Comparative Study of Leading AI Models: Performance, Capabilities, Accuracy, and Applications

Chintan Hirpara1,Khushi Kundariya2,Mahipal Chothani3,Deep Solanki4,Jigar Gajjar5

LokJagruti Kendra University, Ahmedabad, GJ, India[1,2,3]
TechDCybersecurity Limited, Ahmedabad, GJ, India[4,5]

chintanhirpara2707@gmail.com[1], kundariyakhushi0404@gmail.com[2], mahipalchothani5@gmail.com[3], deepsolanki5799@gmail.com[4],cyberian.jg@gmail.com[5]

Abstract:Large language models (LLMs) have progressed from mere academic prototypes to mainstream solutions in education, healthcare, software engineering, and customer service industries. In this paper, we conduct a comparative review of the architecture, reasoning capability, code generation, creative writing capabilities, ethics, and adoption level of five of the most popular conversational artificial intelligence platforms: ChatGPT, DeepSeek, Gemini, Claude, and Grok. Instead of performing our own benchmarking experiments, we synthesize results from the literature reviewed in Section II and scores reported by the AI companies in question on a number of commonly used benchmarking platforms including MMLU, GPQA Diamond, SWE-bench Verified, HumanEval, and GSM8K, supplemented by usage metrics from independent sources. Our comparative analysis shows that none of the examined models outperforms the rest in all respects; in particular, Gemini and DeepSeek score the highest on general knowledge tests, Claude demonstrates the best performance on code-generation and reasoning about long documents, Grok stands out in difficult agentic reasoning tests and real-time information lookup tasks, and ChatGPT remains the most used one. We conclude the paper by outlining the limitations of literature-based comparative analysis and suggesting topics for future research.

Keywords:Artificial Intelligence language models, ChatGPT, DeepSeek, Gemini AI, Claude AI, Grok AI, performance comparison, accuracy evaluation, real-world applications, AI capabilities.

I. INTRODUCTION

1.1 Introduction to AI Language Models

AI has altered human interaction with technology, and AI-based language models are currently the focal point of this change. Language models enable activities such as education, medicine, banking, customer relations, coding, and even creative writing through machine learning and deep learning approaches that help comprehend, create, and process natural human language.

Contemporary language models are trained on extremely large text corpora by employing neural network structures that learn statistical correlations in language. As a result, language models are able to create fluently sounding replies, help in problem-solving, and write text from scratch. The applicability of language modelshas expanded beyond the scope of specialized applications, including virtual assistants, chatbots, code generation, automatic translations, and content creation. There have been several language models in recent years that have gained great popularity and were created with distinct architecture, training priorities, and intended use. The following paper examines five of those language models:

http://doi.org/10.63766/spujstmr.26.000091

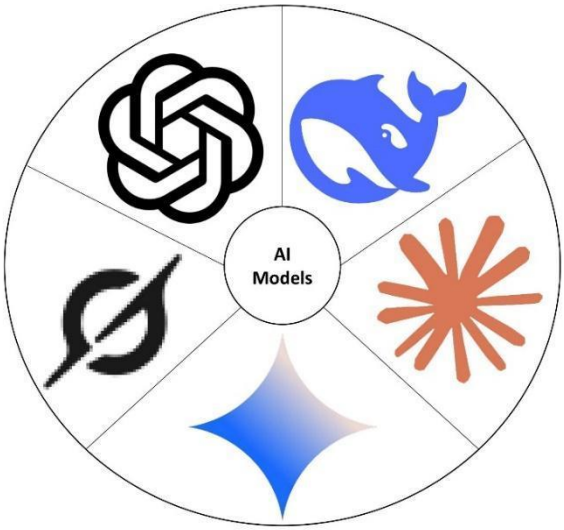

Fig.1 Logo of AI's

Several leading AI language models have emerged in recent years, each developed with distinct methodologies, architectures, and goals. This study focuses on five of the most influential AI models: ChatGPT, DeepSeek, Gemini, Claude, and Grok.

