University of
Zurich^UZH

Katharina O. E. Müller, Chao Feng, Daria Schumm, Weijie Niu, Thomas Grubl, Andy Aidoo, Ahmad Abtahi, Reza Abtahi, Nasim Nezhadsistani, Franciso Enguix, Gökcan Cantali (Edts).

Internet Economics XIX

January 2026

TECHNICAL REPORT – No. IFI-2026.01

University of Zurich
Department of Informatics (IFI)
Binzmühlestrasse 14, CH-8050 Zürich, Switzerland

ifi

B. Stiller et al. (Eds.): Internet Economics XIX: IntEco Seminar
Technical Report No. IFI-2026.01, January 2026
Communication Systems Group (CSG)
Department of Informatics (IFI)
University of Zurich
Binzmühlestrasse 14, CH-8050 Zürich, Switzerland
URL: http://www.csg.uzh.ch/

Introduction

The Department of Informatics (IFI) of the University of Zurich, Switzerland works on research and teaching in the area of computer networks and communication systems. Communication systems include a wide range of topics and drive many research and development activities. Therefore, during the autumn term HS 2025, a new instance of the Internet Economics seminar has been prepared and students as well as supervisors worked on this topic.

Even today, Internet Economics are run rarely as a teaching unit. This observation seems to be a little in contrast to the fact that research on Internet Economics has been established as an important area in the center of technology and economics on networked environments. After some careful investigations it can be found that during the last ten years, the underlying communication technology applied for the Internet and the way electronic business transactions are performed on top of the network have changed. Although, a variety of support functionality has been developed for the Internet case, the core functionality of delivering data, bits, and bytes remained unchanged. Nevertheless, changes and updates occur with respect to the use, the application area, and the technology itself. Therefore, another review of a selected number of topics has been undertaken.

Seminar Operation

Based on well-developed experiences of former seminars, held in different academic environments, all interested students worked on an initially offered set of papers and book chapters. Those relate to the topic titles as presented in the Table of Content below. They prepared a written essay as a clearly focused presentation, an evaluation, and a summary of those topics. Each of these essays is included in this technical report as a separate section and allows for an overview of important areas of concern, sometimes business models in operation, and problems encountered.

In addition, every group of students prepared a slide presentation of approximately 45 minutes to present its findings and summaries to the audience of students attending the seminar and other interested students, research assistants, and professors. Following a general question and answer phase, a student-lead discussion debated open issues and critical statements with the audience.

Local IFI support for preparing talks, reports, and their preparation by students had been granted by Katharina O. E. Müller, Chao Feng, Daria Schumm, Weijie Niu, Thomas Grubl, Andy Aidoo, Ahmad Abtahi, Reza Abtahi, Nasim Nezhadsistani, Franciso Enguix, Gökcan Cantali, and Prof. Burkhard Stiller. In particular, many thanks are addressed to Chao Feng for organizing the seminar and for their strong commitment on getting this technical report ready and quickly published. A larger number of pre-presentation discussions have provided valuable insights in the emerging and moving field of communication systems, both for all groups of students and supervisors. Many thanks to all people contributing to the success of this event, which has happened in a lively group of highly motivated and technically qualified students and people.

Zurich, January 2026

Contents

Chapter 1

The Recent Advances of Retrieval Augmented Generation (RAG) Systems

Joshua Winterflood

Retrieval augmented generation (RAG) systems have emerged over the past 5 years as an ecosystem to assist large language models (LLMs). RAG systems may enhance LLMs in the pre-training, fine-tuning or inference stage. These three main branches form the basis of the recent advances of RAG systems, each addressing a different weakness of LLMs. In this report we will thouroughly investigate the most predominant branch - the inference branch - and discuss how the naive RAG pipeline evolved to modern RAG architectures. This will include a brief introduction of the naive RAG pipeline, where we discuss how embeddings can be used for semantic document retrievals, which are in turn provided to the LLM as context to enhance the LLM's generation phase. Additionally we will pinpoint advanced RAG techniques such as pre- and post-retrieval processes, which are widely employed to address problems which arise due to the retrieval's sensitiveness. Lastly we will investigate how RAG systems can be evaluated, such that a formal measurement of the advances within the RAG ecosystem is at hand.

Contents

1.1 Introduction

Over the past 5 years, retrieval augmented retrieval (RAG) systems have emerged as a leading branch of technology to enhance large language models (LLM) [6]. Prior to the establishment of RAG systems, LLMs performed poorly on realtime-dependent, knowledge-intensive and domain-specific tasks, leading to outdated and generalized responses. To address this issue three branches of RAG have emerged to address the encountered problems in different stages of the LLM deployment: fine-tuning, pre-training and inference. The pre-training branch is the least represented branch and aims to select datasets, which are then fed to the LLM during the training stage. Fine-tuning of LLMs is an intensive task, which can be done completely LLM-sided. However, RAG systems can aid this fine-tuning process, for instance by providing information which is formatted in a specific way [6]. This branch can enhance the LLM in a way, that less training time is needed for the same results and therefore less ressources are needed for the same quality of responses. The inference branch is the most predominant branch, and aims to retrieve specific documents, when given a prompt. These documents are then used alongside with the original query as the context for the LLM's generation phase [8]. By providing the LLM with the retrieved documents, the LLM is then able to generate a response iteratively, where a new token is generated based on all of the previous tokens, the original query and the retrieved documents [8]. The reliance of the stochastic token generation on previous tokens creates the phenomenon of hallucinations, where parts of the response seem reasonable, but are factually incorrect. Augmentation retrieval processes are common methods to mitigate such hallucinations. These processes leverage a judgement stage to asess the generated content so far [6], and perform possibly multiple retrievals based on said judgement. For the retrieval of documents, embeddings are used to facilitate semantic searches. The embedding process itself is a learned technique, using unsupervised constrastive learning [11]. By embedding the query into a high-dimensional vector, the top-k closest vectors from a vector database can be retrieved. Whenever the database contains massive amounts of vectors, approximate nearest neighbor (ANN) search methods are used to accelerate retrieval while maintaining acceptable accuracy. These embedding-based retrieval approaches form the foundation for modern information retrieval and retrieval-augmented generation systems.

1.2 Problem Statement

1.2.1 Limitations of current LLMs architecture

Large Language Models (LLMs) have established themselves as a highly useful tool for various applications. Especially for tasks, where the output quality is hard to measure, the generative characteristic of LLMs yield a robust approach in comparison to rule-based algorithms [7]. Prior to the emergence of LLMs, Recurrent Neural Networks (RNNs) established themselves as an effective aproach to produce coherent text, capturing context to assert the meaning of words [10]. However, due to their sequential processing nature, the performance of RNNs degraded drastically with the introduction of long-range dependencies. Tranformers revolutionized text generation by the means of attention mechanisms, which are able to capture context across multiple documents simultaneously. In the following sections, we pinpoint the gaps of the pre-training, fine-tuning and inference stage of LLMs, which sets the fundament for how RAG systems can be used to asisst LLMs.

1.2.1.1 Pre-Training

The pre-training phase in the development of LLMs encompasses feeding extensive amounts of text data to the model, from which the model learns patterns and gains its general

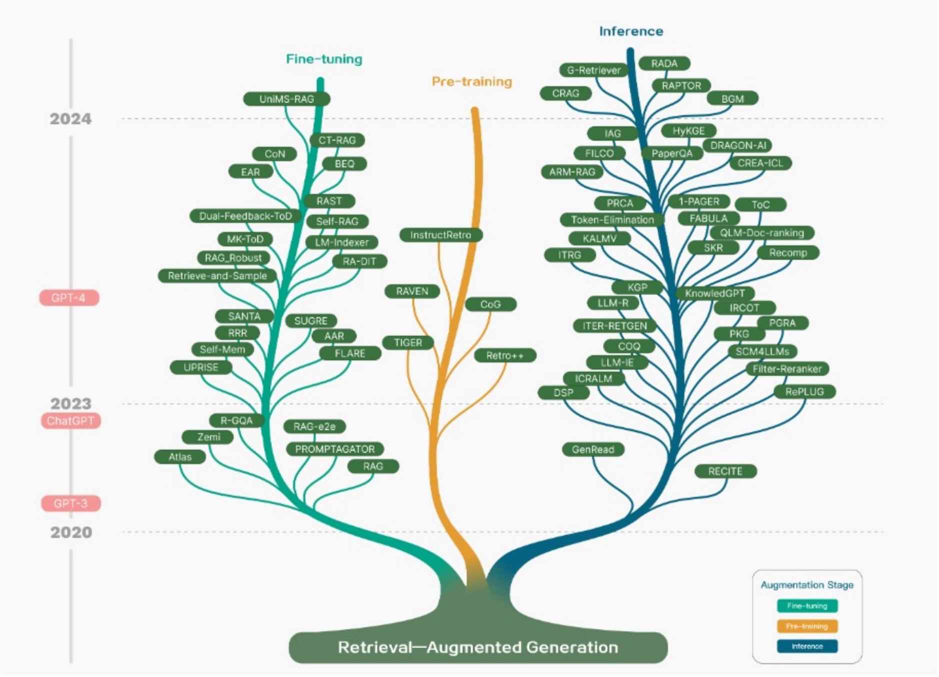


Figure 1.1: Evolution of the RAG ecosystem over the past 5 years and the emergence of the 3 main branches : pre-training, fine-tuning, inference

understanding of language, stored in the form of its parameters. Since this process involves aggregating large amounts of data from the internet, several issues might arise. Firstly, unwanted biases which are present in the collected documents are integrated into the models knowledge, which is reflected in the responses it generates during the inference stage [5](see section 1.2.1.3). Additionally the knowledge present on the internet exhibits a strong tendency towards beginner-level expertise [12]. Here the underlying reason is found within the social structure of beginners and experts within a domain. Beginners make up a large proportion of the entire community, which leads to more frequent discussions on beginner-level topics. Furthermore a beginner is typically in touch with a variety of superficial topics, which leads to even more discussions led on entry-niveau. These phenomenons result in a predominantly superficial, general and naive knowledge base for the model. To address these problematic phenomenons, a filtering strategy is required, to extract relevant information.

1.2.1.2 Fine-Tuning

Fine-tuning is a technique employed after the pre-training phase of LLMs and aims to redirect the model in a specific way. This redirection might take the form of tuning the model in a way, such that its output is formatted in a specific way. Depending on the interest, the model can be trained to produce predominantly tabular, qualitative, quantitative data. Furthermore the length and style of the response can be adjusted to individual needs. Gao et al. [6] states that LLMs struggle to learn new factual information thorough unsupervised fine-tuning.

1.2.1.3 Inference

The inference stage of the LLM is the stage you encounter when prompting a chatbot with a query. It is the stage during which the LLM uses the learned knowledge and patterns to answer your query. The generation of the response happens in a token by token fashion where each token is chosen by the means of a weighted random choice. This weighted random choice of the i-th token is facilitated by the original query, the previous i-1 tokens and retrieved documents if present [8]. This method is prone to halucinations, since the accumulation of the i-1 tokens influences the choice of the next token. Here we define halucinaitons as information which is plausible, but factually incorrect, i.e. verifiably incorrect. Addressing halucinations of LLMs is a topic of current research and we will see how RAG systems can be used to mitigate them. Namely we will grant an overview over augmentation processes [6], which implement a judging phase during which the current state within the generation is asessed.

1.2.1.4 Cost of Training a LLM

Training large language models is expensive, as the establishment of the LLM's knowledge-base usually requires billions of documents for modern models. For instance, the cost of training of Gemini 1, a very recent LLM of Google, is estimated to be in the range from 30 million USD to 190 million USD [2]. This estimation excludes the staff salaries, meaning the estimation is soley concerned with infrastructural cost. This cost is primarily composed of renting GPUs (graphics processing units) and TPUs (tensor processing units), which are specialized hardwares to perform paralell computing [3]. As mentioned previously, paralell computing has revolutionized the approaches used for natural language processing (NLP) tasks and furthermore paralell computing also finds application in deep learning. The vast amounts of energy consumed by these hardwares impose a significant cost as well as for instance the training of GPT-3 consumed as much energy as hundreds of homes annually [3]. Furthermore, the persistance layer requires a lot of space for storing the models parameters. The immensive cost of training a LLM leads to a rigid and slow training regime.

1.3 Approaches

1.3.1 Naive RAG

To discuss the different recent advances of the RAG ecosystem, we first establish a broad overview of the fundamental RAG pipeline. Having emerged in the year of 2020 RAG is an extremely new branch of technology [6]. Within RAG, three main adaptations can be observed: pre-training, fine-tuning and Inference stage. Each of the adpatations aims to improve a different stage of the LLM, as described in the problem statement. In the following sections we introduce the fundamental pipeline, which subsequently allows us to pinpoint how recent advances of RAG addressed a specific problem encounter in the naive implementation.

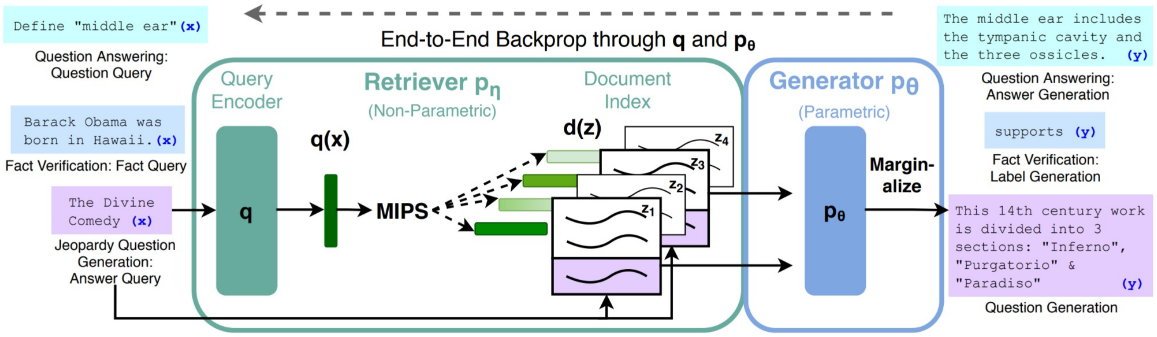

Define "middle ear" (x)
Question Answering: Question Query
Barack Obama was born in Hawaii. (x)
Fact Verification: Fact Query
The Divine Comedy (x)
Jeopardy Question Generation: Answer Query
End-to-End Backprop through q and p_θ
Query Encoder
Retriever p_η
(Non-Parametric)
Document Index
q
q(x)
MIPS
d(z)
z_4
z_3
z_2
z_1
Generator p_θ
(Parametric)
p_θ
Margin-alize
The middle ear includes the tympanic cavity and the three ossicles. (y)
Question Answering: Answer Generation
supports (y)
Fact Verification: Label Generation
This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso" (y)
Question Generation

Figure 1.2: The fundamental RAG pipeline using query encoding to retrieve semantically similar documents and provide them to the generator (LLM) as context

In this pipeline we can observe the following 3 phases :

1.3.1.1 Query Encoding

Query encoding is the process of converting a user's input (like a question or prompt) into a numerical vector in an embedding space. This embedding captures semantic meaning, so sematically similar queries map to nearby points. These query vectors are then used to compare against document embeddings to identify relevant content. By measuring similarity (e.g. via cosine similarity), the system can efficiently find documents that semantically match the query.

1.3.1.2 Document Retrieval

In document retrieval, the system searches a large index of pre-embedded documents to find the most relevant ones to the query vector. It typically uses efficient similarity search, such as approximate nearest neighbor (ANN) algorithms, to return a small subset of documents. These documents act as context or "memory" for the generator, providing background knowledge or factual information. Good retrieval improves both the accuracy and relevance of downstream generation. The usage of an approximate solution is due to the fact that finding the exact top k nearest neighbours of a query vector imposes a great computational expense. In practice, an approximate solution suffices, since the runtime of these queries is more critical then finding the exact solution. Here methods like inverted file index or hierarchical navigable small world are used to find an approximated solution of the top k nearest neighbours of a query vector [9].
Hierarchical navigable small world searches rely on a division of the embedded documents into layers, where each of the layers is traversable in polylogarithmic time.
Inverted file index is a method which relies on a chosen set of centroids C in the embedded euclidean vectorspace. Each centroid c will have a voronoi region associated with it, which is defined as :

$$Voronoi(c) := \{x \in \mathbb{R}^d \mid \|x - c\| \leq \|x - c'\|\} \quad \forall c' \in C$$

Each embedded document will then be assigned to its nearest centroids, by identifying in which voronoi region they are located in. This process will create an association between the a centroid an a list of embedded documents. When querying the vector database with some query vector, in a first step the nearest centroid is identified and in a second step the top k embedded documents associated with the centroid are retrieved. Problematic for the inverted file index approach are queries which lie near the boundary of voronoi regions and sparse vector databases.

1.3.1.3 Generation

The generation component, in our case the LLM, takes the retrieved documents along with the original query as input. It uses them to produce a coherent, contextually grounded output (like an answer, summary, or continuation). The model can "copy" factual detail from the retrieved documents or use them as inspiration while generating. This ensures that the response is both relevant to the query and grounded in up-to-date or factual data.

1.3.2 Advanced RAG

Advanced RAG systems utilize the naive RAG pipeline and enhance it by invoking pre-retrieval and post-retrieval processes [6]. The reason why such processes are widely adopted lies in the sensitive characteristic of the retrieval itself. Small changes of the query which is

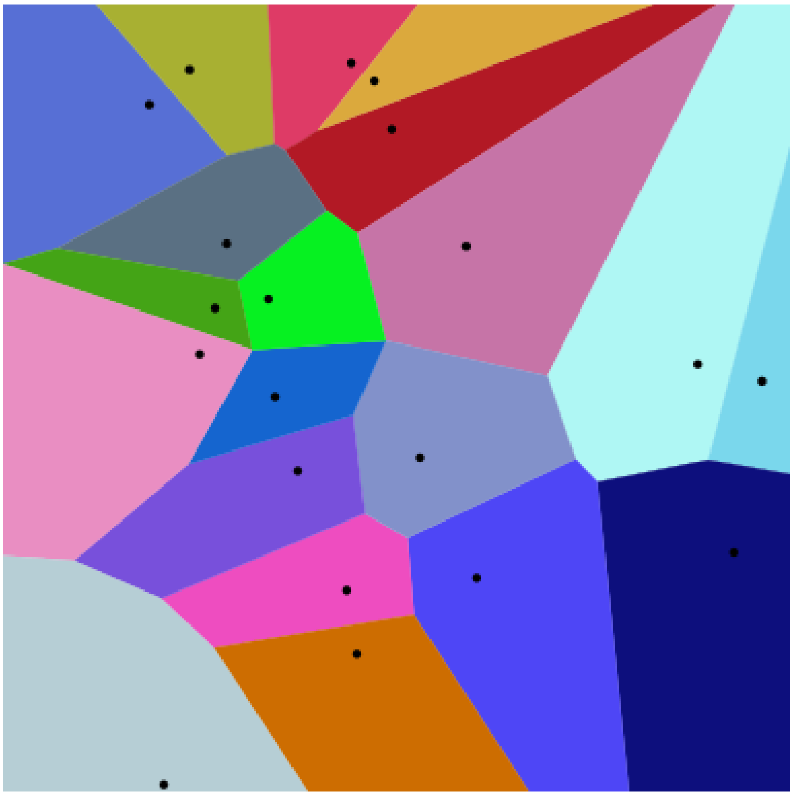


Figure 1.3: Voronoi Cells associated with Centroids, which facilitate an approximate nearest neighbour (ANN) search using the inverted file index method

used for the similarity search may yield a completely different set of retrieved documents. Especially in dense vectordatabases this phenomenon occurs more frequently. Additionally, the system cannot apriori estimate the relevance and length of the retrieved documents, which is why post-retrieval processes are invoked.

1.3.2.1 Pre-retrieval processes

Pre-retrieval processes intend to transform the query in a way such that the retrieved documents are relevant. This usually involves query rewriting, query expansion and query routing.

Query rewriting revolves around the concept that the user's input is not structured in a way such that a lookup will yield the documents of interest. This is due to a common vocabulary mismatch between the user and the embedded documents, since users choose different wordings than those found in formal literature. Furthermore the user's query might contain several semi-independent sections, which do not need to be processes in the same way. If for instance the query involves the comparison of a recent incident to an incident in the past, the two sections of the query can be processes seperatly. Here we may assume that the LLM was trained prior to the recent incident and therefore we need to retrieve the information about the recent incident from an external database. For the incident in the past we may assume that the LLM was trained later on than the incident occured and therefore the LLM has access to this information itself and no retrievals need to be performed. In conclusion, query rewriting identifies the sections for which a retrieval needs to be performed and rewrites those section such that it is more suitable for the lookup.

Query expansion is a technique used to provide the query with additional metadata [1]. This metadata is used to indicate the context in which the query is used. This is crucial since for instance the word "apple" may be a fruit, a company or even a color depending on

the context. Since the embeddings mapping will assign the same vector to the word "apple" in all of the three contexts, expanding the query with additional metadata is essential.
Query routing deals with adjusting the RAG pipeline to a specific query. This process facilitates the concept that not every query or section of a query needs to processed in the same way. In order to implement a query router one needs to first identify the query labels, usually in the range of 10 to 20 labels [4]. These labels should exhibit clear boundaries in a way, such that no query should be assignable to two labels. Subsequently a query classifier is needed, which assigns a label to queries. This classifier may rely on keyword matching, to assign a label based on the presence of a certain set of words in a query. This method cannot capture the context in which a keyword is used, which is why a shift towards an LLM classifier can be observed. Finally based on the label of query a label-specific RAG pipeline is built to process the query or section of the query.

1.3.2.2 Post-retrieval processes

Post-retrieval processes aim to evaluate the relevance of the retrieved documents and compress the documents to provide the LLM with a information-dense context.
Reranking documents is a widely adopted practice where the relevance of the retrieved documents is asessed. The reason why this reranking is important is due to the fact that performing a top k similarity search based on a query might not yield k relevant documents, since the database might not contain as many relevant documents as needed. This phenomenon can be observed especially in sparse databases, where the top k most similar documents are not located very closely to the input query. The goal of reranking the documents is to assign weights to the documents based on their relevance, such that the context provided to the LLM is not impacted by possibility that the searched database may only contain a few relevant documents.
Summary. Summarizing is a technique used after retrieval to compress multiple documents into a shorter, information-dense representation that preserves the most relevant facts for the language model. This reduces noise, removes redundancy, and ensures the model processes only the essential content rather than full passages.
Fusion, by contrast, focuses on integrating information from several retrieved sources through methods such as ranked fusion, passage aggregation, or multi-document merging to create a unified evidence set. While summarization aims at compressing content, fusion aims at combining it to resolve inconsistencies and strengthen shared signals. Together, these processes enhance answer accuracy by providing the model with concise yet coherently integrated information.

1.3.3 Modular RAG

Modular RAG is a further advancement in the RAG ecosystem which is strongly tied to the idea of query routing (see section 1.3.2.1). Here the key concept is abstracting RAG functionalities into modules. These individual modules aim to perform a certain set of instructions to process a query. Futhermore a set of patterns form the basis for how a pipeline is constructed, i.e. patterns are the assembly of multiple modules, whose grouping is oftenly encountered for common queries. In the figure 1.4 the advancement of the RAG pipeline is displayed.

1.4 Evaluation of RAG systems and Reflection

A core question which arises when comparing different approaches of RAG is how to evaluate different approaches. To gain insights of the performance of a RAG system the following evaluation metrics were stated by [6]. According to them, the primary task of

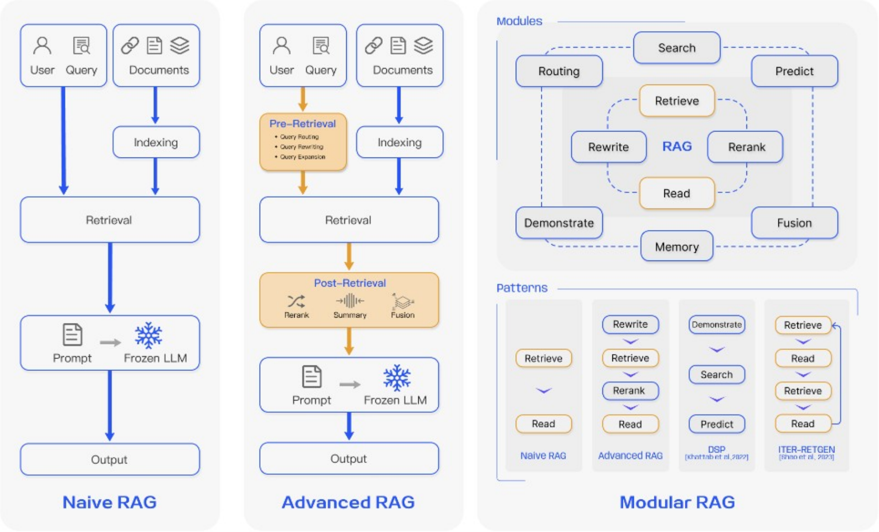


Figure 1.4: RAG Pipeline Advances Overview, displaying the naive, advanced and modular RAG pipeline

RAG remains question answering (QA). Contemporary RAG evaluation focuses on three core quality scores-context relevance, answer faithfulness, and answer relevance-along with four key abilities that measure robustness and adaptability. Context relevance reflects how precisely the system retrieves useful information, while answer faithfulness measures whether generated answers remain grounded in the retrieved context. Answer relevance assesses how directly the response addresses the user's query.

Beyond these scores, RAG systems are evaluated on their noise robustness (handling irrelevant or low-value but related documents), negative rejection (knowing when not to answer due to insufficient evidence), information integration (combining information from multiple sources for complex queries), and counterfactual robustness (ignoring known inaccuracies in retrieved content). Context relevance and noise robustness primarily indicate retrieval quality, whereas the remaining abilities and scores assess generation quality.

Bibliography

[1] Query Expansion in Enhancing Retrieval-Augmented Generation (RAG) | by Sahin Ahmed, Data Scientist | Medium. URL https://medium.com/@sahin.samia/query-expansion-in-enhancing-retrieval-augmented-generation-rag-d41153317383.

[2] The Extreme Cost Of Training AI Models Like ChatGPT and Gemini. URL https://www.forbes.com/sites/katharinabuchholz/2024/08/23/the-extreme-cost-of-training-ai-models/.

[3] (1) The Economics of LLMs: Why Training Costs Are So High ? | LinkedIn. URL https://www.linkedin.com/pulse/economics-llms-why-training-costs-so-high-manas-mallik-ejfuc/.

[4] Building a RAG Router in 2025. A practical guide to routing user… | by Timothé Pearce | Medium. URL https://medium.com/@tim_pearce/building-a-rag-router-in-2025-e0e9d99efe44.

[5] Limitations of LLMs: Bias, Hallucinations, and More. URL https://learnprompting.org/docs/basics/pitfalls.

[6] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey. 2023. URL https://github.com/Tongji-KGLLM/.

[7] Nikitas Karanikolas, Eirini Manga, Nikoletta Samaridi, Vaios Stergiopoulos, Eleni Tousidou, and Michael Vassilakopoulos. Strengths and Weaknesses of LLM-Based and Rule-Based NLP Technologies and Their Potential Synergies. *Electronics 2025, Vol. 14, Page 3064*, 14(15):3064, jul 2025. ISSN 2079-9292. doi: 10.3390/ELECTRONICS14153064. URL https://www.mdpi.com/2079-9292/14/15/3064/htmhttps://www.mdpi.com/2079-9292/14/15/3064.

[8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-Tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 2020. URL https://github.com/huggingface/transformers/blob/master/.

[9] Yu A Malkov and D A Yashunin. Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs. 2016.

[10] Milad Moradi, Ke Yan, David Colwell, Matthias Samwald, and Rhona Asgari. A Critical Review of Methods and Challenges in Large Language Models. 2025. doi: 10.32604/cmc.2025.061263.

[11] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk Microsoft. APPROXIMATE NEAREST NEIGHBOR

NEGATIVE CON-TRASTIVE LEARNING FOR DENSE TEXT RETRIEVAL. 2020. URL `https://aka.ms/ance`.

[12] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less Is More for Alignment. *Advances in Neural Information Processing Systems*, 36, may 2023. ISSN 10495258. URL `https://arxiv.org/pdf/2305.11206`.

Chapter 2

A Comparison of Open-Source and Proprietary Large Language Models

Zhuhao Fan

Large Language Models (LLMs) have become fundamental infrastructure for artificial intelligence applications, yet organizations face a critical decision between proprietary services (e.g., OpenAI's GPT, Anthropic's Claude) and open-source alternatives (e.g., Meta's LLaMA, DeepSeek). This report provides a systematic comparison across three critical dimensions: Public Safety, Personal Privacy, and Commercial & Preference Considerations. Through qualitative analysis of representative models, we examine how each paradigm addresses security threats (including prompt injection, jailbreaking, and data leakage), privacy concerns (data sovereignty, memorization risks, and GDPR compliance), and business trade-offs (cost structures, vendor lock-in, and strategic control). Our analysis reveals fundamental structural differences: proprietary models offer centralized safety mechanisms and operational convenience but require third-party data exposure and create vendor dependency; open-source models enable complete data sovereignty and customization but demand greater technical expertise and infrastructure investment. We find that performance metrics are rapidly converging, making architectural and philosophical differences increasingly decisive. The choice between paradigms reflects classic "build versus buy" considerations, with market segmentation emerging naturally: startups favor proprietary APIs for rapid prototyping, while enterprises increasingly adopt open-source solutions for regulated industries and strategic applications. This work provides evidence-based guidance for researchers, developers, and policymakers navigating the evolving LLM landscape.

Contents

2.1 Introduction and Problem Statement

2.1.1 Basic Mechanisms of Large Language Models

Large Language Models (LLMs) represent a paradigm shift in artificial intelligence, fundamentally transforming how machines understand and generate human language. At their core, modern LLMs are built upon the **Transformer architecture** [21], which introduced the revolutionary *self-attention mechanism*. This mechanism allows the model to weigh the importance of different words in a sequence when processing each token, enabling it to capture complex linguistic relationships and long-range dependencies that were challenging for previous architectures.
The development of LLMs typically follows a two-stage process:

1. **Pre-training**: In this foundational phase, models are trained on massive corpora of text data (often encompassing trillions of tokens) using self-supervised learning objectives. The most common objective is *causal language modeling*, where the model learns to predict the next token in a sequence given the preceding context. Through this process, the model develops a comprehensive understanding of grammar, facts, reasoning patterns, and stylistic variations present in the training data. The scale of this training—in terms of both data size and model parameters (now often exceeding hundreds of billions)—is crucial for emergent capabilities such as reasoning and instruction following.

