Case Report

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

Ranjan Sapkota *, Shaina Raza, Manoj Karkee *

Posted Date: 20 February 2025

doi: 10.20944/preprints202502.1608.v1

Keywords: ChatGPT transparency; DeepSeek transparency; ChatGPT accessibility; DeepSeek accessibility; ChatGPT open source; DeepSeek open source; ChatGPT transparency analysis; DeepSeek transparency analysis; ChatGPT reproducibility; DeepSeek reproducibility; ChatGPT model openness; DeepSeek model openness; ChatGPT research transparency; DeepSeek research transparency; ChatGPT open weight; DeepSeek open weight; ChatGPT data transparency; DeepSeek data transparency; ChatGPT code availability; DeepSeek code availability; ChatGPT bias mitigation; DeepSeek bias mitigation; ChatGPT ethical AI; DeepSeek ethical AI; ChatGPT LLM; DeepSeek LLM; ChatGPT model analysis; DeepSeek model analysis; ChatGPT accessibility standards; DeepSeek accessibility standards; ChatGPT open access; DeepSeek open access

 doi:10.20944/preprints202502.1608.v1

Article

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

Ranjan Sapkota [1,*], Shaina Raza [2] and Manoj Karkee [1,3]

[1] Biological & Environmental Engineering Cornell University, Ithaca, NY 14850, USA
[2] Vector Institute, Ontario, Tronto, ON W1140-108, Canada
[3] Department of Biological Systems Engineering, Washington State University, WA, USA
* Correspondence: rs2672@cornell.edu

Abstract: Despite increasing discussions on open-source Artificial Intelligence (AI), existing research lacks a discussion on the transparency and accessibility of state-of-the-art (SoTA) Large Language Models (LLMs). The Open Source Initiative (OSI) has recently released its first formal definition of open-source software. This definition, when combined with standard dictionary definitions and the sparse published literature, provide an initial framework to support broader accessibility to AI models such as LLMs, but more work is essential to capture the unique dynamics of openness in AI. In addition, concerns about open-washing, where models claim openness but lack full transparency, has been raised, which limits the reproducibility, bias mitigation, and domain adaptation of these models. In this context, our study critically analyzes SoTA LLMs from the last five years, including ChatGPT, DeepSeek, LLaMA, Grok, and others, to assess their adherence to transparency standards and the implications of partial openness. Specifically, we examine transparency and accessibility from two perspectives: open-source vs. open-weight models. Our findings reveal that while some models are labeled as open-source, this does not necessarily mean they are fully open-sourced. Even in the best cases, open-source models often do not report model training data, and code as well as key metrics, such as weight accessibility, and carbon emissions. To the best of our knowledge, this is the first study that systematically examines the transparency and accessibility of over 100 different SoTA LLMs through the dual lens of open-source and open-weight models. The findings open avenues for further research and call for responsible and sustainable AI practices to ensure greater transparency, accountability, and ethical deployment of these models.

Keywords: ChatGPT transparency; DeepSeek transparency; ChatGPT accessibility; DeepSeek accessibility; ChatGPT open source; DeepSeek open source; ChatGPT transparency analysis; DeepSeek transparency analysis; ChatGPT reproducibility; DeepSeek reproducibility; ChatGPT model openness; DeepSeek model openness; ChatGPT research transparency; DeepSeek research transparency; ChatGPT open weight; DeepSeek open weight; ChatGPT data transparency; DeepSeek data transparency; ChatGPT code availability; DeepSeek code availability; ChatGPT bias mitigation; DeepSeek bias mitigation; ChatGPT ethical AI; DeepSeek ethical AI; ChatGPT LLM; DeepSeek LLM; ChatGPT model analysis; DeepSeek model analysis; ChatGPT accessibility standards; DeepSeek accessibility standards; ChatGPT open access; DeepSeek open access

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

 doi:10.20944/preprints202502.1608.v1

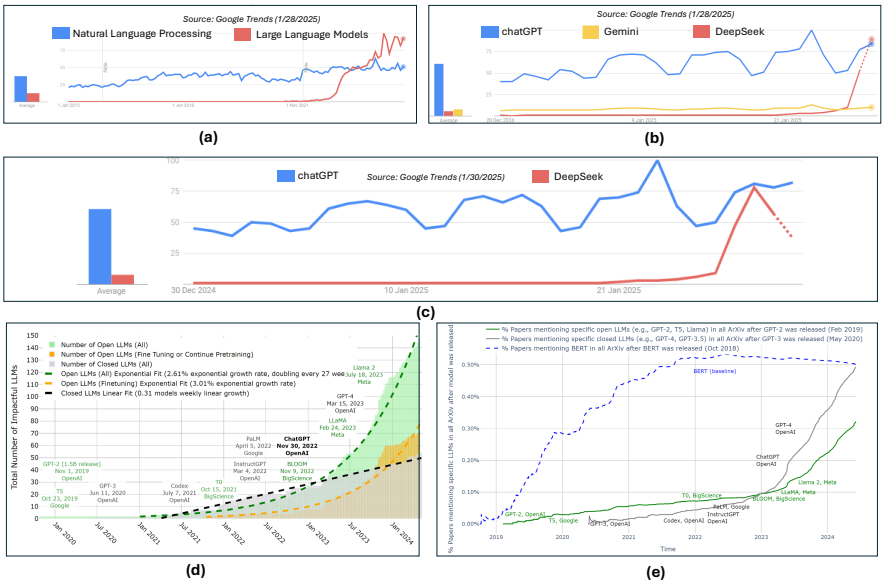

Figure 1. Analysis of NLP/LLM Interest (a) Google Trends showing increasing interest in NLP and LLMs from 2015 to 2025; (b) Global interest for ChatGPT, Gemini, and DeepSeek on January 28 2025, highlighting DeepSeek's rapid rise; (c) ChatGPT and Deepseek global interest on January 30, 2025; (d) Growth rates of open sourced and close closed souced LLMs [1] ; (e) Percentage of arXiv papers mentioning open LLMs or closed LLMs from 2019 onwards, with BERT as a baseline [1]

These technologies have become integral to systems and solutions across a diverse array of sectors, including healthcare [2], finance [3], education [4], and entertainment[5]. Their remarkable capabilities in language understanding and generation have not only revolutionized these industries but have also spurred a new wave of innovation and application development [6,7]. Amidst this rapid expansion, the term "open-source" frequently surfaces within discussions about LLMs [8]. 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, sustainability factors (e.g., CO_2 emissions), or detailed development processes. This gap is also widely discussed in the literature [9] and in numerous tech blogs, including [10], to name a few.

Although proprietary LLMs like OpenAI GPT-series (4/4o) [11] exhibit strong performance, their closed-source nature limits access to API-based interactions. In contrast, open-weight models like Meta LLama-series [12] 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 [13], showcases the flexibility and cost benefits of open-weight models.

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

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 [14]. Initially surpassing ChatGPT in search interest (Figure 1b), its popularity rapidly declined (Figure 1c), reflecting unmet expectations. While DeepSeek-R1 provides

 doi:10.20944/preprints202502.1608.v1

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.

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 goal of this study is to critically examines transparency practices of such "open-weight" LLMs, using DeepSeek-R1 and ChatGPT4o as primary examples, to map the distinctions between open-weight and fully open-source models. By doing so, we aim to:

- Elucidate the terminological ambiguities surrounding "open-source" within the AI domain, specifically distinguishing between truly open-source models and those termed "open-weight" which offer limited transparency.
- Investigate the implications of partial transparency on the reproducibility, community engagement, and ethical dimensions of AI development, emphasizing how these factors influence the practical deployment and trustworthiness of LLMs.
- Propose clearer guidelines and standards to differentiate truly open-source methodologies and models from strategies that merely provide access to pre-trained model weights.

With this study, we seek 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.

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.

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 AI literature. A concise 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.

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

 doi:10.20944/preprints202502.1608.v1

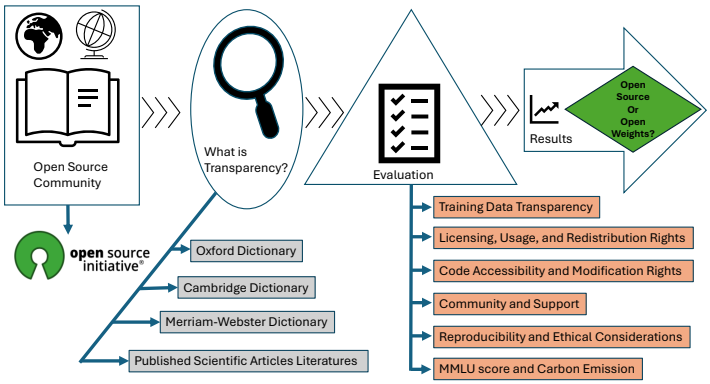

Figure 2. Overview of the methodologies used in evaluating ChatGPT, DeepSeek, and SoTA multimodal LLMs

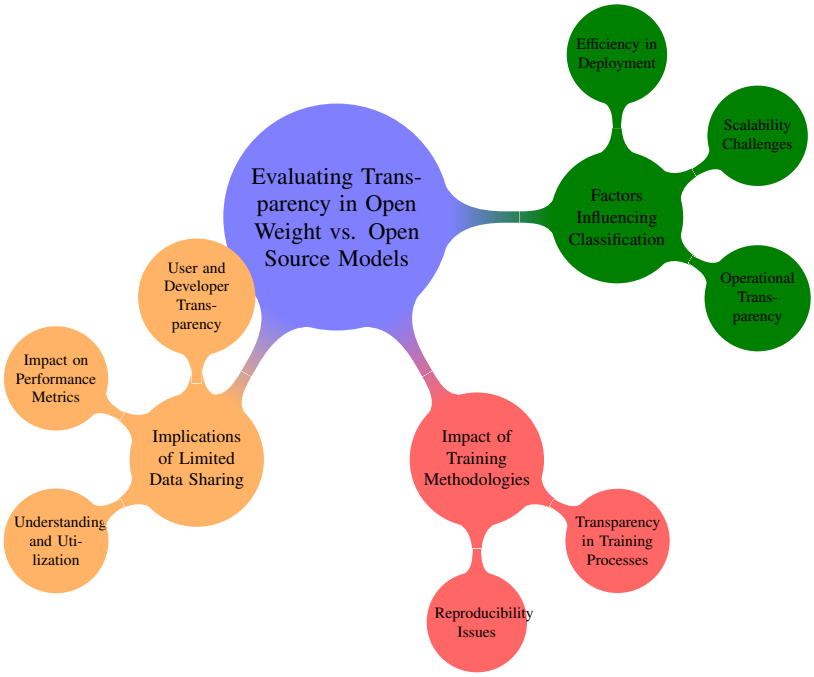

Figure 3. Illustration of an integrative mind map strategy developed to systematically evaluate transparency and accessibility of 111 evaluated LLMs from 2019 to the present. The diagram organizes critical dimensions, including factors influencing model classification, impacts of training methodologies, and consequences of limited data sharing—to comprehensively assess operational efficiency, scalability, and reproducibility challenges in open-weight versus open-source models.

2.2. Criteria for Openness and Transparency

Open-Source LLMs: An open-source LLM provides unrestricted access to its entire codebase, including the model architecture, training data, and the training processes [9]. 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). For example, Meta's LLamA aligns with the open-source paradigm by offering detailed insights into its design and implementation.

The primary goal with open-source models is to ensurecomplete transparency and flexibility. This openness enables comprehensive understanding, recreation, and reproducibility , even though some usage restrictions may still apply. 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

 doi:10.20944/preprints202502.1608.v1

projects, providing access to their training data, code and model weights. With subsequent versions like GPT-3, OpenAI shifted to a closed-source approach, restricting access to the model architecture, code, and weights. This trend continued with GPT-4, which also remains proprietary.

Open-Weight LLMs: Open-weight LLMs make their pre-trained model weights [15], 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 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 [16]. 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.

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. uch access allows any user to use, modify, share, and fully understand the AI system without needing special permissions.

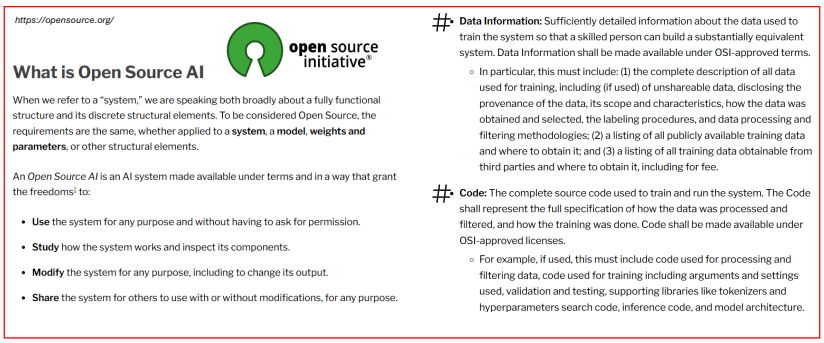

Figure 4. OSI's first official release of the open source definition, which sets foundational criteria/attributes for Openness in AI

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 [17–19]. 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 [20]. They are critical for both protecting creators and enabling users to innovate and adapt software to their needs [21]. For example, the MIT License, highly permissive, allows almost unrestricted use provided the original copyright is included.Similarly, the Apache License 2.0 [2] per-

[2] https://www.apache.org/licenses/LICENSE-2.0

 doi:10.20944/preprints202502.1608.v1

mits 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 1 provides an overview of popular licenses in AI practices, highlighting the varying degrees of permissiveness—from the flexible MIT License [4] to the stricter copyleft provisions of the GNU GPL 3.0 [5].

