumption per-token avoids the messy problem of having to deal with the variable length of input and output token count 'in-the-wild'. Even though per-token functional units are attractive in engineering benchmarks, they presently have limited availability for third-party environmental assessment.

Besides mere access to commercially representative data there are other limitations to this approach. Firstly, per-token energy-intensity is rapidly changing with efficiency improvements in hardware as well as software. Robust estimates of this improvement rates across generations of GPUs, models and middleware (e.g. vLLM optimisations) are, however, absent, but for some sparse examples in the grey literature, such as (L. Lin 2025). This challenge is not unique to per-token estimates. However, given its straightforward reproducibility, it is most obvious here.

Secondly, per-token energy consumption is limited to the raw performance of LLMs and excludes the orchestration that is essential in a production environment. For example, given the increasingly complex scheduling of concurrent inference requests (across GPUs for MoE and while waiting for blocked agents; see section architecture) decoding efficiency depends on intelligent scheduling approaches. Section 4.1

reviewed system architectures such as RAG or orchestration layers, and highlights the need to allocate infrastructure overheads consistently across FUs.

And thirdly, system energy consumption includes overheads from baseline (or “idle”) power consumption which strongly depends on utilisation in a production environment. For example, Open AI has different prices for batch, (high-latency) “Flex”, standard or priority requests (OpenAI 2026) as they vary how they can be scheduled.

The appeal of per-token functional units for text generation might change in the future. For example, agentic AI results in intermediate work (and energy consumption) that might not be tractable with input/output tokens. When evidence grows that the relative balance between the visible and invisible tokens is skewed too much, token-based assessments could become more uncertain and might give way to alternatives.

Based on the way diffusion systems work (Section 4.2), the functional unit for assessments of images and videos is currently best based on resolution, number of frames, and denoising steps. As the technology evolves the semantics of image content might affect denoising steps to a significant degree, possibly motivating a more dynamic representation of user-choices in what image content is being asked for.

Together, these are strong constraints to the use of per-token energy intensity as a functional unit for environmental assessments. While per-token energy intensity is useful to show lower limits of model energy efficiency, and illustrating the raw efficiency improvements over time, per-token energy intensity does not overcome the need for system level data on energy consumption to be reported by commercial operators. So, while lab benchmarks can approximate energy intensity, they are unrepresentative of commercial serving. And aggregated averages, as reported by OpenAI (Altman 2025) or Google (Elsworth et al. 2025) obscure variance and preserve commercial confidentiality and in effect hide the variability that matters for reasoning about environmental impacts.

This is a gap that cannot be filled in the current regime, making the choice of the “right” FU impossible. Based on observations from the domain of general cloud service carbon footprinting we can expect that model providers will become increasingly transparent and granular with reporting the carbon intensity of GenAI services. Around those tools, a cottage industry of hybrid approaches can exist that calibrate proxies based on lab benchmarks and incremental data release by commercial operators.

5.3 Assessment trilemma

The above sections show that the energy and environmental impact of AI services is highly variable and depends on execution details that are largely invisible at the service API. This means it matters strongly who carries out an assessment as privileged data access fundamentally enables the uses of “better” functional units.

Throughout this section, we have characterised LCA assessments as a choice of expressive functional units that provide understanding of task context, allocation approaches that faithfully show variability, and pragmatic access to data.

We can represent this tension as a trilemma (Figure 18) when asking the following questions:

- Granularity (Fidelity to engineering): Do the chosen processes that are modelled and their allocation keys represent the properties that determine the energy consumption or the long-term marginal impact on infrastructure?
- Context (Utility): Does the functional unit represent the `real-world` by measuring the value to a human from a commercially representative system?
- Data Availability (Feasibility): Is the data accessible to the assessor?

We can formulate the tension between pragmatic availability of data, the benefits from representing variability and the need to understand task context for environmental assessments as a trilemma when asking the following questions:

- Granularity (Fidelity to engineering): Do you measure the properties that determine the energy consumption or the long-term marginal impact on infrastructure?
- Context (Utility): Do you represent the `real-world` by measuring the value to a human from a commercially representative system?
- Data Availability (Feasibility): Is the data accessible to the assessor?

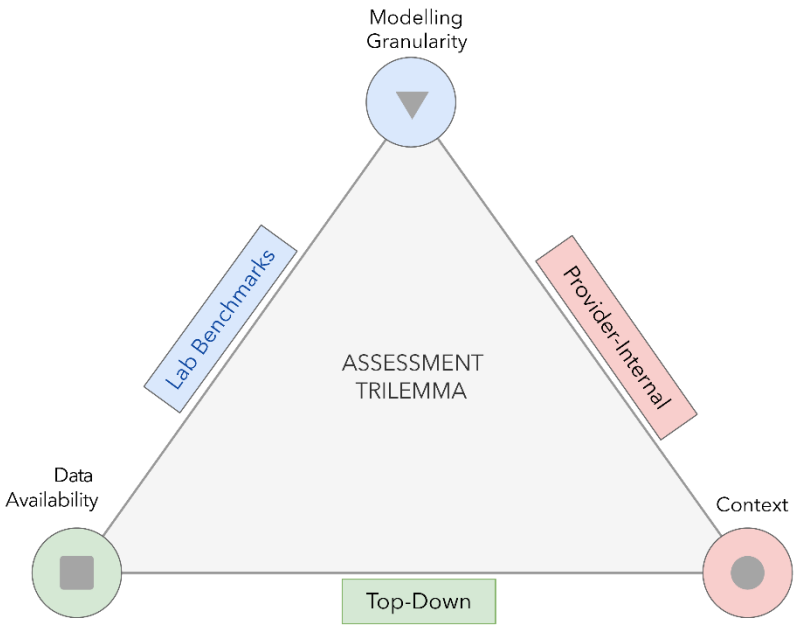


Figure 18: The assessment trilemma. Top: granularity of processes and allocation keys to represent the properties that determine the impacts. Right: context should represent the 'real-world' via functional unit and modelling actual commercial conditions. Left: data availability affects the feasibility.

Any assessment falls into this trilemma. We can illustrate the spectrum of choices via the following examples:

When prioritising modelling granularity and using available data, we could use LCA parameters from lab experiments with open weight models on sample prompts. This would give us the marginal energy consumption per token. However, it lacks context: is not representative of how LLMs are run in commercial environments, nor does it by itself assure that the behaviour of actual users is taken into account.

Alternatively, when prioritising context (i.e. how systems are deployed commercially and how they are actually used) we could use a Top-Down approach and, for example, look at aggregate energy consumption in CSR reports from commercial operators and how they changed between the introduction of GenAI. This is meaningful for macro-economic analysis but does not help to understand how to affect this energy consumption.

Finally, we could aim for maximum granularity as well as full context (how systems are commercially used and how they are used by consumers, or even more so, how the use of an LLM results in substitution or optimisation of other activities – indirect effects) however this is not practically achievable (not pragmatic) because that data is not available.

5.4 Suggested metric and current consumption

The existing literature also struggles with the issues raised above, while aiming to define the most meaningful metric and to subsequently deploy it for concrete assessments. To devise the inference energy of AI models, recent studies have looked at energy *per query* (which can also be referred to as "prompt",

"task", "request", or "inference"). The energy use per query is closely linked to the number of input and output tokens, which are the fundamental drivers of the computational work (and energy) required. However, a "query" can mean many things, with a wide range of input and output tokens.

As discussed below, per-query energy figures can help make the data understandable and relatable for users and policymakers. Their drawbacks, however, are that they hide complexity and nuances, and thus also some of the possible mitigation levers. It can also easily be misused or misinterpreted to understate or overstate the true AI impact. Section 5.4.1 briefly presents relevant literature, while Section 5.4.2 then derives our own proposal.

5.4.1. Deployed metrics and published values

Energy per query

The open source LLM analysed in (Luccioni et al. 2022) has already been introduced in Section 3.4.2. For inference, the small cluster of 16 Nvidia A100 GPUs deployed for inference consumed on average 1664 W. As it served an average of 558 queries per hour, this amounts to **3 Wh / query** on average. These were relatively basic conversational queries; on the other hand, it was also a model deployed relatively early (i.e., 2022) for research purposes, handling a very low number of total requests. The roughly 500 requests per hour amount to about 10,000 daily requests; this is many orders of magnitude lower compared to the billions of requests served daily by large foundational models such as ChatGPT, Gemini, and Claude (Lee 2025). As presented in Section 4.2.2, the latter deploy various optimisations such as batching, which the model analysed by (Luccioni et al. 2022) did not do.

By contrast, Sam Altman claimed mid-2025 that OpenAI requires *on average* **0.34 Wh/query** (Altman 2025). In a recent paper (Elsworth et al. 2025), Google also used energy per query to assess the environmental impact of its Gemini models. The study has comprehensive system boundaries, accounting not only for the main models serving the Gemini app, but also all supporting models for scoring, ranking, or classification. It also takes into the account the rest of the ecosystem and thus not only AI accelerating hardware, but also CPUs, idle machines, and further overhead.

The Google study presents the results for the *median consumption per prompt* as **0.24 Wh / query**, while also stating that "the distribution of energy/prompt metrics can be skewed [...] Part of this skew is driven by small subsets of prompts served by models with low utilization or with high token counts, which consume a disproportionate amount of energy. In such skewed distributions, the arithmetic mean is highly sensitive to these extreme values" (Elsworth et al. 2025). The choice of only presenting the median figure is incomplete and potentially misleading. Gemini can perform many types of tasks, from answering simple text prompts to generating pictures and performing deep research that generates dozens of pages of text. Almost certainly, however, simple prompts are more numerous, which means that the median is most likely the auto-generated AI summary used in Google search (Coroamă 2025a). The more honest interpretation of the resulting value is thus not as "median energy over all of Google's AI queries" but as "average energy for the simplest queries".

To reflect these differences, the "AI Energy Score" benchmark (Hugging Face 2025) evaluates various models based on the same metric (i.e., Wh / 1,000 queries) as well. Next to its primary metric of average energy per 1,000 queries, however, the benchmark also defines 10 types of requests (such as text generation, image classification, or high-definition image generation). The benchmark can then compare the energy performance of various AI models on each of these task categories. To ensure comparability, it uses standardised datasets (e.g., specific sets of 1,000 queries) for each type of task.

(Luccioni et al. 2024) present results of the "AI Energy Score" benchmarking above. The paper shows how the energy required by different types of tasks varies by several orders of magnitude: Text classification and answering questions are the most frugal task types, with an average of only 0.5 Wh and **1 Wh / 1,000 tasks**, respectively. By contrast, with 109 and **477 Wh / 1,000 tasks** on average, image captioning and image generation require 2–3 orders of magnitude more energy and are the most energy-hungry tasks types (among those under scrutiny, that is). But even within the same type of task, the differences can also be significant: For question answering, for example, the spread reached from below 0.1 Wh / 1,000 tasks to over 5 Wh / 1,000 tasks.

Energy per tokens

As discussed earlier, the number of output tokens is a major determinant of energy use per query since the model must generate them sequentially. For a given context, the inference energy correlates roughly linearly to the amount of output tokens (Liu et al. 2024), as each of them needs to be generated while reusing the key value cache, as described in Section 4.2.3.

Seemingly similar to the examples above, Epoch AI uses per-query metric to devise the energy consumption of the GenAI model GPT-4o in February 2025, estimating an energy of about **0.3 Wh / query** (You 2025a). There is an implicit relation to output tokens, however, as the study assumes an *output of 500 tokens*. So the more semantically significant way of conceiving this result is **0.6 mWh / output token**. To compute it, the model estimated the amount of compute required (floating point operations per second, FLOPs) based on the number of active parameters in the model multiplied by the number of tokens generated.

The IEA also notes that generating longer answers requires more compute and is therefore more electricity intensive. Building on third-party estimates such as above and performing plausibility checks with the estimated global AI energy consumption and global number of AI queries, the IEA estimates that text generation might take 0.3 Wh for a small model versus significantly more for larger models or longer outputs (IEA 2025).

Processing input tokens (the “prefill” phase) also consumes energy, which can become an important – and even dominant factor – for queries with very large contexts (e.g., when uploading long documents). While input processing is negligible for short questions, the energy cost scales quadratically with input length. A query with **10,000 input tokens** could increase energy cost to **2.5 Wh**, and **100,000 tokens** to nearly **40 Wh** (You 2025a).

5.4.2. Deriving a token-based energy consumption model for AI inference

Per-query assessments are simple and can be easily conveyed. As argued above, however, they can also be misleading. As SAFE-AI is aimed at a more technical audience, it thus suggests a token-based energy consumption model for AI inference.

Section 5.4.1 argued that both output and input tokens are relevant for the energy consumption of inference. In more detail:

- In the prefill phase, the model must read the entire prompt and build the key value cache for attention. For short contexts, this phase is negligible. Energy costs, however, grow steeply with prompt length, because attention work during prefill scales roughly quadratically with sequence length in standard transformer models (You 2025a; Wang et al. 2025). The marginal energy cost of tokens is thus increasing with context length, optimisations notwithstanding.
- After prefill, output tokens are generated one at a time while reusing the key value cache. As argued above, for a given context, the energy consumption is roughly equal for each new generated output token. However, with longer context, the energy costs still rise as each new token attends over more cached tokens, and memory traffic grows.

Reflecting these considerations, the energy demand for inference can be approximated as

$$E^{inf.}(n_{in}, n_{out}) = E^{pref.}(n_{in}) + E^{dec.}(n_{in}, n_{out}) \approx i * n_{in}^2 + (o * n_{out} + io * n_{in} * n_{out}) \quad (1)$$

where the inference energy $E^{inf.}(n_{in}, n_{out})$ as function of the input (n_{in}) and output (n_{out}) tokens equals the sum of prefill energy $E^{pref.}(n_{in})$ and decode energy $E^{dec.}(n_{in}, n_{out})$. The former correlates approximately quadratically with the number of input tokens, while the second correlates roughly linearly with both the number of output tokens and the product of input and output tokens, as the two jointly make the context grow.

The coefficients i, o, and io are yet to be determined, although some relations are already clear: i, for example, will need to be very small, so that for short contexts, the contribution $i * N_{in}^2$ of input tokens to the overall energy consumption of inference remains negligible as compared to the contribution of the output tokens.

To approximate these coefficients from literature, we proceed as follows:

- According to (You 2025a), a query with 10,000 input tokens could increase energy cost to 2.5 Wh, and one with 100,000 tokens to nearly 40 Wh. It is not clear what the contribution of the output tokens is in these estimates. What is clear, though, is that it is smaller (and probably much smaller) than the 2.5 Wh of a 10,000 token input.

  For the 40 Wh of a 100,000 token input, it is clear that the quadratic term $i * n_{in}^2$ dominates $E^{inf.}(n_{in}, n_{out})$ in Equation 1. If this was the only component of the 40 Wh for $n_{in} = 100{,}000$, this would imply that $i = 4 * 10^{-9} \left[\frac{Wh}{token^2}\right] = 1.44 * 10^{-5} \left[\frac{J}{token^2}\right]$.

  Since the other two terms $o * n_{out}$ and $io * n_{in} * n_{out}$ exist as well, however, and the second one in particular also depends on the large n_{in} (albeit only linearly), a reasonable approximation thus seems

$$i = 3 * 10^{-9} \left[\frac{Wh}{token^2}\right] = 1.08 * 10^{-5} \left[\frac{J}{token^2}\right] \quad (2)$$

- o can be approximated from the value of 0.3 Wh / 500 output tokens generally accepted (Elsworth et al. 2025; You 2025a; IEA 2025) as typical for a very short context today. Given the very short query and tiny value of i computed above, the quadratic term is practically zero, and $io * n_{in} * n_{out}$ will also almost vanish (because io itself is also expected to be much smaller than o. The entire Equation 1 thus reduces to $E^{inf.}(n_{in}, n_{out}) = o * n_{out}$. As the energy is known (0.3 Wh) and N_{out} as well (500), this can be solved to

$$o = 6 * 10^{-4} \left[\frac{Wh}{token}\right] = 2.16 \left[\frac{J}{token}\right] \quad (3)$$

- Finally, coming back to the 40 Wh for a 100,000 input tokens can help addressing the last coefficient io. For this very long context, the quadratic term in Equation 1 is responsible for 30 Wh. EpochAI's example keeps 500 output tokens, so what remains is a linear equation in which all terms but for io are known: $10\,Wh = o * n_{out} + io * n_{in} * n_{out}$, with o computed according to Equation 3 above, $n_{in} = 100{,}000$, and $N_{out} = 500$. Solving it for io yields:

$$io = 1.94 * 10^{-7} \left[\frac{Wh}{token^2}\right] = 6.98 * 10^{-4} \left[\frac{J}{token^2}\right] \quad (4)$$

As expected, the by smallest coefficient is the one of the quadratic equation of the prefill energy. The coefficient of the linear decoding component is 5 orders of magnitude higher, while the coefficient of the term compounding input and output tokens is in between (2 orders of magnitude bigger than i, 3 orders of magnitude below o).

Given their tiny values when expressed in Wh, we suggest using Joules for these coefficients, even though this will imply a conversion to Wh (or kWh, etc) for an overall computation. Section 7.1 showcases how this can be done for a concrete use case. Furthermore, these coefficients – and indeed Equation 1 itself, along with the assumptions it resulted from – are only indicative of a much more complex reality. It is thus not helpful to stick to the precision feigned by the overly precise numbers in

Equations 2–4. We thus suggest rounding those values and their nearest potency of ten, which transforms Equations 1 to

$$E^{inf.}(n_{in}, n_{out}) \approx 10^{-5} \frac{J}{token^2} * n_{in}^2 + 1 \frac{J}{token} * n_{out} + 10^{-3} \frac{J}{token^2} * n_{in} * n_{out} \quad (5)$$

These values are recommended approximations for the coefficients i, o, and io in late 2025 – early 2026; given the rapid efficiency gains in AI, they need to be updated for later studies. With new emerging foundational architectures, the symbolic Equation 1 might also change at a later moment; it should, however, remain longer valid than the values of the coefficients.

5.5 Aggregating external AI usage to AI system level

Section 5.4 derived Equation 5 to approximate the energy consumption of one external API query, as a function of its number of input and output tokens. While all the information needed for the assessment of a single query, this is typically not sufficient for entire AI systems. When introducing the principles of assessment in Figure 5, it has already been stated that the allocation from system to usage level – and, conversely, the aggregations from usage to system level – are not mere multiplications.

In essence, this is what the current section describes: How the aggregation from individual queries to AI system level can be performed for all usages over a time period, which is usually a week or a month (see the discussion on the typical functional unit for AI system in Section 0 below). To this end, Section 5.5.1 first presents a stochastic analysis of external API-based AI queries. Section 5.5.2 continues the analysis to suggest default values for the variances of input and output tokens.

5.5.1. A stochastic analysis of per-query token distribution

When commercially employing API-based external AI services, the AI system provider (and integrator) typically receives from the LLM provider an overview with the number of queries, input tokens, and output tokens over a time period. While the bills are weekly or monthly, the granularity of the overview is normally higher – OpenAI, for example, provides for GPT an hourly disaggregation of queries, input, and output tokens. The granularity, however, is typically not per single query.

Equation 5 is thus not directly deployable in most settings. Even if it were (i.e., per-query logs were available), the complexity of computing the individual energy for tens or hundreds of thousands of prompts is likely not justified, given the prevailing uncertainties as discussed in Section 5.4. On the other hand, having the number of queries as well as input and output tokens over a time period, means that the average number of input tokens and output tokens per query – μ_{in} and μ_{out}, respectively – are trivial to compute.

As Equation 1 shows, however, the energy does not vary linearly with the number of tokens, so working with averages might vastly skew the result, especially if the variance among queries is large. SAFE-AI thus suggests a stochastic analysis.

Across the workload, we treat N_{in} and N_{out} as random variables drawn from some distribution representing the subsequent queries. The expected value for the energy per prompt is noted as $\mathbb{E}[E]$, and represents the average per prompt energy over that distribution (thereby, the bold “E” outside the brackets is the expectation operator). Employing Equation 1, the expected energy can be written as

$$\mathbb{E}[E^{inf.}(n_{in}, n_{out})] = \mathbb{E}[i * n_{in}^2 + (o * n_{out} + io * n_{in} * n_{out})] \quad (6)$$

Given the expected value of a sum is the sum of expectations and that constant coefficients can be brought outside the expectation, Equation 6 can be rewritten as:

$$\mathbb{E}[E^{inf.}(n_{in}, n)] = i * \mathbb{E}[n_{in}^2] + o * \mathbb{E}[n_{out}] + io * \mathbb{E}[n_{in} * n_{out}] \quad (7)$$

Individually, the three terms of the right side of Equation 7, can be treated as follows:

- The middle term, $o * \mathbb{E}[n_{out}]$, is trivial. As it correlates linearly with the output tokens, the expected average value is the average, so it equals to $o * \mu_{out}$.

- To address the first term, we start from the definition of variance for input tokens:

$$\sigma_{in}^2 = \mathbb{E}[(n_{in} - \mu_{in})^2] = \mathbb{E}[n_{in}^2] - 2 * \mu_{in} * \mathbb{E}[n_{in}] + \mathbb{E}[\mu_{in}^2] \quad (8)$$

  Since $\mathbb{E}[n_{in}]$ equals μ_{in} (similar to μ_{out} in the bullet above) and the expectation of a constant (μ_{in}^2) is the constant itself, Equation 8 becomes

$$\sigma_{in}^2 = \mathbb{E}[n_{in}^2] - 2 * \mu_{in}^2 + \mu_{in}^2 = \mathbb{E}[n_{in}^2] - \mu_{in}^2 \quad (9)$$

  Rearranging Equation 9 to solve for $\mathbb{E}[n_{in}^2]$ yields

$$\mathbb{E}[n_{in}^2] = \mu_{in}^2 + \sigma_{in}^2 \quad (10)$$

  which implies that the expected value of the quadratic terms depends on both the mean and the spread, as expected. For the same average number of tokens, more variability increases the energy consumption of the AI system.