2. **Alignment and Fine-tuning**: While pre-trained models acquire broad knowledge, they often require additional tuning to become helpful, harmless, and honest assistants. This is achieved through techniques like:

   - **Supervised Fine-Tuning (SFT)**: Training on high-quality demonstration data to improve task performance
   - **Reinforcement Learning from Human Feedback (RLHF)**: Optimizing model responses based on human preferences [19]
   - **Direct Preference Optimization (DPO)**: A more recent alternative to RLHF that directly optimizes for human preferences

During inference, LLMs generate text auto-regressively—producing one token at a time while conditioning on all previously generated tokens. This generative capability, combined with their extensive knowledge base, enables applications ranging from conversational assistants and code generation to creative writing and complex problem-solving.

2.1.2 Mathematical Foundations

To formalize the mechanisms underlying modern LLMs, we present the core mathematical operations that enable their capabilities.
The fundamental building block of Transformer architectures is the **self-attention mechanism**. Introduced in the seminal work [21], it computes contextualized representations by attending to all positions in the input sequence [4]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{2.1}$$

where $Q \in \mathbb{R}^{n \times d_k}$, $K \in \mathbb{R}^{n \times d_k}$, and $V \in \mathbb{R}^{n \times d_v}$ represent the query, key, and value matrices respectively, n is the sequence length, and d_k is the dimensionality of the key vectors. The scaling factor $\frac{1}{\sqrt{d_k}}$ prevents the dot products from growing too large in magnitude, which

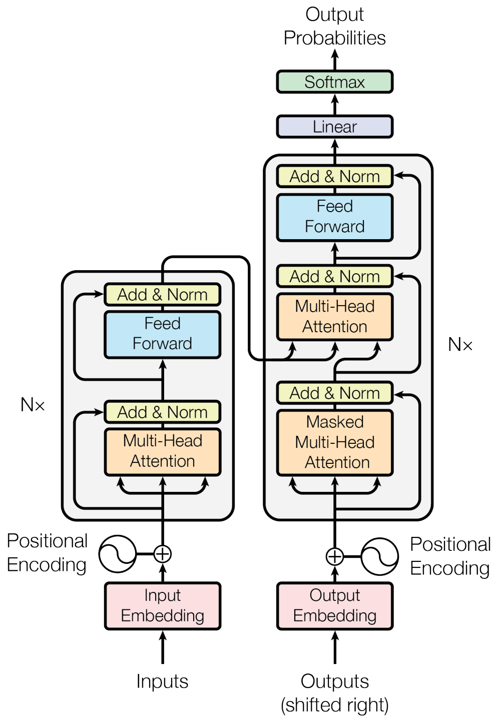


Figure 2.1: The Transformer model architecture. The encoder (left) processes the input sequence through multi-head self-attention and feed-forward layers, while the decoder (right) generates the output sequence using masked self-attention, encoder-decoder attention, and feed-forward layers. Residual connections and layer normalization are applied throughout. Figure adapted from [21].

helps mitigate the issue of the softmax function saturating and thereby stabilizes gradient propagation during training [8].

Figure 2.1 illustrates the complete Transformer architecture, showing how multiple attention heads operate in parallel to capture different aspects of the input relationships. This architecture has become the foundation for modern LLMs, enabling them to process sequences of arbitrary length while maintaining computational efficiency through parallelization.

The **pre-training objective** for causal language modeling is formulated as maximizing the log-likelihood of the training corpus, a fundamental approach in statistical language modeling [26]:

$$\mathcal{L}_{\text{LM}}(\theta) = -\sum_{t=1}^{T} \log P_\theta(w_t \mid w_{<t}) \tag{2.2}$$

where θ represents the model parameters, w_t is the token at position t, and $w_{<t}$ denotes all preceding tokens in the sequence.

For **alignment through Reinforcement Learning from Human Feedback (RLHF)**, the optimization objective balances reward maximization with maintaining proximity to the reference policy [19]:

$$\mathcal{L}_{\text{RLHF}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\log \pi_\theta(y \mid x) \cdot r(x, y)\right] - \beta \cdot \text{KL}(\pi_\theta \| \pi_{\text{ref}}) \tag{2.3}$$

where π_θ is the policy being optimized, π_{ref} is the reference policy (typically the pre-trained model), $r(x, y)$ is the reward function learned from human preferences, and β is a hyperparameter controlling the strength of the KL penalty to prevent the model from deviating too far from the reference distribution.

As shown in Figure 2.2, the RLHF process involves multiple stages of human feedback integration, starting from supervised demonstrations and progressing to preference-based

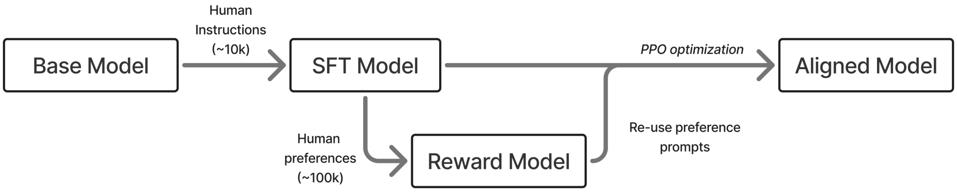


Figure 2.2: Three-stage RLHF: (1) supervised fine-tuning on demonstrations, (2) training a reward model from human preference comparisons, and (3) reinforcement learning (e.g., PPO) to optimize the policy to match human preferences. Figure adapted from [15].

optimization. This multi-stage approach has proven essential for producing helpful, harmless, and honest AI assistants.

2.1.3 Proprietary LLMs

Proprietary LLMs are developed and maintained by commercial organizations that retain exclusive control over their underlying technology. These models, exemplified by OpenAI's GPT series, Anthropic's Claude, and Google's Gemini, are typically accessed through **Application Programming Interfaces (APIs)** or web interfaces. Users interact with these models without direct access to the model weights, architecture details, or training methodologies, which are treated as trade secrets.

The proprietary approach offers several advantages: consistent performance, regular updates and maintenance, integrated safety measures, and ease of use through well-documented APIs. However, this closed nature also raises significant concerns regarding data privacy, vendor lock-in, limited transparency, and dependency on the providing company's business decisions and ethical standards.

2.1.4 Open-Source LLMs

In contrast, open-source LLMs such as Meta's LLaMA series, Falcon, and DeepSeek embrace principles of transparency and accessibility. It is worth noting that while commonly referred to as "open-source', many of these models are more accurately classified as **Open Weights** models, as their licenses often include restrictions on commercial usage or derivative works—a distinction that carries significant legal implications for enterprise compliance. These models publicly release their architecture, training methodologies, and—crucially—their model weights, enabling researchers, developers, and organizations to download, modify, and deploy them on their own infrastructure.

The open-source paradigm fosters innovation through community collaboration, enables independent verification of model capabilities and limitations, and provides users with complete control over their deployment environment. This approach particularly benefits applications requiring data privacy, custom fine-tuning, or integration into specialized workflows. However, it also demands significant technical expertise, computational resources for deployment, and responsibility for implementing appropriate safety measures.

2.1.5 Dimensions of Comparison

The fundamental differences between proprietary and open-source LLMs necessitate a structured framework for comparison. This paper examines these competing paradigms across three critical dimensions that capture their broader implications beyond mere performance metrics:

- **Public Safety**: Evaluating how each approach addresses potential misuse, including the generation of harmful content, dissemination of misinformation, and implementation of safety mechanisms. This dimension examines the trade-offs between centralized control and community-driven safety efforts.
- **Personal Privacy**: Analyzing data handling practices, user control over personal information, and risks of data leakage. This dimension contrasts the privacy implications of cloud-based API services versus local deployment.
- **Commercial and Preference Considerations**: Assessing economic factors including cost structures, business models, strategic flexibility, and market adoption patterns. This dimension explores how different stakeholders balance convenience against control and long-term sustainability.

These dimensions provide a comprehensive framework for understanding the complex trade-offs between the two dominant approaches to LLM development and deployment, enabling more informed decision-making for users, organizations, and policymakers navigating the rapidly evolving AI landscape.

2.2 Related Work

2.2.1 Technical Evolution and Performance Benchmarking of LLMs

The rapid advancement of Large Language Models began with the transformative Transformer architecture, which enabled parallel processing of sequential data through self-attention mechanisms. This breakthrough catalyzed the development of increasingly sophisticated models, from early iterations like GPT-2 to the landmark GPT-3 and beyond. The emergence of proprietary models such as OpenAI's GPT series, Anthropic's Claude, and Google's Gemini established early dominance in performance metrics across various benchmarks including MMLU, GSM8K, and HumanEval.

However, the landscape shifted significantly with the release of open-source alternatives like LLaMA, which demonstrated that high-performance models could be openly available. Recent iterations including LLaMA-3 and DeepSeek have substantially narrowed the performance gap, with comprehensive evaluations showing competitive results on many benchmarks. This *performance convergence* represents a fundamental shift in the LLM ecosystem, making architectural and philosophical differences increasingly relevant compared to raw capability metrics.

2.2.2 Safety, Alignment, and Misuse Prevention

Ensuring the safe deployment of LLMs has emerged as a critical research area. Proprietary model providers have pioneered techniques such as Reinforcement Learning from Human Feedback (RLHF) to align model behavior with human values. Red teaming has become a standard practice for identifying safety vulnerabilities before deployment.

Despite these efforts, adversarial attacks through *jailbreaking* remain a persistent challenge. The open-source paradigm presents a different safety landscape: while models can be deliberately stripped of safety features, the transparency enables community-wide scrutiny and collaborative development of robust defenses. Research has shown that supervised fine-tuning on safety-specific datasets can effectively build harm resistance directly into model weights, though this approach requires careful curation and validation.

2.2.3 Privacy and Data Governance Considerations

Privacy concerns in LLMs span multiple dimensions. A fundamental risk arises from *training data memorization*, where models can regurgitate verbatim sensitive information from their training corpora. This creates particular concern for proprietary models where user interactions may contribute to future training cycles without transparent opt-out mechanisms.

The European Union's GDPR has established the "right to be forgotten" as a legal requirement, yet practical implementation in LLM training remains challenging. Open-source models address these concerns through local deployment, ensuring data never leaves user-controlled environments. This approach is particularly crucial in regulated industries such as healthcare and finance, where data sovereignty is non-negotiable. The ability to completely audit and control data flows represents a significant privacy advantage for open-source implementations.

2.2.4 Business Models and Ecosystem Development

The commercial landscape for LLMs has bifurcated into distinct paradigms. Proprietary models typically follow a Software-as-a-Service (SaaS) model, offering accessibility at the cost of vendor lock-in and recurring expenses. Industry analyses indicate that over 70% of large enterprises are exploring internal LLM deployments to protect intellectual property and maintain strategic control.

Conversely, the open-source ecosystem has fostered innovative business models around hosted services (e.g., Hugging Face, Together AI) that provide the flexibility of open-source with the convenience of managed infrastructure. Market segmentation naturally emerges: startups and individual developers favor proprietary APIs for rapid prototyping, while established enterprises increasingly invest in open-source solutions for long-term strategic advantage. This division reflects fundamental differences in resource availability, risk tolerance, and strategic priorities across organizational types.

2.3 Approach(es)

To conduct a systematic and fair comparison between open-source and proprietary LLMs, this paper adopts a structured, multi-dimensional analytical framework. Rather than focusing solely on narrow performance benchmarks, which are rapidly converging, we seek to evaluate the broader implications and trade-offs of each paradigm. Our methodology consists of two main components: (1) the establishment of key comparison dimensions, and (2) the methodology for gathering and analyzing evidence.

2.3.1 Comparative Framework

We propose a comparative analysis across three critical dimensions that encompass technical, social, and economic considerations: Public Safety, Personal Privacy, and Commercial & Preference. These dimensions were selected because they represent the primary areas of concern and decision-making for policymakers, developers, enterprises, and end-users when choosing between LLM paradigms. They move beyond raw capability to address the real-world impact and sustainability of LLM adoption.

The specific dimensions are defined as follows:

Public Safety. This dimension assesses the potential for LLMs to be misused (e.g., to generate harmful content, disinformation, or facilitate illegal activities) and the effectiveness of the mechanisms in place to mitigate these risks. It examines the ongoing "cat-and-mouse" game between model safeguards and adversarial attacks.

Personal Privacy. This dimension focuses on data sovereignty and control. It evaluates how user data—including prompts, conversations, and documents—is handled, stored, and potentially used for further model training. It also considers the risks of data leakage and the practical implementation of data deletion policies.

Commercial & Preference. This dimension analyzes the economic and strategic implications. It compares cost structures (e.g., subscription fees vs. initial hardware investment), business models (SaaS vs. self-hosted), agility in development, and the risk of vendor lock-in.

2.3.2 Methodology

This study employs a **qualitative comparative analysis** based on synthesis of existing literature, technical documentation, and industry reports. We adopt a case-study approach, examining representative models from each paradigm:

- *Proprietary Models:* OpenAI's GPT series, Anthropic's Claude, and Google's Gemini.
- *Open-Source Models:* Meta's LLaMA series (e.g., LLaMA-2, LLaMA-3), Falcon, and DeepSeek.

For each of the three comparison dimensions—*Public Safety*, *Personal Privacy*, and *Commercial & Preference*—we systematically contrast the architectural choices, deployment models, and operational trade-offs of these representative systems. The analysis draws on:

- Published research on alignment techniques (RLHF, Constitutional AI), adversarial robustness, and red teaming
- Studies on privacy vulnerabilities including data memorization, membership inference attacks, and differential privacy
- Industry analyses of cost structures, business models, and strategic considerations
- Technical documentation of security tools and deployment frameworks

Our goal is not to provide quantitative benchmarks—which rapidly become outdated—but rather to elucidate the *structural differences* between paradigms that shape their respective strengths, limitations, and suitability for different use cases.

2.4 Solutions

2.4.1 Public Safety

The security posture of LLMs varies significantly between proprietary and open-source paradigms, necessitating distinct solutions for each.

2.4.1.1 Security Mechanisms in Proprietary Models

Proprietary models typically employ a centralized, top-down approach to safety, integrating multiple layers of defense:

- **Constitutional AI and Hierarchical Filtering**: Systems evaluate model outputs against predefined ethical principles (a "constitution") and apply hierarchical filters at different abstraction levels before returning results to users [2].
- **Adversarial Training**: Models are hardened through minimax optimization:

  $$\min_{\theta} \max_{\delta \in \Delta} \mathcal{L}(f_{\theta}(x+\delta), y), \tag{2.4}$$

  where δ represents adversarial perturbations constrained to Δ. This technique enhances model resistance against manipulated inputs.
- **Reinforcement Learning from Human Feedback (RLHF)**: Human feedback shapes reward models to align model behavior with preferred outcomes, helping models distinguish between acceptable and harmful responses [19].
- **Output Sanitization Pipelines**: Real-time content classification systems analyze generated text for toxicity, bias, and harmful instructions using specialized classifiers before delivery [12].
- **Red Teaming**: Companies invest in security experts who systematically attempt to breach model defenses, identifying vulnerabilities before malicious actors can exploit them [11].
- **Security Tools Integration**: Proprietary ecosystems often integrate with or offer dedicated security tools that provide prompt injection protection, sensitive data leakage prevention, and hallucination detection. Examples include Lakera Guard and ProtectAI LLM Guard, which have been shown effective in mitigating security threats.

These managed defenses reduce risk for general users. However, they must be continuously updated to counter new jailbreak tactics, creating a perpetual cat-and-mouse dynamic [23].

2.4.1.2 Security Mechanisms in Open-Source Models

Open-source models adopt a fundamentally different, community-driven security approach:

- **Transparent Scrutiny**: Publicly available weights and code enable the global research community to collectively examine models for vulnerabilities and collaboratively develop fixes [3].
- **LoRA Safety Tuning**: Low-Rank Adaptation enables efficient safety fine-tuning through parameter-efficient updates:

  $$W' = W + \Delta W = W + BA, \tag{2.5}$$

  where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ are low-rank matrices with $r \ll \min(d, k)$, allowing safety improvements without full model retraining.
- **Supervised Safety Fine-Tuning**: Developers can proactively fine-tune models on curated safety corpora, building robust defenses directly into model weights rather than relying on external filters.

- **Open-Source Security Tools**: The community provides various security frameworks such as LLM Guard [20] (detecting and sanitizing harmful language, preventing data leakage, resisting prompt injections), Guardrails AI (validating and correcting LLM outputs), and Garak [18] (vulnerability scanner for LLMs). The architectural transparency of open-source models allows organizations to implement and test their own rigorous, auditable guardrails using these specialized tools to counter evolving threats.
- **Ensemble Safety Filters and Customizable Safeguards**: Multiple specialized classifiers can be deployed in parallel for defense-in-depth. Organizations can tailor safety protocols to specific needs and risk environments, potentially exceeding the protections available in proprietary offerings.

The open-source paradigm emphasizes transparency, community verification, and customizable protection mechanisms at the cost of requiring more operational effort.

2.4.1.3 Threat Models and Attack Vectors

Understanding the security landscape requires systematic analysis of attack vectors:

- **Prompt Injection Attacks**: Malicious inputs designed to override safety instructions or manipulate model behavior. These attacks exploit the lack of clear boundaries between instructions and user data in natural language interfaces. Techniques include direct injections, indirect injections [13], and language switching to bypass filters. The proliferation of complex variants, particularly Indirect Prompt Injection, necessitates deep architectural transparency, enabling system-wide security auditing—a key advantage that favors open-source deployment for high-risk applications.
- **Multi-Turn Attacks / Jailbreaking**: Sequences of interactions—including role-playing scenarios, encoded instructions, and multi-turn conversation manipulation—that gradually coax models into unsafe responses. Recent research reports success rates reaching 92.78% against some open-weight models [23].
- **Data Leakage and Training Data Extraction**: LLMs might unintentionally leak sensitive information from their training data [12], including personally identifiable information (PII), confidential corporate data, or internal system prompts if not properly isolated. Adversaries use targeted queries to reconstruct training examples (membership or reconstruction attacks).
- **Model Stealing / Theft**: Attackers repeatedly query models to distill their capabilities into local copies, misusing the model to duplicate its functionality for financial gain or to bypass licensing costs.

Defenses follow a defense-in-depth philosophy: input validation, rate limiting, monitoring, ensemble filtering, adversarial fine-tuning, legal/contractual controls, and air-gapped deployment. The optimal combination depends on deployment context (e.g., internet-facing API vs. on-prem inference).

2.4.2 Personal Privacy

Privacy trade-offs are structural: cloud-hosted, proprietary services centralize data (and hence risk), while local open-source deployments enable stronger data sovereignty.

2.4.2.1 Privacy Limitations in Proprietary Models

Proprietary models face inherent privacy limitations due to their cloud-based nature:

- **Third-Party Data Exposure**: Every prompt and conversation is typically processed on the vendor's servers, which means sensitive information leaves the user's direct control.
- **Unverifiable Data Practices**: Users often cannot independently verify how their data is handled, stored, or whether it is completely deleted upon request, as the infrastructure and processes are not transparent.
- **Data Leakage Risks**: Models may regurgitate verbatim text from their training data, potentially exposing private information from other users who contributed to the training corpus [12]. Industry reports document that data breaches and unauthorized data exposure remain persistent threats [22].

These limitations create fundamental privacy concerns that cannot be fully resolved within the proprietary model framework.

2.4.2.2 Formal Privacy Attack Models

Privacy risks can be formalized through several attack frameworks:

- **Membership Inference:** $A(x, f_\theta) \to \{0, 1\}$, where an attacker predicts whether x was in the training set. Success rates correlate with model overfitting and memorization.
- **Attribute Inference:** $P(\text{attribute} \mid \text{partial\_info}, f_\theta) > P(\text{attribute} \mid \text{partial\_info})$. The model leaks information if its predictions improve attribute inference beyond prior probabilities.
- **Differential Privacy:** This framework provides formal guarantees expressed as (ϵ, δ)-DP: $P[\mathcal{M}(D) \in S] \leq e^\epsilon P[\mathcal{M}(D') \in S] + \delta$, where D and D' are neighboring datasets differing in one record [7].

2.4.2.3 Enhanced Privacy in Open-Source Models

Open-source models provide superior privacy solutions through local deployment:

- **Complete Data Sovereignty**: All data—prompts, documents, and internal secrets—can be processed entirely on the user's own device or private servers, ensuring data never leaves the organization's controlled environment. The critical need for complete data sovereignty and the explicit ability to fulfill the "right to erasure" (GDPR, Article 17) [9] fundamentally limits the deployability of proprietary, cloud-hosted LLMs in highly regulated sectors such as healthcare and finance, where data sovereignty is non-negotiable.
- **Transparent and Auditable Systems**: The open nature of the software allows for independent verification of data handling practices, code inspection, and runtime monitoring [3].
- **Guaranteed Deletion and Flexible Management**: Organizations can enforce and verify complete data deletion when required, with unambiguous ownership rights and clear boundaries about data usage.

For handling sensitive information in fields like healthcare, finance, or legal services, open-source models offer stronger privacy-preserving solutions.

2.4.3 Commercial & Preference

The commercial trade-offs remain: convenience and rapid access versus control and long-term predictability.

2.4.3.1 Proprietary Model Business Framework

Proprietary models often operate on a Software-as-a-Service (SaaS) model:

- **Low Barrier to Entry**: No upfront hardware investment or extensive technical expertise is required for initial use, enabling rapid prototyping and deployment. This is evident in the rapid adoption by startups and individual developers who leverage these APIs to build applications without managing infrastructure [24].
- **Continuous Updates**: Users automatically benefit from the provider's ongoing research and development, receiving model upgrades that enhance capabilities [24].
- **Financial and Strategic Risks**: Recurring subscription costs can scale significantly with usage, leading to unpredictable expenditures [24]. Strategic risks include **vendor lock-in**, where deep integration with a provider's ecosystem (e.g., Microsoft Azure AI) can make switching costly [25], and dependency on the vendor's business continuity.

2.4.3.2 Open-Source Model Value Proposition

Open-source models require greater initial investment but offer superior long-term control:

- **Total Independence and Operational Longevity**: Self-hosted models are not subject to a provider's decision to retire a service, ensuring business continuity and eliminating dependency on external services [6].
- **Complete Customization**: Organizations can fine-tune models extensively on their proprietary data, achieving higher accuracy and relevance for niche applications. This is crucial for enterprises looking to leverage their unique data as a competitive advantage [5].
- **Predictable Long-term Costs**: This represents a shift from **Operational Expenditure (OpEx)**—the pay-as-you-go model of proprietary APIs—to **Capital Expenditure (CapEx)** for hardware acquisition. While requiring substantial initial capital, this structure eliminates unpredictable per-token fees, providing financial predictability and potential cost savings at scale. An IBM study found that 51% of businesses using open-source AI tools reported positive ROI [17].
- **Enhanced Security & Privacy**: On-premises deployment keeps sensitive data within the organization's control, which is a critical factor for regulated industries and enterprises concerned about data privacy [17].

The market naturally segments: startups and individual developers favor proprietary APIs for convenience and low initial cost, while enterprises prefer open-source solutions for control, customization, and data security [24; 6].

2.5 Evaluations and Discussion

2.5.1 A Comparative Analysis Across Three Dimensions

2.5.1.1 Public Safety: The Walled Garden vs. The Open Field

Proprietary models offer centrally-managed safety mechanisms, often integrating advanced security tools. However, this "walled garden" approach can be perpetually challenged by new adversarial attacks, creating a cat-and-mouse dynamic [23]. The effectiveness of safety tools can vary, with evaluations showing differences in their ability to handle malicious prompts while minimizing false positives.

The "open field" approach of open-source models demands more responsibility from the deployer but fosters flexibility and transparency [3]. It allows organizations to directly integrate specialized security toolkits and tailor safety protocols to their specific needs and threat environments. This paradigm can lead to more customizable and potentially auditable security postures.

2.5.1.2 Personal Privacy: Sovereignty vs. Convenience

Proprietary models typically necessitate third-party data exposure through API calls and cloud infrastructure. In contrast, open-source models enable full data sovereignty via local deployment on user-controlled hardware.

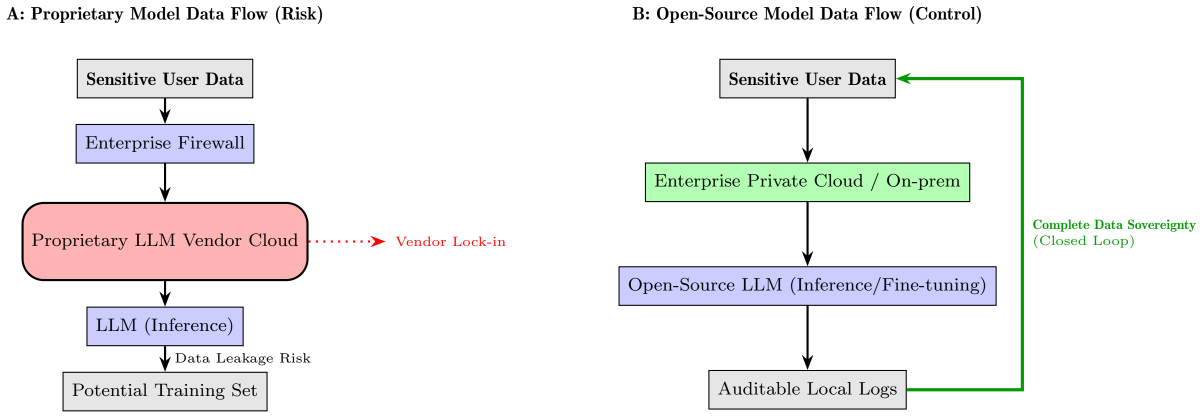

A: Proprietary Model Data Flow (Risk)
Sensitive User Data
Enterprise Firewall
Proprietary LLM Vendor Cloud
Vendor Lock-in
LLM (Inference)
Data Leakage Risk
Potential Training Set
B: Open-Source Model Data Flow (Control)
Sensitive User Data
Enterprise Private Cloud / On-prem
Open-Source LLM (Inference/Fine-tuning)
Auditable Local Logs
Complete Data Sovereignty
(Closed Loop)

Figure 2.3: Data Flow Comparison: Proprietary API vs. Open-Source Deployment.

This distinction is not merely a matter of preference but can represent a compliance imperative for many industries subject to regulations such as GDPR, HIPAA, or other sector-specific data protection requirements. The privacy advantage of open-source models, derived from their local deployment model, is structural and forms a core part of their value proposition, especially for organizations handling sensitive data.

2.5.1.3 Commercial & Preference: Strategic Agility vs. Operational Simplicity

Proprietary APIs lower the barrier to entry and reduce operational overhead but create long-term vendor lock-in and cost uncertainty. Open-source models require significant upfront investment in infrastructure and expertise but grant greater control and predictable long-term costs.

The choice often mirrors classic "build versus buy" decisions in enterprise technology. For core competitive capabilities where differentiation and control are critical, "building" on open-source foundations can be strategically advantageous despite higher initial costs. For non-differentiating functionalities where speed-to-market is paramount, "buying" via proprietary API can be operationally efficient.

This commercial landscape is also influenced by a broader context where enterprise technology buyers face economic uncertainties and are focusing on practical, cost-effective AI solutions that deliver clear returns. The positioning of different organizational types and their strategic choices within this landscape is visualized in Figure 2.4, which provides a practical decision framework for organizations evaluating their LLM deployment strategy.

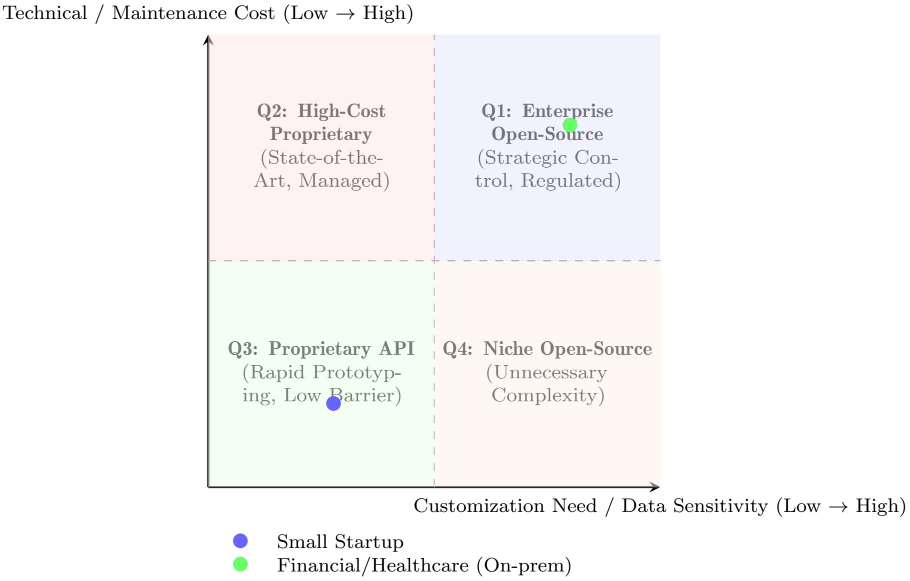

Technical / Maintenance Cost (Low → High)
Q2: High-Cost Proprietary (State-of-the-Art, Managed)
Q1: Enterprise Open-Source (Strategic Control, Regulated)
Q3: Proprietary API (Rapid Prototyping, Low Barrier)
Q4: Niche Open-Source (Unnecessary Complexity)
Customization Need / Data Sensitivity (Low → High)
● Small Startup
● Financial/Healthcare (On-prem)

Figure 2.4: LLM Selection Decision Framework based on Customization and Operational Cost.

2.5.2 Synthesis and Future Trajectories

We observe a complex LLM landscape where the choice between proprietary and open-source involves multifaceted trade-offs. Notably, the boundaries between these paradigms are increasingly blurring: proprietary vendors now offer more flexible deployment options, while open-source models are increasingly available via managed services.

2.5.2.1 Emerging Considerations

Several trends and considerations may shape the future evolution of both paradigms:

- **Application-Level Safety**: There is a growing recognition of the need to evaluate safety at the application level, not just the base model level [1]. Components such as system prompts, retrieval pipelines, and specific guardrails significantly influence the overall safety of deployed LLM applications.