[3] https://creativecommons.org/share-your-work/cclicenses/
[4] https://opensource.org/license/mit

 doi:10.20944/preprints202502.1608.v1

Table 1. 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 [22]	Required	No explicit grant	Unlimited modifications	Must include original notices	-
Apache License 2.0 [23]	Required	Includes patent rights	Modifications documented	Must include original notices	-
GNU GPL 3.0 [24]	Required	-	Derivative works must also be open source	Source code must be disclosed	Strong copyleft
BSD License [25]	Required	No explicit grant	Unlimited modifications	No requirement to disclose source	No endorsement
Creative ML OpenRAIL-M [26]	Required	-	Ethical use guidelines	Must include original notices	Ethical guidelines
CC-BY-4.0 [27]	Credit required	-	Commercial and non-commercial use allowed	Must credit creator	-
CC-BY-NC-4.0 [28]	Credit required	-	Only non-commercial use allowed	Must credit creator	Non-commercial use only
BigScience OpenRAIL-M [29]	Required	-	Ethical use guidelines	Must include original notices	Ethical guidelines
BigCode Open RAIL-M v1 [30]	Required	-	Ethical use guidelines	Must include original notices	Ethical guidelines
Academic Free License v3.0 [31]	Required	Includes patent rights	Unlimited modifications	Must include original notices	-
Boost Software License 1.0 [32]	Required	No explicit grant	Unlimited modifications	Must include original notices	-
BSD 2-clause “Simplified” [33]	Required	No explicit grant	Unlimited modifications	No requirement to disclose source	No endorsement
BSD 3-clause “New” or “Revised” [34]	Required	No explicit grant	Unlimited modifications	No requirement to disclose source	No endorsement

 doi:10.20944/preprints202502.1608.v1

2.2.2. Open Source and Transparency

Following the OSI guidelines, 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 2, which presents a literature review and definitions derived from 10 popular literature defining transparency in AI systems.

Table 2. Unified Definitions of Transparency in AI from the published literatures.

Author and Reference	Definition
Lipton, Z. C. (2018). [35]	"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). [36]	"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). [37]	"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). [38]	"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). [39]	"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). [40]	"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). [41]	"Transparency is offering clear, interpretable explanations for decisions, which necessitates the **availability of training datasets and code** for full interpretability."
Bhatt, et al. (2020). [42]	"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). [43]	"Transparency ensures clear explanations of model behavior, significantly relying on the **availability of training datasets and code** for technical and operational clarity."

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

[6] https://www.oed.com/
[7] https://opensource.org/

 doi:10.20944/preprints202502.1608.v1

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) [44]. 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. 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 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 (e.g., > 3000 citations), were specifically targeted. This criterion was instrumental because highly cited papers typically reflect pivotal developments in the field and are often the genesis of new research trajectories or shifts in scientific paradigms. 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 [45], marking the beginning of a new era in AI by introducing a model architecture based entirely 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 published research 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 112 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 integrated results, illustrated in Figure 5, provide a visual representation of the temporal distribution and 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).

 doi:10.20944/preprints202502.1608.v1

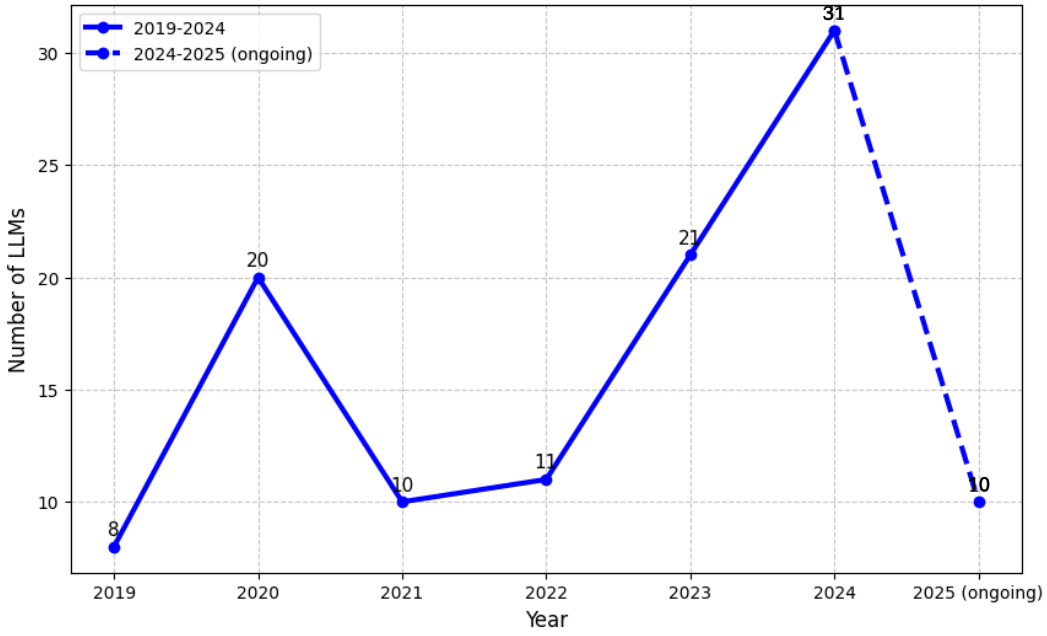

Figure 5. Distribution of 112 research papers analyzed in this study, spanning from 2019 to 2025. The plot reveals a steadily increasing trend in LLM studies, underscoring rapid advancements in transparency and accessibility. Notably, February 2025 signals a new year, suggesting that many more LLMs may soon emerge.

Inclusion Criteria A thorough literature review was conducted to locate transparency 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 [46]. 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 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 their efficiency and scalability in practical applications?

 doi:10.20944/preprints202502.1608.v1

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.

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 reveals that many models marketed or perceived as "open" primarily provide open weights (i.e., publicly available trained parameters) rather than full open-source access (i.e., source code, training data, and detailed methodologies). Table 3 outlines these distinctions across leading multimodal LLMs.

The analysis of the comprehensive table (Table 3), which evaluates 112 LLMs released between 2019 and 2025 interms of release year, training data and key features. Early models, such as GPT-2 [47] and BERT [48], 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 [14] 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 significantly 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, an emerging trend toward greater transparency regarding training methodologies is evident. This evolution reflects an industry-wide shift towards balancing commercial interests with greater accountability and openness in AI research.

Table 3. Detailed specifications of Large Language Models (2019–2025), 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) [47]	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

Continued on next column/page

 doi:10.20944/preprints202502.1608.v1

Continued from previous column/page

Model, Year & Citation	Training Data	Key Features
6.GPT-4o (2024) [49]	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 [50] (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 [14] (2025)	Hybrid dataset of 9.8T tokens from synthetic and organic sources	Mixture of Experts (MoE), enhanced with mathematical reasoning capabilities
15. DeepSeek LLM [51] (2023)	Books+Wiki data up to 2023	Scaling Language Models
16. DeepSeek LLM V2 [52] (2023)	Highly efficient training	MLA, MoE, Lowered costs
17. DeepSeek Coder V2 [53] (2023)	Supports 338 languages	Enhanced coding capabilities
18. DeepSeek V3 [54] (2023)	Advanced MoE architecture	High-performance, FP8 training
19. BERT-Base [48] (2019)	Books+Wiki data collected up to 2019	Masked Language Modeling (MLM)
20. BERT-Large [48] (2019)	Books+Wiki data collected up to 2019	Next Sentence Prediction (NSP)
21. T5-Small [55] (2020)	C4 (Large-scale text dataset)	Text-to-text, encoder-decoder
22. T5-Base [55] (2020)	C4 (Large-scale text dataset)	Text-to-text, scalable, encoder-decoder
23. T5-Large [55] (2020)	C4 (Large-scale text dataset)	Text-to-text, scalable, encoder-decoder
24. T5-3B [55] (2020)	C4 (Large-scale text dataset)	Text-to-text, scalable, encoder-decoder
25. T5-11B [55] (2020)	C4 (Large-scale text dataset)	Text-to-text, scalable, encoder-decoder
26. Mistral 7B [56] (2023)	Compiled from diverse sources totaling 2.4T tokens	Sliding Window Attention (SWA)
27. LLaMA 2 70B [12] (2023)	Diverse corpus aggregated up to 2T tokens	Grouped Query Attention (GQA)
28. CriticGPT [57] (2024)	Human feedback data	Fine-tuned for critique generation
29. Olympus (2024)	40T tokens	Large-scale, proprietary model
30. HLAT [58] (2024)	Not specified	High-performance, task-specific
31. Multimodal-CoT [59] (2023)	Multimodal datasets	Chain-of-Thought reasoning for multimodal tasks

Continued on next column/page

 doi:10.20944/preprints202502.1608.v1

Continued from previous column/page

Model, Year & Citation	Training Data	Key Features
32. AlexaTM 20B [60] (2022)	Not specified	Multilingual, task-specific
33. Chameleon [61] (2024)	9.2T tokens	Multimodal, high-performance
34. Llama 3 70B [62] (2024)	2T tokens	High-performance, open-source
35. LIMA [63] (2024)	Not specified	High-performance, task-specific
36. BlenderBot 3x [64] (2023)	300B tokens	Conversational AI, improved reasoning
37. Atlas [65] (2023)	40B tokens	High-performance, task-specific
38. InCoder [66] (2022)	Not specified	Code generation, task-specific
39. 4M-21 [67] (2024)	Not specified	High-performance, task-specific
40. Apple On-Device model [68] (2024)	1.5T tokens	On-device, task-specific
41. MM1 [69] (2024)	2.08T tokens	Multimodal, high-performance
42. ReALM-3B [70] (2024)	134B tokens	High-performance, task-specific
43. Ferret-UI [71] (2024)	2T tokens	Multimodal, high-performance
44. MGIE [72] (2023)	2T tokens	Multimodal, high-performance
45. Ferret [73] (2023)	2T tokens	Multimodal, high-performance
46. Nemotron-4 340B [74] (2024)	9T tokens	High-performance, task-specific
47. VIMA [75] (2023)	Not specified	Multimodal, high-performance
48. Retro 48B [76] (2023)	1.2T tokens	High-performance, task-specific
49. Raven [77] (2023)	40B tokens	High-performance, task-specific
50. Gemini 1.5 [78] (2024)	Not specified	Multimodal, high-performance
51. Med-Gemini-L 1.0 [79] (2024)	30T tokens	Medical-focused, high-performance
52. Hawk [80] (2024)	300B tokens	High-performance, task-specific
53. Griffin [80] (2024)	300B tokens	High-performance, task-specific
54. Gemma [81] (2024)	6T tokens	High-performance, task-specific
55. Gemini 1.5 Pro [78] (2024)	30T tokens	Multimodal, high-performance
56. PaLi-3 [82] (2023)	Not specified	Multimodal, high-performance
57. RT-X [83] (2023)	Not specified	Robotics-focused, high-performance
58. Med-PaLM M [84] (2024)	780B tokens	Medical-focused, high-performance
59. MAI-1 [85] (2024)	10T tokens	High-performance, task-specific

Continued on next column/page

 doi:10.20944/preprints202502.1608.v1

Continued from previous column/page

Model, Year & Citation	Training Data	Key Features
60. YOCO [86] (2024)	1.6T tokens	High-performance, task-specific
61. phi-3-medium [87] (2024)	4.8T tokens	High-performance, task-specific
62. phi-3-mini [87] (2024)	3.3T tokens	High-performance, task-specific
63. WizardLM-2-8x22B [88] (2024)	Not specified	High-performance, task-specific
64. WaveCoder-Pro-6.7B [89] (2023)	20B tokens	Code-focused, high-performance
65. WaveCoder-Ultra-6.7B [89] (2023)	20B tokens	Code-focused, high-performance
66. WaveCoder-SC-15B [89] (2023)	20B tokens	Code-focused, high-performance
67. OCRA 2 [90] (2023)	Not specified	High-performance, task-specific
68. Florence-2 [91] (2024)	5.4B visual annotations	Multimodal, high-performance
69. Qwen [92] (2023)	3T tokens	High-performance, task-specific
70. SeaLLM-13b [93] (2023)	2T tokens	Multilingual, high-performance
71. Grok-1 [94] (2024)	13.2T tokens	Incorporates humor-enhancing algorithms
72. Phi-4 [87] (2024)	9.8T tokens	Optimized for STEM applications
73. Megatron-LM [95] (2020)	Common Crawl, Wikipedia, Books	Large-scale parallel training, optimized for NVIDIA GPUs
74. Turing-NLG [96] (2020)	Diverse web text	High-quality text generation, used in Microsoft products
75. CTRL [97] (2019)	Diverse web text with control codes	Controlled text generation using control codes
76. XLNet [98] (2019)	BooksCorpus, Wikipedia, Giga5, ClueWeb	Permutation-based training, outperforms BERT on many benchmarks
77. RoBERTa [99] (2019)	BooksCorpus, Wikipedia, CC-News, OpenWebText	Improved BERT with better pretraining techniques
78. ELECTRA [100] (2020)	BooksCorpus, Wikipedia	Replaces masked language modeling with a more efficient discriminative task

