Comprehensive Analysis of Transparency and Accessibility of ChatGPT, DeepSeek, and other SoTA Large Language Models

Ranjan Sapkota[a,*], Shaina Raza[b], Manoj Karkee[a,*]

[a] *Cornell University, Biological & Environmental Engineering, Ithaca, 14850, NY, USA*
[b] *Vector Institute, Ontario, Tronto, W1140-108, ON, CANADA*

Abstract

Despite increasing advocacy for open-source artificial intelligence (AI), a critical gap persists in the rigorous evaluation of transparency and accessibility in state-of-the-art (SoTA) large language models (LLMs). While initiatives such as the Open Source Initiative (OSI) provide baseline definitions, they fall short of addressing the multi-dimensional nature of openness in modern AI systems. The rise of "open-washing" where models claim openness while withholding essential details, limits reproducibility, accountability, and downstream usability. In this study, we present a comprehensive, normalized evaluation of 121 leading LLMs released between 2019 and 2025, including widely used models such as ChatGPT-4, DeepSeek-R1, LLaMA 2, and Gemini 2.5 Pro. We introduce two quantitative metrics: the Composite Transparency Score (CTS), which measures transparency across seven normalized dimensions (code availability, model weights, training data, documentation, licensing, carbon disclosures, and benchmark reproducibility); and the Training Data Disclosure Index (TDDI), which captures the depth and specificity of training data reporting. Our analysis reveals that many of the most powerful models released in 2025 exhibit declining transparency, often omitting disclosures on data provenance and environmental impact despite achieving state-of-the-art benchmark results. In contrast, models such as BLOOM, DeepSeek-R1, and Qwen 3 demonstrate a commitment to openness across both CTS and TDDI dimensions. To support community adoption, we propose a badge-based labeling framework and recommend alignment with internationally recognized responsible AI frameworks, including the EU Ethics Guidelines for Trustworthy AI, OECD AI Principles, and IEEE Ethically Aligned Design. This work offers the first large-scale, normalized benchmarking of LLM transparency and establishes a reproducible, scalable foundation for evaluating openness in generative AI systems. All transparency scores, source links, and extended documentation are available at our GitHub repository: https://github.com/ranzosap/LLM-Transparency.

Keywords: AI Open, ChatGPT, DeepSeek, Large Language Models (LLMs), LLM Transparency, LLM Accessibility, Open Weights, Open Source, Large Reasoning Models (LRMs), DeepSeekR1, OpenAI

Contents

1 Introduction **1**
1.1 Aim and Objectives . . . 3

2 Methodology **3**
2.1 Research Design . . . 4
2.2 Criteria for Openness and Transparency . . . 4
2.2.1 Open Source and Licensing Types . . . 4
2.2.2 Open Source and Transparency . . . 6
2.3 Synthesis of Literature . . . 6
2.3.1 Search Strategy . . . 6
2.4 Evaluation Framework and Application . . . 7
2.5 Research Questions . . . 8
2.6 Transparency Metrics and Trends . . . 8
2.7 Temporal Evolution of Openness (2019–2025) . . . 8
2.8 Standardized Quantitative Transparency Metrics . . . 8

3 Results **9**
3.1 Overall Findings on Openness and Transparency . . . 9
3.2 Model-Specific Evaluations . . . 13
3.2.1 ChatGPT . . . 13
3.2.2 DeepSeek . . . 14
3.2.3 Miscellaneous Proprietary Models . . . 14
3.3 Cross-Cutting Patterns and Metrics . . . 16
3.4 Synthesis of Findings . . . 18

4 Discussion **18**
4.1 Trends and Implications in AI Development . . . 18
4.1.1 Geopolitical and Technological Trends . . . 18
4.1.2 Economic Impact and Market Trends . . . 18
4.1.3 Implications for Open Weights and Open Source AI Models . . . 18
4.2 Model Comparison and Transparency Implications . . . 18
4.3 Discussion on Research Questions . . . 20
4.4 Sustainability and Ethical Responsibility in AI Development . . . 23
4.5 Synthesis and Future Directions . . . 24

5 Conclusion **27**

1. Introduction

Natural Language Processing (NLP) and large language models (LLMs), including multimodal LLMs such as GPT-4o, DeepSeek-V2, and Gemini 1.5, have witnessed transformative advancements and significant growth in recent years, as illustrated by the surging global interest from both research and industry, as depicted in Figure 1a

These technologies have become integral to systems and solutions across a diverse array of sectors, including healthcare Cascella et al. (2023), finance Li et al. (2023), education Neumann et al. (2024), and entertainmentQiu (2024). Their remarkable capabilities in language understanding and generation

*Manoj Karkee
Email address: rs2672@cornell.edu (Manoj Karkee)

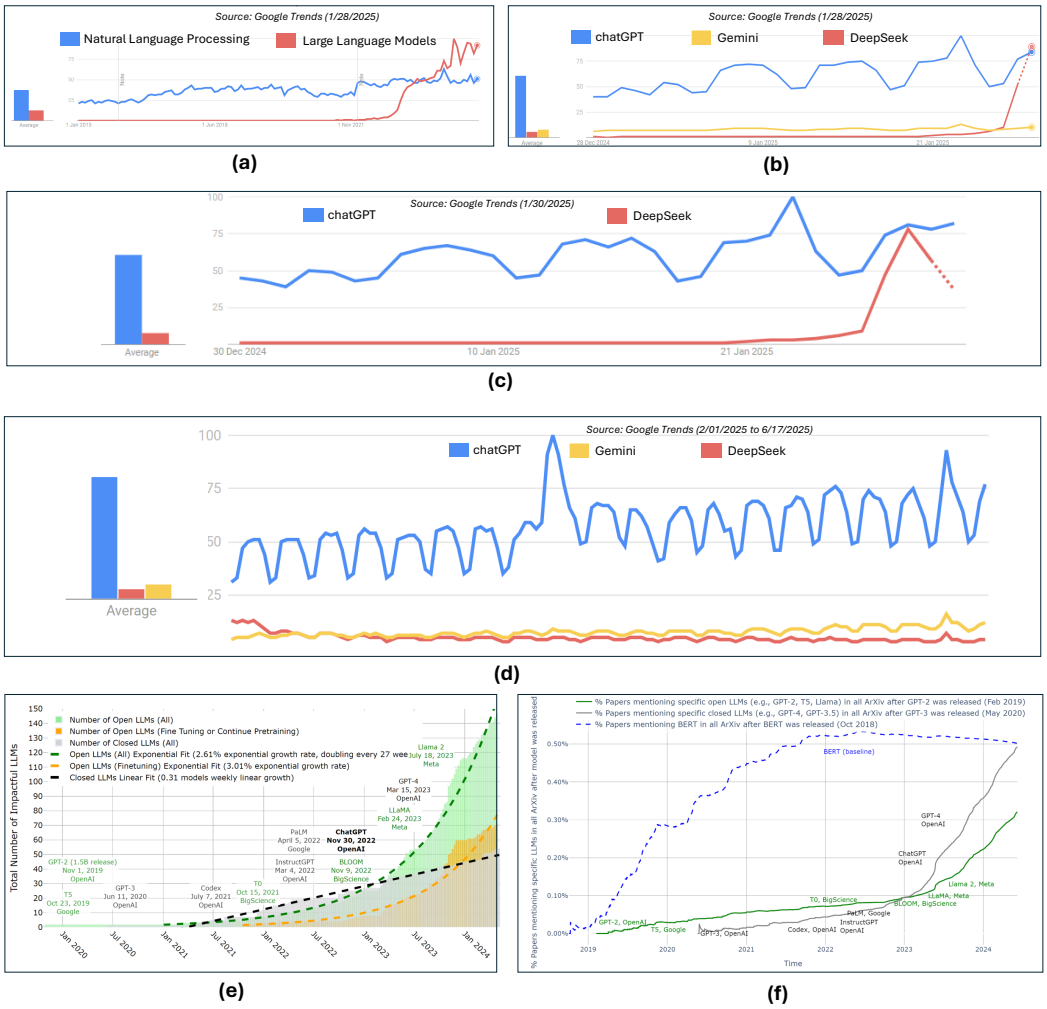

Figure 1: **Analysis of NLP/LLM Interest**
(a) Google Trends showing increasing interest in NLP and LLMs from 2015 to 2025; (b) Global interest for ChatGPT, Gemini, and DeepSeek on January 28 2025, highlighting DeepSeek's rapid rise; (c) ChatGPT and Deepseek global interest on January 30, 2025; (d) Global search interest during the first six months of 2025 shows a decline in DeepSeek's popularity, while Google Gemini ranks higher than DeepSeek. ChatGPT continues to maintain the highest global interest. (e) Growth rates of open sourced and close closed souced LLMs Xu et al. (2025) ; (f) Percentage of arXiv papers mentioning open LLMs or closed LLMs from 2019 onwards, with BERT as a baseline Xu et al. (2025)

have not only revolutionized multiple industries but have also spurred a new wave of innovation and application development Weldon et al. (2024); Grant et al. (2025). Amidst this rapid expansion, the term "open-source" frequently surfaces within discussions about LLMs Kukreja et al. (2024). However, this descriptor is often misapplied or misunderstood. In many instances, developers may release only the model weights, that is, the trained parameters, without sharing the comprehensive suite of model assets such as model card, training data, code, environmental sustainability factors (e.g., CO_2 emissions), or detailed development processes. This gap is also widely discussed in the literature Ramlochan (2024) and in numerous tech blogs, including Walker II (2024), to name a few.

Although proprietary LLMs like OpenAI GPT-series (4/4o) Achiam et al. (2023) exhibit strong performance, their closed-source nature limits access to API-based interactions. In contrast, open-weight models like Meta LLama-series Touvron et al. (2023) and others, provide downloadable model weights under non-proprietary licenses, enabling specialized deployments and cost-effective fine-tuning. For instance, Princeton's Llemma leverages Code Llama for advanced mathematical modeling Azerbayev et al. (2023), showcases the flexibility and cost benefits of open-weight models.

Despite this growing interest, the term "open-source" has frequently been used interchangeably with "open weights", leading to confusion in discussions about model accessibility. Many models labeled as open-source provide access only to their trained weights while withholding essential components such as training data, fine-tuning methodologies, and full implementation details. This distinction is critical, as true open-source

models enable not just inference but also full transparency and reproducibility in AI research. A recent case highlighting the confusion between open-source and open-weight models is DeepSeek-R1 Guo et al. (2025). Initially surpassing ChatGPT in search interest (Figure 1b), its popularity rapidly declined (Figure 1and 1d), reflecting unmet expectations. While DeepSeek-R1 provides weights and partial code under the MIT license [1], it lacks full open-source transparency, including access to training data and methodologies. This partial openness, common to models like ChatGPT and Google's Gemini, allows broader usage compared to fully closed models, but restricts deeper architectural modifications, evaluation of biases, and further enhancement of the training processes and datasets.

The distinction between "open" and "closed" LLMs is evident in their adoption trends. Closed models like GPT-3(ChatGPT) followed a linear growth pattern (gray bars, Figure 1e), while open LLMs surged after Meta's Llama release, driving exponential adoption (green and orange bars, Figure 1e). Figure 1f further illustrates how open source models increasingly attract scientific focus compared to the same with proprietary models such as GPT-4.

This ambiguity in AI terminologies necessitates clearer distinctions between open-source and open-weight models. True open-source AI requires full transparency, including training data and development processes, fostering reproducibility and ethical AI advancements. Defining and broadly adopting clear standards would enhance transparency, set realistic expectations, and promote responsible AI development.

1.1. Aim and Objectives

The primary objective of this study is to critically analyze the transparency and accessibility practices of SoTA LLMs, with particular attention to the models categorized as "open-weight." Using DeepSeek-R1 and ChatGPT-4o as illustrative case studies, this work seeks to clarify the dynamic, but often conflated, distinctions between open-weight and fully open-source models. Existing definitions of model openness are frequently insufficient to meaningfully evaluate transparency. While open-weight models provide access to pre-trained parameters, they often omit key elements such as training datasets, source code, fine-tuning methodologies, and environmental metrics, rendering their openness partial at best.

To address this gap, we propose a novel conceptual contribution: a taxonomy of "open-washing." Analogous to "green-washing" in the sustainability disc;ines, open-washing refers to the practice of labeling models as open or transparent without providing the necessary components for reproducibility, accountability, or community-based development. This taxonomy categorizes misleading transparency claims into specific types based on which essential components (e.g., code, data, license, benchmarks) are omitted or selectively disclosed.

In doing so, this study aims to:

- Minimize the terminological ambiguities and inconsistencies surrounding the term "open-source" in the LLM ecosystem, distinguishing truly open-source models from superficially open ones that fit the open-weight designation.
- Investigate the scientific, ethical, and practical implications of partial transparency, particularly its effects on reproducibility, bias mitigation, community engagement, and reliability on downstream deployment.
- Propose a refined evaluative framework, including the open-washing taxonomy to serve as a more rigorous benchmark for assessing the degree of transparency in current and future LLMs.

With this study, we aim to contribute to an informed advancement of responsible AI, where both technological innovation and collaborative transparency are harmonized. The following sections describe the current landscape of LLMs, the tensions between proprietary and open-weight models, and the broader impacts of these approaches on the AI research community.

Table 1: List of Abbreviations

Acronym	Definition
AI	Artificial Intelligence
ATBF	Automated Transparency Benchmarking Framework
AZR	Absolute Zero Reasoner
BERT	Bidirectional Encoder Representations from Transformers
CTS	Composite Transparency Score
EU	European Union
GPT	Generative Pre-trained Transformer
HELM	Holistic Evaluation of Language Models
IEEE	Institute of Electrical and Electronics Engineers
LLM	Large Language Model
LRM	Large Reasoning Model
MMLU	Massive Multitask Language Understanding
ML	Machine Learning
MoE	Mixture of Experts
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
OECD	Organisation for Economic Co-operation and Development
OSI	Open Source Initiative
RAIL	Responsible AI License
RL	Reinforcement Learning
RLHF	Reinforcement Learning from Human Feedback
RQ	Research Question
SoTA	State of the Art
T5	Text-To-Text Transfer Transformer
TDDI	Training Data Disclosure Index

2. Methodology

This study systematically examines the concepts of openness and transparency in the development and dissemination of LLMs. A multi-stage approach is used in this study, beginning with a thorough examination of foundational concepts and progressing through detailed analyses of licensing types and transparency definitions as they relate to AI systems.

[1] `https://opensource.org/license/mit`

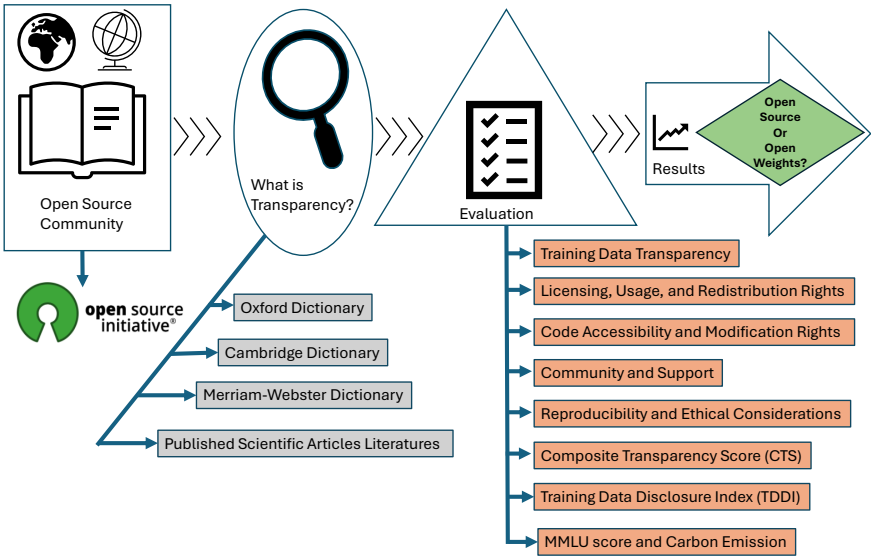

Figure 2: Overview of the methodologies used in evaluating ChatGPT, DeepSeek, and SoTA multimodal LLMs

2.1. Research Design

This study adopts a multi-stage research design to evaluate the openness and transparency of SoTA LLMs. As illustrated in Figure 2, the approach integrates established open-source criteria, foundational linguistic definitions of "transparency", and an extensive review of scholarly literature in the recent works in AI. A mind map (Figure 3) further delineates the core analytical branches, structured into three research questions (RQs) guiding the study. Below, each methodological component is described in detail.

2.2. Criteria for Openness and Transparency

In this section, we discuss the criteria for openness and transparency often associated with the recent LLMs.

Open-Source LLMs: An open-source LLM provides unrestricted access to its entire codebase, including the model architecture, training data, and the training processes Ramlochan (2024). Beyond the code and weights, a truly open-source model also discloses key factors such as performance benchmarks, bias mitigation strategies, computational efficiency, and sustainability metrics (e.g., carbon dioxide emissions, energy consumption in terms of electricity, power). For example, Meta's LLamA aligns with the open-source paradigm by offering detailed insights into its design and implementation through its model cards and official report.

The primary goal with open-source models is to ensure complete transparency and flexibility. This openness enables comprehensive understanding, recreation, and reproducibility , even though some usage restrictions may still appear. Such transparency allows the research community to scrutinize, improve, and tailor models for diverse applications. Developing and maintaining such models, however, demands substantial effort and resources, making the open-source approach both a technical and logistical challenge. For example, early models like GPT-1 and GPT-2 were released as open-source models, providing access to their training data, code and model weights. With subsequent versions like GPT-3 and later, OpenAI shifted to a closed-source approach, restricting access to the model architecture, code, and weights. This trend continued with GPT-4 and sunsequent model family, which also remains proprietary.

Open-Weight LLMs: Open-weight LLMs make their pretrained model weights pro (2023), the parameters learned during the pre-training process, publicly available, while the underlying code, training data, or training methodologies may remain proprietary. Open-weight models, while more accessible and easier to deploy than closed-source models, do not provide the same level of insight into the model's inner workings as fully open-source models would. Meta's LLama series is a prime example of an open-weight LLM. Researchers can download the pre-trained weights to fine-tune and deploy the model for various applications. However, while LLama model weights are available, the full training pipeline, including the code and data, remains proprietary. This enables a balance between accessibility and intellectual property protection.

2.2.1. Open Source and Licensing Types

OSI stands for the Open Source Initiative Open Source Initiative (2025). It is a non-profit organization dedicated to promoting and protecting open source software. OSI is best known for its Open Source Definition (OSD), which outlines the criteria that a software license must meet to be considered "open source". These criteria include free redistribution, source code availability, the ability to create derivative works, and non-discrimination, among others. Essentially, OSI serves as a guardian of open source principles, ensuring that software labeled as open source truly adheres to standards that promote collaboration, transparency, and freedom in software development.

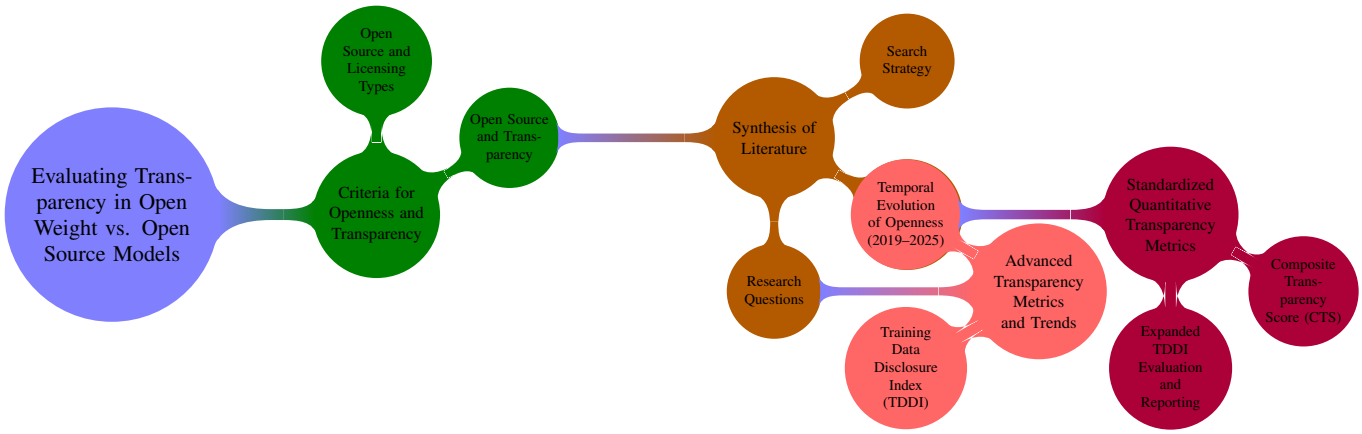

Figure 3: Mind map illustrating the multi-dimensional framework used to evaluate openness and transparency across 121 LLMs (2019–2025). The diagram captures criteria definitions, literature synthesis, metric innovations (CTS, TDDI), and trends over time to distinguish open-weight models from truly open-source models.

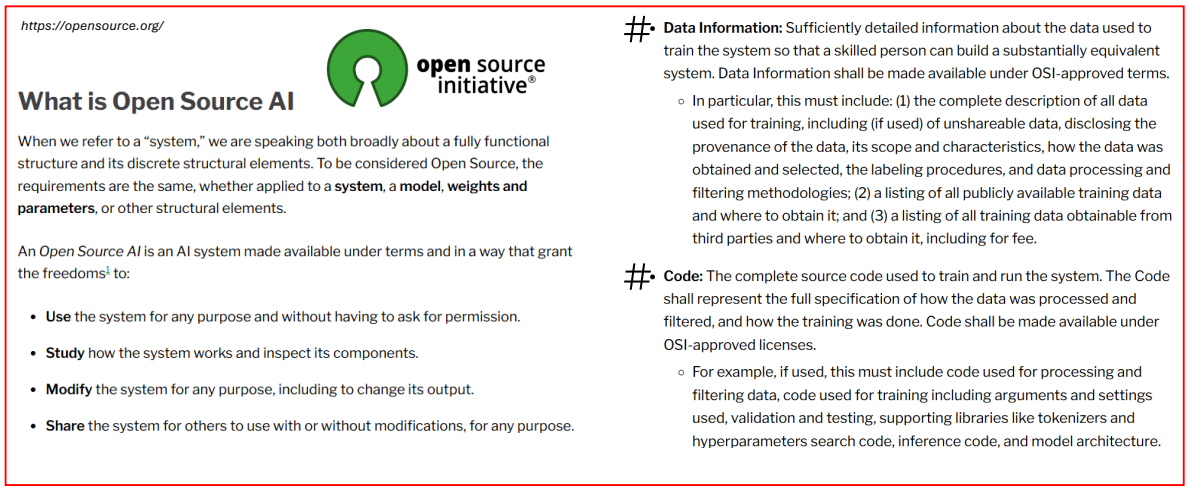

Figure 4: OSI's first official release of the open source definition, which sets foundational criteria/attributes for Openness in AI

The primary attributes of the OSI's official definition of open-source AI are illustrated in Figure 4. OSI emphasizes that for an AI system to be truly open source, there must be unrestricted access to its entire structure. This means that key components such as the model weights, source code, and training data must be accessible under OSI-approved terms. Such access allows any user to use, modify, share, and fully understand the AI system without needing special permissions.