- **Practical AI and Business Alignment**: Enterprises are increasingly focused on "Practical AI", seeking technologies that offer clear, measurable returns and align with specific business needs and industry requirements, often amidst economic uncertainty.

- **Specialized Security Tools**: The ecosystem of LLM security tools, both proprietary and open-source, continues to evolve, offering more sophisticated means to address vulnerabilities like prompt injections, data leakage, and model hallucinations.

- **Emerging Architectures**: Novel architectures such as Mixture-of-Experts (MoE) with sparse activation, Mixture-of-Depths for dynamic computation allocation, and State Space Models (e.g., Mamba) may offer computational advantages and challenge the current Transformer-dominated landscape. Additionally, the rise of **Small Language**

Models (SLMs) significantly lowers the hardware barrier for local deployment, serving as a catalyst for SMEs to adopt open-source solutions without massive infrastructure investment.

- **Regulatory Evolution**: Emerging AI regulations (EU AI Act, potential US frameworks) will fundamentally reshape the LLM ecosystem, potentially imposing new compliance burdens on open-source models while creating regulatory moats and stricter safety requirements for proprietary models.

2.5.3 Limitations of this Study

This analysis is based on the rapidly evolving state of the art. The pace of innovation in large language models and their associated security tools means that specific capabilities, offerings, and the threat landscape can change significantly over short time periods [3]. Furthermore, the analysis often focuses on prominent models and widely discussed tools, and may not capture the full diversity of the LLM ecosystem, including highly specialized or regional solutions.

Evaluations of security tools and model performance are often snapshots in time [11]. The dynamic nature of both model development and adversarial techniques necessitates continuous evaluation rather than one-time assessment for real-world deployment decisions.

2.6 Summary and Conclusions

This systematic comparison reveals fundamental trade-offs between open-source and proprietary LLM paradigms that extend beyond performance metrics to encompass architectural, philosophical, and strategic considerations.

2.6.1 Key Findings

Public Safety: Centralized Control vs. Community Vigilance. Proprietary models implement managed safety through techniques like Constitutional AI and RLHF, providing consistent baseline protections but engaging in perpetual cat-and-mouse dynamics with adversarial attacks. Open-source models enable transparent scrutiny and customizable defenses, allowing organizations to tailor security protocols to specific threat environments through community-driven verification and defense-in-depth strategies.

Personal Privacy: Structural Sovereignty vs. Operational Convenience. The privacy distinction is architectural: proprietary models necessitate third-party data exposure through cloud APIs, creating inherent risks of data leakage and unverifiable deletion practices. Open-source models enable complete data sovereignty through local deployment, providing decisive advantages for regulated industries where data control represents a compliance imperative rather than mere preference.

Commercial Considerations: Strategic Control vs. Operational Efficiency. Proprietary APIs lower barriers to entry through SaaS models, offering operational efficiency but creating vendor lock-in risks and unpredictable scaling costs. In contrast, open-source models require substantial initial investment but provide long-term cost predictability, complete customization, and operational independence. Market segmentation naturally emerges, with startups favoring proprietary solutions for rapid prototyping while enterprises adopt open-source for core competitive capabilities.

2.6.2 Decision Framework

Organizations should prioritize **proprietary models** when:

- Rapid prototyping and time-to-market are critical
- Technical expertise for deployment is limited
- Data sensitivity is low and regulatory requirements minimal
- Automatic updates and managed infrastructure are valued over control

Open-source models are preferable when:

- Data sovereignty and privacy are non-negotiable (regulated industries)
- Long-term cost predictability and vendor independence are strategic priorities
- Extensive customization and fine-tuning on proprietary data are required
- Technical capabilities for self-hosting exist

2.6.3 Future Outlook

The LLM ecosystem exhibits dual trajectories: performance metrics are converging while strategic approaches are diverging. Emerging hybrid solutions blur traditional boundaries, with proprietary vendors offering flexible deployment options and open-source models available through managed services. Future competitive advantage will be determined by ecosystem suitability rather than individual model capabilities, with regulatory frameworks, specialized architectures, and application-level safety evaluation shaping the evolving landscape.

The optimal strategy increasingly involves portfolio approaches—leveraging proprietary APIs for non-differentiating rapid experimentation while deploying open-source solutions for strategic core capabilities and sensitive data handling. This strategic pivot toward open-source control is evidenced by market data: high-performing organizations, for example, are over 40% more likely to utilize open-source models to maintain autonomy and reduce vendor lock-in [16]. Furthermore, the democratization of fine-tuning techniques, such as Low-Rank Adaptation (LoRA) [14], drastically lowers the computational barrier for bespoke model customization, further accelerating the open-source trajectory for niche and domain-specific applications.

The emergence of comprehensive regulatory frameworks, such as the EU AI Act [10], will increasingly favor models that offer auditable transparency and full architectural control, thereby strengthening the open-source value proposition for high-risk applications. These regulatory developments, combined with the growing emphasis on data sovereignty under frameworks like GDPR [9], position open-source models as the preferred choice for organizations operating in highly regulated environments.

Bibliography

[1] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

[2] Yuntao Bai et al. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

[3] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, et al. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

[4] Stefania Cristina. The Transformer Attention Mechanism. https://machinelearningmastery.com/the-transformer-attention-mechanism/, 2023.

[5] Larry Dignan. Enterprise ai: It's all about the proprietary data. https://www.constellationr.com/blog-news/insights/enterprise-ai-its-all-about-proprietary-data, 2025.

[6] Yulia Dmitrievna and Eduard Parsadanyan. The 11 best open-source llms for 2025. https://blog.n8n.io/open-source-llm/, 2025.

[7] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

[8] DZ. From Attention to Self-Attention and Transformers: a Mathematical Introduction. https://dzdata.medium.com/from-attention-to-self-attention-and-transformers-a-mathematical-introduction-45ed053a1ed3, 2024.

[9] European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation). Official Journal of the European Union, 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj. Article 17: Right to erasure ('right to be forgotten').

[10] European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

[11] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022. URL https://arxiv.org/abs/2209.07858.

[12] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for*

Computational Linguistics: EMNLP 2020, pages 3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL `https://aclanthology.org/2020.findings-emnlp.301/`.

[13] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection, 2023. URL `https://arxiv.org/abs/2302.12173`.

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL `https://arxiv.org/abs/2106.09685`.

[15] Nathan Lambert. Reinforcement learning from human feedback, 2025. URL `https://arxiv.org/abs/2504.12501`.

[16] McKinsey & Company. The state of ai in 2025: Agents, innovation, and transformation. Technical report, McKinsey Global Institute, October 2025. URL `https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai`.

[17] Anabelle Nicoud. Open-source ai in 2025: Smaller, smarter and more collaborative. `https://www.ibm.com/think/news/2025-open-ai-trends`, 2025.

[18] NVIDIA. Garak: A framework for large language model red teaming. GitHub Repository, 2024. URL `https://github.com/NVIDIA/garak`. Open-source vulnerability scanner for LLMs.

[19] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. URL `https://arxiv.org/abs/2203.02155`.

[20] Protect AI. Llm guard: The security toolkit for llm interactions. GitHub Repository, 2024. URL `https://github.com/protectai/llm-guard`. Open-source toolkit for securing LLM applications.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[22] Verizon Business. 2024 Data Breach Investigations Report. Technical report, Verizon Communications Inc., 2024. URL `https://www.verizon.com/business/resources/reports/2024-dbir-data-breach-investigations-report.pdf`. Comprehensive annual report on data breaches and security incidents.

[23] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail?, 2023. URL `https://arxiv.org/abs/2307.02483`.

[24] Rachel Whitener. The top ai models and trends shaping saas in 2025. `https://www.cloudzero.com/blog/top-ai-models/`, 2025.

[25] Albert Yu. 16 top enterprise ai vendors to consider in 2025. `https://www.shakudo.io/blog/top-enterprise-ai-vendors-to-consider`, 2025.

[26] Aston Zhang, Zachary C Lipton, Mu Li, and Alexander J Smola. *Dive into deep learning.* Cambridge University Press, 2023.

Chapter 3

Clickbait Capitalism – The Economics of Misinformation Online

Elliot Jonsson

This report examines how economic incentives, political strategies, and new technologies shape the spread of misinformation in today's digital environment. It explains how clickbait capitalism rewards attention rather than accuracy, creating conditions where disinformation can grow both as a political tool and as a profitable product. Russia's influence campaigns illustrate how state actors use this system to polarise societies, with Sweden serving as an example of a country that is both vulnerable and resilient due to its digital openness. The report also explores how large language models change the structure of propaganda by enabling automated and scalable content creation. Several countermeasures are discussed, including fact checking, regulation, economic incentives for platforms, and improvements in digital competence. The analysis shows that no single solution is sufficient. Instead, resilience depends on a combination of policy, technology, and education, as well as an understanding of how people interpret and react to information.

Contents

3.1 Introduction and Problem Statement

In this report, I use the term clickbait capitalism to describe the economic system that drives the modern internet. It is a system in which human attention is monetised. Digital platforms such as Facebook, YouTube, TikTok, and X do not earn revenue by producing content. Instead, they profit by keeping users engaged for as long as possible. Their business model depends on advertising income, which grows with every view, click, and share. As a result, online content that provokes curiosity, emotion, or outrage spreads more widely than accurate information [11]. This dynamic encourages platforms to optimise for engagement rather than truth. It also creates a structural bias. Content that triggers strong reactions becomes more visible and therefore more profitable. The outcome is an attention economy where the most sensational material wins, regardless of whether it is true. In this sense, clickbait capitalism forms the economic foundation that allows misinformation to thrive.

Within this attention-driven system, I argue that disinformation has become a profitable product. The same incentives that drive viral marketing now reward false or exaggerated stories because they generate engagement. Entire networks of low-quality content farms and ”fake news” sites operate with the primary aim of monetising deception through ad clicks or affiliate revenue, as documented in research on disinformation economies and influence operations [6]. In this view, disinformation is not simply a byproduct of the digital age. It functions as a deliberate commercial strategy that turns propaganda, manipulation, and outrage into commodities traded in the global information market.

Propaganda, which once was the domain of state-run media has found a new ecosystem within this economy. Today, it no longer requires a centralised state apparatus to spread. It can be distributed through the same commercial platforms that host viral entertainment. State actors, extremist groups, and private influencers all exploit the same attention-based logic by blending political messaging with clickbait culture, as documented in research on contemporary disinformation strategies [11]. The result is a modern hybrid information environment in which propaganda is both political and profitable. Platforms benefit financially from traffic and engagement, while political actors benefit strategically from shaping public perception.

Research by the RAND Corporation, a U.S.-based policy research institute [6] shows that Russia frequently uses digital platforms to spread large volumes of emotional and polarising content. This pattern makes Russia one of the clearest examples of how state actors can exploit attention-based systems for strategic influence. Instead of relying only on traditional state media, Russian actors employ digital techniques that allow their messages to circulate quickly and reach broad audiences. Russia therefore offers a strong case for examining how political motives can align with the economic incentives that shape today’s information environment.

One implication is that the emergence of large language models (LLMs) adds a new technological layer to this economy of misinformation. These systems are trained on massive datasets to predict and generate text, and they can produce convincing articles, comments, and images at scale and at minimal cost. As a result, the supply of misinformation can become virtually infinite while its production cost approaches zero. The economic logic remains the same: engagement generates profit. What changes is that AI accelerates and automates this process. One risk is that truth becomes economically uncompetitive. Authentic journalism is slow and costly, while synthetic disinformation is instant and scalable.

At the same time, LLMs are not inherently harmful. Their impact depends on how people use them. The same persuasive and conversational abilities that can be used for propaganda can also be used for positive purposes. Research shows that long and personalised conversations with AI systems can reduce belief in conspiracy theories over time [10]. This suggests that such systems can support reflection and help people reconsider false beliefs when used responsibly. These tools can be influenced by biased data, political goals, or commercial incentives. However, they can also be used to support fact-checking, reduce conflict, and improve public understanding. The challenge is therefore not only technical. It is also human and structural.

According to a report by the Swedish Centre for Eastern European Studies (SCEEUS) [9] and Kragh [5], Sweden illustrates this tension well. It is a small and digitally open democracy where the distance between citizens and politicians is short, which makes it both resilient and vulnerable. Russian influence operations have targeted Swedish debates on membership in the North Atlantic Treaty Organization (NATO), migration, and energy policy. The reports note that Sweden's size and transparency give such efforts symbolic and political value.

This study examines how economic incentives in clickbait capitalism interact with recent developments in AI. It focuses on how financial incentives, human behaviour, and AI tools together can sustain and scale disinformation. The study also explores how these forces can be strategically used by state actors, including Russia.

3.2 Related Work

Many researchers have shown that misinformation online is not only a communication problem but also an economic one. A well-known study from the Massachusetts Institute of Technology [11] found that false news spreads faster and wider than true news on social media. The reason is that false stories often sound more emotional, surprising, or entertaining, which makes people more likely to share them. Every click, view, and share creates advertising revenue for digital platforms. As a result, these engagements become more valuable than accuracy. This means that online systems are built to reward attention rather than truth. Scholars often describe this as part of the ”attention economy”, a digital market where human attention is the main product. In this system, misinformation is no longer just an accident. It becomes a practice that is economically rewarded.

Another group of studies focuses on how Russia uses information as a political tool. Researchers in political communication and security studies describe this as a form of information warfare that builds on older Soviet strategies known as ”active measures”. Reports from the Swedish Institute of International Affairs [9] show that Russian influence campaigns target both local and foreign audiences, especially in Europe and the United States. Their goal is often to weaken trust in democratic institutions, increase social division, and promote ideas that support Russian interests. In Sweden, for example, Russian media and online networks have focused on topics such as NATO membership, gas prices, and migration [9]. These topics are chosen because they create emotional reactions such as fear, anger, or uncertainty. Such reactions increase engagement and visibility online. The research shows that Russian propaganda takes advantage of the same attention-based logic that drives the online economy.

More recently, researchers have started to study how artificial intelligence changes the way misinformation spreads. Large language models (LLMs) can now create realistic and persuasive text almost instantly. This makes it easier and cheaper to produce and share false or misleading information. Scholars often mention three main risks: bias, hallucination, and content automation [2] [6] [8]. Bias happens because AI systems learn from online data that already contains errors or stereotypes. Hallucination means that AI can produce information that sounds true but is completely made up. In a recent paper, Anqi Shao [8] describes AI hallucinations as a new kind of misinformation that comes from how these models predict words, not from human intention. Finally, researchers warn about ”AI content farms”. These systems produce large amounts of low-quality text

automatically, often just to attract clicks and advertising revenue. This combination of low cost and high speed makes misinformation more profitable and harder to control.
Together, these studies show that misinformation spreads because of how the digital economy works, not just because of human behaviour. False content is profitable, political propaganda uses the same system to reach people, and new AI tools make it all faster and cheaper. This research provides the background for understanding how clickbait capitalism connects economics, politics, and technology in today's information environment.

3.3 Approaches: Russia and Propaganda Online

Research shows that Russia uses a broad combination of traditional influence methods and modern online techniques when it conducts information operations. RAND describes this communication model as the firehose of falsehood, meaning that Russia spreads large amounts of messages at high speed without a clear focus on consistency or factual accuracy. The strategy builds on volume, repetition, emotional framing, and constant exposure. In this approach the goal is not always to convince people of one clear version of events. Instead it aims to create confusion and weaken trust in information in general [6].

The strategy builds on volume, repetition, emotional framing, and constant exposure. In this approach the goal is not always to convince people of one clear version of events. Instead, it aims to create confusion and weaken trust in information in general. This pattern has been described in the literature as similar to the idea of ”flooding the zone with misinformation”, where a large amount of often contradictory content makes it harder for audiences to know what to trust [4]. Such an environment benefits actors seeking influence by replacing clarity with noise and uncertainty [6].

Alongside this communication model, Russia uses a mix of tools that function together in the current online environment. RAND notes that state controlled media such as RT and Sputnik continue to present narratives that support the Russian state. They operate as international news outlets. However, they often highlight stories that frame Europe and the United States as unstable or aggressive. At the same time, covert online activity takes place through organised groups that operate accounts on social media and attempt to shape public discussions. These coordinated accounts include bot networks that automatically react to posts. They also include human-operated profiles that take part in debates in order to amplify certain messages. Troll factories like the Internet Research Agency in St Petersburg are known examples of such activity. Their purpose is to place politically charged stories into ordinary online conversations in a way that makes the messages appear spontaneous and locally produced [6].

Beyond these general methods, Russia also targets specific countries where geopolitical and symbolic gains are higher. Sweden is one of them. For decades, Sweden maintained military non-alignment. In parallel, it acted as a key partner for the NATO and the European Union. This gave Sweden an unusual position: formally outside military alliances yet deeply integrated in Western political and security structures. From a Russian perspective, influencing Swedish opinion has therefore offered both geopolitical and symbolic value. Weakening trust in Swedish institutions or shaping perceptions around defence issues can indirectly affect the balance of security in the Baltic region.

Sweden also carries symbolic weight as a stable and well-functioning democracy. In Russian state media, Sweden often appears as a reference point in debates about European values, migration, and social cohesion. Challenging Sweden's image or portraying it as unstable can support broader narratives that question the strength of Western democracies. This symbolic dimension makes Sweden useful not only as a regional target but also as an example in Russia's international communication.

At the same time, Sweden's domestic debates offer openings that foreign actors can exploit. Issues such as migration, crime, energy policy, and NATO membership have been polarising topics in recent years. These subjects tend to produce strong emotional reactions and attract significant attention online, which makes them ideal targets in the attention-driven logic of modern platforms. Russian influence campaigns often amplify existing tensions rather than introduce entirely new themes. By repeating stories that emphasise division or uncertainty, they can strengthen narratives that already circulate within Swedish political discourse.

The SCEEUS report, authored by Martin Kragh [5] [9] shows how these methods have been directed at Sweden in recent years. According to the report, Sweden has long been an important country in Russian security thinking, and this pattern continues today. Russian influence efforts often focus on issues that already generate debate within Swedish society. Topics such as NATO membership, migration, crime, and energy policy are repeatedly used because they trigger emotional reactions and can divide public opinion. Kragh presents several examples of how Russian actors have used both official channels and covert online activity to shape discussions around these issues. The report also notes that Russian campaigns sometimes involve forgeries, cyber intrusions, or targeted narratives in connection with major political events. One example is the increased focus on Sweden in Russian state media during the years when Sweden moved closer to NATO cooperation [5]

Taken together, the RAND analysis, the SCEEUS report, and Sweden's geopolitical situation show how Russia adapts a broad communication strategy to a specific national context. The firehose model, state-controlled media, troll networks, and coordinated online accounts form a layered system that can be directed at countries where the symbolic and strategic payoff is high. Sweden fits this pattern. Its long tradition of military non-alignment combined with close cooperation with NATO and the EU gives it strategic relevance in the Baltic region. Its reputation as a stable democracy gives it symbolic value in Russian narratives about European decline. Simultaneously, Sweden's open digital environment and its polarising debates on issues such as migration, crime, energy policy, and NATO membership provide entry points for influence efforts. The choice of topics is often based on identifying existing tensions in a target society and then repeating stories that reinforce those tensions. In this way, Russia's general methods and Sweden's specific vulnerabilities reinforce each other, combining older political goals with modern digital tools that make influence faster, cheaper, and harder to detect. This dynamic helps explain why Sweden continues to be a recurring focus of Russian information operations despite its small population.

3.4 Large Language Models and Propaganda

Large language models have introduced a major shift in the way information is created and circulated. According to the Stanford Foundation Models report, these systems are built by training on enormous collections of text from the internet as well as from books articles and other written sources. During training, the model repeatedly predicts the next word in a sentence and through this process it learns patterns structures and regularities in human language [2].

Because the training data covers a wide range of topics and styles, the model can later produce coherent text, respond to questions, summarise long documents, and imitate different voices. This also means that the model does not understand truth in any human sense. It mirrors patterns rather than evaluating them. This distinction is important because it shows why LLMs can reproduce misinformation, even when the developers did not intend it. This gives large language models a degree of fluency that makes their output appear natural and convincing, even when no human writer is involved.

The fact that these models rely on large internet datasets also makes them vulnerable to the broader information environment. Models trained on digital content absorb not only useful knowledge but also the distortions, bias, and errors that exist online. If certain narratives are repeated very frequently the model learns them as statistical patterns. In practice, this creates an incentive structure where the most persistent voices in the online environment exert the strongest influence on future models. This gives organised propaganda networks disproportionate power compared to ordinary users, because a model can reproduce claims that reflect what is common in the data, even if those claims are misleading. Importantly this process does not require direct interference with the system. The influence appears through the data itself and through the way the model learns to imitate the language it has seen.

This makes large language models sensitive to what researchers describe as data contamination. Bommasani et al. [2] explain that models trained on the public internet cannot avoid being shaped by the quality of the information that circulates there. If false or biased content becomes widespread, it can enter the training material and become part of the model's behaviour. Recent research by Anthropic further demonstrates that even small, targeted samples of problematic data can disproportionately influence model outputs, highlighting how vulnerable training processes are to manipulation and poisoning [1]. In the context of propaganda this creates a situation where repeated narratives may later reappear through the model in more polished and persuasive form. Shao [8] describes this as a feedback effect since earlier distortions can be reinforced by the model and later reintroduced to the online environment.

A second important development concerns scale. Large language models make it possible to produce text at a speed and volume that were not achievable through human labour alone. Once a model has been trained it can generate unlimited amounts of text at very low cost. Tasks which previously required large teams, such as writing comments, creating posts, or producing opinion pieces, can now be done by a small group with access to generative tools. This shift lowers the cost of influence operations and removes many of the practical constraints that previously limited the spread of propaganda. In economic terms it increases the supply of misinformation at almost no marginal cost. These tools can create messages in many languages, adapt tone and style to different audiences, and respond instantly to ongoing discussions. As a result, influence operations can move from manual work to continuous and automated production.

This automation also changes the nature of online identities. Large language models can generate synthetic personas that appear consistent, realistic, and human. This blurs the line between genuine public debate and artificial participation. When synthetic personas are mixed with real citizens the very idea of a democratic conversation becomes harder to maintain. They can imitate conversational habits preferences and writing style in a way that makes detection difficult. Such systems can operate across multiple platforms and interact in ways that give the impression of genuine public engagement. In the context of propaganda this allows actors to spread political messages in a way that blends seamlessly into ordinary online activity [2].

These capabilities introduce several risks. The first concerns credibility. Once AI generated text becomes indistinguishable from human writing users can no longer rely on typical cues to judge whether a message comes from an individual or from a generative system [8]. The result is an environment where trust becomes a scarce resource. Even accurate information may be doubted simply because users cannot verify its origin. This uncertainty weakens trust in online information as a whole. A second risk involves amplification. Because models can produce content at such high volume false or misleading information can circulate faster than fact-checking or content moderation can respond. Traditional countermeasures such as removing harmful posts or verifying sources are challenged by the speed and scale of automated production.

A further risk is feedback contamination. If large language models are trained on material that includes their own earlier outputs the model may gradually learn from versions of itself. Bommasani et al. [2] note that this can create cycles in which earlier mistakes or biases reappear in new forms. This dynamic matters because it shows how misinformation can accumulate over time. The issue is not one mistake but the gradual drift of the entire information ecosystem away from reliable grounding. Over time this feedback loop can distort the model's behaviour in a way that is difficult to correct. Shao [8] raises similar concerns in relation to hallucination, which refers to the tendency of a model to generate statements that sound plausible but are not supported by any actual evidence. If hallucinated content appears online and is later included in new training data the model may reinforce its own inaccuracies.

Finally there are questions of responsibility. When propaganda is produced by automated systems rather than human writers it becomes more difficult to identify the source of influence. This lack of accountability undermines democratic institutions which depend on the ability to trace responsibility and assign blame. Propaganda without authorship becomes harder to regulate and easier to deny. The person who develops the model, the organisation that deploys it, and the actors who insert strategic content into the information environment all contribute to the outcome. Yet no single individual controls the entire process. Bommasani et al. [2] argue that this lack of accountability is one of the central challenges of large language models since it complicates efforts to manage their social impact.

Figure 3.1 gives a simple visual overview of how feedback contamination can appear inside a large language model. The outer area represents the full training data that the model learns from. Inside this space there are smaller areas that show different types of overlap between new material and earlier sources. The light red field illustrates everything that could be contaminated. The darker centre shows rephrased samples which are harder to detect because the words are new but the meaning is almost the same. The left side of the figure shows how surface level methods such as n gram overlap can identify direct repetition. The right side shows how similarity searches based on embeddings can reveal deeper connections between texts, even when the wording has changed. The figure also shows that some contaminated material may remain inside the training data, even after attempts to clean it. This illustrates why contamination can be difficult to remove. Some patterns sit inside the training material in subtle ways and can re appear when the model generates new text.

Together, these developments show that large language models do not simply change how propaganda is created. They change the structure of propaganda itself. The influence of information becomes continuous automated and integrated into the same digital spaces where people carry out their everyday communication. It operates with a speed and scale that make traditional protective measures less effective. For this reason, understanding the relationship between training data model behaviour and the surrounding information environment is essential for assessing the role of large language models in modern propaganda and for identifying how democratic societies can respond.

3.5 Solutions and Countermeasures

There is no single solution to the problems created by misinformation and AI-driven propaganda. Instead, several approaches need to work together to reduce the impact of false content and strengthen the resilience of democratic societies. These approaches involve technological tools, policy interventions, economic incentives, and education. None of them can solve the problem alone, but each addresses a different part of the system that allows misinformation to spread.

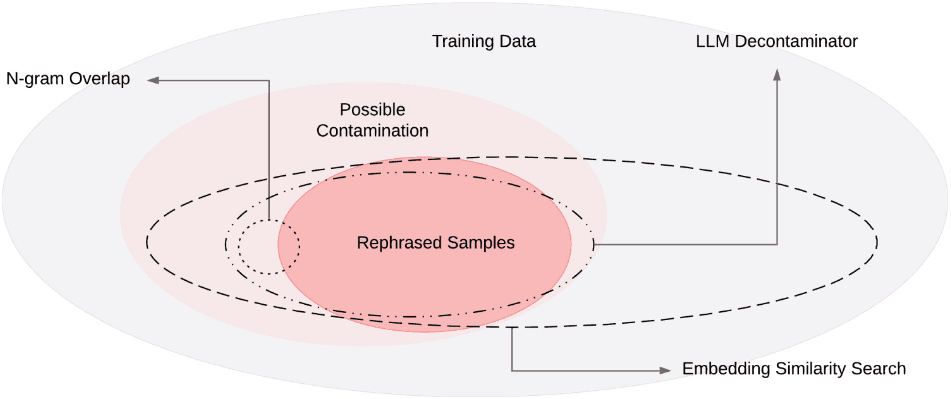


Figure 3.1: Feedback contamination in LLMs

One of the most common responses is fact-checking. Verifying information and exposing false claims is an important part of any strategy against misinformation. The challenge is that fact-checking becomes harder as the volume of online content continues to grow. Automated systems can create thousands of posts, articles, and comments in very little time, which makes traditional verification slow in comparison. Tools that claim to detect AI-generated text are often presented as a solution, but the technology is still too unreliable to depend on. Language models evolve quickly, which means detectors tend to fall behind and sometimes classify human writing as synthetic. Because of this, it is not realistic to expect an AI system to reliably determine which information is true or false. Instead, fact-checking needs to be supported by human oversight and by clearer rules about transparency so that information can be traced, understood, and evaluated.

A second dimension concerns the economic incentives behind online platforms. As long as digital companies earn money from engagement, they indirectly profit from content that spreads quickly, even if that content is misleading. Outrage and sensational stories generate attention, and attention generates advertising revenue. This means that the economic structure of the platforms creates little motivation to stop disinformation. To change this dynamic, platforms need to face economic consequences when harmful content spreads. Possible measures include financial penalties or requirements that reduce the profitability of disinformation. Greater transparency about how algorithms select and promote content may also help limit engagement-driven incentives. In other words, the system needs to shift towards rewarding accuracy and trust rather than emotional reactions and clicks.

Policy and regulation form a third area where significant efforts are already underway. At the European level, the Digital Services Act introduces obligations for large online platforms to reduce systemic risks connected to disinformation. It requires transparency regarding how algorithms operate and demands regular assessments of how harmful content spreads. If companies fail to address these risks, they may face sanctions. Regulation does not replace the need for fact checking or platform responsibility, but it establishes clearer expectations and ensures that companies cannot ignore the problem. In Sweden, the Civil Contingencies Agency works to increase public awareness through campaigns on source criticism, educational programs, and analysis of foreign information influence. These initiatives aim to strengthen society's ability to identify manipulation and to make the information environment more transparent.

Education and media literacy represent another key element of any long-term, bottom-up solution. Even with strong regulation and improved platform policies, users need the ability to judge information for themselves. Developing digital competence helps people recognise manipulative content, understand how algorithms shape what they see, and identify signals of coordinated influence. The European Digital Competence Framework outlines what digital literacy should include and offers guidance on how these skills can be taught. Expanding such frameworks in schools and universities may help create a more informed and critical population. This is especially important because the success of disinformation depends on user behaviour. If people stop clicking on misleading content and avoid sharing unverified claims, the financial incentive behind misinformation weakens. In this way, improving media literacy not only protects individuals but also reduces the return on investment for those who attempt to spread propaganda.

Taken together, these approaches show that the challenge of misinformation cannot be solved through technology alone. Automated detectors and fact-checking tools play a role, but they cannot compensate for economic structures that reward attention or for political actors that intentionally manipulate the information environment. Regulation can create new responsibilities for platforms and demand greater transparency, while education helps citizens navigate a complex digital landscape. When combined, these strategies can make societies more resilient by reducing the spread of false content and strengthening trust in public information. The aim is not to eliminate misinformation entirely but to limit its influence and make it less profitable for the actors who depend on it.

3.6 Evaluation and Discussion

Evaluating the current landscape of propaganda and misinformation shows that no single factor explains its impact. Instead, it is the interaction between political strategy, economic incentives, and technological change that shapes the outcome. Russia's methods remain effective not because they are especially sophisticated, but because they fit the logic of today's digital environment. High volume messaging, emotional framing, and repetition work well in a system built to reward engagement. The firehose approach does not require every claim to be convincing; it only needs to create confusion and weaken trust.