Continued on next column/page

 doi:10.20944/preprints202502.1608.v1

Continued from previous column/page

Model, Year & Citation	Training Data	Key Features
79. ALBERT [101] (2019)	BooksCorpus, Wikipedia	Parameter reduction techniques for efficient training
80. DistilBERT [102] (2019)	BooksCorpus, Wikipedia	Distilled version of BERT, smaller and faster
81. BigBird [103] (2020)	BooksCorpus, Wikipedia, PG-19	Sparse attention mechanism for handling long sequences
82. Gopher [104] (2021)	MassiveText dataset (2.5T tokens)	Focused on scaling laws and model performance
83. Chinchilla [105] (2022)	MassiveText dataset (1.4T tokens)	Optimized for compute-efficient training
84. PaLM [106] (2022)	Diverse web text, books, code	Pathways system for efficient training, multilingual support
85. OPT [107] (2022)	Diverse web text	Open-source alternative to GPT-3
86. BLOOM [108] (2022)	ROOTS corpus (1.6T tokens)	Multilingual, open-source, collaborative effort
87. Jurassic-1 [109] (2021)	Diverse web text	High-quality text generation, API-based access
88. Codex [110] (2021)	Code repositories (e.g., GitHub)	Specialized in code generation and understanding
89. T0 [111] (2021)	Diverse NLP datasets	Zero-shot task generalization
90. UL2 [112] (2022)	Diverse web text	Unified pretraining for diverse NLP tasks
91. GLaM [113] (2021)	Diverse web text	Sparse mixture of experts (MoE) architecture
92. ERNIE 3.0 [114] (2021)	Chinese and English text	Knowledge-enhanced pretraining
93. GPT-NeoX [115] (2022)	The Pile (825GB dataset)	Open-source, large-scale, efficient training
94. CodeGen [116] (2022)	Code repositories (e.g., GitHub)	Specialized in code generation
95. FLAN-T5 [117] (2022)	Diverse NLP datasets	Instruction fine-tuning for better generalization
96. mT5 [118] (2020)	mC4 dataset (101 languages)	Multilingual text-to-text transfer
97. Reformer [119] (2020)	Diverse web text	Efficient attention mechanism for long sequences
98. Longformer [120] (2020)	BooksCorpus, Wikipedia	Efficient attention for long documents

Continued on next column/page

 doi:10.20944/preprints202502.1608.v1

Continued from previous column/page

Model, Year & Citation	Training Data	Key Features
99. DeBERTa [121] (2021)	BooksCorpus, Wikipedia	Disentangled attention mechanism
100. T-NLG [122] (2020)	Diverse web text	High-quality text generation
101. Switch Transformer [123] (2021)	Diverse web text	Sparse mixture of experts (MoE)
102. WuDao 2.0 [124] (2021)	Chinese and English text	Largest Chinese language model
103. LaMDA [125] (2021)	Diverse dialogue data	Specialized in conversational AI
104. MT-NLG [96] (2021)	Diverse web text	High-quality text generation
105. GShard [126] (2020)	Diverse web text	Sparse mixture of experts (MoE)
106. T5-XXL [55] (2020)	C4 dataset	Large-scale text-to-text transfer
107. ProphetNet [127] (2020)	BooksCorpus, Wikipedia	Future token prediction for better sequence modeling
108. DialoGPT [128] (2020)	Reddit dialogue data	Specialized in conversational AI
109. BART [129] (2020)	BooksCorpus, Wikipedia	Denoising autoencoder for text generation
110. PEGASUS [130] (2020)	C4 dataset	Pre-training with gap-sentences for summarization
111. UniLM [131] (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

Further comparative analysis indicates that the evolution of LLMs is marked by continuous refinement of critical features essential for practical applications. Models such as T5-XXL [55] 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 computational efficiency, reproducibility, and bias mitigation. Additionally, evolving training methodologies—from basic text-to-text transfer to sophisticated hybrid approaches—have resulted in models that are increasingly adept at 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 governance. The emergence of open-weight models, such as those from DeepSeek and ChatGPT, illustrates a deliberate strategy to balance accessibility with proprietary innovation. These findings in the Table 3 suggest that future LLMs will continue to build on these innovations, paving the way for more transparent, efficient, and ethically responsible AI systems [105].

Furthermore, in appendix section, Table 5 provides a comprehensive overview of these 111 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 [47], ChatGPT-3.5, ChatGPT-4) largely adopt proprietary licenses, restricting access to their training data, code, and methodologies. In contrast, models like BERT [48] and certain DeepSeek variants (e.g., DeepSeek-R1 [14]) 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

 doi:10.20944/preprints202502.1608.v1

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 [55] and LLaMA 2 [12], where a mix of open and proprietary strategies reflects divergent priorities—commercial viability versus reproducibility. This heterogeneous licensing landscape, as corroborated by studies e.g., [14,48,49,53,55,56], underlines the challenges in balancing innovation, transparency, and community engagement in modern LLM development.

 doi:10.20944/preprints202502.1608.v1

Table 5. 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 [47]	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 [49]	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 [50]	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 [14]	Apache 2.0	✓	64	8192	64/8	671B	128K	90.8%	44
15	DeepSeek LLM [51]	Proprietary	✓	24	2048	16	67B	2048	N/A	44
16	DeepSeek LLM V2 [52]	Proprietary	No	Not specified	Not specified	Not specified	236B	128K	78.5%	✗
17	DeepSeek Coder V2 [53]	Proprietary	No	Not specified	Not specified	Not specified	236B	128K	79.2%	✗
18	DeepSeek V3 [54]	Proprietary	No	Not specified	Not specified	Not specified	671B	128K	75.7%	✗
19	BERT-Base [48]	Apache 2.0	✓	12	768	12	110M	512	67.2%	0.652
20	BERT-Large [48]	Apache 2.0	✓	24	1024	16	340M	512	69.3%	0.652
21	T5-Small [55]	Apache 2.0	Yes	6/6	512	8	60M	512	✗	✗
22	T5-Base [55]	Apache 2.0	Yes	12/12	768	12	220M	512	35.9%	✗
23	T5-Large [55]	Apache 2.0	Yes	24/24	1024	16	770M	512	40%	✗
24	T5-3B [55]	Apache 2.0	Yes	24/24	1024	32	3B	512	✗	✗
25	T5-11B [55]	Apache 2.0	Yes	24/24	1024	128	11B	512	48.6%	T5-11B
26	Mistral 7B [56]	Apache 2.0	✓	32	4096	32	7.3B	8K	62.5%	✗
27	LLaMA 2 70B [12]	Llama 2	✓	80	8192	64	65.2B	4K	68.9%	291.42
28	CriticGPT [57]	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 [58]	Proprietary	×	Not specified	Not specified	Not specified	7B	Not specified	✗	✗
31	Multimodal-CoT [59]	Proprietary	×	Not specified	Not specified	Not specified	Not specified	Not specified	✗	✗

Continued on next page

 doi:10.20944/preprints202502.1608.v1

Table 5 – continued from previous page

No.	Model	License	Weights	Layers	Hidden	Heads	Params	Context	MMLU score	Carbon Emitted (tCO2eq)
32	AlexaTM 20B [60]	Proprietary	×	Not specified	Not specified	Not specified	20B	Not specified	✗	✗
33	Chameleon [61]	Proprietary	×	Not specified	Not specified	Not specified	34B	Not specified	✗	✗
34	Llama 3 70B [62]	Llama 3	✓	Not specified	Not specified	Not specified	70B	Not specified	82.0%	1900
35	LIMA [63]	Proprietary	×	Not specified	Not specified	Not specified	65B	Not specified	✗	✗
36	BlenderBot 3x [64]	Proprietary	×	Not specified	Not specified	Not specified	150B	Not specified	✗	✗
37	Atlas [65]	Proprietary	×	Not specified	Not specified	Not specified	11B	Not specified	47.9%	✗
38	InCoder [66]	Proprietary	×	Not specified	Not specified	Not specified	6.7B	Not specified	✗	✗
39	4M-21 [67]	Proprietary	×	Not specified	Not specified	Not specified	3B	Not specified	✗	✗
40	Apple On-Device model [68]	Proprietary	×	Not specified	Not specified	Not specified	3.04B	Not specified	✗	✗
41	MM1 [69]	Proprietary	×	Not specified	Not specified	Not specified	30B	Not specified	✗	✗
42	ReALM-3B [70]	Proprietary	×	Not specified	Not specified	Not specified	3B	Not specified	✗	✗
43	Ferret-UI [71]	Proprietary	×	Not specified	Not specified	Not specified	13B	Not specified	✗	✗
44	MGIE [72]	Proprietary	×	Not specified	Not specified	Not specified	7B	Not specified	✗	✗
45	Ferret [73]	Proprietary	×	Not specified	Not specified	Not specified	13B	Not specified	✗	✗
46	Nemotron-4 340B [74]	Proprietary	×	Not specified	Not specified	Not specified	340B	Not specified	✗	✗
47	VIMA [75]	Proprietary	×	Not specified	Not specified	Not specified	0.2B	Not specified	✗	✗
48	Retro 48B [76]	Proprietary	×	Not specified	Not specified	Not specified	48B	Not specified	✗	✗
49	Raven [77]	Proprietary	×	Not specified	Not specified	Not specified	11B	Not specified	✗	✗
50	Gemini 1.5 [78]	Proprietary	×	Not specified	Not specified	Not specified	Not specified	Not specified	90%	✗
51	Med-Gemini-L 1.0 [79]	Proprietary	×	Not specified	Not specified	Not specified	1500B	Not specified	✗	✗
52	Hawk [80]	Proprietary	×	Not specified	Not specified	Not specified	7B	Not specified	✗	✗
53	Griffin [80]	Proprietary	×	Not specified	Not specified	Not specified	14B	Not specified	✗	✗

Continued on next page

 doi:10.20944/preprints202502.1608.v1

Table 5 – continued from previous page

No.	Model	License	Weights	Layers	Hidden	Heads	Params	Context	MMLU score	Carbon Emitted (tCO2eq)
54	Gemma [81]	Proprietary	×	Not specified	Not specified	Not specified	7B	Not specified	64.3%	✗
55	Gemini 1.5 Pro [78]	Proprietary	×	Not specified	Not specified	Not specified	1500B	Not specified	✗	✗
56	PaLi-3 [82]	Proprietary	×	Not specified	Not specified	Not specified	6B	Not specified	✗	✗
57	RT-X [83]	Proprietary	×	Not specified	Not specified	Not specified	55B	Not specified	✗	✗
58	Med-PaLM M [84]	Proprietary	×	Not specified	Not specified	Not specified	540B	Not specified	✗	✗
59	MAI-1 [85]	Proprietary	×	Not specified	Not specified	Not specified	500B	Not specified	✗	✗
60	YOCO [86]	Proprietary	×	Not specified	Not specified	Not specified	3B	Not specified	✗	✗
61	phi-3-medium [87]	Proprietary	×	Not specified	Not specified	Not specified	14B	Not specified	✗	✗
62	phi-3-mini [87]	Proprietary	×	Not specified	Not specified	Not specified	3.8B	Not specified	✗	✗
63	WizardLM-2-8x22B [88]	Proprietary	×	Not specified	Not specified	Not specified	141B	Not specified	✗	✗
64	WaveCoder-Pro-6.7B [89]	Proprietary	×	Not specified	Not specified	Not specified	6.7B	Not specified	✗	✗
65	WaveCoder-Ultra-6.7B [89]	Proprietary	×	Not specified	Not specified	Not specified	6.7B	Not specified	✗	✗
66	WaveCoder-SC-15B [89]	Proprietary	×	Not specified	Not specified	Not specified	15B	Not specified	✗	✗
67	OCRA 2 [90]	Proprietary	×	Not specified	Not specified	Not specified	7B, 13B	Not specified	✗	✗
68	Florence-2 [91]	Proprietary	×	Not specified	Not specified	Not specified	Not specified	Not specified	✗	✗
69	Qwen [92]	Proprietary	×	Not specified	Not specified	Not specified	72B	Not specified	✗	✗
70	SeaLLM-13b [93]	Proprietary	×	Not specified	Not specified	Not specified	13B	Not specified	✗	✗
71	Grok-1 [94]	Apache 2.0	✓	64	6144	48/8	314B	8K	N/A	x
72	Phi-4 [87]	MIT	✓	48	3072	32	14B	16K	71.2%	x
73	Megatron-LM [95]	Custom	No	72	3072	32	8.3B	2048	✗	✗
74	Turing-NLG [96]	Proprietary	No	78	4256	28	17B	1024	✗	✗
75	CTRL(Conditional Transformer Language Model) [97]	Apache 2.0	✓	48	1280	16	1.6B	256	✗	✗
76	XLNet [98]	Apache 2.0	✓	24	1024	16	340M (Base), 1.5B (Large)	512	✗	0.652