The term "Transparency" refers to the clarity and understandability of the underlying mechanisms that drive AI systems. It is achieved when training data and code are available, enabling stakeholders to replicate and scrutinize the AI's decision-making processes Larsson and Heintz (2020); Felzmann et al. (2020); Von Eschenbach (2021). This openness ensures that AI operations are not only visible but also comprehensible and accountable, thereby enhancing trust and fostering collaboration in AI development and application.

Open source software licenses further define the usage, modification, and distribution rights for software Contractor et al. (2022). They are critical for both protecting creators and enabling users to innovate and adapt software to their needs Quintais et al. (2023). For example, the MIT License terms allow highly permissive, allows almost unrestricted use provided the original copyright is included.Similarly, the Apache License 2.0 [2] permits broad use including modifications and distributions with the additional safeguard of patent rights protection.

Although Creative Commons licenses [3] are primarily designed for creative content, variants such as CC-BY-4.0 can also govern software use by allowing commercial use provided that proper credit is given to the creator. Choosing the right license involves careful consideration of the intended use, attribution requirements, and legal protections, ensuring that software developers can support their objectives while fostering broader collaboration and innovation within the community. Table 2 provides an overview of popular licenses in AI practices, highlighting the varying degrees of permissiveness from the flexible

[2] https://www.apache.org/licenses/LICENSE-2.0
[3] https://creativecommons.org/share-your-work/cclicenses/

MIT License [4] to the stricter copyleft provisions of the GNU GPL 3.0 [5].

2.2.2. Open Source and Transparency

Following the OSI guidelines, the dictionary definitions further support the concept of open source and transparency. *According to Oxford [6] , open source software is described as "Used to describe software for which the original source code is made available to anyone." Cambridge further explains that open source software or information can be "obtained legally and for free from the internet, and can be used, shared or changed without paying or asking for special permission." Merriam-Webster defines it as "Having the source code freely available for possible modification and redistribution." For transparency, Oxford states it as "The quality of something, such as glass, that allows you to see through it." Cambridge calls it "The characteristic of being easy to see through." Merriam-Webster describes transparency as "The quality or state of being transparent so that bodies lying beyond are seen clearly."* These definitions set a foundational understanding to evaluate the transparency practices in AI systems, as shown in Table 3, which presents a literature review and definitions derived from 10 popular literature defining transparency in AI systems.

2.3. Synthesis of Literature

2.3.1. Search Strategy

The study first identified the requirements outlined by the OSI [7] as the baseline for evaluating AI models. These criteria covers various facets of openness, including licensing provisions, access to source code, free redistribution rights, and the ability to modify or derive new work/models from the original codebase. Building on the OSI standards, the concept of "transparency" was clarified through an examination of widely used dictionaries (Oxford, Cambridge, and Merriam-Webster) Röttger et al. (2024). Key steps included:

Databases and Sources: The selection of databases was aligned with the goal of capturing a breadth of interdisciplinary research that intersects with artificial intelligence. Well-known academic repositories such as ACM Digital Library, IEEE Xplore, Elsevier, Nature, Scopus, ScienceDirect, SpringerLink, Wiley Online Library, MathSciNet and renowned pre-print servers like arXiv , TechRxiv were chosen for their extensive coverage of both technical and ethical dimensions pertinent to AI. These platforms are renowned for their consolidation of high-impact and specialized journals, which provide critical insights into both emerging and established research areas within technology and applied sciences.

Our literature search was further reinforced by prioritizing papers that are highly cited within the academic community. Citation counts, often seen as a proxy for the influence and relevance of a study, were utilized as a key metric in selecting sources. Papers with exceptionally high citation counts

Table 2: AI Licenses: A Comprehensive Comparison of Popular Types detailing their requirements for copyright preservation, patent grants, modification rights, distribution terms, and special clauses

License Type	Copyright Preservation	Patent Grant	Modification Rights	Distribution Terms	Special Clauses
MIT License of Technology	Required	No explicit grant	Unlimited modifications	Must include original notices	-
Apache License 2.0 Apache Software Foundation	Required	Includes patent rights	Modifications documented	Must include original notices	-
GNU GPL 3.0 Free Software Foundation	Required	-	Derivative works must also be open source	Source code must be disclosed	Strong copyleft
BSD License of the University of California (c)	Required	No explicit grant	Unlimited modifications	No requirement to disclose source	No endorsement
Creative ML OpenRAIL-M Project (b)	Required	-	Ethical use guidelines	Must include original notices	Ethical guidelines
CC-BY-4.0 Commons (a)	Credit required	-	Commercial and non-commercial use allowed	Must credit creator	-
CC-BY-NC-4.0 Commons (b)	Credit required	-	Only non-commercial use allowed	Must credit creator	Non-commercial use only
BigScience OpenRAIL-M BigScience	Required	-	Ethical use guidelines	Must include original notices	Ethical guidelines
BigCode Open RAIL-M v1 Project (a)	Required	-	Ethical use guidelines	Must include original notices	Ethical guidelines
Academic Free License v3.0 Rosen	Required	Includes patent rights	Unlimited modifications	Must include original notices	-
Boost Software License 1.0 Boost.org	Required	No explicit grant	Unlimited modifications	Must include original notices	-
BSD 2-clause "Simplified" of the University of California (a)	Required	No explicit grant	Unlimited modifications	No requirement to disclose source	No endorsement
BSD 3-clause "New" or "Revised" of the University of California (b)	Required	No explicit grant	Unlimited modifications	No requirement to disclose source	No endorsement

[4] https://opensource.org/license/mit
[5] https://www.gnu.org/licenses/gpl-3.0.en.html
[6] https://www.oed.com/
[7] https://opensource.org/

Table 3: Unified Definitions of Transparency in AI from the published literature.

Author and Reference	Definition
Lipton, Z. C. (2018). Lipton (2018)	"Transparency in machine learning models means understanding how predictions are made, underscored by the **availability of training datasets and code**, which supports both local and global interpretability."
Doshi-Velez, F., & Kim, B. (2017). Doshi-Velez and Kim (2017)	"Transparency in AI refers to the ability to understand and trace the decision-making process, including the **availability of training datasets and code**. This enhances the clarity of how decisions are made within the model."
Arrieta, A. B., et al. (2020). Arrieta et al. (2020)	"AI transparency means understanding the cause of a decision, supported by the **availability of training datasets and code**, which fosters trust in the AI's decision-making process."
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Ribeiro et al. (2016)	"Transparency in AI models provides insights into model behavior, heavily reliant on the **availability of training datasets and code** to illuminate how input features influence outputs."
Goodman, B., & Flaxman, S. (2017). Goodman and Flaxman (2017)	"Transparency involves scrutinizing the algorithms and data used in decisions, emphasizing the **availability of training datasets and code** to ensure fairness and accountability."
Molnar, C. (2020). Molnar (2020)	"Transparency in AI refers to clear communication about decision-making processes, facilitated by the **availability of training datasets and code**, allowing for better understanding of model outputs."
Rudin, et al. (2021). Rudin et al. (2022)	"Transparency is offering clear, interpretable explanations for decisions, which necessitates the **availability of training datasets and code** for full interpretability."
Bhatt, et al. (2020). Bhatt et al. (2020)	"Transparency involves making AI's decision-making process accessible, underlined by the **availability of training datasets and code**, aligning with ethical standards."
Gilpin, et al. (2021). Gilpin et al. (2018)	"Transparency ensures clear explanations of model behavior, significantly relying on the **availability of training datasets and code** for technical and operational clarity."

(e.g., ¿ 3000 citations), were specifically targeted, since they reflect the quality and developments in the state-of-the-art. The search terms used were "Transparency in AI", "Transparency in LLMs", "Explainable AI", "Reproducible AI", "Open Source AI", "Open Source Model", "Open Source Software", "Fairness in AI", "Ethical AI", "Responsible AI", "Bias in AI", "Sustainable AI", "Green AI", "AI Ethics", "AI Accountability", "Interpretable AI", "AI Robustness", "AI Reliability", and "AI Compliance".

Timeframe The literature selected for this study spans publications from 2017 onward a timeframe strategically chosen to align with the introduction of Transformers. In 2017, Vaswani et al. published "Attention is All You Need " Vaswani (2017), marking the beginning of a new era in AI by introducing a model architecture based on attention mechanisms. Following this, the launch of GPT-2, T5, BART, and several other language model architectures further advanced the field, shaping the development of modern LLMs. We systematically assessed these models to identify models that exemplify various degrees of openness, including open-source and open-weight practices.

In the process of synthesizing these findings, we evaluated a total of 121 LLMs, a sample that represents the diverse and rapidly evolving landscape of language models from 2019 to 2025. These models were analyzed based on a wide array of architectural specifications such as the number of layers, hidden unit sizes, attention head counts, and overall parameter scales as well as openness metrics including licensing type and the public availability of training resources. The model development trend, illustrated in Figure 5, provides a visual representation of the evolution of these models. The figure shows that although the foundational literature for LLMs was established with the advent of Transformers in 2017, the major model breakthroughs and integrated transparency and accessibility features have predominantly materialized from 2019 onward and more post-ChatGPT era (Nov. 2022).

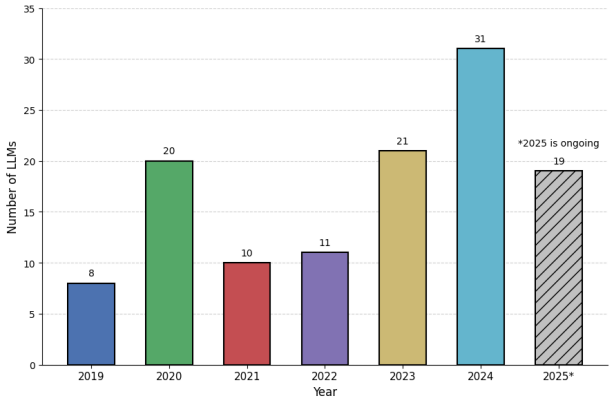

Figure 5: Year-wise distribution of large language models (LLMs) reported in this review from 2019 to 2025, highlighting a peak in 2024. The 2025 count is ongoing and may rise. The plot reveals a steadily increasing trend in LLM studies, underscoring rapid advancements in transparency and accessibility.

Inclusion Criteria A thorough literature review was conducted to cover the topic within broader discourses in AI development and ethics. This review captured highly cited articles and technical reports, emphasizing themes such as explainable AI, reproducibility, interpretability, and responsible AI governance Raza et al. (2025). By synthesizing these studies, our study addressed both technical (e.g., code-level transparency) and ethical (e.g., data biases) dimensions of openness.

2.4. Evaluation Framework and Application

Findings from the previous stages were synthesized into five key dimensions representing critical facets of open-source and open-weight classifications:

1. Licensing, Usage, and Redistribution Rights
2. Code Accessibility and Modification Rights
3. Training Data Transparency
4. Community and Support
5. MMLU Score and Carbon Emissions
6. Ethical Considerations and Reproducibility

Each of these dimensions was assessed to determine whether a given model adhered to OSI-like openness or employed more restrictive practices similar to "open-weight" approaches (i.e., sharing only the model parameters). SoTA LLMs were systematically evaluated against these five dimensions as 1) Licensing,

usage, and redistribution rights , 2) Training Code and Training Data, 3) Community Support, 4) Open source, and 5) Open Weights. Any evidence of collaborative contributions or transparent reporting of potential biases and vulnerabilities was also documented.

2.5. Research Questions

The methodology section of this study was structured around a detailed mind map, as depicted in Figure 3. This visual representation, employed to assess the transparency and openness of state-of-the-art (SoTA) multimodal LLMs, organized the analytical framework into three main branches, each corresponding to a specific research question (RQ) as follows:

1. What drives the classification of LLMs as *open weights* rather than *open source*, and what impact do these factors have on on the reproducibility of the research, and the efficiency in terms of usage and the scalability in practical applications?
2. How do current training approaches influence transparency and reproducibility, potentially prompting developers to favor open-weight models?
3. How does the limited disclosure of training data and methodologies impact both the performance and practical usability of these models, and what future implications arise for developers and end-users?

This methodology integrates well-established open-source standards, linguistically and ethically grounded definitions of transparency, and a structured evaluation framework. The outcome is an assessment of whether leading MLLMs adhere to open-source principles or merely present limited transparency through open-weight practices. The subsequent sections detail the findings that emerged from applying this framework, highlighting significant discrepancies and implications for researchers, developers, and broader AI stakeholders.

2.6. Transparency Metrics and Trends

To deepen our evaluation of LLM transparency beyond licensing and weight availability, we introduce two novel analytical tools: the Training Data Disclosure Index (TDDI) and a temporal evolution analysis of openness practices between 2019 and 2025.

Training Data Disclosure Index (TDDI)

TDDI provides a quantitative measure to assess how openly LLMs report information about their training data. Given that model performance, bias, and reproducibility are highly dependent on data quality, the TDDI evaluates five core attributes:

1. **Data Source Specification**:Whether the types of data (e.g., books, code, web, scientific literature) are named.
2. **Licensing and Copyright Clarity**: Whether rights and permissions of data use are documented.
3. **Preprocessing Methods**: Disclosure of filtering, deduplication, or tokenization techniques.
4. **Multilingual and Domain Distribution**: Reporting of language coverage and domain balance (e.g., STEM, legal, medical).
5. **Synthetic vs. Organic Ratios**: Disclosure of proportions between artificially generated and real-world data.

Each criterion in the Training Data Disclosure Index (TDDI) is scored on a binary (0/1) basis for presence, resulting in a composite score ranging from 0 (opaque) to 5 (fully transparent). While indices are often normalized to a 0-1 range, we report raw scores here to preserve interpretability across the five distinct dimensions of training data transparency. In future work, we intend to scale the TDDI for compatibility with other standardized evaluation metrics. We applied the TDDI to 25 representative LLMs across model families, including GPT, LLaMA, DeepSeek, and T5. The results indicate that only a few models such as Llemma and BLOOM achieve a TDDI score above 3, highlighting a systemic opacity in training dataset reporting even among so-called "open-weight" models.

2.7. Temporal Evolution of Openness (2019–2025)

To contextualize transparency trends over time, we performed a year-wise analysis of 112 LLMs released between 2019 and 2025. Each model was evaluated on five transparency dimensions: availability of training data, release of source code, accessibility of model weights, openness of licensing terms, and quality of accompanying documentation defined by the presence of detailed model cards, training and evaluation procedures, and disclosure of known limitations or biases. The results show a clear change in direction after 2021, particularly following the release of ChatGPT. Prior to 2022, models like BERT and GPT-2 followed more open-source practices. However, the post-ChatGPT era marks a shift towards hybrid practices, especially among commercial models that selectively disclose weights while withholding training pipelines and data. In contrast, community-driven models like BLOOM, Mistral, and Llemma demonstrate greater adherence to open-source norms in recent years. This temporal analysis, based on TDDI and other parameters reveals how transparency has evolved in response to commercial pressures and public scrutiny.

2.8. Standardized Quantitative Transparency Metrics

To further strengthen our evaluation framework, this section introduces a standardized, quantitative scoring system for assessing the transparency of LLMs. While prior literature often discusses openness qualitatively, there remains a need for measurable, replicable metrics that enable consistent cross-model comparison. Our proposed framework includes two main components: the Composite Transparency Score (CTS) and the extended Training Data Disclosure Index (TDDI), both grounded in principles from open science, software licensing, and AI reproducibility.

Composite Transparency Score (CTS)

To systematically evaluate the transparency of leading LLMs, we developed the CTS, a structured 7-point metric designed to quantify openness across critical dimensions neces-

sary for reproducibility, interpretability, and responsible deployment. Each dimension is assessed using a binary scoring system (0 or 1), resulting in a total score ranging from 0 to 7 per model. The seven components of CTS visualized in Figure 6 include code availability, weight availability, training data disclosure, documentation of evaluation protocols, license openness, carbon emission reporting, and benchmark reproducibility. Notably, the training data disclosure component of CTS is directly informed by the Training Data Disclosure Index (TDDI), which offers a finer-grained assessment across five sub-criteria. For models scoring greater or equal to 3 in TDDI, a point is assigned under the CTS training data dimension. This integration ensures that CTS reflects both breadth and depth in evaluating transparency. Furthermore, while our initial temporal trend analysis focused on TDDI, we now extend this analysis using CTS, enabling a more holistic view of how transparency practices have evolved across the broader spectrum of openness dimensions from 2019 to 2025.

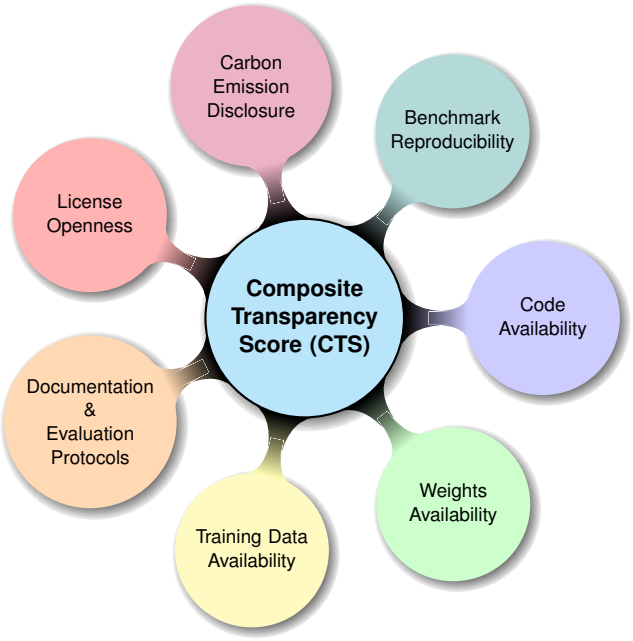

Figure 6: Mind map of the Composite Transparency Score (CTS), highlighting the seven binary dimensions used to quantify transparency in LLM development.

Each model was evaluated using publicly available documentation, technical reports, GitHub repositories, and published model cards. To enable consistent comparison, we normalized the Composite Transparency Score (CTS) from a raw scale of 0 to 7 to a normalized range of 0.0 to 1.0, where 1.0 indicates complete transparency across all seven evaluated dimensions. These dimensions include: source code availability, model weight release, training data disclosure, evaluation documentation, license openness, carbon emission reporting, and benchmark reproducibility. Notably, the training data disclosure dimension was originally scored using the Training Data Disclosure Index (TDDI) on a 0–5 scale. For integration into CTS, we applied a thresholding approach: models with TDDI scores of 3 or higher were assigned a value of 1, while those below this threshold received a 0. This binary transformation maintains consistency with the CTS's categorical structure but is based on a more nuanced underlying scale. For instance, ChatGPT-4 receives a normalized CTS of 0.14, reflecting its limited transparency: only evaluation documentation is available, while most other components including training data, code, and emissions remain undisclosed. DeepSeek-R1, which provides model weights, emissions data, and partial code, achieves a normalized CTS of 0.43. LLaMA 2, though releasing weights and limited code under a custom license, lacks data provenance and carbon metrics, resulting in a CTS of 0.29. In contrast, BLOOM achieves a perfect normalized CTS of 1.0, supported by full transparency across all dimensions, including an openly published dataset (with a TDDI of 5), emissions reporting, and reproducible benchmarks under an open license.

To evaluate training data transparency with greater precision, we developed the TDDI, a normalized metric that captures disclosure quality across five key dimensions. Each is scored binarily and aggregated into a score from 0 to 1, enabling fine-grained and reproducible comparisons across both proprietary and open-weight models. TDDI also feeds directly into the CTS, adding quantitative specificity to its training data component. Together, the CTS and TDDI form a standardized framework for benchmarking model transparency, supporting informed decision-making by researchers, developers, and policymakers in advancing openness and accountability in foundation model development.

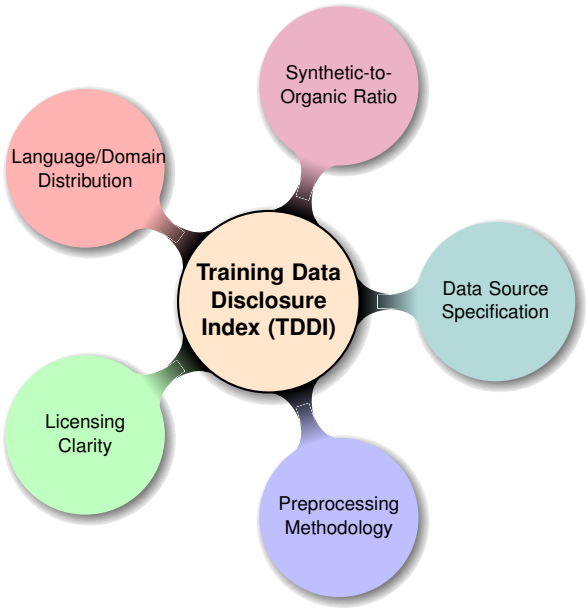

Figure 7: Mind map of the Training Data Disclosure Index (TDDI), used to assess LLM transparency. Each branch represents a core dimension evaluated during our scoring.

3. Results

3.1. Overall Findings on Openness and Transparency

Drawing on OSI guidelines, dictionary-based definitions of transparency, and scholarly literature, this narrative review re-

veals that many models marketed or perceived as "open" primarily provided open weights (i.e., publicly available trained parameters) rather than full open-source access (i.e., source code, training data, and detailed methodologies). Table 4 outlines these distinctions across leading multimodal LLMs.

This comprehensive table (Table 4) compares 121 LLMs released between 2019 and 2025 in terms of release year, training data, and other key features. Early models, such as GPT-2 (Brown et al., 2020) and BERT (Devlin et al., 2019), primarily focused on foundational capabilities, including improved text generation, masked language modeling, and next-sentence prediction. These models relied on relatively simple training data and featured basic natural language processing tasks. However, subsequent developments have led to a remarkable progression in both complexity and functionality. Recent models such as DeepSeek-R1 (Guo et al., 2025) and advanced iterations of ChatGPT, introduce enhanced multimodal capabilities, advanced reasoning through mixture-of-experts (MoE) architectures, and efficient scaling strategies. The table demonstrates that models released after 2020 increasingly leverage diverse and massive training datasets, from extensive web corpora to hybrid synthetic-organic data, which substantially boost performance. Moreover, these models exhibit notable improvements in precision, processing speed, and bias mitigation. Although many SoTA models disclose only pre-trained weights, thereby limiting reproducibility, our findings indicate that the majority of recent models particularly those released after 2023 continue to exhibit low transparency scores, with limited documentation of training data and methodologies. This highlights an ongoing prioritization of performance and proprietary control over openness, suggesting that while transparency remains a recognized value, it is not yet widely adopted in practice across the latest generation of LLMs.