- Finally, for the third term, we proceed similarly to above, first introducing the covariance and in the end rearranging the equation to derive $\mathbb{E}[n_{in} * n_{out}]$ as a function of the covariance. The covariance of n_{in} and n_{out} is defined as

$$Cov(n_{in}, n_{out}) = \mathbb{E}[(n_{in} - \mu_{in}) * (n_{out} - \mu_{out})] = \mathbb{E}[n_{in} * n_{out} - \mu_{out} * n_{in} - \mu_{in} * n_{out} + \mu_{in} * \mu_{out}] \quad (11)$$

  As expectations of constants are the constants themselves, and – as argued above – $\mathbb{E}[n_{out}] = \mu_{out}$ and $\mathbb{E}[n_{in}] = \mu_{in}$, Equation 11 can be rewritten as

$$Cov(n_{in}, n_{out}) = \mathbb{E}[n_{in} * n_{out}] - 2 * \mu_{in} * \mu_{out} + \mu_{in} * \mu_{out} = \mathbb{E}[n_{in} * n_{out}] - \mu_{in} * \mu_{out} \quad (12)$$

  Rearranging Equation 12 yields the desired form

$$\mathbb{E}[n_{in} * n_{out}] = \mu_{in} * \mu_{out} + Cov(n_{in}, n_{out}) \quad (13)$$

  Finally, we introduce the correlation coefficient

$$\rho = \frac{Cov(n_{in}, n_{out})}{\sigma_{in} * \sigma_{out}}$$

which normalises the covariance in the interval $\rho \epsilon [-1, 1]$ for later simplicity. With it, Equation 7 can be rewritten as

$$\mathbb{E}[n_{in} * n_{out}] = \mu_{in} * \mu_{out} + \rho * \sigma_{in} * \sigma_{out} \tag{14}$$

Using the remarks from the first bullet as well as the results from Equations 10 and 14 means that Equation 14 can be rewritten as

$$\mathbb{E}[E^{inf.}(n_{in}, n_{out})] = i * (\mu_{in}^2 + \sigma_{in}^2) + o * \mu_{out} + io * (\mu_{in} * \mu_{out} + \rho * \sigma_{in} * \sigma_{out}) \tag{15}$$

The expected value from Equation 15 only depends on

- the generic coefficients i, o, and io, which were extensively discussed in Section 5.4.2,
- the means μ_{in} and μ_{out}, trivially computed from the known overall number of queries and tokens,
- and
    - either the covariance, or
    - (usually preferred) the correlation coefficient ρ as well as the two variances σ_{in} and σ_{out}.

Section 5.5.2 proceeds to discuss the two variances and their correlation coefficient.

5.5.2. Approximating the variances of the token counts and their correlation coefficient

Let n be a token count per query (i.e., either n_{in} or n_{out}). As already known, its mean is $\mu = \mathbb{E}[n]$. As token counts are always positive and often have rights tails (as some queries can get substantially longer than most), we assume a default lognormal distribution of token counts: $n \sim LogNormal(m, s)$, which means that $\ln n$ is normally distributed with mean m and standard deviation s:

$$\ln n \sim N(m, s^2)$$

Hereby, m and s are mean and standard deviation of the normal distribution $\ln n$ (while μ and σ are the mean and standard deviation of the lognormal distribution of tokens n). The connections between them are given by Equations 16 and 17 (Dransfield and Brightwell 2003):

$$\mu = \exp\left(m + \frac{1}{2}s^2\right) \tag{16}$$

$$\sigma^2 = (\exp(s^2) - 1) * \exp(2m + s^2) \tag{17}$$

Squaring both sides of Equation 16 yields

$$\mu^2 = \exp\left(m + \frac{1}{2}s^2\right)^2 = \exp(2m + s^2) \tag{18}$$

and substituting the right side of Equation 18 into Equation 17 leads to

$$\sigma^2 = \mu^2 * (\exp(s^2) - 1) \tag{19}$$

The quantiles of a lognormal distribution, on the other hand, satisfy the following:

$$Q_p = \exp(m + s * z_p)$$

where z_p is the standard normal quantile (i.e., the inverse cumulative distribution function, CDF). For any quantile, the central quantile ratio r is:

$$r = \frac{Q_{1-\alpha}}{Q_\alpha} = \frac{\exp(m + s * z_{1-\alpha})}{\exp(m + s * z_\alpha)} = \exp\big(s * (z_{1-\alpha} - z_\alpha)\big) = \exp(s * \Delta z) \tag{20}$$

Extracting the logarithm on both sides and solving Equation 20 for s yields:

$$s = \frac{\ln r}{\Delta z} \Longrightarrow s^2 = \left(\frac{\ln r}{\Delta z}\right)^2 \tag{21}$$

Using the thus-computed s^2 in Equation 19 leads to the following form

$$\sigma^2 = \mu^2 * \left(\exp\left(\left(\frac{\ln r}{\Delta z}\right)^2\right) - 1\right) \tag{22}$$

In the symbolic equation 22,

- z_α and $z_{1-\alpha}$ are standard normal percentiles; their values can be obtained from tables. For $\alpha = 0.05$, for example,

$$z_{0.05} = \Phi^{-1}(0.05) \approx -1.645, z_{0.95} = \Phi^{-1}(0.95) \approx +1.645 \Longrightarrow \Delta z \approx 3.29 \tag{23}$$

- The central quantile ratio r, on the other hand, must be measured or estimated for the same percentiles. As we assume no per-query logs are available (otherwise, the stochastic analysis would not have been needed), an estimation must be performed. These estimates depend a lot on the application and general guidance is difficult to provide. For a RAG chatbot over internal documents in a company, where the context of up to the last 10 messages is sent, sensible default ranges for the central 90% span could be $r_{in} = 4 - 10$, while the spread for the output should be smaller, $r_{out} = 3 - 6$.

Using as an example $r = 5$ in Equation 22 yields as result

$$\sigma^2 = \mu^2 * \left(\exp\left(\left(\frac{\ln 5}{3.2897}\right)^2\right) - 1\right) \approx 0.2704\,\mu^2 \Longrightarrow \sigma \approx 0.52\mu \tag{24}$$

Using such assumptions of the spread of the lognormal distribution, Equation 22 can be employed to estimate σ_{in} and σ_{in} based on the respective means. They can be then deployed back in Equation 15.

The only remaining parameter is then the correlation coefficient ρ between n_{in} and n_{out}. Usually, there is some weak correlation between the length of the query and the length of the answer. The strength of this correlation, however, also depends on the exact type of application and its usage patterns. A few examples include:

- A relatively low correlation of $\rho = 0.1 - 0.3$ when many queries are lookup questions, in which the answer is short even if context is long, or if the usage context enforces a conciseness policy.

- A high correlation of 0.4 – 0.7 could for example appear if answers often summarise the retrieved context (“explain / synthesise”) or for frequent open-ended questions, which tend to trigger longer, more thorough responses.
- If nothing at all is known about the correlation between queries and answers, a likely solid middle ground is $\rho = 0.3$.

6 Suggested assessment workflow for an AI service provider

Building on the insights and principles from the three previous chapters, the current chapter suggests an assessment workflow for *AI systems*, who may deploy both internal or external ML models. This suggested workflow thus focuses on the middle column from Figure 15, being aimed at AI service providers.

This is the level most in need of principles and guidance: Assessments on this level are required for company reporting, disclosures and communication, possible mitigation measures, comparisons across different systems, and as basis for computing the footprint of individual AI usages. The workflow is discussed in detail throughout this chapter; Figure 19 provides its overview for easier navigation.

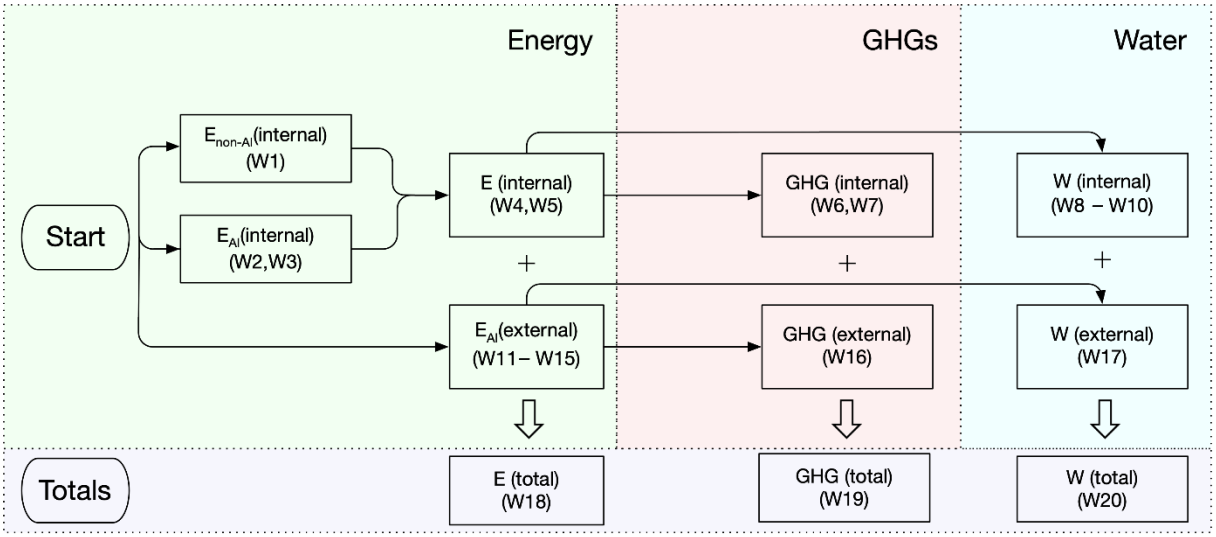


Figure 19: Overview of the SAFE-AI generic assessment workflow of AI systems. Equation numbers in parentheses (W nn) correspond to the workflow equation numbers established along this chapter.

6.1 System boundaries and functional unit

As discussed throughout Section 4.1, AI systems can deploy both internal and/or external AI models. Internal models (i.e. those depicted in blue in Section 4.1) are treated together with the system orchestration and all the other system-internal components (i.e., those depicted in green in Section 4.1) in Section 6.2. API calls to external models belong within the system boundaries of assessment as well; they are discussed in Section 6.3.

The **functional unit** for AI system assessment must reflect overall service provisioning over a reasonable period of time. Due to large inter-daily variations (e.g., between working days and weekends), a day is for most applications too short a period. An year, on the other hand, is often too coarse: systems might substantially change over one year or cease to exist altogether. SAFE-AI thus suggests a week or – in most cases even better – a month as the ideal assessment period. The suggested functional unit for AI system assessment thus is

Functional unit (AI system): Provisioning by one AI system of all its services for one month.

The focus of the primary data collection should lie on the operational (i.e., use phase) footprint of the DC-based servers hosting the AI system. We suggest this principle both for manageability and due to the outstanding importance of these factors along the corresponding two dimensions – environmental lifecycle and device categories.

6.2 Assessing provider-internal components

The assessment of system-internal components is conceptually fairly straightforward, requiring however an understanding of various details and principles discussed in earlier chapters (which will be summarised here) as well as some attention to detail.

As discussed from the very outset in Section 1.2.1, SAFE-AI argues that between energy and GHGs, energy should be the main focus of an assessment, and the corresponding GHGs should be derived at the very end from the energy consumption. Correspondingly, the suggested workflow focuses on energy, while employing location-based carbon intensity values to subsequently compute the related GHGs. Why market-based calculations are not encouraged has been addressed in detail in Section 3.4.3 for AI models; but the arguments listed there are equally valid for AI systems as well.

6.2.1. Energy

The individual assessment steps are as follows:

- Identify all servers (which can be physical servers but are more likely virtual machines) supporting the *AI system*.
    - This should include all internally deployed AI models, the orchestration layer, databases, and all the other ancillary services.
    - It is helpful to request a system architecture first (or, if not available, to develop one in conjunction with the system developers or operators). Having a solid understanding of the architecture prior to the assessment, increases the chances of a comprehensive assessment that does not oversee some of the system components.
    - The number of virtual machines (VMs) supporting the system might be dynamic depending, for example, on the dynamic utilisation of the system (being able to dynamically react to changes in the system's utilisation, avoiding both over- and under-provisioning, being good system design). The assessment should thus either continuously track which servers are being deployed or be able to suggest defensible averages.
    - Finally, while probably not typical, it is conceivable that the servers running the AI system reside in different data centres; for example, the system-internal AI models could reside in an AI-focused DC, while orchestration, databases, and all other modules reside in a regular general-compute DC. The physical boundaries of the assessment should match the logical boundaries of the system architecture.
- Deploy tools for the server-side measurement of energy consumption
    - These can be either physical sensors (e.g., measuring rack-level power drain, if the entire rack is dedicated to the AI system under scrutiny) or virtual sensors that report VM power consumption.
- Sum the individual energy consumptions of all servers dedicated to the AI system over the assessment period
    - Do not forget servers that might have been active for a shorter period within the entire period of assessment.
    - The suggested assessment period is 1 month, as discussed in Section 0.
    - Servers running the internal AI models should be distinguished from the servers running all the other components of the AI ecosystem for reasons that will be clear shortly:

$$E_{FU}\left(\forall Serv_{non-AI}^{op.}\right) = \sum_{\forall i \in \{non-AI\}} \int_{FU\ start}^{FU\ end} P(Serv_i) \qquad (W1)$$

$$E_{FU}\left(\forall Serv_{AI}^{inf.}\right) = \sum_{\forall j \in \{AI\}} \int_{FU\ start}^{FU\ end} P\left(Serv_j\right) \qquad (W2)$$

where the energy consumption of all servers over the period of assessment $E_{FU}(\forall Serv)$ is computed by first integrating the power of each server $P(Serv_i)$ over the time of assessment (i.e., the time period of the functional unit), and then summing up the thus-resulting energies for all servers involved in the system. Thereby, the assessment distinguishes between

- servers running non-AI components (such as the orchestration, databases, and further parts of the ecosystem), and
- servers running the internal AI models, if any.

- Account for the non-deployment phases of internal AI models
  - As shown in Figure 15, the AI service provider (i.e., system developer) can directly assess the deployment phase of the internally deployed AI models. Unless these have been also developed internally (which is rarely the case), there is no visibility – and quite some uncertainty – on the impact of the phases preceding it, as discussed in Section 5.1.2.
  - The discussion in Sections 3.4.1 and 3.4.5 as well as their summary in Table 4, however, show that both training – and potentially even more so derisking training runs as well as basic research and the development of intermediate, unreleased models (which need to be attributed to the released models) – can be substantial and should not be ignored.
  - Given the generally unsolvable epistemic and ontological uncertainties, assumptions are necessary. Given the values in Table 4, SAFE-AI suggests scaling up the AI energy in servers by a factor of 1.5, i.e., adding an overhead of 50% for these preceding AI model lifecycle phases:

$$E_{FU}\left(\forall Serv_{AI}^{op.}\right) = ailc * E_{FU}\left(\forall Serv_{AI}^{inf.}\right) \qquad (W3)$$

where the factor $ailc$ is employed to extrapolate the inference energy of internal AI servers $E_{FU}\left(\forall Serv_{AI}^{inf.}\right)$ to their overall operational lifecycle energy over the entire machine learning lifecycle $E_{FU}\left(\forall Serv_{AI}^{op.}\right)$.

- Add operational non-AI and AI energy
  - Operational non-AI energy and the operational AI energy (including the invisible AI model lifecycle stages computed according to Equation W3) are summed up.
  - The non-AI energy was not scaled up accordingly, since the development of regular software such as databases or applications is generally not such an energy-intensive process as the research and development of AI models

$$E_{FU}(\forall Serv^{op.}) = E_{FU}\left(\forall Serv_{non-AI}^{op.}\right) + E_{FU}\left(\forall Serv_{AI}^{op.}\right) \qquad (W4)$$

- Account for non-IT consumption in data centres
    - For both physical and virtual servers (which themselves run on physical machines in DCs), an overhead needs to be added to account for the non-IT power consumption attributable to the system analysed.
    - In data centres, the power usage effectiveness (PUE) is an established metric that relates the overall power consumption (i.e., IT + non-IT power consumption) of a DC to the IT power consumption. The closer to 1.0 the PUE value is, the less energy is waste in a DC for non-IT tasks, in particular for cooling. Large and efficient colocation DC providers have currently PUE values around 1.2 – 1.25.

$$E_{FU}(AI_{int.}) = E_{FU}(DC^{op.}) = PUE * E_{FU}(\forall Serv^{op.}) \quad (W5)$$

6.2.2. Greenhouse gases

Energetically, these were the most important steps. Now the focus turns to computing the GHG footprint of the internal AI system components.

- Derive data centre GHGs from the energy consumption.
    - If the data centre operator deploys on-site power generation used directly for the DC (this can be either renewable or – as is more and more common for AI data centre operators – power generated in gas turbines (Coroamă and Dumbravă 2026)), the corresponding share must be computed using the specific carbon intensity of that source.
    - For the rest, which is grid-based power, location-based accounting should be used. Section 3.4.3 discussed why market-based accounting is not encouraged. The carbon intensity of the grid mixed is thus used for this share.
    - Strictly speaking, the amount of electricity consumption that has been added in Equation W3 to account for the research and development phases of internal AI models should be treated differently. The third-party training is likely to have occurred elsewhere geographically, so another (e.g., US or worldwide) carbon intensity should be applied for that share of $E_{FU}(DC)$. In practice, however, this imprecision is likely minor compared to the other uncertainties of the workflow, so for simplicity this distinction can be ignored.

$$GHG_{FU}(DC^{op.}) = E_{FU}(DC^{op.}) * \left(s_{on-site} * ci_{on-site} + (1 - s_{on-site}) * ci_{grid}\right) \quad (W6)$$

where the share of power generated on-site $s_{on-site} * E_{FU}(DC^{op.})$ is multiplied by its carbon intensity $ci_{on-site}$, and the rest of electricity supplied by the grid $(1 - s_{on-site}) * E_{FU}(DC^{op.})$ is multiplied by the location-based (i.e., grid) carbon intensity ci_{grid}, the sum of the two yielding the operational carbon footprint of the AI system for the functional unit.

    - The carbon intensities –ci_{grid} and, exceptionally, where appropriate, $ci_{on-site}$ – should be the full lifecycle emission factors, i.e., including the production of the power plants.
- Account for production phase of DC microelectronics and for further device categories
    - As analysed in Sections 3.4.2 and 3.4.4, both the production phase of microelectronics and the end devices (through their production and operation alike) contribute with relatively small amounts to the overall carbon footprint of an AI model.

- Both contribute to the overall energy as well, but to an even lesser part than their GHG share. Hence, they were omitted from the energy discussion in Section 6.2.1.
- The contribution of networks, on the other hand, is negligible for carbon as well (see Section 3.4.4).
- Given their overall relatively modest contribution, SAFE-AI suggests a single scaling coefficient for both the production of DC devices and the contribution of end devices over their entire environmental lifecycle. Currently, a reasonable default for this coefficient is between 1.05 – 1.1.
- In future, either the decarbonisation of the operational phase or much more energy efficient ML models could yield the shares of microelectronics production or of end devices relatively more important. If this happens, this single coefficient should be revisited. It may then be necessary not only to adapt the coefficient but to introduce two distinct ones, for the production of microelectronics and for end devices, respectively. For now, however, using just this single coefficient, the corresponding equation is:

$$GHG_{FU}(AI_{int.}^{\forall LC}) = envlc * GHG_{FU}(DC^{op.}) \qquad (W7)$$

where the coefficient $envlc$ is used to extrapolate the operation DC emissions not only for the entire environmental lifecycle (as its name suggests), but also to other categories of devices – in essence, to end devices. Given the overview from Table 4 and the discussion in Section 3.4.5, currently, a low coefficient of 1.03 – 1.05 seems reasonable. This step can thus also be skipped.

6.2.3. Water

For the assessment of water, the workflow is based on the principles discussed in Section 3.5, and in particular the default values for the WUE and water intensity of electricity from Section 3.5.7.

- Derive data centre water consumption via the WUE from its energy consumption

$$W_{FU}^{on-site}(DC^{op.}) = WUE_{DC} * E_{FU}(DC^{op.}) \qquad (W8)$$

where the water footprint in the DC running the AI system $W_{FU}(DC^{op.})$ equals the DCs average water usage effectiveness along the duration of the functional unit times the entire (i.e., of all servers) energy consumption of the system over that period. If no precise data on the WUE is available, a value of 1.0 (i.e., 1 litre per kWh) is a reasonable default.

- Reflecting Equation W6 for GHGs, derive the water consumption due to electricity production via the location-based water intensity of electricity and the on-site electricity generation (if any)

$$W_{FU}^{indirect}(DC^{op.}) = E_{FU}(DC^{op.}) * \left(s_{on-site} * wi_{on-site} + (1 - s_{on-site}) * wi_{grid}\right) \qquad (W9)$$

where the share of power generated on-site $s_{on-site} * E_{FU}(DC^{op.})$ is multiplied by its water intensity $wi_{on-site}$, and the rest of electricity supplied by the grid $(1 - s_{on-site}) * E_{FU}(DC^{op.})$ is multiplied by the location-based (i.e., grid) water intensity wi_{grid}, the sum of the two yielding the indirect water footprint of the AI system for the functional unit.

If no values are known, $wi_{on-site} = wi_{grid} = 1.8$ are solid default values. In case of on-site electricity production via gas turbines, $wi_{on-site} = 0$ is a good default value, as – despite their many

sustainability flaws – gas turbines at least do typically not induce any water consumption (Coroamă and Dumbravă 2026).

- Finally, add the two together

$$W_{FU}(DC^{op.}) = W_{FU}^{on-site}(DC^{op.}) + W_{FU}^{indirect}(DC^{op.}) \quad (W10)$$

If there is no known on-site power generation (the typical case), Equation W9 becomes a simple multiplication of energy consumption with the water intensity of electricity, the independent computation of on-site and upstream water can be skipped, and Equation W10 directly computed as

$$\boldsymbol{W_{FU}(DC^{op.})} = W_{FU}^{on-site}(DC^{op.}) + W_{FU}^{indirect}(DC^{op.}) = (\boldsymbol{WUE_{DC} + wi_{on-site}}) * \boldsymbol{E_{FU}(DC^{op.})} \quad (W10')$$

6.3 Assessing system-external AI models

The steps in assessing provider-external models are based on the per-usage assessment presented in Section 5.4 and the stochastic analysis used for aggregation on system level from Section 5.5. The current section is written to be self-contained and directly applicable. For a more profound understanding of the logic behind many of its recommendations, refer back to Sections 5.4 and 5.5.

6.3.1. Energy

The individual assessment steps are as follows:

- Identify all external models employed by an AI system over the period of assessment.
  - For each so-identified external model, perform all the subsequent assessment steps.
- Compute the average number of input and output tokens per query, μ_{in} and μ_{out}.
  - From the logs received from the external AI model provider, identify the total number of queries, input tokens, and output tokens via

$$\mu_{in} = \frac{Q}{N_{in}}; \quad \mu_{out} = \frac{Q}{N_{out}} \quad (W11)$$

where Q is the number of queries over the length of the functional unit, and N_{in} and N_{out} are the total number of input and output tokens, respectively.

- Presuming a lognormal distribution (see Section 5.5) for input and output token lengths (N_{in} and N_{out}), estimate the standard variations σ_{in} and σ_{out}.
  - Compute both variances (i.e., the square root of standard deviations) while deploying Equation W12, which repeats Equation 22 from Section 5.5:

$$\sigma^2 = \mu^2 * \left(\exp\left(\left(\frac{\ln r}{z_{1-\alpha} - z_{\alpha}} \right)^2 \right) - 1 \right) \quad (W12)$$

  - In Equation W12:
    - z_{α} and $z_{1-\alpha}$ represent quantiles of a standard normal distribution:
      - for $\alpha = 0.05$, for example, $z_{0.05} \approx -1.645$ and $z_{0.95} \approx 1.645$,
      - for $\alpha = 0.025$, $z_{0.025} \approx -1.96$ and $z_{0.p75} \approx 1.96$.