Continued on next page

 doi:10.20944/preprints202502.1608.v1

Table 5 – continued from previous page

No.	Model	License	Weights	Layers	Hidden	Heads	Params	Context	MMLU score	Carbon Emitted (tCO2eq)
77	RoBERTa [99]	MIT	✓	24	1024	16	355M	512	✗	✗
78	ELECTRA [100]	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) [101]	Apache 2.0	✓	12 (Base), 24 (Large)	768 (Base), 1024 (Large)	12 (Base), 16 (Large)	12M (Base), 18M (Large)	512	✗	0.652
80	DistilBERT [102]	Apache 2.0	✓	6	768	12	66M	512	✗	0.652
81	BigBird [103]	Apache 2.0	✓	12 (Base), 24 (Large)	768 (Base), 1024 (Large)	12 (Base), 16 (Large)	110M (Base), 340M (Large)	4096	✗	✗
82	Gopher [104]	Proprietary	No	80	8192	128	280B	2048	60%	✗
83	Chinchilla [105]	Proprietary	No	80	8192	128	70B	2048	✗	✗
84	PaLM [106]	Proprietary	No	118	18432	128	540B	8192	69.3%	✗
85	OPT (Open Pretrained Transformer) [107]	Non-commercial	✓	96	12288	96	175B	2048	✗	✗
86	BLOOM [108]	Responsible AI License	✓	70	14336	112	176B	2048	90%	✗
87	Jurassic-1 [109]	Proprietary	No	76	12288	96	178B	2048	67.5	✗
88	Codex [110]	Proprietary	No	96	12288	96	12B	4096	✗	✗
89	T0 (T5 for Zero-Shot Tasks) [111]	Apache 2.0	✓	24	1024	16	11B	512	✗	✗
90	UL2 (Unifying Language Learning Paradigms) [112]	Apache 2.0	✓	32	4096	32	20B	2048	✗	✗
91	GLaM (Generalist Language Model) [113]	Proprietary	No	64	8192	128	1.2T (sparse)	2048	✗	✗
92	ERNIE 3.0 [114]	Proprietary	No	48	4096	64	10B	512	✗	✗
93	GPT-NeoX [115]	Apache 2.0	✓	44	6144	64	20B	2048	33.6	✗
94	CodeGen [116]	Apache 2.0	✓	32	4096	32	16B	2048	✗	✗
95	FLAN-T5 [117]	Apache 2.0	✓	24	1024	16	11B	512	52.5	552
96	mT5 (Multilingual T5) [118]	Apache 2.0	✓	24	1024	16	13B	512	52.4	552
97	Reformer [119]	Apache 2.0	✓	12	768	12	150M	64K	✗	552
98	Longformer [120]	Apache 2.0	✓	12	768	12	150M	4096	✗	552
99	DeBERTa [121]	MIT	✓	12	768	12	1.5B	512	✗	552
100	T-NLG (Turing Natural Language Generation) [122]	Proprietary	No	78	4256	28	17B	1024	✗	✗
101	Switch Transformer [123]	Apache 2.0	✓	24	4096	32	1.6T (sparse)	2048	✗	✗
102	WuDao 2.0 [124]	Proprietary	No	128	12288	96	1.75T	2048	86.4%	✗
103	LaMDA [125]	Proprietary	No	64	8192	128	137B	2048	86%	552
104	MT-NLG [96]	Proprietary	No	105	20480	128	530B	2048	67.5%	284

Continued on next page

 doi:10.20944/preprints202502.1608.v1

Table 5 – continued from previous page

No.	Model	License	Weights	Layers	Hidden	Heads	Params	Context	MMLU score	Carbon Emitted (tCO2eq)
105	GShard [126]	Proprietary	No	64	8192	128	600B	2048	✗	4.3%
106	T5-XXL [55]	Apache 2.0	✓	24	1024	16	11B	512	48.6%	✗
107	ProphetNet [127]	MIT	✓	12	768	12	300M	512	✗	✗
108	DialoGPT [128]	MIT	✓	24	1024	16	345M	1024	25.81%	552
109	BART [129]	MIT	✓	12	1024	16	406M	1024	✗	✗
110	PEGASUS [130]	Apache 2.0	✓	16	1024	16	568M	512	✗	✗
111	UniLM [131]	MIT	✓	12	768	12	340M	512	✗	✗

 doi:10.20944/preprints202502.1608.v1

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 frequently 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 [132,133].

Further analysis of Table 5 (Refer to appendix section) reveals significant 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 carbon emissions (e.g., 552 tCO_2eq for several variants and 1,035 tCO_2eq for GPT-4o [49]). In contrast, earlier models such as GPT-2 lack these performance benchmarks, reflecting evolving evaluation standards. The DeepSeek family shows a promising balance: 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 trends are evident in the T5 series [55] and LLaMA 2 [12], which have progressively incorporated larger, more diverse datasets alongside performance improvements. Studies e.g., [14,53,104,105,112**?**] indicate that while performance enhancements are significant, the associated environmental costs necessitate a shift toward more energy-efficient architectures and transparent reporting practices.

Table 5 also details the architectural specifications that underpin model performance. The GPT family, including 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 [49]. 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 [55] and LLaMA 2 [12] further illustrate a trend toward optimizing architectural design for scalability, efficiency, and energy conservation. These models reveal a shift from sheer scale towards balanced configurations that emphasize reproducibility and ethical considerations. Several works e.g., [12,48,53,55,56,105] 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 can reconcile performance, efficiency, and openness in next-generation LLMs.

Furthermore, Table 5 (Refer to appendix section) presents an extended analysis of 111 LLMs from 2019 to 2025, emphasizing architectural specifications and openness metrics. Early models such as GPT-2 [47] and BERT [48] laid the groundwork with moderate layer counts, hidden sizes, and attention head configurations. As the field evolved, later models—particularly within the ChatGPT and DeepSeek families [14,53], 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 [12,55,105]. Overall, the extended analysis highlights an industry-wide progression from simpler architectures with limited transparency to highly engineered systems that balance commercial interests with technical rigor and ethical considerations.

3.2. Model-Specific Evaluations

ChatGPT

GPT-4 [134] and ChatGPT [11] are proprietary models with limited architectural transparency: their training datasets, fine-tuning protocols, and structural details (e.g., layer configurations, attention mechanisms) remain undisclosed. While GPT-4's technical report outlines high-level capabilities[135],

 doi:10.20944/preprints202502.1608.v1

it omits reproducibility-critical specifics such as pre-training corpus composition, hyperparameters, and energy consumption metrics, reflecting a priority on commercial secrecy, which limits its scientific openness. Similarly, ChatGPT's API-based access restricts users to input-output interactions without exposing model internals [136], thus creating a "black box" system that lacks transparency and does not allow third-party modifications.

ChatGPT adopt 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 [137,138]. 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 is misleading because true open-weight standards—such as parameter accessibility (e.g., Llama 2) or training code disclosure (e.g., BLOOM [139]), are absent, underscoring the competing priorities between commercial control and open scientific collaboration in modern AI ecosystems. The ChatGPT's version's:

- **GPT-2:** [47] 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 protocols. While permitting commercial use and modification, the absence of detailed preprocessing 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 [50].
- **Default ChatGPT-3.5:** Default ChatGPT-3.5 [50] 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 [50] 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 [49] 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 [49] 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.

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 [14]. 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 [140,141]. 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.

 doi:10.20944/preprints202502.1608.v1

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 models:

- **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 [14], challenges conventional scaling paradigms, as evidenced by its disruptive impact on GPU market dynamics [51–54]. 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 [51].
- **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 [52].
- **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 [53].
- **DeepSeek V3**: DeepSeek V3 uses proprietary FP8 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. [54]

Miscellaneous Proprietary Models

Meta's Llama The Llemma language model [13], 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%).

 doi:10.20944/preprints202502.1608.v1

However, Llemma is also categorized as open-weights rather than fully open-source due to incomplete transparency in its development pipeline [13]. While the Proof-Pile-2 dataset is released [8], it excludes subsets like Lean theorem-proving data and lacks detailed documentation on 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, comprising Ultra (1.56T parameters), Pro (137B), and Nano (3.2B/17.5B) variants, operates under an open-weights paradigm as well, where pretrained model parameters are accessible via APIs but remain proprietary and non-modifiable [142]. The architecture integrates multimodal fusion mechanisms, including cross-modal attention layers and sparsely activated mixture-of-experts (MoE) blocks, trained on a corpus of 12.5 trillion text tokens, 3.2 billion images, and 1.1 billion video-audio pairs. While technical documentation outlines innovations such as dynamic token routing for modality-specific computations and TPUv5-optimized distributed training, critical reproducibility details such as the MoE router logic, TPU compiler configurations, and multimodal alignment loss functions are not shared. Furthermore, the training dataset composition that includes spanning web documents (50%), code repositories (18%), and proprietary media (32%) lacks granular metadata, obscuring data provenance and ethical sourcing practices. This partial transparency enables limited third-party deployment (e.g., Gemini Nano on Pixel devices) but restricts independent assessment of biases or safety protocols, as weights are encrypted and inference-only.

Gemini is classified as open-weights rather than open-source due to the constraints in transparency and licensing. Its proprietary Google license prohibits weight modification, redistribution, and commercial use cases competing with Google services, diverging from open-source standards like Apache 2.0 that permit unrestricted adaptation. The model's technical report omits hyperparameters critical for replication, such as Gemini Ultra's learning rate schedule (0.00000625), Pro's 4.8-bit quantization thresholds, and Nano's knowledge distillation ratios from larger variants. Additionally, the reinforcement learning from human feedback (RLHF) pipeline—including reward model architectures and safety filtering protocols, is described only abstractly, preventing independent validation of alignment strategies. While Gemini's API-accessible weights support downstream applications (e.g., Android's on-device AI), the absence of training code (e.g., JAX/FLAX implementations), dataset indices, and infrastructure blueprints encourages dependency on Google's proprietary ecosystem. This lack of transparency safeguards Google's business secrecy including multimodal data pipelines, similar to those of others discussed above, illustrating the inherent tradeoff in corporate AI between accessibility and competitive secrecy. For these Google models, full open-source status would necessitate

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

 doi:10.20944/preprints202502.1608.v1

disclosing training methodologies, infrastructure configurations, and dataset curation frameworks, steps incompatible with Google's commercial AI roadmap.

Mistral AI Mistral AI's models, including Mistral 7B and Mixtral 8x7B, can also be classified as open-weights due to their release of model parameters and architectural blueprints under the Apache 2.0 license, which permits commercial use, modification, and redistribution [56]. These models employ advanced architectures such as grouped-query attention (GQA) and sliding window attention (SWA) to optimize inference efficiency, with Mistral 7B trained on 2.4 trillion tokens of multilingual data. However, critical components necessary for full reproducibility—including the composition of the training dataset, hyperparameter configurations (e.g., learning rate schedules, batch sizes), and reinforcement learning from human feedback (RLHF) pipelines—remain undisclosed. This selective transparency extends to licensing: advanced models like Codestral-22B are governed by the Mistral Non-Production License (MNPL), which restricts commercial deployment without explicit agreements, creating a tiered accessibility framework. While Mistral provides inference code and quantized weight variants (e.g., GGUF, AWQ), the absence of training infrastructure details (e.g., TPU/GPU cluster configurations) and proprietary fine-tuning datasets limits independent replication of their results.

Mistral AI's models diverge from open-source standards through three systemic constraints: (1) data opacity, as the training corpus—spanning web texts, code repositories, and multilingual sources—lacks detailed documentation, preventing analysis of potential bias or ethical sourcing; (2) methodological gaps, where architectural innovations like SWA (window size: 4,096 tokens) are disclosed, but RLHF reward models and distributed training protocols remain proprietary; and (3) restrictive licensing, including the prohibition of commercial use by MNPL, which does not follow open-source principles where modification rights would be granted to users/developers. This hybrid approach, as mentioned before, tries to balance regional and companies's interests in protecting their commercial competitiveness while enabling, at least marginally, community-driven adaptations (e.g., fine-tuning for code generation). Full open-source compliance would necessitate releasing training data, infrastructure blueprints, and unconstrained licenses—compromises incompatible with Mistral's market strategy, which prioritizes controlled innovation over unrestricted 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 [87,143]. 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 optimizations (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

 doi:10.20944/preprints202502.1608.v1

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.