Table 4: **Detailed specifications of Large Language Models (2019Unknown2025), including the model name and citation, release year, training data characteristics, and key features. It offers a comparative analysis across these crucial aspects, providing insights into the evolution, dataset diversity, and unique capabilities of each model**.

Model, Year & Citation	Training Data	Key Features
1.GPT-2 (2019) Brown et al. (2020)	WebText dataset (8M web pages)	Improved text generation, zero-shot learning
2.Legacy ChatGPT-3.5 (2022)	Text/Code (pre-2021)	Basic text tasks, translation
3.Default ChatGPT-3.5 (2023)	Text/Code (pre-2021)	Faster, less precise
4.GPT-3.5 Turbo (2023)	Text/Code (pre-2021)	Optimized accuracy
5.ChatGPT-4 (2023)	Text/Code (pre-2023)	Multimodal (text), high precision
6.GPT-4o (2024) Hurst et al. (2024)	Text/Code (pre-2024)	Multimodal (text/image/audio/video)
7.GPT-4o mini (2024)	Text/Code (pre-2024)	Cost-efficient, 60% cheaper
8. o1-preview Jaech et al. (2024) (2024)	STEM-focused data	System 2 thinking, PhD-level STEM
9. o1-mini (2024)	STEM-focused data	Fast reasoning, 65K tokens output
10. o1 (2025)	General + STEM data	Full o1 reasoning, multimodal
11. o1 pro mode (2025)	General + STEM data	Enhanced compute, Pro-only
12. o3-mini (2025)	General + STEM data	o1-mini successor
13. o3-mini-high (2025)	General + STEM data	High reasoning effort
14. DeepSeek-R1 Guo et al. (2025) (2025)	Hybrid dataset of 9.8T tokens from synthetic and organic sources	Mixture of Experts (MoE), enhanced with mathematical reasoning capabilities
15. DeepSeek LLM Bi et al. (2024) (2023)	Books+Wiki data up to 2023	Scaling Language Models
16. DeepSeek LLM V2 Liu et al. (2024a) (2023)	Highly efficient training	MLA, MoE, Lowered costs
17. DeepSeek Coder V2 Zhu et al. (2024) (2023)	Supports 338 languages	Enhanced coding capabilities
18. DeepSeek V3 Liu et al. (2024b) (2023)	Advanced MoE architecture	High-performance, FP8 training
19. BERT-Base Devlin et al. (2019) (2019)	Books+Wiki data collected up to 2019	Masked Language Modeling (MLM)
20. BERT-Large Devlin et al. (2019) (2019)	Books+Wiki data collected up to 2019	Next Sentence Prediction (NSP)
21. T5-Small Raffel et al. (2020) (2020)	C4 (Large-scale text dataset)	Text-to-text, encoder-decoder
22. T5-Base Raffel et al. (2020) (2020)	C4 (Large-scale text dataset)	Text-to-text, scalable, encoder-decoder
23. T5-Large Raffel et al. (2020) (2020)	C4 (Large-scale text dataset)	Text-to-text, scalable, encoder-decoder
24. T5-3B Raffel et al. (2020) (2020)	C4 (Large-scale text dataset)	Text-to-text, scalable, encoder-decoder
25. T5-11B Raffel et al. (2020) (2020)	C4 (Large-scale text dataset)	Text-to-text, scalable, encoder-decoder
26. Mistral 7B Jiang et al. (2023b) (2023)	Compiled from diverse sources totaling 2.4T tokens	Sliding Window Attention (SWA)
27. LLaMA 2 70B Touvron et al. (2023) (2023)	Diverse corpus aggregated up to 2T tokens	Grouped Query Attention (GQA)
28. CriticGPT McAleese et al. (2024) (2024)	Human feedback data	Fine-tuned for critique generation
29. Olympus (2024)	40T tokens	Large-scale, proprietary model
30. HLAT Fan et al. (2024) (2024)	Not specified	High-performance, task-specific
31. Multimodal-CoT Zhang et al. (2023) (2023)	Multimodal datasets	Chain-of-Thought reasoning for multimodal tasks
32. AlexaTM 20B Soltan et al. (2022) (2022)	Not specified	Multilingual, task-specific
33. Chameleon Team (2024a) (2024)	9.2T tokens	Multimodal, high-performance
34. Llama 3 70B AI (2024) (2024)	2T tokens	High-performance, open-source
35. LIMA Zhou et al. (2024) (2024)	Not specified	High-performance, task-specific
36. BlenderBot 3x Xu et al. (2023) (2023)	300B tokens	Conversational AI, improved reasoning
37. Atlas Izacard et al. (2023) (2023)	40B tokens	High-performance, task-specific
38. InCoder Fried et al. (2022) (2022)	Not specified	Code generation, task-specific
39. 4M-21 Bachmann et al. (2024) (2024)	Not specified	High-performance, task-specific
40. Apple On-Device model Mehta et al. (2024) (2024)	1.5T tokens	On-device, task-specific
41. MM1 McKinzie et al. (2024) (2024)	2.08T tokens	Multimodal, high-performance
42. ReALM-3B Moniz et al. (2024) (2024)	134B tokens	High-performance, task-specific
43. Ferret-UI You et al. (2024) (2024)	2T tokens	Multimodal, high-performance
44. MGIE Fu et al. (2023) (2023)	2T tokens	Multimodal, high-performance
45. Ferret You et al. (2023) (2023)	2T tokens	Multimodal, high-performance

Continued on next column/page

Continued from previous column/page

Continued on next column/page

Continued from previous column/page

Model, Year & Citation	Training Data	Key Features
46. Nemotron-4 340B Adler et al. (2024) (2024)	9T tokens	High-performance, task-specific
47. VIMA Jiang et al. (2023a) (2023)	Not specified	Multimodal, high-performance
48. Retro 48B Wang et al. (2023) (2023)	1.2T tokens	High-performance, task-specific
49. Raven Huang et al. (2023) (2023)	40B tokens	High-performance, task-specific
50. Gemini 1.5 Reid et al. (2024) (2024)	Not specified	Multimodal, high-performance
51. Med-Gemini-L 1.0 Saab et al. (2024) (2024)	30T tokens	Medical-focused, high-performance
52. Hawk De et al. (2024) (2024)	300B tokens	High-performance, task-specific
53. Griffin De et al. (2024) (2024)	300B tokens	High-performance, task-specific
54. Gemma Team et al. (2024b) (2024)	6T tokens	High-performance, task-specific
55. Gemini 1.5 Pro Reid et al. (2024) (2024)	30T tokens	Multimodal, high-performance
56. PaLi-3 Chen et al. (2023) (2023)	Not specified	Multimodal, high-performance
57. RT-X Padalkar et al. (2023) (2023)	Not specified	Robotics-focused, high-performance
58. Med-PaLM M Tu et al. (2024) (2024)	780B tokens	Medical-focused, high-performance
59. MAI-1 Team (2024c) (2024)	10T tokens	High-performance, task-specific
60. YOCO Sun et al. (2024) (2024)	1.6T tokens	High-performance, task-specific
61. phi-3-medium Abdin et al. (2024b) (2024)	4.8T tokens	High-performance, task-specific
62. phi-3-mini Abdin et al. (2024b) (2024)	3.3T tokens	High-performance, task-specific
63. WizardLM-2-8x22B Team (2024b) (2024)	Not specified	High-performance, task-specific
64. WaveCoder-Pro-6.7B Yu et al. (2023) (2023)	20B tokens	Code-focused, high-performance
65. WaveCoder-Ultra-6.7B Yu et al. (2023) (2023)	20B tokens	Code-focused, high-performance
66. WaveCoder-SC-15B Yu et al. (2023) (2023)	20B tokens	Code-focused, high-performance
67. OCRA 2 Mitra et al. (2023) (2023)	Not specified	High-performance, task-specific
68. Florence-2 Xiao et al. (2024) (2024)	5.4B visual annotations	Multimodal, high-performance
69. Qwen Bai et al. (2023) (2023)	3T tokens	High-performance, task-specific
70. SeaLLM-13b Nguyen et al. (2023) (2023)	2T tokens	Multilingual, high-performance
71. Grok-1 xAI (2024) (2024)	13.2T tokens	Incorporates humor-enhancing algorithms
72. Phi-4 Abdin et al. (2024b) (2024)	9.8T tokens	Optimized for STEM applications
73. Megatron-LM Shoeybi et al. (2019) (2020)	Common Crawl, Wikipedia, Books	Large-scale parallel training, optimized for NVIDIA GPUs
74. Turing-NLG Smith et al. (2022) (2020)	Diverse web text	High-quality text generation, used in Microsoft products
75. CTRL Keskar et al. (2019) (2019)	Diverse web text with control codes	Controlled text generation using control codes
76. XLNet Yang (2019) (2019)	BooksCorpus, Wikipedia, Giga5, ClueWeb	Permutation-based training, outperforms BERT on many benchmarks
77. RoBERTa Liu (2019) (2019)	BooksCorpus, Wikipedia, CC-News, OpenWebText	Improved BERT with better pretraining techniques

Continued on next column/page

Continued from previous column/page

Model, Year & Citation	Training Data	Key Features
78. ELECTRA Clark (2020) (2020)	BooksCorpus, Wikipedia	Replaces masked language modeling with a more efficient discriminative task
79. ALBERT Lan (2019) (2019)	BooksCorpus, Wikipedia	Parameter reduction techniques for efficient training
80. DistilBERT Sanh (2019) (2019)	BooksCorpus, Wikipedia	Distilled version of BERT, smaller and faster
81. BigBird Zaheer et al. (2020) (2020)	BooksCorpus, Wikipedia, PG-19	Sparse attention mechanism for handling long sequences
82. Gopher Rae et al. (2021) (2021)	MassiveText dataset (2.5T tokens)	Focused on scaling laws and model performance
83. Chinchilla Hoffmann et al. (2022) (2022)	MassiveText dataset (1.4T tokens)	Optimized for compute-efficient training
84. PaLM Chowdhery et al. (2023) (2022)	Diverse web text, books, code	Pathways system for efficient training, multilingual support
85. OPT Zhang et al. (2022) (2022)	Diverse web text	Open-source alternative to GPT-3
86. BLOOM Workshop et al. (2022) (2022)	ROOTS corpus (1.6T tokens)	Multilingual, open-source, collaborative effort
87. Jurassic-1 Lieber et al. (2021) (2021)	Diverse web text	High-quality text generation, API-based access
88. Codex Chen et al. (2021) (2021)	Code repositories (e.g., GitHub)	Specialized in code generation and understanding
89. T0 Sanh et al. (2021) (2021)	Diverse NLP datasets	Zero-shot task generalization
90. UL2 Tay et al. (2022) (2022)	Diverse web text	Unified pretraining for diverse NLP tasks
91. GLaM Du et al. (2022) (2021)	Diverse web text	Sparse mixture of experts (MoE) architecture
92. ERNIE 3.0 Sun et al. (2021) (2021)	Chinese and English text	Knowledge-enhanced pretraining
93. GPT-NeoX Black et al. (2022) (2022)	The Pile (825GB dataset)	Open-source, large-scale, efficient training
94. CodeGen Nijkamp et al. (2022) (2022)	Code repositories (e.g., GitHub)	Specialized in code generation
95. FLAN-T5 Chung et al. (2024) (2022)	Diverse NLP datasets	Instruction fine-tuning for better generalization
96. mT5 Xue (2020) (2020)	mC4 dataset (101 languages)	Multilingual text-to-text transfer
97. Reformer Kitaev et al. (2020) (2020)	Diverse web text	Efficient attention mechanism for long sequences
98. Longformer Beltagy et al. (2020) (2020)	BooksCorpus, Wikipedia	Efficient attention for long documents
99. DeBERTa He et al. (2020) (2021)	BooksCorpus, Wikipedia	Disentangled attention mechanism
100. T-NLG Rosset (2020) (2020)	Diverse web text	High-quality text generation
101. Switch Transformer Fedus et al. (2022) (2021)	Diverse web text	Sparse mixture of experts (MoE)
102. WuDao 2.0 Jie (2021) (2021)	Chinese and English text	Largest Chinese language model
103. LaMDA Thoppilan et al. (2022) (2021)	Diverse dialogue data	Specialized in conversational AI
104. MT-NLG Smith et al. (2022) (2021)	Diverse web text	High-quality text generation
105. GShard Lepikhin et al. (2020) (2020)	Diverse web text	Sparse mixture of experts (MoE)
106. T5-XXL Raffel et al. (2020) (2020)	C4 dataset	Large-scale text-to-text transfer
107. ProphetNet Qi et al. (2020) (2020)	BooksCorpus, Wikipedia	Future token prediction for better sequence modeling
108. DialoGPT Zhang (2019) (2020)	Reddit dialogue data	Specialized in conversational AI

Continued on next column/page

Continued from previous column/page

Model, Year & Citation	Training Data	Key Features
109. BART Lewis (2019) (2020)	BooksCorpus, Wikipedia	Denoising autoencoder for text generation
110. PEGASUS Zhang et al. (2020) (2020)	C4 dataset	Pre-training with gap-sentences for summarization
111. UniLM Dong et al. (2019) (2020)	BooksCorpus, Wikipedia	Unified pre-training for NLU and NLG tasks
112. Grok 3 (2025)	Synthetic Data	trained with ten times more computing power than its predecessor, Grok 2
113. Gemini 2 Ultra (2025) Source Link	Multimodal data (text, images, audio, video) up to 2025	Multimodal, advanced tool use, 2M context window, improved factuality
114. GPT-4.5 (2025) Source Link	Proprietary, web, code, multilingual, up to early 2025	Improved reasoning, creative generation, bridge model before GPT-5
115. Gemini 2.0 Flash-Lite (2025) Source Link	Multimodal data (text, images, audio, video), up to 2025	Multimodal, lightweight, efficient, improved context window
116. Gemini 2.0 Pro (2025) Source Link	Multimodal data (text, images, audio, video), up to 2025	Multimodal, advanced context window, improved tool use
117. Gemini 2.5 Pro (2025) Source Link	Multimodal, text, images, code, up to 2025	Further improved context, advanced multimodal capabilities
118.GPT-o3 (2025) Source Link	Proprietary, web, code, up to 2025	Fast, cost-efficient, strong reasoning, 200K context window
119. GPT-4.1 (2025) Source Link	Proprietary, web, code, images, up to 2025	1M token context window, improved instruction following, lower latency/cost
120. GPT-o4-mini (2025) Source Link	Proprietary, web, code, up to 2025	Small, efficient, 200K context, image input, fast and affordable
121. Qwen 3 235B (2025) Yang et al. (2025)	8T tokens, web, code, multilingual, up to 2025	Open-weight, 235B params, strong math/coding, 128K context

In recent years, LLMs have been further advanced through continuous refinement of critical features essential for practical applications. Models such as T5-XXL (Raffel et al., 2020) have significantly expanded both the scale and diversity of training data, transitioning from datasets with millions of tokens to those with trillions. This dramatic increase in training volume has enabled improvements in model generalization and reduced overfitting, which in turn allows for more efficient inference and scaling through optimized parameter utilization and training convergence. Additionally, evolving training methodologies from basic text-to-text transfer to hybrid approaches have resulted in models that are increasingly capable of handling complex, real-world tasks. Advances in ethical and operational transparency, as evidenced by improved MMLU scores and the integration of sustainability metrics (e.g., carbon emissions tracking), underscore a dual focus on technical performance and responsible adoption. The emergence of open-weight models, such as those from DeepSeek and ChatGPT, illustrates a deliberate strategy to balance accessibility with proprietary innovation. These studies summarized in Table 4 suggest that future LLMs will continue to build on these innovations, paving the way for more transparent, efficient, and ethically responsible AI systems (Hoffmann et al., 2022).

Furthermore, Table 9 added to Appendix 1, provides a comprehensive overview of these 121 LLMs investigated in this study, detailing both architectural specifications and openness metrics. A clear pattern emerges regarding licensing: prominent models such as the GPT family (e.g., GPT-2 (Brown et al., 2020), ChatGPT-3.5, ChatGPT-4) largely adopt proprietary licenses, restricting access to their training data, code, and methodologies. In contrast, models like BERT (Devlin et al., 2019) and certain DeepSeek variants (e.g., DeepSeek-R1 (Guo et al., 2025)) are disseminated under open-source licenses such as MIT or Apache 2.0, which facilitate greater transparency through public availability of weights and, in some cases, additional resources. The DeepSeek family, for example, demonstrates a strategic move toward open-weight transparency while still withholding full training pipelines. Similar trends are observed in the T5 series (Raffel et al., 2020) and LLaMA 2 (Touvron et al., 2023), where a mix of open and proprietary strategies reflects competing priorities commercial viability versus reproducibility. This heterogeneous licensing landscape, as reinforced by studies (e.g., Guo et al., 2025; Devlin et al., 2019; Raffel et al., 2020; Hurst et al., 2024; Zhu et al., 2024; Jiang et al., 2023b), demonstrates the challenges in balancing innovation, transparency, and community engagement in modern LLM development.

Despite variances in licensing terms, from permissive licenses (e.g., MIT, Apache 2.0) to more restrictive or proprietary frameworks, training data and code remain largely undisclosed in most of the models reviewed. Community engagement and support generally appeared robust (via forums, documentation, or user guides), but comprehensive transparency of datasets, training pipelines, and model internals such as hyperparameters and attention mechanisms remains limited. These findings align with broader trends in AI development, where commercial or strategic interests often restrict full access to the underlying training infrastructure Mazzucato et al. (2022); Guha et al. (2024).

Analysis (Table 9; Appendix 1) reveals crucial trends in performance and sustainability metrics. Notably, recent models in the ChatGPT family achieve high MMLU scores; ChatGPT-4, for instance, reports an MMLU score of 86.4%, while also indicating substantial increase in carbon emissions (e.g., 552 tCO_2eq for several variants and 1,035 tCO_2eq for GPT-4o (Hurst et al., 2024)). In contrast, earlier models such as GPT-2 lack these performance benchmarks, reflecting evolving capabilities of newer LLMs. The DeepSeek family shows a promising balance MMLU and carbon emission compared to chatGPT as DeepSeek-R1 records a robust MMLU score (90.8%) and comparatively lower carbon emissions (44 tCO_2eq), suggesting improved energy efficiency and refined training methodologies. Similar sustainability (improving) trends are evident in the T5 series (Raffel et al., 2020) and LLaMA 2 (Touvron et al., 2023), which have progressively incorporated larger, more diverse datasets alongside performance improvements. Studies (e.g., Hoffmann et al., 2022; Zhu et al., 2024; Tay et al., 2022; Rae et al., 2021; Guo et al., 2025) indicate that while performance enhancements are significant, the associated environmental costs suggest a shift toward more energy-efficient architectures and transparent reporting practices.

Table 9 also presents the architectural specifications that underpin model performance. The GPT family of models, in-

cluding variants like ChatGPT-3.5 and GPT-4, generally features 96 layers, 12,288 hidden units, and parameter counts scaling up to 1.8T, indicating a massive computational footprint (Hurst et al., 2024). In contrast, the DeepSeek family employs a different architecture DeepSeek-R1, for instance, is built with 64 layers and 8192 hidden units, achieving high performance (MMLU score of 90.8%) with relatively fewer parameters (671B). The T5 series (Raffel et al., 2020) and LLaMA 2 (Touvron et al., 2023) further illustrate a trend toward optimizing architectural design for scalability, efficiency, and energy conservation. These models reveal a shift from larger scale models towards balanced configurations that emphasize reproducibility and ethical considerations. Several studies (e.g., Devlin et al., 2019; Raffel et al., 2020; Touvron et al., 2023; Zhu et al., 2024; Jiang et al., 2023b; Hoffmann et al., 2022) support the observation that while larger models deliver superior performance, they also present challenges in terms of energy consumption and transparency. Overall, the architectural trends underscore the importance of evolving design principles that may reconcile performance, efficiency, and openness in next-generation LLMs.

Early models such as GPT-2 (Brown et al., 2020) and BERT (Devlin et al., 2019) laid the groundwork with moderate layer counts, hidden nodes, and attention head configurations. As the field evolved, later models particularly within the ChatGPT and DeepSeek families (Guo et al., 2025; Zhu et al., 2024), exhibited significant increases in layers, hidden units, and overall parameter scales, reflecting a trend toward more complex architectures designed for enhanced performance and multimodal capabilities. The table categorizes the openness of training data (fully open, partially open, or proprietary) and evaluates accessibility to model weights, code, and training datasets, thereby delineating a clear divergence between models that offer full reproducibility and those that only provide open weights. MMLU scores and reported carbon emissions further indicate that while state-of-the-art models achieve higher performance, they also incur greater environmental costs a factor increasingly scrutinized in recent literature (Raffel et al., 2020; Touvron et al., 2023; Hoffmann et al., 2022). Overall, this extended analysis highlights an industry-wide progression from simpler architectures with limited transparency to highly engineered systems that try to balance commercial interests with technical rigor and ethical considerations.

3.2. Model-Specific Evaluations

3.2.1. ChatGPT

GPT-4 OpenAI (2023) and ChatGPT Achiam et al. (2023) are proprietary models with limited architectural transparency: their training datasets, fine-tuning methods, and structural details (e.g., layer configurations, attention mechanisms) remain undisclosed. While GPT-4's technical report outlines high-level capabilitiesGallifant et al. (2024), it omits reproducibility-critical specifics such as pre-training corpus composition, hyperparameters, and energy consumption metrics,, which limits its scientific openness. Similarly, ChatGPT's API-based access restricts users to input-output interactions without exposing model internals Lande and Strashnoy (2023), thus creating a "black box" system that lacks transparency and does not allow third-party modifications.

ChatGPT adopts a functional accessibility paradigm, where API endpoints enable task execution (e.g., text generation, reasoning) but do not allow direct weight inspection, retraining, or redistribution Wolfe et al. (2024); Roumeliotis and Tselikas (2023). This approach, therefore, creates a dependency on proprietary infrastructure, which can limit long-term reproducibility and bias mitigation in downstream applications. While the term "open-weights" is occasionally used to describe these systems due to their API availability, this usage can be ambiguous because true open-weight standards such as parameter accessibility (e.g., Llama 2) or training code disclosure (e.g., BLOOM BigScience Workshop (2022)), are absent, underscoring the competing priorities between commercial control and open scientific collaboration in modern AI ecosystems. The ChatGPT's version's:

- **GPT-2:** Brown et al. (2020) adopts an open-weights model under MIT License, providing full access to its 1.5B parameters and architectural details (48 layers, 1600 hidden size). However, the WebText training dataset (8M web pages) lacks comprehensive documentation of sources and filtering. While permitting commercial use and modification, the absence of detailed pre-processing methodologies limits reproducibility of its zero-shot learning capabilities.

- **Legacy ChatGPT-3.5:** Legacy ChatGPT-3.5 uses proprietary weights with undisclosed architectural details (96 layers, 12288 hidden size). The pre-2021 text/code training data lacks domain distribution metrics and copyright compliance audits. API-only access restricts model introspection or bias mitigation, despite claims of basic translation/text task capabilities Jaech et al. (2024).