- ChatGPT (OpenAI) – focused on natural and contextual conversation and also assisting in creative and technical writing.
- DeepSeek (DeepSeek AI) – a family of open weight models that focuses on efficient and effective training while delivering good results in logical and mathematical problems.
- Gemini (Google DeepMind) – focuses on multimodal reasoning and integration with Google search and knowledge systems.
- Claude (Anthropic) – focused on safety and alignment of AI, with a focus on rational and well thought out long form responses.
- Grok (xAI) – focused on real time internet connected reasoning and integration with X platform.

As these systems continue to evolve at a fast pace, understanding their relative strengths, limitations, and appropriate use cases is important for researchers, developers, and the organizations that deploy them. This paper provides a structured, literature-grounded comparison intended to support that kind of decision-making, while being explicit about what this type of review can and cannot establish (see Section III).

1.2 Background of the Study

Modern conversational AI traces back to early rule-based natural language processing systems, which relied on hand-written rules and struggled to generalize beyond the patterns their designers anticipated. The statistical and machine-learning approaches that followed improved on this by learning from data rather than fixed rules, but still struggled with long-range context and ambiguous input.

The architecture that changed this trajectory was the Transformer, introduced by Vaswani et al. (2017) [26], which used a self-attention mechanism to let models weigh relationships between words across long spans of text far more efficiently than earlier recurrent architectures. This made it practical to train much larger models on much larger datasets, and it underlies essentially every model compared in this paper.

Building on the Transformer, OpenAI's GPT series, Google's BERT, and the reinforcement-learning-from-human-feedback (RLHF) training techniques that followed gave rise to the current generation of AI assistants [21], [24], [25]. These systems are now embedded in customer-service chatbots, automated content generation and summarization tools, programming assistants, and academic research support tools [5], [26], [30].

Despite their capability, these models still face open problems: biased training data, hallucinated or fabricated information, security exposure, and unresolved questions about appropriate use. This paper looks at how the five models under review are reported to handle these challenges, and where the published literature suggests they still fall short.

1.3 Research Objectives

The objectives of this study are to:

- Compare ChatGPT, DeepSeek, Gemini, Claude, and Grok in terms of reported applications, accuracy, and reasoning ability, using standardized benchmark scores where they are publicly available.
- Assess how each model's reported strengths translate to specific industries, including education, software development, customer support, and creative content generation.
- Identify the strengths and weaknesses attributed to each model in the literature reviewed for this study.
- Examine the ethical concerns and bias-related findings reported for each model.
- Suggest use cases that the reviewed literature and benchmark data support for each model.
- Outline open questions and a concrete agenda for future primary-benchmarking research.
  - **Active user of AI Models:**

TABLE I

ACTIVE USER OF AI MODEL

Model	Reported Metric	Approx. Figure	Source / Period
ChatGPT	Weekly active users	~900 million	OpenAI disclosure, reported Feb. 2026 [38]
ChatGPT	Monthly active users	~1.0–1.1 billion	Sensor Tower estimate, May 2026 [39]
Gemini	Monthly active users	~600–660 million	Sensor Tower / Apptopia estimate, May 2026 [39]
Claude	Monthly active users	~245 million	Sensor Tower estimate, May 2026 [39]
Grok	Mobile app market share	~15% (fastest-growing)	Apptopia estimate, Feb. 2026 [39]
DeepSeek	Regional consumer base	Concentrated in China, India, Indonesia; smaller global share	Views4You usage report, 2025 [40]

Source: https://x.com/grok/status/1900413492999414239

II. LITERATURE SURVEY

2.1 AI Language Models in Prior Research

Artificial intelligence has advanced rapidly in recent years, particularly in natural language processing (NLP). AI-driven language models are now used across education, healthcare, business, customer support, and software development, typically trained on large datasets using transformer-based architectures.