Dolly 2.0 Dolly 2.0 stands out in the landscape of LLMs by being fully open-sourced, a rarity in a field where most models are typically released with "open weights" only (https://huggingface.co/databricks/dolly-v2-12b). Unlike its predecessors and many contemporaries, Dolly 2.0 extends openness beyond just the model weights to include its training code and the specific dataset used for its instruction tuning. This comprehensive approach to transparency is facilitated by Databricks' use of a permissive CC-BY-SA license for the model and dataset, and an Apache 2.0 license for the code. These licenses allow for both commercial use and community adaptation, encouraging innovation and customization without the typical restrictions imposed by proprietary models or models with limited openness discussed earlier. Full openness of Dolly 2.0 addresses a critical gap in the AI community, where the utility of LLMs has often been hampered by limited access to the underlying tools that could allow users to understand, replicate, and build upon pre-existing work. By open-sourcing the entire ecosystem of Dolly 2.0, which includes weights, code, and training data, Databricks sets a new precedent for transparency and adaptability in AI development. This approach not only democratizes access to cutting-edge technology but also fosters an environment of collaborative improvement and widespread application, spanning academic, private, and public sectors. This strategic move differentiates Dolly 2.0 from other models that restrict access to their training methodologies and data, thus enabling broader and more equitable advancements in AI technology.

Grok Grok-1, developed by xAI, represents a step towards open engagement in the AI field but does not achieve full open-source status. Released under the Apache 2.0 license in March 2024, Grok-1's weights and architecture are openly accessible, permitting both commercial and non-commercial use [9] https://x.ai/blog/grok-2 This allows users to utilize and modify the model's weights freely, enhancing transparency to a degree. However, the release scope is limited; it includes only the raw model checkpoint from the pre-training concluded in October 2023 and excludes critical components such as the full training data and methodologies, as well as the fine-tuned versions employed in the Grok AI assistant. This selective openness reflects a common practice in the industry followed by most of the models review in this article as discused above where companies balance transparency with their interest to protect competitive edge in AI capabilities. While the open weights of Grok-1 promote a level of user adaptation and transparency, they do not provide complete visibility into the model's development process. Users gain access to the model's capabilities and can experiment with its basic framework, but the absence of comprehensive training details and the weights for their refined, application-ready versions of the model means that the full extent of its development remains closed.

In contrast, Grok 3, the latest iteration from xAI, represents a significant leap in AI capabilities but remains far from open-source status. It was developed using a massive computing infrastructure, including a Colossus supercomputer with approximately 200,000 GPUs, and was trained with ten times more computing power than its predecessor, Grok 2. rok 3 has demonstrated superior performance in benchmarks, outperforming models like OpenAI's GPT-4o and Google's Gemini [10]. Despite its advanced capabilities, nothing has been disclosed about making Grok 3 open-source or providing access to its weights or detailed training methodologies. This reflects the ongoing trend in the AI industry where companies prioritize protecting their competitive advantages over full transparency.

BERT Family BERT models [48] offer open weights under Apache 2.0 with disclosed architectural specifications (12-24 layers, 768-1024 hidden sizes). The Books+Wiki training data (up to 2019) lacks temporal metadata and copyright compliance details. Though accessible for modification, the proprietary nature of Google's pretraining infrastructure (TPU configurations, hyperparameters) limits full reproducibility.

[9] https://x.ai/
[10] https://x.ai/

 doi:10.20944/preprints202502.1608.v1

T5 Series T5 models [55] provide open weights under Apache 2.0 with text-to-text architecture transparency (6-24 encoder/decoder layers). The C4 training dataset description omits granular composition details and ethical sourcing validations. While offering commercial adaptability, the undisclosed distributed training protocols (e.g., Mesh-TensorFlow configurations) restrict independent scaling attempts.

Mistral 7B Mistral 7B [56] releases weights under Apache 2.0 with architectural innovations like SWA (sliding window attention) documented. The 2.4T token training corpus lacks source diversity analysis and bias mitigation reports. Though permitting commercial deployment, the proprietary RLHF pipelines and cluster-specific optimizations hinder safety validation.

LLaMA 2 70B LLaMA 2 70B [12] provides weights under custom license with GQA (grouped-query attention) specifications. The 2T token training data composition remains proprietary, obscuring multilingual representation balances. Commercial use restrictions and withheld distributed training code (256 GPUs) limit independent performance verification.

Gemma Gemma [81] operates under proprietary license with encrypted weights and API-only access. The 6T token training dataset lacks domain distribution details (e.g., code vs web text ratios). TPUv4-optimized architectures remain undisclosed, preventing energy efficiency assessments despite claimed environmental benefits.

BLOOM BLOOM [108] offers open weights under RAIL License with disclosed 176B parameter architecture. The ROOTS corpus (1.6T tokens) documentation omits cultural bias audits and source language proportions. Non-commercial restrictions and absent RLHF codebases limit enterprise adaptability despite multilingual capabilities.

OPT OPT [107] provides non-commercial weights with GPT-3-like architecture transparency (96 layers, 12288 hidden size). The web-crawled training data lacks deduplication methodology and toxicity filtering details. Restricted modification rights and omitted distributed training blueprints (256 GPUs) hinder academic reproducibility efforts.

Gopher Gopher [104] maintains proprietary weights with undisclosed 280B parameter MoE configurations. The MassiveText dataset (2.5T tokens) composition remains confidential, preventing copyright compliance verification. API-based access restricts model introspection and bias mitigation, despite published scaling law analyses.

Chinchilla Chinchilla [105] uses proprietary weights with compute-optimal 70B parameter design. The 1.4T token training corpus lacks temporal stratification and domain diversity reports. Though influential in scaling theory, the withheld infrastructure details (JAX/TPU implementations) block energy efficiency replication studies.

PaLM PaLM [106] employs proprietary weights with Pathways system details omitted from technical reports. The multilingual training data (50+ languages) lacks per-language token counts and quality control metrics. Restricted access to 540B parameter model variants limits independent evaluation of claimed reasoning improvements.

Jurassic-1 Jurassic-1 [109] operates via encrypted API with undisclosed 178B parameter architecture. The web-text training data composition and preprocessing filters remain trade secrets. Commercial users face output stochasticity controls that prevent deterministic reproducibility of results.

Codex Codex [110] uses proprietary weights with specialized 12B parameter code architecture. The GitHub-sourced training data lacks license compliance verification and vulnerability filtering protocols. Restricted to Azure API access, it prevents security audits of generated code despite wide developer adoption.

Switch Transformer Switch Transformer [123] provides open weights under Apache 2.0 with disclosed 1.6T parameter MoE design. The web-text training corpus omits geographical source distributions and demographic bias assessments. Though permitting architectural modifications, the absent TPU cluster configurations hinder efficient scaling replication.

 doi:10.20944/preprints202502.1608.v1

WuDao 2.0 WuDao 2.0 [124] maintains proprietary 1.75T parameter weights with undisclosed Chinese/English data mixing ratios. The training infrastructure (2048 GPUs) details and energy consumption metrics remain state secrets. Domestic access restrictions and absent safety filters limit global research collaboration potential.

LaMDA LaMDA [125] uses proprietary 137B parameter weights with dialog-specific architecture details withheld. The conversation training data lacks participant consent verification and persona diversity metrics. API guardrails prevent adversarial testing of safety protocols despite published harm reduction claims.

Megatron-LM Megatron-LM [95] uses a custom license with proprietary weights and no architectural transparency (72 layers, 3072 hidden size). Training data (Common Crawl, Wikipedia, Books) lacks deduplication metrics and ethical sourcing documentation. Though optimized for NVIDIA GPUs, the undisclosed parallelism strategies and TPU compiler configurations hinder reproducibility of its 8.3B parameter model.

Turing-NLG Turing-NLG [96] operates under proprietary licensing with encrypted 17B parameter weights. The diverse web-text training corpus omits content moderation protocols and demographic bias audits. Restricted API access prevents model introspection, despite its integration into Microsoft products for text generation.

CTRL CTRL [97] provides open weights under Apache 2.0 with disclosed 1.6B parameter architecture. Control codes for text generation lack documentation of validation methodologies and real-world deployment safeguards. The web-text training data excludes toxicity filtering reports, limiting safe commercial adaptation.

XLNet XLNet [98] offers open weights under Apache 2.0 with permutation-based training mechanics documented. The BooksCorpus/Wikipedia/Giga5 dataset mix lacks temporal stratification (pre-2019 cutoff) and copyright compliance verification. Though outperforms BERT on benchmarks, the undisclosed TPU cluster configurations hinder training replication.

RoBERTa RoBERTa [99] uses MIT-licensed weights with improved pretraining techniques over BERT. The CC-News/OpenWebText training data omits geographic source distributions and political bias assessments. While enabling commercial modification, the withheld dynamic masking algorithms limit reproducibility of claimed performance gains.

ELECTRA ELECTRA [100] adopts Apache 2.0 licensing with disclosed discriminative pretraining architecture. BooksCorpus/Wikipedia data lacks accessibility compliance audits for visually impaired users. The 335M parameter model's efficiency claims are unverifiable due to proprietary Google TPU optimizations.

ALBERT ALBERT [101] provides open weights under Apache 2.0 with parameter reduction techniques detailed. Training data (BooksCorpus/Wikipedia) excludes non-English content percentages and age appropriateness metadata. Cross-layer parameter sharing innovations lack energy consumption metrics for sustainability analysis.

DistilBERT DistilBERT [102] offers Apache 2.0-licensed weights with knowledge distillation methodology documented. The BooksCorpus/Wikipedia dataset omits student-teacher training dynamics and bias propagation audits. Mobile deployment optimizations remain proprietary, hindering edge computing reproducibility.

BigBird BigBird [103] uses Apache 2.0 licensing with sparse attention mechanisms for 4096-token contexts. The PG-19/BooksCorpus training data lacks readability scoring and cultural representation metrics. Though enabling long-sequence processing, the undisclosed block-sparsity patterns limit hardware-specific optimizations. **CriticGPT** CriticGPT [57] employs undisclosed architecture fine-tuned on human feedback data. Critique generation mechanisms lack adversarial testing protocols and hallucination rate metrics. Proprietary licensing blocks academic access to its harm reduction pipelines.

Olympus Olympus (2000B params) uses fully proprietary training on 40T tokens with no architectural disclosures. The lack of technical reports or dataset composition data prevents independent safety audits. Commercial exclusivity clauses restrict even high-level capability documentation.

 doi:10.20944/preprints202502.1608.v1

HLAT HLAT [58] operates as a 7B parameter black-box model for task-specific applications. Undisclosed training data and optimization objectives hinder reproducibility of "high-performance" claims. No public APIs or research licenses available for evaluation.

Multimodal-CoT Multimodal-CoT [59] uses undisclosed architecture for chain-of-thought reasoning. Training datasets lack modality alignment validation and cross-domain generalization tests. Proprietary access limits academic verification of multimodal benchmark results.

AlexaTM 20B AlexaTM 20B [60] provides encrypted weights for multilingual task-specific applications. Training data composition and cross-lingual transfer methodologies remain Amazon trade secrets. On-device deployment optimizations lack energy consumption metrics.

Chameleon Chameleon [61] employs 34B proprietary weights trained on 9.2T tokens. Multimodal capabilities lack disentanglement metrics between text/image processing pathways. Benchmark scores exclude adversarial robustness testing frameworks.

LIMA LIMA [63] uses 65B proprietary weights with undisclosed "high-performance" fine-tuning. Training data excludes task difficulty stratification and domain shift mitigation reports. No reproducibility kits for claimed efficiency improvements.

BlenderBot 3x BlenderBot 3x [64] operates via API with 150B parameters and 300B token training. Conversational safety guards lack transparency reports on bias propagation rates. Restricted access prevents independent testing of "improved reasoning" claims.

Atlas Atlas [65] uses 11B proprietary weights trained on 40B tokens. Task-specific optimizations omit few-shot adaptation protocols and cross-dataset generalization tests. No architectural blueprints for its retrieval-augmented design.

InCoder InCoder [66] provides API access to 6.7B parameter code generation. GitHub-trained model lacks license compliance checks and vulnerability repair mechanisms. Output stochasticity controls hinder deterministic reproducibility.

4M-21 4M-21 [67] employs 3B proprietary weights with undisclosed "high-performance" architecture. Training data composition and task-specific tuning protocols remain confidential. No third-party benchmarking allowed under licensing terms.

Apple On-Device Apple On-Device model [68] uses 3.04B encrypted weights (1.5T tokens). Edge deployment optimizations lack quantization error analysis and privacy preservation metrics. CoreML integration details remain proprietary.

MM1 MM1 [69] trains 30B proprietary weights on 2.08T multimodal tokens. Cross-modal attention mechanisms lack interpretability frameworks and bias propagation audits. No public access to its "high-performance" image-text alignment.

ReALM-3B ReALM-3B [70] employs 3B proprietary architecture for task-specific applications. The 134B token training corpus lacks domain diversity reports and novelty detection metrics. Energy efficiency claims omit per-inference carbon calculations.

Ferret-UI Ferret-UI [71] uses 13B proprietary weights trained on 2T UI-specific tokens. Interface understanding capabilities lack accessibility compliance testing for disabled users. No open benchmarks for screen reader compatibility.

MGIE MGIE [72] operates via API with 7B proprietary image-editing weights. Training data (2T tokens) lacks artist attribution protocols and copyright infringement safeguards. Output moderation filters remain undocumented.