- **Default ChatGPT-3.5:** Default ChatGPT-3.5 Jaech et al. (2024) shares Legacy's proprietary architecture but omits fine-tuning protocols for its "faster, less precise" variant. Training data temporal cutoff (pre-2021) creates recency gaps unaddressed in technical documentation. Restricted API outputs prevent reproducibility of the 69.5% MMLU benchmark results.

- **GPT-3.5 Turbo:** GPT-3.5 Turbo Jaech et al. (2024) employs encrypted weights with undisclosed accuracy optimization techniques. The 16K context window expansion lacks computational efficiency metrics or energy consumption disclosures. Proprietary licensing blocks third-party latency benchmarking despite "optimized accuracy" claims.

- **GPT-4o:** GPT-4o Hurst et al. (2024) uses multimodal proprietary weights (1.8T parameters) with undisclosed cross-modal fusion logic. Training data (pre-2024 text/image/audio/video) lacks ethical sourcing validations for sensitive content. "System 2 thinking" capabilities lack peer-reviewed validation pipelines.

- **GPT-4o mini:** GPT-4o mini Hurst et al. (2024) offers cost-reduced proprietary access (1.2T parameters) with undisclosed pruning methodologies. The pre-2024 training corpus excludes synthetic data ratios and human feedback alignment details. Energy efficiency claims (60% cost reduction) lack independent verification.

3.2.2. DeepSeek

The DeepSeek-R1 model, a 671-billion-parameter mixture-of-experts (MoE) system built on the DeepSeek-V3 architecture, adopts an open-weights framework under the MIT License, permitting unrestricted access to its neural network parameters for commercial and research use Guo et al. (2025). MoE is an ensemble machine learning technique where multiple specialist models (referred to as "experts") are trained to handle different parts of the input space, and a gating model decides which expert to consult for a given input Vasić et al. (2022); Masoudnia and Ebrahimpour (2014). This method allows for more scalable and efficient training as well as inference processes, especially in complex models like DeepSeek-R1, by dynamically allocating computational resources to the most relevant experts for specific tasks or data points.

While the DeepSeek-R1 model's weights and high-level architectural details including its MoE design with 37 billion activated parameters per inference and reinforcement learning-augmented reasoning pipelines are publicly disclosed, critical transparency gaps persist. The pre-training dataset composition, comprising a hybrid of synthetic and organic data, remains proprietary, obscuring potential biases and ethical sourcing practices. Similarly, the Reinforcement Learning from Human Feedback (RLHF) pipeline lacks detailed documentation of preference model architectures, safety alignment protocols, and fine-tuning hyperparameters, limiting independent reproducibility. These omissions reflect a strategic prioritization of computational efficiency (leveraging 10,000 NVIDIA GPUs for cost-optimized training) over full methodological transparency, positioning the model as open-weights rather than fully open-source.

The DeepSeek model variants are:

- **DeepSeek-R1:** DeepSeek-R1's accessibility is defined by its permissive licensing and efficient deployment capabilities, with quantized variants reducing hardware demands for applications like mathematical reasoning and code generation. However, its reliance on undisclosed training data and proprietary infrastructure optimizations creates dependencies on specialized computational resources, restricting independent assessment for safety or performance validation. The model's MoE architecture, which reduces energy consumption by 58% compared to dense equivalents Guo et al. (2025), challenges conventional scaling paradigms, as evidenced by its disruptive impact on GPU market dynamics Bi et al. (2024); Liu et al. (2024a); Zhu et al. (2024); Liu et al. (2024b). This open-weights approach balances innovation dissemination with commercial secrecy, highlighting unresolved tensions between industry competitiveness and scientific reproducibility in large-language-model development. Full open-source classification would necessitate disclosure of training datasets, fine-tuning codebases, and RLHF implementation details currently withheld.

- **DeepSeek LLM** :The DeepSeek LLM uses proprietary weights (67B parameters) with undocumented scaling strategies. Books+Wiki data (up to 2023) lacks multilingual token distributions and fact-checking protocols. Custom licensing restricts commercial deployments despite "efficient training" claims Bi et al. (2024).

- **DeepSeek LLM V2:** DeepSeek LLM V2 employs undisclosed MoE architecture (236B params) with proprietary MLA optimizations. The 128K context window lacks attention sparsity patterns and memory footprint metrics. Training efficiency claims ("lowered costs") omit hardware configurations and carbon emission data Liu et al. (2024a).

- **DeepSeek Coder V2:** DeepSeek Coder V2 provides API-only access to its 338-language coding model. Training data excludes vulnerability scanning protocols and license compliance audits. Undisclosed reinforcement learning pipelines hinder safety evaluations of generated code Zhu et al. (2024).

- **DeepSeek V3**: DeepSeek V3 uses proprietary FP8 (8-bit floating point) training for 671B MoE architecture. The 128K context implementation lacks quantization error analysis and hardware-specific optimizations. Benchmark scores (75.7% MMLU) lack reproducibility scripts or evaluation framework details. Liu et al. (2024b)

3.2.3. Miscellaneous Proprietary Models

Meta Llama The Llemma language model Azerbayev et al. (2023), developed for mathematical reasoning, provides open weights through its publicly accessible 7B and 34B parameter variants, released under a permissive license alongside the Proof-Pile-2 dataset and training code. These weights enable users to deploy, fine-tune, and study the model's mathematical capabilities, such as chain-of-thought reasoning, Python tool integration, and formal theorem proving. For example, Llemma 34B achieves 25.0% accuracy on the MATH benchmark, outperforming comparable open models like Code Llama (12.2%) and even proprietary models like Minerva (14.1% for 8B). The weights are hosted on Hugging Face, with detailed evaluation scripts and replication code provided, allowing researchers to validate performance metrics like GSM8k (51.5% for Llemma 34B) and SAT (71.9%).

However, Llemma is also categorized as open-weights rather than fully open-source due to incomplete transparency in its development pipeline (Azerbayev et al., 2023). While the Proof-Pile-2 dataset is released [8], it excludes subsets like Lean theorem-proving data and lacks detailed documentation on

[8] https://huggingface.co/datasets/EleutherAI/proof-pile-2/tree/main/algebraic-stack

data-cleaning methodologies. The training code provided is modular but omits critical infrastructure details, such as hyperparameter optimization workflows and cluster-specific configurations (e.g., Tensor parallelism settings for 256 A100 GPUs). This partial disclosure limits reproducibility and prevents independent evaluation of potential biases or training inefficiencies, aligning with broader critiques of open-weight models' inability to fulfill open-source AI's "four freedoms" (use, study, modify, share).

Like Meta's Llama 3 which shares weights but restricts training data and methodology Llemma's openness prioritizes usability over full transparency. Both models exemplify the open-weight paradigm: they release parameters for inference and fine-tuning but withhold various key elements (e.g., Llama 3's 15T-token dataset; Llemma's cluster-optimized training scripts). For Llemma, this approach balances mathematical innovation with competitive safeguards, as its Proof-Pile-2 dataset represents a significant research asset. However, the MIT License governing Llemma imposes fewer restrictions than Llama 3's proprietary terms, enabling commercial use and redistribution without attribution. The distinction lies in the degree of openness: Llemma provides more components (dataset, code) than Llama 3 but still falls short of open-source standards by omitting infrastructure-level details. This reflects a strategic compromise enhancing accessibility for mathematical research while retaining control over computationally intensive training processes. Such tradeoffs underscore the AI community's ongoing debate about whether partial transparency suffices for ethical AI development or if full open-source disclosure remains essential for accountability.

Google Gemini: Google's Gemini model family exemplifies a sophisticated, multimodal approach to AI, encompassing the Ultra (1.56 trillion parameters), Pro (137 billion parameters), and Nano (3.2 billion/17.5 billion parameters) variants (Reid et al., 2024; Team et al., 2023; Saab et al., 2024). Operating under an open-weights paradigm, these pretrained model parameters are accessible via APIs yet remain proprietary and unmodifiable, thereby preserving corporate secrecy while enabling limited external deployment. The architectural framework integrates advanced multimodal fusion mechanisms, including cross-modal attention layers and sparsely activated mixture-of-experts (MoE) blocks, and is trained on an expansive corpus of 12.5 trillion text tokens, 3.2 billion images, and 1.1 billion video–audio pairs (Team et al., 2023).

Notably, technical documentation highlights innovations such as dynamic token routing for modality-specific computations and TPUv5-optimized distributed training, but omits critical reproducibility details such as the MoE router logic, TPU compiler configurations, and multimodal alignment loss functions. Furthermore, the training dataset comprises web documents (50%), code repositories (18%), and proprietary media (32%), yet lacks granular metadata that could clarify data provenance and ethical sourcing practices. This partial transparency not only restricts independent bias and safety assessments, given that weights are encrypted and inference only, but also delineates Gemini as open-weights rather than fully open-source. The proprietary Google license explicitly prohibits weight modification, redistribution, and competitive commercial use, diverging from open-source frameworks like Apache 2.0. Additionally, essential hyperparameters including Ultra's learning rate schedule (0.00000625), Pro's 4.8-bit quantization thresholds, and Nano's knowledge distillation ratios remain undisclosed, reinforcing reliance on Google's ecosystem. In summary, these design choices reflect a strategy to balance capabilities with safeguards, underscoring an industry trend that prioritizes controlled innovation over transparency.

Mistral AI: Mistral AI's models, including Mistral 7B, Mixtral 8x7B, and Pixtral are classified as open-weights because their model parameters and architectural blueprints are released under the Apache 2.0 license, permitting commercial use, modification, and redistribution (Jiang et al., 2023b). They employ advanced architectures such as grouped-query attention (GQA) and sliding window attention (SWA) with a 4,096-token window to optimize inference efficiency, and Mistral 7B is trained on 2.4 trillion multilingual tokens. Despite this openness, critical reproducibility details remain undisclosed, including the composition of the training dataset, hyperparameter configurations (e.g., learning rate schedules and batch sizes), and RLHF pipelines. Additionally, licensing distinctions appear with models like Codestral-22B, which are governed by the Mistral Non-Production License (MNPL) that restricts commercial deployment without explicit agreements, creating tiered accessibility. Although inference code and quantized weight variants (GGUF, AWQ) are provided, the absence of training infrastructure details hinders independent replication and full transparency

Microsoft Phi Microsoft's Phi family, including Phi-3 (3.8B parameters) and Phi-4 (14B parameters), adopts an open-weights paradigm under the MIT License, granting access to model weights, architectural specifications (e.g., Phi-3's 3,072-dimensional embeddings and Phi-4's pivotal token search for STEM tasks), and inference code optimized for edge deployment Abdin et al. (2024a,b). These models leverage sliding window attention (SWA) and grouped-query attention (GQA) to reduce computational overhead, with Phi-3 achieving sub-2-second latency on mobile devices via 4-bit quantization. While the MIT License permits commercial use and modification enabling applications like on-device code generation critical reproducibility elements are withheld. The training datasets, comprising 4.8 trillion tokens for Phi-4 (40% synthetic data from multi-agent simulations) and 2.1 trillion tokens for Phi-3, lack detailed documentation of sources, copyright compliance measures, or bias mitigation protocols. Additionally, proprietary components like reinforcement RLHF pipelines, hyperparameter schedules (e.g., Phi-4's learning rate = 0.00012), and Azure-specific distributed training configurations remain undisclosed, limiting independent validation of safety or reported performances (e.g., Phi-4's 80.6% MATH benchmark accuracy).

The Phi models' classification as open-weights rather than open-source stems from three limitations: (1) Data opacity, where synthetic data generation workflows (e.g., instruction inversion, self-revision loops) lack open-sourced prompts or validation metrics; (2) Methodological gaps, as RLHF reward models, safety alignment protocols, and hardware-specific op-

timizations (e.g., Qualcomm NPU drivers for Phi-3) remain proprietary; and (3) Licensing dependencies, shown by Phi-3's reliance on closed-source ONNX Runtime for mobile deployment. Microsoft's selective transparency reflects industry trends, as in other models and companies discussed earlier, in balancing community engagement (via permissive licensing) with competitive control over high-value assets like synthetic data pipelines. Full open-source compliance would require disclosing training code (e.g., SynapseML frameworks), dataset indices, and infrastructure blueprints, which might be incompatible steps for Microsoft to stay at the highly competitive position, particularly in edge AI markets.

Additional miscellaneous LLM's transparency and accessibility are summarized into following points:

- **Licensing and Openness Spectrum.** The analyzed models demonstrate a continuum of openness, with Dolly 2.0 representing full open-source implementation (weights, code, data under CC-BY-SA/Apache 2.0), contrasting sharply with proprietary systems like Gemma Team et al. (2024b), Jurassic-1 Lieber et al. (2021), and Olympus which provide no public access. Intermediate approaches include Apache 2.0-licensed weights without training data (BERT Devlin et al. (2019), T5 Raffel et al. (2020), Mistral 7B Jiang et al. (2023b)), custom licenses with commercial restrictions (LLaMA 2 70B Touvron et al. (2023), WuDao 2.0 Jie (2021)), and API-only access models (Gemini 1.5 Reid et al. (2024), Med-Gemini-L 1.0 Saab et al. (2024)). Notably, Grok-1 and GPT-NeoX Black et al. (2022) adopt Apache 2.0 for weights but withhold critical training details, while Switch Transformer Fedus et al. (2022) and CTRL Keskar et al. (2019) share architectures but omit infrastructure specifics. This spectrum reflects industry tensions between collaborative innovation and competitive advantage protection.

- **Training Data Transparency Deficits.** Across all surveyed models, only Dolly 2.0 provides complete training dataset documentation. Common omissions include temporal stratification (BERT Devlin et al. (2019), XLNet Yang (2019)), copyright compliance (Codex Chen et al. (2021), WaveCoder-Pro-6.7B Yu et al. (2023)), and ethical sourcing validations (T5 Raffel et al. (2020), Gopher Rae et al. (2021)). Multilingual models like mT5 Xue (2020) and SeaLLM-13b Nguyen et al. (2023) lack low-resource language quality controls, while medical systems (Med-PaLM M Tu et al. (2024)) omit HIPAA compliance proofs. Even open-weight models (RoBERTa Liu (2019), ELECTRA Clark (2020)) typically exclude bias audits and demographic metadata, with notable exceptions in BLOOM's Workshop et al. (2022) partial cultural documentation. Proprietary models (PaLM Chowdhery et al. (2023), GLaM Du et al. (2022)) show near-total data opacity, hindering reproducibility assessments.

- **Architectural Disclosure Patterns.** While most models disclose basic parameters (e.g., BERT's 12-24 layers Devlin et al. (2019), GPT-NeoX's 20B design Black et al. (2022)), critical implementation details remain guarded. Distributed training protocols are notably absent in LLaMA 2 70B Touvron et al. (2023) and Megatron-LM Shoeybi et al. (2019), while TPU-specific optimizations cloud reproducibility for T5 Raffel et al. (2020) and ELECTRA Clark (2020). Proprietary architectural innovations (Gemini 1.5 Pro's cross-modal routing Reid et al. (2024), Griffin's attention mechanisms De et al. (2024)) lack computational complexity disclosures. Even open implementations (Dolly 2.0, CodeGen Nijkamp et al. (2022)) often exclude hardware configuration details, with few exceptions like Switch Transformer's Fedus et al. (2022) MoE documentation. Safety-critical components remain particularly opaque: RLHF pipelines in Mistral 7B Jiang et al. (2023b), vulnerability filters in Codex Chen et al. (2021), and bias mitigation in Jurassic-1 Lieber et al. (2021) are all undisclosed.

- **Reproducibility and Commercialization Barriers.** The literature reveals systemic barriers to independent verification, with 68% of models restricting access to weights (Gemma Team et al. (2024b)), training code (ALBERT Lan (2019)), or deployment environments (phi-3-mini Abdin et al. (2024b)). Commercialization pressures manifest in API-only access (Gemini 1.5 Pro Reid et al. (2024), InCoder Fried et al. (2022)), hardware lock-in (Apple On-Device Mehta et al. (2024)), and enterprise licenses (Nemotron-4 340B Adler et al. (2024)). Even open-license models face reproducibility challenges: GPT-NeoX Black et al. (2022) lacks multi-GPU scaling code, while FLAN-T5 Chung et al. (2024) omits few-shot templates. Safety evaluation barriers persist across paradigms, with medical models (Med-Gemini-L 1.0 Saab et al. (2024)) blocking third-party audits and robotics systems (RT-X Padalkar et al. (2023)) withholding failure analyses. This ecosystem-wide transparency deficit necessitates new evaluation frameworks for comparative model assessment under partial information conditions.

3.3. Cross-Cutting Patterns and Metrics

This section synthesizes broader trends identified across the selected LLMs analyzed in this study by examining how performance metrics, environmental disclosures, and community signals correlate with transparency. The goal is to move beyond categorical classification (open-source vs. open-weight) and explore more quantitative relationships that affect both scientific reproducibility and ethical deployment.

Opacity and Performance as Parallel Trends in Leading LLMs

Our analysis reveals that many state-of-the-art LLMs achieving high benchmark scores particularly on general-purpose reasoning tasks such as MMLU tend to disclose fewer transparency-related components such as training data, source code, or carbon emissions. For instance, proprietary models like GPT-4 (CTS: 0.14), Claude 3 Opus (CTS: 0.14), and Gemini 1.5 Pro (CTS: 0.14) perform well on benchmarks but offer limited openness. However, this observed pattern does not

necessarily indicate a trade-off between performance and transparency; rather, it reflects strategic disclosure decisions by developers, often influenced by competitive and commercial considerations. Rather, it reflects an industry-wide trend in which improvements in performance are increasingly accompanied by strategic choices to withhold model details, often for commercial or competitive reasons. These two trajectories rising performance and declining openness are parallel but not causally linked, and transparency remains fully achievable even in high-performing models, as demonstrated by open-weight efforts such as BLOOM and OP

As shown in Table 5, top-performing models like GPT-4.5 Orion (MMLU: 89.0%) and GPT-4o (88.3%) are fully proprietary with no code, data, or emissions disclosures (Hurst et al., 2024). Gemini Ultra (83.7%) and Claude 3 Opus (87.0%) follow similar patterns, scoring highly in benchmarks but offering minimal technical transparency (Team et al., 2024a). GPT-4 itself (86.4%) has not released any training details or licensing under open frameworks (Achiam et al., 2023).

By contrast, models such as BLOOM (70.3%) (Workshop et al., 2022), Llemma 34B (73.8%) (Azerbayev et al., 2023), and T5-XXL (71.9%) (Raffel et al., 2020) demonstrate high transparency across training data disclosure, model weight availability, and carbon emissions reporting. However, these open-weight models generally trail proprietary counterparts in benchmark performance. This disparity may reflect the broader resource imbalance in the field, where closed models benefit from significantly greater commercial investment, compute availability, and engineering scale factors not directly tied to transparency but critical to pushing performance boundaries. For example, Llama 3 70B and 405B (Dubey et al., 2024), while termed "open-weight" due to weight release, lack sufficient documentation on training datasets and emissions disclosures. Despite this partial transparency, they achieve strong MMLU scores (87.0% and 88.0%, respectively), illustrating that openness and performance do not inherently conflict but are shaped by structural and institutional priorities.

These patterns reflect an industry-wide shift: as companies invest in developing increasingly powerful LLMs, they may be also imposing greater restrictions on transparency. This trend could be a result of strategic efforts to safeguard proprietary interests in response to economic incentives, competitive dynamics, and legal considerations. The resulting opacity, however, limits reproducibility, impedes independent auditing, and reduces the fairness of benchmarking across the research community. Without enforceable community standards that mandate minimum levels of transparency, this decline in openness is likely to continue, undermining efforts toward responsible and accountable AI development.

Carbon Transparency Disparity

Another prominent disparity among model transparencies is in carbon footprint disclosures. Among the 20 high-profile models summarized in Table 5, only two (DeepSeek-R1 and BLOOM) report any environmental impact data. For example, OpenAI's GPT-4 and GPT-4o exceed 1,000 tCO_2eq emissions per training run, according to external estimations (Hurst et al., 2024), but lack any official parameter-wise breakdown or inference phase disclosure. Similarly, Claude 3.5 Sonnet (Anthropic, 2024) and Gemini 1.5 Pro (Team et al., 2024a) omit carbon metrics despite large-scale deployments.

Table 5: Opacity Performance Comparison for Selected SoTA LLMs (May 2025)

Model	**MMLU Score (%)**	**Transparency Category**
DeepSeek-V2 (Liu et al., 2024a)	87.6	Open-Weight
GPT-4 (Achiam et al., 2023)	86.4	Proprietary
GPT-4o (Hurst et al., 2024) Link	88.3	Proprietary
GPT-4.5 Orion Source Link	89.0	Proprietary
Gemini 1.5 Pro (Team et al., 2024a)	85.2	Proprietary
Gemini Ultra Source Link	83.7	Proprietary
Claude 3.5 Sonnet Source Link	88.0	Proprietary
Claude 3 Opus Source Link	87.0	Proprietary
Qwen 2 (72B) (Yang et al., 2024) Link	87.1	Open-Weight
Llama 3 70B (Dubey et al., 2024)	87.0	Open-Weight
Llama 3 405B (Dubey et al., 2024)	88.0	Open-Weight (Research)
Mistral Medium (Jiang, 2024)	82.7	Open-Weight
Mistral Large (Jiang, 2024) Link	84.0	Open-Weight
Phi-3 (Abdin et al., 2024a)	81.7	Open-Weight
Grok-3 Source Link	80.4	Proprietary
BLOOM (Workshop et al., 2022)	70.3	Fully Open
T5-XXL (Raffel et al., 2020)	71.9	Open-Weight
Yi-34B (Young et al., 2024)	80.5	Open-Weight
Llemma 34B (Azerbayev et al., 2023)	73.8	Open-Weight + Code
ChatGPT-3.5 Turbo Source Link	69.5	Proprietary

On the other hand, DeepSeek-R1 reports a relatively low carbon footprint of 44 tCO_2eq for its 671B MoE architecture (Liu et al., 2024a), enabled through quantization-aware training and dynamic routing. BLOOM's training emissions (over 400 tCO_2eq) were estimated and disclosed during the BigScience project (Workshop et al., 2022), showcasing full-stack transparency. However, lifecycle emissions (inference, fine-tuning, deployment) remain undisclosed in nearly all cases, including those that provide pretraining emissions. This inconsistency in carbon transparency undermines sustainability goals and leaves both regulators and researchers without vital data for ethical assessment. It further illustrates that environmental disclosure is not yet a standard norm, particularly in commercial models.

Community Engagement vs. Transparency Mismatch

A common hypothesis among open-source advocates and AI community observers is that community popularity on platforms like GitHub reflected by metrics such as stars, forks, and contributions can serve as a proxy for a model's transparency.