- r is the corresponding central quantile ratio of query token lengths. For the two examples above, for example, it represents the ratio between the token lengths at the upper and the lower margin of the 90% or 95% central intervals, respectively.
- If no precise data is available, the following intervals can serve as orientation for most applications: $r_{in} = 4 - 10$, $r_{out} = 3 - 6$, with reasonable defaults: $r_{in} = 7$ and $r_{out} = 4$.

- Estimate the correlation coefficient ρ between N_{in} and N_{out}.
  - Usually, a moderate positive correlation between the two can be assumed. A solid default value is $\rho = 0.3$.
  - For short lookup questions or enforced briefness, a lower correlation $\rho = 0.1 - 0.3$ is reasonable.
  - For other tasks such as summarising provided texts without a specified answer length, a higher correlation can be assumed $\rho = 0.4 - 0.7$.
- Using these values, compute the expected inference energy per query over the time of the functional unit via Equation W13, which repeats Equation 15 from Section 5.5:

$$\mathbb{E}[E^{inf.}(n_{in}, n_{out})] = i * (\mu_{in}^2 + \sigma_{in}^2) + o * \mu_{out} + io * (\mu_{in} * \mu_{out} + \rho * \sigma_{in} * \sigma_{out}) \quad (W13)$$

  - As discussed in Section 5.4, good current approximations for the coefficients i, o, and io in Equation W13 are:
    - $i = 10^{-5} \frac{J}{token^2}$
    - $o = 1 \frac{J}{token}$
    - $io = 10^{-3} \frac{J}{token^2}$
- Aggregate this per-query value to all inferences of the model over the functional unit according to the trivial equation:

$$E_{FU}^{inf.}(M_i) = Q * \mathbb{E}[E^{inf.}(n_{in}, n_{out})] \quad (W14)$$

  where M_i is the i[th] external AI model employed by the AI system, and Q the number of queries over the entire length of the functional unit.
- As opposed to the internal components assessed in Section 6.2, many overheads no longer need to be accounted for:
  - As the values of the coefficients i, o, and io coefficients were derived from the analysis of foundational model providers, they entail already the PUE.
  - End devices have already been accounted for while discussing the internal components in Section 6.2; doing so again would lead to double counting.
  - The production of the microelectronics could be accounted for here as well, but especially for frontier models, its contribution will be even tinier than discussed previously; for simplicity, it can thus be ignored here.
  - The training phase is included in some of these numbers, but not in others.
- Account for non-deployment phases of internal AI models

- The only overhead generally not included is that of derisking training runs plus energy expenditure not directly related any released model (e.g., such as basic research and the development of intermediate, unreleased models), as discussed in Sections 3.4.1 and 3.4.5.
- To account for them and thus compute the operational energy for the entire AI model lifecycle of model M_i, we use the same coefficient $ailc$ as already done in Equation W3 for the internal AI models, insofar they exist. As above a good default value is 1.5:

$$E_{FU}^{op.}(M_i) = ailc * E_{FU}^{inf.}(M_i) \quad (W15)$$

6.3.2. Greenhouse gases

Finally, this energy consumption can also be transformed into greenhouse gases. As in Section 6.2, SAFE-AI also suggests the usage of location-based power mix for this transformation. The grid, however, is no longer the one where the AI system provider is located, but where (at least presumably) the external AI model is located. The corresponding equation is:

$$GHG_{FU}(M_i) = ci_{remote} * E_{FU}^{op.}(M_i) \quad (W16)$$

where ci_{remote} is the estimated carbon intensity of the remote grid where the AI model operated. In doubt, this is most likely an US data centre – or perhaps a European or a Chinese one. If absolutely nothing is known, using an average global carbon intensity of electricity – or the average US one – are good proxies.

6.3.3. Water

For external models, nothing is typically known about the external DC hosting them (and much less the DCs that had hosted their training). As neither the WUE nor the water intensity are known, the workflow no longer distinguishes between the two:

- Derive the total water consumption (both direct and indirect) from the electricity consumption of the external models

$$W_{FU}(M_i) = ((WUE_{remote} + wi_{remote}) * E_{FU}^{op.}(M_i) \quad (W17)$$

  where WUE_{remote} and wi_{remote} are the estimated water usage effectiveness of the remote DC and the estimate water intensity of the remote grid, respectively. If no data are known, a sum of 2.8 (1.0 for the WUE and 1.8 for the water intensity) are suitable defaults.

6.4 Bringing it all together

The final steps are to add together the energy and carbon values of the internal and external AI system components (the latter for all external AI models employed). Thus, the total energy of the AI system over the course of the functional unit is

$$E_{FU}(AI_{syst.}) = E_{FU}(AI_{int.}) + E_{FU}(AI_{ext.}) = E_{FU}(AI_{int.}) + \sum_{\forall i \in ext.AI} E_{FU}^{op.}(M_i) \quad (W18)$$

where $E_{FU}(AI_{syst.})$ is the overall energy of the AI system over the functional unit, $E_{FU}(AI_{int.})$ the overall energy of the internal components per Equation W5, and $E_{FU}(AI_{ext.})$ the overall energy of the external AI models employed by the system. Their energy is the sum of the operational energy of each model employed by the AI system, $E_{FU}^{op.}(M_i)$, as per Equation W15.

Similarly, the GHGs of internal system components and external models also need to be added together. The total GHG emissions of the AI system over the course of the functional unit are:

$$GHG_{FU}(AI_{syst.}) = GHG_{FU}(AI_{int.}^{\forall LC}) + GHG_{FU}(AI_{ext.}) = GHG_{FU}(AI_{int.}^{\forall LC}) + \sum_{\forall i \in ext.AI} GHG_{FU}(M_i) \qquad (W19)$$

where $GHG_{FU}(AI_{syst.})$ are the overall GHG emissions of the AI system over the functional unit, $GHG_{FU}(AI_{int.}^{\forall LC})$ the overall energy of the internal components per Equation W10, and $GHG_{FU}(AI_{ext.})$ the overall energy of the external AI models employed by the system. Their GHGs are the sum of the GHGs of each model employed by the AI system, $GHG_{FU}(M_i)$, as per Equation W16.

Finally, the water of internal system components and external models is computed similarly:

$$W_{FU}(AI_{syst.}) = W_{FU}(DC^{op.}) + W_{FU}(AI_{ext.}) = W_{FU}(DC^{op.}) + \sum_{\forall i \in ext.AI} W_{FU}(M_i) \qquad (W20)$$

where $W_{FU}(AI_{syst.})$ is the overall water consumption of the AI system over the functional unit, $W_{FU}(DC^{op.})$ the overall water consumption of the internal components as we Equation W10, and $W_{FU}(AI_{ext.})$ the overall water consumption of the external AI models employed by the system. This second part equals the sum of the individual water consumptions of each model employed by the AI system, $W_{FU}(M_i)$, as per Equation W17.

7 Use cases

The development of the SAFE-AI framework was accompanied by two use cases. These were meant from the outset to validate the conceptual framework, demonstrating its application and revealing potential weaknesses and required corrections or additions. Especially the first use case presented in Section 7.1 did so exceptionally: By providing detailed access to its architecture and fine-granular data, it helped to better highlight the factors determining environmental outcomes and develop a more practical framework for future concrete assessments.

The two use cases had an additional benefit: They were used not only to validate the framework for direct effects, but also to put them in ad-hoc relation to possible indirect effects of AI. The distinction between the two types of effects was made in the very beginning of this report, in Section 1.2.1. That section also argued how uncertain indirect effects are. Methods for their robust assessment do not exist yet (Bieser et al. 2024), and the development of such methods would be far beyond the scope of this report. Despite all the caveats, ad-hoc assessments for the relatively limited scope of the two use cases are nevertheless possible, as will be shown below.

7.1 Zü-Re: Sustainability chatbot of the city of Zurich

The city of Zurich ("Stadt Zürich") recently expanded its support for the public via a sustainability-oriented chatbot called Zü-Re (Stadt Zürich 2025). The chatbot, which was launched in March 2025, can answer a variety of questions by city residents that are relevant to both the circular economy and local businesses. These may revolve around e.g. sharing, passing things on, or repairing, such as pinpointing towards the next second-hand clothing store, bicycle repair shop, or the citywide sharing app.

We perform the assessment of this system's energy consumption and GHG emissions according to the workflow put forward in Chapter 6. As suggested there, the functional unit is the provisioning of all the chatbot's services over one month. Although we followed the deployment for several months (between March and October 2025), we chose as period of assessment the month of July 2025.

Section 7.1.1 presents the first step of assessment, the identification of services. Section 7.1.2 presents an overview of the system's usage, which motivates the system architecture, makes its understanding simpler, and serves as sanity check for the assessment model. Section 7.1.3 presents the assessment of the system-internal components of the ecosystem (i.e., orchestration and all ancillary layers); as will be discussed, there are no internal AI models to assess. Section 7.1.4 then presents the assessment of the externally employed AI models, and Section 7.1.5 brings it all together.

7.1.1. Chatbot architecture

As discussed in Section 6.2.1, the first step of assessment is the identification of all servers supporting the AI system. We did not have a system architecture from the outset. In 2 workshops and several further calls with the system's designer, we thus developed a far-reaching architecture of the system's architecture. Understanding the importance of developing such detailed understanding informed the corresponding SAFE-AI recommendation.

The Zü-Re chatbot has been implemented by an external partner on behalf of the city of Zurich, who also hosts the system. Its main components are shown in Figure 20. It is a typical RAG architecture, reflecting the principles outlined in Section 4.1.2 and Figure 12. Compared to that relatively generic introduction, Figure 20 shows more details, which are necessary for a rigorous assessment. The authors of the report, in fact, had access to even more architectural details; the report strikes a balance between protecting the intellectual property of the system designers, while at the same time revealing enough details to warrant transparency and reproducibility).

Zü-Re's architecture consists of several main components. They are organised and presented below according to the layer-architecture of the developers, with the colours in Figure 20 chosen to fit the generic architecture in Figure 12.

- Layer 1 contains generic security and decryption components, which are shared among the numerous apps hosted by the developer, whether they are own apps or developed for customers such as the city of Zurich. The services share two virtual machines (VMs) and comprise:
    - a generic firewall (pfSense),
    - an SSL proxy handling decryption,
    - a more specific web application firewall, which analyses the decrypted packets, and
    - a load balancing component among subsequent VMs.
- Layer 2 comprises the “main app”, which equates to the orchestration layer of an AI system, as defined in Section 4.1. The main app orchestrates the application flow, communication with databases (DBs) and LLMs as well as a couple of further internal services. Flexibly deployed on 1 – 10 VMs, depending on demand (i.e., amount of user requests), this layer 2 comprises next to the main app:
    - the in-session storage, which memorises the chat history between user and LLM and appends the last 20 messages to any new request,
    - services that keep statistics and switch between the different languages the bot can understand,
    - a module that generates the website (e.g., the welcome message, which is each time slightly different), and
    - a crawler that searches for relevant information in internal documents but can also perform web searches, if required.
- Layer 3 contains the vector DB crucial for the RAG approach, and which is so efficient that 1VM is sufficient. In the context of Zü-Re, its usage was as follows:
    - the initialisation of the vector DB, i.e., the processing of the initial chunks, took around some 4h of dedicated VM usage, while
    - later regular additions to the vector DB run in parallel with its usage to retrieve relevant context on the same VM.
- Layer 4 harbours a MySQL DB, which stores website content and configurations, various backend statistics, the history of chats, etc. Given the more data-heavy nature of SQL DBs, it is designed for 4-20 VMs.

All these components run on the servers of the development partner, which are physically located in a colocation DC in the greater Zurich area. This influences the grid mix, as will be discussed below.

Additionally, there are two LLMs being used by the main app when necessary, a lighter one for simple requests and a larger one for more complex requests. At the moment of writing, these two are GPT-4o mini and GPT-4.1, respectively. Their assessment is presented when discussing the assessment of external AI models in Section 7.1.4.

7.1.2.System usage

The following is a short and simplified overview of the system’s usage, which focuses on the processes relevant to the overall energy consumption. This step is not explicitly required by the assessment workflow in Chapter 6. It is also not necessarily required for an energy or environmental assessment. It does, however, belong to the understanding of the system’s architecture and its typical usage, and is thereby an implicit sanity check while performing the assessment.

Additionally, understanding the system’s usage patterns of the two employed external models, will help shaping in Section 7.1.4 the assumptions for the variances of each these two.

1. When a user opens the chatbot’s URL, the static part of the webpage is loaded from the corresponding Layer 2 server. The welcome message, however, is not static but generated using

GPT-4.1, as the chatbot's operators wanted it to appear fresh and each time different. Figure 21 provides an example.

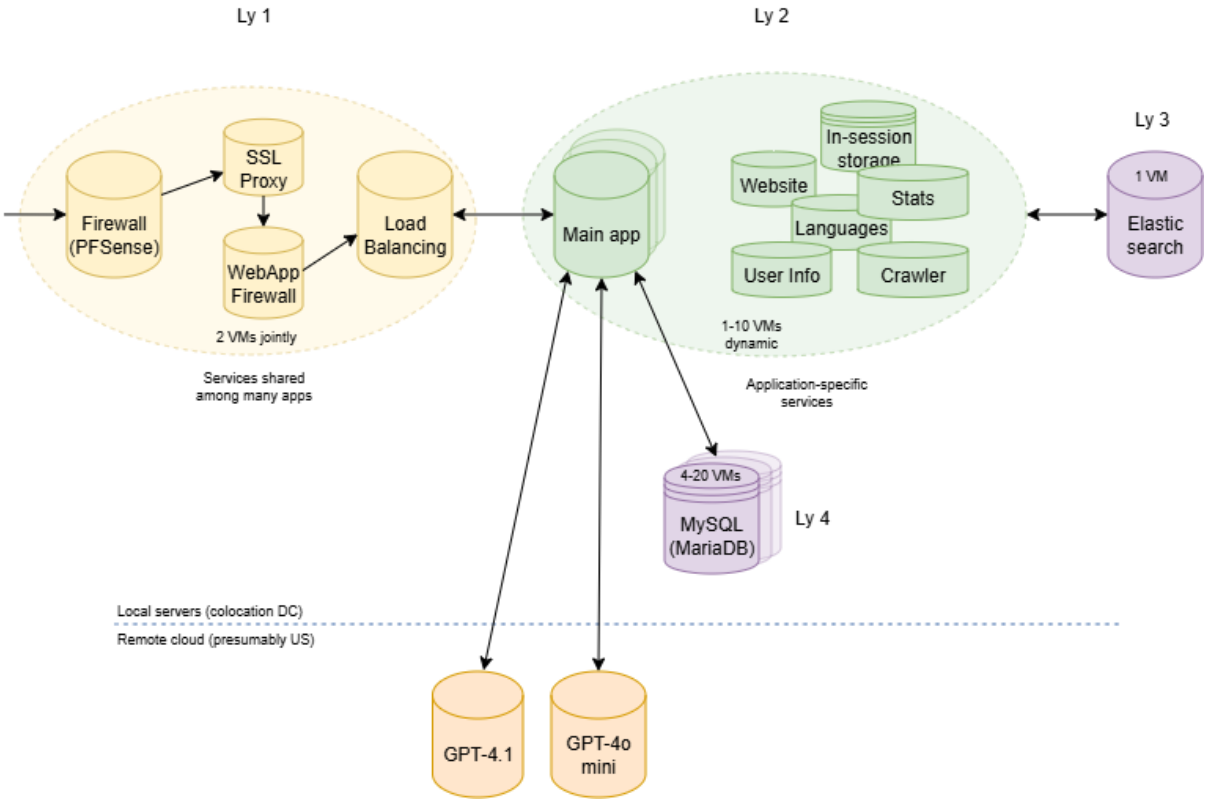


Figure 20: Overview of the RAG architecture surrounding the Zü-Re sustainability chatbot of the city of Zurich.

2. After the user's request, the system categorises the conversation into one of several conversation types, which all induce specific system behaviours. Conversation types include, for example,
    a. Smalltalk, when the user for example asks "what's up?"
    b. Question about the chatbot itself, e.g. "which questions can you help me with?"
    c. Technical questions, which are the chatbot's actual aim, such as "What should I do with my old smartphone?" or "Are there still shoemakers who know to replace worn-out shoe soles?"
3. Depending on the likely conversation type, the main app exhibits different behaviours. From the examples above,
    a. Smalltalk is directed to GPT-4.1.
    b. Questions about the chatbot itself are directed to Elasticsearch.
    c. Technical questions, meanwhile, require a more complex treatment:
        i. To guarantee an exhaustive search, search terms and the central topics are extracted from the user's query, translated if necessary (Layer 2 language server) and are sent to GPT-4o mini for synonym searches.
        ii. The vector DB is then consulted for relevant answers, i.e., answers that are semantically similar to either the original terms or their synonyms.

iii. Finally, the filtered list of answers is sent together with the chat history to GPT-4.1 to formulate an answer.

4. After the answer is being sent to the user, the chat continues back from step 2 if the user desires to go on.

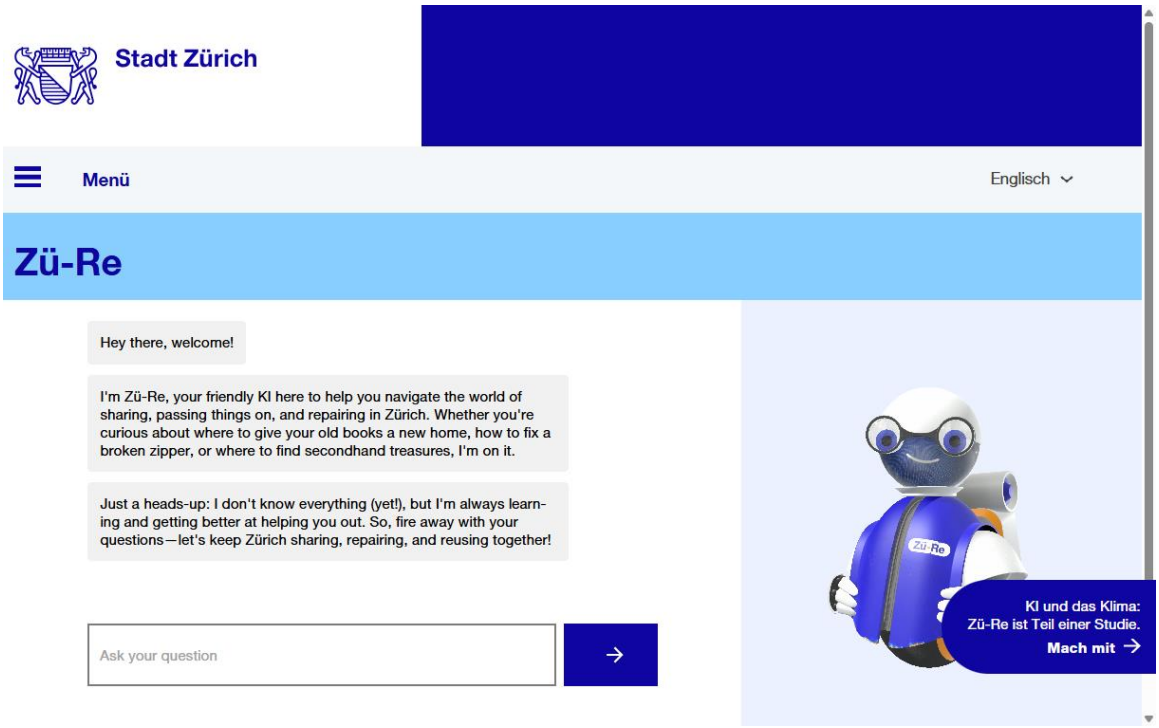


Figure 21: Example for one possible welcome screen of the ZüRe chatbot pilot project. Between April – October 2025, the text in the lower right corner was raising the users' attention to our study, encouraging them to take part.

7.1.3. Assessing Zü-Re-internal components

As described in the beginning of this chapter, we applied the SAFE-AI framework to estimate the direct energy impact of the Zü-Re chatbot during one month of usage, namely July 2025. As Section 6.2.1 recommends, power consumption and usage data were obtained from monitoring tools installed on the VMs as follows:

- access logs from Nginx and Apache servers for the number of requests for each architecture layer,
- pf Logs for the firewall,
- vnStat for the traffic volume,
- collectd for CPU, RAM, disk I/O and network throughput, and
- powertop and powercap for the VM power consumption.

For Layer 1 of the architecture (see Figure 20), power consumption data was not available. However, the generic security and load balancing services on Layer 1, which run on 2 VMs, are shared among dozens of services of the software partner who developed the chatbot for the city of Zurich. As already discussed in Section 2.2 and shown in Figure 5, their contribution is thus likely marginal is neglected.

The rest of the collected data resulted in the inventory presented in Table 7. Although able to accommodate more intense usage (thus foreseeing up to 10 VMs for the main application / orchestration layer and up to 20 VMs for the SQL database), the system required over the course of the entire month only its minimum design hardware (i.e., 1 VM for orchestration and 4 VMs for the SQL DB).

As per workflow step W1, the total energy consumption of all 3 layers over this month was 448.84 kWh. Attributing a small consumption to the (non-measured but shared among many different applications, as discussed above) Layer 1 yields the following concrete instantiation UC1 (i.e., use case equation 1) of W1:

$$E_{07/25}\left(\forall Serv_{non-AI}^{op.}\right) = 450\ kWh \qquad (UC1)$$

Table 7: Attributes of the Zü-Re AI ecosystem: the number of VMs the system was designed for and the actually required ones, the number of internal requests per layer (corresponding to 1107 chat occurrences over the time of monitoring), and the total energy consumption of all VMs in each layer. All data refers to July 2025.

	Layer 2	Layer 3	Layer 4
# VMs possible (design)	1 – 10	1	4 – 20
# VMs actual	1	1	4
# requests for layer	105,115	63,069	147,161
Compute energy [**kWh**]	**57.70**	**122.93**	**268.21**

As the system's architecture in Figure 20 shows, the Zü-Re chatbot does not employ any internal AI models. The instantiations of W2 and W3 (i.e., the expansion of internal AI model inference to the entire operational energy) are thus trivial:

$$E_{07/25}\left(\forall Serv_{AI}^{inf.}\right) = 0 \qquad (UC2)$$

$$E_{07/25}\left(\forall Serv_{AI}^{op.}\right) = ailc * E_{07/25}\left(\forall Serv_{AI}^{inf.}\right) = ailc * 0 = 0 \qquad (UC3)$$

Correspondingly, the W4-sum of internal non-AI and AI energy equals the non-AI energy:

$$E_{07/25}(\forall Serv^{op.}) = E_{07/25}\left(\forall Serv_{non-AI}^{op.}\right) + 0 = 450\ kWh \qquad (UC4)$$

Instantiating W5 for a PUE value of 1.2, which is typical for large data centres around Zurich (Coroamă 2025b) yields:

$$\boldsymbol{E_{07/25}(AI_{int.})} = E_{07/25}(DC^{op.}) = PUE * E_{07/25}(\forall Serv^{op.}) = 1.2 * 450\ kWh = \mathbf{540\ kWh} \qquad (UC5)$$

As for the GHGs of system-internal components, to our knowledge, there is no on-site power production in the corresponding colocation DC. Thus, when instantiating W6, $s_{on-site} = 0$ and $1 - s_{on-site} = 1$. Swiss electricity production is very clean. All sources (i.e., nuclear, large hydro, and further renewables) are low-carbon, and the carbon intensity of production (from a lifecycle perspective) is only around 20 g CO_2 / kWh. The consumption mix, however, also contains some imported electricity, so the location-based carbon intensity of consumption (i.e., ci_{grid} in W6) is: $ci_{CH} = 59\ g\ \frac{CO_2}{kWh}$.