At the same time, it is difficult to measure how effective these methods actually are. Influence does not leave clear traces, and people rarely shift their opinions in ways that can be directly linked to one message or campaign. Public attitudes are shaped through a mix of personal beliefs, political events, and social context, which makes it challenging to isolate the precise impact of Russian activity. As a result, the visibility of propaganda online does not necessarily equal influence. Nonetheless, the alignment between Russia's strategy and the incentives of digital platforms – fast distribution, emotional content, and minimal verification – suggests that these methods can exploit existing vulnerabilities in democratic societies. Even when the exact scale of their effect is uncertain, the structural fit between strategy and platform dynamics remains significant.

Sweden provides a useful example of how these forces interact. As a small, open, and highly digitalised society, Sweden is exposed to influence efforts because information spreads quickly and public debate is easily accessible. At the same time, the country also has strong institutions, high levels of education, and a relatively trusted media system. This combination makes Sweden both vulnerable and resilient. External actors can reach Swedish audiences with little effort, but Swedish society also has the capacity to analyse and resist such attempts. In this sense, Sweden illustrates a broader challenge faced by many democracies: openness brings strength, but also risk.

The rise of large language models adds a new layer of uncertainty. These systems can scale up propaganda efforts and accelerate the spread of misleading information, but it is still unclear whether their long-term impact can be controlled. One open question is whether LLM-driven propaganda can be limited, or whether society is entering an arms race between automated influence systems and increasingly advanced detection tools. Another question concerns responsibility. Commercial AI companies such as OpenAI, Google, and Meta now operate technologies that can shape public understanding on a global scale.

Their decisions about training data, safety measures, and transparency directly influence how vulnerable societies are to manipulation.

There is also the issue of how small, digitally open countries will be affected. Nations like Sweden may face a higher level of risk, simply because their information environments are easy to access and their populations are active online. If propaganda becomes fully automated, the scale of influence attempts may grow faster than the capacity of institutions to respond.

Finally, I argue that the discussion must acknowledge a fundamental limitation: the human brain does not function like a computer. People do not process information by calculating probabilities or scanning datasets. Instead, they rely on emotions, social cues, and cognitive shortcuts. Propaganda succeeds when it exploits these tendencies, and AI-driven content may intensify this effect by producing messages that feel personal and believable. This means that technological solutions alone will never be enough. Any response must also consider how people interpret and react to information, because the ultimate target of propaganda is not the system, but the human mind.

3.7 Conclusion

This report shows that the relationship between propaganda, technology, and the digital economy is more complex than it first appears. Large language models and automated systems have changed the speed and volume of online communication. Yet the main driving forces remain human. Political motives, commercial interests, and cognitive habits still shape how information is created and shared. AI amplifies these patterns but does not produce them on its own. Understanding this interaction is essential for assessing the challenges that modern societies face.

One clear conclusion is that the economic structures of the online environment often work against the protection of public debate. Digital platforms mainly earn money from user engagement. Engagement tends to increase when content triggers strong emotional reactions. This creates a structural conflict of interest. Efforts to limit harmful or misleading content can directly reduce platform profits. Content moderation is expensive to implement. Reducing the visibility of sensational content can also lower advertising revenue. Investigative reporting by Reuters [7] shows that a significant share of Meta's income may come from fraudulent advertising. Financial reporting by CNBC [3] further indicates that such ads were projected to account for around 10 percent of Meta's total revenue in 2024. These findings show how economic incentives can discourage strict enforcement and instead allow harmful content to persist. As a result, technological development often focuses on automation and scale rather than trust and accuracy. Without meaningful changes to the economic incentives that shape the digital environment, technical solutions alone are unlikely to lead to lasting improvements in the quality of public discourse.

Another conclusion is that large language models blur the line between genuine human communication and artificial content. These systems can produce writing that feels personal and confident, which makes them effective tools in influence operations. Yet the impact depends on human choices. The models do not decide which narratives should be pushed. They do not decide which conflicts should be amplified. It is the people who use them who make these decisions. This highlights a broader point in the report: the main problem is rarely the tool itself, but the intention behind it.

The human mind is also a crucial part of this picture. People react to emotion long before they react to facts. Stories, identities, and symbols shape how we interpret information. Propaganda has always taken advantage of this. AI-generated messages can intensify the effect because they can be tailored to feel relevant and familiar. This means that critical thinking and media literacy remain the strongest forms of defence. Technology can support these skills, but it cannot replace them. Individuals still need to question sources and recognise messages that target emotion rather than reason.

The examples discussed in the report show how propaganda adapts to the structure of modern information systems. Russia uses a combination of older influence traditions and new digital tools, and these methods fit well into the attention-driven logic of online platforms. Sweden's position as a small, open, and highly digital society shows how geopolitical interests, domestic debates, and platform incentives can come together to create both risk and resilience. Strong institutions, high education levels, and a trusted media landscape can limit many attempts at manipulation. At the same time, the country's openness and political climate create opportunities for external actors. This combination shows that resistance to influence depends on more than technology; it grows out of the political and social structure of a society.

Looking ahead, influence operations will likely become more automated and more difficult to detect. Large language models are already used in these efforts, and their role will only grow as the technology becomes faster and cheaper to use. Future risks include a situation where synthetic content becomes so common that people can no longer trust anything they see online. There is also a growing concern that propaganda systems and defensive systems will develop in parallel, creating a kind of race in which each side tries to outpace the other. Another uncertainty is the possibility that AI models will interact with each other in unpredictable ways, potentially allowing misinformation to spread without direct human intention. These trends raise a core question: can democratic institutions and regulatory systems adapt quickly enough to manage a communication landscape that evolves at the speed of computation? Progress will require cooperation between governments, researchers, digital platforms, and citizens. Above all, it will require recognition that technology tends to magnify human choices rather than replace them.

The challenge of misinformation is not only a struggle against false content. It is a question of how societies organise their information systems and how they hold commercial actors accountable. It is also a question of how individuals understand the world around them. Large language models will continue to evolve, and strategies for manipulation will evolve with them. This makes it necessary to strengthen the human abilities that no machine can provide. Critical thinking, transparency, and democratic norms remain the core of any defence. By addressing the economic conditions that reward harmful content and by investing in the resilience of citizens, democratic societies can protect the integrity of public debate at a time when influence can be produced as quickly as a line of text.

Bibliography

[1] Anthropic, UK AI Security Institute, and Alan Turing Institute. A small number of samples can poison llms of any size. `https://www.anthropic.com/research/small-samples-poison`, 2025.

[2] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Sanjeev Arora, Samuel von Arx, Michael S. Bernstein, et al. On the opportunities and risks of foundation models. Technical report, Stanford CRFM, 2021. URL `https://crfm.stanford.edu/assets/report.pdf`.

[3] CNBC. Meta reportedly projected 10 percent of 2024 sales came from scam and fraudulent ads. `https://www.cnbc.com/2025/11/06/meta-reportedly-projected-10percent-of-2024-sales-came-from-scam-fraud-ads.html`, 2025.

[4] Thomas H. Costello, Gordon Pennycook, and David G. Rand. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):1164–1165, 2024. doi: 10.1126/science.adq1814.

[5] Martin Kragh. Russian information influence operations towards sweden. Technical report, UI/SCEEUS, Stockholm, 2024. In Swedish.

[6] Christopher Paul and Miriam Matthews. The russian firehose of falsehood propaganda model. Technical report, RAND Corporation, Santa Monica, CA, 2016. URL `https://www.rand.org/pubs/perspectives/PE198.html`.

[7] Reuters Investigations. Meta is earning a fortune from a deluge of fraudulent ads, documents show. `https://www.reuters.com/investigations/meta-is-earning-fortune-deluge-fraudulent-ads-documents-show-2025-11-06/`, 2025.

[8] A. Shao. Beyond misinformation: A conceptual framework for studying ai hallucinations in (science) communication, 2025. URL `https://arxiv.org/abs/2504.13777`. arXiv:2504.13777.

[9] Swedish Institute of International Affairs (UI/SCEEUS). Russian information influence operations towards sweden. Technical report, UI/SCEEUS, 2024. URL `https://www.ui.se/globalassets/ui.se-eng/publications/sceeus/2024-publications/russian-information-influence-operations-towards-sweden.pdf`.

[10] Ekeoma Uzogara et al. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 2025. doi: 10.1126/science.adq1814.

[11] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018. URL `https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308`.

Chapter 4

Exploring Governance in Self-Sovereign Identity Wallets

Saksham Joshi

Self-Sovereign Identity (SSI) seeks to decentralize digital identity by granting individuals control over their personal data and credentials. While existing research extensively examines the cryptographic foundations, standards, and technical architectures of SSI systems, considerably less attention has been paid to the governance of SSI wallets—the primary interface through which users experience self-sovereign identity in practice. This report addresses this gap by analyzing governance structures in SSI wallets and assessing how they support or undermine the principles of decentralization and user self-sovereignty.

The study first reviews relevant literature on SSI foundations, decentralized governance, usability trade-offs, and wallet security to establish the research gap. It then introduces a conceptual framework based on four governance dimensions: control, transparency, accountability, and resilience. Using a qualitative, comparative methodology, the framework is applied to existing SSI wallet ecosystems to examine governance concentration, decision-making authority, and institutional dependencies.

The analysis reveals that while SSI wallets enable operational control over credentials, governance authority frequently remains centralized within foundations, corporations, or public institutions. Open-source development and protocol decentralization do not necessarily result in decentralized governance, and usability, interoperability, and regulatory requirements often reinforce centralization. The report concludes that governance—rather than cryptographic design alone—represents a critical determinant of practical self-sovereignty in SSI wallets. It argues that sustainable SSI ecosystems require explicit, transparent, and resilient governance arrangements that acknowledge and manage the trade-offs between decentralization, usability, and accountability.

Contents

4.1 Introduction

Self-Sovereign Identity (SSI) represents a shift away from traditional, centralized identity systems by enabling individuals to control and manage their digital identities directly. Instead of depending on governments, corporations, or digital platforms to authenticate users, SSI distributes trust and gives user ownership over their credentials and cryptographic keys. Within this architecture, the SSI wallets server is the core user-facing component, functioning as the primary interface through which individuals store, present, and manage verifiable credentials. Because every SSI interaction-whether issuance, verification, or presentation-happens through the wallet, its design fundamentally shapes how users experience autonomy and control.

Despite this promise of decentralization, SSI wallets introduce a structural tension: although users control their keys and credentials, the wallet software itself is governed by the developers or organizations that build and maintain it. These actors determine which features are supported, how standards are implemented, where updates are deployed, and what data may be collected. This creates a hidden dependency that can reintroduce forms of centralization, even within a system designed to eliminate them. The contradiction is significant because, without transparent governance, users cannot reliably assess whether the wallet truly operates in their interest, preserves their privacy, or upholds the principles of self-sovereignty.

Existing academic work on SSI largely focuses on technical components-such as decentralized identifiers (DIDs), verifiable credential protocols, cryptographic primitives, interoperability frameworks, and privacy-preserving verification mechanisms. In contrast, governance structures behind SSI wallets remain comparatively under-examined. Questions of transparency, accountability, decision-making power, update control, and the socio-technical dynamics of wallet maintenance receive far less attention, despite being crucial for evaluating whether SSI systems genuinely achieve decentralization in practice. Governance therefore represents a critical research gap.

This report addresses this gap by examining SSI wallet governance through the lens of transparency. The central research question guiding the study is: **To what extent do governance structures in SSI wallets support or undermine the principles of decentralization and user self-sovereignty?** The analysis further considers who controls wallet updates and standards adoption, how transparent and accountable existing governance models are, and which governance structures might strengthen user autonomy and trust.

The purpose of this report is to investigate how transparency(or lack thereof) in wallet governance affects user control, decentralization, and trust within SSI ecosystems. By comparing governance approaches across different types of wallet and identifying the risks posed by opaque decision-making, the report proposes a conceptual framework for more transparent, accountable, and resilient wallet governance. Strengthening governance is essential not only for the credibility of SSI systems but also to ensure that self-sovereignty is realized in practice rather than remaining an idealistic design goal.

4.2 Literature Review

4.2.1 SSI Foundations and Technical Architecture

Self-Sovereign Identity (SSI) is introduced in the literature as a response to the structural limitations of centralized digital identity systems, which concentrate control over identity data in the hands of institutions and platform providers. Foundational work defines SSI as an identity model in which individuals retain direct control over their identifiers and credentials, enabling greater autonomy, privacy, and portability across digital contexts without reliance on centralized authorities [1]. A core principle underlying SSI is the reduction of unnecessary data disclosure and the minimization of trust placed in intermediaries during identity verification.

From a technical standpoint, SSI architectures are largely standardized around a small set of interoperable components. Decentralized Identifiers (DIDs) provide globally unique, cryptographically verifiable identifiers that are controlled by the identity subject rather than issued or managed by centralized registries [10]. Verifiable Credentials (VCs) define a structured data model and associated cryptographic mechanisms that allow trusted entities to issue attestations about identity attributes, which holders can later present to verifiers in a tamper-evident and privacy-preserving manner [11]. These mechanisms are typically supported by public-key cryptography, digital signatures, and, in many implementations, distributed ledger technologies that enable identifier resolution and trust anchoring.

Across surveys and systematic reviews, SSI ecosystems are consistently described through the interaction of three primary roles: issuers, holders, and verifiers. Issuers attest to claims by issuing credentials, holders manage and selectively disclose these credentials, and verifiers assess their validity without requiring direct access to the underlying personal data. Comparative studies emphasize that standardized protocols and data models allow interoperability across SSI implementations while preserving user privacy and reducing dependence on centralized identity infrastructures [5][4].

Within this technical framing, wallets are introduced as the software artifacts responsible for managing cryptographic keys, storing decentralized identifiers, and holding verifiable credentials on behalf of users. Wallets are therefore acknowledged as essential operational components within SSI systems. However, existing literature predominantly discusses wallets in functional and technical terms—such as credential storage, protocol compliance, and security properties—while offering limited analysis of how wallet software is developed, updated, or governed in practice [5][12].

4.2.2 Governance in Blockchain and Decentralized Systems

Governance has emerged as a central research topic in blockchain and decentralized systems, addressing the mechanisms through which collective decisions are made, rules are enforced, and system evolution is coordinated without centralized authority. Existing literature conceptualizes governance as a combination of technical mechanisms and social processes that determine how changes to protocols, rules, and organizational structures are proposed, evaluated, and implemented [2]. In decentralized environments, governance is often presented as a core challenge due to the absence of formal hierarchies and the need to balance decentralization with effective coordination.

A significant portion of this literature focuses on governance models implemented through Decentralized Autonomous Organizations (DAOs). DAO governance mechanisms typically rely on token-based voting, delegation, and proposal systems to enable stakeholder participation in decision-making [2]. Studies identify a range of governance approaches, including on-chain governance—where decisions are encoded and executed directly by smart contracts—and off-chain governance, where deliberation and coordination occur through informal channels such as forums, working groups, or core development teams [2]. While these mechanisms aim to enhance transparency and decentralization, empirical analyses highlight persistent challenges such as voter apathy, governance capture by large token holders, and the concentration of decision-making power among a small subset of participants.

Recent governance reviews further emphasize that decentralization in governance is often more limited in practice than suggested by formal mechanisms. Although DAOs provide participatory frameworks, many critical decisions are influenced by core developers, multisignature committees, or informal leadership structures that operate outside formal voting processes [2]. As a result, governance outcomes may depend less on broad stakeholder participation and more on technical expertise, economic power, or social influence. These findings suggest that transparency and accountability in decentralized governance cannot be assumed solely on the basis of open voting mechanisms.

Despite the maturity of governance research at the protocol and organizational levels, this literature largely treats governance as a property of blockchain networks or DAOs rather than of user-facing software. Wallets, which mediate user interaction with decentralized systems and shape how governance decisions are experienced in practice, are not analyzed as governance actors within existing frameworks. Consequently, governance studies provide valuable theoretical insights into decentralization, participation, and control, but offer limited guidance on how governance is exercised at the interface level where users interact with decentralized identity systems.

4.2.3 Usability vs Decentralization Trade-offs

A recurring theme across the SSI and broader blockchain literature is the tension between decentralization and usability. While decentralized architectures aim to maximize user autonomy and minimize reliance on trusted intermediaries, they often impose significant cognitive and technical burdens on end users. Surveys of SSI systems consistently identify usability as a critical non-functional requirement, noting that complex key management, recovery procedures, and unfamiliar interaction patterns can hinder adoption despite strong privacy and security guarantees [12][5].

In the context of SSI, decentralization requires users to manage cryptographic keys, control identifiers, and make informed decisions about credential disclosure. Comparative studies highlight that these responsibilities, while essential for self-sovereignty, introduce usability challenges that are not present in centralized identity systems [12]. As a result, many SSI implementations adopt design choices that trade strict decentralization for improved user experience, such as simplified recovery mechanisms, abstracted key management, or reliance on trusted infrastructure components. These choices are frequently justified as necessary for practical deployment and user adoption.

Related literature emphasizes that usability pressures often lead to the reintroduction of centralized elements within ostensibly decentralized systems. Surveys and taxonomies of SSI solutions show that wallets frequently embed assumptions about trusted services, default configurations, or predefined workflows to reduce user complexity [5] [4]. While these design decisions can enhance accessibility, they may also shift control away from users in subtle ways, particularly when users lack visibility into how wallet software operates or evolves over time.

Importantly, existing studies treat usability trade-offs primarily as design or engineering challenges rather than governance issues. Decisions about which trade-offs to prioritize—such as convenience over decentralization or automation over user control—are rarely analyzed in terms of who makes these decisions, how they are justified, or whether users have meaningful influence over them. Consequently, the literature recognizes the existence of usability–decentralization trade-offs but does not examine how governance structures within wallets institutionalize these trade-offs and shape long-term user autonomy.

4.2.4 Wallet Security and Threat Models

Security considerations form a substantial part of the existing literature on SSI wallets and related blockchain-based systems. Academic surveys and systematic reviews analyze wallets primarily through the lens of technical threat models, focusing on risks associated with cryptographic key management, credential storage, recovery mechanisms, and interaction with potentially malicious environments [4][12][5]. These studies frame wallets as critical security components, as compromise of wallet integrity can directly result in identity theft, credential misuse, or loss of access to identity data.

A dominant concern across the literature is private key management. SSI wallets require users to securely generate, store, and protect cryptographic keys that control identifiers and credentials. Surveys highlight threats such as key loss, phishing attacks, malware, and insecure backup practices, all of which can undermine the security guarantees of SSI systems even when underlying protocols are correctly implemented [12][5]. To mitigate these risks, many wallet implementations introduce recovery mechanisms, such as social recovery, custodial backups, or trusted recovery agents. While these mechanisms improve resilience against key loss, they may also introduce additional trust assumptions and potential attack surfaces.

Beyond key management, the literature also examines privacy-related risks associated with wallet usage. Studies note that improper handling of metadata, network interactions, or credential presentation patterns can lead to unintended information leakage, enabling correlation or tracking of users across contexts [4][5]. Wallets are therefore evaluated in terms of their ability to support selective disclosure, unlinkability, and minimal data exposure. However, these evaluations remain largely technical, focusing on protocol compliance rather than operational practices.

Importantly, existing threat models generally assume that wallet software behaves as intended and that updates or design changes are inherently benign. Security analyses rarely consider who controls wallet updates, how security-critical decisions are made, or how emergency interventions are governed in practice. As a result, wallets are treated as static technical artifacts rather than evolving software systems embedded within governance structures.

4.2.5 Synthesis and Research Gap

The literature reviewed above demonstrates that research on Self-Sovereign Identity has reached a high level of maturity with respect to technical architecture, cryptographic mechanisms, and ecosystem design. Foundational works establish clear principles of decentralization, user control, and privacy, while standards such as Decentralized Identifiers and Verifiable Credentials define interoperable technical building blocks for SSI systems [1] [10] [11]. Survey and review papers further analyze SSI implementations in terms of functional requirements, security properties, and usability considerations, providing comprehensive overviews of the current state of the ecosystem [5][4][12].

At the same time, governance has emerged as a well-developed area of inquiry within the broader blockchain literature. Studies of decentralized governance and DAOs examine how decision-making power is distributed, how protocol changes are coordinated, and how transparency and accountability are maintained in decentralized systems [2]. These works highlight persistent governance challenges, including concentration of power, informal control structures, and the gap between formal decentralization and practical decision-making. However, governance research remains largely focused on protocols and organizations rather than on user-facing software components.

When considering usability and security, the literature recognizes that SSI wallets sit at the intersection of competing objectives. Usability studies document the trade-offs between decentralization and user convenience, while security analyses develop detailed threat models addressing key management, recovery, and privacy risks [4][12][5]. Yet, these discussions frame wallets primarily as technical artifacts whose behavior is defined by design choices and protocol compliance, rather than as evolving systems shaped by ongoing governance decisions.

Taken together, this body of literature reveals a significant gap. Although SSI wallets are the primary interface through which users experience self-sovereign identity in practice, their governance structures remain largely unexamined. Decisions regarding wallet development, feature inclusion, update mechanisms, and standards adoption are typically made by foundations or development organizations operating in a centralized manner, even as the surrounding identity architecture aspires toward decentralization. The literature does not systematically analyze how such governance arrangements affect transparency, nor how they shape the extent to which users exercise meaningful control over their identity data. This gap motivates the present study, which identifies an overlooked problem in existing SSI research and develops a conceptual framework for analyzing governance in SSI wallets. By focusing on transparency in wallet governance, the report seeks to clarify how centralized decision-making at the wallet level can undermine the principles of decentralization and user self-sovereignty that SSI systems are intended to uphold.

4.3 Conceptual Framework

To systematically analyze governance in Self-Sovereign Identity (SSI) wallets, this report introduces a conceptual framework that operationalizes governance at the wallet level. The framework is motivated by the research gap identified in Section 2, namely that while SSI architectures aim to decentralize identity control, wallet governance remains insufficiently examined despite wallets serving as the primary interface through which users experience SSI in practice. Rather than treating wallets as neutral technical artifacts, the framework conceptualizes them as socio-technical systems whose governance structures shape user autonomy, transparency, and trust.

The framework comprises four governance dimensions—Control, Transparency, Accountability, and Resilience—which together capture the key aspects through which governance manifests in SSI wallets. These dimensions are derived from recurring concerns in the literature on SSI architecture, decentralized governance, usability trade-offs, and wallet security, and they provide a structured basis for comparative and analytical evaluation.

4.3.1 Control

Control refers to the distribution of decision-making power over the evolution and operation of an SSI wallet. This includes who determines changes to the wallet codebase, who controls update mechanisms, and who decides which decentralized identifier (DID) methods, credential formats, or standards are supported. Although SSI principles emphasize user control over identity data, control at the wallet level may remain centralized if key decisions are made exclusively by foundations, core development teams, or sponsoring organizations. Within this framework, control captures the extent to which users can meaningfully influence wallet behavior beyond mere usage, and whether governance authority is distributed or concentrated.

4.3.2 Transparency

Transparency concerns the visibility and openness of governance processes related to SSI wallets. This dimension examines whether decisions about wallet development, updates, and feature inclusion are documented, publicly communicated, and accessible to users. Transparent governance enables users to understand how and why changes occur, assess potential risks, and evaluate whether wallet behavior aligns with SSI principles. In the absence of transparency, governance decisions may remain opaque, limiting users' ability to make informed choices and undermining trust in the wallet as an instrument of self-sovereign identity.

4.3.3 Accountability

Accountability addresses responsibility and answerability within wallet governance structures. This dimension considers who bears responsibility when governance decisions lead to security failures, privacy violations, or loss of user autonomy. In decentralized identity systems, accountability can be difficult to establish due to diffuse roles and informal decision-making processes. The framework therefore examines whether clear mechanisms exist for assigning responsibility, addressing user grievances, and responding to governance failures. Accountability is critical for evaluating whether SSI wallets provide not only technical self-sovereignty but also institutional safeguards for users.

4.3.4 Resilience

Resilience refers to the robustness of wallet governance against capture, coercion, or systemic failure. This includes the ability of governance structures to withstand concentration of power, external pressure from regulators or sponsors, and disruptions such as developer withdrawal or infrastructure failure. Resilient governance mechanisms reduce reliance on single points of control and enhance the long-term sustainability of SSI wallets. Within the framework, resilience captures whether governance arrangements can adapt to challenges while preserving decentralization and user trust.

4.3.5 Application

Together, these four dimensions provide an analytical lens for evaluating SSI wallets in a structured and comparable manner. By examining how different wallets allocate control, ensure transparency, enforce accountability, and maintain resilience, the framework enables systematic analysis of governance practices at the wallet level. This approach allows the report to move beyond purely technical evaluation and to assess whether SSI wallets, in practice, align with the principles of decentralization and self-sovereignty that underpin the SSI paradigm.

Table 4.1: Governance Dimensions and Analytical Focus

Governance Dimension	**Analytical Focus**
Control	Who decides updates, standards, and core functionality of the SSI wallet
Transparency	Visibility and documentation of governance processes and decision-making
Accountability	Allocation of responsibility and available recourse when failures or disputes occur
Resilience	Robustness of governance structures against capture, coercion, or systemic breakdown

4.4 Methodology

This study adopts a qualitative, comparative governance analysis to examine how governance is operationalized within Self-Sovereign Identity (SSI) wallets. Given the conceptual nature of the research question and the absence of primary data collection or fieldwork, the methodology focuses on systematic analysis of existing SSI wallet implementations through the analytical framework introduced in Section 3. This approach allows for structured comparison across wallets while maintaining alignment with the principles and goals of SSI.

4.4.1 Research Design

The research design is exploratory and comparative. Rather than evaluating wallets in terms of performance or technical efficiency, the analysis focuses on governance characteristics and decision-making structures. SSI wallets are treated as socio-technical systems whose governance models influence user autonomy, transparency, and trust. By applying a consistent set of governance dimensions across multiple cases, the methodology enables identification of patterns, similarities, and divergences in wallet governance practices.

4.4.2 Case Selection

The study selects three to five representative SSI wallets to capture variation across governance models and institutional contexts. Wallets are chosen to reflect diversity along the following dimensions:

- Open-source community-driven wallets, where governance is primarily exercised by distributed developer communities.
- Corporate-backed wallets, developed and maintained by private organizations.
- Institutionally or government-backed wallets, where governance is linked to public-sector or semi-public entities.

This selection strategy ensures that the analysis is not limited to a single governance paradigm and allows comparison of how different institutional arrangements shape wallet governance in practice.

4.4.3 Analytical Procedure

Each selected wallet is evaluated using the four governance dimensions defined in the conceptual framework: Control, Transparency, Accountability, and Resilience. For each dimension, qualitative indicators are examined, such as:

- who controls the wallet codebase and update mechanisms,
- how governance decisions are communicated and documented,
- whether responsibility for failures or disputes is clearly defined,
- and how governance structures respond to risks such as capture, coercion, or organizational failure.

The analysis relies on publicly available documentation, governance statements, development practices, and observable decision-making processes associated with each wallet. Applying the same analytical lens to each case ensures methodological consistency and comparability.

4.4.4 Comparative Analysis

The analysis relies on publicly available documentation, governance statements, development practices, and observable decision-making processes associated with each wallet. Applying the same analytical lens to each case ensures methodological consistency and comparability.

4.4.5 Methodological Limitations

This methodology is subject to several limitations. First, the analysis is based on publicly available information and does not incorporate insights from internal decision-making processes or stakeholder interviews. Second, governance practices may evolve over time, and the analysis reflects governance structures at a specific point. Despite these limitations, the comparative framework provides a systematic and replicable approach for analyzing governance in SSI wallets and offers a foundation for future empirical or longitudinal research.

4.5 Governance Analaysis of SSI Wallet

This section applies the conceptual framework introduced above to empirically analyze governance practices in selected Self-Sovereign Identity (SSI) wallets. Following the methodology outlined in Section 4, the analysis evaluates wallet governance along four dimensions—Control, Transparency, Accountability, and Resilience—in order to assess how governance structures align with the principles of decentralization and user self-sovereignty. The section begins with an in-depth case analysis of the Sovrin ecosystem as a foundational SSI implementation.

4.5.1 Sovrin Wallet Governance Analysis

Sovrin represents one of the earliest and most influential implementations of Self-Sovereign Identity, positioning itself as a global public utility for decentralized identity. The Sovrin ecosystem is built on top of Hyperledger Indy and is governed by a formal and comprehensive governance architecture administered by the Sovrin Foundation, a non-profit organization established to steward the Sovrin Network [7]. This governance structure provides a useful case for examining how formalized governance operates within an SSI ecosystem and how it affects wallet-level self-sovereignty.

4.5.1.1 Control

Control within the Sovrin ecosystem is primarily exercised through the Sovrin Foundation and its associated governance bodies. The Sovrin Governance Framework explicitly defines the roles, responsibilities, and authorities involved in operating the Sovrin Network, including decision-making over ledger governance, steward participation, and policy enforcement [3][8]. While SSI principles emphasize user control over identifiers and credentials, strategic and technical decisions—such as updates to governance policies, ledger rules, and network participation criteria—are centrally coordinated and ultimately approved by the Foundation's governing bodies. Wallet implementations interacting with the Sovrin Network therefore operate within a governance environment where core infrastructural decisions remain foundation-led rather than user-driven.

4.5.1.2 Transparency

Transparency is a comparatively strong dimension within Sovrin's governance model. Governance documents, including the Ecosystem Governance Framework and Utility Governance Framework, are publicly available and structured to provide detailed explanations of governance processes, policies, and institutional roles [3][8]. Additionally, the Sovrin Governance Framework Working Group allows for public review and comment on governance documents, enabling community visibility into proposed changes and revisions [9]. However, while documentation is openly accessible, transparency primarily operates at the institutional and policy level rather than at the level of individual wallet users, who may have limited practical engagement with governance deliberations.

4.5.1.3 Accountability

Accountability within the Sovrin ecosystem is formalized through legal and organizational mechanisms. The Sovrin Trust Assurance Framework defines criteria for compliance and conformance among network participants, establishing responsibility structures for stewards, transaction endorsers, and other actors [6]. The Sovrin Foundation assumes overarching responsibility for maintaining the integrity and public-interest orientation of the network. Nevertheless, accountability remains institution-centric: users of SSI wallets have limited direct recourse or participatory mechanisms to influence governance outcomes or address governance-related failures beyond institutional complaint or compliance channels.