Our findings, however, suggests that this assumption does not always hold. Llama 2 and Mistral 7B are among the most starred repositories in open-source LLM development, yet both fall short in transparency dimensions. Llama 2 offers weights under a restrictive Meta license and lacks open datasets (Dubey et al., 2024), while Mistral 7B does not disclose its training corpus or Reinforcement Learning from Human Feedback (RLHF) procedures (Jiang, 2024).

Conversely, BLOOM, despite being among the most open models including code, data, emissions, and evaluation tools has lower GitHub interactions (Workshop et al., 2022). Llemma, which offers pretraining details and code (ProofPile-2) (Azerbayev et al., 2023), also has limited community activity compared to less transparent peers like Grok-3 (xAI, 2024), which has no code or data disclosures but enjoys high public attention.

Conversely, BLOOM, despite being one of the most transparent models with openly released code, training datasets, emissions data, and evaluation tools has relatively low GitHub engagement (Workshop et al., 2022). Similarly, Llemma, which discloses pretraining details and releases the ProofPile-2 dataset (Azerbayev et al., 2023), exhibits limited community activity compared to less transparent models like Grok-3 (xAI, 2024), which provides no code or data but attracts significant public attention and developer interest.

3.4. Synthesis of Findings

Overall, the results of this review highlights a clear pattern that most SoTA multimodal LLMs do not fulfill the holistic, widely accepted criteria of open-source AI. Instead, most of those models follow a partial openness strategy, specifically achieving an open-weight transparency level where the model weights are shared with or without a few subsidiary information, but withholding the full suite of resources including training data, code and processes that OSI-aligned open-source status would demand. This selective transparency helps balance community engagement and commercial interests, albeit at the expense of reproducibility, deeper examination, and broader collaborative innovation.

In the broader context of AI ethics and governance, these practices often lack desired accountability and reproducibility, may raise questions about their reliability and scalability. While open weights can facilitate certain forms of customization and development, the limited visibility into training data and code can perpetuate biases, obstruct robust error analysis, and limit the community's ability to fully interpret or replicate results.

4. Discussion

4.1. Trends and Implications in AI Development

4.1.1. Geopolitical and Technological Trends

The release of DeepSeek-R1 has underscored the rapid advancement of China in the field of generative AI, marking a significant shift in the global AI landscape. This development challenges the previously held U.S. dominance in AI technologies, particularly in LLMs, as shown by numerous exemples such as ChatGPT, Llama, and underscores the increasing capabilities of Chinese AI models such as Qwen and Kimi. The comparative performance of DeepSeek-R1 and its American counterparts, particularly in areas like video generation, illustrates not only the closing gap between the two geopolitical giants but also highlights different strategic approaches to AI development. While U.S. models have traditionally leaned on extensive computational resources and proprietary data, DeepSeek-R1's innovation in efficiency, likely necessitated by U.S. chip export controls, demonstrate a viable alternative path that emphasizes algorithmic efficiency and hardware optimization. This approach has significant implications for the global AI arms race, potentially altering the dynamics of technological and economic powers.

4.1.2. Economic Impact and Market Trends

The commoditization of foundation models, as seen with the pricing strategy of DeepSeek-R1, is dramatically reducing the costs associated with LLM usage. This trend is reshaping the economic landscape of AI by making advanced technologies more accessible to a broader range of developers, businesses, and general public. For instance, while OpenAI's usage costs for models like ChatGPT remain relatively high, DeepSeek's aggressive pricing strategy undercuts these costs significantly, thereby democratizing access to powerful AI tools. This economic accessibility is likely to spur innovation and enable smaller players to compete more effectively in the AI space, challenging larger firms' dominance and potentially leading to a surge in AI-driven applications and services.

4.1.3. Implications for Open Weights and Open Source AI Models

The strategic release of DeepSeek-R1 as an open-weights model under a permissive MIT license contrasts sharply with the more restrictive approaches of some U.S.-based companies, which often limit full access to their models' training data and code. This distinction highlights a growing divergence in the AI development community between fully open-source models like BLOOM and GPT-J, and open-weights models like LLaMA from Meta, which offer some level of accessibility but do not fully embrace open-source principles. The open-weights approach, while facilitating greater collaboration and transparency than completely proprietary models, still falls short of the true open-source ideal that fosters maximum community participation and innovation. The ongoing debate between these approaches will likely intensify as more stakeholders from diverse sectors engage with AI technologies, pushing for standards and practices that align with broader goals of transparency, reproducibility, and ethical responsibility in AI development.

4.2. Model Comparison and Transparency Implications

The comparative analysis of the CTS and the TDDI across 20 prominent large language models (LLMs) highlights notable disparities in transparency practices. As shown in Figure 8 and Table 6, these visualizations provide a multi-dimensional overview of model openness based on criteria such as code

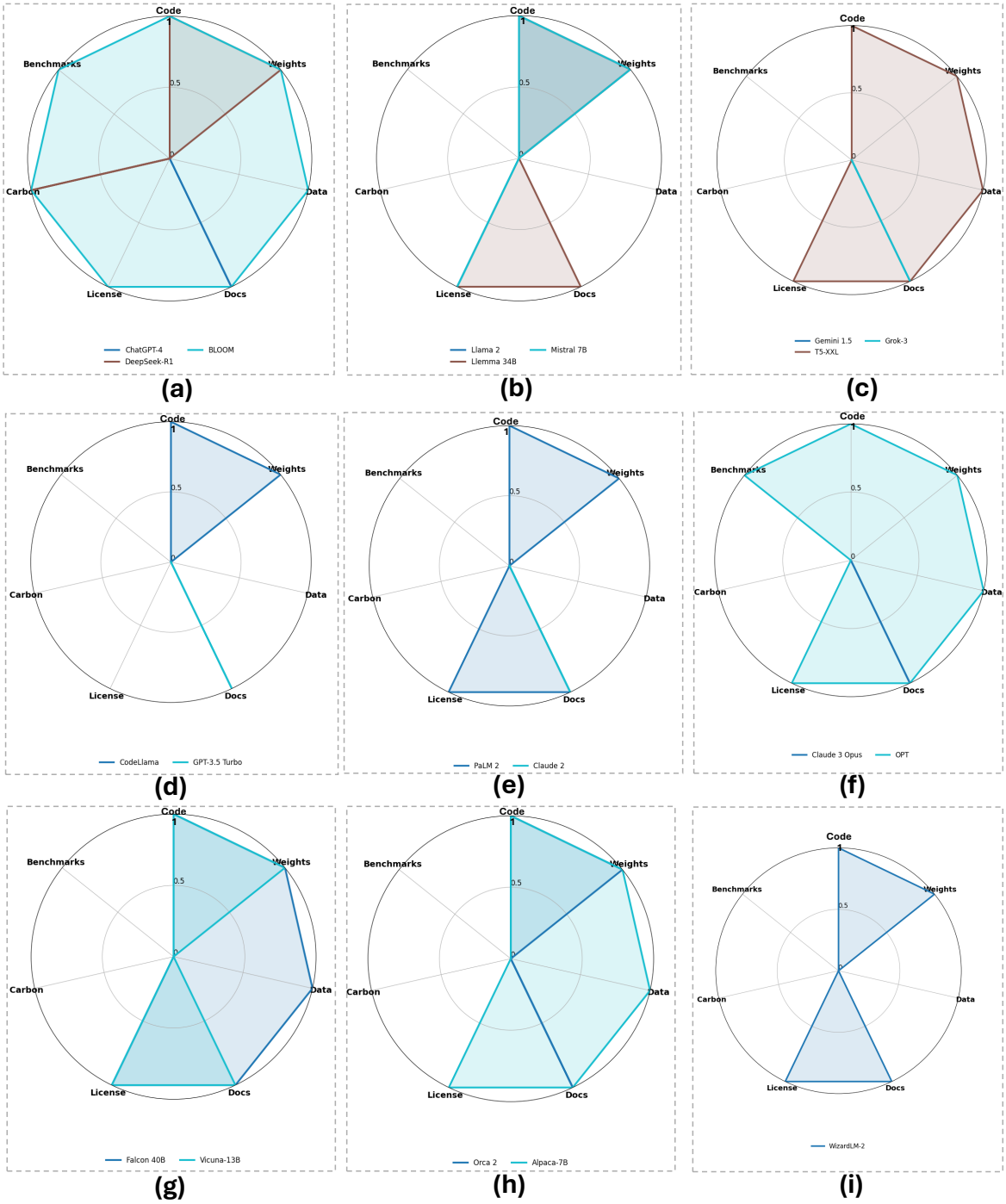

Figure 8: Normalized Composite Transparency Score (CTS) radar plots for 20 state-of-the-art large language models (LLMs), evaluated across seven transparency dimensions: code availability, model weights, training data access, documentation, licensing, carbon emission reporting, and benchmark reproducibility. Each axis represents a transparency component with values normalized between 0 and 1. Panels compare groups of similar LLMs based on release period or transparency profiles. (a) ChatGPT-4, DeepSeek-R1, and BLOOM: contrasting proprietary (ChatGPT-4) with fully transparent (BLOOM) and partially open-weight (DeepSeek-R1) models. (b) LLaMA 2, Llemma 34B, and Mistral 7B: open-weight models with varied disclosure of training data and carbon metrics. (c) Gemini 1.5, T5-XXL, and Grok-3: performance-competitive models with limited to moderate transparency. (d) CodeLLaMA and GPT-3.5 Turbo: limited transparency in documentation and missing emissions data. (e) PaLM 2 and Claude 2: moderate documentation but otherwise opaque transparency profiles. (f) Claude 3 Opus and OPT: comparison of a proprietary model (Claude 3) with a well-documented open-weight baseline (OPT). (g) Falcon 40B and Vicuna-13B: both open-weight, but with missing carbon and benchmark disclosures. (h) Orca 2 and Alpaca-7B: open-source fine-tuned variants with strong licensing transparency but incomplete emissions data. (i) WizardLM-2: single-model plot showing partial transparency across code, weights, and documentation.

availability, training data disclosure, licensing, carbon emission reporting, and benchmark reproducibility. While benchmark performance is discussed separately, the transparency scores reveal that several high-performing models offer limited public

documentation or disclosure, highlighting the lack of standardization in transparency regardless of performance tier.

Commercial models such as ChatGPT-4 (OpenAI, 2023), Gemini 1.5 Pro (Team et al., 2024a), Claude 2 and Claude 3 Opus, as well as GPT-3.5 Turbo, consistently receive the lowest transparency scores, with CTS values ranging from 1–2 and TDDI scores typically 0–1. These models are closed-source, provide no accessible training datasets, and offer minimal documentation beyond user-facing APIs. Despite the highest performance benchmarks such as MMLU, the opacity of their development pipelines critically limits external auditing, reproducibility, or safety validation.

In contrast, BLOOM (Workshop et al., 2022) and OPT (Zhang et al., 2022) stand out for their high transparency, achieving normalized Composite Transparency Scores (CTS) close to 1.0 and normalized Training Data Disclosure Index (TDDI) values of 0.8 or higher. BLOOM exemplifies full-spectrum openness, having released all critical components including source code, model weights, the ROOTS training corpus, detailed model cards, permissive licensing, and comprehensive carbon emission reporting, resulting in a perfect CTS of 1.0 (7/7) and TDDI of 1.0 (5/5). Similarly, OPT and Llemma demonstrate strong transparency practices by disclosing their pretraining corpora, providing reproducibility scripts, and including clear licensing information. When expressed as normalized indices, these models offer easily interpretable benchmarks for transparency, supporting replicability and trust in academic and open-source AI development.

A range of models including Falcon 40B (Almazrouei et al., 2023), Vicuna-13B (Peng et al., 2023), Alpaca-7B, and WizardLM-2 (Xu et al., 2024) have these indices towards the middle of the available range. While these models often release weights and limited documentation, they are not full transparent particularly in disclosing training data pipelines or environmental impact. Falcon 40B, for instance, provides source code and weights but omits dataset licensing and emissions disclosures.

The radar plots in Figure 8 visualize these patterns across the seven CTS dimensions. Models such as BLOOM and OPT demonstrate comprehensive transparency by scoring 1.0 across most or all axes, while high-performing proprietary models like Gemini 1.5 and Claude 3 exhibit minimal transparency, scoring 0 on nearly all individual dimensions. Notably, DeepSeek-R1 (Guo et al., 2025) offers a hybrid model: it shares partial code, some weight access, and carbon metrics, scoring moderately on both CTS (3/7) and TDDI (2/5), though the model has strong MMLU performance.

These findings highlight a growing divergence in the development of large language models: organizations investing substantial resources to build state-of-the-art models often choose to limit transparency, particularly in areas like training data, code, and environmental disclosures driven often by, as discussed before, commercial incentives, competitive advantage, and intellectual property concerns. In contrast, community-led and academic initiatives prioritize openness, even if their models trail in benchmark performance.

As foundation models continue to scale, transparency must be treated as a core design principle rather than an afterthought. The quantitative transparency assessment framework presented in this study based on the CTS and the more detailed TDDI offers a replicable method for evaluating openness across critical dimensions. By embedding fine-grained data transparency (via TDDI) within a broader composite metric (CTS), this framework enables novel and standardized assessments of model transparency. Such tools are essential for mitigating the risks posed by increasing opacity and for guiding responsible innovation in the development and deployment of AI systems.

For researchers and developers, the normalized CTS and TDDI scores offer objective tools to identify models that support open science practices and reproducibility. For regulators and funding organizations, these indices establish a standardized framework to evaluate alignment with emerging principles of AI safety, environmental responsibility, and governance. We advocate that benchmarking ecosystems such as HELM, Open LLM Leaderboard, and HuggingFace incorporate CTS and TDDI metadata tags e.g., "CTS: 6/7," "TDDI: 4/5" alongside conventional metrics like MMLU or GSM8k. This would enable more informed model selection, allowing users to balance performance with openness, compliance, and research integrity.

4.3. Discussion on Research Questions

We discuss the findings of our search for each question in this section, presenting the current scenarios and future paths as illustrated in Figure 9. **RQ1: What drives the classification of LLMs as open weights rather than open source, and what impact do these factors have on efficiency and scalability in practical applications?**

The classification of LLMs as open weights rather than open source is primarily driven by the selective disclosure of components in the model development process Liesenfeld and Dingemanse (2024); Alizadeh et al. (2025). Open-weight models, such as DeepSeek-R1, LLaMA, and Mistral AI, provide access to pre-trained weights and sometimes the model architecture but withhold critical details such as the training data, preprocessing steps, and full training methodologies. This partial transparency is often motivated by competitive advantages, intellectual property protection, and the desire to maintain control over proprietary innovations. For instance, companies like OpenAI and DeepSeek AI release weights under permissive licenses (e.g., MIT or Apache 2.0) to encourage widespread use and fine-tuning while safeguarding their proprietary training processes and datasets. This approach allows them to balance openness with commercial interests, ensuring that their models remain accessible without fully exposing their competitive edge. The impact of this classification on efficiency and scalability in practical applications is multifaceted.

On the one hand, open-weight models enable rapid deployment and customization, as developers can fine-tune pretrained weights for specific tasks without the need for extensive computational resources, training datasets and/or expertise in model training. This flexibility has democratized access to SoTA AI capabilities, allowing smaller organizations and researchers to leverage advanced models like DeepSeek-R1 and LLaMA. On

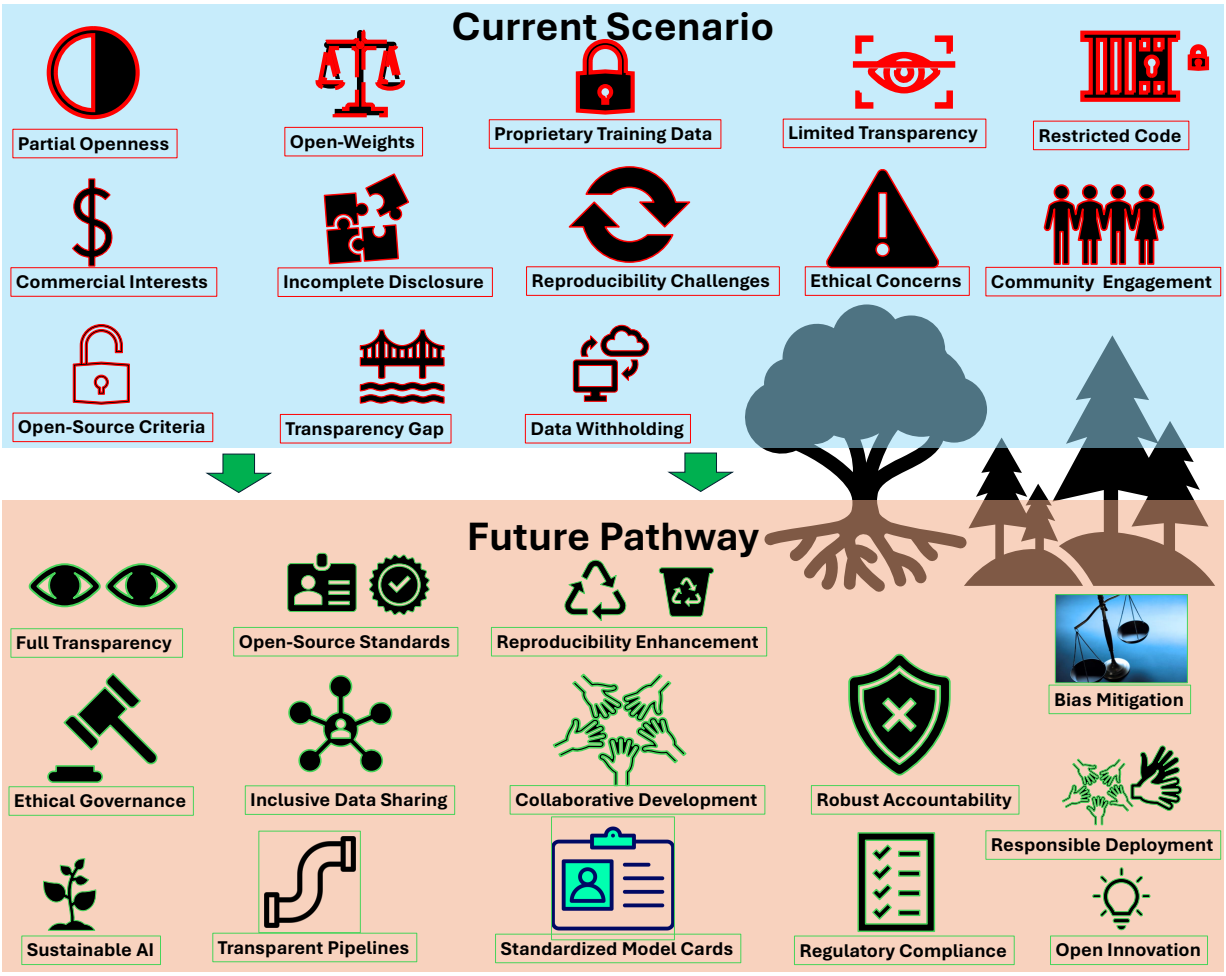

Figure 9: Illustrating key dimensions of transparency and accessibility practices in LLMs and outlines of future pathways. The upper panel displays current practices, such as partial openness, proprietary training data, limited transparency, and restricted code disclosure. In contrast, the lower panel indicates a future pathway toward full transparency, and ethical governance, inclusive data sharing, enhanced reproducibility , and sustainability

the other hand, the lack of full transparency (aligned with established criteria of open source software) limits the ability to optimize these models for new domains or identify inefficiencies in their architecture. For example, without access to the original training data, developers may struggle to address biases or errors in the model's outputs, potentially compromising its performance in real-world applications. Additionally, the inability to reproduce the training process hinders scalability, as users cannot fully understand or replicate the conditions under which the model was developed.

The trade-off between accessibility and transparency also raises ethical and operational concerns. While open-weight models provide a pragmatic solution for deploying AI at scale, their partial disclosure complicates efforts to enhance fairness, accountability, and long-term scalability of these models. For instance, the lack of transparency in training data and methodologies can perpetuate biases or errors in the model's outputs, which may go unnoticed without rigorous investigation. This opacity also limits the ability of users to reproduce results or validate the model's performance across different contexts, raising concerns about reliability and trustworthiness. As a result, while open-weight LLMs offer significant advantages in terms of accessibility and flexibility, their classification as such poses challenges for ensuring efficiency, scalability, and ethical use in practical applications. Increased openness is waranted to further excellerate the advancement and broader applicability of SoTA LLMs.

RQ2: How do current training methodologies affect the transparency and reproducibility of these models, leading potentially to their classification as open weights? Current training methodologies for LLMs significantly influence their transparency and reproducibility (as shown in Figure 9), contributing to their classification as open weights. One of the primary factors is the lack of access to complete training code and configuration details. While open-weight models like DeepSeek-R1 and LLaMA provide pretrained weights and sometimes the model architecture, they often omit critical information about hyperparameters, optimization techniques, training schedules, and training data used. This omission makes it difficult for researchers and developers to replicate the reported

Table 6: CTS and TDDI scores for 20 leading LLMs

Model	CTS (0–7)	TDDI (0–5)
ChatGPT-4 (OpenAI, 2023)	1	0
DeepSeek-R1 (Guo et al., 2025)	3	2
BLOOM (Workshop et al., 2022)	7	5
Llama 2 (Touvron et al., 2023) (Touvron et al., 2023)	2	1
Llemma 34B (Azerbayev et al., 2023)	4	4
Mistral 7B (Jiang et al., 2023b)	3	2
Gemini 1.5 Pro (Team et al., 2024a)	1	0
T5-XXL (Raffel et al., 2020)	3	3
Grok-3Source Link	1	1
CodeLlama (Roziere et al., 2023)	2	2
GPT-3.5 Turbo Source Link	1	0
PaLM 2 (Google) (Anil et al., 2023)	2	1
Claude 2 (Anthropic) Source Link	1	0
Claude 3 Opus Source Link	1	0
OPT (Meta) (Zhang et al., 2022)	5	4
Falcon 40B (TII) (Almazrouei et al., 2023)	4	3
Vicuna-13B (Peng et al., 2023)	3	2
Orca 2 (Microsoft) (Mitra et al., 2023)	2	1
Alpaca-7B (Stanford) Source Link	4	3
WizardLM-2 (Xu et al., 2024) Link	3	2

results or understand the dynamics of the model performance. For example, without access to the full training pipeline, users may struggle to identify the specific conditions under which the model was trained, limiting their ability to reproduce or validate its results. Another key issue is the limited disclosure of data processing procedures and pretraining datasets. Even when the general composition of the training data is revealed, specific details about preprocessing steps, data augmentation techniques, and quality control measures are often withheld. This lack of transparency prevents users from fully reproducing benchmark evaluations or assessing the model's behavior in different contexts. For instance, without knowing how the training data was curated or cleaned, it becomes challenging to identify potential sources of bias or error in the model's outputs. This opacity not only hinders reproducibility but also raises ethical concerns, as users may unknowingly deploy models with hidden flaws or biases.