Together with the result from UC5, this leads to

$$GHG_{07/25}(DC^{op.}) = E_{07/25}(DC^{op.}) * ci_{CH} = 540\ kWh * 59\ g\ \frac{CO_2}{kWh} = 31.86\ kg\ CO_2 \qquad (UC6)$$

Instantiating W7 to also account for the small contribution of end devices yields

$$\boldsymbol{GHG_{07/25}\left(AI_{int.}^{\forall LC}\right)} = envlc * GHG_{07/25}(DC^{op.}) = 1.05 * 31.86 = \boldsymbol{33.45\ kg\ CO_2} \qquad (UC7)$$

As for water, there is even less information available, so we use the default $WUE = 1.0$ and $wi = 1.8$ as described in Section 6.2.3. As there is also no on-site generation, Equation W9 is as simple as Equation 8, yielding:

$$W_{07/25}^{on-site}(DC^{op.}) = WUE_{DC} * E_{07/25}(DC^{op.}) = 1.0\frac{l}{kWh} * 540\ kWh = 540\ litres \qquad (UC8)$$

$$W_{07/25}^{indirect}(DC^{op.}) = wi_{grid} * E_{07/25}(DC^{op.}) = 1.8\frac{l}{kWh} * 540\ kWh = 972\ litres \qquad (UC9)$$

These two now need to be added together according to W10. Alternatively, Equation W10' could have been applied directly:

$$\boldsymbol{W_{07/25}(DC^{op.})} = (WUE_{DC} + wi_{on-site}) * E_{07/25}(DC^{op.}) = 2.8\frac{l}{kWh} * 540\ kWh = \boldsymbol{1,512\ litres} \qquad (UC10')$$

7.1.4. Assessing external LLM models

As shown in Figure 20 and discussed in Section 7.1.2 while presenting the system's usage, the Zü-Re chatbot uses two external LLMs, GPT-4.1 for more complex tasks and GPT-4o-mini for simpler ones. This fulfils the first assessment step defined in Section 6.3.1, the identification of all external models employed by the AI system.

As further discussed in Section 6.3.1, the next step is gathering the total number of queries, input tokens, and output tokens over the course of the functional unit, for all externally employed models. Table 8 shows this overview.

Table 8: The two models employed by the Zü-Re chatbot, and their total number of queries, input and output tokens during July 2025.

	Queries [Q]	Input tokens [N_{in}]	Output tokens [N_{out}]
GPT-4.1	27,705	35,316,734	3,859,335
GPT-4o-mini	7,966	1,288,763	99,299

As requested by SAFE-AI, we now perform the assessment of each of these models. For efficiency, we discuss both in parallel, expanding the indices in Equations W8 – W13 to also show which of the two models they refer to, "4.1" or "4o".

First, we deploy the values in Table 8 to compute the means of tokens per query according to W11:

$$\begin{aligned} \mu_{in,4.1} &= 1275; \quad \mu_{out,4.1} = 139 \\ \mu_{in,4o} &= 162; \quad \mu_{out,4.1} = 12 \end{aligned} \qquad (UC11)$$

To compute the standard deviations, we proceed as follows:

- As suggested in Section 6.3.1, we presume lognormal distributions of tokens for both models, and both for input as well as output tokens.
- We choose $\alpha = 0.05$, thus looking at the distribution of the 90% central intervals in both cases.
- Correspondingly, in both cases, $z_{0.05} \approx -1.645$ and $z_{0.95} \approx 1.645$, and thus $\Delta z \approx 3.29$.

Additionally, as described in in Section 7.1.2

- and also shown in Table 8 through its much more intense usage, GPT-4.1 is the main model employed by Zü-Re, in particular for formulating both meaningful, system-intended answers as well as for small talk. It is also used to generate the welcome page. There is thus a fair amount of spread in its output tokens, and we thus presume $r_{out,4.1} = 4.5$.
- Furthermore, GPT-4.1 inputs are both RAG-based and receive from the session storage up to the last 20 conversations. This could indicate quite a large spread of input tokens. However, looking at statistics of the chat logs, it became evident that most chats were only a few interactions long. All considering, we presume a larger, but relatively moderate $r_{in,4.1} = 6$.
- By contrast, GPT-4o-mini is mainly used for retrieving synonyms to keywords identified in the users' queries. This makes both the inputs and the outputs quite uniform and the choose low values for both, i.e., $r_{in,4o} = 2$ and $r_{out,4o} = 1.5$.

Using all these values in W12 yields:

$$\sigma_{in,4.1}^2 = 1275^2 * \left(\exp\left(\left(\frac{\ln 6}{3.29}\right)^2\right) - 1\right) \cong 561285 \Longrightarrow \sigma_{in,4.1} \cong 749$$
$$\sigma_{out,4.1}^2 = 139^2 * \left(\exp\left(\left(\frac{\ln 4.5}{3.29}\right)^2\right) - 1\right) \approx 4491 \Longrightarrow \sigma_{out,4.1} \approx 67$$

$$\sigma_{in,4o}^2 = 162^2 * \left(\exp\left(\left(\frac{\ln 2}{3.29}\right)^2\right) - 1\right) \cong 1191 \Longrightarrow \sigma_{in,4o} \approx 34.5$$
$$\sigma_{out,4o}^2 = 12^2 * \left(\exp\left(\left(\frac{\ln 1.5}{3.29}\right)^2\right) - 1\right) \approx 2.2 \Longrightarrow \sigma_{in,4o} \approx 1.5 \qquad (UC12)$$

Furthermore, for GPT-4.1, we presume the default the correlation coefficient ρ between N_{in} and N_{out} as suggested in Section 6.3.1, $\rho_{4.1} = 0.3$. As inputs and outputs for GPT-4o-mini are not only more homogeneous but also quite correlated, we assume a much higher correlation coefficient $\rho_{4o} = 0.7$.

After these considerations, we are now able to instantiate W13 for the two models:

$$\mathbb{E}\left[E_{4.1}^{inf.}\right] = 10^{-5} * (1275^2 + 749^2) + 1 * 139 + 10^{-3} * (1275 * 139 + 0.3 * 749 * 67) \approx 353\,J$$
$$\mathbb{E}\left[E_{4o}^{inf.}\right] = 10^{-5} * (162^2 + 34.5^2) + 1 * 12 + 10^{-3} * (162 * 12 + 0.7 * 34.5 * 1.5) \cong 14\,J \qquad (UC13)$$

Before continuing with the assessment, a few characteristics of the results in UC13 are worth highlighting:

- In this use case, even the energetically more expensive LLM queries are with about 353 Joules equivalent to approximately 0.1 Wh. This is on average three times less energy than the average 0.3 Wh estimated for the same GPT model by (You 2025a). Given that the average number of GPT-4.1 output tokens in our use case is $\mu_{out,4.1} = 139$, and thus about 3.5 times shorter than the 500 output tokens considered in (You 2025a), comes as a confirmation of the robustness of our model.
- There is a 25-fold difference between the more energetically expensive GPT-4.1 queries and the energetically less expensive GPT-4o-mini queries. This shows that brevity and homogeneity of queries are indeed energy consumption mitigation methods. They are, of course, not always possible.
- For these models, energy is dominated by output tokens. As shown in the upper row of UC13, from the 353 Joules of the GPT-4.1 query,
    - 139 Joules come from the linear term $o * \mu_{out}$,

- another 192 Joules are due to the combination of input and output tokens, i.e., $io * (\mu_{in} * \mu_{out} + \rho * \sigma_{in} * \sigma_{out})$, itself dominated by the mean and standard deviation of the output tokens,
- while only 22 Joules are due to the quadratic $i * (\mu_{in}^2 + \sigma_{in}^2)$ term.

Coming back to the assessment workflow, the AI usage energy consumptions can now be aggregated to the functional unit while instantiating W14:

$$E_{07/25}^{inf.}(GPT-4.1) = Q_{4.1} * \mathbb{E}\left[E_{4.1}^{inf.}\right] = 27{,}705 * 353\,J \cong 9{,}780{,}000\,J \approx 2.7\,kWh$$
$$E_{07/25}^{inf.}(GPT-4o) = Q_{4o} * \mathbb{E}\left[E_{4o}^{inf.}\right] = 7{,}966 * 14\,J \cong 112{,}000\,J \cong 0.031\,kWh \quad (UC14)$$

Accounting via W15 for the overhead of research and unreleased models yields the final energy result:

$$\boldsymbol{E_{07/25}^{op.}(GPT-4.1)} = 1.5 * E_{FU}^{inf.}(GPT-4.1) = 1.5 * 2.7\,kWh \cong \boldsymbol{4\,kWh}$$
$$\boldsymbol{E_{07/25}^{op.}(GPT-4o)} = 1.5 * E_{FU}^{inf.}(GPT-4o) \cong 1.5 * 0.03\,kWh = \boldsymbol{0.045\,kWh} \quad (UC15)$$

For the transformation to GHGs, we use the average carbon intensity of the US grid, as instructed in Section 6.3.2. Given the models deployed were both OpenAI's GPT models, it stands to reason to assume the inference takes place in US-based data centres. The corresponding carbon intensity is $ci_{US} = 385\ g\frac{CO_2}{kWh}$. Using this value in W16 yields

$$\boldsymbol{GHG_{07/25}(GPT-4.1) = ci_{US} * E_{07/25}^{op.}(GPT-4.1) = 385\ g\frac{CO_2}{kWh} * 4\,kWh \approx 1.5\,kg\,CO_2}$$
$$\boldsymbol{GHG_{07/25}(GPT-4o) = ci_{US} * E_{07/25}^{op.}(GPT-4o) = 385\ g\frac{CO_2}{kWh} * 0.045\,kWh \approx 0.017\,kg\,CO_2}\ (UC16)$$

To derive the water consumption and as nothing is known of the remote DCs running the models, we use W17 with the default values $\boldsymbol{WUE = 1.0}$ and $\boldsymbol{wi = 1.8}$ as described in Section 6.3.3:

$$\boldsymbol{W_{07/25}(GPT-4.1) = ((WUE_{remote} + wi_{remote}) * E_{07/25}^{op.}(GPT-4.1) = 2.8\frac{l}{kWh} * 4\,kWh = 11.2\,l}$$
$$\boldsymbol{W_{07/25}(GPT-4o) = ((WUE_{remote} + wi_{remote}) * E_{07/25}^{op.}(GPT-4o) = 2.8\frac{l}{kWh} * 0.045\,kWh = 0.1\,l}\ (UC17)$$

7.1.5. Overall direct effects of Zü-Re

Finally, the results of the internal components and externally employed LLMs need to be added together. Section 6.4 provided the straightforward way of doing so:

- For energy, W17 needs to be instantiated as follows:

$$\boldsymbol{E_{07/25}(AI_{syst.})} = E_{07/25}(AI_{int.}) + \sum_{\forall i \in ext.AI} E_{07/25}^{op.}(M_i) = 540\,kWh + 4kWh = \boldsymbol{544\,kWh} \quad (UC18)$$

- Similarly, for GHGs, W19 is instantiated:

$$\boldsymbol{GHG_{07/25}(AI_{syst.})} = GHG_{07/25}(AI_{int.}) + \sum_{\forall i \in ext.AI} GHG_{07/25}^{op.}(M_i) = (33.5 + 1.5)kg\,CO_2 = \boldsymbol{35\,kg\,CO_2}\ (UC19)$$

- Finally, for water, W20 is instantiated:

$$W_{07/25}(AI_{syst.}) = W_{07/25}(DC^{op.}) + \sum_{\forall i \in ext.AI} W_{07/25}(M_i) = (1{,}512 + 11)\ litres \sim 1{,}500\ litres \qquad (UC20)$$

7.1.6. Interpretation

Quite a few interesting insights can be gained from this case study:

- For Zü-Re, the overall energy and GHG impact are dominated by the orchestration ("main app") and all the other local ecosystem components. They amount to **more than 99% of total energy consumption** and, despite the dirtier electricity used by the external models, still to more than 95% of overall GHGs. And this although the SAFE-AI framework postulates a 50% overhead for the footprint of AI models (whether internal or external) to account for basic research and the development of intermediate, unreleased models.
- This ratio can partly be explained by the relatively under-utilised (and thus inefficient) local components, while the remote foundational models are highly efficient. As discussed in Section 7.1.3, although the Zü-Re system was designed for 1–10 VMs on Layer 2 (main app) and 4–20 VMs on Layer 4 (SQL DB), the minimum of 1 and 4 VMs, respectively, were sufficient. And even these servers were likely an overprovision: Over the analysed period (July 2025), 1107 meaningful chats were held with the system, i.e., a daily average of about 35-36 chats. While API calls to the remote models scale linearly with usage, the minimum fixed costs of the local components were amortised over a relatively low usage.
- Nevertheless, this case study shows that only taking into account the impact of the AI models and ignoring the rest of the AI ecosystem can lead to massive understatements of the overall impact – for Zü-Re, this would have been 2 orders of magnitude of understatement. The SAFE-AI focus on the entire AI system thus appears justified.
- With about 550 kWh, Zü-Re did not consume a substantial – but also not a negligible – amount of energy over one month. Thanks to the very clean Swiss electricity mix (even from a consumption perspective, which includes some dirtier imported power), the 35 kg CO_2, on the other hand, are arguably entirely negligible.
- More relevant, perhaps, is the footprint per chat, which is not to be mistaken for an external AI query. A chat is an entire chat session by one user. Over the course of the functional unit (i.e., July 2025), there were 1107 meaningful chat sessions with Zü-Re. "Meaningful" is hereby defined as more than just writing "hello" or some gibberish into the chat window. Distributing the monthly footprint over these chats means about 0.5 kWh / chat session and about 3 g CO_2/chat session.
- As nothing is known about the WUE and the water intensity of electricity either locally or remotely, the same default values have been used in both. Unsurprisingly then, the same ratio of electricity consumption propagates for the water consumption, with the local component dominating. This result is to be taken with two grains of salt: one, due to these insecurities. The other, because water consumption has a local impact relevant in the context of possible water scarcity, as argued in Section 3.5.5. The same consumption in water-rich Switzerland (locally sometimes called "the water castle Switzerland") is very different from some DC in, say, Arizona.
- In step UC6 above, the lifecycle carbon intensity of the Swiss grid mix of 59 g CO_2 / kWh was used, according to the principles of using location-based and not market-based values. The question, however, is what the granularity of locality should be. The local DC is located within the Swiss canton of Zurich (ZH). With 49 g CO_2 / kWh, the cantonal carbon intensity is lower than the national average (EKZ 2026). Using ci_{ZH} and not the national ci_{CH} would have correspondingly lowered the result of (UC6) to 26.5 kg CO_2 and that of UC19 to 28 kg CO_2.

7.2 Indirect effects of Zü-Re and overall assessment

Although the SAFE-AI framework focuses on direct impacts, an aim of SAFE-AI from the outset was to also study the indirect impacts of the case studies. This is of both theoretical and practical significance. In the context of a sustainability chatbot such as Zü-Re, it is particularly important to assess whether the environmental price of the direct impact was worth the potential gains from fostering circular economy practices; in other words, whether the net impact introduced in Figure 1 (i.e., the sum of direct and indirect impacts) is beneficial or detrimental.

In the context of the sustainability chatbot, beneficial indirect impacts are expected to result mainly from behavioural changes of the users leading to more circular economy actions such as increased repair, reuse, share, etc. – in other words, lift more of the users' actions in the hierarchy of circularity as e.g. conceptualised by (Kirchherr et al. 2017): refuse – rethink – reduce – reuse – repair – refurbish – re-manufacture – repurpose – recycle – recover energy.

The plan was to estimate the extent of these behavioural changes via two complementary means:

- by an automated evaluation of the history of individual chats with the chatbot, which would indicate the topics users were interested in and the answers provided by the chatbot, thus yielding indications of possible behavioural changes, and
- voluntary answers to a survey sent to those users opting in, which explicitly asks them whether the answers of the chatbot provided useful and/or novel insights to them, which might have led to behavioural changes.

We discuss these two instruments in more detail in the two subsections below, starting with the survey.

7.2.1. Survey method and results

To take part in the survey, the users were asked to explicitly opt in (as shown in the lower right corner of the welcome screen from Figure 21) by voluntarily providing us their email address. Those who did opt in, were sent a link to the short survey 2 weeks after their initial chatbot interaction.

The alternative would have been to ask users to answer the survey immediately after chatbot interaction. This would have probably increased participation, perhaps even substantially, as an email sent 2 weeks later can easily be overseen, ignored, or initially postponed and then never looked at again. The users would also have been able to respond whether the chatbot's answers were new to them. Whether these ultimately would lead to a behavioural change would have remained subject to speculation though, as in the moment of the initial interaction the change could obviously not have happened yet.

Given this trade-off, it was a conscious decision to sacrifice participation in favour of outcome certainty (as potential behavioural changes would have time to manifest over the two weeks until the email came). Such certainty would provide us important grounding for the statistical analysis of the domains addressed with the bot during chats. Using a very conservative extrapolation from the survey's results to the domains addressed during the chats would allow for a reasonable (and conservative) estimate of the effect.

To keep the survey as short as possible, if only contained a few essential questions, as follows:

- two 5-point ordinal scales on the *novelty* and the *usefulness* of the chatbots answers, respectively,
- two questions on the *outcome*, i.e.,
    - "Did the chatbot lead to a behavioural change on your part?" [yes/no], and
    - "How did you make use of the information and recommendations from Zü-Re? Alternatively, why did you decide not to use them? Which products or services were involved? What specific actions did you take, and what was the outcome?"
- one question on the likely counterfactual: "Please briefly describe what you would likely have done without the Zü-Re recommendation."

Unfortunately, however, survey participation was far lower than our worst-case expectations: For the entire 6 months of the project and thousands of chatbot users, only a few dozen persons left their email addresses for participation. From them, only 13 ended up answering the survey.

Among these, as shown in Table 9, both novelty and usefulness were rated neutral to positive, with no negative ratings. The self-assessment indicates likely behavioural changes in about 50% of the cases. While generally promising, these results suffer, however, from three fundamental flaws:

- *No statistical significance* due to the very low participation.
- *Selection bias* among the respondents (a flaw that would have persisted even for a higher response rate).
- *Uncertainty of the counterfactual*: While the outcome is certain, the counterfactual is inherently hypothetical. While the users' self-assessment of the counterfactual can represent a good indication, the alternative remains ultimately ontologically uncertain.

Table 9: Outcome of the single-choice questions of the Zü-Re survey.

	Not at all / No	Rather not	So-so	Rather yes	Very / Yes
Novelty	–	–	5	3	5
Usefulness	–	–	4	5	4
Behaviour change	6	n/a	n/a	n/a	7

Some of the textual answers were also revealing, covering a large bandwidth, from truly helpful and likely circularity-enhancing effects to rather trivial insights that could have likely be gathered from Google searches as well, e.g.:

- "Information on **second-hand shops for outdoor clothing**, and acted on the information provided → Would have **sent it for disposal instead** of giving it to a second-hand shop."
- "I posted a question and the responses were helpful. For example, regarding where to get rid of old scarves; it was suggested that I donate them to a **charity** → did not act yet".
- "It was about the new **recycling centre**. We went there and everything was fine – but I think I would have found it **quicker** on the website! :)"
- "I looked for and found a **volunteer role** at the Salvation Army **charity shop**. I've also donated quite a lot of things there. It makes me happy. I've also started recycling plastic separately using the new Migros bags → would have used Google AI instead."

7.2.2. Evaluation of the chat collection

Over an assessment period of 6 months (3 March – 3 September 2025), 7751 meaningful chats were held with the Zü-Re chatbot. The categorisation into meaningful chats was automatically performed at runtime, as described in Section 7.1.2.

In an AI-supported analysis, we defined the following categories of products and services these chats focused on:

- **Electrical and household appliances** (toaster, computer, smartphone, washing machine, frying pan, water filter cartridges, citrus press, pressure washer)
- **Everyday household waste** (paper, cardboard, plastic packaging, Tetrapak, paint residue, glass, food waste, oil, PET bottles, corks, aerosol cans, fluorescent tubes, polystyrene, batteries, printer cartridges, garden waste)
- **Textiles** (clothing, bedding/linen)
- **Miscellaneous everyday items** (books, pens)

- **Sports equipment** (rucksacks, climbing gear, hiking boots)
- **Bicycles** (bike, e-bike, cargo bike, road bike, carbon bike frame)
- **Children's items** (pushchair, children's clothing)
- **Furniture** (mattress, sofa, wardrobe, garden furniture, chairs, wooden table)
- **Building materials** (windows, concrete)
- **Others** (further products or services)

The automated, AI-performed evaluation of the chat collection, revealed the distribution reflected in Figure 22. As one chat could revolve around several topics, multiple categorisations of a single chat were possible.

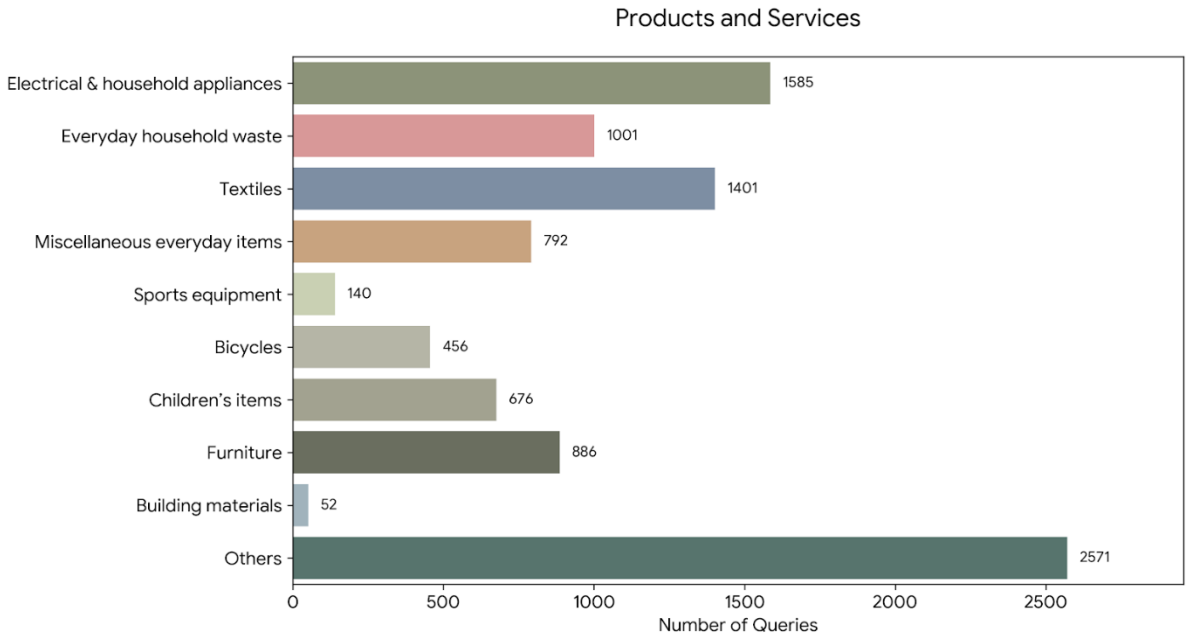


Figure 22: Main categories of products of services that occurred in the chats with Zü-Re, together with their usage distribution. A single chat could comprise more than one category.