4.5.1.4 Resilience

From a resilience perspective, Sovrin's governance model emphasizes stability and regulatory compliance through formalized roles and legal agreements. While this structure enhances resistance to arbitrary changes and provides continuity, it also introduces potential single points of governance concentration. The reliance on the Sovrin Foundation as the central coordinating entity raises questions about resilience against organizational failure, capture, or external pressure. Although the network is technically decentralized, governance resilience is closely tied to the ongoing viability and neutrality of the Foundation itself [7][8].

4.5.1.5 Interim Assessment

Overall, the Sovrin case illustrates a governance model characterized by high formalization and transparency, but centralized control and accountability at the institutional level. While Sovrin's governance architecture aligns with SSI goals in terms of public documentation and rule-based operation, it also demonstrates how wallet-level self-sovereignty can be constrained by foundation-led governance structures operating above the wallet interface.

4.5.2 Comparative Governance Assessment

This subsection situates the Sovrin governance model in relation to the normative principles of Self-Sovereign Identity, as articulated in the literature and standards reviewed in Section 2. Rather than comparing Sovrin to other wallets at this stage, the analysis evaluates the degree of alignment and tension between Sovrin's governance structures and core SSI objectives, including decentralization, user autonomy, and transparency.

From a governance perspective, Sovrin exhibits a strong commitment to formalization and rule-based operation. The existence of a comprehensive governance framework, public documentation, and legally defined roles contributes positively to transparency and predictability in network operations [3][8][6]. Compared to many decentralized systems that rely on informal or opaque governance mechanisms, Sovrin's approach reduces ambiguity regarding authority and responsibility.

However, when evaluated against the principle of user self-sovereignty, notable tensions emerge. While SSI architectures emphasize user control over identifiers and credentials, governance authority within the Sovrin ecosystem remains concentrated at the institutional level. Strategic decisions regarding network rules, policy updates, and participation criteria are coordinated and approved by the Sovrin Foundation and its governing bodies, rather than by wallet users themselves [3][8]. As a result, user sovereignty is primarily exercised at the operational level (credential management) rather than at the governance level.

In terms of accountability, Sovrin's governance model assigns responsibility through formal organizational structures and compliance frameworks, such as the Trust Assurance Framework [6]. This enhances institutional accountability but offers limited mechanisms for direct user participation or redress in governance processes. Accountability is therefore upward-facing—toward stewards and governing bodies—rather than outward-facing toward end users of SSI wallets.

With respect to resilience, Sovrin's governance framework prioritizes stability and regulatory alignment. While this contributes to robustness against arbitrary changes, it also introduces dependencies on the continued neutrality and operational viability of the Sovrin Foundation. Consequently, governance resilience is partially constrained by centralized institutional reliance, creating potential vulnerabilities in scenarios of organizational failure or external pressure.

Taken together, this assessment suggests that Sovrin represents a governance model that is transparent and institutionally accountable, yet centralized in terms of control. This configuration highlights a broader pattern within SSI ecosystems: while technical architectures may enable decentralized identity, governance structures at the wallet and network level can reintroduce centralized authority, thereby shaping the practical limits of self-sovereignty.

4.6 Analysis

This section synthesizes the empirical findings presented in Section 5 to derive broader insights into governance dynamics in Self-Sovereign Identity (SSI) wallets. Rather than reiterating case-specific observations, the analysis focuses on cross-cutting patterns, structural tensions, and recurring trade-offs that shape wallet governance in practice. These insights directly address the research question by evaluating how governance structures support or undermine decentralization and user self-sovereignty.

4.6.1 Governance Concentration as a Systemic Risk

Across examined SSI wallet ecosystems, governance concentration emerges as the primary structural risk to decentralization. Even when underlying identity protocols are decentralized, decision-making authority over wallet development, updates, and compliance is frequently centralized within foundations, corporations, or public institutions. This concentration enables coordination and stability but simultaneously creates dependencies that limit user influence over the evolution of wallet functionality. As a result, decentralization at the protocol level does not necessarily translate into decentralized governance at the wallet level, where strategic control remains consolidated.

4.6.2 Open-Source Does Not Imply Decentralized Governance

A recurring assumption in decentralized systems is that open-source development inherently enables decentralized governance. The empirical analysis challenges this assumption. While open-source SSI wallets provide transparency and auditability, governance power often resides with a small group of maintainers who control code review, merging rights, and release cycles. Users may benefit from visibility into development processes, yet they typically lack formal mechanisms to influence governance outcomes. This distinction highlights that openness of code is a necessary but insufficient condition for governance decentralization.

4.6.3 Usability Pressures Reinforce Centralization

Usability considerations significantly shape wallet governance decisions. Corporate-backed SSI wallets tend to prioritize user experience, streamlined onboarding, and simplified recovery mechanisms. While these design choices reduce friction and encourage adoption, they frequently rely on centralized infrastructure, controlled update channels, or custodial recovery models. The analysis indicates that usability pressures often justify governance centralization as a pragmatic trade-off, reinforcing the tension between ease of use and self-sovereign control. In practice, user convenience is frequently privileged over participatory governance.

4.6.4 Accountability Gains at the Cost of Sovereignty

Government-backed or institutionally anchored SSI wallets exhibit strong accountability characteristics, including clear legal responsibility, regulatory alignment, and formal oversight. However, these strengths are accompanied by reduced decentralization and increased state or institutional control. Accountability is achieved through hierarchical governance structures rather than distributed participation, limiting user autonomy over governance decisions. This pattern underscores a fundamental tension within SSI: mechanisms that enhance accountability may simultaneously constrain self-sovereignty when implemented through centralized authority.

4.6.5 Fully Decentralized Wallet Governance

The comparative analysis suggests that fully decentralized governance in SSI wallets is rare and difficult to sustain. Governance requires coordination, conflict resolution, and long-term maintenance—functions that are challenging to distribute without reintroducing centralized control. Interoperability requirements, compliance obligations, and security responsibilities further incentivize centralized governance structures. Consequently, SSI wallets tend to adopt hybrid arrangements that balance decentralization claims with practical governance constraints.

4.6.6 Interoperability

Interoperability, a core objective of SSI ecosystems, unintentionally strengthens governance centralization. Wallets must align with dominant standards, reference implementations, and certification regimes to remain interoperable. These alignment pressures often concentrate influence in standard-setting bodies, foundations, or early ecosystem leaders. As a result, governance authority becomes centralized not through explicit control, but through dependency on shared infrastructure and standards compliance.

4.6.7 Interim Conclusion

Taken together, these findings indicate that governance—not cryptography or protocol design—is the decisive factor shaping the practical limits of self-sovereignty in SSI wallets. While existing governance models enable stability, usability, and accountability, they frequently do so by constraining user participation and decentralizing control only at a technical level. This analysis highlights the need for governance models that explicitly address these trade-offs rather than assuming that decentralization emerges automatically from open protocols.

4.7 Pathways Toward Sustainable SSI Wallet Governance

Building on the empirical findings and analytical insights presented in Sections 5–7, this section outlines potential pathways toward more sustainable governance in Self-Sovereign Identity (SSI) wallets. Rather than proposing a fully specified governance architecture, the discussion remains exploratory and conceptual, focusing on design implications that address the structural tensions identified in current SSI wallet governance models.

A first implication concerns the need to distinguish explicitly between technical decentralization and governance decentralization. The analysis demonstrates that decentralized protocols alone do not ensure decentralized decision-making at the wallet level. Sustainable SSI wallet governance therefore requires mechanisms that extend self-sovereignty beyond credential control to include visibility and influence over wallet evolution. This suggests that governance should be treated as a first-class design concern, rather than an implicit byproduct of open standards or open-source development.

Second, the findings indicate that transparency must be complemented by participation. While many SSI wallet ecosystems provide extensive documentation and public governance artifacts, transparency alone does not empower users if decision-making authority remains concentrated. Pathways toward sustainability may involve lightweight participatory mechanisms—such as structured community consultations, public rationale for governance decisions, or non-binding user feedback processes—that enhance legitimacy without imposing excessive coordination costs.

Third, accountability mechanisms appear most effective when they balance institutional responsibility with user-facing safeguards. Foundation-led and government-backed wallets offer clarity in responsibility and regulatory alignment, yet often limit user recourse. A more sustainable approach may involve separating operational accountability (e.g., security incidents, compliance) from strategic governance, thereby preserving clear responsibility while reducing concentration of long-term control.

Fourth, the analysis highlights the importance of governance resilience in evolving SSI ecosystems. Reliance on single organizations—whether foundations, corporations, or public authorities—creates vulnerabilities to capture, policy shifts, or organizational failure. Sustainable governance pathways may therefore emphasize redundancy, role separation, and gradual decentralization of decision-making authority as ecosystems mature, rather than attempting full decentralization from the outset.

Finally, interoperability requirements suggest that coordination is unavoidable in SSI wallet governance. Rather than resisting this reality, sustainable governance models should make coordination explicit and transparent, clarifying who sets standards, how changes propagate, and how wallets can exit or contest governance arrangements. Such clarity can mitigate the risks of de facto centralization driven by hidden dependencies.

Taken together, these pathways do not prescribe a single optimal governance model. Instead, they emphasize that sustainable SSI wallet governance is likely to be hybrid, adaptive, and context-dependent—balancing decentralization ideals with usability, security, and regulatory constraints. Recognizing and explicitly addressing these trade-offs is a necessary step toward aligning wallet governance with the broader goals of self-sovereign identity.

4.8 Conclusion

This report examined governance in Self-Sovereign Identity (SSI) wallets as a critical yet underexplored dimension of decentralized identity systems. While SSI architectures are designed to shift control of identity data to individuals, the analysis demonstrated that wallets—the primary interface through which users interact with SSI—often remain governed by centralized actors. This creates a structural tension between the normative goals of self-sovereignty and the practical realities of wallet governance.

Through a structured literature review and a comparative governance analysis, the report showed that existing SSI research largely focuses on cryptographic protocols, standards, and credential formats, while paying limited attention to governance at the wallet level. By introducing a conceptual framework based on control, transparency, accountability, and resilience, the study provided a systematic lens for evaluating how governance structures shape user autonomy in practice. Application of this framework revealed recurring patterns of governance concentration, even in ecosystems built on decentralized protocols.

The findings highlight that open-source development and protocol decentralization do not automatically translate into decentralized governance. Usability pressures, interoperability requirements, regulatory considerations, and security responsibilities frequently incentivize centralized decision-making within foundations, corporations, or public institutions. As a result, user self-sovereignty is often realized at the operational level—through control over credentials—while remaining constrained at the governance level.

At the same time, the analysis suggests that fully decentralized wallet governance is difficult to achieve and may not always be desirable. Governance models that emphasize stability, accountability, and compliance offer important benefits, particularly for identity infrastructures operating in legally and socially sensitive contexts. The challenge, therefore, is not to eliminate governance structures, but to design them in ways that are transparent, resilient, and aligned with the principles of self-sovereign identity.

By framing wallet governance as a first-order concern in SSI system design, this report contributes to a more nuanced understanding of decentralization in practice. It argues that without explicit attention to governance, SSI risks reproducing centralized power dynamics at the wallet layer, undermining its core promise. Future research could extend this work by empirically examining user perceptions of wallet governance, comparing governance practices across a broader set of wallets, or exploring mechanisms for participatory governance that balance decentralization with usability and accountability.

Bibliography

[1] Christopher Allen. The path to self-sovereign identity. Blog post, April 2016. URL https://www.lifewithalacrity.com/article/the-path-to-self-soverereign-identity/. Life With Alacrity.

[2] Jungsuk Han, Jongsub Lee, and Tao Li. A review of dao governance: Recent literature and emerging trends. Technical Report 1044/2025, ECGI Working Paper Series in Finance, March 2025. URL https://ssrn.com/abstract_id=5074046. Available at SSRN.

[3] Sankarshan Mukhopadhyay and Line Kofoed. Sovrin ecosystem governance framework v3.1: Primary document. Technical Report Version 3.1, Sovrin Foundation, August 2023. URL https://sovrin.org/library/sovrin-ecosystem-governance-framework/. Approved by the Sovrin Board of Trustees.

[4] Alan Sherriff, Kaliya Young, and Michael Shea. Editorial: Establishing self-sovereign identity with blockchain. *Frontiers in Blockchain*, 5:955868, 2022. doi: 10.3389/fbloc.2022.955868. URL https://www.frontiersin.org/articles/10.3389/fbloc.2022.955868.

[5] Reza Soltani, Uyen Trang Nguyen, and Aijun An. A survey of self-sovereign identity ecosystem. *Security and Communication Networks*, 2021:1–39, 2021. doi: 10.1155/2021/8873429. URL https://onlinelibrary.wiley.com/doi/10.1155/2021/8873429.

[6] Sovrin Foundation. Sovrin trust assurance framework v1. Technical Report Version 1, Sovrin Foundation, March 2019. URL https://sovrin.org/library/sovrin-trust-assurance-framework/. Controlled Document of the Sovrin Governance Framework v2, approved by the Sovrin Board of Trustees.

[7] Sovrin Foundation. Sovrin: A global utility for self-sovereign identity. Project documentation and technical overview, 2025. URL https://sovrin-foundation.github.io/sovrin/. Accessed December 31, 2025.

[8] Sovrin Governance Framework Working Group. Sovrin utility governance framework v3.1: Primary document. Technical Report Version 3.1, Sovrin Foundation, August 2023. URL https://sovrin.org/library/sovrin-utility-governance-framework/. Approved by the Sovrin Board of Trustees.

[9] Sovrin Governance Framework Working Group. Sovrin governance framework working group meeting page. Meeting agendas and notes (archival documentation), 2023. URL https://sovrin.org/library/sovrin-governance-framework-working-group/. Includes agendas and minutes from 2017–2023; accessed December 31, 2025.

[10] Manu Sporny, Dave Longley, Markus Sabadello, Drummond Reed, Orie Steele, and Christopher Allen. Decentralized identifiers (dids) 1.0: Core architecture, data model,

and representations. W3c recommendation, World Wide Web Consortium (W3C), 2022. URL `https://www.w3.org/TR/did-1.0/`.

[11] Manu Sporny, Dave Longley, David Chadwick, and Ivan Herman. Verifiable credentials data model v2.0. W3c recommendation, World Wide Web Consortium (W3C), May 2025. URL `https://www.w3.org/TR/vc-data-model-2.0/`.

[12] Razieh Nokhbeh Zaeem, Kai Chih Chang, Teng-Chieh Huang, David Liau, Wenting Song, Aditya Tyagi, Manah M. Khalil, Michael R. Lamison, Siddhartha Pandey, and K. Suzanne Barber. Blockchain-based self-sovereign identity: Survey, requirements, use-cases, and comparative study. Technical Report UTCID Report #21-06, University of Texas at Austin, Center for Identity, August 2021. URL `https://identity.utexas.edu/sites/default/files/2021-08/Blockchain-Based%20Self-Sovereign%20Identity-%20Survey,%20Requirements,%20Use-Cases,%20and%20Comparative%20Study.pdf`.

Chapter 5

Payoff-Driven Consensus: Incentive Design for Multi-Agent Federated Reinforcement Learning (RL)

Pierre Obermaier

Federated reinforcement learning (FRL) [1; 2] extends federated learning (FL) [3; 4] to sequential decision-making problems. In FRL, multiple agents collaboratively train a reinforcement learning policy without sharing raw data. Each agent learns from its local environment and periodically exchanges model updates with a central server or with other agents in a peer-to-peer manner. This approach preserves data privacy and ownership, as in FL, but introduces new challenges: agents may represent self-interested organizations with divergent goals, resulting in misaligned incentives. For example, free-riding on the contributions of others may arise [5]. This report examines algorithmic and economic incentive mechanisms, including reward shaping, reputation systems, and contract-based methods, for aligning individual agents with the collective objective, facilitating payoff-driven consensus on a global policy under which cooperation is individually rational [6].

Contents

5.1 Introduction

In recent years, federated learning (FL) has attracted attention as a decentralized, collaborative approach that enables multiple parties to train a shared model while preserving privacy [4]. Instead of sharing the raw data used for model training directly, the raw data is kept private, and the aggregated, locally computed updates [3] are exchanged.

While FL has been successfully deployed in supervised learning settings, many real-world tasks involve sequential decision-making and interactive environments. Reinforcement learning (RL) provides a natural framework for such tasks, and the integration of federated learning with reinforcement learning has given rise to federated reinforcement learning (FRL) [2]. FRL allows multiple agents, often from different organizations, to learn local policies while contributing to a shared global model. Applications include autonomous driving, robotics, energy systems, finance, healthcare, and other sectors where data is sensitive.

Despite having significant benefits, FRL also faces fundamental economic and strategic challenges. Participating agents often have heterogeneous objectives, computational constraints, and privacy requirements, leading them not to behave cooperatively. This opens the door to misaligned incentives, such as free-riding behavior, strategic withholding of updates, or even manipulation of contributions to maximize private benefits and gain a competitive edge. Such issues have been documented in both classic FL systems and also in FRL systems, where sequential interaction may amplify these effects [5].

Achieving cooperation in FRL systems requires mechanisms that realign individual incentives with the collective learning objective, ensuring that cooperation is individually rational for agents. This need motivates the study of incentive mechanisms that achieve *payoff-driven consensus*, in which adherence to a shared global policy constitutes a symmetric Nash equilibrium, and no agent benefits from unilateral deviation from it. The remainder of this work establishes the foundation for federated reinforcement learning, reviews the current state of the art in incentive mechanisms, and discusses their contributions to stable collaborative learning in heterogeneous, privacy-preserving, multi-agent environments.

5.2 Related Work

Federated Learning was introduced by McMahan et al. as a decentralized approach to training shared models by aggregating locally computed updates [3]. Yang et al. [7] (2019) generalize the FL framework as a privacy-preserving decentralized collaborative-learning technique across organizations. They also define a categorization framework based on the data's distribution characteristics, horizontal FL, vertical FL, and federated transfer learning. In *horizontal federated learning (HFL)*, or sample-based FL, the datasets share the same feature space. Google's first proposed FL approach [3] is an example of an HFL approach. *Vertical federated learning (VFL)*, or feature-based federated learning, is applicable when the organizations' datasets contain the same sample ID space (i.e., user base) but differ in feature space, so the datasets may have overlapping entities but collect different features about the entities. *Federated transfer learning (FTL)* applies when datasets differ in sample and feature space.

Reinforcement learning (RL) is a subset of ML in which agents learn by interacting with their environment [1] to solve a task, often a sequential decision-making problem. By extending the FL paradigm to RL, the idea of federated reinforcement learning (FRL) emerged [2; 8]. Although related to multi-agent reinforcement learning (MARL) [9] in that they are both distributed RL approaches, FRL puts the privacy-preserving aspect of FL at its core.

While these works establish the technical foundations of FL and FRL, they generally assume cooperative agents aligned with a single collective goal. In a realistic multi-agent scenario, such as between different organizations, this may not be the case. Organizations, or rather their agents, may behave strategically or self-interestedly to achieve their respective goals, undermining the learning process towards a collective goal. Furthermore, free-rider, trust, and coordination problems may emerge. Recent research acknowledges these issues and explores solutions. Park et al. (2023) [6] propose a payoff-mechanism design to achieve cooperation in a multi-agent decision-making problem. Meng et al. (2024) [5] model FL in a competitive market, analyzing free-riding behavior and incentive compatibility. Haupt et al. (2024) [10] demonstrate how formal contracts can realign agents to a socially optimal equilibrium in a MARL setting. These emerging approaches show that aligning local payoffs with the collective objective, achieving *payoff-driven consensus*, is central to maintaining cooperation in distributed learning.

5.3 Background

The previous section briefly summarizes the history of federated reinforcement learning, referring to foundational research and providing a sense of the timeline. This section provides the formal definitions and notation for FL and FRL used throughout the remainder of this report, and it takes a deeper look at federated learning and federated reinforcement learning to understand the challenges that arise and to grasp where payoff-driven consensus comes into play. It also establishes how these two frameworks differ from similar or related concepts.

5.3.1 Federated Learning

Federated learning (FL), as proposed by McMahan et al. [4], is similar to distributed computing, where multiple processing nodes are connected via a network so that each node can process a different part of a common task. The main goal of distributed computing is to accelerate task completion, whereas FL focuses on building a global model without privacy leakage. Formally, in FL, there can be N collaborators, each with their own private data $\{D_1, \dots, D_N\}$. A naive approach would be to pool the data and train a collective model M_{SUM} using the pooled dataset $D = \{D_1 \cup D_2 \cup \dots \cup D_N\}$. However, this would leak the private data of the collaborating entities. So, in federated learning, the aim is to train a model M_{FED} without compromising the data of each entity $D_i, \forall i \in N$, while achieving accuracy similar to that of M_{SUM}. Using V_{SUM} to denote the accuracy of model M_{SUM}, and analogously V_{FED} for M_{FED}, the difference in performance between these two models can be written as:

$$|V_{SUM} - V_{FED}| \leq \delta, \delta \in \mathbb{R} \geq 0$$

The federated learning algorithm is said to have δ-accuracy loss.

FL can be further classified into categories by the distribution characteristics of the data $D_i, \forall i \in N$ as introduced by Yang et al. [7]. The dataset D_i for an entity is a matrix, with each row representing a sample, each column representing a feature, and, optionally, a label. Let the feature space be denoted by X, the label data space by Y, and the sample ID space by I. The subsequent sections will cover how I, X, and Y are used to categorize FL approaches.

5.3.1.1 Horizontal Federated Learning

Horizontal FL, also called sample-based FL, applies when the different datasets share the same feature space X, but different sample space I. For example, the Zürcher Kantonalbank and the Basler Kantonalbank, both regional banks in Switzerland, may have different sets of users I in their respective regions, with little or no overlap. Due to their very similar business, their feature and label space are mostly the same as shown in figure 5.1. Therefore, horizontal FL can be summarized as follows:

$$X_i = Xj, Yi = Yj, I_i \neq Ij, \forall D_i, D_j, i \neq j.$$

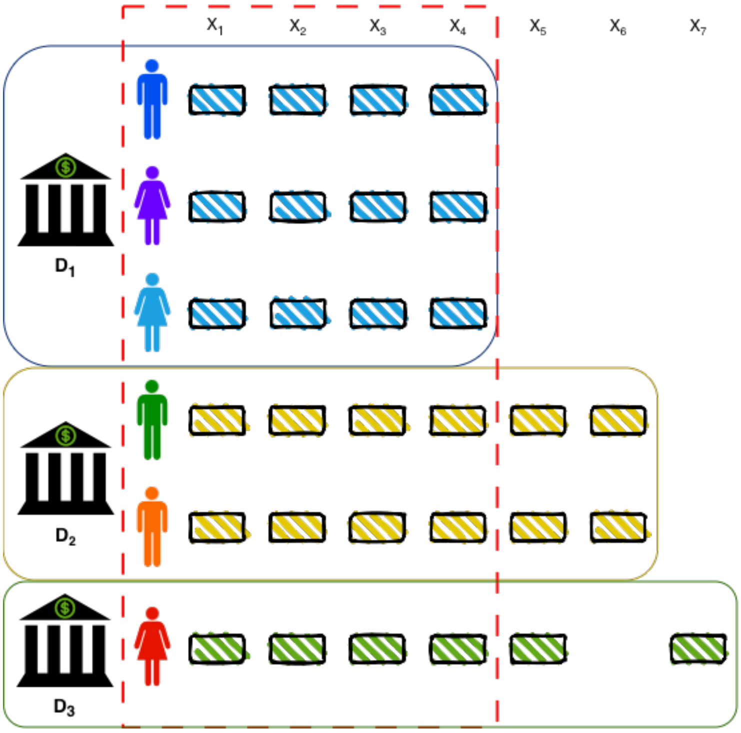


Figure 5.1: A horizontal federated learning scenario

5.3.1.2 Vertical Federated Learning

Vertical FL, also called feature-based FL, applies when the datasets share the same sample space I but differ in the feature X or label Y spaces. For example, the Zürcher Kantonalbank and the Zürcher Verkehrsverbund, a regional bank and the public transport operator of the same region. As many residents of that region (Zurich) will be customers of both companies, the intersection of their sample spaces I will be large. However, these companies will collect very different data about their customers. While the bank will collect customers' revenue and expenditure, the public transport operator will have ticket purchasing information, making their feature space very different. This is shown in figure 5.2. Vertical FL is therefore summarized as:

$$X_i \neq X_j, Y_i \neq Y_j, I_i = I_j, \forall D_i, D_j, i \neq j.$$

5.3.1.3 Federated Transfer Learning

Federated transfer learning applies when datasets differ in both sample and feature space. Consider the Zürcher Kantonalbank again, but this time take the public transport operator of Singapore, SMRT (Singapore Mass Rapid Transit). They have different feature spaces due to their different businesses, unlike the vertical FL scenario. Due to geographical restrictions, the two companies will also have vastly different customer bases, making the intersection of their sample space small as demonstrated in 5.3. Federated transfer learning is summarized as:

$$X_i \neq X_j, Y_i \neq Y_j, I_i \neq I_j, \forall D_i, D_j, i \neq j.$$

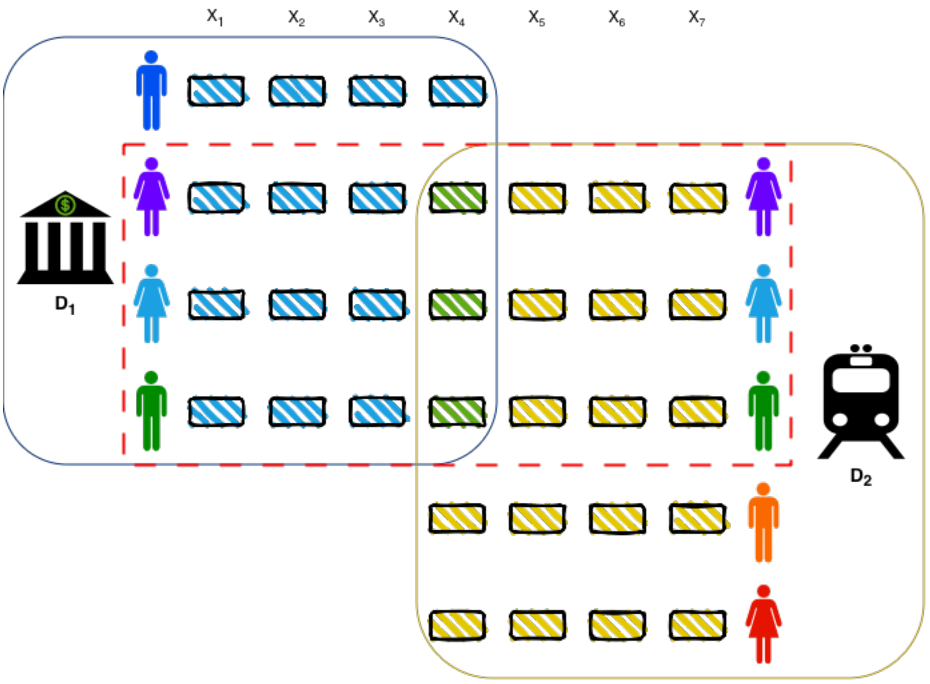


Figure 5.2: A vertical federated learning scenario

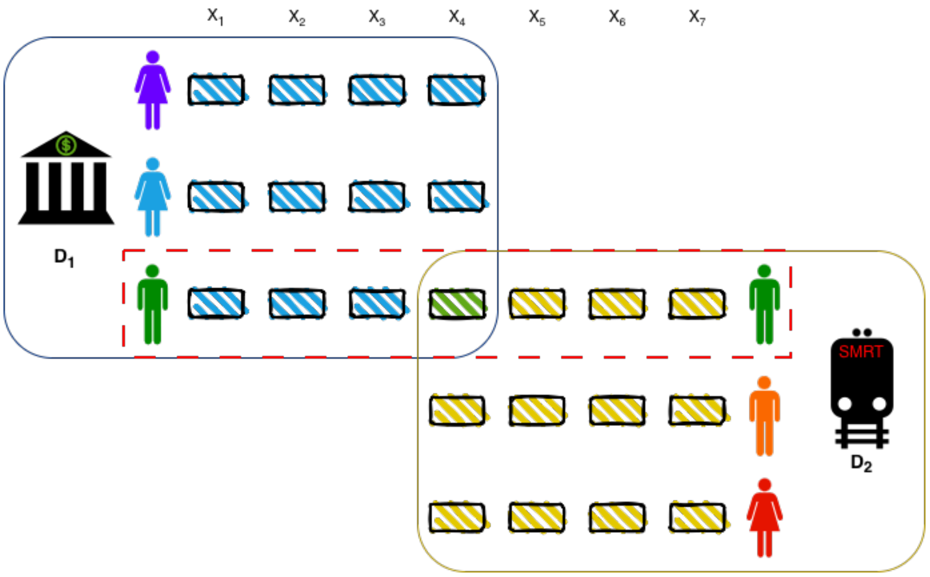


Figure 5.3: A federated transfer learning scenario

5.3.2 Reinforcement Learning

In traditional supervised ML tasks, a pre-prepared dataset of samples and their labels, from which the algorithm shall learn to predict new (previously unseen) samples, is provided. On the other hand, unsupervised ML tasks also receive a dataset of samples that do not contain a label to be predicted. Instead, the ML algorithm should find a structure or pattern in the data, thereby learning to perform, e.g., classification or clustering. In reinforcement learning (RL), no dataset is provided. Instead, the algorithm, commonly called an agent, autonomously interacts with the environment, trying to find the best course of action given the environment's current state. The goal is to obtain an *intelligent agent*, an agent that is autonomous and reacts proactively [11]. Formally, at time step t, the agent takes an action a_t to move from state s_t to state s_{t+1}. For any action taken, the agent gets a reward $R_t = r(s_t, a_t)$. The agent's goal is to learn, using trial and error, what the best course of action is based on the cumulative reward

$$G_t = \sum_{k=t}^{T-1} R_{k+1}.$$

T is a so-called terminating state $s_T \in \mathcal{S}$, where $\mathcal{S}$ is the set of all possible states. Completing a sequence of actions from the initial state s_0 and reaching the terminating state s_T is called an *episode.* Tasks with a well-defined terminating state are called episodic tasks. Games are often episodic, with a clear initial state, the starting configuration, discrete actions between state transitions of the players, and conditions under which the game ends, the terminating states. An example is the board game Go, in which an RL agent, called AlphaGo Zero, famously outperformed professional Go players of the time [12].