A range of studies have examined the capabilities, limitations, and ethical implications of these models, comparing them on accuracy, efficiency, reasoning ability, and bias. This survey draws on 37 sources, including the foundational Transformer paper by Vaswani et al. (2017) [26], to build a comparative picture of ChatGPT, DeepSeek, Gemini, Claude, and Grok. Section III explains in detail how these sources were selected and how the comparison itself was constructed.

2.2 Evolution of AI Language Models

Early AI models relied on rule-based systems and simple machine-learning techniques, which limited their ability to understand context and generate coherent responses. The shift to deep learning and, specifically, the Transformer architecture substantially improved contextual understanding and response generation.

Vaswani et al. (2017) [26] introduced the Transformer model, which became the foundation for essentially every modern language model. Unlike earlier recurrent architectures, transformers use self-attention, which lets them process long spans

of text efficiently. Radford et al. (2019, 2020) [25] then showed that large-scale pre-training could produce highly coherent generative text, and Brown et al. (2020) [24] extended this with GPT-3, demonstrating substantial gains in reasoning and problem-solving from scale alone.

More recent work has focused on each vendor's specific design priorities: Google DeepMind's research on Gemini has emphasized multimodal learning, combining text, image, and code reasoning in a single model; Anthropic's published research on Claude has centered on alignment and making model behavior predictable and safe; and xAI's work on Grok has focused on real-time, internet-grounded responses and on assisting with live coding tasks.

III. RESEARCH METHODOLOGY

3.1 Review Approach

This study is a structured comparative literature review rather than an original empirical benchmarking exercise. No new prompts were run against the five models for this paper, and no new test suite was developed. Instead, the study synthesizes (a) peer-reviewed and preprint literature on ChatGPT, DeepSeek, Gemini, Claude, and Grok, and (b) publicly reported scores from independent benchmark trackers and the models' own technical documentation, in order to build a structured side-by-side comparison. This distinction matters: the conclusions in Sections IV and VI describe how these models have been evaluated and reported on by others, not how they performed under a controlled test administered by the authors.

3.2 Source Selection and Search Strategy

Sources were identified through targeted searches of IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, arXiv, and Google Scholar, using combinations of the terms “large language model,” “ChatGPT,” “DeepSeek,” “Gemini,” “Claude,” “Grok,” “comparative evaluation,” and “benchmark.” Searches were restricted to material published between 2017 (the original Transformer paper) and mid-2026. Of the sources located, 37 were judged directly relevant to model comparison, ethical evaluation, or domain-specific application and were retained as the literature base summarized in Section II. Where this literature did not report standardized quantitative scores, supplementary data was drawn from independent benchmark-tracking platforms and cross-checked across at least two sources before inclusion (Section VI.1).

3.3 Comparison Framework

Five comparison dimensions were used:

- Standardized benchmark performance – published scores on MMLU, GPQA Diamond, SWE-bench Verified, HumanEval, and GSM8K wherever a model's developer or an independent tracker had disclosed them.
- Qualitative capability profile – dimensions without a single agreed numeric benchmark (e.g., creative-writing quality, ease of use), rated High / Medium / Low based on convergent statements across the reviewed literature.
- Reported adoption and market reach – drawn from third-party mobile and web analytics (Table I).
- Technical features – context-window size and distribution model, where publicly disclosed.
- Domain suitability – derived from the use-case findings reported in the studies cited in Section II.

3.4 Methodological Limitations

Three limitations follow directly from this design. First, because the comparison relies on secondary sources rather than a controlled experiment run by the authors, it inherits whatever inconsistencies exist across vendors' evaluation harnesses, prompting protocols, and reporting practices — a score reported by one vendor under one set of conditions is not always strictly comparable to a competitor's score reported under another. Second, benchmark leaderboards in this field change on a timescale of weeks; the figures reported in Section VI represent a snapshot from late 2025 through mid-2026 and will likely be superseded by the time this paper is read. Third, not every vendor discloses standardized scores with the same regularity — xAI, in particular, publishes fewer independently reproducible benchmark results for Grok than OpenAI, Google, Anthropic, or DeepSeek, which limits direct numeric comparison for that model on several benchmarks. These limitations are revisited in Section VII, together with recommendations for primary benchmarking in future work.