Similarly, we also performed an automated categorisation of the system's resulting recommendations. As Figure 23 shows, quite a few of these recommendations were in line with circular economy principles: Swapping was only recommended once, but borrowing or renting (instead of buying) hundreds of times. Passing on, selling, donating as well as repairing (usually with concrete advices where the repairs can be performed in the neighbourhood) were recommended thousands of times, on par with disposal.

The indirect effects of Zü-Re's recommendations need to be computed as a difference between the as-is outcome and the (hypothetical) outcome of the counterfactual. For behavioural changes towards circular economy actions, this requires two conditions to be met: that a behavioural change did take place and that it would not have happened in absence of the chatbot.

The chatbots action recommendations do not imply that the action will actually be undertaken. Nor does the recommendation of a chatbot say anything about the counterfactual. Perhaps the user was already aware of those recommendations, and they turned out to be trivial. Or the user might be aware of better options than the bot's recommended actions. Or they may be outdated, too cumbersome, etc. It is even possible that the recommendations worsen the environmental impact of the user's actions.

We had hoped the ex-post survey would provide some complementary grounding to the analysis of domains and recommendations presented here. Given its limitations listed in Section 7.2.1 right above

Table 9, however, no robust estimate is possible. In a purely hypothetical exercise, we thus present in Section 7.2.3 below the sort of behavioural changes required to offset the direct impact and discuss their plausibility.

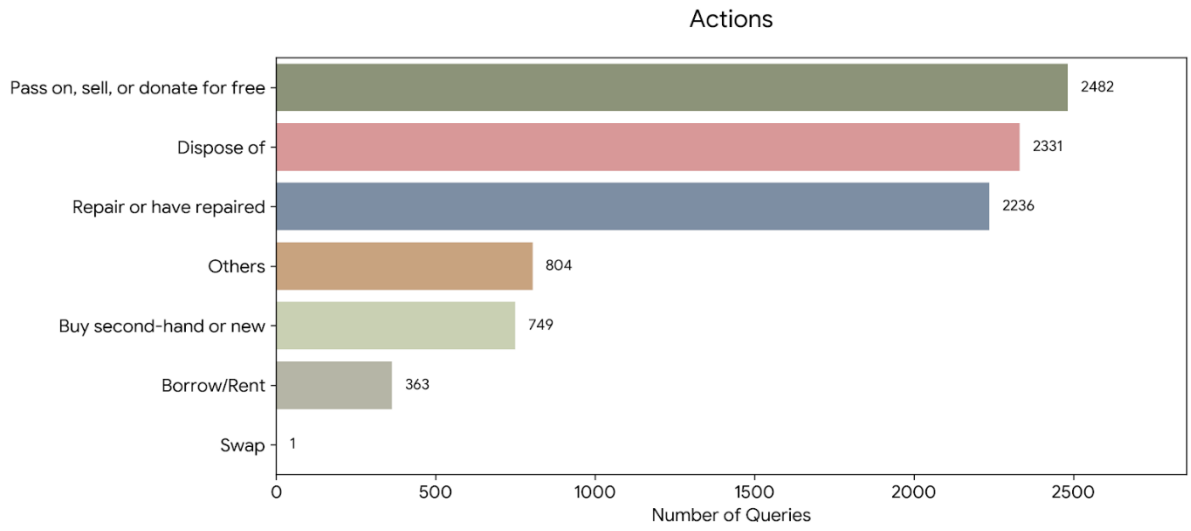


Figure 23: Zü-Re's recommended action categories, and their distribution.

7.2.3. Indirect effects required to offset the direct ones

As shown in Figure 22, one of the categories most often occurring are textiles. As a result, many system recommendations pinpointed users towards shops where they may, for example, have the zipper of an overcoat repaired, or towards options for passing them on or donating them to charities, who then either sell them or pass them on themselves. The carbon footprint of jacket production is around 30 kg CO_2eq (Carbonfact 2026); for an overcoat or high-quality waterproof jacket, it is likely substantially higher. And without a functional zipper, a jacket or an overcoat are not usable. As the lifespan of a waterproof jacket is 3-5 years (Cotswold 2025), prolonging its useful lifetime by one year can avoid around 10 kg CO_2eq.

Another product often addressed in the chats are bicycles and e-bikes. According to the LCA database "ecoinvent" their production footprint reaches from about 140 kg CO_2eq for a regular bicycle to almost 200 kg CO_2eq for an e-bike. Assuming a lifespan of around 7-10 years for the two, expanding the useful lifespan of a bicycle by repairing or donating it is likely to avoid around 20 kg CO_2eq.

As derived in Section 7.1.5, the direct impact of Zü-Re was around 35 kg CO_2eq per month. The distribution of products and services from Figure 22 is for 6 months; normalising it to one month, it can thus be said that if among the 235 monthly chats on textiles and 90 monthly chats on bicycles, the chatbots recommendations lead to one sweater and one bicycle being worn and used for one year more each, this roughly offsets the bots direct emissions.

7.3 Dora: Clinical AI agent by Ufonia

This case study examines the AI-assisted digital health solution Dora (Ufonia 2025), developed by Ufonia Ltd, Oxford, UK, as a further example of the application of the SAFE-AI framework to assess both the direct environmental impacts of an AI system and the indirect effects it might enable in a real-world service context.

Whereas the Zü-Re chatbot primarily influences environmental outcomes *indirectly* through changes in user behaviour and information use, the Dora system enables emission reductions mainly through service optimisation and substitution, by avoiding carbon-intensive physical healthcare activities such as face-to-face appointments and patient travel. While both case studies include direct and indirect effects, the Zü-Re case study focuses on illustrating the direct energy costs of running a complex generative AI architecture. The Dora case study, in contrast, is an example of a tightly bound net impact analysis.

The Dora system is analysed within the National Health Service (NHS) cataract surgery pathway, a high-volume clinical service characterised by a substantial number of patient interactions and a significant contribution from patient travel and on-site clinical activity. Dora is used to deliver autonomous, voice-based clinical conversations via telephone, supporting pre-surgery assessments and reminders as well as post-surgery follow-up and the collection of patient-reported outcome measures. Its introduction enables partial **optimisation** by reduction of travel to face-to-face outpatient appointments and **substitution** of letters with phone calls. Additionally, there is time-rebound from freeing up staff from routine staff-led telephone consultations. The analysis presented here builds on primary data gathered and an existing preliminary analysis by the Centre for Sustainable Healthcare (Higham 2023) during September 2023. We focus on the most significant optimisation effects and the direct impacts of the AI system.

From a SAFE-AI perspective, the assessment is conducted primarily at the system and usage levels. The direct environmental impacts considered correspond to the operation of the AI-enabled service, including server use, staff activity, and patient device interaction. Indirect effects are assessed in terms of emission changes resulting from changes to the clinical pathway, most notably reductions in physical outpatient visits and associated travel, consumables, and building energy use.

The scope of the case study is limited to energy use and associated greenhouse gas emissions, drawing on empirical data from an existing evaluation of the cataract pathway before and after the introduction of Dora (Meinert et al. 2024). The analysis is not intended as a full consequential life-cycle assessment of healthcare delivery, nor does it assess broader behavioural or long-term systemic effects. Results are therefore context-specific and should be interpreted as illustrative of the SAFE-AI framework rather than as generalisable conclusions about AI in healthcare.

7.3.1. System composition and usage

Dora is designed to deliver autonomous, voice-based clinical conversations via standard telephone networks. The system aims to increase healthcare capacity to meet rising demand, including for cataract surgery, where demand is expected to rise by 25% over the next decade and, by substituting selected physical interactions, has the potential to alter the carbon intensity of service delivery.

The solution has been introduced into the cataract surgery pathway at four NHS sites in England (“Site 1-4”). The intervention targets routine cataract surgeries, which are among the most common procedures performed in the NHS, with over 450,000 operations undertaken by the NHS annually in England, and 680,000 overall, including outsourced treatments (Donachie et al. 2025). Most cases succeed without complications or further treatment. This has already led to optimisation of remote patient support over the phone. The Dora AI-enabled system further extends this optimisation.

Patients, however, who are ineligible for the AI pathway, such as those with complex medical comorbidities, complicated surgical needs, or those unable to complete a call in English, remain on the traditional nurse-led or face-to-face pathway.

Pre-surgery phase

In the pre-operative phase, Dora is integrated as a triage and preparation tool (see Figure 24). In the reference (i.e., traditional) pathway, between 5 and 37% of patients (varying between hospitals) decide not to continue to surgery after having already travelled to the pre-surgery outpatient appointment ("PS drop-out").

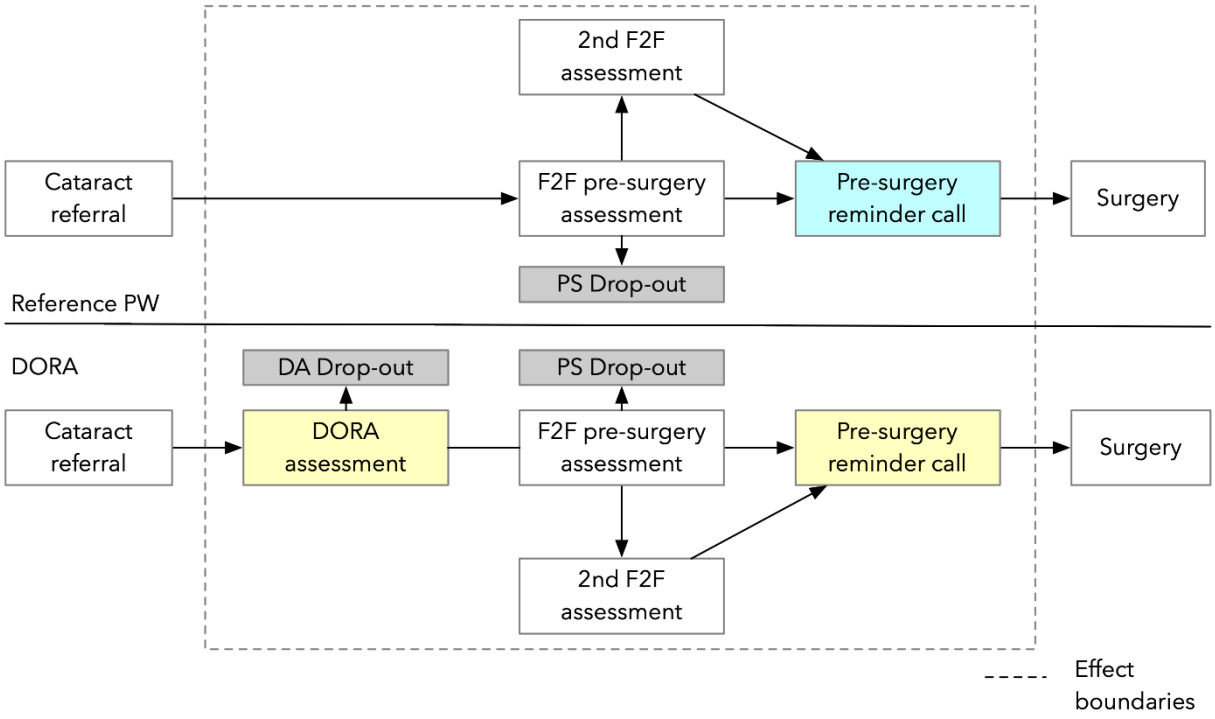


Figure 24: Pre-surgery cataract pathways – reference pathway and Dora AI-enabled system. The main optimisation effect is in drop-outs before the F2F pre-surgery assessment. An additional effect is reduced 2nd F2F assessments as patients arrive better prepared at the at the first F2F assessment.

Helping patients to make this decision before travelling is the main optimisation effect of Dora in the pre-surgery phase. All referred patients receive a "pre-surgery assessment" call from Dora before attending their physical face-to-face (F2F) appointment. This initial screening allows the system to identify patients who are unsure about proceeding. Typically, 6% of patients flag issues during this call and are diverted to a human nurse for further discussion. Half of these patients do not proceed (exit at "DA drop-out"). Consequently, the attendance rate for the subsequent physical assessment improves, with 97% of referred patients attending their F2F appointment as compared to the reference pathway. Additionally, Dora performs a "pre-surgery reminder call" in the days leading up to the operation, replacing the manual call traditionally made by a hospital administrator. During this interaction, the AI reminds patients of necessary items to bring, such as medication lists, which reduces the likelihood of appointments being cancelled or repeated due to patient unpreparedness.

Post-surgery phase

In the post-operative phase (see Figure 25), Dora replaces some of the routine human-staffed interactions to monitor recovery. All eligible patients receive a post-surgery follow-up call from the AI system. Patients who "raise no concerns" at this automated clinical review are discharged immediately, while those who flag potential concerns are escalated to a face-to-face review or a nurse-led consultation. Finally, 12 weeks after surgery, the system conducts a call to collect Patient Reported Outcome Measures (PROMs), substituting the traditional method of sending paper feedback forms via post.

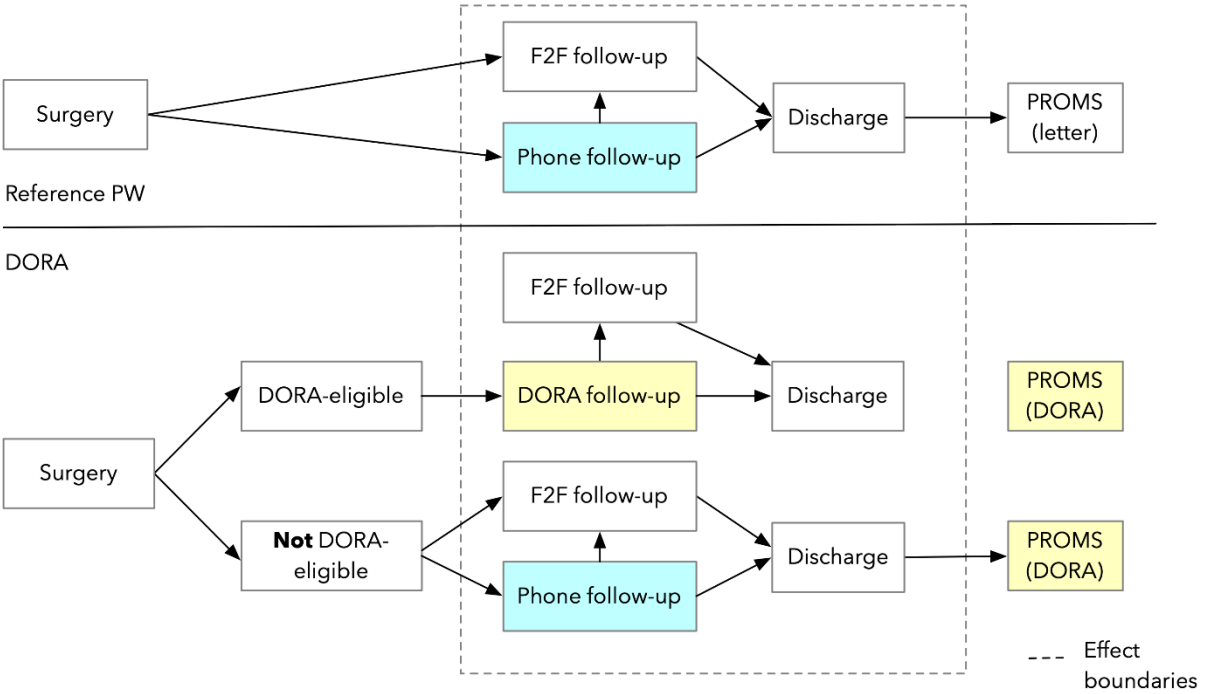


Figure 25: Post-surgery cataract pathways – reference pathway and Dora AI-enabled system.

The main optimisation lies in the reduction of F2F review appointments. Here, Dora standardises the escalation process, which leads to reductions in some settings but slight increases in others where the previous human triage might have been less rigorous or different in nature.

7.3.2. Direct environmental impact

The assessment of Dora's direct environmental footprint follows a bottom-up process-based analysis in line with the SAFE-AI methodology. The approach quantifies the GHG emissions associated with every digital interaction facilitated by the AI system, aggregating impacts from server energy consumption, staff operations, and patient device usage.

As shown in Figure 26, the technical infrastructure supporting Dora relies on a combination of internal cloud-based servers and external commercial API services, alongside standard telecommunications networks. Specifically, speech-to-text transcription uses a blend of commercial API services, while text-to-speech output also relies on an external commercial API. The internal components – a custom NLP pipeline for intent and entity extraction and a conversation machine learning model – are hosted on Ufonia's own infrastructure. Patients use their telephones to interact with the system and receive SMS as reminders.

At the time of this evaluation the servers had a capacity of approximately 13,440 calls per week, though utilisation across the four sites averaged 440 calls per week during the study period. The operational footprint includes the energy and embodied emissions of the Ufonia staff (IT support and operations) who oversee the system, largely working from home using laptops and monitors. On the user side, the system utilises the patient's own smartphone or landline for voice connectivity and Short Message Service (SMS) for reminders, without requiring patients to install specific applications or utilise high-bandwidth data.

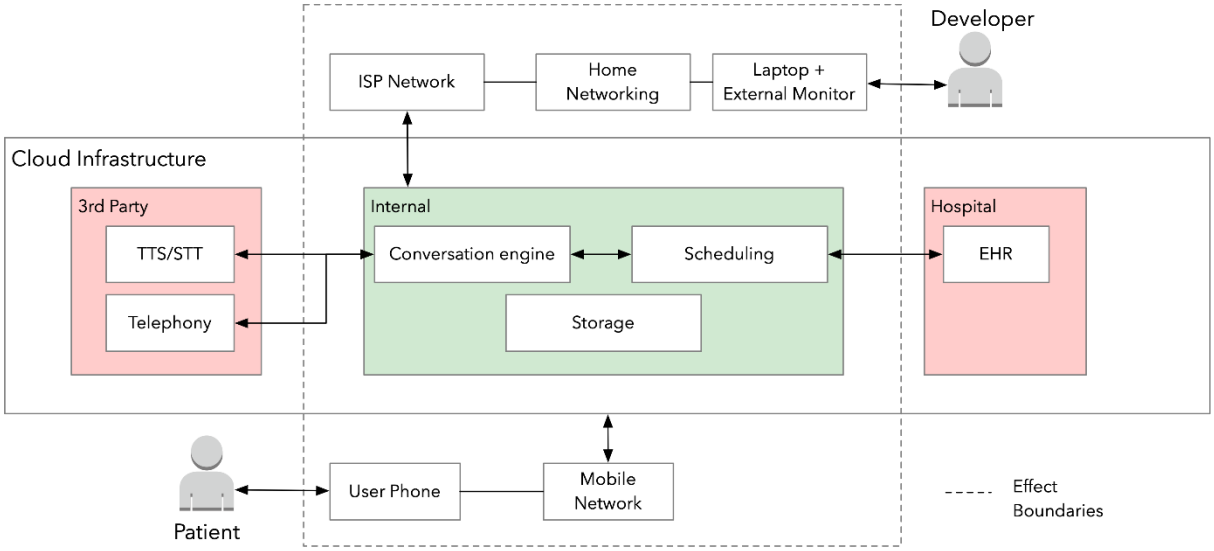


Figure 26: Dora system architecture showing hybrid internal/external AI components. Speech-to-text (STT) and text-to-speech (TTS) processing rely on external commercial APIs, while intent extraction and dialogue management use internally developed models. The current carbon analysis captures only the internal Ufonia components; emissions from external API providers are not fully accounted for.

The existing carbon footprint analysis considers the impacts along four system parts with varying life cycle scopes:

- **Server use**: Only impacts from electricity consumption are considered. Servers account for the majority of the direct impact, contributing between 57% and 63% of the estimated GHG emissions per call. The emissions are estimated based on the financial cost of the servers, assuming 15% of the spend covers electricity and that the energy mix is 67% green and 33% grey based on cloud provider data. Notably, the servers operate below their capacity of 13,440 calls per week (with actual usage around 440 calls per week).

  This figure does not include the energy consumption of external commercial APIs used for speech-to-text and text-to-speech processing, which handle audio streams for every call, and for which no data had been collected. The carbon footprint of these external services is thus not captured in this analysis, representing a limitation in the system boundary.

- **Ufonia staff**: This category captures the environmental cost of the IT support and operations team, including the human reviewers who quality-assure a subset of calls. It includes emissions from homeworking energy consumption, network usage, and the embodied carbon of hardware laptops and 2 LED monitors, assuming a five-year lifespan. Networking was modelled with 21W in total over a 7.5h working day, including home WiFi.

- **Patient User Devices**: The analysis accounts for the embodied emissions, electricity, and network usage of the smartphones used by patients to receive the AI calls.

- **Ancillary Items**: The footprint includes a text message reminder sent to patients before every call and a physical one-page information leaflet provided for the pre-surgery and post-surgery interactions.

- **External API use** (not quantified): Every Dora call requires real-time audio processing via external commercial speech-to-text and text-to-speech APIs. The energy consumption and associated emissions of these third-party services fall outside the current analysis boundary. Depending on the providers used and their infrastructure, this could represent a material addition to the per-call carbon footprint.

The total carbon footprint per Dora call varies by interaction type, driven primarily by call duration and the inclusion of physical materials:

- **Post-surgery follow-up**: This interaction carries the highest footprint at 150 g CO_2eq. per call. This is attributed to its longer average duration of 7 minutes and the inclusion of the physical information leaflet.
- **Pre-surgery assessment**: This call generates 130 g CO_2eq., based on an average duration of 6.8 minutes.
- **Pre-surgery reminder**: With a shorter duration of 5.6 minutes, this call results in 120 g CO_2eq.
- **PROMs call**: As the shortest interaction at 5.1 minutes, the outcomes survey has the lowest footprint of 110 g CO_2eq.

These figures establish the direct environmental cost of the AI system, ranging between 0.11 – 0.15 CO_2eq. per interaction.

7.3.3. Indirect effects

The dominant indirect environmental effect associated with the deployment of Dora arises from fewer F2F outpatient appointments and associated travel. In the reference cataract pathway, post-operative reviews involve nurse-led or patient-initiated telephone consultations that in some instances are followed by physical appointments, resulting in patient travel. With Dora, eligible patients receive an automated post-operative assessment by telephone, and only those who report symptoms or fail predefined checks are escalated to face-to-face review.