In the real world, tasks generally cannot be modeled neatly as discrete. Nonetheless, RL has been successfully deployed to real-world tasks, as evidenced by recent publications, such as drone racing by researchers at the University of Zurich that beat world champions [13] or the numerous publications on autonomous driving [14; 15]. Because time is continuous, the state space is continuous as well. In autonomous driving, the state may include the vehicle's location, current velocity, trajectory, and more, subject to when exactly the sensor data are measured and processed. Additionally, there may not be a clear terminating state. The "episode" ends only when the vehicle is completely turned off. Otherwise, the algorithm runs continuously while the car is still driving. Intermittent stops at a red light or for turning are just that, intermittent. For continuous tasks, the previous definition of cumulative reward is insufficient. Thus, a discounted cumulative reward calculation is introduced with a discount factor $0 \leq \gamma \leq 1$:

$$G_t = \sum_{t=0}^{\infty} \gamma^t R_{t+1}.$$

The discount factor γ balances the importance of early actions with that of future actions. With a discount factor of $\gamma = 0$ only the reward of the first action is counted, while future rewards are all discounted to 0. A discount factor of $\gamma = 1$ will result in all rewards being considered and added to the cumulative sum. When using a discount factor of $\gamma = 1$, the cumulative reward G_t is unbounded, which can lead to numerical instability of the algorithm. Therefore, a discount factor of $\gamma < 1$ is common in practice and in the mathematical analysis of RL algorithms.

The agent learns which action to take in each state to maximize the cumulative reward G_t. These state-based decisions are called the *policy* π. To maximize the reward, the algorithm needs some sense of the future rewards that it can obtain. In RL, there are two major methods to do so, using the state value function $V_\pi(s)$ or the action value function

$Q_\pi(s,a)$. The state value function $V_\pi(s)$ is the expected reward of following policy π after reaching state s, while the state action value function $Q_\pi(s,a)$ is the expected reward obtained when action a is taken in the current state s and then following the policy π. Formally, the functions are defined as follows:

$$V_\pi(s) = \mathbb{E}\left[G_t|s\right], \forall s \in \mathcal{S}$$

$$Q_\pi(s,a) = \mathbb{E}\left[G_t|s,a\right], \forall s \in \mathcal{S}, a \in \mathcal{A}.$$

Similarly to $\mathcal{S}$ being the set of all possible states s, $\mathcal{A}$ is the set of all possible actions a. The performance of the RL algorithm depends on the accuracy of its approximation of the value functions $V_\pi(s)$ or $Q_\pi(s,a)$. By iteratively refining the approximation using value or policy iteration until convergance, the goal of finding an optimal policy π^*, which is the one that maximizes the cumulative reward, can be achieved:

$$\pi^{=} argmax_\pi V_\pi(s), \forall s \in \mathcal{S}$$

5.3.2.1 Multi-Agent Reinforcement Learning

In multi-agent reinforcement learning (MARL), multiple agents $N = \{1, \ldots, n\}$ operate simultaneously in a shared environment. Each agent observes part or all of the global state and may coordinate their actions either through explicit communication over a network or implicitly via the environment. The actions of individual agents may alter the environment, thereby modifying the global state. Formally, a MARL setting is defined by a set of agents N, a global state space $\mathcal{S}$, and a joint action space $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. In fully cooperative settings, the instantaneous reward at time t is given by

$$r_t = r(s_t, a_t) = \sum_{i \in N} r_i(s_t, a_{i,t}).$$

Each agent $i \in N$ follows a policy $\pi_i(a_i \mid s)$, and under decentralized execution the joint policy factorizes as

$$\pi(a \mid s) = \prod_{i \in N} \pi_i(a_i \mid s).$$

Similar to FL, MARL systems are often deployed in a distributed manner. Autonomous driving is a multi-agent domain in which agents, human or artificial, interact with each other within a shared environment. In practice, however, such shared environments typically arise only during deployment. Due to safety, cost, and feasibility constraints, RL agent training for real-world tasks primarily occurs in simulation, with real-world deployment once the agent demonstrates sufficient, robust performance [15].

5.3.3 Federated Reinforcement Learning

Federated Reinforcement Learning (FRL) extends the federated learning paradigm to sequential decision-making environments in which multiple distributed agents interact with their own local environments and collectively aim to improve a shared policy or value function. Each agent $i \in \{1, \ldots, n\}$ observes local states $s_{i,t}$, selects actions $a_{i,t}$ according to a policy $\pi_i(a_{i,t}|s_{i,t})$, and receives local rewards $R_{i,t} = r_{i,t}(s_{i,t}, a_{i,t})$. Agents periodically communicate model updates or policy parameters to a coordinating server, or directly to one another in peer-to-peer systems, for aggregation into a global policy π_G. This global policy is then redistributed for improvement in the next local learning rounds. In contrast to MARL, it explicitly addresses collaborative and distributed training.

While FRL is related to MARL, it differs fundamentally in its assumptions. FRL enables training agents in isolated, potentially heterogeneous environments under privacy and ownership constraints. At the same time, FRL introduces additional challenges beyond FL, including temporal dependencies, exploration–exploitation trade-offs, and non-stationary learning dynamics. These characteristics make it challenging to aggregate locally learned policies or value functions, both strategically and algorithmically [2].

5.4 Payoff-Driven Consensus

Having gained insight into how federated learning evolved and was eventually applied to reinforcement learning, giving rise to federated reinforcement learning, helps in understanding some of the benefits of FRL: the ability to collaboratively train a shared model using data from all collaborators while retaining data privacy in a distributed, potentially decentralized way. Having more data available generally leads to a better overall machine learning model [16]. So FL or FRL enables training of better models than any one collaborator could train on their own, especially in domains where samples are scarce or sensitive. As an example, hospitals, and by extension society, would benefit hugely by having a better model detecting cancer based on imaging (X-ray, CT, MRI, etc.) data of patients. Autonomous driving is a task for which FRL specifically shows promise [14]. With the promise of a better-performing model, one could conclude that agents, or their organization, would naturally cooperate and collaborate in learning such a shared model.

However, participating agents operate under heterogeneous local objectives, such as maximizing task-specific performance based on their private data distributions, minimizing computational or communication costs, or protecting proprietary information. Such heterogeneity has been widely observed in federated learning systems and is known to create strategic incentives for free-riding, withholding updates, or manipulating contributions, leading to self-interested behavior rather than cooperation (e.g., Yang et al., 2019 [7], Meng et al., 2024 [5]). In FRL, these issues are further amplified by sequential decision-making, where local reward structures and environment dynamics may differ substantially between agents. As a result, cooperative behavior cannot be taken for granted. These challenges motivate the need for mechanisms that ensure that cooperation in FRL is not only algorithmically feasible, but also individually rational for all agents involved. To align agents and have individual cooperation emerge rationally, *payoff-driven consensus* is needed. Payoff-driven consensus is achieved once it is not profitable for any agent to unilaterally deviate from the globally shared policy, since doing so would not yield a higher payoff (expected cumulative reward G_t). Deviating becomes irrational.

Formally, let $N = \{1, \dots, n\}$ denote the set of agents. Each agent i has an admissible policy set Π_i, and a joint policy profile is denoted $\pi = (\pi_1, \dots, \pi_n)$. The expected payoff for agent i under profile π is given by a utility function $U_i : \Pi_1 \times \cdots \times \Pi_n \to \mathbb{R}$. A consensus policy is one in which all agents adopt the same policy, i.e., $(\pi_G, \dots, \pi_G)$.

A consensus policy π_G has achieved payoff-driven consensus if it forms a symmetric Nash equilibrium. Concretely, π_G satisfies

$$U_i(\pi_G, \pi_G) \geq U_i(\pi_i, \pi_G) \qquad \forall i \in N,\ \forall \pi_i \in \Pi_i.$$

Here, $U_i(\pi_i, \pi_{-i})$ denotes the payoff to agent i when it uses policy π_i and all other agents use the joint policy π_{-i}. In particular, $U_i(\pi_G, \pi_G)$ if all agents follow the global consensus policy π_G.

How can payoff-driven consensus be achieved? Let's have a look at three different incentive mechanisms, reward shaping, reputation systems, and contract-based methods to understand how they can guide agents to cooperate in the collaborative learning process.

5.4.1 Reward Shaping

Reward shaping is a technique in which the original reward function is augmented to provide an extra reward signal, leading to a new formulation $r' = r + F$, where r is the original reward signal from the environment and F is the shaping reward function. In classic RL settings, domain knowledge and heuristics are encoded into the shaping reward function to guide algorithms to learn faster and better [17]. Reward shaping may introduce policy inconsistency. Potential-based reward shaping (PBRS) [18] proposes to solve this by introducing a potential function $\phi : \mathcal{S} \to \mathbb{R}$, defining the shaping reward function as:

$$F(s, s') = \gamma\phi(s') - \phi(s).$$

In PBRS, the shaping reward function F is defined as the difference in potential of two consecutive states $s, s' \in \mathcal{S}$, where γ is the discount factor. It is the first approach to guarantee policy invariance. Intuitively, the agent receives positive shaping rewards for moving "uphill" in potential and negative rewards for moving "downhill" without altering the global optimum.

While reward shaping is used in classic RL settings to speed up convergence, in FRL it can serve as a tool for incentive design. Hu et al. (2021) propose a learning algorithm called federated reward shaping (FRS), which employs potential-based reward shaping within the FRL framework [19]. It provides a mechanism to encode collaborative structure without sharing raw data and iteratively learns a potential function in the "reward shaping update stage" during federated learning rounds. Eventually, the potential function learns to capture globally beneficial behavior of the federation. In the "policy update stage," the agent now receives a shaping reward for displaying the globally desired behavior, in addition to its local reward. By incorporating the reward signal into the RL algorithm, the agent is incentivized to learn this globally desired behavior, which may lead to cross-pollination of skills. The resulting policy can become more broadly valuable across agents, promoting cooperation. However, organizations may choose to ignore the reward signal. Therefore, reward shaping alone cannot solve free-riding or manipulative behavior.

5.4.2 Reputation Systems

Reputation systems constitute a second class of incentive mechanisms designed to promote cooperation in FRL. In contrast to reward shaping, where incentives are embedded in the agent's reward signal, reputation-based methods use meta-level incentives derived from the agent's behavioral history. The idea is simple: participation in FRL involves repeated interactions, and agents can build trust by demonstrating continued cooperation.

Each agent $i \in \{1, \ldots, n\}$ is assigned a reputation score ρ_i. ρ_i reflecting its trust score, based on criteria such as the reliability of the agent (frequency of participation, dropout rate), contribution quality, compliance with protocol (truthfulness, absence of malicious updates), and other conceivable metrics. The reputation score ρ_i is continuously updated across rounds of federated training.

Agents, respectively the organizations behind them, may be granted priority access to the global model, their contributions to learning may be weighted more heavily, or they may be given other benefits if they possess a high reputation score ρ_i. Agents with low reputation scores, on the other hand, may have their contributions weighted lower, be excluded from the federation entirely (effectively withdrawing access to the shared global model), or be penalized in other ways. From a game-theoretic standpoint, a reputation system transforms the one-shot incentive structure into an infinitely repeated game, in which cooperation can become a subgame-perfect equilibrium. This is similar to the transition from the prisoner's dilemma to its iterated variation. In such a setting, deviation yields immediate short-term gains but causes long-term losses through reputation decay.

This mechanism is especially relevant in FRL, as its data-privacy-preserving properties make it very challenging to detect low-effort or malicious contributions. Al-Maslami et al. (2024) introduced *Reputation-Aware Multi-Agent DRL*, which weights local updates using reputation scores to ensure robustness and fairness in hierarchical FL systems [20]. By weighting the updates, a clear relationship between an agent's reputation and its impact on the globally shared policy emerges. Agents may still try to undermine collaborative learning using malicious or low-effort updates. However, the resulting reduction in agents' reputations can decrease their contributions until they become negligible. In contrast, cooperative agents are incentivized to contribute more, as greater influence over global policy aligns it more closely with their self-interest. Reputation systems also mitigate free-riding behavior, but cannot eliminate it entirely. Because reputation relies on a history of interaction with the federation, an organization may join to obtain an initial shared policy and subsequently continue training independently without contributing further. Although such agents may eventually be excluded, they may already have derived significant benefit by that point.

5.4.3 Contract-Based Mechanisms

Contract-based mechanisms represent the most explicit and enforceable form of incentive alignment in multi-agent federated reinforcement learning. While reward shaping and reputation system influence agent behavior indirectly, contracts directly specify the rules, obligations, payments, and penalties that govern participation in the federation. A contract is a formal agreement between organizations that may specify expected contributions (e.g., number or quality of updates, computational resources), permissible model manipulations (no adversary updates), penalties for deviations or malicious behavior, bonus payments based on contributions, and rights to access the shared global model. Contracts are highly flexible. Bonuses may include lump-sum payments or revenue-sharing agreements, while penalties may consist of exclusion from the federation, reduced/delayed access to the global model, or monetary fines. Contract tiers are also possible, in which an organization can choose different levels of contributions with corresponding rewards, obligations, and penalties.

Such contracts finally enable the elimination of free-riding. For instance, a contract may require a minimum level of contribution per training round, with failure to comply resulting in predefined penalties such as monetary fines, forfeiture of escrowed payments, or loss of access to future model updates. In this setting, free-riding becomes strictly dominated, as the cost of non-cooperation outweighs the benefit of accessing the shared policy. Similarly, contracts can deter malicious updates by specifying penalties for detected manipulation attempts, transforming adversarial behavior into a costly strategy rather than a low-risk attack.

Contracts also allow for flexible incentive structures that account for heterogeneity among participants. Tiered participation schemes can be defined, enabling organizations to choose different contribution levels with corresponding rewards, access rights, and obligations. Bonus mechanisms may include lump-sum payments, revenue-sharing agreements, or preferential access to the global model, while penalties may range from delayed updates to exclusion from the federation. By tying access and rewards directly to verifiable contributions, contract-based mechanisms align cooperative behavior with organizational self-interest.

Recent publications show the viability of such approaches. Park and Barreiro-Gomez (2023) apply mechanism design to multi-agent task allocation, demonstrating how payoff mechanisms can guide agents towards socially optimal equilibria [6]. Haupt et al. (2024) further show how formal contracts can mitigate social dilemmas in MARL by designing transfer payments that make cooperation individually rational [10]. Contracts are particularly valuable in settings where participants are competing organizations, providing a powerful complement to algorithmic and reputation-based incentives by enabling enforceable cooperation.

5.5 Conclusion

This report examined the emerging problem of incentive alignment in a multi-agent federated reinforcement learning (FRL) setting. While FRL promises substantial benefits, most notably improved model performance under strict privacy constraints, cooperation among self-interested agents cannot be taken for granted. Heterogeneous objectives, free-riding behavior, and competitive strategic goals can destabilize the learning process and undermine model quality.

To address these challenges, payoff-driven consensus is needed, in which adherence to the shared global policy is individually rational for each participant. Achieving payoff-driven consensus requires carefully designed incentive mechanisms that realign agents' local payoffs with the collective objective. This report explored the use of reward shaping, reputation systems, and contract-based mechanisms to achieve consensus.

Reputation systems and contract-based mechanisms primarily govern cooperation at the institutional or organizational level, whereas reward shaping provides a direct, algorithmic way of incentivizing agents by modifying the reward function. Using smart contracts, both reputation- and contract-based incentives could also be implemented algorithmically and, in some settings, incorporated into the reward design itself. Designing such mechanisms, particularly in combination, is inherently non-trivial, as it requires careful consideration of strategic behavior, stability, and incentive compatibility within federated reinforcement learning pipelines.

Achieving payoff-driven consensus is therefore not merely a technical challenge, but a fundamentally strategic and economic one. Successful federated reinforcement learning deployments across self-interested organizations must integrate incentive design, algorithms, and governance mechanisms to ensure that cooperation emerges as the rational and best course of action.

Bibliography

[1] Sutton, Richard S., and Andrew Barto. *Reinforcement Learning: An Introduction.* Second edition, The MIT Press, 2020. Adaptive Computation and Machine Learning.

[2] Qi, Jiaju, et al. *Federated Reinforcement Learning: Techniques, Applications, and Open Challenges.* arXiv:2108.11887, arXiv, 24 Oct. 2021, `https://doi.org/10.48550/arXiv.2108.11887`.

[3] McMahan, H. Brendan, et al. *Communication-Efficient Learning of Deep Networks from Decentralized Data.* arXiv:1602.05629, arXiv, 26 Jan. 2023, `https://doi.org/10.48550/arXiv.1602.05629`.

[4] Konečný, Jakub, et al. *Federated Learning: Strategies for Improving Communication Efficiency.* arXiv:1610.05492, arXiv, 30 Oct. 2017, `https://doi.org/10.48550/arXiv.1610.05492`.

[5] Meng, Jiajun, et al. *Federated Learning and Free-Riding in a Competitive Market.* arXiv:2410.12723, arXiv, 16 Oct. 2024, `https://doi.org/10.48550/arXiv.2410.12723`.

[6] Park, Shinkyu, and Julian Barreiro-Gomez. *Payoff Mechanism Design for Coordination in Multi-Agent Task Allocation Games.* arXiv:2306.02278, arXiv, 18 Sept. 2023, `https://doi.org/10.48550/arXiv.2306.02278`.

[7] Yang, Qiang, et al. *Federated Machine Learning: Concept and Applications.* ACM Transactions on Intelligent Systems and Technology, vol. 10, no. 2, Mar. 2019, pp. 1–19, `https://doi.org/10.1145/3298981`.

[8] Zhuo, Hankz Hankui, et al. *Federated Deep Reinforcement Learning.* arXiv:1901.08277, arXiv, 9 Feb. 2020, `https://doi.org/10.48550/arXiv.1901.08277`.

[9] Busoniu, Lucian, et al. *A Comprehensive Survey of Multiagent Reinforcement Learning.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 38, no. 2, Mar. 2008, pp. 156–72, `https://doi.org/10.1109/TSMCC.2007.913919`.

[10] Haupt, Andreas, et al. *Formal Contracts Mitigate Social Dilemmas in Multi-Agent Reinforcement Learning.* Autonomous Agents and Multi-Agent Systems, vol. 38, no. 2, Oct. 2024, p. 51, `https://doi.org/10.1007/s10458-024-09682-5`.

[11] Wooldridge, Michael, and Nicholas R. Jennings. *Intelligent Agents: Theory and Practice.* The Knowledge Engineering Review, vol. 10, no. 2, June 1995, `https://doi.org/10.1017/S0269888900008122`.

[12] Silver, David, et al. *Mastering the Game of Go without Human Knowledge.* Nature, vol. 550, no. 7676, Oct. 2017, pp. 354–59, `https://doi.org/10.1038/nature24270`.

[13] Kaufmann, Elia, et al. *Champion-Level Drone Racing Using Deep Reinforcement Learning.* Nature, vol. 620, no. 7976, Aug. 2023, pp. 982–87, https://doi.org/10.1038/s41586-023-06419-4.

[14] Liang, Xinle, et al. *Federated Transfer Reinforcement Learning for Autonomous Driving.* arXiv:1910.06001, arXiv, 14 Oct. 2019, https://doi.org/10.48550/arXiv.1910.06001.

[15] Kiran, B. Ravi, et al. *Deep Reinforcement Learning for Autonomous Driving: A Survey.* arXiv:2002.00444, arXiv, 23 Jan. 2021. arXiv.org, https://doi.org/10.48550/arXiv.2002.00444.

[16] Hestness, Joel, et al. *Deep Learning Scaling Is Predictable, Empirically.* arXiv:1712.00409, arXiv, 1 Dec. 2017. arXiv.org, https://doi.org/10.48550/arXiv.1712.00409.

[17] Hu, Yujing, et al. *Learning to Utilize Shaping Rewards: A New Approach of Reward Shaping.* arXiv:2011.02669, arXiv, 5 Nov. 2020, https://doi.org/10.48550/arXiv.2011.02669.

[18] Andrew Y Ng, Daishi Harada, and Stuart Russell. *Policy invariance under reward transformations: Theory and application to reward shaping.* In Proceedings of the 16th International Conference on Machine Learning (ICML'99), pages 278–287, 1999.

[19] Hu, Yiqiu, et al. *Reward Shaping Based Federated Reinforcement Learning.* IEEE Access, vol. 9, 2021, pp. 67259–67, https://doi.org/10.1109/ACCESS.2021.3074221.

[20] Al-Maslamani, Noora Mohammed, et al. *Reputation-Aware Multi-Agent DRL for Secure Hierarchical Federated Learning in IoT.* IEEE Open Journal of the Communications Society, vol. 4, 2023, pp. 1274–84, https://doi.org/10.1109/OJCOMS.2023.3280359.

[21] Olfati-Saber, Reza, et al. *Consensus and Cooperation in Networked Multi-Agent Systems.* Proceedings of the IEEE, vol. 95, no. 1, Jan. 2007, pp. 215–33, https://doi.org/10.1109/JPROC.2006.887293.

Chapter 6

The Role of Explainable Recommender Systems in Internet Economics

Marcelina Suszczyk

Recommender systems constitute a key technological foundation of many internet-based platforms and strongly influence economic interactions between users and service providers. While recent advances in machine learning have led to highly accurate recommendation models, these systems often exhibit limited transparency. This lack of interpretability raises concerns regarding user trust, system accountability, and regulatory compliance. Explainable Recommender Systems aim to address these issues by providing understandable explanations for personalized recommendations. This report provides an overview of explainable recommender systems, motivated by the shift from traditional, more transparent approaches to modern black-box models. It introduces core recommender system paradigms, including content-based, collaborative, knowledge-based, and hybrid methods, before discussing how explainability can be incorporated through different explanation types and technical implementations. The report then examines challenges in assessing explanation quality, highlighting trade-offs between technical accuracy and human interpretability as well as methodological limitations in evaluation. Practical applications are analyzed from the perspectives of system designers, end users, and business owners. Finally, the report addresses economic evaluation, emphasizing business-oriented metrics alongside traditional accuracy measures and outlining current gaps and challenges in assessing the economic impact of explainable recommender systems.

Contents

6.1 Background and Motivation

6.1.1 The Rise of Explainability in AI and Recommendation Systems

Recommender systems are widely used to support decision-making in environments characterized by large and complex choice spaces, such as online retail (Amazon, Zalando), media streaming (Netflix, Spotify), and social platforms (Instagram, TikTok). By presenting personalized suggestions, these systems reduce information overload for users while simultaneously enabling platforms to increase engagement and revenue. Burke defines a recommender system as ”any system that produces individualized recommendations as output or has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options” [1].

6.1.2 The Need for Transparency and User Trust

With the increasing adoption of complex machine learning models recommendation processes have become more difficult to interpret. This so-called black-box problem limits the ability of users to understand why certain items are suggested and complicates validation and debugging for system designers. From an economic perspective, a lack of transparency can negatively affect trust, long-term user retention, and compliance with legal requirements [4]. These challenges motivate the integration of explainability into recommender systems.

6.1.3 From Traditional to Explainable Recommenders

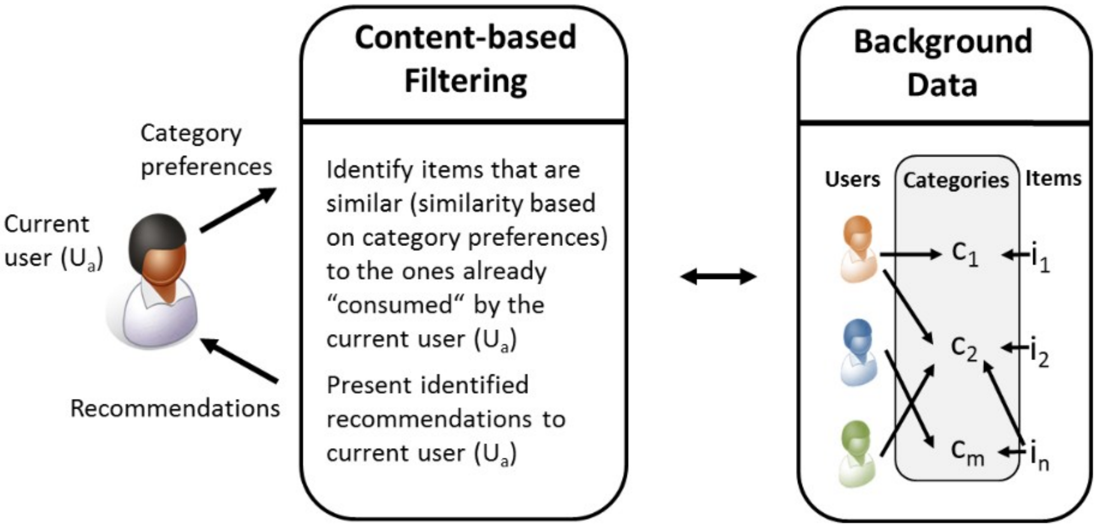
As recommender systems have evolved, a shift can be observed from relatively transparent, heuristic-based approaches toward increasingly complex machine learning models optimized for predictive accuracy. Early content-based and neighborhood-based collaborative filtering techniques allowed recommendations to be linked to observable user or item similarities. In contrast, modern systems often rely on deep learning architectures, whose internal decision processes are difficult to interpret even for experts. This development has intensified concerns related to transparency, accountability, and user trust, particularly in economically relevant online environments where recommendations directly influence decision-making. Explainable recommender systems have emerged as a response to these challenges, seeking to restore interpretability while preserving the performance advantages of advanced models.

6.2 Recommender System Approaches

To meaningfully discuss explainability in recommender systems, it is essential to first understand the classical recommendation approaches on which many modern systems are built. Different recommender paradigms vary substantially in how recommendations are generated, the types of data they rely on, and the degree to which their decision processes are inherently interpretable. These underlying characteristics directly influence which explainability methods are feasible and how explanations can be constructed.

6.2.1 Content-Based Filtering

Content-based recommender systems suggest items that are similar to those a user has previously preferred, based on explicit item features. These systems allow relatively straightforward explanations, as recommendations can be directly linked to observable attributes. This approach is visualized in figure 6.1.



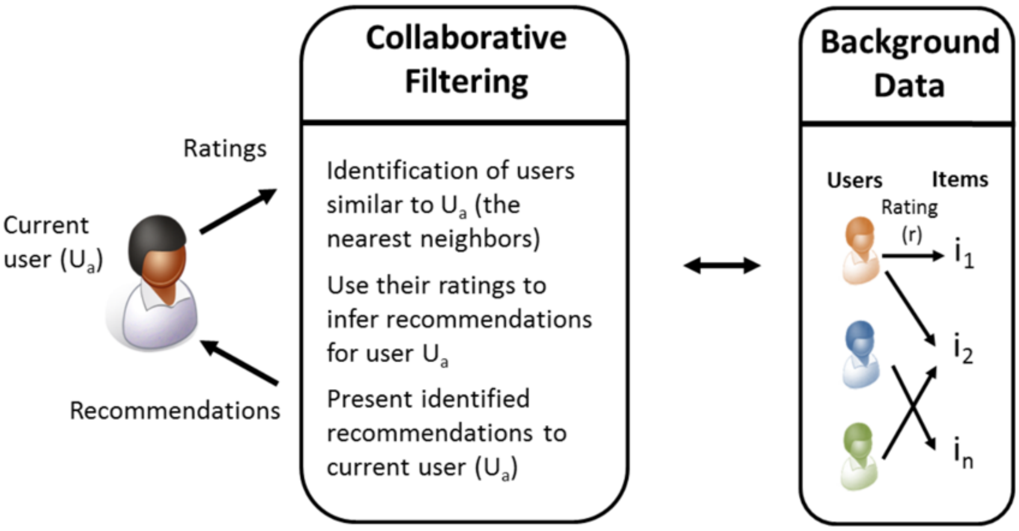
Identify items that are similar (similarity based on category preferences) to the ones already “consumed“ by the current user (U_a)

Present identified recommendations to current user (U_a)

Figure 6.1: Content-based filtering [5]

6.2.2 Collaborative Filtering

Collaborative filtering exploits patterns in collective user behavior to identify similarities between users or items. While highly effective in many domains, the reliance on latent representations often reduces interpretability. This approach is visualized in figure 6.2.



Identification of users similar to U_a (the nearest neighbors)

Use their ratings to infer recommendations for user U_a

Present identified recommendations to current user (U_a)

Figure 6.2: Collaborative filtering [5]

6.2.3 Knowledge-Based Filtering

Knowledge-based approaches rely on explicit domain knowledge and constraints to generate recommendations, as seen in figure 6.3. They are particularly suitable for domains with infrequent interactions or high decision complexity.

6.2.4 Hybrid Methods

Hybrid recommender systems combine multiple techniques to compensate for individual weaknesses and improve robustness. There are different approaches of creating a hybrid recomendation, summarised in figure 6.4.