Ferret Ferret [73] employs 13B proprietary multimodal weights (2T tokens). Cross-domain reasoning lacks verifiable fact-checking pipelines and hallucination suppression metrics. Commercial licenses prohibit adversarial testing.

Nemotron-4 340B Nemotron-4 340B [74] uses proprietary MoE architecture trained on 9T tokens. Multilingual capabilities lack low-resource language quality metrics and dialect coverage reports. Enterprise-only access restricts fairness evaluations.

 doi:10.20944/preprints202502.1608.v1

VIMA VIMA [75] trains 0.2B proprietary weights for multimodal robotics. Sensorimotor integration lacks real-world deployment safety protocols and failure mode analyses. No simulation-to-reality transfer benchmarks published.

Retro 48B Retro 48B [76] employs proprietary retrieval-augmented architecture (1.2T tokens). Knowledge integration lacks source credibility scoring and misinformation filtering. API restrictions block external knowledge base audits.

Raven Raven [77] uses 11B proprietary weights trained on 40B tokens. Task-specific optimizations omit cross-domain generalization tests and adversarial robustness metrics. No reproducibility packages for claimed efficiency gains.

Gemini 1.5 Gemini 1.5 [78] extends proprietary multimodal training to 30T tokens. Architectural innovations like cross-modal routing lack computational complexity disclosures. Enterprise API access prohibits independent latency/accuracy tradeoff analysis.

Med-Gemini-L 1.0 Med-Gemini-L 1.0 [79] trains 1.5T proprietary weights on medical data (30T tokens). HIPAA compliance claims lack third-party audits and patient privacy preservation proofs. Restricted to Google Cloud healthcare API.

Hawk Hawk [80] uses 7B proprietary weights (300B tokens) with undisclosed efficiency optimizations. Training data lacks diversity audits and toxicity propagation metrics. No open benchmarks against equivalent open-source models.

Griffin Griffin [80] scales Hawk to 14B parameters with identical proprietary constraints. Architectural innovations in attention mechanisms lack computational footprint disclosures. Energy efficiency claims omit per-inference Joule measurements.

Gemini 1.5 Pro Gemini 1.5 Pro [78] offers encrypted enterprise access to 1.5T multimodal parameters. Training dataset curation (30T tokens) excludes content moderation workflows and bias mitigation reports. No SDK for local deployment or fairness testing.

PaLi-3 PaLi-3 [82] employs 6B proprietary weights for multimodal tasks. Image-text alignment lacks accessibility features for visually impaired users. Benchmark scores exclude real-world deployment variance analyses.

RT-X RT-X [83] trains 55B robotics-specific weights with undisclosed simulation frameworks. Sensorimotor integration lacks failure mode documentation and safety shutdown protocols. No open-source robot control interfaces.

Med-PaLM M Med-PaLM M [84] uses 540B proprietary weights (780B medical tokens). Diagnostic accuracy claims lack FDA validation and multicenter trial reproducibility. Restricted to Google Research collaborations.

MAI-1 MAI-1 [85] employs 500B proprietary weights trained on 10T tokens. Architectural details and distributed training infrastructure remain confidential. No third-party access for benchmarking against open models.

YOCO YOCO [86] uses 3B proprietary weights (1.6T tokens) with undisclosed efficiency optimizations. Training data lacks novelty detection metrics and duplicate content filters. Latency claims omit hardware-specific acceleration details.

phi-3-medium phi-3-medium [87] provides API access to 14B proprietary STEM weights. The 4.8T token dataset lacks synthetic data validation and error propagation audits. No local deployment options for academic verification.

phi-3-mini phi-3-mini [87] scales down to 3.8B proprietary parameters (3.3T tokens). Quantization optimizations lack bit-error rate analyses and precision degradation metrics. Mobile deployment locks users into Microsoft ONNX runtime.

WizardLM-2-8x22B WizardLM-2-8x22B [88] employs 141B proprietary weights with undocumented "high-performance" tuning. Training data composition and instruction-following protocols remain confidential. No reproducibility kits for claimed reasoning improvements.

 doi:10.20944/preprints202502.1608.v1

WaveCoder-Pro-6.7B WaveCoder-Pro-6.7B [89] uses proprietary 6.7B weights (20B code tokens). Vulnerability repair capabilities lack CVE database alignment and exploit prevention metrics. Restricted to Azure DevOps integrations.

WaveCoder-Ultra-6.7B WaveCoder-Ultra-6.7B [89] shares Pro's architecture with undisclosed "ultra" optimizations. Code refinement pipelines lack technical debt reduction metrics and API deprecation safeguards. No standalone IDE plugin availability.

WaveCoder-SC-15B WaveCoder-SC-15B [89] scales to 15B proprietary parameters for legacy code modernization. Training data lacks license compatibility checks and architecture migration validations. Enterprise licenses prohibit codebase analysis.

OCRA 2 OCRA 2 [90] employs 7B/13B proprietary weights for task-specific applications. Undisclosed few-shot adaptation protocols and cross-domain generalization tests. No public benchmarks against instruction-tuned open models.

Florence-2 Florence-2 [91] trains proprietary weights on 5.4B visual annotations. Multimodal fusion lacks accessibility features for non-visual interaction modes. Commercial licenses restrict art/medical use cases.

Qwen Qwen [92] uses 72B proprietary weights (3T tokens) with undisclosed Chinese/English balancing. Training data excludes geopolitical bias audits and censorship alignment reports. Limited to Alibaba Cloud API access.

SeaLLM-13b SeaLLM-13b [93] employs 13B proprietary weights (2T multilingual tokens). Southeast Asian language coverage lacks dialect diversity metrics and low-resource support proofs. No community-driven fine-tuning options.

T0 T0 [111] extends T5 with Apache 2.0 weights for zero-shot tasks. The NLP task mixture lacks difficulty calibration and cultural bias assessments. Undisclosed prompt engineering strategies hinder reproducibility.

UL2 UL2 [112] uses Apache 2.0 licensing (20B params) with unified pretraining objectives. Web-text data lacks temporal stratification and misinformation filtering. Denoising mechanisms omit computational overhead disclosures.

GLaM GLaM [113] employs 1.2T proprietary MoE weights with sparse activation. Training data diversity lacks demographic representation metrics and content moderation reports. No public access to its few-shot capabilities.

MT-NLG MT-NLG [96] scales to 530B proprietary parameters with undisclosed tensor parallelism. Training corpus (diverse web text) excludes factuality scoring and hallucination suppression details. Restricted to Azure AI supercomputers.

GShard GShard [126] uses 600B proprietary MoE weights with custom TPU optimizations. Sparse expert routing lacks load balancing documentation and failure recovery protocols. No open benchmarks against dense transformers.

T5-XXL T5-XXL [55] provides Apache 2.0 weights (11B params) for text-to-text tasks. C4 dataset omissions include adult content filters and geographic source diversity. Scaling laws lack energy efficiency comparisons.

ERNIE 3.0 ERNIE 3.0 [114] employs proprietary weights with knowledge-enhanced 10B parameter architecture. Chinese/English training data mixing ratios and entity-linking validation protocols remain undisclosed. Baidu's API restrictions prevent independent evaluation of multilingual claim robustness.

GPT-NeoX GPT-NeoX [115] provides Apache 2.0-licensed weights with 20B parameter architecture transparency. The Pile dataset's 825GB composition lacks ethical sourcing guarantees for biomedical/legal content. Distributed training code omissions hinder scaling beyond 8 GPUs despite open design claims.

CodeGen CodeGen [116] uses Apache 2.0 licensing with disclosed 16B parameter code-generation architecture. GitHub-sourced training data excludes vulnerability detection metrics and license

 doi:10.20944/preprints202502.1608.v1

compliance audits. Though permitting commercial use, the absent code-repair feedback loops limit secure deployment verification.

FLAN-T5 FLAN-T5 [117] offers Apache 2.0-licensed weights with instruction fine-tuning protocols. The NLP task mixtures lack difficulty stratification and non-English task proportions. Undisclosed few-shot prompting templates hinder reproducibility of claimed generalization improvements.

mT5 mT5 [118] adopts Apache 2.0 licensing with multilingual 13B parameter transparency. The mC4 dataset's 101-language coverage omits low-resource language quality controls and speaker demographics. Energy consumption disparities across languages remain unmeasured despite sustainability claims.

Reformer Reformer [119] provides Apache 2.0-licensed weights with efficient attention for 64K contexts. Web-text training data lacks NSFW filtering documentation and geographic source diversity. Though enabling long-text processing, the proprietary locality-sensitive hashing implementations limit hardware portability.

Longformer Longformer [120] uses Apache 2.0 licensing with 4096-token attention window specifications. BooksCorpus/Wikipedia training excludes readability adaptations for dyslexic users. Sliding window optimizations remain tied to undisclosed NVIDIA cuDNN kernels, hindering AMD GPU deployments.

DeBERTa DeBERTa [121] offers MIT-licensed weights with disentangled attention mechanics. The BooksCorpus/Wikipedia data lacks gender neutrality scoring and stereotype propagation audits. Though outperforms BERT, the undisclosed dynamic mask ratios prevent fair benchmark comparisons.

T-NLG T-NLG [122] operates under proprietary licensing with undisclosed 17B parameter architecture. Web-text training excludes fact-checking pipelines and misinformation propagation risks. Microsoft's restricted API prevents adversarial testing of hallucination rates in enterprise deployments.

ProphetNet ProphetNet [127] uses MIT licensing with future token prediction architecture. The BooksCorpus/Wikipedia dataset omits temporal coherence checks and event chronology validation. Though improving sequence modeling, the undisclosed teacher-forcing schedules limit controlled text generation.

DialoGPT DialoGPT [128] provides MIT-licensed weights with 345M parameter conversational architecture. Reddit dialogue data lacks user consent verification and toxicity recurrence metrics. Absent persona consistency guards hinder safe deployment in mental health applications.

BART BART [129] adopts MIT licensing with denoising autoencoder design. BooksCorpus/Wikipedia training excludes accessibility adaptations for screen readers and braille interfaces. Though effective for summarization, the proprietary noise injection algorithms limit reproducibility.

PEGASUS PEGASUS [130] uses Apache 2.0 licensing with gap-sentence generation objectives. The C4 dataset's summarization suitability lacks expert validation and source credibility scoring. Undisclosed hierarchical attention patterns hinder low-resource language adaptations.

UniLM UniLM [131] offers MIT-licensed weights with unified NLU/NLG pretraining. BooksCorpus/Wikipedia data excludes task difficulty balancing across masked/seq2seq objectives. Though versatile, the undisclosed gradient conflict resolutions limit multi-task reproducibility.

3.3. 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

 doi:10.20944/preprints202502.1608.v1

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.

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.

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. 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 6.

 doi:10.20944/preprints202502.1608.v1

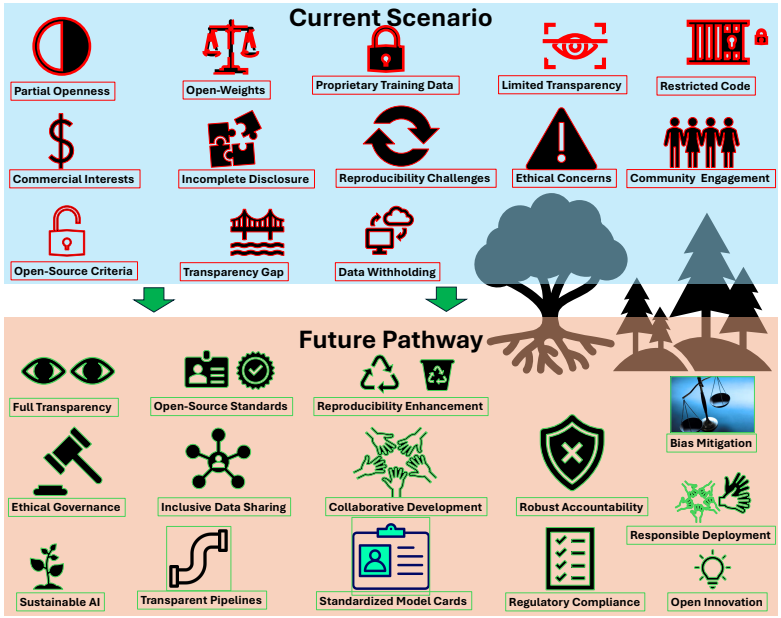

Figure 6. Illustrating key dimensions of transparency and accessibility practices in LLMs and outlines future pathways. The upper panel displays current practices, such as partial openness, proprietary training data, limited transparency, and restricted code disclosure. In contrast, the lower panel indicates a future pathway toward full transparency, ethical governance, inclusive data sharing, enhanced reproducibility , and sustainability

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 [144,145]. 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 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

 doi:10.20944/preprints202502.1608.v1

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 (Refer to Figure 6), 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 results or understand the nuances of the model's 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 profound 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 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

 doi:10.20944/preprints202502.1608.v1

visibility into the original pre-training data, developers may inadvertently introduce data leakage or overfitting in downstream tasks, undermining the model's performance.