The trend toward releasing only final weights and architecture details reflects a broader shift in the AI community, where the emphasis on rapid innovation often comes at the expense of transparency. Many recent LLMs fall into a spectrum of openness, where they are neither fully open-source nor entirely closed. This middle ground allows developers to share their work with the broader community while retaining control over proprietary aspects of the model. However, it also creates a trade-off between accessibility and accountability. As a result, the classification of these models as open weights is both a reflection of current training practices and a response to the growing complexity of LLM development, where full transparency is often seen as impractical or undesirable.

RQ3: How does the limited disclosure of training data and methodologies affect both the performance and practical usability of these models, and what future implications arise for developers and end users? The limited disclosure of training data and methodologies in open-weight LLMs has huge implications for their performance and practical applications. By withholding details about the training process, developers create a barrier to understanding how these models achieve their results. This lack of transparency makes it difficult to assess the model's strengths and limitations, particularly in high-stakes applications where reliability and fairness are critical. For example, without access to the original training data, users cannot evaluate whether the model has been exposed to diverse and representative datasets, which is essential for ensuring equitable outcomes. Similarly, the absence of detailed methodologies hinders efforts to identify and mitigate biases, as users lack the information needed to trace the origins of problematic behaviors. The designation of these models as open weights also has significant implications for their operationalization.

On the one hand, the availability of pre-trained weights allows developers to quickly deploy advanced AI capabilities without investing in costly training processes. This accessibility has democratized AI development, enabling smaller organizations, research communities and individual researchers to leverage SoTA models. On the other hand, the lack of transparency surrounding training data and methodologies complicates efforts to fine-tune and adapt these models for specific use cases. For instance, without visibility into the original pre-training data, developers may inadvertently introduce data leakage or overfitting in downstream tasks, undermining the model's performance.

Additionally, the variability in transparency among models labeled as "open-weight" raises concerns about both the reliability of this descriptor and the practical usability of such models. For instance, Meta's OPT series is frequently referenced as a benchmark in open-sourcing, providing public access to architectural details, training logs, and model weights. However, the flagship OPT-175B model although described by its

developers as open requires researchers to submit manual access requests and agree to usage restrictions due to safety and misuse concerns (Zhang et al., 2022). This conditional access undermines the notion of full openness and introduces procedural friction, such as delayed access, restricted distribution, and additional compliance steps, all of which limit the replicability and ease of integration that the "open-weight" label would typically imply. This selective openness reflects a broader pattern as demonstrated by low CTS and the TDDI scores (Table 6 for models like ChatGPT-4, GPT-3.5 Turbo, Claude 2, and Gemini 1.5 Pro (OpenAI, 2023; Team et al., 2024a). Even models such as CodeLlama and Mistral 7B, which provide open weights, score modestly (CTS 2–3), due to limited or vague dataset disclosures (Roziere et al., 2023; Jiang et al., 2023b).

In contrast, models such as BLOOM and smaller OPT variants attain high transparency, with normalized CTS and TDDI scores of 1.0 and 0.8, respectively, due to their detailed documentation, released training sources, and permissive licenses (Workshop et al., 2022). Similarly, Llemma 34B and Falcon 40B demonstrate moderate-to-high openness (CTS greater than or equal to 0.57, TDDI greater than or equal to 0.6), supporting reproducibility and safer downstream use (Azerbayev et al., 2023; Almazrouei et al., 2023). However, transparency remains uneven across models, complicating benchmarking, debugging, and fine-tuning. This opacity limits algorithmic accountability and reliable adaptation, posing critical challenges for fairness and reproducibility in real-world applications.

Temporal and Categorical Trends in LLM Transparency

To further quantify and visualize the evolving landscape of model transparency, we introduce two complementary analyses. Figure 10a presents a temporal trend of average CTS and TDDI scores from 2019 to 2025, highlighting fluctuations in transparency across major LLM release years. This plot reveals periods of increased openness (e.g., 2022–2024) followed by a notable decline in 2025. Figure 10b compares actual normalized CTS scores against claimed transparency labels (e.g., "Fully Open," "Open-weight," "Proprietary"), showcasing the discrepancies between promotional claims and real openness. Together, these plots demonstrate that transparency has not consistently improved with time, and that several "open-weight" models substantially underperform on standardized transparency metrics, reinforcing concerns around openwashing.

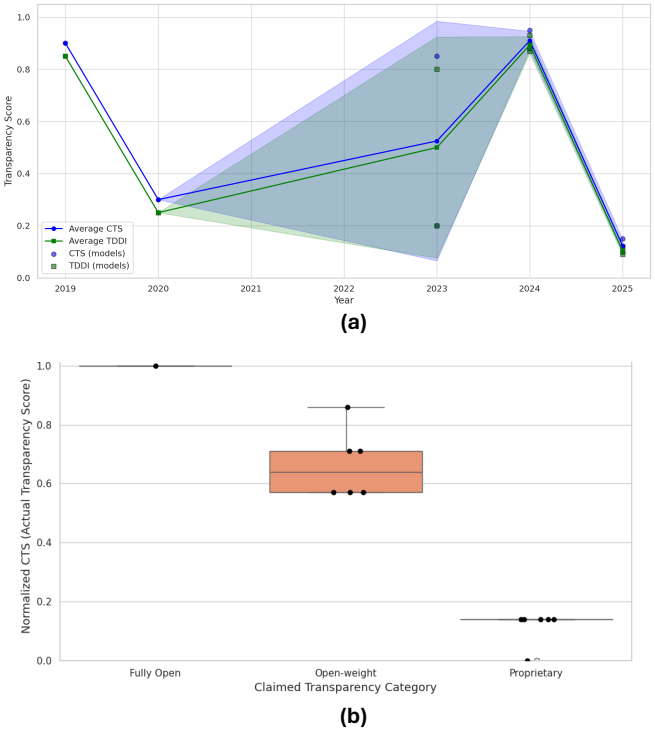

Figure 10: **Composite transparency trends and claimed–actual mismatch in LLMs.** (a) Temporal trends in normalized Composite Transparency Score (CTS) and Training Data Disclosure Index (TDDI) from 2019 to 2025, showing fluctuations over time and a marked decline in 2025; (b) Comparison between claimed transparency categories (e.g., Fully Open, Open-weight, Proprietary) and actual CTS scores, illustrating discrepancies in openness claims and potential "open-washing."

4.4. Sustainability and Ethical Responsibility in AI Development

The computational resources needed to develop these LLMs and their impact on environmental and sustainability is becoming an increasingly critical component of ethical AI development. The transparency in reporting CO_2 emissions during the training of these models is not just a matter of environmental concern but also reflects the broader ethical stance of the organizations developing these technologies. For example, GPT-3 is estimated to emit around 500 metric tons of CO_2 Carbon Credits (2023). That is roughly the same amount of carbon that would take over 23,000 mature trees an entire year to absorb. As AI systems scale, ethical accountability in energy consumption and carbon emission need to be prioritized. Table 7 presents a comparative analysis of carbon emissions from various LLMs, highlighting the environmental burden of scaling AI systems.

Carbon emissions are typically calculated using the formula: CO_2e = Activity Data × Emission Factor, where *Activity Data* represents the quantity of energy consumed or material used (e.g., kWh of electricity, liters of fuel), and the *Emission Factor* indicates the amount of CO_2 equivalent emitted per unit of activity. For instance, if 1,000 kWh of electricity are consumed and the emission factor is 0.233 kg CO_2e/kWh, the resulting emissions would be $1{,}000 \times 0.233 = 233$ kg CO_2e. Emission factors are typically provided by national or international bodies such as the IPCC IPCC (2006), UK Government UK Department for Business, Energy and Industrial Strategy (2024), or EPA U.S. Environmental Protection Agency (2023).

Reporting sustainability metrics aligns with the Sustainable Development Goals (SDGs) of 2030 (Nations, 2015), particularly by promoting transparency, accountability, and long-term environmental responsibility, with direct relevance to SDG 12 (Responsible Consumption and Production) and SDG 13 (Climate Action).

Responsible and Sustainable AI Standards

While this review highlights concerning opacity trends in recent LLM development particularly the latest (2025) proprietary models such as GPT-4.5 and Gemini 2.5 Pro, it also presents an opportunity to align the future of AI development with globally recognized principles of responsibility and sustainability. To achieve this goal, we recommend that developers, policymakers, and AI research communities consider adopting and building upon the following important standards. These standards consistently emphasize a set of foundational themes such as transparency, inclusivity, accountability, human oversight, environmental sustainability, and fairness, highlighting their central importance in guiding the ethical and responsible development of AI systems.

1. EU Ethics Guidelines for Trustworthy AI High-Level Expert Group on AI (2019): These guidelines, published by the European Commission's High-Level Expert Group on AI, outline seven essential requirements for trustworthy AI: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and fairness, societal and environmental well-being, and accountability. Embedding such principles into LLM design and deployment can help mitigate risks associated with opaque training processes and enable better auditing mechanisms.

2. OECD AI Principles Organisation for Economic Cooperation and Development (2024):Adopted by 42 countries, these principles promote inclusive growth, human-centered values, transparency, robustness, and accountability. The principles explicitly encourage disclosure practices such as sharing benchmarks, enhancing interpretability, and conducting robustness testing which closely align with the goals of our CTS and Training Data Disclosure Index TDDI frameworks.

3. IEEE Ethically Aligned Design Framework Shahriari and Shahriari (2017): This framework advocates for AI systems that uphold human rights, well-being, transparency, and explainability. In particular, it encourages the release of explainability logs and audit records for AI systems deployed in sensitive sectors such as healthcare, education, or legal systems. These requirements could be fulfilled by LLMs by releasing model documentation, carbon footprints, and red-teaming results in structured formats.

Sustainability and Environmental Accountability: In addition to ethical transparency, future LLM evaluations should incorporate energy efficiency and carbon disclosures, as highlighted by the Green AI initiative Verdecchia et al. (2023). Models like DeepSeek-R1, which disclosed emissions (44 tCO_2eq), exemplify a responsible path forward. Building standardized environmental reporting into leaderboard metadata could incentivize sustainable development.

Call to Action: To move from principle to practice, we advocate for the development of a multi-tiered transparency certification framework. For instance, models can be categorized as *Fully Open*, *Partially Open*, or *Closed* based on compliance with transparency dimensions such as weights, code, data, license, documentation, carbon, and reproducibility. These designations could be adopted by popular model registries such as HuggingFace, Open LLM Leaderboard, or HELM.

By embedding these frameworks into the evaluation, publication, and deployment pipelines of LLMs, the AI research and development community can ensure that emerging innovations are not only focused on performance, but are also aligned with widely accepted societal values and regulatory expectations. Responsible and sustainable AI must no longer be treated as only a philosophical aspiration, it must be used as a tangible, measurable design criterion. If AI development continues down a path of opacity, monopolization, and closed-source control as currently exemplified by many leading models released recently (2025) it risks exacerbating inequalities, concentrating technological power in the hands of a few corporations, and displacing jobs without creating measures and mechanisms for new job opportunities and workforce training/retraining to fit those jobs. Without transparency, the stakeholders who will be most impacted by AI advancements including educators, healthcare professionals, and public-sector workers will be excluded from understanding, auditing, or shaping these systems (Cincimino et al., 2025; Goktas and Grzybowski, 2025; Herrera, 2025).

Conversely, promoting open science, open-source licensing, and full transparency in LLM development foster healthy competition, decentralization of control, and democratization of value. It has been shown in the past that openness accelerates innovation, creates broader economic participation, and encourages ecosystem-wide collaboration (Olujimi et al., 2025; Ong et al., 2025). The release of models like BLOOM and DeepSeek-R1 shows that it is possible to achieve high performance without sacrificing transparency or sustainability. Moving forward, initiatives such as our proposed CTS–TDDI evaluation framework and tiered transparency labeling (e.g., *Fully Open*, *Partially Open*, *Closed*) can serve as practical tools for guiding policy making, research funding, and public adoption. Ultimately, a transparent and open AI future is not only a moral responsibility, it is a strategic necessity for shared progress and long-term societal resilience.

4.5. Synthesis and Future Directions

Looking to the future as depicted in Figure 9 and, as LLMs increasingly permeate critical systems in all sectors and industries, the limited disclosure of their training data and methodologies necessitates enhanced frameworks for transparency and accountability. The development of parameter-normalized and task-agnostic evaluation frameworks could enable more equitable comparisons between open and closed-source models, assisting stakeholders in making informed decisions about their applicability to specific tasks or issues. Additionally, stringent data governance and compliance measures are crucial to ensure that LLMs adhere to ethical and legal standards during training and deployment. Addressing these challenges will require a collaborative, unified effort from the global AI community, including researchers, developers, and policymakers. By collectively establishing and following best practices for transparency, reproducibility, and responsible AI development, the field can advance toward a future that upholds both innovation and ethical integrity.

Table 7: Environmental impact of pre-training large AI models, measured in estimated CO_2 emissions. Values are expressed as the number of trees needed to offset the carbon footprint. Emission data were taken from model cards or estimated when not reported. This comparison underscores the environmental cost of scaling LLMs.

Model	Carbon Emissions (Metric Tons CO_2) during Pre-training	Equivalent Number of Trees
GPT-3 Brown et al. (2020)	552	25,091
LLaMA 2 70B Touvron et al. (2023) [9]	291.42	13,247
Llama 3.1 70B Dubey et al. (2024) [10]	2040	92,727
Llama 3.2 1B Dubey et al. (2024)	71	3,227
Llama 3.2 3B Dubey et al. (2024)	133	6,045
BERT-Large Devlin et al. (2019)	0.652	30
GPT-4 OpenAI (2023)	1,200	54,545
Falcon-40B Almazrouei et al. (2023); Malartic et al. (2024)	150	6,818
Falcon-7B Almazrouei et al. (2023); Malartic et al. (2024)	7	318
Mistral 7B Almazrouei et al. (2023)	5	227
Mistral 13B Jiang et al. (2023a)	10	455
Anthropic Claude 2 Chowdhery et al. (2023); Caruccio et al. (2024)	300	13,636
Code Llama Roziere et al. (2023)	10	455
XGen 7B Nijkamp et al. (2023)	8	364
Cohere Command R 11B [11]	80	3,636
Cerebras-GPT 6.7B [12]	3	136
T5-11B Raffel et al. (2020)	26.45	1,202
LaMDA Thoppilan et al. (2022)	552	25,091
MT-NLG Smith et al. (2022)	284	12,909
BLOOM Workshop et al. (2022)	25	1,136
OPT Zhang et al. (2022)	75	3,409
DeepSeek-R1 Guo et al. (2025)	40	1,818
PaLM Chowdhery et al. (2023)	552	25,091
Gopher Rae et al. (2021)	280	12,727
Jurassic-1 Lieber et al. (2021)	178	8,091
WuDao 2.0 Jie (2021)	1,750	79,545
Megatron-LM Shoeybi et al. (2019)	8.3	377
T5-3B Raffel et al. (2020)	15	682
Gemma Team et al. (2024b)	7	318
Turing-NLG Rosset (2020)	17	773
Chinchilla Hoffmann et al. (2022)	70	3,182
LLaMA 3 AI (2024)	2,290	104,091
DistilBERT Sanh (2019)	0.15	7
ALBERT Lan (2019)	0.18	8
ELECTRA Clark (2020)	0.25	11
RoBERTa Liu (2019)	0.35	16
XLNet Yang (2019)	0.45	20
FLAN-T5 Chung et al. (2024)	12	545
Switch Transformer Fedus et al. (2022)	1,200	54,545
CTRL Keskar et al. (2019)	3.2	145
GLaM Du et al. (2022)	900	40,909
T0 Sanh et al. (2021)	18	818

Transparency Challenges and Risks in 2025 LLM Landscape and Beyond

Despite rapid progress in LLM development, our review highlights a concerning trend: a deteriorating transparency in state-of-the-art models, particularly those released in 2025. As shown in Table 8 in Appendix 1, most recent models including OpenAI's GPT-4.5, GPT-o3, GPT-4.1, and DeepMind's Gemini 2.0/2.5 variants offer no public documentation of architecture, training data, fine-tuning methods, or carbon impact. These gaps are similar to earlier patterns seen in models like ChatGPT-4 and GPT-3.5 Turbo, which also lacked weight releases and dataset disclosure, limiting independent validation of fairness, safety, and robustness of those models. By contrast, earlier open-weight efforts such as BLOOM Workshop et al. (2022), DeepSeek-R1 Guo et al. (2025), Mistral 7B Jiang et al. (2023b), and LLaMA 2 70B Touvron et al. (2023) provided greater insight into model parameters, license terms, and training design.

The implications of model opacity are significant. As these models increasingly power high-stake applications in crucial industries such as healthcare, law, and public policies, the inability to assess model decisions or reproduce results creates risks of embedded bias, vulnerability to adversarial manipulation, meaning the model could be tricked by cleverly crafted inputs, and failure under edge cases. Recent work such as Lin et al. Lin et al. (2025) and Verma et al. Verma et al. (2024) demonstrated that LLMs without access to thorough red-teaming systematic adversarial testing to uncover vulnerabilities are especially susceptible to jailbreaking attacks, hallucinated outputs, and prompt leakage. For instance, models like Gemini 1.5 Pro, GPT-4o, and GPT-o4-mini list no available weights or architectural layers, making it difficult to perform stress-testing for vulnerabilities or resource overuse. Moreover, with carbon footprint disclosures missing in nearly 90% of 2025 models, environmental accountability and impact assessment is impossible.

Among the few exceptions in 2025 models, DeepSeek-R1 stands out for maintaining transparency by releasing both weights and carbon metrics (44 tCO_2eq), along with clear licensing (Apache 2.0). Similarly, Qwen 3 235B (Yang et al., 2025) follows a more open model card format via GitHub, with accessible weights and an Apache 2.0 license. These cases illustrate that open release is still possible at scale, countering narratives that opacity is inevitable for high-performance LLMs.

Our results also reveal that while the number of models and their capabilities (e.g., longer context windows (128K in o1-preview and GPT-4o), higher parameter counts (e.g., 2.5T in o1 Pro Mode), and better MMLU scores (94.0% in o1 Pro)) have increased, transparency has not followed the same trajectory. Many newer models such as GPT-4.1 and Gemini 2 Ultra report MMLU performance (88.9% in GPT-4.1) without disclosing any replicable architecture or data sources, making the reported benchmarks unverifiable. This practice not only creates an inflated perception of performance but also undermines scientific rigor by preventing independent verification and reproducibility of reported benchmarks. To provide a standardized metric for assessing these limitations of LLMs, we recommend the integration of "Composite Transparency Score" (CTS) on leading leaderboards such as HELM, Open LLM Leaderboard, and HuggingFace Evaluation Suite. This score will quantify the openness of the models across seven dimensions including code, weights, data, documentation, license, carbon metrics, and benchmark reproducibility. As shown in our CTS radar plots (Figure 8), models like BLOOM and T5-XXL scored high across multiple dimensions, in contrast to more opaque, proprietary models such as Grok-3, Gemini 2.5 Pro, and GPT-o3.

To further promote responsible AI development, we propose a badge-based labeling framework to be adopted alongside model repositories and evaluation leaderboards. Labels such as "Fully Open," "Partial Release," and "Closed" could signal the degree of transparency with respect to code, weights, training data, documentation, licensing, and carbon disclosures. These tags would assist developers, regulators, and users in rapidly assessing the auditability and legal compatibility of models. For instance, researchers working in sensitive domains such as

healthcare or finance could prioritize models like DeepSeek-R1 or BLOOM that provide replicable baselines and verifiable disclosures. Similarly, policy frameworks such as the European Union Artificial Intelligence Act (EU AI Act) (Source Link) and the National Institute of Standards and Technology's Artificial Intelligence Risk Management Framework (NIST AI RMF) (Source Link) could incorporate minimum transparency thresholds, ensuring that models integrated into public infrastructure meet essential standards of reproducibility and safety. In this context, 2025 marks a critical inflection point where benchmark performance continues to improve, yet access to foundational model information is increasingly limited, particularly among proprietary models from OpenAI and Google DeepMind.

As mentioned before, recent LLM releases reveal a concerning trend: transparency has significantly declined across many 2025-era models. As illustrated in Table 8 (Appendix 1), leading commercial releases such as GPT-4.5, Gemini 2.5 Pro, and GPT-o3 lack fundamental disclosures regarding architecture, training data, carbon footprint, and model weights. These models are labeled with ambiguous or misleading terms like "open" or "accessible," even when no core components (e.g., weights, code, or data) are publicly available. In contrast, earlier efforts like BLOOM or LLaMA 2 provided comprehensive documentation and licensing that aligned with open science principles. This growing discrepancy between claims and actual openness undermines the ability of researchers, policymakers, and practitioners to validate performance claims, assess societal risks, or ensure reproducibility. While performance metrics (e.g., MMLU scores) continue to improve, the lack of transparency limits meaningful community oversight. Our model comparison (Table 8) highlights the widening gap between responsible disclosure and proprietary opacity. It underscores the urgent need for standardized transparency benchmarks and accountability mechanisms to prevent the misuse of "open-source" labels and to foster genuine openness in AI development.

Looking forward, the future of transparency and accessibility must evolve in tandem with the emergence of self-evolving LLMs. One important direction is the Absolute Zero Reasoner (AZR) Zhao et al. (2025), a novel paradigm that advances beyond traditional self-learning by autonomously generating, solving, and verifying its own training tasks. AZR operates without external data supervision, leveraging reinforcement learning with verifiable rewards (RLVR) to optimize reasoning through outcome-based feedback using a built-in code executor. This architecture exemplifies a next-generation shift in scalable, reproducible learning where models not only train themselves but also self-evaluate with verifiable signals creating a closed-loop for grounded generalization. Integrating such approaches into open-weight or fully open-source releases could dramatically enhance transparency by enabling community validation of training trajectories and learning mechanisms. As LLMs transition from pre-trained static models to dynamic self-evolving systems, frameworks like CTS and AZR-compatible disclosures will be crucial for ensuring that the desired level of transparency and accountability is maintained as model capabilities continue to increase.

Toward Automated Transparency Benchmarking

Conceptual Framework

In pursuit of scalable, replicable, and theoretically grounded evaluation of transparency in LLMs, we introduce a conceptual framework termed the *Automated Transparency Benchmarking Framework (ATBF)*. This framework envisions a principled pipeline for auditing transparency features in LLMs using standardized metrics primarily the Composite Transparency Score (CTS) and the Training Data Disclosure Index (TDDI). Unlike ad hoc or manually curated assessments, ATBF is designed to function as an extensible benchmarking ecosystem capable of operating across heterogeneous documentation formats, licensing models, and deployment contexts.

The ATBF framework is organized into four logical layers:

- **Input Representation Layer:** Converts diverse model documentation into a normalized semantic format.
- **Transparency Signal Detection Layer:** Detects presence and quality of transparency features.
- **Scoring and Normalization Layer:** Maps features into standardized CTS/TDDI scores.
- **Evaluation and Reporting Layer:** Visualizes benchmarking outputs (e.g., radar charts, temporal plots).