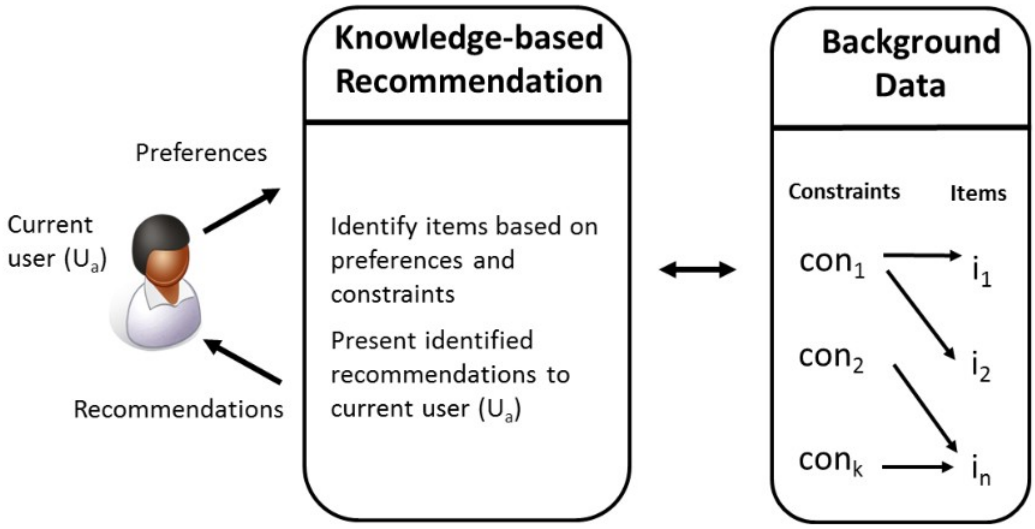
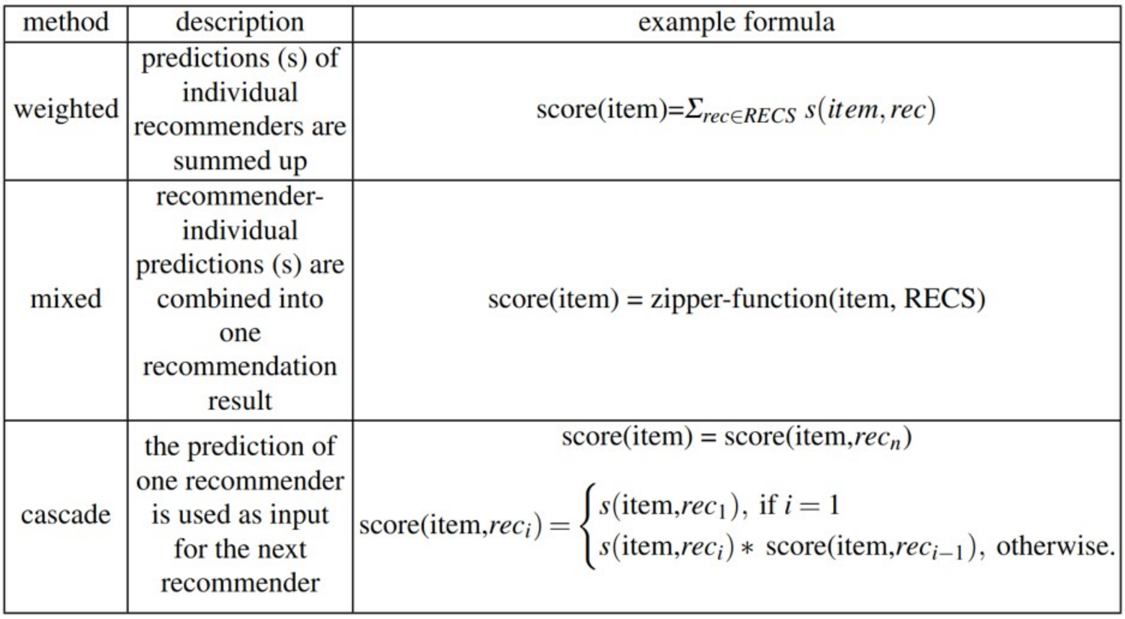

Knowledge-based Recommendation
Identify items based on preferences and constraints
Present identified recommendations to current user (U_a)
Background Data
Constraints
Items
con_1
con_2
con_k
i_1
i_2
i_n
Preferences
Current user (U_a)
Recommendations

Figure 6.3: Knowledge-based filtering [5]



method	description	example formula
weighted	predictions (s) of individual recommenders are summed up	score(item)=$\Sigma_{rec \in RECS}\ s(item, rec)$
mixed	recommender-individual predictions (s) are combined into one recommendation result	score(item) = zipper-function(item, RECS)
cascade	the prediction of one recommender is used as input for the next recommender	score(item) = score(item,rec_n) score(item,rec_i) $= \begin{cases} s(\text{item}, rec_1), & \text{if } i = 1 \\ s(\text{item}, rec_i) * \text{score}(\text{item}, rec_{i-1}), & \text{otherwise.} \end{cases}$

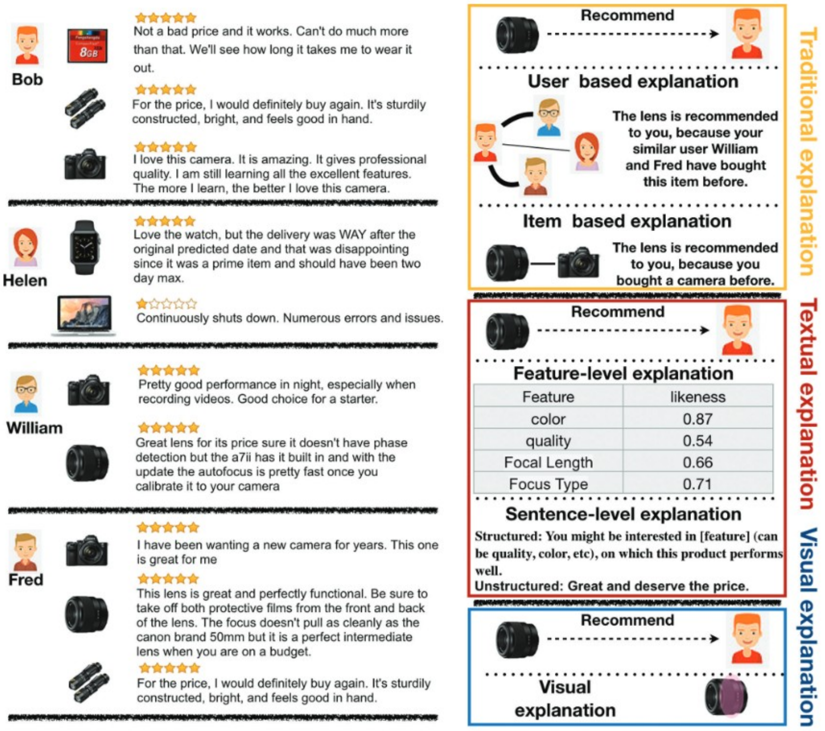
Figure 6.4: Hybrid methods - recommendation approaches based on combining different methods [5]

6.3 Explainable Recommender System Solutions

Explainable recommender systems extend traditional models by providing justifications for recommendation outcomes. Explanations may be generated using internal model information or external post-hoc techniques. Zhang and Chen distinguish explanations based on both content and presentation format, as presented in figure 6.5.

6.3.1 Explanation Types

Common explanation forms include user- or item-based references, textual and sentence-level explanations, feature-level explanations, visual highlights, and social explanations that reference peer behavior.



Not a bad price and it works. Can't do much more than that. We'll see how long it takes me to wear it out.

For the price, I would definitely buy again. It's sturdily constructed, bright, and feels good in hand.

I love this camera. It is amazing. It gives professional quality. I am still learning all the excellent features. The more I learn, the better I love this camera.

Love the watch, but the delivery was WAY after the original predicted date and that was disappointing since it was a prime item and should have been two day max.

Continuously shuts down. Numerous errors and issues.

Pretty good performance in night, especially when recording videos. Good choice for a starter.

Great lens for its price sure it doesn't have phase detection but the a7ii has it built in and with the update the autofocus is pretty fast once you calibrate it to your camera

I have been wanting a new camera for years. This one is great for me

This lens is great and perfectly functional. Be sure to take off both protective films from the front and back of the lens. The focus doesn't pull as cleanly as the canon brand 50mm but it is a perfect intermediate lens when you are on a budget.

For the price, I would definitely buy again. It's sturdily constructed, bright, and feels good in hand.

The lens is recommended to you, because your similar user William and Fred have bought this item before.

The lens is recommended to you, because you bought a camera before.

Feature	likeness
color	0.87
quality	0.54
Focal Length	0.66
Focus Type	0.71

Structured: You might be interested in [feature] (can be quality, color, etc), on which this product performs well.
Unstructured: Great and deserve the price.

Figure 6.5: Different types of recommendation explanations [2]

6.3.2 Technical Implementation

Explainability can be implemented using factorization models with interpretable latent dimensions, deep learning models augmented with attention mechanisms, or model-agnostic post-hoc explanation techniques. Each approach involves trade-offs between transparency, predictive performance, and system complexity.

6.4 Assessing Explanation Quality

Assessing the quality of explanations in recommender systems is a non-trivial task, as explanations must satisfy both technical correctness and human-centered interpretability requirements. Unlike traditional recommender evaluation, which focuses primarily on predictive accuracy, explanation quality involves subjective and contextual dimensions that are difficult to capture with a single metric. As a result, evaluation frameworks typically combine technical, behavioral, and perceptual criteria [2].

6.4.1 Technical Accuracy vs. Human Interpretability

A central tension in explainable recommender systems lies in the trade-off between technical accuracy and human interpretability. From a technical perspective, an explanation is considered accurate if it faithfully reflects the internal decision logic of the underlying recommendation model. Faithfulness ensures that explanations are not misleading and correctly represent the factors that influenced the recommendation outcome. Recent work emphasizes that explanations which are plausible but not faithful may increase user acceptance while simultaneously obscuring true model behavior, thereby undermining transparency and accountability [8]. However, technically accurate explanations are often difficult for non-expert users to understand, particularly when they rely on complex interactions learned by deep models. Human interpretability, in contrast, prioritizes simplicity, coherence, and alignment with users' mental models [2]. Explanations that reference intuitive concepts such as item features, past user behavior, or social signals are generally easier to process, even if they provide only a partial view of the underlying model. This creates a fundamental design challenge: explanations must balance fidelity to the model with cognitive accessibility for users, without sacrificing either dimension entirely.

6.4.2 Challenges and Limitations in Evaluation

Despite growing methodological maturity, evaluating explanation quality remains subject to several challenges. First, explanation effectiveness is highly context-dependent and varies across users, domains, and tasks. An explanation that is beneficial in an e-commerce setting may be ineffective or even distracting in safety-critical or high-stakes domains. Second, user studies are costly, time-consuming, and difficult to reproduce, which limits their applicability for large-scale and longitudinal evaluation. Another limitation concerns the risk of confounding persuasive effects with genuine explanatory value. Explanations that increase user engagement may do so by persuasion rather than by improving user understanding, raising ethical and methodological concerns. Furthermore, there is no universally accepted ground truth for explanation quality, making objective comparisons between explanation techniques inherently difficult. As a result, evaluation frameworks must clearly state their assumptions, objectives, and target user groups. In addition, empirical evaluation is constrained by the limited availability of real-world performance data. Commercial platforms typically treat detailed information about recommender system performance, user responses to explanations, and internal evaluation metrics as proprietary. This lack of transparency restricts independent validation and limits the generalizability of results obtained from academic prototypes or controlled experiments. Consequently, many evaluation studies rely on publicly available datasets or simulated environments, which may only partially reflect the complexity and economic relevance of large-scale commercial systems.

6.5 Practical Applications

Explainable recommender systems have transitioned from a primarily research-oriented topic to a component of operational analytics and interactive user services. Their real-world applications span system design, user experience, and business strategy. By anchoring discussion in established research, this section demonstrates how explainability is leveraged across these domains, and where theoretical insights intersect with practical constraints.

6.5.1 System Designers and Developers

From a technical perspective, explainability serves as a tool for model inspection, refinement, and debugging. Traditional recommender evaluations focused predominantly on accuracy and ranking metrics, yet such system-centric metrics often fail to capture whether a model behaves as intended with respect to diverse stakeholders and contexts [6]. Explainable systems provide developers with insights into latent factors, feature contributions, or interaction patterns, making it easier to detect erroneous behavior or unintended biases. As Jannach et al. note, explanations also illuminate the interaction between algorithmic components and feature engineering choices, clarifying how design decisions affect outcomes in practice [7]. Furthermore, systems that support interactive explanation interfaces enable iterative refinement and cross-disciplinary collaboration, as designers can reason about model behavior with non-technical stakeholders, an important factor in large-scale deployments. In this context, explanations can be operationalized both at the model level (intrinsic explanations, which are built into model architectures) and at the post-hoc level (external methods that generate human-friendly narratives from model outputs). Hybrid evaluation frameworks that combine such model introspection with empirical user study feedback are recommended to ensure that explanations serve both engineering and end-use purposes [1].

6.5.2 End Users

From the end-user perspective, explainable recommender systems primarily affect trust, perceived transparency, and decision confidence [4]. Explanations can help users understand why certain items are recommended, which may reduce feelings of manipulation or loss of control in highly personalized environments. By linking recommendations to past interactions, preferences, or explicit item attributes, explanations support users in forming more informed decisions. Moreover, explainability can improve user engagement by enabling more meaningful interaction with the system, such as refining preferences or correcting incorrect assumptions. At the same time, explanations must be designed carefully to avoid cognitive overload or confusion [2]. Overly complex or technical explanations may reduce usability, while overly simplified explanations risk being perceived as uninformative or misleading. Consequently, explanation design must account for diverse user expectations, levels of expertise, and contextual needs.

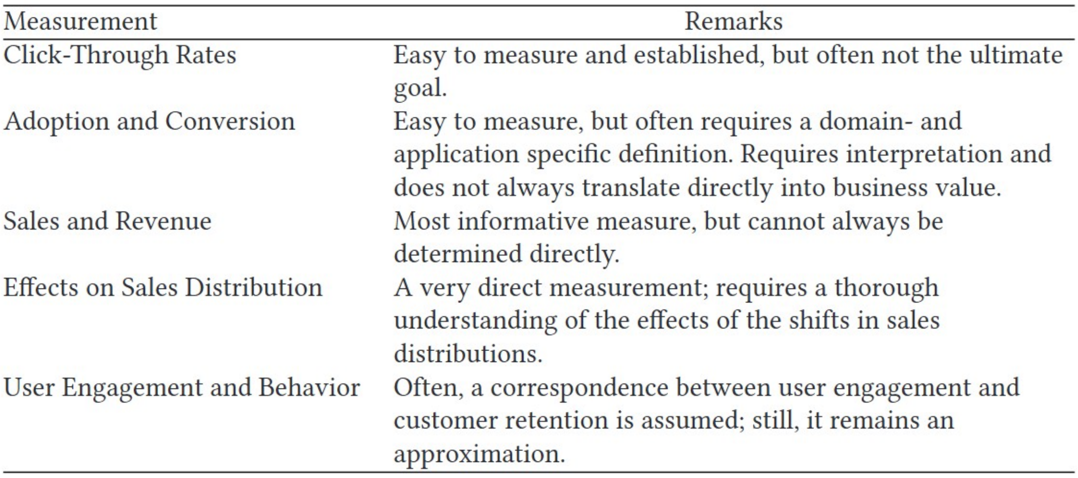
6.5.3 Business Owners

For business owners and platform operators, explainable recommender systems have direct economic implications. Transparent recommendations can increase user trust and satisfaction, which in turn may positively influence engagement, conversion rates, and long-term customer retention. In competitive digital markets, explainability can therefore serve as a differentiating factor that enhances perceived service quality [8]. In addition, explainable recommender systems support compliance with emerging legal and regulatory requirements related to transparency and accountability. Providing understandable explanations can reduce legal risk and strengthen a platform's public legitimacy. However, business owners must also consider potential downsides, such as the disclosure of strategic information about recommendation logic or the possibility that explanations reduce persuasive effectiveness. As a result, commercial deployments often involve a careful balance between transparency, competitive advantage, and economic performance.

6.6 Economic Evaluation

6.6.1 Evaluation Metrics

To understand the impact of recommender systems from the economic perspective, it is crucial to not only consider conventional accuracy-oriented measures, but also the impact on a platform's objectives. Research mostly focuses on measuring the impact on the end user's perspective, assuming that satisfied users will drive the value creation for the provider. However direct impact in terms of the business value is also an important quality that has to be considered. To address this area, De Biasio et al. [8] surveyed the literature incorporating monetary and behavioral metrics such as conversion rates, click-through rates, revenue uplift, profit margins, average order value, and customer lifetime value to quantify how recommendations contribute to economic goals. The measures to assess the economic value of recommenders are listed and explained in figure 2.6.



Measurement	Remarks
Click-Through Rates	Easy to measure and established, but often not the ultimate goal.
Adoption and Conversion	Easy to measure, but often requires a domain- and application specific definition. Requires interpretation and does not always translate directly into business value.
Sales and Revenue	Most informative measure, but cannot always be determined directly.
Effects on Sales Distribution	A very direct measurement; requires a thorough understanding of the effects of the shifts in sales distributions.
User Engagement and Behavior	Often, a correspondence between user engagement and customer retention is assumed; still, it remains an approximation.

Figure 6.6: Measures to assess the economic value of recommenders [9]

6.6.2 Risks, costs and challenges

Although the evaluation of traditional recommender systems constitutes a well-established research area, the systematic assessment of explainable recommender systems remains comparatively underexplored. One possible explanation is that such evaluations are methodologically complex and resource-intensive, with outcomes that vary substantially across application domains and user groups. Existing work primarily focuses on the evaluation of specific explanation techniques ([10]) or on measuring the effects of explanations from a software engineering perspective ([11]). However, a review of the available literature did not reveal publications that explicitly analyze the impact of explainable recommender systems on economic performance measures.

6.7 Summary and Outlook

Advances in recommendation methodologies have led to substantial improvements in both predictive accuracy and the range of application domains for recommender systems. However, these improvements have often come at the expense of transparency and traceability for users and customers. While traditional recommender approaches have been extensively studied and evaluated using well-established quantitative metrics, comparable progress in the systematic evaluation of explainable recommender systems remains limited. At the same time, recent regulatory developments within the European Union emphasize requirements for transparency and, in some cases, a right to explanation for automated decision-making systems. These regulatory pressures are likely to increase the practical relevance of explainable recommender systems and provide additional motivation for further research in this area, as explanations may become a necessary component of deployed recommendation technologies.

Bibliography

[1] Burke, Robin: *Hybrid Recommender Systems: Survey and Experiments*, User Modeling and User-Adapted Interaction, Vol. 12, DOI: 10.1023/A:1021240730564, 2002.

[2] Yongfeng Zhang, Xu Chen: *Explainable Recommendation: A Survey and New Perspectives*, Foundations and Trends in Information Retrieval, Vol. 14, No. 1, pp 1-101. DOI: 10.1561/1500000066, 2020.

[3] Herlocker, Jonathan L., Joseph A. Konstan, John Riedl.: *Explaining collaborative filtering recommendations.* Proceedings of the 2000 ACM conference on Computer supported cooperative work, 2000.

[4] Zhao, Y., Wang, Y., Liu, Y., Cheng, X., Aggarwal, C. C., Derr, T.: *Fairness and Diversity in Recommender Systems: A Survey*, ACM Transactions on Intelligent Systems and Technology, 16(1), 1-28., 2025

[5] Felfernig, A., Jeran, M., Ninaus, G., Reinfrank, F., Reiterer, S., Stettinger, M.: *Basic Approaches in Recommendation Systems*, DOI: 10.1007/978-3-642-45135-5_2, 2014

[6] Shani, G., Gunawardana, A.: *A Survey of Accuracy Evaluation Metrics of Recommendation Tasks*, Journal of Machine Learning Research 10.12, 2009

[7] Gedikli, F., Dietmar J., Mouzhi G.: *How should I explain? A comparison of different explanation types for recommender systems*, International Journal of Human-Computer Studies 72.4: 367-382, 2014

[8] De Biasio, A., Navarin, N. and Jannach, D.: *Economic recommender systems - a systematic review*, Electronic Commerce Research and Applications, 63, p.101352, 2024

[9] Jannach, D. and Jugovac, M.: *Measuring the business value of recommender systems*, ACM Transactions on Management Information Systems (TMIS), 10(4), pp.1-23, 2019

[10] Herlocker, J.L., Konstan, J.A. and Riedl, J.: *Explaining collaborative filtering recommendations*, In Proceedings of the 2000 ACM conference on Computer supported cooperative work (pp. 241-250), 2000

[11] Kulesza, T., Burnett, M., Wong, W.K. and Stumpf, S.: *Principles of explanatory debugging to personalize interactive machine learning*, In Proceedings of the 20th international conference on intelligent user interfaces (pp. 126-137), 2015

Chapter 7

The Cost of Achieving Green AI: Is It Worth It or Not?

Tristan Hein

The rapid expansion of artificial intelligence (AI) has brought unprecedented computational demand and a growing environmental footprint. This paper evaluates whether pursuing Green AI practices is worth the economic and technical cost. We review the Green vs. Red AI distinction, quantify impacts across energy, carbon, and water, and assess policy drivers. Drawing on peer-reviewed literature, standards, and credible corporate or international reports, we analyze economic trade-offs, present case studies, and discuss limitations and rebound effects. We conclude that Green AI is both feasible and necessary: efficiency, clean energy, and transparent reporting can materially reduce impacts without stalling innovation, provided incentives and governance align.

Contents

7.1 Introduction

From 2012 to 2018, the compute used in milestone AI training runs grew by roughly 300,000 times, an exponential surge with a doubling time of only a few months [1]. This dramatic growth in model size and training effort has enabled new state-of-the-art results, but it comes at a cost: escalating energy consumption and environmental footprint. Training a single large deep learning model can consume enough electricity to emit hundreds of kilograms to tons of CO_2. In one notable analysis, a Transformer-based NLP model's training was estimated to produce five times the carbon emissions of an average car's lifetime [22]. Aggregated across the exploding use of AI, these impacts raise serious concerns. Sector-wide assessments warn that, absent interventions, data center electricity demand—of which AI is a major contributor—could more than double by 2030, further straining global sustainability efforts [12]. In short, ever-growing compute appetite has made the question of AI's environmental cost impossible to ignore.

In response, the concept of "Green AI" has emerged as a rallying call to align AI innovation with efficiency and sustainability goals [21]. Researchers have begun advocating that energy use, carbon emissions, and even water consumption become important metrics of progress, alongside traditional metrics like accuracy [7; 21]. This involves encouraging methods that achieve the same results with less compute and transparently reporting resource usage so that improvements in efficiency are rewarded. A cultural shift is underway. For example, Henderson et al. propose standardized reporting of energy and CO_2 for ML experiments [7], and new benchmarks now include power measurements to foster competition on efficiency, not just speed or accuracy. At the same time, industry and policymakers are increasing the pressure to green the AI ecosystem. Major tech companies have announced 24/7 carbon-free energy goals for their data centers, and international standards bodies have introduced metrics such as PUE and WUE to quantify data center sustainability [13]. Regulatory frameworks are following suit. The European Union's Corporate Sustainability Reporting Directive, for example, mandates that organizations disclose their carbon and energy footprints [3], which implicitly includes large AI workloads. Collectively, these developments reflect a broad recognition that the benefits of AI must be balanced against its environmental costs.

Given this backdrop, this paper examines the cost of achieving Green AI in terms of technical, economic, and potential performance trade-offs and asks: "Is it worth it or not?" We approach this question by synthesizing current strategies for reducing AI's environmental impact and evaluating when these efforts yield net positive outcomes. The discussion spans advances in efficient model design and hardware, operational optimizations such as smart scheduling and siting of compute, and the role of transparency, benchmarks, and policy in driving change. Our goal is to provide a neutral, up-to-date overview of the state of Green AI, highlighting the conditions under which pursuing greener practices is truly worthwhile, where the trade-offs or limitations lie, and how recent literature assesses the overall value proposition.

7.2 Theoretical Background and Related Work

7.2.1 Red AI vs. Green AI

Artificial Intelligence (AI) research has often pursued accuracy by scaling computation: larger models, bigger datasets, and longer training runs. This *Red AI* paradigm treats compute as an elastic input, even when marginal accuracy returns diminish and costs rise [21]. From 2012 to 2018, compute used in milestone training runs grew roughly 300,000 times, doubling approximately every 3.4 months, which illustrates how scale became the default path to progress rather than efficiency [1].

Green AI elevates efficiency across energy, cost, and time, along with environmental impact such as carbon emissions and water use, as first-class objectives alongside accuracy [21; 7]. The goal is to design and report results in ways that compare quality together with resource use, encouraging compute-aware innovation and reproducible science [7]. While the benefits are clear, there are trade-offs. Efficiency-focused work may require engineering time, validation to avoid accuracy regressions, and compatibility with specific hardware and tooling. Some techniques, such as aggressive quantization, can introduce small quality losses that must be weighed against lifecycle savings. Importantly, when strong results are achievable with lower compute budgets, market concentration can decrease. This enables small and medium-sized enterprises and academia to participate more easily at the research frontier [2].

7.2.2 Measurement, reporting, and metrics

This subsection clarifies which metrics are relevant and how they connect to broader system-level performance indicators. At the workload level, three quantifiable measures are most important:

- **Energy (kWh):** Electricity consumed by training and inference, ideally with detailed power traces over time [7].
- **Carbon (kgCO_2e):** Emissions estimated by combining energy usage with the grid's location- and time-specific carbon intensity, including any declared assumptions [7].
- **Water (L):** Consumption from both direct cooling at the facility and indirect use from electricity generation. These figures vary strongly by region and generation technology [16].

Facility Key Performance Indicators help standardize reporting context. Power Usage Effectiveness is calculated as total facility power divided by IT power, while Water Usage Effectiveness expresses water use per unit of IT energy [13]. Reporting workload-level metrics alongside KPIs like cooling type, PUE or WUE ranges, and geographic region improves comparability and interpretability across systems [7; 13]. A thorough report should mention hardware type and quantity, numerical precision, training duration, utilization rate, cloud region, and job timing, followed by energy, carbon, and water metrics along with assumptions and methodologies [7; 15].

Benchmarks and shared norms complement numerical metrics. For example, MLPerf's power-aware benchmarks enable standardized comparisons across software and hardware stacks [20]. Open reports such as BLOOM institutionalize environmental metadata and impact accounting, setting a reference point for future disclosures [26].

7.2.3 Compute trends and scaling laws

Exponential growth in available compute has driven many recent breakthroughs in AI [1], though many models are not trained compute-optimally. For a fixed budget of floating point operations, smaller models trained on more data tokens can outperform much larger but under-trained models. The comparison between Chinchilla and GPT-3 exemplifies this, showing similar or better accuracy at far lower cost and energy consumption [8]. Several factors explain the wide variance in footprint between models:

- **Hardware:** Performance-per-watt improves with specialized accelerators, high memory bandwidth, and optimized computation kernels [19].
- **Numerical precision:** Mixed or low-precision training and inference reduce energy demands, often without accuracy loss when properly validated [19].

- **Architecture:** Efficient model designs, including sparse or mixture-of-experts structures, can reduce the amount of compute needed for a given accuracy level [19].
- **Siting and scheduling:** Choosing low-carbon data center regions, aligning workloads with greener time windows, and using facilities with strong PUE and WUE scores all reduce operational impact for the same task [19; 13].

Model families such as MobileNet and EfficientNet demonstrate that careful design choices can deliver competitive accuracy with significantly reduced compute requirements, aligning capability with energy efficiency goals [9; 23]. In summary, compute-optimal training and system-level co-design strategies enable the field to progress more intelligently rather than relying solely on scale.

7.3 Environmental, Economic, and Policy Impacts of Green AI

7.3.1 Economic trade-offs and incentives

Adopting Green AI typically exchanges modest up-front effort, such as engineering time for model compression or serving optimization and hardware selection, for lifecycle savings. Recent disclosures suggest that training accounts for approximately 20 to 40 percent of energy use in Machine Learning (ML), while inference contributes about 60 to 70 percent. As a result, even small efficiency gains during inference can lead to significant reductions in operating expenses at scale [12]. Compute-optimal training improves accuracy per unit of computation, reducing costs while maintaining target quality [8]. Because high compute and capital requirements can be barriers to entry, improving efficiency and disclosing costs and impacts transparently can reduce market concentration and allow small and medium-sized enterprises and academia to participate more competitively [2]. Furthermore, instruments like the Corporate Sustainability Reporting Directive (CSRD) and internal carbon pricing help convert environmental externalities into actionable management metrics, influencing siting, scheduling, and facility-level decisions [3; 7].

7.3.2 Environmental impacts: energy, carbon, and water

This subsection introduces the three core environmental accounts and explains why each is important. Energy use is the primary driver and links directly to both carbon emissions and water withdrawals through the electricity supply and cooling infrastructure. While many strategies reduce all three impacts simultaneously, site-specific factors such as grid composition, local climate, and water stress levels influence outcomes.

Energy (kWh). Training and serving large models require substantial electricity. Projections indicate that data centre demand will continue to grow through 2030, with AI being a major contributor [12]. High energy use increases operational costs and may stress local power systems. Effective strategies to reduce energy use include compute-optimal training, compression methods such as quantization, pruning, or distillation, the use of sparse or mixture-of-experts architectures, and serving techniques that prioritize throughput efficiency [19; 8].

Carbon (kgCO$_2$e). Emissions for the same workload can vary significantly depending on hardware generation, siting decisions, timing, and cooling overhead [19]. Aligning workloads with low-carbon electricity supply, ideally through 24/7 renewable matching, and using

Table 7.1: Environmental accounts and typical methods (illustrative).

Account	**What it measures**	**Primary methods**	**Example indicators**
Energy	Electricity used by training/inference	Compute-optimal training; compression/quantization; sparsity/MoE; serving/throughput tuning; utilization	kWh/job; kWh/-query; perf-per-watt [19; 8]
Carbon	CO_2e from electricity mix and overhead	Low-carbon siting; carbon-aware scheduling; efficient cooling with low PUE; clean PPAs and 24/7 CFE	kgCO_2e/job; kgCO_2e/query; PUE; clean energy share [12; 19]
Water	Direct cooling and upstream generation	Low-WUE cooling; reclaimed water; cool-hour scheduling; basin-aware siting	L/kWh (WUE); potable vs non-potable share [13; 16]

carbon-aware scheduling improves comparability across systems and can meaningfully reduce total emissions [12; 7].

Water (L). Cooling infrastructure and power generation can impose substantial water demands. The specific cooling technology and the water stress level of the local watershed both play major roles in determining impact [16]. Using reclaimed or non-potable water, selecting low-WUE technologies, and scheduling jobs during cooler periods can reduce withdrawals. Reporting water usage alongside energy use and other KPIs enhances transparency and context [13; 16].

Lifecycle and rebound. As operations decarbonize, embodied emissions from hardware manufacturing and the risk of rebound effects become more relevant. When efficiency improvements lower operational costs, the resulting increased usage can offset sustainability gains [19; 12]. Managing total environmental impact requires moving beyond intensity-based metrics like emissions per kilowatt-hour and instead budgeting total energy, carbon, and water use over time. Mitigation strategies include extending hardware lifetimes, modular system upgrades, target utilization levels, and governance that tracks cumulative rather than relative metrics.

7.3.3 Policy, standards, and governance: EU vs US vs Asia

Policy frameworks and standards play an important role in shaping incentives, guiding infrastructure choices, and promoting consistent reporting. Disclosure mandates, common metrics, and permitting regimes influence how organizations design and operate AI systems. At the same time, benchmarks and review processes help shape academic and industry culture.

European Union. The updated Energy Efficiency Directive (Directive (EU) 2023/1791) mandates that large data centres disclose information such as total energy use, PUE, WUE, renewable energy share, and waste-heat recovery to a centralized EU database [4]. The Corporate Sustainability Reporting Directive extends these requirements to a wide range of firms and introduces new obligations for tracking environmental performance

Table 7.2: Regional approaches to Green AI governance (indicative).