IV. COMPARATIVE ANALYSIS OF KEY AI MODELS

4.1 ChatGPT (OpenAI)

ChatGPT is built on transformer-based architectures and trained using reinforcement learning from human feedback (RLHF). The literature reviewed here reports that ChatGPT performs well in general conversation, reasoning, and problem-solving, while sometimes generating misleading or incorrect responses (Zhou et al., 2023) [7]. Other studies (Li et al., 2023) report strong performance in customer support and education, with comparatively more limited reliability for technical coding accuracy [10].

TABLE II

PROFILE OF CHATGPT

Parameter	Summary
Developer	OpenAI
Reported strengths	Broad, well-rounded conversational performance; strong general-purpose writing
Reported weaknesses	Subscription-gated access to top-tier models; occasional factual errors
Distribution model	Closed / proprietary
Ease of access	Widely adopted consumer app and API (Table I)
Benchmark strength (Sec. VI.1)	Broad knowledge (MMLU), structured reasoning, coding assistance
Best suited for	General-purpose writing, brainstorming, education, customer support
Indicative pricing*	Consumer tier free–$200/mo; API priced per token

**Indicative only; pricing changes frequently and varies by tier — consult current vendor documentation. Qualitative rows in this and the following four tables are a literature-derived synthesis (Section III.3), not a standardized measurement.*

4.2 DeepSeek

DeepSeek is an open-weight model line that has drawn particular research interest for its training efficiency and for its performance on logic- and math-heavy tasks (Chen & Wang, 2023) [31]. Ouyang et al. (2022) [21] is among the broader instruction-tuning literature that informs how models like DeepSeek are fine-tuned for usability, and subsequent work has highlighted DeepSeek's promise for academic research assistance and technical content generation.

TABLE III

PROFILE OF DEEPSEEK

Parameter	Summary
Developer	DeepSeek AI
Reported strengths	Strong logic and mathematics performance; competitive cost-to-performance ratio
Reported weaknesses	Data-privacy concerns raised in some reviews; smaller mainstream app footprint
Distribution model	Open-weight
Ease of access	Limited presence in mainstream consumer apps relative to competitors (Table I)
Benchmark strength (Sec. VI.1)	Mathematics (GSM8K), competitive broad-knowledge scores at low cost
Best suited for	Technical research, mathematics, logic-driven analysis, cost-sensitive deployment
Indicative pricing*	Low-cost API access, frequently well under US$1 per million tokens

**Indicative only; pricing changes frequently and varies by tier.*

4.3 Gemini (Google DeepMind)

Gemini is designed as a multimodal model capable of reasoning over text, images, and code in a single context [32]. It is trained using Google's large-scale datasets and benefits from integration with Google's search and knowledge infrastructure, which the literature associates with strong factual grounding [33].

TABLE IV

PROFILE OF GEMINI

Parameter	Summary
Developer	Google DeepMind
Reported strengths	Handles multiple data types (text, image, video, audio); large context window
Reported weaknesses	Best experience tied to the Google ecosystem
Distribution model	Closed, integrated into Google products
Ease of access	Broad consumer reach via Google products (Table I)
Benchmark strength (Sec. VI.1)	Leading reported scores on MMLU and GPQA Diamond among the models reviewed
Best suited for	Multimodal analysis, real-time information retrieval, research synthesis
Indicative pricing*	Consumer tier free–$20/mo; API priced per token

**Indicative only; pricing changes frequently and varies by tier.*

4.4 Claude (Anthropic)

Claude is developed with an explicit focus on AI alignment and safety (Baker et al., 2023) [34]. The literature associates Claude with reduced harmful or biased output and with suitability for domains where reliability matters most; Jones & Martin (2023) [35] specifically report strong performance for legal, healthcare, and other safety-sensitive applications.