Patient travel is the largest contributor to the carbon footprint of outpatient appointments. Travel is reduced in the pre-operative phase through earlier identification of patients who do not wish to proceed to surgery, and a reduction in repeat F2F pre-surgery assessments. However, the savings in the pre-op phase are smaller than in the post-op phase. Overall, the indirect environmental benefits of Dora are driven primarily by structural changes to the care pathway, rather than by behavioural change, resulting in short and well-defined causal chains and comparatively low uncertainty in the estimated emission reductions.

Impacts in the reference pathway: The overall impact of a single F2F appointment is between 12.24 – 12.78 kg CO_2eq., according to the Centre for Sustainable Healthcare. From this, patient travel is the dominant factor, contributing 61% to 64% (approximately 7.79 kgCO2e on average) of the total emissions per appointment. This is based on strong data from patient surveys specific questions during Dora calls. The data includes the distance and travel modality of patients and combined with carbon intensities from (DEFRA 2023).

Consumables used during F2F appointments include items such as PPE, eye drops, and medical wipes and account for approximately 18% (2.19 – 2.27 kg CO_2eq.) of the impact, using environmentally extended input output analysis (EEIOA).

Hospital overheads: In an attributional context, the energy required to heat, light, and power the reception, nurse rooms, and clinic areas adds another 11% to 12% (1.32 – 1.58 kg CO_2eq.). However, in the Dora case, these overheads should not be factored into a counterfactual analysis, as they are not meaningfully affected from the avoidance of a subset of F2F appointments along the pathway because the building will continue to be operated and staff will continue to commute there.

7.3.4. Net effect of the Dora system

For the net reduced emissions, the additional energy consumption of the Dora system added to all patients' pathways must be weight against the savings for the patients who avoid travel. Table 10 summarises the relative changes between the traditional F2F pathway and the Dora pathway.

Table 10: Relative changes in number of Face-to-Face appointments and follow-up phone calls across the four hospitals where the Dora system was evaluated. Showing the reference cataract pathway and the Dora-augmented pathway.

Appointment type			F2F pre-surgery assessment OPA	F2F post-surgery follow-up OPA	Post surgery phone follow-ups
Number of appointments per 100 cataract surgeries	**Site 1**	Reference PW	111.1	33.3	66.7
		Dora	107.8	24.8	33.8
		Difference	-3%	-26%	-49%
	Site 2	Reference PW	111.1	20.2	79.8
		Dora	107.8	21.8	40.3
		Difference	-3%	8%	-50%
	Site 3	Reference PW	158.7	46	0.6
		Dora	154	30.9	12.9
		Difference	-3%	-33%	2148%
	Site 4	Reference PW	105.2	32.1	4.2
		Dora	102.1	31.3	10.5
		Difference	-3%	-3%	153%

In the **pre-surgery** phase F2F appointments were reduced by 3%. At 0.13 kg CO2eq. per call and 10 kg CO2eq. per F2F appointments (12.24 kg minus short-term invariable hospital overheads), the savings of roughly 17 kg CO_2eq. (3 x 10 kg CO_2eq. – 100 x 0.13 kg CO_2eq.) are small in this phase.

Larger total emission reductions are observed in the **post-surgery** pathway. While there was an overall reduction of escalations of post-op reviews to F2F appointments, the relative change differed between all hospitals, with one hospital observing an increase in F2F appointments. In this context it must be mentioned that hospital trusts operate different models regarding who initiates a post-op review call – either a nurse (Site 1 & 2) or the patient (Site 3 & 4).

In “Nurse-Led" Sites (1 & 2), the results were mixed.

- At Site 1, the Dora pathway significantly reduced F2F appointments by 26% (from 33.3 down to 24.8 per 100 surgeries) compared to the traditional nurse-led model. This suggests the AI was “stricter” than the nurses at this site in discharging patients remotely.
- At Site 2, however, the Dora pathway actually *increased* F2F appointments by 8% (from 20.2 up to 21.8 per 100 surgeries). This implies that the nurses were previously escalating fewer patients than the AI did, or that the AI's safety protocols required it to be more cautious than the human nurses were.

In both "Patient-Initiated" Sites (3 & 4) after implementation of Dora a reduction of F2F appointments was observed: At Site 3, F2F appointments dropped by 33%, while at Site 4, F2F appointments dropped by only 3%.

The difference in emissions between the AI-enabled pathway and reference pathway is much more pronounced in the post-surgery phase. When considering avoided emissions from travel and F2F appointments per visit of about 10 kg CO_2eq. a reduction of 26% resulted in around 260 kg CO_2eq. compared to 15 kg CO_2eq. from additional Dora calls (0.15 kg CO_2eq. per call, because calls are longer than in the pre-op phase).

7.3.5. Discussion and conclusion

For the hospitals where F2F post-surgery appointments were relatively high (Site 1: 33%, Site 3: 46% in reference pathway), the introduction of Dora brought a meaningful reduction (Site 1: 25%, Site 3: 31%). Here, the introduction of the AI-enabled solution removed discretion that defaulted to higher-carbon choices and standardise low-carbon execution. In this way, the introduction of Dora changed the escalation-step from a behavioural decision to a structural one. While in the case of Zü-Re, decision support is given without changing the default (and thus resulting in uncertain substitution effects), Dora results in more consistent escalations.

In the other two hospitals, F2F appointment rates were already near or lower (Site 2: 20%, Site 4: 32%) than the rates in DORA regimes in the 'high-F2F' hospitals. Thus, the introduction of Dora only had a moderate substitution effect, or even an induction effect (resulting in additional impacts).

While it is not possible to predict the magnitude of indirect effects with precision before deployment, it is possible to assess whether an AI system has high, medium, or low substitution potential by analysing the reference baseline activity and whether the AI alters defaults rather than merely informing decisions. This latter point affects the decarbonisation potential of chatbots. As (Cook et al. 2025) remark, the existing literature on using AI for decarbonisation identifies mainly automation (e.g., in energy systems), while chatbots require humans to act on the knowledge they present.

System boundary limitations. The direct carbon footprint figures reported (0.11 – 0.15 kg CO_2eq. per call) capture only Ufonia's internal infrastructure. Because Dora relies on external commercial APIs for both speech-to-text and text-to-speech processing—services that handle streamed audio for every patient interaction—the true direct footprint is higher than reported. Without transparency from API providers on per-request energy consumption, this gap cannot be quantified precisely, but it represents a methodological limitation that should be acknowledged when comparing Dora's footprint to other AI systems or when assessing net environmental benefits.

Rebound. Demand for cataract surgery in the UK is currently not met by the NHS and about one third are outsourced to private suppliers. Optimisation along the pathway therefore has a good chance of rebound within the cataract pathway itself, by reducing this backlog. This would have a beneficial societal effect, while erasing the potential environmental effect of an AI system, as analysed above. There is thus the opportunity for time rebound within the wider ophthalmology sector resulting in shorter waiting lists and thus increased health care provision for the general society, albeit together with a slightly increased carbon impacts, for example from hospital consumables.

Conclusion. While both Zü-Re and DORA exhibit indirect environmental effects, the uncertainty associated with these effects differs substantially: Zü-Re relies on long and behaviourally mediated consequence chains, whereas DORA's impacts arise from short, structurally conditioned optimisation effects. Both systems, however, share a reliance on external commercial AI APIs: Zü-Re for generative language models, and Dora for speech processing services. This similarity is relevant when comparing the direct environmental footprints of the two systems.

8 Conclusions and outlook

8.1 Core contributions of SAFE-AI

8.1.1. Moving beyond the “black box” view of AI impacts

SAFE-AI shifts environmental assessment away from a black-box treatment of AI (which studies generative AI models in isolation), by providing an analysis where and how energy and other impacts arise. Rather than treating model as the only object of assessment, it considers the architectural drivers of resource demand across hardware, data-centre operation, and the surrounding software system.

SAFE-AI formalises the interplay between ML models, the AI system that orchestrates them, and the way that the system is actually used. The framework considers these as complementary but inseparable layers. This is a necessary broadening of perspective to enable environmental assessments for the wider sector of AI-enabled software.

SAFE-AI structures assessment across three distinct, but tightly coupled, levels. At the *model level*, it assesses impacts associated with the development and existence of an ML model, including training and the allocation of associated overheads. At the *system level*, it assesses the broader application ecosystem that mediates user interaction, including orchestration layers, databases, retrieval mechanisms (e.g., RAG), security components, and hosting choices. At the *usage level*, it allocates impacts to individual interactions (for example, a single query), which is essential for accountability, procurement decisions, and any credible pathway to eco-labelling of AI services.

Across the 3 layers, SAFE-AI takes an architectural perspective. On the *model level*, it considers the essential drivers of energy consumption in the transformer architecture. On the *system level*, it shows how energy demand is shaped not only by model choice, but also by architectural decisions such as retrieval pipelines and orchestration logic. This architecture-driven perspective is particularly important for modern AI deployments in which non-model components are important contributors and where one user request can trigger multiple downstream calls and tool invocations. Finally, on *usage level*, it proposes a stochastic analysis to leverage per-query token distribution for a system-wide energy consumption analysis.

8.1.2. Consolidating the state of the art and addressing recurring confusions

A further contribution is the synthesis and clarification of concepts that are repeatedly confused in the AI footprinting discourse. SAFE-AI consolidates existing literature and provides terminology and structuring that help explain why published numbers diverge, why system boundaries matter, and where uncertainty is unavoidable.

SAFE-AI introduces the assessment trilemma, which formalises the relationship between three types of constraint in AI footprinting: an assessor typically cannot maximise granularity (engineering fidelity), context (real-world representativeness), and data availability (feasibility given access constraints) at the same time. This trilemma explains why lab benchmarks can be precise yet unrepresentative, why top-down corporate reports can be contextually meaningful yet diagnostically weak, and how damaging the lack of transparency model providers is to accuracy of assessments.

SAFE-AI also clarifies that the ML model lifecycle (e.g., research, pre-training, fine-tuning, inference) and the environmental lifecycle (hardware production, operation, end-of-life) of supporting devices are independent, which has previously been misrepresented. Each ML lifecycle stage entails processes across the full environmental lifecycle of the enabling devices and infrastructure. This distinction matters because conflating these lifecycles can lead to systematic omissions, such as treating inference as “use phase only” while silently excluding embodied impacts that must be allocated to that inference.

The framework also provides a structured account of AI water footprints and highlights two recurring issues. First, it documents that estimates are often not comparable because “withdrawal” and “consumption” are used inconsistently across sources, arguing that only consumption is truly relevant. Secondly,

it distinguishes direct water consumption for DC cooling from indirect water use associated with electricity generation, noting that the latter can significantly exceed direct cooling water in some contexts.

8.1.3. A workflow for energy, GHG, and water assessment of AI systems

Finally, one of the most important contributions of the framework is to establish an assessment workflow for AI systems. It encompasses the assessment of the entire AI ecosystem, and of both internal and external ML models.

This workflow is particularly relevant for AI system developers that employ external AI foundation models. Starting from sparse and generic information that each such system provider receives as part of the billing from the foundation model provider, SAFE-AI deploys a stochastic analysis that allows for a more robust assessment and aggregation from *usage level* to *system level*.

The workflow comprises 20 equations that together lead to robust assessments of an AI system's energy consumption as well as related GHG emissions and water consumption. These equations aim to find a way out of the assessment trilemma, striking a balance between granularity, context, and availability of data. When enough data is available (e.g., on whether on-site power generation is deployed or the exact greenhouse gas or water intensities of electricity), they also yield highly granular results.

When only few details are known, or the assessed system is not so complex (e.g., if there is no on-site power generation or no internal ML models), several of these equations become trivial (e.g., multiplication by zero), and the assessment reduces to about 12 non-trivial equations. As basis for the GHG and water estimates, the energy assessment is mandatory; if not both GHGs and water are required, the assessment reduces even further.

Some of the impacts, especially those related to overheads for fundamental AI research and unreleased models – but also, for example, the influence of end devices – are highly uncertain yet likely too important to be ignored. The SAFE-AI workflow thus explicitly considers them and suggests some reasonable default values for the PUE, the overhead of fundamental AI research and unreleased models, and the overhead of end devices.

These overheads are not listed indiscriminately, but where they are likely to be material: e.g., the PUE for the local DC of the AI system but not the remote DCs of external AI models (as it is likely already included), or the end device overhead for GHGs (due to device production) but not water (not existing) or energy (likely negligible).

Finally, the workflow also provides default values that can be used where no precise data are available: for the coefficients used to compute the per-query energy depending on input and output tokens as well as the WUE of data centres and the water intensity of electricity. For the GHG calculations, it also advices to use the US or worldwide average carbon intensity of electricity, if no precise information on the DC location is available.

8.2 Core conclusions for the environmental assessment of AI

C1. Assess AI systems in context: A central conclusion is that assessing isolated models is insufficient for product carbon footprints of AI services. The integration of models with orchestration layers, retrieval components (including RAG), databases, and operational tooling requires inclusive system boundaries, or it risks producing results that are incomplete, non-comparable across services, or misleading for procurement and policy.

C2. Hidden overheads can dominate marginal usage impacts: AI systems contain overheads that are not visible at the user interface but still contribute to the footprint, including baseline (idle) power, shared infrastructure, logging and safety layers, and retrieval and database operations. In addition, agentic AI patterns introduce the risk of "unbounded compute trees," in which a single user request triggers chains of sub-tasks, tool calls, and follow-up model invocations. This makes energy and impact per user interaction highly dependent on context, orchestration design and guardrails, rather than on the nominal model alone.

C3. Data visibility is the primary barrier to robust assessments: SAFE-AI concludes that the dominant constraint on accuracy is the hierarchy of information access. Frontier model providers hold the primary data needed for rigorous accounting, while downstream deployers and independent assessors often must rely on secondary estimates and proxies.

C4. Inference impacts are variable and context-dependent; "per query" numbers are not stable properties: There is no single "cost per query" that can be treated as a stable characteristic of an AI service. Energy use depends on prompt and context length, output length, model architecture (including MoE), hardware utilisation, batching, and scheduling. For text-based GenAI models, output tokens are the most significant driver of energy consumption. However, for long context tasks, input processing can become dominant, and system features such as RAG retrieval pipelines can materially change the overall footprint. Consequently, reporting must either represent this variability explicitly or clearly document the assumptions.

C5. Inference dominates, but sector-wide training and development remains important and not sufficiently understood: While training can be energy-intensive, aggregate inference for widely used models surpass training over a model's lifetime. Moreover, AI research and development activities including experimentation or failed training runs appear substantial and need to better understood to avoid systematic understatement. Any accounting that reports training only as the final successful run, without allocating development overheads, risks biasing results downwards.

C6. Functional units and allocation form central methodological challenges: A further conclusion is that the key methodological difficulty lies in defining functional units and allocation principles that are simultaneously meaningful, comparable, and feasible under assessment constraints. Because many impacts are shared across users and services, and because workloads are heterogeneous, allocation choices determine what per-usage results actually mean.

C7. For now, tokens are a necessary evil of a generic metric: Given the heterogeneity of possible queries, the technically-oriented token numbers are a necessary evil for a robust generic assessment method, in particular when aggregating from usage to system level. Context-dependent, this metric should then ideally be transformed into a functional unit that is semantically more meaningful in the application context.

C8. Indirect effects are important, yet contextual and difficult to assess. While direct effects are themselves riddled with uncertainties and methodological challenges, conceptually they are straightforward. By contrast, indirect effects are often complex, subtle, and intertwined. Their assessment is thus highly contextual and difficult to generalise, and as such not a part of the SAFE-AI framework. Two case studies have nevertheless shown how such assessments can be performed: Relatively straightforward when the effects are limited and the counterfactual well defined (which is rather the exception), and scenario-based when effects are more complex and the (hypothetical) counterfactual hard to establish. The case studies have also shown how AI may bring about environmental benefits in two sectors of application: circular economy and healthcare, respectively.

8.3 Outlook and open questions

An important current development is the growth of agentic AI. While its basic architecture has been presented in Section 4.1.3, agentic AI systems have been excluded from further considerations, and in particular from the assessment workflow from Chapter 6. The conceptual issue with agentic AI is their recursion, which is conceptually (albeit of course not practically) boundless. Assessments of agentic AI systems thus requires new principles, which SAFE-AI is ill equipped for. A fundamental examination of agentic AI systems seems to the authors the most important future research direction.

At the same time, generative AI models are also continuously developing. Image and video generation increasingly no longer use transformer architectures but diffusion models, for which tokens have a different meaning. Diffusion models have been briefly addressed in Section 4.2.5, but it has not been thoroughly examined whether, and to which extent, the SAFE-AI principles can be deployed for diffusion models as well. It is likely that some changes are required; however, unlike agentic AI systems,

fundamental changes to the workflow might not be required. How large these required adaptations are – in particular whether they are fundamental or marginal – is also an open question that deserves further attention.

Training has an important role in the assessment of AI energy consumption not only due to its bulk energy cost, but additionally through its potential to affect downstream inference efficiency. Strategically, data curation and deduplication decisions during training can significantly influence both training cost and downstream inference efficiency by reducing redundant computation. From an assessment perspective, training energy should therefore be understood not only as a one-off cost to be amortised over model usage, but also as a design phase that critically shapes the long-term energy profile of AI systems. These important considerations have not at all been touched in this study.

The assessment of water impact has focused on the water consumption as more relevant indicator than water withdrawal. Water, however, has a highly local impact, which depends on the water scarcity. Framing the water impact of AI – and of data centres more generally – in terms of water scarcity, should be a high research priority.

Finally, the SAFE-AI assessment workflow establishes various default values for several types of coefficients: i) those deployed in the energy assessment of one inference based on the number of tokens, ii) the energy or GHG overheads due to the PUE, fundamental AI research and unreleased models, and end devices as well as iii) the water usage effectiveness of DCs and the water intensity of electricity generation. When the general knowledge progresses and better defaults can be defined, these defaults must be updated.

Especially the coefficients used to model the per-query energy consumption are subject to rapid change. The energy efficiency of the accelerated computing hardware deployed in AI computations is improving by about 50% per year (Coroamă et al. 2025). New algorithmic paradigms, model architectures, and data representations can be even more disruptive, as e.g. (DeepSeek 2025) has shown. Yearly algorithmic efficiency gains are probably on par with or higher than hardware efficiency, and possibly as high as 300% (Amodei 2025). As the two efficiency gains combine, the per-token efficiency is rapidly growing. While this trend is countered by ever more powerful and complex models using longer contexts and delivering increasingly sophisticated results, and the overall energy usage of AI is consequently growing, the per-token energy is nevertheless decreasing rapidly. Exploring how these coefficients could be parametrised, so they would as a result be automatically updated with passing time, would be a worthwhile endeavour.

Acknowledgements

The authors wish to thank Roland Brüniger and Dr. Michael Moser (SFOE) for accompanying the project. They both provided valuable feedback along the entire project as well as to earlier drafts of this manuscript. Additionally, many thanks to Dr. Aisling Higham (Ufonia) as well as Anja Tamburini, Daniela Diener, and Selina Galliker (all from the city of Zürich) and Dimo Notarfrancesco (Staay AG) for their support with data and discussions on the respective case studies. Anja Tamburini and Dimo Notarfrancesco in particular have been always available discussion partners and valuable sources of data throughout the entire project.

References

Altman, Sam. 2025. ‘The Gentle Singularity’. *Sam Altman*, June 10. https://blog.samaltman.com/the-gentle-singularity.

Amazon Sustainability. 2026. *Water Stewardship - Amazon Sustainability*. https://sustainability.aboutamazon.com/natural-resources/water.

Amodei, Dario. 2025. *On DeepSeek and Export Controls*. January 28. https://darioamodei.com/on-deepseek-and-export-controls.html.

Azevedo, Dan, Christian Belady, and Jack Pouchet. 2011. *Water Usage Afectiveness (WUE): A Green Grid Data Center Sustainability Metric*. White Paper No. 35. The Green Grid. https://airatwork.com/wp-content/uploads/The-Green-Grid-White-Paper-35-WUE-Usage-Guidelines.pdf.

Baliga, J., R. Ayre, K. Hinton, W. V. Sorin, and R. S. Tucker. 2009. ‘Energy Consumption in Optical IP Networks’. *Journal of Lightwave Technology* 27 (13): 2391–403. https://doi.org/10.1109/JLT.2008.2010142.

Baliga, J., R. W. A. Ayre, K. Hinton, and R. S. Tucker. 2011. ‘Green Cloud Computing: Balancing Energy in Processing, Storage, and Transport’. *Proceedings of the IEEE* 99 (1): 149–67. https://doi.org/10.1109/JPROC.2010.2060451.

Baylor, Denis, Eric Breck, Heng-Tze Cheng, et al. 2017. ‘TFX: A TensorFlow-Based Production-Scale Machine Learning Platform’. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA), KDD ’17, August 13, 1387–95. https://doi.org/10.1145/3097983.3098021.

Berthelot, Adrien, Eddy Caron, Mathilde Jay, and Laurent Lefèvre. 2024. ‘Estimating the Environmental Impact of Generative-AI Services Using an LCA-Based Methodology’. *Procedia CIRP*, 31st CIRP Conference on Life Cycle Engineering, vol. 122 (January): 707–12. https://doi.org/10.1016/j.procir.2024.01.098.

Berthelot, Adrien, Eddy Caron, Mathilde Jay, and Laurent Lefèvre. 2025. ‘Understanding the Environmental Impact of Generative AI Services’. *Commun. ACM* 68 (7): 46–53. https://doi.org/10.1145/3725984.

Bieser, Jan C. T., Vlad C. Coroamă, Pernilla Bergmark, and Matthias Stürmer. 2024. ‘The Greenhouse Gas (GHG) Reduction Potential of ICT: A Critical Review of Telecommunication Companies’ GHG Enablement Assessments’. *Journal of Industrial Ecology* n/a (n/a). https://doi.org/10.1111/jiec.13524.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, et al. 2022. ‘On the Opportunities and Risks of Foundation Models’. arXiv:2108.07258. Preprint, arXiv, July 12. https://doi.org/10.48550/arXiv.2108.07258.

Bordage, Frédéric, Lorraine de Montenay, Etienne Lees-Perasso, et al. 2021. *Digital Technologies in Europe: An Environmental Life Cycle Approach*. https://extranet.greens-efa.eu/public/media/file/1/7388.

Börjesson Rivera, Miriam, Cecilia Håkansson, Åsa Svenfelt, and Göran Finnveden. 2014. ‘Including Second Order Effects in Environmental Assessments of ICT’. *Environmental Modelling & Software* 56 (June): 105–15. https://doi.org/10.1016/j.envsoft.2014.02.005.

Bornstein, Matt, and Rajko Radovanovic. 2023. ‘Emerging Architectures for LLM Applications’. *Andreessen Horowitz*, June 20. https://a16z.com/emerging-architectures-for-llm-applications/.