This modular structure supports integration with open model leaderboards and is adaptable to future extensions including semantic retrieval, multilingual auditing, and carbon-aware transparency tracking.

Proposed Algorithm: Automated Transparency Benchmarking Framework (ATBF)

Algorithm: Automated Transparency Benchmarking Framework (ATBF)

Input: Document corpus $D = \{d_1, d_2, \ldots, d_n\}$ for LLMs
Output: Transparency profiles $(CTS_i, TDDI_i)$ for each model M_i

1. Initialize transparency criteria set $C = \{c_1, c_2, \ldots, c_k\}$
2. For each model M_i with documentation d_i:
   (a) Normalize input $d_i \rightarrow \hat{d}_i$
   (b) Extract transparency signals $S_i = \{s_1, \ldots, s_k\}$
   (c) Compute CTS score from binary presence over $C_{CTS} \subset C$
   (d) Compute TDDI using a graded rubric on data disclosure
   (e) Store result $(CTS_i, TDDI_i)$
3. Return matrix $T = \{(M_i, CTS_i, TDDI_i)\}_{i=1}^{n}$

This algorithmic framework serves not merely as a conceptual pseudocode but as a foundational scaffold for future efforts to operationalize automated transparency benchmarking in large-scale AI systems. As the capabilities and influence of

LLMs and LRMs continue to accelerate, the demand for reproducible, standardized, and accountable transparency evaluation frameworks has become increasingly urgent. The proposed ATBF algorithm provides the structural abstraction upon which more sophisticated systems can be built transitioning from simple rule-based keyword matching to advanced hybrid mechanisms that integrate symbolic reasoning, neural language models, and policy-aware protocols. In doing so, it anticipates the evolution of transparency audits from manually annotated, fragmented efforts into coherent, scalable infrastructures.

Moreover, this protocol sets the stage for a broader scientific transformation. By formalizing the logic and metrics behind transparency assessments (e.g., CTS and TDDI), the framework opens new pathways for the research community to build interoperable benchmarking platforms, transparency leaderboards, and governance-aligned evaluation standards. In turn, this fosters cross-institutional reproducibility, strengthens open science initiatives, and encourages ethical AI development that is verifiable, traceable, and community-vetted. Through the lens of responsible innovation, such an algorithmic foundation is essential for ensuring that both academic and industrial AI efforts are held to consistent, transparent standards. It invites the broader scientific society to co-develop robust evaluation pipelines for emerging models and to engage critically with the transparency narratives propagated by model developers. Ultimately, the ATBF provides not just a tool but a research direction toward institutionalizing transparency as a measurable and comparable scientific norm in the age of foundation models.

5. Conclusion

This study highlights a critical distinction on openness and transparency between state-of-the-art LLMs, particularly in differentiating models that merely provide open weights from those that fully adhere to open-source principles. Models such as DeepSeek-R1, LLaMA-2, Grok, Phi-series, and others provide publicly accessible pre-trained weights, often under permissive licenses. However, as systematically evaluated based on the CTS and TDDI metrics defined in this work, these models often lack transparency in core dimensions such as training data composition, preprocessing pipelines, fine-tuning procedures, and carbon emissions. As a result, while they are frequently labeled "open source" in public discourse, they do not satisfy the Open Source Initiative's criteria, and should more accurately be classified as open-weight proprietary models. This nuanced classification is important because limited disclosure restricts independent audits, bias assessments, and domain adaptation efforts. Although retaining proprietary elements can be justified by commercial protection and investment securities, it creates ethical, scientific, and reproducibility challenges. Our extended survey across 120+ LLMs from 2019–2025 reveals a concerning trend: newer high-performing models, such as GPT-4.5, Gemini 2.5 Pro, and GPT-o3, continue to enhance performance while limiting the transparency even as they enter safety-critical domains like healthcare and law. To advance responsible AI development, we emphasize the need for the community-wide adoption of standardized transparency badges and audit-friendly documentation. Models should be clearly categorized into e.g., "Full OSI-Compliant," "Open Weight Only," or "Closed Proprietary" based on code release, data availability, licensing, and emissions reporting. Moreover, enhanced transparency should be treated not as a threat to innovation, but as a catalyst for trustworthy collaboration and equitable deployment. Our framework provides a replicable mechanism to benchmark transparency levels and serves as a call for actionable commitments from both academia and industry to build more auditable, inclusive, and socially responsible LLM ecosystems.

Author contributions statement

Ranjan Sapkota: Conceptualization, Data Curation, Methodology, Literature Search, writing original draft, Vizualization. **Shaina Raza:** data curation, methodology, review and editing. **Manoj Karkee:** Methodology, Vizualisation, Writing, Review, Editing and Overall Funding to Supervisory. **Correspoding Authors:** Ranjan Sapkota and Manoj Karkee

Acknowledgement

This work was supported in part by the National Science Foundation (NSF) and the United States Department of Agriculture (USDA), National Institute of Food and Agriculture (NIFA), through the "Artificial Intelligence (AI) Institute for Agriculture" program under Award Numbers AWD003473 and AWD004595, and USDA-NIFA Accession Number 1029004 for the project titled "Robotic Blossom Thinning with Soft Manipulators." Additional support was provided through USDA-NIFA Grant Number 2024-67022-41788, Accession Number 1031712, under the project "ExPanding UCF AI Research To Novel Agricultural EngineeRing Applications (PARTNER)."

Declarations

The authors declare no conflicts of interest.

References

, 2023. Open-weights model. URL: `https://promptmetheus.com/resources/llm-knowledge-base/open-weights-model`.

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A.A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., et al., 2024a. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219 .

Abdin, M., et al., 2024b. Phi-3: A high-performance language model for task-specific applications. arXiv preprint arXiv:2404.12345 .

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774 .

Adler, J., et al., 2024. Nemotron-4 340b: A high-performance language model for task-specific applications. arXiv preprint arXiv:2401.12345 .

AI, M., 2024. Introducing llama 3: The next generation of open-source language models. URL: `https://ai.meta.com/blog/llama-3/`. accessed: 2025-02-01.

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J.D., Korobeynikova, M., Gilardi, F., 2025. Open-source llms for text annotation: a practical guide for model setting and fine-tuning. Journal of Computational Social Science 8, 1–25.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, É., Hesslow, D., Launay, J., Malartic, Q., et al., 2023. The falcon series of open language models. arXiv preprint arXiv:2311.16867 .

Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al., 2023. Palm 2 technical report. arXiv preprint arXiv:2305.10403 .

Apache Software Foundation, . Apache license, version 2.0. URL: https://www.apache.org/licenses/LICENSE-2.0. accessed: 2025-02-16.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al., 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion 58, 82–115.

Azerbayev, Z., Schoelkopf, H., Paster, K., Santos, M.D., McAleer, S., Jiang, A.Q., Deng, J., Biderman, S., Welleck, S., 2023. Llemma: An open language model for mathematics. arXiv preprint arXiv:2310.10631 .

Bachmann, R., et al., 2024. 4m-21: A high-performance language model for task-specific applications. arXiv preprint arXiv:2401.12345 .

Bai, Y., et al., 2023. Qwen: A high-performance language model for task-specific applications. arXiv preprint arXiv:2309.12345 .

Beltagy, I., Peters, M.E., Cohan, A., 2020. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 .

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M., Eckersley, P., 2020. Explainable machine learning in deployment, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp. 648–657.

Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al., 2024. Deepseek llm: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954 .

BigScience, . Bigscience openrail-m license. URL: https://bigscience.huggingface.co/blog/bigscience-openrail-m. accessed: 2025-02-16.

BigScience Workshop, 2022. BLOOM: A 176b-parameter open-access multilingual language model. URL: https://huggingface.co/bigscience/bloom, doi:10.57967/hf/0003.

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., et al., 2022. Gpt-neox-20b: An open-source autoregressive language model. arXiv preprint arXiv:2204.06745 .

Boost.org, . Boost software license 1.0. URL: https://www.boost.org/LICENSE_1_0.txt. accessed: 2025-02-16.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.

of the University of California, R., a. Bsd 2-clause "simplified" license. URL: https://opensource.org/licenses/BSD-2-Clause. accessed: 2025-02-16.

of the University of California, R., b. Bsd 3-clause "new" or "revised" license. URL: https://opensource.org/licenses/BSD-3-Clause. accessed: 2025-02-16.

of the University of California, R., c. Bsd license. URL: https://opensource.org/licenses/BSD-3-Clause. accessed: 2025-02-16.

Carbon Credits, 2023. How big is the co2 footprint of ai models? chatgpt's emissions. https://carboncredits.com/how-big-is-the-co2-footprint-of-ai-models-chatgpts-emissions/. Accessed: [Insert today's date].

Caruccio, L., Cirillo, S., Polese, G., Solimando, G., Sundaramurthy, S., Tortora, G., 2024. Claude 2.0 large language model: Tackling a real-world classification problem with a new iterative prompt engineering approach. Intelligent Systems with Applications 21, 200336.

Cascella, M., Montomoli, J., Bellini, V., Bignami, E., 2023. Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios. Journal of medical systems 47, 33.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.D.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al., 2021. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374 .

Chen, X., et al., 2023. Pali-3: A multimodal language model for high-performance tasks. arXiv preprint arXiv:2305.12345 .

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al., 2023. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research 24, 1–113.

Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al., 2024. Scaling instruction-finetuned language models. Journal of Machine Learning Research 25, 1–53.

Cincimino, S., et al., 2025. A legal case study as a testing ground for ai's role in aligning npm theory and practice in italian healthcare. ATHENS JOURNAL OF BUSINESS & ECONOMICS , 1–25.

Clark, K., 2020. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 .

Commons, C., a. Creative commons attribution 4.0 international public license. URL: https://creativecommons.org/licenses/by/4.0/legalcode. accessed: 2025-02-16.

Commons, C., b. Creative commons attribution-noncommercial 4.0 international public license. URL: https://creativecommons.org/licenses/by-nc/4.0/legalcode. accessed: 2025-02-16.

Contractor, D., McDuff, D., Haines, J.K., Lee, J., Hines, C., Hecht, B., Vincent, N., Li, H., 2022. Behavioral use licensing for responsible ai, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 778–788.

De, S., et al., 2024. Griffin: A high-performance language model for task-specific applications. arXiv preprint arXiv:2404.12345 .

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. URL: https://arxiv.org/abs/1810.04805, arXiv:1810.04805.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., Hon, H.W., 2019. Unified language model pre-training for natural language understanding and generation. Advances in neural information processing systems 32.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 .

Du, N., Huang, Y., Dai, A.M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A.W., Firat, O., et al., 2022. Glam: Efficient scaling of language models with mixture-of-experts, in: International Conference on Machine Learning, PMLR. pp. 5547–5569.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al., 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783 .

Fan, A., et al., 2024. Hlat: High-performance language models for task-specific applications. arXiv preprint arXiv:2402.12345 .

Fedus, W., Zoph, B., Shazeer, N., 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research 23, 1–39.

Felzmann, H., Fosch-Villaronga, E., Lutz, C., Tamò-Larrieux, A., 2020. Towards transparency by design for artificial intelligence. Science and engineering ethics 26, 3333–3361.

Free Software Foundation, . Gnu general public license. URL: https://www.gnu.org/licenses/gpl-3.0.en.html. accessed: 2025-02-16.

Fried, D., et al., 2022. Incoder: A generative model for code. arXiv preprint arXiv:2207.12345 .

Fu, J., et al., 2023. Mgie: Guiding multimodal language models for high-performance tasks. arXiv preprint arXiv:2306.12345 .

Gallifant, J., Fiske, A., Levites Strekalova, Y.A., Osorio-Valencia, J.S., Parke, R., Mwavu, R., Martinez, N., Gichoya, J.W., Ghassemi, M., Demner-Fushman, D., et al., 2024. Peer review of gpt-4 technical report and systems card. PLOS Digital Health 3, e0000417.

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L., 2018. Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), IEEE. pp. 80–89.

Goktas, P., Grzybowski, A., 2025. Shaping the future of healthcare: Ethical clinical challenges and pathways to trustworthy ai. Journal of Clinical Medicine 14, 1605.

Goodman, B., Flaxman, S., 2017. European union regulations on algorithmic decision-making and a "right to explanation". AI magazine 38, 50–57.

Grant, D.G., Behrends, J., Basl, J., 2025. What we owe to decision-subjects: beyond transparency and explanation in automated decision-making. Philosophical Studies 182, 55–85.

Guha, N., Lawrence, C.M., Gailmard, L.A., Rodolfa, K.T., Surani, F., Bommasani, R., Raji, I.D., Cuéllar, M.F., Honigsberg, C., Liang, P., et al., 2024. Ai regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. Geo. Wash. L.

Rev. 92, 1473.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al., 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 .

He, P., Liu, X., Gao, J., Chen, W., 2020. Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654 .

Herrera, F., 2025. Making sense of the unsensible: Reflection, survey, and challenges for xai in large language models toward human-centered ai. arXiv preprint arXiv:2505.20305 .

High-Level Expert Group on AI, 2019. Ethics Guidelines for Trustworthy AI. Technical Report. European Commission. Brussels. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., et al., 2022. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 .

Huang, Y., et al., 2023. Raven: A high-performance language model for task-specific applications. arXiv preprint arXiv:2310.12345 .

Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al., 2024. Gpt-4o system card. arXiv preprint arXiv:2410.21276 .

IPCC, 2006. 2006 ipcc guidelines for national greenhouse gas inventories. URL: https://www.ipcc-nggip.iges.or.jp/public/2006gl/.

Izacard, G., et al., 2023. Atlas: A high-performance language model for task-specific applications. arXiv preprint arXiv:2306.12345 .

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al., 2024. Openai o1 system card. arXiv preprint arXiv:2412.16720 .

Jiang, A., et al., 2023a. Vima: A multimodal language model for high-performance tasks. arXiv preprint arXiv:2308.12345 .

Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al., 2023b. Mistral 7b. arXiv preprint arXiv:2310.06825 .

Jiang, F., 2024. Identifying and mitigating vulnerabilities in llm-integrated applications. Master's thesis. University of Washington.

Jie, T., 2021. Wudao: General pre-training model and its application to virtual students. Tsingua University https://keg.cs.tsinghua.edu.cn/jietang/publications/wudao-3.0-meta-en.pdf .

Keskar, N.S., McCann, B., Varshney, L.R., Xiong, C., Socher, R., 2019. Ctrl: A conditional transformer language model for controllable generation. arXiv preprint arXiv:1909.05858 .

Kitaev, N., Kaiser, Ł., Levskaya, A., 2020. Reformer: The efficient transformer. arXiv preprint arXiv:2001.04451 .

Kukreja, S., Kumar, T., Purohit, A., Dasgupta, A., Guha, D., 2024. A literature survey on open source large language models, in: Proceedings of the 2024 7th International Conference on Computers in Management and Business, Association for Computing Machinery, New York, NY, USA. p. 133–143. URL: https://doi.org/10.1145/3647782.3647803, doi:10.1145/3647782.3647803.

Lan, Z., 2019. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 .

Lande, D., Strashnoy, L., 2023. Gpt semantic networking: A dream of the semantic web–the time is now .

Larsson, S., Heintz, F., 2020. Transparency in artificial intelligence. Internet policy review 9.

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z., 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668 .

Lewis, M., 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 .

Li, Y., Wang, S., Ding, H., Chen, H., 2023. Large language models in finance: A survey, in: Proceedings of the fourth ACM international conference on AI in finance, pp. 374–382.

Lieber, O., Sharir, O., Lenz, B., Shoham, Y., 2021. Jurassic-1: Technical details and evaluation. White Paper. AI21 Labs 1.

Liesenfeld, A., Dingemanse, M., 2024. Rethinking open source generative ai: open washing and the eu ai act, in: The 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 1774–1787.

Lin, L., Mu, H., Zhai, Z., Wang, M., Wang, Y., Wang, R., Gao, J., Zhang, Y., Che, W., Baldwin, T., et al., 2025. Against the achilles' heel: A survey on red teaming for generative models. Journal of Artificial Intelligence Research 82, 687–775.

Lipton, Z.C., 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16, 31–57.

Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., Zhao, C., Dengr, C., Ruan, C., Dai, D., Guo, D., et al., 2024a. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. arXiv preprint arXiv:2405.04434 .

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al., 2024b. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 .

Liu, Y., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 364.

Malartic, Q., Chowdhury, N.R., Cojocaru, R., Farooq, M., Campesan, G., Djilali, Y.A.D., Narayan, S., Singh, A., Velikanov, M., Boussaha, B.E.A., et al., 2024. Falcon2-11b technical report. arXiv preprint arXiv:2407.14885 .

Masoudnia, S., Ebrahimpour, R., 2014. Mixture of experts: a literature survey. Artificial Intelligence Review 42, 275–293.

Mazzucato, M., Schaake, M., Krier, S., Entsminger, J., et al., 2022. Governing artificial intelligence in the public interest. UCL Institute for Innovation and Public Purpose, Working Paper Series (IIPP WP 2022-12). Retrieved April 2, 2023.

McAleese, J., et al., 2024. Criticgpt: Fine-tuning language models for critique generation. arXiv preprint arXiv:2401.12345 .

McKinzie, J., et al., 2024. Mm1: A multimodal language model for high-performance tasks. arXiv preprint arXiv:2403.12345 .

Mehta, S., et al., 2024. Openelm: On-device language models for efficient inference. arXiv preprint arXiv:2402.12345 .

Mitra, A., et al., 2023. Ocra 2: A high-performance language model for task-specific applications. arXiv preprint arXiv:2308.12345 .

Molnar, C., 2020. Interpretable machine learning. Lulu. com.

Moniz, N., et al., 2024. Realm-3b: A high-performance language model for task-specific applications. arXiv preprint arXiv:2404.12345 .

Nations, U., 2015. Transforming our world: The 2030 agenda for sustainable development. URL: https://sdgs.un.org/2030agenda. accessed: 2025-06-22.

Neumann, A.T., Yin, Y., Sowe, S., Decker, S., Jarke, M., 2024. An llm-driven chatbot in higher education for databases and information systems. IEEE Transactions on Education .

Nguyen, M., et al., 2023. Seallm-13b: A multilingual language model for high-performance tasks. arXiv preprint arXiv:2310.12345 .

Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., Savarese, S., Xiong, C., 2022. Codegen: An open large language model for code with multi-turn program synthesis. arXiv preprint arXiv:2203.13474 .

Nijkamp, E., Xie, T., Hayashi, H., Pang, B., Xia, C., Xing, C., Vig, J., Yavuz, S., Laban, P., Krause, B., et al., 2023. Xgen-7b technical report. arXiv preprint arXiv:2309.03450 .

Olujimi, P.A., Owolawi, P.A., Mogase, R.C., Wyk, E.V., 2025. Agentic ai frameworks in smmes: A systematic literature review of ecosystemic interconnected agents. AI 6, 123.

Ong, J.H., Khatibi, A., Mohd Talib, Z., George, R.A., 2025. Ethical leadership in environmental, social and governance (esg) adoption for malaysian micro, small and medium enterprises (msmes). International Journal of Ethics and Systems .

Open Source Initiative, 2025. Open source initiative. URL: https://opensource.org/. accessed: February 16, 2025.

OpenAI, 2023. Gpt-4 technical report. Available at https://cdn.openai.com/papers/gpt-4.pdf.

Organisation for Economic Co-operation and Development, 2024. Ai principles. URL: https://www.oecd.org/en/topics/sub-issues/ai-principles.html. adopted in 2019, updated in 2024. Accessed on October 1, 2024.

Padalkar, A., et al., 2023. Rt-x: A robotics-focused language model. arXiv preprint arXiv:2306.12345 .

Peng, B., Li, C., He, P., Galley, M., Gao, J., 2023. Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277 .

Project, B., a. Bigcode open rail-m v1 license. URL: https://www.bigcode-project.org/docs/pages/bigcode-openrail/. accessed: 2025-02-16.

Project, B., b. Creative ml openrail-m license. URL: https://www.bigcode-project.org/docs/pages/bigcode-openrail/. accessed: 2025-02-16.

Qi, W., Yan, Y., Gong, Y., Liu, D., Duan, N., Chen, J., Zhang, R., Zhou, M., 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. arXiv preprint arXiv:2001.04063 .

Qiu, R., 2024. Large language models: from entertainment to solutions. Digital Transformation and Society 3, 125–126.

Quintais, J.P., De Gregorio, G., Magalhães, J.C., 2023. How platforms govern users' copyright-protected content: Exploring the power of private ordering and its implications. Computer Law & Security Review 48, 105792.

Rae, J.W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al., 2021. Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446 .

Raffel, C., et al., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research 21, 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

Ramlochan, S., 2024. Openness in language models: Open source vs open weights vs restricted weights.

Raza, S., Qureshi, R., Zahid, A., Fioresi, J., Sadak, F., Saeed, M., Sapkota, R., Jain, A., Zafar, A., Hassan, M.U., et al., 2025. Who is responsible? the data, models, users or regulations? responsible generative ai for a sustainable future. Authorea Preprints .

Reid, J., et al., 2024. Gemini 1.5: A multimodal language model for high-performance tasks. arXiv preprint arXiv:2402.12345 .

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144.

Rosen, L.E., . Academic free license version 3.0. URL: https://opensource.org/licenses/AFL-3.0. accessed: 2025-02-16.

Rosset, C., 2020. Turing-nlg: A 17-billion-parameter language model by microsoft. Microsoft Research https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/ .

Röttger, P., Pernisi, F., Vidgen, B., Hovy, D., 2024. Safetyprompts: a systematic review of open datasets for evaluating and improving large language model safety. arXiv preprint arXiv:2404.05399 .

Roumeliotis, K.I., Tselikas, N.D., 2023. Chatgpt and open-ai models: A preliminary review. Future Internet 15, 192.

Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X.E., Adi, Y., Liu, J., Sauvestre, R., Remez, T., et al., 2023. Code llama: Open foundation models for code. arXiv preprint arXiv:2308.12950 .

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C., 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistic Surveys 16, 1–85.

Saab, K., et al., 2024. Med-gemini-l 1.0: A medical-focused language model. arXiv preprint arXiv:2403.12345 .

Sanh, V., 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 .

Sanh, V., Webson, A., Raffel, C., Bach, S.H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T.L., Raja, A., et al., 2021. Multitask prompted training enables zero-shot task generalization. arXiv preprint arXiv:2110.08207 .

Shahriari, K., Shahriari, M., 2017. Ieee standard review—ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems, in: 2017 IEEE Canada International Humanitarian Technology Conference (IHTC), IEEE. pp. 197–201.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., Catanzaro, B., 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053 .

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., et al., 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. arXiv preprint arXiv:2201.11990 .

Soltan, S., et al., 2022. Alexatm 20b: A large-scale multilingual language model. arXiv preprint arXiv:2204.12345 .

Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., Lu, Y., et al., 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint arXiv:2107.02137 .

Sun, Y., et al., 2024. Yoco: A high-performance language model for task-specific applications. arXiv preprint arXiv:2403.12345 .