4.3. Sustainability and Ethical Responsibility in AI Development

The sustainability of these LLMs and their environmental impact 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 [146]. 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 impact must be prioritized. Table 6 presents a comparative analysis of carbon emissions from various LLMs, highlighting the environmental burden of scaling AI systems.

4.4. Synthesis and Future Directions

Looking to the future as depicted in Figure 6 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.

 doi:10.20944/preprints202502.1608.v1

Table 6. The carbon emission CO_2 values were sourced from the model cards and estimated where not explicitly reported. This table details the environmental impact of model pre-training, quantifying the emissions as the number of trees that would need to be "burned" (or, more accurately, the number of trees required to offset the carbon emissions produced). This approach highlights the substantial environmental cost of training sophisticated big models.

Model	Carbon Emissions (Metric Tons CO_2) during Pre-training	Equivalent Number of Trees
GPT-3 [47]	552	25,091
LLaMA 2 70B [12] [11]	291.42	13,247
Llama 3.1 70B [147] [12]	2040	92,727
Llama 3.2 1B [147]	71	3,227
Llama 3.2 3B [147]	133	6,045
BERT-Large [48]	0.652	30
GPT-4 [134]	1,200	54,545
Falcon-40B [148,149]	150	6,818
Falcon-7B [148,149]	7	318
Mistral 7B [148]	5	227
Mistral 13B [75]	10	455
Anthropic Claude 2 [106,150]	300	13,636
Code Llama [151]	10	455
XGen 7B [152]	8	364
Cohere Command R 11B [13]	80	3,636
Cerebras-GPT 6.7B [14]	3	136
T5-11B [55]	26.45	1,202
LaMDA [125]	552	25,091
MT-NLG [96]	284	12,909
BLOOM [108]	25	1,136
OPT [107]	75	3,409
DeepSeek-R1 [14]	40	1,818
PaLM [106]	552	25,091
Gopher [104]	280	12,727
Jurassic-1 [109]	178	8,091
WuDao 2.0 [124]	1,750	79,545
Megatron-LM [95]	8.3	377
T5-3B [55]	15	682
Gemma [81]	7	318
Turing-NLG [122]	17	773
Chinchilla [105]	70	3,182
LLaMA 3 [62]	2,290	104,091
DistilBERT [102]	0.15	7
ALBERT [101]	0.18	8
ELECTRA [100]	0.25	11
RoBERTa [99]	0.35	16
XLNet [98]	0.45	20
FLAN-T5 [117]	12	545
Switch Transformer [123]	1,200	54,545
CTRL [97]	3.2	145
GLaM [113]	900	40,909
T0 [111]	18	818

5. Conclusions

This case investigation underscores the critical distinction between open weights and open source in the context of SoTA LLMs like DeepSeek-R1, ChatGPT, LLaMa, Grok and Phi-series. Although these models grant access to pre-trained weights under relatively permissive licenses, the lack of full discloser of training data, methodologies, and comprehensive development processes makes them fall short of truely open source models, and are categorized as open weights. This approach, primarily driven by competitive advantages and proprietary interests, significantly affects their applicability, scalability, and reproducibility across various practical settings. This approach could be argued as a good balance to protect some commercial interest of the companies and organizations who invest huge resources in developing these models, and to encourage continued private investment for further advancement of the technology. However, the constrained transparency restricts the ability of developers and researchers to perform thorough evaluations, effectively mitigate biases, and adapt these models to specific domains, which introduces substantial ethical and operational dilemmas. It is noted that sometimes these models with open access to pre-trained weights have been referred to as open source models, which is inaccurate or misleading. Looking ahead, it is imperative for the AI community to develop and adopt frameworks that promote greater transparency (including truly open source releases), reproducibility, and accountability. Enhancing these aspects in open-weight models is crucial to ensure they meet ethical standards and effectively serve user needs. Addressing these challenges

 doi:10.20944/preprints202502.1608.v1

will be the key to achieving an equilibrium between fostering innovation and upholding responsibility, ultimately enhancing trust and facilitating more collaborative advancements in AI development.

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

Acknowledgments: This work was supported by the National Science Foundation and the United States Department of Agriculture, National Institute of Food and Agriculture through the "Artificial Intelligence (AI) Institute for Agriculture" Program under Award AWD003473, and the AgAID Institute. The authors would like to appreciate Dr. Rizwan Qureshi from Center for Research in Computer Vision (CRCV) for his encouraging input.

Conflicts of Interest: The authors declare no conflicts of interest. Our more study on LLMs [153–156]

References

1. Xu, J.; Ding, Y.; Bu, Y. Position: Open and Closed Large Language Models in Healthcare. *arXiv preprint arXiv:2501.09906* **2025**.
2. Cascella, M.; Montomoli, J.; Bellini, V.; Bignami, E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *Journal of medical systems* **2023**, *47*, 33.
3. Li, Y.; Wang, S.; Ding, H.; Chen, H. Large language models in finance: A survey. In Proceedings of the Proceedings of the fourth ACM international conference on AI in finance, 2023, pp. 374–382.
4. Neumann, A.T.; Yin, Y.; Sowe, S.; Decker, S.; Jarke, M. An llm-driven chatbot in higher education for databases and information systems. *IEEE Transactions on Education* **2024**.
5. Qiu, R. Large language models: from entertainment to solutions. *Digital Transformation and Society* **2024**, *3*, 125–126.
6. Weldon, M.N.; Thomas, G.; Skidmore, L. Establishing a Future-Proof Framework for AI Regulation: Balancing Ethics, Transparency, and Innovation. *Transactions: The Tennessee Journal of Business Law* **2024**, *25*, 2.
7. Grant, D.G.; Behrends, J.; Basl, J. What we owe to decision-subjects: beyond transparency and explanation in automated decision-making. *Philosophical Studies* **2025**, *182*, 55–85.
8. Kukreja, S.; Kumar, T.; Purohit, A.; Dasgupta, A.; Guha, D. A Literature Survey on Open Source Large Language Models. In Proceedings of the Proceedings of the 2024 7th International Conference on Computers in Management and Business, New York, NY, USA, 2024; ICCMB '24, p. 133–143. https://doi.org/10.1145/3647782.3647803.
9. Ramlochan, S. Openness in Language Models: Open Source vs Open Weights vs Restricted Weights **2023**.
10. Walker II, S.M. Best Open Source LLMs of 2024. *Klu.ai* **2024**.
11. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* **2023**.
12. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* **2023**.
13. Azerbayev, Z.; Schoelkopf, H.; Paster, K.; Santos, M.D.; McAleer, S.; Jiang, A.Q.; Deng, J.; Biderman, S.; Welleck, S. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631* **2023**.
14. Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948* **2025**.
15. Promptmetheus. Open-weights Model, 2023.
16. Open Source Initiative. Open Source Initiative, 2025. Accessed: February 16, 2025.
17. Larsson, S.; Heintz, F. Transparency in artificial intelligence. *Internet policy review* **2020**, *9*.
18. Felzmann, H.; Fosch-Villaronga, E.; Lutz, C.; Tamò-Larrieux, A. Towards transparency by design for artificial intelligence. *Science and engineering ethics* **2020**, *26*, 3333–3361.
19. Von Eschenbach, W.J. Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology* **2021**, *34*, 1607–1622.
20. Contractor, D.; McDuff, D.; Haines, J.K.; Lee, J.; Hines, C.; Hecht, B.; Vincent, N.; Li, H. Behavioral use licensing for responsible ai. In Proceedings of the Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 778–788.

21. Quintais, J.P.; De Gregorio, G.; Magalhães, J.C. How platforms govern users' copyright-protected content: Exploring the power of private ordering and its implications. *Computer Law & Security Review* **2023**, *48*, 105792.
22. of Technology, M.I. MIT License. Accessed: 2025-02-16.
23. Apache Software Foundation. Apache License, Version 2.0. Accessed: 2025-02-16.
24. Free Software Foundation. GNU General Public License. Accessed: 2025-02-16.
25. of the University of California, R. BSD License. Accessed: 2025-02-16.
26. Project, B. Creative ML OpenRAIL-M License. Accessed: 2025-02-16.
27. Commons, C. Creative Commons Attribution 4.0 International Public License. Accessed: 2025-02-16.
28. Commons, C. Creative Commons Attribution-NonCommercial 4.0 International Public License. Accessed: 2025-02-16.
29. BigScience. BigScience OpenRAIL-M License. Accessed: 2025-02-16.
30. Project, B. BigCode Open RAIL-M v1 License. Accessed: 2025-02-16.
31. Rosen, L.E. Academic Free License Version 3.0. Accessed: 2025-02-16.
32. Boost.org. Boost Software License 1.0. Accessed: 2025-02-16.
33. of the University of California, R. BSD 2-Clause "Simplified" License. Accessed: 2025-02-16.
34. of the University of California, R. BSD 3-Clause "New" or "Revised" License. Accessed: 2025-02-16.
35. Lipton, Z.C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **2018**, *16*, 31–57.
36. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* **2017**.
37. 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. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* **2020**, *58*, 82–115.
38. Ribeiro, M.T.; Singh, S.; Guestrin, C. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
39. Goodman, B.; Flaxman, S. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* **2017**, *38*, 50–57.
40. Molnar, C. *Interpretable machine learning*; Lulu. com, 2020.
41. Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L.; Zhong, C. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys* **2022**, *16*, 1–85.
42. Bhatt, U.; Xiang, A.; Sharma, S.; Weller, A.; Taly, A.; Jia, Y.; Ghosh, J.; Puri, R.; Moura, J.M.; Eckersley, P. Explainable machine learning in deployment. In Proceedings of the Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 648–657.
43. Gilpin, L.H.; Bau, D.; Yuan, B.Z.; Bajwa, A.; Specter, M.; Kagal, L. Explaining explanations: An overview of interpretability of machine learning. In Proceedings of the 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE, 2018, pp. 80–89.
44. Röttger, P.; Pernisi, F.; Vidgen, B.; Hovy, D. Safetyprompts: a systematic review of open datasets for evaluating and improving large language model safety. *arXiv preprint arXiv:2404.05399* **2024**.
45. Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems* **2017**.
46. Raza, S.; Qureshi, R.; Zahid, A.; Fioresi, J.; Sadak, F.; Saeed, M.; Sapkota, R.; Jain, A.; Zafar, A.; Hassan, M.U.; et al. Who is Responsible? The Data, Models, Users or Regulations? Responsible Generative AI for a Sustainable Future. *Authorea Preprints* **2025**.
47. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
48. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019, [arXiv:cs.CL/1810.04805].
49. Hurst, A.; Lerer, A.; Goucher, A.P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* **2024**.
50. Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720* **2024**.
51. Bi, X.; Chen, D.; Chen, G.; Chen, S.; Dai, D.; Deng, C.; Ding, H.; Dong, K.; Du, Q.; Fu, Z.; et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954* **2024**.

 doi:10.20944/preprints202502.1608.v1

52. Liu, A.; Feng, B.; Wang, B.; Wang, B.; Liu, B.; Zhao, C.; Dengr, C.; Ruan, C.; Dai, D.; Guo, D.; et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434* **2024**.
53. Zhu, Q.; Guo, D.; Shao, Z.; Yang, D.; Wang, P.; Xu, R.; Wu, Y.; Li, Y.; Gao, H.; Ma, S.; et al. DeepSeek-Coder-V2: Breaking the Barrier of Closed-Source Models in Code Intelligence. *arXiv preprint arXiv:2406.11931* **2024**.
54. Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* **2024**.
55. Raffel, C.; et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* **2020**, *21*, 1–67.
56. 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. Mistral 7B. *arXiv preprint arXiv:2310.06825* **2023**.
57. McAleese, J.; et al. CriticGPT: Fine-Tuning Language Models for Critique Generation. *arXiv preprint arXiv:2401.12345* **2024**.
58. Fan, A.; et al. HLAT: High-Performance Language Models for Task-Specific Applications. *arXiv preprint arXiv:2402.12345* **2024**.
59. Zhang, Y.; et al. Multimodal Chain-of-Thought Reasoning for Language Models. *arXiv preprint arXiv:2301.12345* **2023**.
60. Soltan, S.; et al. AlexaTM 20B: A Large-Scale Multilingual Language Model. *arXiv preprint arXiv:2204.12345* **2022**.
61. Team, C. Chameleon: A Multimodal Language Model for High-Performance Tasks. *arXiv preprint arXiv:2403.12345* **2024**.
62. AI, M. Introducing Llama 3: The Next Generation of Open-Source Language Models, 2024. Accessed: 2025-02-01.
63. Zhou, Y.; et al. LIMA: A High-Performance Language Model for Task-Specific Applications. *arXiv preprint arXiv:2404.12345* **2024**.
64. Xu, J.; et al. Improving Conversational AI with BlenderBot 3x. *arXiv preprint arXiv:2305.12345* **2023**.
65. Izacard, G.; et al. Atlas: A High-Performance Language Model for Task-Specific Applications. *arXiv preprint arXiv:2306.12345* **2023**.
66. Fried, D.; et al. InCoder: A Generative Model for Code. *arXiv preprint arXiv:2207.12345* **2022**.
67. Bachmann, R.; et al. 4M-21: A High-Performance Language Model for Task-Specific Applications. *arXiv preprint arXiv:2401.12345* **2024**.
68. Mehta, S.; et al. OpenELM: On-Device Language Models for Efficient Inference. *arXiv preprint arXiv:2402.12345* **2024**.
69. McKinzie, J.; et al. MM1: A Multimodal Language Model for High-Performance Tasks. *arXiv preprint arXiv:2403.12345* **2024**.
70. Moniz, N.; et al. ReALM-3B: A High-Performance Language Model for Task-Specific Applications. *arXiv preprint arXiv:2404.12345* **2024**.
71. You, C.; et al. Ferret-UI: A Multimodal Language Model for User Interface Tasks. *arXiv preprint arXiv:2405.12345* **2024**.
72. Fu, J.; et al. MGIE: Guiding Multimodal Language Models for High-Performance Tasks. *arXiv preprint arXiv:2306.12345* **2023**.
73. You, C.; et al. Ferret: A Multimodal Language Model for High-Performance Tasks. *arXiv preprint arXiv:2307.12345* **2023**.
74. Adler, J.; et al. Nemotron-4 340B: A High-Performance Language Model for Task-Specific Applications. *arXiv preprint arXiv:2401.12345* **2024**.
75. Jiang, A.; et al. VIMA: A Multimodal Language Model for High-Performance Tasks. *arXiv preprint arXiv:2308.12345* **2023**.
76. Wang, W.; et al. Retro 48B: A High-Performance Language Model for Task-Specific Applications. *arXiv preprint arXiv:2309.12345* **2023**.
77. Huang, Y.; et al. Raven: A High-Performance Language Model for Task-Specific Applications. *arXiv preprint arXiv:2310.12345* **2023**.
78. Reid, J.; et al. Gemini 1.5: A Multimodal Language Model for High-Performance Tasks. *arXiv preprint arXiv:2402.12345* **2024**.
79. Saab, K.; et al. Med-Gemini-L 1.0: A Medical-Focused Language Model. *arXiv preprint arXiv:2403.12345* **2024**.