Aspect	**European Union**	**United States**	**Asia (e.g., Singapore)**
Policy approach	Binding transparency standards; KPI reporting to EU database [4; 13]	Disclosure-oriented; voluntary benchmarks and corporate pledges [25; 20]	Efficiency gating for new capacity under DC–CFA pilot [11]
Metrics focus	PUE, WUE, renewable share, and heat reuse [4; 13]	Aggregate GHG and energy use; no federal PUE or WUE mandates [25]	PUE targets of 1.3 or lower at full IT load, with siting requirements [10]
Design goals	Environmental awareness via harmonized standards and targets for climate-neutral data centres by 2030 [4]	Technology- and market-driven efficiency with regional variation in policy enforcement [20]	Strict capacity controls linked to efficiency and sustainability certifications [10; 11]

[3]. The ISO/IEC 30134 standard series supports these efforts by defining consistent KPI terminology and measurement guidance [13].

United States. The Securities and Exchange Commission (SEC) has proposed rules that would formalize climate-related disclosures, including greenhouse gas emissions and energy use, for publicly listed companies [25]. Broader adoption of Green AI is supported through voluntary mechanisms, such as the MLPerf power benchmarks, corporate climate pledges, and regional permitting practices rather than binding federal mandates [20].

Asia (selected). In Singapore, new data-centre development has resumed under strict energy-efficiency conditions through the Data Centre Call for Application (DC–CFA) pilot. Facilities must meet stringent requirements such as PUE of 1.3 or lower at full IT load and Green Mark Platinum certification [10; 11]. Other countries in Asia have adopted similar PUE targets or promoted renewable energy integration as part of national digital infrastructure planning. At the global level, the International Telecommunication Union's Y.3001 framework includes environmental sustainability and energy consumption as core design principles for next-generation networks [14].

7.4 Case Studies and Practical Examples

7.4.1 Large corporations

Many hyperscale AI infrastructure operators such as Google, Microsoft, Amazon, and Meta have set ambitious environmental goals and invested in custom technologies to reduce the footprint of their data centers and AI workloads. A prominent example is Google's commitment to power its operations with 100 percent carbon-free energy *at all times* by 2030 [5]. This 24/7 carbon-free energy initiative goes beyond annual offsets and entails matching each hour of electricity use with local renewable generation, thereby eliminating fossil-based energy from its supply [5]. In practice, Google increased the share of hourly carbon-free power for its data centers from 61 percent in 2019 to about 67 percent in 2020, and some sites already run above 90 percent CFE [5]. Achieving 24/7 CFE globally will require a combination of energy innovations, such as advanced storage and load shifting, as well as policy support given the intermittency of renewables [5]. Google's 2020 white paper outlines this strategy and emphasizes the importance of *temporal matching*, meaning energy usage aligns with clean energy availability rather than relying solely on carbon offsets.

Another key tactic among tech giants is carbon-aware scheduling. This involves timing flexible computing tasks, such as non-urgent model training or batch processing, to run when and where low-carbon electricity is plentiful. For instance, Google Cloud introduced a feature that allows customers to choose regions based on real-time CFE percentages, helping shift workloads to cleaner grids. Microsoft has also experimented with delaying workloads during peaks in grid carbon intensity as part of its aim to reduce emissions. These strategies, combined with purchasing agreements for renewables, help large firms reduce the carbon intensity of their AI operations.

Hyperscalers are also improving efficiency through vertical integration of their hardware and facilities. They design custom AI accelerators such as Google's TPUs and Amazon's Inferentia chips that deliver more performance per watt than standard processors. By optimizing chips for machine learning tasks, energy use for training and inference can drop significantly. Google reports two- to five-fold improvements in energy efficiency by using specialized ML hardware instead of general-purpose CPUs [19]. Additionally, companies invest heavily in data center facility efficiency. A common metric is Power Usage Effectiveness (PUE), defined by ISO/IEC 30134-2 as the ratio of total facility power to IT equipment power. State-of-the-art hyperscale data centers achieve PUE values around 1.1 to 1.2, meaning almost 90 percent of energy goes directly to computing. In contrast, older enterprise data centers might have PUE of 2.0 or higher, where only 50 percent of energy is used for computing and the rest is lost to cooling and overhead [13; 19]. Techniques such as advanced cooling, including evaporative cooling or liquid immersion, and AI-driven environmental control have pushed efficiency to these levels. However, some of these methods raise concerns about water usage. For example, evaporative cooling can consume millions of liters of water at a large campus. To address this, companies are pledging "water-positive" operations. Both Microsoft and Google aim to replenish more water than they withdraw by 2030 and are exploring innovations like grey water reuse and switching to air cooling in cooler climates [17]. Recent studies highlight that AI's water footprint is becoming significant. Training a single large model can directly consume hundreds of thousands of liters of water for cooling, with even more consumed indirectly through electricity generation [16]. These developments have prompted calls for spatial and temporal workload shifting to reduce water use. For instance, jobs might be run at night or during cooler seasons to minimize evaporative losses.

From a broader perspective, the impact of these corporate sustainability efforts is evident but not consistent. Industry-wide data indicate that while efficiency per unit of computation has improved, total data center energy use continues to rise due to the growth in demand. In 2024, data centers consumed roughly 415 TWh, representing about 1.5 percent of global electricity use. This figure has been growing at around 12 percent per year, with AI being a key driver. The International Energy Agency projects that global data center electricity consumption could roughly double by 2030 to approximately 940 TWh in its baseline scenario, reaching about 3 percent of worldwide electricity use [12]. Leading firms argue that efficiency measures and the adoption of renewables will keep AI's carbon footprint under control or even net-positive over time [19]. For example, Google noted that machine learning workloads have remained under 15 percent of its total energy use in recent years by applying a suite of best practices such as efficient models, custom hardware, effective cooling, and optimized workload placement [19]. Nonetheless, the tech sector's absolute emissions remain significant and are still increasing, raising questions about the credibility of "carbon-neutral" or "net-zero" claims.

Indeed, third-party audits have exposed gaps between Big Tech's promises and their actual progress. The Corporate Climate Responsibility Monitor 2023, for example, judged many net-zero pledges in the sector to lack transparency and integrity [18]. Common issues

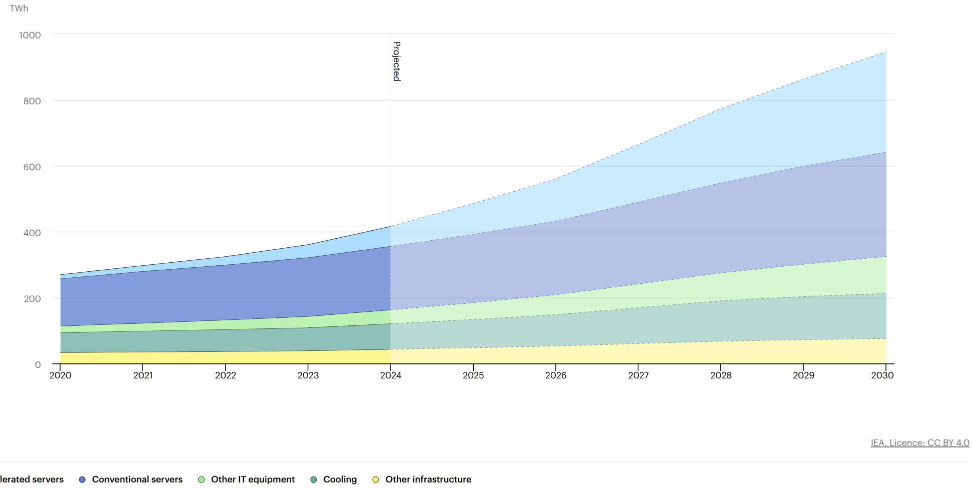


Figure 7.1: Projected global data centre electricity consumption by equipment type (2020–2030), showing steep growth in compute, networking, and cooling loads. Source: IEA (2024) [12].

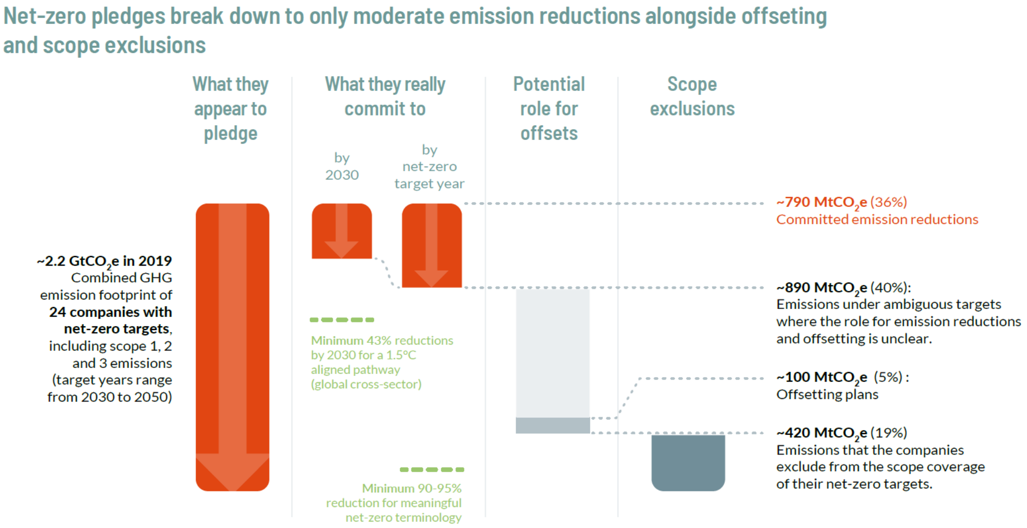


Figure 7.2: Breakdown of corporate net-zero pledges by actual reductions, offsetting, and exclusions. Adapted from NewClimate Institute (2023) [18].

included the use of vague carbon offsets to claim neutrality instead of reducing emissions directly, and target scopes that omit supply chain or product-use emissions [18]. These findings suggest that while hyperscalers are ahead of smaller players in adopting Green AI practices, there remain significant caveats. It is not yet guaranteed that their efficiency gains and renewable investments will outpace the growth in AI demand. Continued public scrutiny and standardized reporting will be critical to ensure that green claims translate into tangible environmental benefits.

7.4.2 SMEs and startups

Compared to trillion-dollar hyperscalers, small and medium-sized enterprises and startups have fewer resources for custom hardware or dedicated facilities. Green AI efforts in this cohort therefore tend to focus on software efficiency and smart use of existing infrastructure. One common strategy is model size reduction through techniques such as network pruning, quantization, and knowledge distillation. By pruning redundant parameters or quantizing with lower precision arithmetic, companies can compress models to use less memory and energy with minimal impact on accuracy. For instance, 8-bit or 4-bit quantized versions of neural networks can run much faster and more efficiently than 32-bit versions, enabling deployment on power-constrained devices. Distillation goes further by training a smaller student model to replicate a large model's behavior, often achieving comparable accuracy with a fraction of the compute requirements [15]. These approaches are increasingly accessible via open-source tools, allowing even small startups to significantly reduce the carbon footprint of their AI workloads. Empirical results have shown that well-optimized compact models, such as MobileNets or EfficientNet, can reduce inference energy consumption by an order of magnitude or more compared to naively large models, especially for vision and mobile applications [9; 23]. By prioritizing efficiency-first architectures, SMEs not only reduce environmental impact but also save on cloud compute costs, which is a critical business advantage.

Another path toward greener AI among smaller players is the use of edge computing and on-device AI to avoid continuous cloud server usage. When inference runs locally on user devices such as phones or IoT sensors, it distributes the energy load and often eliminates the need for power-intensive data center queries. This is only feasible if models are lightweight and the devices are already powered, for example in the case of a smartphone that is regularly charged. By offloading specific tasks to edge hardware, startups can reduce the volume of data processed in central servers, indirectly lowering aggregate cloud energy use. There are trade-offs since not all AI tasks can be efficiently performed on the edge, and the total energy impact requires holistic evaluation. Still, for use cases such as mobile vision, audio processing, or personalization, edge deployment can offer advantages in both latency and sustainability [9].

When SMEs do rely on cloud infrastructure, they increasingly adopt a green cloud approach by selecting providers and configurations that minimize emissions. Many cloud platforms now publish region-specific carbon intensity data. This allows SMEs to schedule jobs in locations with high renewable energy shares or during off-peak hours when clean energy is more available. Supporting tools are emerging to aid this transition. For example, Microsoft Azure's Emissions Impact Dashboard and startups like Clockwork.io offer insights into the CO_2 impact of cloud usage. A culture of emissions transparency is also taking root among tech startups, partly inspired by academic initiatives. Following the call by Lacoste et al. [15] for carbon accountability in machine learning research, several startups have begun publishing training emissions or integrating carbon metrics into documentation. The BigScience project's open report of the 50-ton CO_2 footprint for training the BLOOM language model in 2022 set a precedent. Even nonprofit collaborations can lead in disclosure and establish best practices that startups may follow [26].

Nonetheless, SMEs face constraints in pursuing Green AI. Unlike large firms, they may lack access to efficient hardware or the ability to choose compute locations. A startup might be tied to a specific cloud provider due to customer requirements or platform compatibility. The upfront time needed to rewrite code for efficiency or set up tracking systems may be difficult to justify with limited staff. Very small companies may also lack bargaining power to request clean energy contracts beyond default offerings. Still, many SMEs find that efficiency aligns with their business objectives. Reducing computation lowers cloud bills and often improves responsiveness, both of which benefit users. As a result, Green AI practices are spreading organically in the startup ecosystem, driven by both practical advantages and a growing sense of responsibility.

7.4.3 NGOs, consortia, and academia

Outside of industry, non-governmental organizations, multi-stakeholder consortia, and academic institutions play critical roles in advancing Green AI through guidance, standards, and accountability. At the highest level, global organizations have begun embedding environmental goals into AI ethics frameworks. UNESCO's *Recommendation on the Ethics of Artificial Intelligence*, adopted in November 2021, identified sustainability and environmental responsibility as core principles [24]. Endorsed by 193 member states, the document urges that AI be developed and deployed in ways that avoid environmental degradation and resource depletion. This provides an ethical foundation for policymakers to support greener AI practices.

Meanwhile, industry and academic coalitions are developing more concrete tools. The Green Software Foundation, launched in 2021, introduced the Software Carbon Intensity (SCI) specification [6]. This standard defines a methodology for calculating software-related emissions normalized per unit of output, such as per inference or transaction. An SCI score enables consistent measurement and comparison of emissions across software systems, including AI applications. It also provides developers with an actionable metric to reduce emissions by optimizing code or selecting cleaner execution environments. Although still early in adoption, SCI is backed by major firms and research bodies and may evolve into an international standard. In parallel, ISO/IEC JTC1 is preparing the first global standard for sustainable AI. This framework is expected to incorporate multi-metric footprints, including carbon and water, as core indicators.

NGOs and research consortia also serve as watchdogs and accountability agents. The NewClimate Institute, for example, has published audits exposing gaps in corporate climate strategies [18]. Groups such as Carbon Market Watch and Greenpeace have scrutinized energy sourcing and policy lobbying by tech firms, pushing for transparent and substantive commitments. In academia, open-source tools and reporting norms are gaining traction. Henderson et al. [7] created the Experiment Impact Tracker, which logs energy and emissions from machine learning training. They also advocated for standard reporting of compute usage, energy, and carbon in research papers. This practice is increasingly adopted by journals and conferences. For instance, NeurIPS now includes environmental impact in its broader impacts checklist, and Green AI workshops have expanded significantly since 2019 [15; 21]. Multi-institution collaborations such as Climate Change AI and the BigScience consortium provide infrastructure to share best practices and build tools for efficient machine learning. These collective efforts promote a culture in which sustainability is integral to responsible AI development.

In summary, collaboration between civil society, standard bodies, and academia is essential to push Green AI from aspiration to norm. These actors complement industry efforts by introducing benchmarks, creating public pressure, and making tools and knowledge widely available. Their involvement ensures that sustainability in AI becomes a widely shared objective across the research and technology landscape.

7.5 Evaluation and Discussion

7.5.1 Synthesis: Conditions Under Which Green AI Is Worth the Cost

We have surveyed various strategies for mitigating AI's environmental footprint, spanning from engineering solutions to policy and cultural shifts. But under what conditions do these Green AI measures truly pay off, delivering net-positive outcomes environmentally and societally?

First and foremost, Green AI is most worth the cost when technical feasibility aligns with substantive impact. In other words, when efficiency gains or emissions reductions are large enough to justify the effort invested. Empirical evidence is encouraging: innovations in model design, hardware, and data center operations have demonstrated one to two orders of magnitude improvements in energy efficiency. Google's 4M best practices framework, for example, achieved up to 1000 times emissions reduction by combining efficient models, machines, mechanization, and map optimization [19]. Similarly, Hoffmann et al. showed that a 70B-parameter model trained compute-optimally on four times more data outperformed a 175B model trained conventionally, using far less compute overall [8]. Another favorable condition is the availability of clean energy and sustainable infrastructure. If AI workloads are powered by carbon-free electricity, especially during times of renewable surplus, their climate impact shrinks dramatically. Carbon-aware scheduling and regional job placement can help leverage this flexibility, although policy and market signals must align to support such shifts.

Economic and strategic incentives also strengthen the case for Green AI. Rising energy costs, carbon pricing, and regulatory frameworks like the EU's CSRD [3] make efficiency financially prudent. Strategically, being perceived as a sustainability leader can attract talent and customers. Moreover, Schwartz et al. argue that prioritizing efficiency democratizes AI research by lowering entry barriers [21].

Importantly, Green AI should not come at the expense of innovation. Encouragingly, many Green AI practices such as architecture optimization or compute-efficient training have led to breakthroughs rather than trade-offs. A culture of energy-aware metrics and transparency, as advocated by Henderson et al. [7] and Lacoste et al. [15], allows researchers to optimize both impact and insight.

7.5.2 Limitations and Risks: Rebound, Partial Metrics, Evidence Gaps

Despite progress, Green AI faces several limitations and risks. One is the rebound effect, where increased efficiency enables broader usage and ultimately raises total consumption. The IEA projects that data center energy use could double by 2030, driven in part by expanding AI workloads [12]. Without managing total demand, efficiency alone may worsen the problem.

Second, current practices often focus narrowly on carbon and energy, overlooking other environmental dimensions such as water usage and embodied emissions. AI's water footprint from both cooling and electricity generation can be substantial [16]. If data centers rely on clean power but consume significant volumes of freshwater in arid regions, sustainability claims may be misleading. Similarly, emissions from manufacturing hardware like TPUs, GPUs, and supporting infrastructure are often excluded. Patterson et al. note that most assessments, including theirs, only consider operational emissions [19].

Transparency is another concern. Many net-zero claims rely heavily on offsets or vague reporting. The NewClimate Institute's 2023 audit found widespread deficiencies in corporate climate pledges [18]. Even when firms purchase renewable energy credits, they may still operate on fossil grids during peak hours. Google's push for hourly 24/7 carbon-free energy seeks to close this gap, but most companies are far from achieving that standard.

Lastly, we must ask whether current Green AI efforts are sufficient to address the overall scale of the challenge. Efficiency gains of two or five times are meaningful, but compute demand is growing by a factor of ten or more every few years. Bender et al. caution that ever-larger models could offset environmental gains [2]. Without rethinking what counts as progress in AI, we risk accelerating consumption rather than reducing it. The field must embrace the goal of achieving more with less, as envisioned by Patterson et al. [19].

In short, while Green AI provides viable pathways to sustainability, we must remain vigilant about rebound effects, narrow metrics, and unverifiable claims. A more holistic and transparent approach is necessary to realize transformative change.

7.5.3 Recommendations

Building on our analysis, we propose several concrete actions for researchers, industry practitioners, and policymakers to accelerate the transition to truly sustainable AI:

Adopt standardized metrics and disclosure practices. The AI field should align around a set of clear Green AI indicators and reporting guidelines. Just as accuracy and throughput are standard benchmarks, we need consistent metrics for energy efficiency and emissions. Key indicators such as PUE and WUE from the ISO 30134 standards should be routinely reported by cloud providers and for major training runs [13]. At the model level, researchers can report energy per training task or CO_2 per inference. Journals and conferences should encourage this through submission checklists. On the corporate side, AI-related energy use and emissions should appear in sustainability reports. Regulatory frameworks are beginning to mandate this. The EU's Energy Efficiency Directive 2023 requires large data centers to disclose such metrics publicly [4], while the CSRD mandates broader environmental disclosures [3]. Early adoption of these practices can improve accountability and track progress effectively.

Prioritize efficiency in R&D and system design. Efficiency optimization should be a central objective in research and engineering. This includes investment in model compression, architecture search, and low-complexity algorithms. Compute-optimal training, such as that exemplified by Chinchilla [8], helps avoid unnecessary resource use. Techniques such as sparsity and quantization should become common in deployment [19]. On the hardware side, energy-efficient chips and AI accelerators should be prioritized. Research should also explore models that function with less data or operate at lower numerical precision. As Green AI matures, funding, recognition, and publication incentives should support sustainable innovation.

Embed carbon awareness into ML operations. Modern ML operations should include energy and emissions tracking, as well as carbon-aware scheduling. Open-source tools and commercial dashboards can estimate emissions by location and workload. These should be integrated into experiment tracking systems to allow automatic emissions reporting. Carbon-aware scheduling allows non-urgent jobs to be deferred until renewable availability is higher. Google's map optimization approach demonstrated up to tenfold emissions savings from selecting cleaner locations and times for job execution [19]. In self-managed environments, similar results can be achieved with batteries or intelligent load balancing. MLOps pipelines should also minimize waste by deactivating idle resources, deleting redundant checkpoints, and streamlining storage.

Align incentives through procurement and policy. Customers procuring AI services should include sustainability criteria in vendor evaluations. This might involve prioritizing providers with hourly renewable matching, low PUE, or detailed emissions transparency. These choices create demand for greener infrastructure. Governments can support this transition through targeted funding, regulatory updates, and procurement policies. For example, regulations could require disclosure of AI compute emissions or incentivize use of waste-heat recovery and other efficiencies. Publicly funded AI efforts should themselves meet sustainability standards. International coordination can further establish global baselines and support adoption in emerging economies.

In conclusion, the success of Green AI depends on aligning technical, operational, and policy efforts. With standardized metrics, efficiency-focused development, integrated emissions awareness, and supportive incentives, the AI ecosystem can scale responsibly. The goal is not just to make AI green in principle, but to embed sustainability as a default mode of practice.

7.6 Conclusion

Recent literature suggests that, under the right conditions, Green AI can deliver substantial net benefits by curbing the environmental footprint of AI systems without compromising progress. A key insight is that many efficiency interventions yield outsized gains. Improved model architectures, algorithmic optimizations, and more efficient hardware can reduce energy usage by orders of magnitude while maintaining comparable accuracy [19]. These benefits are especially pronounced for large-scale training and deployment scenarios, where resource demands are highest and even small improvements translate to significant absolute savings. Operational choices such as the location and timing of computation can also substantially reduce emissions; studies report up to 5–10× reductions when workloads are matched to cleaner power grids [19]. When best practices are implemented systematically, the efficiency gains often outweigh upfront costs, making Green AI worthwhile in many practical settings. Projections indicate that scaling efficiency measures and clean energy adoption will be critical for keeping AI's total energy demand manageable [12].

However, the value of Green AI is not guaranteed. Efficiency can lower costs and thus spur wider AI deployment, potentially offsetting gains through rebound effects. The impact depends heavily on how practices are adopted at scale. Moreover, while carbon and energy have received most attention, the water footprint of AI remains underexplored. As Li et al. point out, managing water consumption is critical in regions facing scarcity, yet most assessments exclude this dimension [16]. Some sustainability improvements also come with trade-offs. For example, quantization may slightly reduce model accuracy. These compromises must be carefully evaluated to ensure that sustainability goals are not achieved at the expense of core AI performance.

Overall, the emerging consensus across recent research and policy work is cautiously optimistic. Standard metrics and transparent reporting are seen as foundational. As energy and emissions tracking becomes integrated into ML research and AI operations, it becomes easier to compare systems and identify areas for improvement [7]. Policy instruments such as the CSRD mandate disclosures that bring AI's resource use into public view [3]. Together, technical innovation, operational awareness, and supportive governance form a foundation for scaling AI sustainably. Green AI is no longer a niche concern—it is becoming a defining challenge for the field. With appropriate attention to transparency, efficiency, and system-level coordination, the AI community can continue advancing capabilities while mitigating environmental costs.

Bibliography

[1] Dario Amodei and Danny Hernandez. Ai and compute. OpenAI Analysis, 2018. URL https://openai.com/research/ai-and-compute. Accessed 2025-10-23.

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623, 2021. doi: 10.1145/3442188.3445922. URL https://s10251.pcdn.co/pdf/2021-bender-parrots.pdf. Accessed 2025-10-25.

[3] European Union. Corporate sustainability reporting directive (directive (eu) 2022/2464). Official Journal of the European Union, 2022. URL https://eur-lex.europa.eu/eli/dir/2022/2464/oj. Accessed 2025-10-11.

[4] European Union. Directive (eu) 2023/1791 on energy efficiency (recast of the energy efficiency directive). Official Journal of the European Union, 2023. URL https://eur-lex.europa.eu/eli/dir/2023/1791/oj/eng. Accessed 2025-10-15.

[5] Google. 24/7 carbon-free energy: Methodology and strategy. White Paper / Blog Summary, 2020. URL https://www.gstatic.com/gumdrop/sustainability/24x7-carbon-free-energy-data-centers.pdf. Accessed 2025-10-18.

[6] Green Software Foundation. Software carbon intensity (sci) specification. Technical Specification, 2021. URL https://sci.greensoftware.foundation/. Accessed 2025-10-20.

[7] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43, 2020. URL http://jmlr.org/papers/v21/20-312.html. Accessed 2025-10-24.

[8] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, and et al. Training compute-optimal large language models. *arXiv*, 2022. URL https://arxiv.org/abs/2203.15556. Accessed 2025-10-16.

[9] Andrew G. Howard, Menglong Zhu, Bo Chen, and et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. In *arXiv*, 2017. URL https://arxiv.org/abs/1704.04861. Accessed 2025-10-14.

[10] Infocomm Media Development Authority (IMDA) and Singapore Economic Development Board (EDB). Annex a: Summary of pilot dc-cfa key parameters & criteria. Official PDF, 2022. URL https://www.imda.gov.sg/-/media/imda/files/news-and-events/media-room/media-releases/2022/07/annex-a---summary-of-pilot-dc-cfa-key-parameters-and-criteria.pdf. Specifies Green Mark Platinum and PUE ≤ 1.3 at 100% IT load; application window closed 21 Nov 2022; accessed 30 Oct 2025.

[11] Infocomm Media Development Authority (IMDA) and Singapore Economic Development Board (EDB). Launch of pilot data centre – call for application (dc-cfa) to support sustainable growth of dcs. Press release, July 2022. URL `https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2022/launch-of-pilot-data-centre---call-for-application-to-support-sustainable-growth-of-dcs`. Last updated 26 Jul 2024; accessed 30 Oct 2025.

[12] International Energy Agency. Data centres and data transmission networks: Analysis and forecasts 2024. IEA Report, 2024. URL `https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks`. Accessed 2025-10-17.

[13] ISO/IEC. Iso/iec 30134 series: Data centre key performance indicators (pue, wue, etc.). International Standard, 2016. URL `https://www.iso.org/standard/63451.html`. Series incl. 30134-2 (PUE), 30134-9 (WUE).

[14] ITU-T. Recommendation itu-t y.3001: Future networks: Objectives and design goals. International Telecommunication Union, 2011. URL `https://www.itu.int/rec/T-REC-Y.3001`. Accessed 2025-10-26.

[15] Alexandre Lacoste, Alexandra Sasha Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. arXiv:1910.09700, 2019. URL `https://arxiv.org/abs/1910.09700`. Accessed 2025-10-10.

[16] Pengfei Li, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. Making ai less 'thirsty'. *Communications of the ACM*, 68(7):54–61, 2025. doi: 10.1145/3724499. URL `https://arxiv.org/abs/2304.03271`. Accessed 2025-10-19.

[17] Microsoft. Microsoft will be carbon negative by 2030. Official Announcement / Sustainability Report, 2020. URL `https://blogs.microsoft.com/blog/2020/01/16/microsoft-will-be-carbon-negative-by-2030/`. Accessed 2025-10-23.

[18] NewClimate Institute. Corporate climate responsibility monitor 2023. Independent Assessment Report, 2023. URL `https://newclimate.org/resources/corporate-climate-responsibility-monitor-2023`. Accessed 2025-10-21.

[19] David Patterson, Joseph Gonzalez, Quoc V. Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. arXiv:2104.10350, 2021. URL `https://arxiv.org/abs/2104.10350`. Accessed 2025-10-13.

[20] Vijay Janapa Reddi, Christine Cheng, Sergey Kanev, and et al. Mlperf inference benchmark. In *ISCA 2020*, pages 446–459, 2020. doi: 10.1145/3396474.3397011. URL `https://parsa.epfl.ch/course-info/cs723/papers/mlperf_inference.pdf`. Accessed 2025-10-28.

[21] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020. doi: 10.1145/3381831. URL `https://arxiv.org/abs/1907.10597`. Accessed 2025-11-01.

[22] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv*, 2019. URL `https://arxiv.org/abs/1906.02243`. Accessed 2025-10-11.

[23] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6105–6114, 2019. URL `http://proceedings.mlr.press/v97/tan19a.html`. Accessed 2025-10-15.

[24] UNESCO. Recommendation on the ethics of artificial intelligence. Adopted by UNESCO General Conference, 2021. URL `https://unesdoc.unesco.org/ark:/48223/pf0000381137`. Accessed 2025-10-22.

[25] U.S. Securities and Exchange Commission. The enhancement and standardization of climate-related disclosures for investors (proposed rule). SEC Fact Sheet/Release No. 33-11042, 2022. URL `https://www.sec.gov/rules/proposed/2022/33-11042.pdf`. Accessed 2025-10-12.

[26] BigScience Workshop. Bloom: A 176b parameter open-access multilingual language model (training report and environmental accounting). Project Report, 2023. URL `https://bigscience.huggingface.co/blog/bloom`. Accessed 2025-10-12.