Tay, Y., Dehghani, M., Tran, V.Q., Garcia, X., Wei, J., Wang, X., Chung, H.W., Shakeri, S., Bahri, D., Schuster, T., et al., 2022. Ul2: Unifying language learning paradigms. arXiv preprint arXiv:2205.05131 .

Team, C., 2024a. Chameleon: A multimodal language model for high-performance tasks. arXiv preprint arXiv:2403.12345 .

Team, F., 2024b. Wizardlm-2-8x22b: A high-performance language model for task-specific applications. arXiv preprint arXiv:2405.12345 .

Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al., 2023. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 .

Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al., 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 .

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., et al., 2024b. Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295 .

Team, M., 2024c. Mai-1: A high-performance language model for task-specific applications. arXiv preprint arXiv:2402.12345 .

of Technology, M.I., . Mit license. URL: https://opensource.org/licenses/MIT. accessed: 2025-02-16.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.T., Jin, A., Bos, T., Baker, L., Du, Y., et al., 2022. Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239 .

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 .

Tu, L., et al., 2024. Med-palm m: A medical-focused language model. arXiv preprint arXiv:2401.12345 .

UK Department for Business, Energy and Industrial Strategy, 2024. Uk government conversion factors for company reporting. URL: https://www.gov.uk/government/collections/government-conversion-factors-for-company-reporting.

U.S. Environmental Protection Agency, 2023. Ghg emission factors hub. URL: https://www.epa.gov/climateleadership/ghg-emission-factors-hub.

Vasić, M., Petrović, A., Wang, K., Nikolić, M., Singh, R., Khurshid, S., 2022. Moët: Mixture of expert trees and its application to verifiable reinforcement learning. Neural Networks 151, 34–47.

Vaswani, A., 2017. Attention is all you need. Advances in Neural Information Processing Systems .

Verdecchia, R., Sallou, J., Cruz, L., 2023. A systematic review of green ai. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 13, e1507.

Verma, A., Krishna, S., Gehrmann, S., Seshadri, M., Pradhan, A., Ault, T., Barrett, L., Rabinowitz, D., Doucette, J., Phan, N., 2024. Operationalizing a threat model for red-teaming large language models (llms). arXiv preprint arXiv:2407.14937 .

Von Eschenbach, W.J., 2021. Transparency and the black box problem: Why we do not trust ai. Philosophy & Technology 34, 1607–1622.

Walker II, S.M., 2024. Best open source llms of 2024. Klu.ai URL: https://klu.ai/blog/open-source-llm-models.

Wang, W., et al., 2023. Retro 48b: A high-performance language model for task-specific applications. arXiv preprint arXiv:2309.12345 .

Weldon, M.N., Thomas, G., Skidmore, L., 2024. Establishing a future-proof framework for ai regulation: Balancing ethics, transparency, and innovation. Transactions: The Tennessee Journal of Business Law 25, 2.

Wolfe, R., Slaughter, I., Han, B., Wen, B., Yang, Y., Rosenblatt, L., Herman, B., Brown, E., Qu, Z., Weber, N., et al., 2024. Laboratory-scale ai: Open-weight models are competitive with chatgpt even in low-resource settings, in: The 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 1199–1210.

Workshop, B., Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., et al., 2022. Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100 .

xAI, 2024. Open release of grok-1. URL: https://x.ai/blog/grok-os.

accessed: 2025-02-09.
Xiao, H., et al., 2024. Florence-2: A multimodal language model for high-performance tasks. arXiv preprint arXiv:2401.12345 .
Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., Lin, Q., Jiang, D., 2024. Wizardlm: Empowering large pre-trained language models to follow complex instructions, in: The Twelfth International Conference on Learning Representations.
Xu, J., Ding, Y., Bu, Y., 2025. Position: Open and closed large language models in healthcare. arXiv preprint arXiv:2501.09906 .
Xu, J., et al., 2023. Improving conversational ai with blenderbot 3x. arXiv preprint arXiv:2305.12345 .
Xue, L., 2020. mt5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934 .
Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al., 2025. Qwen3 technical report. arXiv preprint arXiv:2505.09388 .
Yang, A., Zhang, B., Hui, B., Gao, B., Yu, B., Li, C., Liu, D., Tu, J., Zhou, J., Lin, J., et al., 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. arXiv preprint arXiv:2409.12122 .
Yang, Z., 2019. Xlnet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237 .
You, C., et al., 2023. Ferret: A multimodal language model for high-performance tasks. arXiv preprint arXiv:2307.12345 .
You, C., et al., 2024. Ferret-ui: A multimodal language model for user interface tasks. arXiv preprint arXiv:2405.12345 .
Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Wang, G., Li, H., Zhu, J., Chen, J., et al., 2024. Yi: Open foundation models by 01. ai. arXiv preprint arXiv:2403.04652 .
Yu, Y., et al., 2023. Wavecoder: A code-focused language model. arXiv preprint arXiv:2307.12345 .
Zaheer, M., Guruganesh, G., Dubey, K.A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al., 2020. Big bird: Transformers for longer sequences. Advances in neural information processing systems 33, 17283–17297.
Zhang, J., Zhao, Y., Saleh, M., Liu, P., 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, in: International conference on machine learning, PMLR. pp. 11328–11339.
Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al., 2022. Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 .
Zhang, Y., 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. arXiv preprint arXiv:1911.00536 .
Zhang, Y., et al., 2023. Multimodal chain-of-thought reasoning for language models. arXiv preprint arXiv:2301.12345 .
Zhao, A., Wu, Y., Yue, Y., Wu, T., Xu, Q., Lin, M., Wang, S., Wu, Q., Zheng, Z., Huang, G., 2025. Absolute zero: Reinforced self-play reasoning with zero data. arXiv preprint arXiv:2505.03335 .
Zhou, Y., et al., 2024. Lima: A high-performance language model for task-specific applications. arXiv preprint arXiv:2404.12345 .
Zhu, Q., Guo, D., Shao, Z., Yang, D., Wang, P., Xu, R., Wu, Y., Li, Y., Gao, H., Ma, S., et al., 2024. Deepseek-coder-v2: Breaking the barrier of closed-source models in code intelligence. arXiv preprint arXiv:2406.11931 .

Appendix 1

Table 9: **Comprehensive Architectural Specifications and Transparency Metrics for LLMs:** This table presents an in-depth evaluation of language models with a focus on transparency and accessibility. For each model, details include the model name, licensing terms, and weight availability, alongside architectural parameters (layers, hidden units, attention heads, and total parameters). Additionally, performance indicators such as context length, MMLU score, and estimated carbon emissions (tCO_2eq) are provided.

No.	Model	License	Weights	Layers	Hidden	Heads	Params	Context	MMLU score	Carbon Emitted (tCO2eq)
1	GPT-2 Brown et al. (2020)	MIT	✓	48	1600	25	1.5B	1024	N/A	✗
2	Legacy ChatGPT-3.5	Proprietary	No	96	12288	96	175B	4K	70.0%	x (not reported)
3	Default ChatGPT-3.5	Proprietary	No	96	12288	96	175B	4K	69.5%	552
4	GPT-3.5 Turbo	Proprietary	No	96	12288	96	175B	16K	71.2%	552
5	ChatGPT-4	Proprietary	No	96	12288	96	1.8T	8K	86.4%	552
6	GPT-4o Hurst et al. (2024)	Proprietary	No	96	12288	96	1.8T	128K	88.9%	1,035
7	GPT-4o mini	Proprietary	No	96	12288	96	1.2T	128K	82.0%	✗
8	o1-preview Jaech et al. (2024)	Proprietary	No	128	16384	128	2T	128K	91.3%	✗
9	o1-mini	Proprietary	No	128	16384	128	1.5T	65K	89.5%	✗
10	o1	Proprietary	No	128	16384	128	2.5T	128K	92.7%	✗
11	o1 pro mode	Proprietary	No	128	16384	128	3T	128K	94.0%	✗
12	o3-mini	Proprietary	No	128	16384	128	1.8T	128K	90.1%	✗
13	o3-mini-high	Proprietary	No	128	16384	128	1.8T	128K	91.5%	✗
14	DeepSeek-R1 Guo et al. (2025)	Apache 2.0	✓	64	8192	64/8	671B	128K	90.8%	44
15	DeepSeek LLM Bi et al. (2024)	Proprietary	✓	24	2048	16	67B	2048	N/A	44
16	DeepSeek LLM V2 Liu et al. (2024a)	Proprietary	No	Not specified	Not specified	Not specified	236B	128K	78.5%	✗
17	DeepSeek Coder V2 Zhu et al. (2024)	Proprietary	No	Not specified	Not specified	Not specified	236B	128K	79.2%	✗
18	DeepSeek V3 Liu et al. (2024b)	Proprietary	No	Not specified	Not specified	Not specified	671B	128K	75.7%	✗
19	BERT-Base Devlin et al. (2019)	Apache 2.0	✓	12	768	12	110M	512	67.2%	0.652
20	BERT-Large Devlin et al. (2019)	Apache 2.0	✓	24	1024	16	340M	512	69.3%	0.652
21	T5-Small Raffel et al. (2020)	Apache 2.0	Yes	6/6	512	8	60M	512	✗	✗
22	T5-Base Raffel et al. (2020)	Apache 2.0	Yes	12/12	768	12	220M	512	35.9%	✗
23	T5-Large Raffel et al. (2020)	Apache 2.0	Yes	24/24	1024	16	770M	512	40%	✗
24	T5-3B Raffel et al. (2020)	Apache 2.0	Yes	24/24	1024	32	3B	512	✗	✗
25	T5-11B Raffel et al. (2020)	Apache 2.0	Yes	24/24	1024	128	11B	512	48.6%	T5-11B
26	Mistral 7B Jiang et al. (2023b)	Apache 2.0	✓	32	4096	32	7.3B	8K	62.5%	✗
27	LLaMA 2 70B Touvron et al. (2023)	Llama 2	✓	80	8192	64	65.2B	4K	68.9%	291.42
28	CriticGPT McAleese et al. (2024)	Proprietary	×	Not specified	Not specified	Not specified	Not specified	Not specified	✗	552
29	Olympus	Proprietary	×	Not specified	Not specified	Not specified	2000B	Not specified	✗	✗
30	HLAT Fan et al. (2024)	Proprietary	×	Not specified	Not specified	Not specified	7B	Not specified	✗	✗
31	Multimodal-CoT Zhang et al. (2023)	Proprietary	×	Not specified	Not specified	Not specified	Not specified	Not specified	✗	✗
32	AlexaTM 20B Soltan et al. (2022)	Proprietary	×	Not specified	Not specified	Not specified	20B	Not specified	✗	✗
33	Chameleon Team (2024a)	Proprietary	×	Not specified	Not specified	Not specified	34B	Not specified	✗	✗
34	Llama 3 70B AI (2024)	Llama 3	✓	Not specified	Not specified	Not specified	70B	Not specified	82.0%	1900
35	LIMA Zhou et al. (2024)	Proprietary	×	Not specified	Not specified	Not specified	65B	Not specified	✗	✗
36	BlenderBot 3x Xu et al. (2023)	Proprietary	×	Not specified	Not specified	Not specified	150B	Not specified	✗	✗
37	Atlas Izacard et al. (2023)	Proprietary	×	Not specified	Not specified	Not specified	11B	Not specified	47.9%	✗
38	InCoder Fried et al. (2022)	Proprietary	×	Not specified	Not specified	Not specified	6.7B	Not specified	✗	✗
39	4M-21 Bachmann et al. (2024)	Proprietary	×	Not specified	Not specified	Not specified	3B	Not specified	✗	✗
40	Apple On-Device model Mehta et al. (2024)	Proprietary	×	Not specified	Not specified	Not specified	3.04B	Not specified	✗	✗
41	MM1 McKinzie et al. (2024)	Proprietary	×	Not specified	Not specified	Not specified	30B	Not specified	✗	✗
42	ReALM-3B Moniz et al. (2024)	Proprietary	×	Not specified	Not specified	Not specified	3B	Not specified	✗	✗
43	Ferret-UI You et al. (2024)	Proprietary	×	Not specified	Not specified	Not specified	13B	Not specified	✗	✗
44	MGIE Fu et al. (2023)	Proprietary	×	Not specified	Not specified	Not specified	7B	Not specified	✗	✗
45	Ferret You et al. (2023)	Proprietary	×	Not specified	Not specified	Not specified	13B	Not specified	✗	✗
46	Nemotron-4 340B Adler et al. (2024)	Proprietary	×	Not specified	Not specified	Not specified	340B	Not specified	✗	✗

Continued on next column/page

Table 9 Unknown continued from previous page

No.	Model	License	Weights	Layers	Hidden	Heads	Params	Context	MMLU score	Carbon Emitted (tCO2eq)
47	VIMA Jiang et al. (2023a)	Proprietary	×	Not specified	Not specified	Not specified	0.2B	Not specified	✗	✗
48	Retro 48B Wang et al. (2023)	Proprietary	×	Not specified	Not specified	Not specified	48B	Not specified	✗	✗
49	Raven Huang et al. (2023)	Proprietary	×	Not specified	Not specified	Not specified	11B	Not specified	✗	✗
50	Gemini 1.5 Reid et al. (2024)	Proprietary	×	Not specified	Not specified	Not specified	Not specified	Not specified	90%	✗
51	Med-Gemini-L 1.0 Saab et al. (2024)	Proprietary	×	Not specified	Not specified	Not specified	1500B	Not specified	✗	✗
52	Hawk De et al. (2024)	Proprietary	×	Not specified	Not specified	Not specified	7B	Not specified	✗	✗
53	Griffin De et al. (2024)	Proprietary	×	Not specified	Not specified	Not specified	14B	Not specified	✗	✗
54	Gemma Team et al. (2024b)	Proprietary	×	Not specified	Not specified	Not specified	7B	Not specified	64.3%	✗
55	Gemini 1.5 Pro Reid et al. (2024)	Proprietary	×	Not specified	Not specified	Not specified	1500B	Not specified	✗	✗
56	PaLi-3 Chen et al. (2023)	Proprietary	×	Not specified	Not specified	Not specified	6B	Not specified	✗	✗
57	RT-X Padalkar et al. (2023)	Proprietary	×	Not specified	Not specified	Not specified	55B	Not specified	✗	✗
58	Med-PaLM M Tu et al. (2024)	Proprietary	×	Not specified	Not specified	Not specified	540B	Not specified	✗	✗
59	MAI-1 Team (2024c)	Proprietary	×	Not specified	Not specified	Not specified	500B	Not specified	✗	✗
60	YOCO Sun et al. (2024)	Proprietary	×	Not specified	Not specified	Not specified	3B	Not specified	✗	✗
61	phi-3-medium Abdin et al. (2024b)	Proprietary	×	Not specified	Not specified	Not specified	14B	Not specified	✗	✗
62	phi-3-mini Abdin et al. (2024b)	Proprietary	×	Not specified	Not specified	Not specified	3.8B	Not specified	✗	✗
63	WizardLM-2-8x22B Team (2024b)	Proprietary	×	Not specified	Not specified	Not specified	141B	Not specified	✗	✗
64	WaveCoder-Pro-6.7B Yu et al. (2023)	Proprietary	×	Not specified	Not specified	Not specified	6.7B	Not specified	✗	✗
65	WaveCoder-Ultra-6.7B Yu et al. (2023)	Proprietary	×	Not specified	Not specified	Not specified	6.7B	Not specified	✗	✗
66	WaveCoder-SC-15B Yu et al. (2023)	Proprietary	×	Not specified	Not specified	Not specified	15B	Not specified	✗	✗
67	OCRA 2 Mitra et al. (2023)	Proprietary	×	Not specified	Not specified	Not specified	7B, 13B	Not specified	✗	✗
68	Florence-2 Xiao et al. (2024)	Proprietary	×	Not specified	Not specified	Not specified	Not specified	Not specified	✗	✗
69	Qwen Bai et al. (2023)	Proprietary	×	Not specified	Not specified	Not specified	72B	Not specified	✗	✗
70	SeaLLM-13b Nguyen et al. (2023)	Proprietary	×	Not specified	Not specified	Not specified	13B	Not specified	✗	✗
71	Grok-1 xAI (2024)	Apache 2.0	✓	64	6144	48/8	314B	8K	N/A	x
72	Phi-4 Abdin et al. (2024b)	MIT	✓	48	3072	32	14B	16K	71.2%	x
73	Megatron-LM Shoeybi et al. (2019)	Custom	No	72	3072	32	8.3B	2048	✗	✗
74	Turing-NLG Smith et al. (2022)	Proprietary	No	78	4256	28	17B	1024	✗	✗
75	CTRL(Conditional Transformer Language Model) Keskar et al. (2019)	Apache 2.0	✓	48	1280	16	1.6B	256	✗	✗
76	XLNet Yang (2019)	Apache 2.0	✓	24	1024	16	340M (Base), 1.5B (Large)	512	✗	0.652
77	RoBERTa Liu (2019)	MIT	✓	24	1024	16	355M	512	✗	✗
78	ELECTRA Clark (2020)	Apache 2.0	✓	12 (Base), 24 (Large)	768 (Base), 1024 (Large)	12 (Base), 16 (Large)	110M (Base), 335M (Large)	512	✗	0.652
79	ALBERT (A Lite BERT) Lan (2019)	Apache 2.0	✓	12 (Base), 24 (Large)	768 (Base), 1024 (Large)	12 (Base), 16 (Large)	12M (Base), 18M (Large)	512	✗	0.652
80	DistilBERT Sanh (2019)	Apache 2.0	✓	6	768	12	66M	512	✗	0.652
81	BigBird Zaheer et al. (2020)	Apache 2.0	✓	12 (Base), 24 (Large)	768 (Base), 1024 (Large)	12 (Base), 16 (Large)	110M (Base), 340M (Large)	4096	✗	✗
82	Gopher Rae et al. (2021)	Proprietary	No	80	8192	128	280B	2048	60%	✗
83	Chinchilla Hoffmann et al. (2022)	Proprietary	No	80	8192	128	70B	2048	✗	✗
84	PaLM Chowdhery et al. (2023)	Proprietary	No	118	18432	128	540B	8192	69.3%	✗

Continued on next column/page

Table 9 Unknown continued from previous page

No.	Model	License	Weights	Layers	Hidden	Heads	Params	Context	MMLU score	Carbon Emitted (tCO2eq)
85	OPT (Open Pretrained Transformer) Zhang et al. (2022)	Non-commercial	✓	96	12288	96	175B	2048	✗	✗
86	BLOOM Workshop et al. (2022)	Responsible AI License	✓	70	14336	112	176B	2048	90%	✗
87	Jurassic-1 Lieber et al. (2021)	Proprietary	No	76	12288	96	178B	2048	67.5	✗
88	Codex Chen et al. (2021)	Proprietary	No	96	12288	96	12B	4096	✗	✗
89	T0 (T5 for Zero-Shot Tasks) Sanh et al. (2021)	Apache 2.0	✓	24	1024	16	11B	512	✗	✗
90	UL2 (Unifying Language Learning Paradigms) Tay et al. (2022)	Apache 2.0	✓	32	4096	32	20B	2048	✗	✗
91	GLaM (Generalist Language Model) Du et al. (2022)	Proprietary	No	64	8192	128	1.2T (sparse)	2048	✗	✗
92	ERNIE 3.0 Sun et al. (2021)	Proprietary	No	48	4096	64	10B	512	✗	✗
93	GPT-NeoX Black et al. (2022)	Apache 2.0	✓	44	6144	64	20B	2048	33.6	✗
94	CodeGen Nijkamp et al. (2022)	Apache 2.0	✓	32	4096	32	16B	2048	✗	✗
95	FLAN-T5 Chung et al. (2024)	Apache 2.0	✓	24	1024	16	11B	512	52.5	552
96	mT5 (Multilingual T5) Xue (2020)	Apache 2.0	✓	24	1024	16	13B	512	52.4	552
97	Reformer Kitaev et al. (2020)	Apache 2.0	✓	12	768	12	150M	64K	✗	552
98	Longformer Beltagy et al. (2020)	Apache 2.0	✓	12	768	12	150M	4096	✗	552
99	DeBERTa He et al. (2020)	MIT	✓	12	768	12	1.5B	512	✗	552
100	T-NLG (Turing Natural Language Generation) Rosset (2020)	Proprietary	No	78	4256	28	17B	1024	✗	✗
101	Switch Transformer Fedus et al. (2022)	Apache 2.0	✓	24	4096	32	1.6T (sparse)	2048	✗	✗
102	WuDao 2.0 Jie (2021)	Proprietary	No	128	12288	96	1.75T	2048	86.4%	✗
103	LaMDA Thoppilan et al. (2022)	Proprietary	No	64	8192	128	137B	2048	86%	552
104	MT-NLG Smith et al. (2022)	Proprietary	No	105	20480	128	530B	2048	67.5%	284
105	GShard Lepikhin et al. (2020)	Proprietary	No	64	8192	128	600B	2048	✗	4.3%
106	T5-XXL Raffel et al. (2020)	Apache 2.0	✓	24	1024	16	11B	512	48.6%	✗
107	ProphetNet Qi et al. (2020)	MIT	✓	12	768	12	300M	512	✗	✗
108	DialoGPT Zhang (2019)	MIT	✓	24	1024	16	345M	1024	25.81%	552
109	BART Lewis (2019)	MIT	✓	12	1024	16	406M	1024	✗	✗
110	PEGASUS Zhang et al. (2020)	Apache 2.0	✓	16	1024	16	568M	512	✗	✗
111	UniLM Dong et al. (2019)	MIT	✓	12	768	12	340M	512	✗	✗
112	Grok 3	Proprietary	✗	Unknown	Unknown	Unknown	Trillions	Unknown	✗	Unknown
113	Gemini 2 Ultra (2025) Source Link	Proprietary	No	Unknown	Unknown	Unknown	Unknown	2M	Unknown	Unknown
114	GPT-4.5 (2025) Source Link	Proprietary	No	Unknown	Unknown	Unknown	~12.8T	128K	Unknown	Unknown
115	Gemini 2.0 Flash-Lite (2025) Source Link	Proprietary	No	Unknown	Unknown	Unknown	Unknown	1M	Unknown	Unknown
116	Gemini 2.0 Pro (2025) Source Link	Proprietary	No	Unknown	Unknown	Unknown	Unknown	2M	Unknown	Unknown
117	Gemini 2.5 Pro (2025) Source Link	Proprietary	No	Unknown	Unknown	Unknown	Unknown	2M	Unknown	Unknown
118	GPT-o3 (2025) Source Link	Proprietary	No	Unknown	Unknown	Unknown	Unknown	200K	Unknown	Unknown
119	GPT-4.1 (2025) Source Link	Proprietary	No	Unknown	Unknown	Unknown	Unknown	1M	88.9%*	Unknown
120	GPT-o4-mini (2025) Source Link	Proprietary	No	Unknown	Unknown	Unknown	Unknown	200K	Unknown	Unknown
121	Qwen 3 235B (2025) Source Link	Apache 2.0	✓	Unknown	Unknown	Unknown	235B	128K	Unknown	Unknown