Brandtzaeg, Petter Bae, and Asbjørn Følstad. 2017. ‘Why People Use Chatbots’. In *Internet Science*, edited by Ioannis Kompatsiaris, Jonathan Cave, Anna Satsiou, et al. Springer International Publishing. https://doi.org/10.1007/978-3-319-70284-1_30.

Bremer, Christina, George Kamiya, Pernilla Bergmark, Vlad C. Coroamă, Eric R. Masanet, and Reid Lifset. 2023. *Assessing Energy and Climate Effects of Digitalization: Methodological Challenges and Key Recommendations*. nDEE Framing Paper Series. Research Coordination Network on the Digital Economy and the Environment. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4459526.

Brohan, Anthony, Noah Brown, Justice Carbajal, et al. 2023. ‘RT-1: Robotics Transformer for Real-World Control at Scale’. arXiv:2212.06817. Preprint, arXiv, August 11. https://doi.org/10.48550/arXiv.2212.06817.

BSI Knowledge. 2011. *Specification for the Assessment of the Life Cycle Greenhouse Gas Emissions of Goods and Services*. PAS 2050:2011. https://knowledge.bsigroup.com/products/specification-for-the-assessment-of-the-life-cycle-greenhouse-gas-emissions-of-goods-and-services.

Carbonfact. 2026. *Carbon Footprint of Jackets*. https://www.carbonfact.com/carbon-footprint/jackets.

Carion, Nicolas, Laura Gustafson, Yuan-Ting Hu, et al. 2025. ‘SAM 3: Segment Anything with Concepts’. arXiv:2511.16719. Preprint, arXiv, November 20. https://doi.org/10.48550/arXiv.2511.16719.

Chase, Harrison. 2024. ‘What Is a “Cognitive Architecture”?’ *LangChain*, July 6. https://blog.langchain.com/what-is-a-cognitive-architecture/.

Chatterji, Aaron, Thomas Cunningham, David J. Deming, et al. 2025. ‘How People Use ChatGPT’. Working Paper No. 34255. Working Paper Series. National Bureau of Economic Research, September. https://doi.org/10.3386/w34255.

Cheung, Dana. 2024. *Introducing the Emerging LLM Tech Stack*. March 15. https://www.codesmith.io/blog/emerging-llm-tech-stack.

Comscore. 2016. ‘Part 2: Why the Power of Habit Drives Power Law Distributions in Mobile App Usage’. *Comscore, Inc.*, September 16. https://www.comscore.com/Insights/Blog/Part-2-Why-the-Power-of-Habit-Drives-Power-Law-Distributions-in-Mobile-App-Usage.

Cook, Joseph, Romain Jacob, Jo Lindsay Walton, Adrien Berthelot, Asim Hussain, and Daniel Schien. 2025. ‘Beyond Counting Carbon: AI Environmental Assessments Struggle to Inform Net Impact Decisions’. Preprint, ETH Zurich. https://doi.org/10.3929/ethz-c-000789254.

Coroamă, Vlad C. 2021. *Investigating the Inconsistencies among Energy and Energy Intensity Estimates of the Internet – Metrics and Harmonising Values*. No. 67656. Swiss Federal Office of Energy SFOE. https://www.aramis.admin.ch/Default?DocumentID=67656.

Coroamă, Vlad C. 2025a. ‘Doubts about Using Median Prompt as Metric’. *Linked In*, November 12. https://www.linkedin.com/feed/update/urn:li:activity:7394057436141121536/.

Coroamă, Vlad C. 2025b. *Temperature Optimisation in Data Centres*. Swiss Federal Office of Energy SFOE.

Coroamă, Vlad C., Pernilla Bergmark, Mattias Höjer, and Jens Malmodin. 2020. ‘A Methodology for Assessing the Environmental Effects Induced by ICT Services: Part I: Single Services’. *Proceedings of the 7th International Conference on ICT for Sustainability*, June 21, 36–45. https://doi.org/10.1145/3401335.3401716.

Coroamă, Vlad C., and Oana Dumbravă. 2026. *Circular Economy Synergies and Trade-Offs in Data Centres*. Swiss Federal Office of Energy SFOE.

Coroamă, Vlad C., and Lorenz M. Hilty. 2014. 'Assessing Internet Energy Intensity: A Review of Methods and Results'. *Environmental Impact Assessment Review* 45 (February): 63–68. https://doi.org/10.1016/j.eiar.2013.12.004.

Coroamă, Vlad C., Simon Hinterholzer, Kejsi Progni, Oana Dumbravă, and Ralph Hintemann. 2025. *Energy Efficiency of Servers: Past and Possible Future Trends*. IEA 4E TCP Efficient, Demand Flexible Networked Appliances (EDNA). https://www.iea-4e.org/wp-content/uploads/2025/05/EDNA-EE-of-servers-FINAL.pdf.

Coroamă, Vlad C., and Daniel Pargman. 2020. 'Skill Rebound: On an Unintended Effect of Digitalization'. *Proceedings of the 7th International Conference on ICT for Sustainability* (New York, NY, USA), ICT4S2020, June 21, 213–19. https://doi.org/10.1145/3401335.3401362.

Coroama, Vlad C., Daniel Schien, Chris Preist, and Lorenz M. Hilty. 2015. 'The Energy Intensity of the Internet: Home and Access Networks'. In *ICT Innovations for Sustainability*, edited by Lorenz M. Hilty and Bernard Aebischer, vol. 310. Advances in Intelligent Systems and Computing. Springer International Publishing. https://doi.org/10.1007/978-3-319-09228-7_8.

Corona, Blanca, Li Shen, Denise Reike, Jesús Rosales Carreón, and Ernst Worrell. 2019. 'Towards Sustainable Development through the Circular Economy—A Review and Critical Assessment on Current Circularity Metrics'. *Resources, Conservation and Recycling* 151 (December): 104498. https://doi.org/10.1016/j.resconrec.2019.104498.

Cotswold. 2025. 'When To Replace Your Waterproof Jacket'. Cotswold Outdoor. https://www.cotswoldoutdoor.com/the-knowledge/walking/when-to-replace-your-waterproof.html.

Davy, Benjamin. 2021. 'Building an AWS EC2 Carbon Emissions Dataset'. *Teads Engineering*, September 23. https://medium.com/teads-engineering/building-an-aws-ec2-carbon-emissions-dataset-3f0fd76c98ac.

De Koninck, Michiel De. 2024. 'How LLMs Access Real-Time Data from the Web'. *ML6team*, May 23. https://blog.ml6.eu/how-llms-access-real-time-data-from-the-web-a85ae544dc01.

DeepSeek. 2025. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. https://github.com/deepseek-ai/DeepSeek-R1/blob/main/DeepSeek_R1.pdf.

DEFRA. 2023. 'Greenhouse Gas Reporting: Conversion Factors 2023'. GOV.UK, Department for Energy Security and Net Zero, June 28. https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2023.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding'. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, edited by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Donachie, Paul Hennry John, Beth Barnes, Mike Burdon, and John C. Buchan. 2025. *National Cataract Audit for the 2023 NHS Year: 01 April 2023 to 31 March 2024*. National Ophthalmology Database Audit. The Royal College of Ophthalmologists. https://nodaudit.org.uk/sites/default/files/2025-07/NOD%20Cataract%20Audit%208th%20Annual%20Report%202025%20Final.pdf.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, et al. 2020. ‘An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale’. Paper presented at International Conference on Learning Representations. October 2. https://openreview.net/forum?id=YicbFdNTTy.

Dransfield, Bob, and Bob Brightwell. 2003. ‘The Lognormal Distribution’. *Influential Points*. https://influentialpoints.com/Training/the-log_normal_distribution.htm.

Ecoinvent association. 2025. ‘Market for Transport, Freight, Sea, Container Ship, Heavy Fuel Oil [Dataset 21045, Cutoff Version 3.11]’. https://ecoquery.ecoinvent.org/3.11/cutoff/dataset/21045.

EKZ. 2026. *EKZ*. https://www.ekz.ch/de/angebote/strom.html.

Elsworth, Cooper, Keguo Huang, David Patterson, et al. 2025. ‘Measuring the Environmental Impact of Delivering AI at Google Scale’. arXiv:2508.15734. Preprint, arXiv, August 21. https://doi.org/10.48550/arXiv.2508.15734.

EPRI. 2024. *Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption*. https://www.epri.com/research/products/3002028905.

Falk, Sophia, David Ekchajzer, Thibault Pirson, et al. 2025. ‘More than Carbon: Cradle-to-Grave Environmental Impacts of GenAI Training on the Nvidia A100 GPU’. arXiv:2509.00093. Preprint, arXiv, August 27. https://doi.org/10.48550/arXiv.2509.00093.

Farzan, Mireille, and Susanna Kallio. 2024. *A Transparent and Standards-Based Way to Assess the Environmental Impact of AI Systems*. Nokia. https://onestore.nokia.com/asset/214115.

Fava, James A., Andrea Smerek, Almut B. Heinrich, and Laura Morrison. 2014. ‘The Role of the Society of Environmental Toxicology and Chemistry (SETAC) in Life Cycle Assessment (LCA) Development and Application’. In *Background and Future Prospects in Life Cycle Assessment*, edited by Walter Klöpffer. Springer Netherlands. https://doi.org/10.1007/978-94-017-8697-3_2.

Fedus, William, Barret Zoph, and Noam Shazeer. 2022. ‘Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity’. arXiv:2101.03961. Preprint, arXiv, June 16. https://doi.org/10.48550/arXiv.2101.03961.

Fernandez, Jared, Clara Na, Vashisth Tiwari, Yonatan Bisk, Sasha Luccioni, and Emma Strubell. 2025. ‘Energy Considerations of Large Language Model Inference and Efficiency Optimizations’. arXiv:2504.17674. Preprint, arXiv, April 24. https://doi.org/10.48550/arXiv.2504.17674.

Freitag, Charlotte, Mike Berners-Lee, Kelly Widdicks, Bran Knowles, Gordon S. Blair, and Adrian Friday. 2021. ‘The Real Climate and Transformative Impact of ICT: A Critique of Estimates, Trends, and Regulations’. *Patterns* 2 (9). https://doi.org/10.1016/j.patter.2021.100340.

Friday, Adrian, Christina Bremer, Oliver Bates, Christian Remy, Srinjoy Mitra, and Jan Tobias Muehlberg. 2024. ‘The Belief in Moore’s Law Is Undermining ICT Climate Action’. arXiv:2411.17391. Preprint, arXiv, November 27. https://doi.org/10.48550/arXiv.2411.17391.

Google. 2024. ‘What Is Retrieval-Augmented Generation (RAG)?’ *Google Cloud*. https://cloud.google.com/use-cases/retrieval-augmented-generation.

Google Cloud. 2025. ‘What Is MLOps?’ *Google*. https://cloud.google.com/discover/what-is-mlops.

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. ‘The Llama 3 Herd of Models’. arXiv:2407.21783. Preprint, arXiv, November 23. https://doi.org/10.48550/arXiv.2407.21783.

Gu, Albert, and Tri Dao. 2024. ‘Mamba: Linear-Time Sequence Modeling with Selective State Spaces’. arXiv:2312.00752. Preprint, arXiv, May 31. https://doi.org/10.48550/arXiv.2312.00752.

Hada, Rishav. 2025. ‘How to Build an Ideal Tech Stack for LLM Applications’. *Future AGI*, June 24. https://futureagi.com/blogs/llm-application-tech-stack-2025.

Higgins, Andrew. 2024. ‘What Is Water Usage Effectiveness (WUE) in Data Centers?’ *Interconnections - The Equinix Blog*, November 13. https://blog.equinix.com/blog/2024/11/13/what-is-water-usage-effectiveness-wue-in-data-centers/.

Higham, Aisling. 2023. ‘Automated Clinical Conversations across the Cataract Pathway with an Artificial Intelligence (AI) Conversation Agent: A UK Regional Service Evaluation Protocol’. Preprint, medRxiv, June 15. https://doi.org/10.1101/2023.06.14.23291399.

Hilty, Lorenz M. 2008. *Information Technology and Sustainability: Essays on the Relationship between Information Technology and Sustainable Development*. BoD – Books on Demand.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel. 2020. ‘Denoising Diffusion Probabilistic Models’. arXiv:2006.11239. Preprint, arXiv, December 16. https://doi.org/10.48550/arXiv.2006.11239.

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, et al. 2022. ‘Training Compute-Optimal Large Language Models’. arXiv:2203.15556. Preprint, arXiv, March 29. https://doi.org/10.48550/arXiv.2203.15556.

Holzapfel, Peter, Vanessa Bach, and Matthias Finkbeiner. 2023. ‘Electricity Accounting in Life Cycle Assessment: The Challenge of Double Counting’. *The International Journal of Life Cycle Assessment* 28 (7): 771–87. https://doi.org/10.1007/s11367-023-02158-w.

Horner, Nathaniel C., Arman Shehabi, and Inês L. Azevedo. 2016. ‘Known Unknowns: Indirect Energy Effects of Information and Communication Technology’. *Environmental Research Letters* 11 (10): 103001. https://doi.org/10.1088/1748-9326/11/10/103001.

Howard, Jeremy, and Sebastian Ruder. 2018. ‘Universal Language Model Fine-Tuning for Text Classification’. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Iryna Gurevych and Yusuke Miyao. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1031.

HPE. 2021. ‘HPE Product Carbon Footprint – HPE ProLiant DL345 Gen10 Plus Server Data Sheet’. PSNow, October. https://www.hpe.com/psnow/doc/a50005151enw.

Hugging Face. 2025. *AI Energy Score*. https://huggingface.github.io/AIEnergyScore/.

IEA. 2025. *Energy and AI*. https://www.iea.org/reports/energy-and-ai.

ISO. 2006a. *DIN EN ISO 14040*. Beuth Verlag GmbH, October. https://doi.org/10.31030/1555059.

ISO. 2006b. *ISO 14044: Environmental Management — Life Cycle Assessment — Requirements and Guidelines*. International Organization for Standardization. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/03/84/38498.html.

ISO. 2018. *ISO 14067: Greenhouse Gases — Carbon Footprint of Products — Requirements and Guidelines for Quantification*. https://www.iso.org/standard/71206.html.

ISO/IEC. 2023. *ISO/IEC 5338:2023*. December. https://www.iso.org/standard/81118.html.

ITU. 2024. *AI and the Environment - International Standards for AI and the Environment*. https://www.itu.int/dms_pub/itu-t/opb/env/T-ENV-ENV-2024-1-PDF-E.pdf.

Jegen, Maya. 2024. 'Life Cycle Assessment: From Industry to Policy to Politics'. *The International Journal of Life Cycle Assessment* 29 (4): 597–606. https://doi.org/10.1007/s11367-023-02273-8.

JRC. 2010. *ILCD Handbook: General Guide for Life Cycle Assessment - General Guidance*. https://eplca.jrc.ec.europa.eu/uploads/ILCD-Handbook-General-guide-for-LCA-DETAILED-GUIDANCE-12March2010-ISBN-fin-v1.0-EN.pdf.

Jumper, John, Richard Evans, Alexander Pritzel, et al. 2021. 'Highly Accurate Protein Structure Prediction with AlphaFold'. *Nature* 596 (7873): 583–89. https://doi.org/10.1038/s41586-021-03819-2.

Kamiya, George. 2020. *The Carbon Footprint of Streaming Video: Fact-Checking the Headlines*. December 11. https://www.iea.org/commentaries/the-carbon-footprint-of-streaming-video-fact-checking-the-headlines.

Kamiya, George, and Vlad C. Coroamă. 2025. *Data Centre Energy Use: Critical Review of Models and Results*. IEA 4E TCP Efficient, Demand Flexible Networked Appliances (EDNA). https://www.iea-4e.org/wp-content/uploads/2025/01/Data-Centre-Energy-Use-Critical-Review-of-Models-and-Results.pdf.

Kaur, Jagreet. 2025. 'Agentic AI Infrastructure Stack for Agentic Systems'. *Xenonstack*, October 8. https://www.xenonstack.com/blog/ai-agent-infrastructure-stack.

Khattab, Omar, Arnav Singhvi, Paridhi Maheshwari, et al. 2023. 'DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines'. arXiv:2310.03714. Preprint, arXiv, October 5. https://doi.org/10.48550/arXiv.2310.03714.

Kirchherr, Julian, Denise Reike, and Marko Hekkert. 2017. 'Conceptualizing the Circular Economy: An Analysis of 114 Definitions'. *Resources, Conservation and Recycling* 127 (December): 221–32. https://doi.org/10.1016/j.resconrec.2017.09.005.

Kwon, Woosuk. 2025. 'vLLM: An Efficient Inference Engine for Large Language Models'. University of California, Berkeley. https://www2.eecs.berkeley.edu/Pubs/TechRpts/2025/EECS-2025-192.pdf.

Lacoste, Alexandre, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. 'Quantifying the Carbon Emissions of Machine Learning'. arXiv:1910.09700. Preprint, arXiv, November 4. https://doi.org/10.48550/arXiv.1910.09700.

LangChain. 2023. 'Tutorial: ChatGPT Over Your Data'. *LangChain Blog*, February 6. https://blog.langchain.com/tutorial-chatgpt-over-your-data/.

Lee, Robert A. 2025. 'ChatGPT vs. Google Gemini Statistics 2026: Head-to-Head AI Trends'. *SQ Magazine*, August 28. https://sqmagazine.co.uk/chatgpt-vs-google-gemini-statistics/.

Lei, Nuoa, Jun Lu, Arman Shehabi, and Eric Masanet. 2025. 'The Water Use of Data Center Workloads: A Review and Assessment of Key Determinants'. *Resources, Conservation and Recycling* 219 (June): 108310. https://doi.org/10.1016/j.resconrec.2025.108310.

Lei, Nuoa, and Eric Masanet. 2022. 'Climate- and Technology-Specific PUE and WUE Estimations for U.S. Data Centers Using a Hybrid Statistical and Thermodynamics-Based Approach'.

Resources, Conservation and Recycling 182 (July): 106323. https://doi.org/10.1016/j.resconrec.2022.106323.

Lemarchal, Julien, and Reyan Laifa. 2025. 'The Complete Guide to the AWS AI Agentic Ecosystem'. *Devoteam*. https://www.devoteam.com/expert-view/aws-ai-agentic-ecosystem/.

Lepikhin, Dmitry, HyoukJoong Lee, Yuanzhong Xu, et al. 2020. 'GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding'. arXiv:2006.16668. Preprint, arXiv, June 30. https://doi.org/10.48550/arXiv.2006.16668.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, et al. 2020. 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks'. *Advances in Neural Information Processing Systems* 33: 9459–74. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Li, Pengfei, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. 2025. 'Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models'. arXiv:2304.03271. Preprint, arXiv, January 15. https://doi.org/10.48550/arXiv.2304.03271.

Lin, Chin-Yew. 2004. 'ROUGE: A Package for Automatic Evaluation of Summaries'. *Proceedings of the Workshop on Text Summarization Branches Out*. https://aclanthology.org/W04-1013.pdf.

Lin, Leonard. 2025. 'Power Usage and Energy Efficiency'. *Llm-Tracker*, May 5. https://llm-tracker.info/_TOORG/Power-Usage-and-Energy-Efficiency.

Liu, Di, Meng Chen, Baotong Lu, et al. 2024. 'RetrievalAttention: Accelerating Long-Context LLM Inference via Vector Retrieval'. arXiv:2409.10516. Preprint, arXiv, December 31. https://doi.org/10.48550/arXiv.2409.10516.

Luccioni, Alexandra Sasha, Boris Gamazaychikov, Theo Alves da Costa, and Emma Strubell. 2025. 'Misinformation by Omission: The Need for More Environmental Transparency in AI'. arXiv:2506.15572. Preprint, arXiv, June 18. https://doi.org/10.48550/arXiv.2506.15572.

Luccioni, Alexandra Sasha, Boris Gamazaychikov, Sara Hooker, et al. 2024. 'Light Bulbs Have Energy Ratings — so Why Can't AI Chatbots?' *Nature* 632 (8026): 736–38. https://doi.org/10.1038/d41586-024-02680-3.

Luccioni, Alexandra Sasha, Sylvain Viguier, and Anne-Laure Ligozat. 2022. 'Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model'. arXiv:2211.02001. Preprint, arXiv, November 3. https://doi.org/10.48550/arXiv.2211.02001.

Makonin, Stephen, Laura U. Marks, Radek Przedpelski, Alejandro Rodriquez-Silva, and Ramy ElMallah. 2022. 'Calculating the Carbon Footprint of Streaming Media: Beyond the Myth of Efficiency'. Paper presented at Eighth Workshop on Computing within Limits 2022. *Computing within Limits*, June 21. https://doi.org/10.21428/bf6fb269.7625cc76.

Martineau, Kim. 2023. 'What Is Retrieval-Augmented Generation (RAG)?' *IBM Research*, August 22. https://research.ibm.com/blog/retrieval-augmented-generation-RAG.

Masanet, Eric, Nuoa Lei, and Jonathan Koomey. 2024. 'To Better Understand AI's Growing Energy Use, Analysts Need a Data Revolution'. *Joule* 8 (9): 2427–36. https://doi.org/10.1016/j.joule.2024.07.018.

Meinert, Edward, Madison Miline-Ives, Ernest Lim, et al. 2024. 'Accuracy and Safety of an Autonomous Artificial Intelligence Clinical Assistant Conducting Telemedicine Follow-up Assessment

for Cataract Surgery'. *eClinicalMedicine* 73 (102692). https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(24)00271-2/fulltext.

Merritt, Rick. 2025. 'What Is Retrieval-Augmented Generation, Aka RAG?' *NVIDIA Blog*, January 31. https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/.

Mistral AI. 2025. *Our Contribution to a Global Environmental Standard for AI*. July 22. https://mistral.ai/news/our-contribution-to-a-global-environmental-standard-for-ai.

Mytton, David. 2020a. 'Assessing the Suitability of the Greenhouse Gas Protocol for Calculation of Emissions from Public Cloud Computing Workloads'. *Journal of Cloud Computing* 9 (1): 45. https://doi.org/10.1186/s13677-020-00185-8.

Mytton, David. 2020b. 'Hiding Greenhouse Gas Emissions in the Cloud'. *Nature Climate Change* 10 (8): 701. https://doi.org/10.1038/s41558-020-0837-6.

Nvidia. 2025. *Product Carbon Footprint (PCF) Summary for NVIDIA HGX H100*. https://images.nvidia.com/aem-dam/Solutions/documents/HGX-H100-PCF-Summary.pdf.

O'Brien, Isabel. 2024. 'Data Center Emissions Probably 662% Higher than Big Tech Claims. Can It Keep up the Ruse?' Technology. *The Guardian*, September 15. https://www.theguardian.com/technology/2024/sep/15/data-center-gas-emissions-tech.

OpenAI. 2025. 'API Pricing'. https://openai.com/api/pricing/.

OpenAI. 2026. *Pricing*. https://developers.openai.com/api/docs/pricing.