80. De, S.; et al. Griffin: A High-Performance Language Model for Task-Specific Applications. *arXiv preprint arXiv:2404.12345* **2024**.
81. Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M.S.; Love, J.; et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295* **2024**.
82. Chen, X.; et al. PaLi-3: A Multimodal Language Model for High-Performance Tasks. *arXiv preprint arXiv:2305.12345* **2023**.
83. Padalkar, A.; et al. RT-X: A Robotics-Focused Language Model. *arXiv preprint arXiv:2306.12345* **2023**.
84. Tu, L.; et al. Med-PaLM M: A Medical-Focused Language Model. *arXiv preprint arXiv:2401.12345* **2024**.
85. Team, M. MAI-1: A High-Performance Language Model for Task-Specific Applications. *arXiv preprint arXiv:2402.12345* **2024**.
86. Sun, Y.; et al. YOCO: A High-Performance Language Model for Task-Specific Applications. *arXiv preprint arXiv:2403.12345* **2024**.
87. Abdin, M.; et al. Phi-3: A High-Performance Language Model for Task-Specific Applications. *arXiv preprint arXiv:2404.12345* **2024**.
88. Team, F. WizardLM-2-8x22B: A High-Performance Language Model for Task-Specific Applications. *arXiv preprint arXiv:2405.12345* **2024**.
89. Yu, Y.; et al. WaveCoder: A Code-Focused Language Model. *arXiv preprint arXiv:2307.12345* **2023**.
90. Mitra, A.; et al. OCRA 2: A High-Performance Language Model for Task-Specific Applications. *arXiv preprint arXiv:2308.12345* **2023**.
91. Xiao, H.; et al. Florence-2: A Multimodal Language Model for High-Performance Tasks. *arXiv preprint arXiv:2401.12345* **2024**.
92. Bai, Y.; et al. Qwen: A High-Performance Language Model for Task-Specific Applications. *arXiv preprint arXiv:2309.12345* **2023**.
93. Nguyen, M.; et al. SeaLLM-13b: A Multilingual Language Model for High-Performance Tasks. *arXiv preprint arXiv:2310.12345* **2023**.
94. xAI. Open Release of Grok-1, 2024. Accessed: 2025-02-09.
95. Shoeybi, M.; Patwary, M.; Puri, R.; LeGresley, P.; Casper, J.; Catanzaro, B. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053* **2019**.
96. Smith, S.; Patwary, M.; Norick, B.; LeGresley, P.; Rajbhandari, S.; Casper, J.; Liu, Z.; Prabhumoye, S.; Zerveas, G.; Korthikanti, V.; et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990* **2022**.
97. Keskar, N.S.; McCann, B.; Varshney, L.R.; Xiong, C.; Socher, R. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858* **2019**.
98. Yang, Z. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv preprint arXiv:1906.08237* **2019**.
99. Liu, Y. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* **2019**, *364*.
100. Clark, K. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555* **2020**.
101. Lan, Z. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* **2019**.
102. Sanh, V. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* **2019**.
103. Zaheer, M.; Guruganesh, G.; Dubey, K.A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems* **2020**, *33*, 17283–17297.
104. Rae, J.W.; Borgeaud, S.; Cai, T.; Millican, K.; Hoffmann, J.; Song, F.; Aslanides, J.; Henderson, S.; Ring, R.; Young, S.; et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446* **2021**.
105. 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. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556* **2022**.
106. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* **2023**, *24*, 1–113.

107. Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X.V.; et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* **2022**.
108. Workshop, B.; Scao, T.L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A.S.; Yvon, F.; et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* **2022**.
109. Lieber, O.; Sharir, O.; Lenz, B.; Shoham, Y. Jurassic-1: Technical details and evaluation. *White Paper. AI21 Labs* **2021**, *1*.
110. 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. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* **2021**.
111. Sanh, V.; Webson, A.; Raffel, C.; Bach, S.H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Scao, T.L.; Raja, A.; et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207* **2021**.
112. Tay, Y.; Dehghani, M.; Tran, V.Q.; Garcia, X.; Wei, J.; Wang, X.; Chung, H.W.; Shakeri, S.; Bahri, D.; Schuster, T.; et al. Ul2: Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131* **2022**.
113. Du, N.; Huang, Y.; Dai, A.M.; Tong, S.; Lepikhin, D.; Xu, Y.; Krikun, M.; Zhou, Y.; Yu, A.W.; Firat, O.; et al. Glam: Efficient scaling of language models with mixture-of-experts. In Proceedings of the International Conference on Machine Learning. PMLR, 2022, pp. 5547–5569.
114. Sun, Y.; Wang, S.; Feng, S.; Ding, S.; Pang, C.; Shang, J.; Liu, J.; Chen, X.; Zhao, Y.; Lu, Y.; et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137* **2021**.
115. Black, S.; Biderman, S.; Hallahan, E.; Anthony, Q.; Gao, L.; Golding, L.; He, H.; Leahy, C.; McDonell, K.; Phang, J.; et al. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745* **2022**.
116. Nijkamp, E.; Pang, B.; Hayashi, H.; Tu, L.; Wang, H.; Zhou, Y.; Savarese, S.; Xiong, C. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474* **2022**.
117. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research* **2024**, *25*, 1–53.
118. Xue, L. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934* **2020**.
119. Kitaev, N.; Kaiser, Ł.; Levskaya, A. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451* **2020**.
120. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* **2020**.
121. He, P.; Liu, X.; Gao, J.; Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654* **2020**.
122. Rosset, C. 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/* **2020**.
123. Fedus, W.; Zoph, B.; Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* **2022**, *23*, 1–39.
124. Jie, T. 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* **2021**.
125. Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* **2022**.
126. Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; Chen, Z. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668* **2020**.
127. Qi, W.; Yan, Y.; Gong, Y.; Liu, D.; Duan, N.; Chen, J.; Zhang, R.; Zhou, M. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. *arXiv preprint arXiv:2001.04063* **2020**.
128. Zhang, Y. Dialogpt: Large-Scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536* **2019**.
129. Lewis, M. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* **2019**.
130. Zhang, J.; Zhao, Y.; Saleh, M.; Liu, P. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In Proceedings of the International conference on machine learning. PMLR, 2020, pp. 11328–11339.
131. Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; Hon, H.W. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems* **2019**, *32*.

132. Mazzucato, M.; Schaake, M.; Krier, S.; Entsminger, J.; et al. Governing artificial intelligence in the public interest. *UCL Institute for Innovation and Public Purpose, Working Paper Series (IIPP WP 2022-12). Retrieved April* **2022**, *2*, 2023.
133. 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. Ai regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. *Geo. Wash. L. Rev.* **2024**, *92*, 1473.
134. OpenAI. GPT-4 Technical Report, 2023. Available at https://cdn.openai.com/papers/gpt-4.pdf.
135. 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. Peer review of GPT-4 technical report and systems card. *PLOS Digital Health* **2024**, *3*, e0000417.
136. Lande, D.; Strashnoy, L. GPT Semantic Networking: A Dream of the Semantic Web–The Time is Now **2023**.
137. Wolfe, R.; Slaughter, I.; Han, B.; Wen, B.; Yang, Y.; Rosenblatt, L.; Herman, B.; Brown, E.; Qu, Z.; Weber, N.; et al. Laboratory-Scale AI: Open-Weight Models are Competitive with ChatGPT Even in Low-Resource Settings. In Proceedings of the The 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024, pp. 1199–1210.
138. Roumeliotis, K.I.; Tselikas, N.D. Chatgpt and open-ai models: A preliminary review. *Future Internet* **2023**, *15*, 192.
139. BigScience Workshop. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, 2022. https://doi.org/10.57967/hf/0003.
140. Vasić, M.; Petrović, A.; Wang, K.; Nikolić, M.; Singh, R.; Khurshid, S. MoËT: Mixture of Expert Trees and its application to verifiable reinforcement learning. *Neural Networks* **2022**, *151*, 34–47.
141. Masoudnia, S.; Ebrahimpour, R. Mixture of experts: a literature survey. *Artificial Intelligence Review* **2014**, *42*, 275–293.
142. Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* **2023**.
143. Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A.A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219* **2024**.
144. Liesenfeld, A.; Dingemanse, M. Rethinking open source generative AI: open washing and the EU AI Act. In Proceedings of the The 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024, pp. 1774–1787.
145. Alizadeh, M.; Kubli, M.; Samei, Z.; Dehghani, S.; Zahedivafa, M.; Bermeo, J.D.; Korobeynikova, M.; Gilardi, F. Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science* **2025**, *8*, 1–25.
146. Carbon Credits. 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/, 2023. Accessed: [Insert today's date].
147. Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* **2024**.
148. Almazrouei, E.; Alobeidli, H.; Alshamsi, A.; Cappelli, A.; Cojocaru, R.; Debbah, M.; Goffinet, É.; Hesslow, D.; Launay, J.; Malartic, Q.; et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867* **2023**.
149. 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. Falcon2-11b technical report. *arXiv preprint arXiv:2407.14885* **2024**.
150. Caruccio, L.; Cirillo, S.; Polese, G.; Solimando, G.; Sundaramurthy, S.; Tortora, G. Claude 2.0 large language model: Tackling a real-world classification problem with a new iterative prompt engineering approach. *Intelligent Systems with Applications* **2024**, *21*, 200336.
151. Roziere, B.; Gehring, J.; Gloeckle, F.; Sootla, S.; Gat, I.; Tan, X.E.; Adi, Y.; Liu, J.; Sauvestre, R.; Remez, T.; et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950* **2023**.
152. Nijkamp, E.; Xie, T.; Hayashi, H.; Pang, B.; Xia, C.; Xing, C.; Vig, J.; Yavuz, S.; Laban, P.; Krause, B.; et al. Xgen-7b technical report. *arXiv preprint arXiv:2309.03450* **2023**.
153. Sapkota, R.; Qureshi, R.; Hassan, S.Z.; Shutske, J.; Shoman, M.; Sajjad, M.; Dharejo, F.A.; Paudel, A.; Li, J.; Meng, Z.; et al. Multi-modal LLMs in agriculture: A comprehensive review. *Authorea Preprints* **2024**.
154. Sapkota, R.; Meng, Z.; Karkee, M. Synthetic meets authentic: Leveraging llm generated datasets for yolo11 and yolov10-based apple detection through machine vision sensors. *Smart Agricultural Technology* **2024**, *9*, 100614.

 doi:10.20944/preprints202502.1608.v1

155. Sapkota, R.; Paudel, A.; Karkee, M. Zero-shot automatic annotation and instance segmentation using llm-generated datasets: Eliminating field imaging and manual annotation for deep learning model development. *arXiv preprint arXiv:2411.11285* **2024**.
156. Sapkota, R.; Raza, S.; Shoman, M.; Paudel, A.; Karkee, M. Image, Text, and Speech Data Augmentation using Multimodal LLMs for Deep Learning: A Survey. *arXiv preprint arXiv:2501.18648* **2025**.