Paccou, Rémi, and Fons Wijnhoven. 2024. *Artificial Intelligence and Electricity: A System Dynamics Approach*. https://www.se.com/ww/en/insights/sustainability/sustainability-research-institute/artificial-intelligence-electricity-system-dynamics-approach/?campaign_objective=awareness&mcl_name=sustainability.

Park, Joon Sung, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. 'Generative Agents: Interactive Simulacra of Human Behavior'. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (New York, NY, USA), UIST '23, October 29, 1–22. https://doi.org/10.1145/3586183.3606763.

Patterson, David, Joseph Gonzalez, Urs Hölzle, et al. 2022. 'The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink'. *Computer* 55 (7): 18–28. https://doi.org/10.1109/MC.2022.3148714.

Peer, Rebecca A. M., Emily Grubert, and Kelly T. Sanders. 2019. 'A Regional Assessment of the Water Embedded in the US Electricity System'. *Environmental Research Letters* 14 (8): 084014. https://doi.org/10.1088/1748-9326/ab2daa.

Peham, Thomas. 2024. 'Can GPT-4 Browse The Internet?' *Otterly.AI Blog - Best AI Search Monitoring Solution*, May 3. http://otterly.ai/blog/can-gpt-4-browse-the-internet/.

Pohl, Johanna, Lorenz M. Hilty, and Matthias Finkbeiner. 2019. 'How LCA Contributes to the Environmental Assessment of Higher Order Effects of ICT Application: A Review of Different Approaches'. *Journal of Cleaner Production* 219 (May): 698–712. https://doi.org/10.1016/j.jclepro.2019.02.018.

Preist, Chris, Dan Schien, Paul Shabajee, Stephen Wood, and Christopher Hodgson. 2014. 'Analyzing End-to-End Energy Consumption for Digital Services'. *Computer* 47 (5): 92–95. https://doi.org/10.1109/MC.2014.110.

Rainio, Oona, Jarmo Teuho, and Riku Klén. 2024. ‘Evaluation Metrics and Statistical Tests for Machine Learning’. *Scientific Reports* 14 (1): 6086. https://doi.org/10.1038/s41598-024-56706-x.

Ramos, Rafael. 2025. ‘Measuring AI Model Performance: Tokens per Second, Model Sizes, and Inferencing Tools’. *OpenMetal IaaS*, May 20. https://openmetal.io/resources/blog/ai-model-performance-tokens-per-second/.

Ren, Shaolei, Bill Tomlinson, Rebecca W. Black, and Andrew W. Torrance. 2024. ‘Reconciling the Contrasting Narratives on the Environmental Impact of Large Language Models’. *Scientific Reports* 14 (1): 26310. https://doi.org/10.1038/s41598-024-76682-6.

Roberts, Dan. 2025. ‘9 Years to AGI? OpenAI’s Dan Roberts Reasons About Emulating Einstein’. 9 Years to AGI? OpenAI’s Dan Roberts Reasons About Emulating Einstein, May 8. https://www.youtube.com/watch?v=_rjD_2zn2JU.

Rolnick, David, Priya L. Donti, Lynn H. Kaack, et al. 2022. ‘Tackling Climate Change with Machine Learning’. *ACM Computing Surveys* 55 (2): 42:1-42:96. https://doi.org/10.1145/3485128.

Samsi, Siddharth, Dan Zhao, Joseph McDonald, et al. 2023. ‘From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference’. arXiv.Org, October 4. https://arxiv.org/abs/2310.03003v1.

Schaubroeck, Thomas. 2023. ‘Relevance of Attributional and Consequential Life Cycle Assessment for Society and Decision Support’. *Frontiers in Sustainability* 4 (July). https://doi.org/10.3389/frsus.2023.1063583.

Schien, Daniel, Vlad C. Coroama, Lorenz M. Hilty, and Chris Preist. 2015. ‘The Energy Intensity of the Internet: Edge and Core Networks’. In *ICT Innovations for Sustainability*, edited by Lorenz M. Hilty and Bernard Aebischer, vol. 310. Advances in Intelligent Systems and Computing. Springer International Publishing. https://doi.org/10.1007/978-3-319-09228-7_9.

Schien, Daniel, Paul Shabajee, Huseyin Burak Akyol, Luke Benson, and Angeliki Katsenou. 2024a. ‘Assessing the Carbon Reduction Potential for Video Streaming from Short-Term Coding Changes’. *2024 16th International Conference on Quality of Multimedia Experience (QoMEX)*, June, 22–28. https://doi.org/10.1109/QoMEX61742.2024.10598286.

Schien, Daniel, Paul Shabajee, Huseyin Burak Akyol, Luke Benson, and Angeliki Katsenou. 2024b. ‘Assessing the Carbon Reduction Potential for Video Streaming from Short-Term Coding Changes’. *2024 16th International Conference on Quality of Multimedia Experience (QoMEX)*, June, 22–28. https://doi.org/10.1109/QoMEX61742.2024.10598286.

Schien, Daniel, Paul Shabajee, Louise Krug, Greg McSorley, and Chris Preist. 2025. ‘Causal Allocation of Fixed Impacts in Product Systems: Assessing the Effect of Data Demand on Network Energy Consumption’. *Journal of Industrial Ecology* 29 (5): 1618–31. https://doi.org/10.1111/jiec.70057.

Schien, Daniel, Paul Shabajee, Mike Yearworth, and Chris Preist. 2013. ‘Modeling and Assessing Variability in Energy Consumption During the Use Stage of Online Multimedia Services’. *Journal of Industrial Ecology* 17 (6): 800–813. https://doi.org/10.1111/jiec.12065.

Schmid, Nicolas, Vlad C. Coroamă, Oana Dumbravă, et al. 2025. *Carbon leakage in AI-driven data center growth? An assessment of drivers and barriers to the localization of data center operations and investments with respect to carbon pricing policies*. No. 68/2025. Umweltbundesamt. https://www.umweltbundesamt.de/publikationen/carbon-leakage-in-ai-driven-data-center-growth.

Schmidhuber, Jürgen. 2015. ‘Deep Learning in Neural Networks: An Overview’. *Neural Networks* 61 (January): 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.

Schneider, Ian, Hui Xu, Stephan Benecke, et al. 2025. ‘An Introduction to Life-Cycle Emissions of AI Hardware’. *IEEE Micro*, 1–10. https://doi.org/10.1109/MM.2025.3592568.

Schwartz, Roy, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. ‘Green AI’. *Communications of the ACM*, ahead of print, November 17. New York, NY, USA. https://doi.org/10.1145/3381831.

Shehabi, Arman, Sarah J. Smith, Alex Hubbard, et al. 2024. *2024 United States Data Center Energy Usage Report*. LBNL-2001637. https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf.

Shinn, Noah, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. ‘Reflexion: Language Agents with Verbal Reinforcement Learning’. *Advances in Neural Information Processing Systems* 36 (December): 8634–52. https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Sotos, Mary. 2015. *GHG Protocol Scope 2 Guidance*. World Resources Institute. https://ghgprotocol.org/scope-2-guidance.

Stadt Zürich. 2025. ‘Zü-Re I KI-Chatbot’. https://zue-re.stadt-zuerich.ch/.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. ‘Energy and Policy Considerations for Deep Learning in NLP’. arXiv:1906.02243. Preprint, arXiv, June 5. https://doi.org/10.48550/arXiv.1906.02243.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2020. ‘Energy and Policy Considerations for Modern Deep Learning Research’. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (09): 09. https://doi.org/10.1609/aaai.v34i09.7123.

Stryker, Cole. 2025. ‘What Is Agentic AI?’ *IBM*, February 24. https://www.ibm.com/think/topics/agentic-ai.

Thomas, Lillianne, and Ryan Avery. 2023. *Transfer Learning, Fine-Tuning and Hyperparameter Tuning — Deep Learning with TensorFlow*. Development Seed. https://developmentseed.org/tensorflow-eo-training-2/docs/Lesson7c_transfer_learning_hyperparam_opt.html.

Tozzi, Christopher. 2025. ‘A Guide to Data Center Water Usage Effectiveness (WUE) and Best Practices’. *DataCenter Knowledge*, January 17. https://www.datacenterknowledge.com/cooling/a-guide-to-data-center-water-usage-effectiveness-wue-and-best-practices.

Typedef. 2025. ‘13 LLM Adoption Statistics: Critical Data Points for Enterprise AI Implementation in 2025’. *Typedef Blog*, October 7. https://typedef.ai/resources/llm-adoption-statistics.

Ufonia. 2025. ‘Dora: We Automate Routine Care Pathways’. https://www.ufonia.com/uk/what-we-do.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. 2017. ‘Attention Is All You Need’. Paper presented at Neural Information Processing Systems. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. https://doi.org/10.48550/arXiv.1706.03762.

Verdecchia, Roberto, June Sallou, and Luís Cruz. 2023. ‘A Systematic Review of Green AI’. *WIREs Data Mining and Knowledge Discovery* 13 (4): e1507. https://doi.org/10.1002/widm.1507.

Vries, Alex de. 2023. ‘The Growing Energy Footprint of Artificial Intelligence’. *Joule* 7 (10): 2191–94. https://doi.org/10.1016/j.joule.2023.09.004.

Wang, Haonan, Xuxin Xiao, Mingyu Yan, et al. 2025. ‘A Systematic Characterization of LLM Inference on GPUs’. arXiv:2512.01644. Preprint, arXiv, December 1. https://doi.org/10.48550/arXiv.2512.01644.

WBCSD and WRI. 2004. *A Corporate Accounting and Reporting Standard*. https://ghgprotocol.org/corporate-standard.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, et al. 2023. ‘Chain-of-Thought Prompting Elicits Reasoning in Large Language Models’. arXiv:2201.11903. Preprint, arXiv, January 10. https://doi.org/10.48550/arXiv.2201.11903.

Wilkins, Joe. 2025. ‘Sam Altman Admits That Saying “Please” and “Thank You” to ChatGPT Is Wasting Millions of Dollars in Computing Power’. *Futurism*, April 19. https://futurism.com/altman-please-thanks-chatgpt.

Williams, Eric. 2011. ‘Environmental Effects of Information and Communications Technologies’. *Nature* 479 (7373): 354–58. https://doi.org/10.1038/nature10682.

WRI and WBCSD. 2011. *Product Life Cycle Accounting and Reporting Standard*. https://ghgprotocol.org/sites/default/files/standards/Product-Life-Cycle-Accounting-Reporting-Standard_041613.pdf.

Wu, Carole-Jean, Ramya Raghavendra, Udit Gupta, et al. 2022. ‘Sustainable AI: Environmental Implications, Challenges and Opportunities’. arXiv:2111.00364. Preprint, arXiv, January 9. https://doi.org/10.48550/arXiv.2111.00364.

Yao, Shunyu, Jeffrey Zhao, Dian Yu, et al. 2022. ‘ReAct: Synergizing Reasoning and Acting in Language Models’. Paper presented at The Eleventh International Conference on Learning Representations. September 29. https://openreview.net/forum?id=WE_vluYUL-X.

You, Josh. 2025a. ‘How Much Energy Does ChatGPT Use?’ *Epoch AI*, February 7. https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use.

You, Josh. 2025b. ‘Most of OpenAI’s 2024 Compute Went to Experiments’. *Epoch AI*, October 10. https://epoch.ai/data-insights/openai-compute-spend.

A Environmental lifecycle assessment

The discipline of industrial environmental impact assessment traces its beginnings back to the 1970s, both to the scientific origins driven by the Society of Environmental Toxicology and Chemistry (SETAC) (Fava et al. 2014) and a manufacturing and engineering lineage from the late 1960s (Jegen 2024) with the goal to assess and improve environmental performance of products and inform environmental decision-making. The two disciplines converged under the term life cycle assessment (LCA) in the 1990s and were subsequently standardised via the ISO 1404x set of standards (ISO 2006a, 14040, 2006b, 14044).

Central to the ontology of LCA is the segmentation of environmentally relevant aspects of a product or service into a connected network of `unit processes' during which environmentally relevant inputs of energy and materials are converted into environmentally relevant flows, such as greenhouse gases and chemical flows to land, water and air. The aspiration is for these networks to

a) comprehensively cover everything economic activity directly connected to a product – from raw material extraction through manufacturing, distribution, use, and end-of-life – and

b) to form a closed loop, within which materials are retained in a cycle.

The study of opportunities to close any open loops has led to the field of circular economy (Corona et al. 2019).

The quantification of flows to the environment is normalised relative to the flow of product (and by-products) through the concept of `exchanges'. For example, if a container shipping company (transporting, among others, GPUs or CPUs for datacentres) has bought diesel for their fleet of cargo ships that jointly have travelled x-thousand kilometres and transported y-thousand tonnes of cargo, then they might normalise the diesel consumption relative to 1 tonnes-kilometre of container shipping (Ecoinvent association 2025).

Originally developed for physical products, the principles of LCA can be consistently applied to services as well. Even though we are particularly interested in the assessment of digital services, we continue to refer to 'products' during this section as they relate easier to the necessarily physical nature of exchanges.

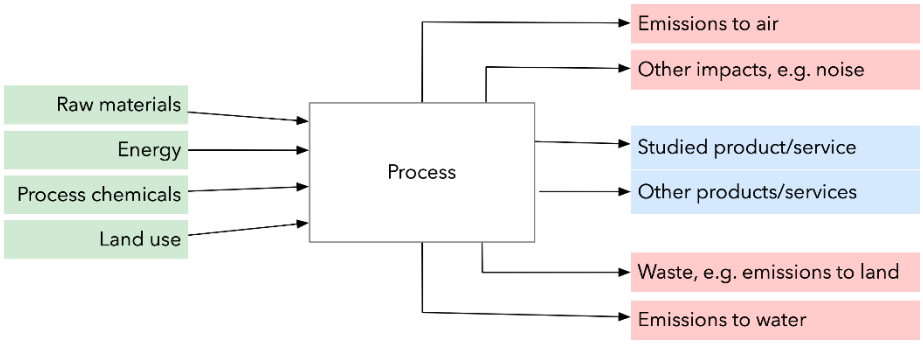


Figure 27: LCA process model with input and output flows to eco and technosphere.

The conceptual framing of the unit process has direct epistemological implications for LCA. The representation of unit processes in a network grounds the discipline in the construction of models that abstract from actual economy and industry through formulation of typical, yet variable, and thus uncertain exchanges. While intuitive, familiar and established, we can recognise that the approach to model services as a network of individual processes is not the only option. For example, an alternative approach (with its own complications) could be the observation of the environment in its entirety.

Further, the choice how to normalise flows of input materials and energy to product and environment (as exchanges) has profound implications on the interpretation of the output of LCAs. Two different lines of reasoning how processes relate to environmental impact are relevant to distinguish here: attributional and consequential (Schaubroeck 2023) modelling. The former analyses energy and chemical flows observed within the given system boundaries of the product system. For example, an attributional assessment of sending 1 GB of data through a network would assign a share of the total energy consumption by a network in some time frame (for example per month) in proportion of 1GB of the total data volume transported. This analytical approach has the appeal that can be evidenced with observational (measured) data, yet is retrospective. There are, however, some shortcomings with this – because the energy consumption by networks does not vary in the short term with demand (Schien et al. 2025) such an attributional assessment could not be used to evaluate the effects of short-term changes to to demand. This is common shortcoming with attributional assessments that generally applies to products system with fixed costs (Schaubroeck 2023).

An alternative approach, a so-called *consequential* assessment, is to explicitly study how *changes* to a product system would affect environmental impacts. This usually requires modelling, rather than evaluation by experiment. Because environmental impact is related to efficiency, and because efficiency, in turn, depends strongly on the scale at which they are implemented. In the context of networks, such a consequential assessment would study how demand increases the capacity required by network devices [ibid], and how changes in capacity can translate into energy savings. While consequential assessments are geared to be relevant for decision making, through their necessary speculative nature, they typically cannot be evidenced with observational data and are thus often considered more uncertain (Schaubroeck 2023). This uncertainty increases, the further into the future the consequences of a decision should be considered. For this reason, consequential approaches in the ICT sector tend to look at the short term (Schien et al. 2024b).

In established markets engineering decisions, by definition, tend to not result in structural changes. The ILCD handbook refers to these as ‘micro’ decision contexts (JRC 2010). For these, the use of attributional footprints is appropriate. However, in the case of rapidly-changing GenAI services, engineering decision and innovations contribute directly to a new or better-quality capabilities that affect user-behaviour and change markets at rapid pace. In the language of the ILCD Handbook, this means that GenAI engineering decisions should be considered as meso-level decisions (with structural implications) and be evaluated in a consequential manner. However, consequential assessments usually require the consideration of market data, which is not commonly publicly available. In this report we thus, follow standard practice and apply attributional logic for the modelling work, and highlight consequential considerations where appropriate.

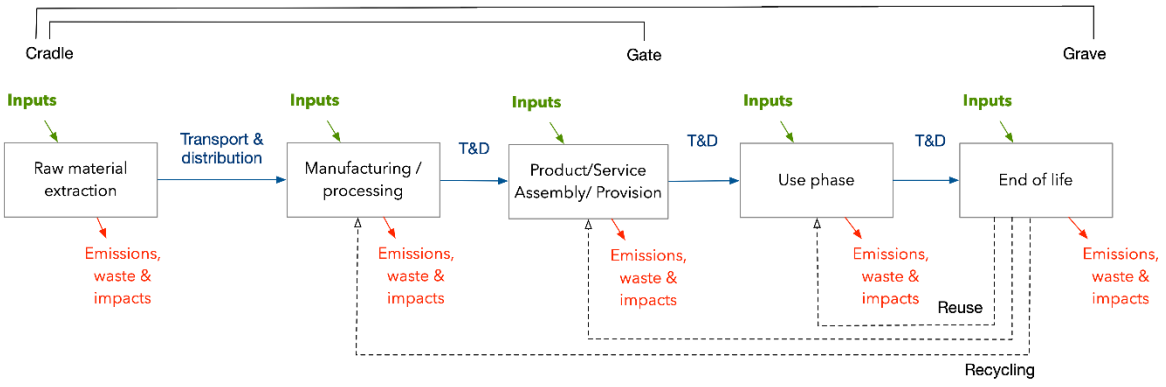


Figure 28: LCA Stages including aggregated process flows and reuse and recycling circularity principles.

As argued above, LCA conceives environmental impact as a network of processes converting energy and material flows into products. Depending on the boundaries placed around these networks, LCA can be applied to calculate environmental performance of individual products and services, all the way to

entire organisations. These are usually overlapping, since organisational processes required for a specific product or service are also part of the processes of the organization footprint as a whole. Both product and organisational assessments have been standardised. They are briefly presented below.

Organisational assessment

For the assessment of organisations, the most important standard is the GHG protocol (WBCSD and WRI 2004); describing the environmental impact of business XYZ *as a whole*. One of the most widely known concepts established by the GHG protocol is to convention to categorise emissions into scopes, namely "scope 1" (e.g. combustion of fossil fuel in owned generators), "scope 2" (electricity consumption), and "scope 3" (GHG emissions upstream – suppliers -- and down-stream in the supply chain -- consumers). Organisational GHG footprinting is a legal requirement for many organisations, for example for:

- mandatory reporting, such as the EU Corporate Sustainability Reporting Directive (CSRD) or the UK's Streamlined Energy and Carbon Reporting (SECR; previously the Carbon Reduction Commitment Energy Efficiency Scheme)
- Voluntary disclosure, via the Carbon Disclosure Project (CDP), or
- Environmental target setting, such as the Science-Based Targets initiative (SBTi)

Through these and other uses, corporate GHG accounting is a routine input into corporate social responsibility and other non-financial reporting.

Product assessment

Assessments of specific products and services do not use the concept of scope. Instead, the LCA boundary logic of 'cradle-to-gate' 'cradle-to-grave' are established by the product carbon footprint (PCF) standards; most importantly (BSI Knowledge 2011; WRI and WBCSD 2011; ISO 2018). However, the concept of scope 3 is frequently used informally to describe impact upstream (supplier) or downstream (consumer) from a business carrying out a PCF. The assessment of upstream scope 3 impacts is a central challenge for the assessment of cloud-based GenAI services (see Section 4.3).

PCF can be used for a wide set of applications:

- Internal management and design; Hotspot analysis, Eco-design,
- Use for product comparison (external) and eco-design (internal)
- Procurement decisions

Important underlying principles are:

- Relevance – the modelling must reflect the product's real function and impacts. This principle underlies the attention placed in PCF on defining the product function via the so-called functional unit.
- Consistency
- Completeness
- Accuracy
- Transparency

Functional unit (FU)

In this context, the functional unit (FU) represents a "quantified performance of a product system for use as a reference unit" (ISO 2018). The choice of FU needs to be relevant as a description of the actual

use of the product in the market. Without further information, between two alternative FUs, the one describing the more likely use is more relevant.

The FU has significant impact on the outcome of an assessment as environmental and product flows are normalised relative to it. In a basic attributional assessment, it could for example be assumed that the impact of an AI service generating one image is roughly 1/n of the total impact of n images generated in one month. However, in a consequential assessment a distinction might be made between a request outside and during periods of peak demand, and thus exchanges (energy consumption and embodied impacts) would be differently normalized to the service.

The functional unit plays a differently important role as the basis for any potential comparison of environmental performance with alternative products. A functional unit can support comparability of the environmental performance of alternative services if it describes *what the user receives* rather than what the system consumes. This can refer for example to “one pair of dried hands” when comparing paper hand towels with electric driers. In the digital domain, this could be e.g. “delivery of 1 hour of HD video content to one end-user device at 1080p resolution” instead of “1 GB of data streamed”. This example shows that the quality of the service is relevant when comparing two product systems.

The specificity of a functional unit definition depends on the purpose of a study:

- If the goal is a comparison of a service *with itself* (e.g., hotspot analysis and design support of a single product) a less precise, yet consistently applied functional unit is sufficient. For example, screen brightness (the luminescence) is a key determinant of energy consumption by display devices. The consequence of this could be to include the brightness of a scene as part of the functional unit of a streaming or an AI-image-generation service, if the goal is to understand hotspots across the entire value chain.
- However, if the goal is to *compare different services* (e.g., for environmental performance comparison or eco-labelling), then specificity of the functional unit definition is more important; and even more so, if the comparison should be made between independently conducted assessments.

Further, the definition of a functional unit for comparison must trade-off between comparability (by specifying attributes that are most deterministic on the environmental performance) and generality (as a precise definition narrows the set of use cases that are being described).

Between two otherwise identical functional units, the one with the more parsimonious set of attributes is preferable. However, the knowledge of what properties are most deterministic of the environmental impact might change as the environmental assessment progresses and should thus be allowed to evolve during an assessment.

There is also a practical side to this, as it is often difficult to quantify a quality attribute. For example, even if statistically consistent, the user’s judgement of the accuracy of a random GenAI response is subjective, and two users might prefer different responses that might appear to have the same factual content.

Finally, when analysing the benefits of substituting an AI service for another, non-AI service (such as a GenAI service substituting human text generation), the challenges identified above are compounded. This last point, however, is outside the scope of this work.