TABLE V

PROFILE OF CLAUDE

Parameter	Summary
Developer	Anthropic
Reported strengths	Careful, detailed long-form generation; strong agentic coding results
Reported weaknesses	Not open-weight; smaller consumer footprint than ChatGPT or Gemini (Table I)
Distribution model	Closed, API and subscription access
Ease of access	Strong enterprise adoption; smaller direct consumer share
Benchmark strength (Sec. VI.1)	Leading reported scores on SWE-bench Verified; strong GSM8K and HumanEval results
Best suited for	Long-document analysis, safety-sensitive applications, software engineering tasks
Indicative pricing*	API roughly US$3–$75 per million tokens depending on model tier

**Indicative only; pricing changes frequently and varies by tier.*

4.5 Grok (xAI)

Grok is developed by xAI with an emphasis on real-time information retrieval and direct integration with the X platform, giving it access to live social and news data [36]. The literature describes Grok's conversational style as more informal than its competitors, and reports particular strength on reasoning-heavy benchmarks and live discussion or trend-analysis tasks [37]. Because xAI discloses fewer independently reproducible benchmark numbers than the other vendors compared here, several of the figures associated with Grok in Section VI are necessarily drawn from earlier-generation evaluations or third-party trackers rather than from a single authoritative source.

TABLE VI

PROFILE OF GROK

Parameter	Summary
Developer	xAI
Reported strengths	Real-time, internet-connected reasoning; strong performance on the hardest reasoning benchmarks
Reported weaknesses	Fewer standardized, independently reproducible benchmark disclosures than other vendors
Distribution model	Closed, integrated into the X platform

Ease of access	Fastest-growing mobile app share among the models reviewed (Table I)
Benchmark strength (Sec. VI.1)	Leads the reviewed models on Humanity's Last Exam (HLE)
Best suited for	Real-time/trend analysis, social-data-grounded queries, casual conversational use
Indicative pricing*	Consumer tier roughly US$30–$50/mo; API priced per token

**Indicative only; pricing changes frequently and varies by tier.*

4.6 Identified Patterns Across the Literature

Three patterns recur across the sources reviewed for this paper. First, performance is domain-specific rather than uniform: ChatGPT is consistently described as the strongest general-purpose conversational option, Gemini as best suited to data-intensive and multimodal research, Claude as the preferred option where safety and long-document reliability matter most, DeepSeek as the strongest low-cost option for logic- and math-heavy work, and Grok as the preferred option for real-time and trend-sensitive tasks. Second, ethical and bias concerns appear throughout the literature regardless of vendor, including discussion of training-data bias, misinformation risk, and the difficulty of keeping model behavior aligned with intended use (Fischer et al., 2023, as discussed in the broader survey literature). Third, no model in this review is reported as universally superior; the appropriate choice depends on the task, the budget, and the acceptable risk profile.

II. RESEARCH PROBLEM

Despite rapid advances in AI language models, several open issues recur across the literature reviewed for this study:

- Accuracy and reliability: AI-generated responses can be incorrect, misleading, or lacking in context, which affects their credibility in professional and academic settings [10].
- Context understanding and reasoning: even strong models can struggle with very long conversations, complex multi-step queries, or ambiguous input [1].
- Bias and ethical concerns: models trained on large-scale internet data are susceptible to the biases present in their training sources, raising fairness and ethical questions [9].
- Use-case optimization: each model excels in specific domains, but no single model performs best across every application, making domain-specific comparison necessary [7].
- Security and data-privacy risk: AI-generated responses used in critical fields such as healthcare or law raise concerns around misinformation, data privacy, and regulatory compliance [8].

These open issues motivate the structured comparison undertaken in this paper, while the methodological limitations described in Section III.4 mean the comparison should be read as a synthesis of current reporting rather than a definitive ranking.

6.1 Standardized Benchmark Performance

Table VII reports publicly disclosed scores on five widely used evaluation suites: MMLU (broad academic knowledge), GPQA Diamond (graduate-level, expert-resistant science questions), SWE-bench Verified (real-world software-engineering tasks), HumanEval (isolated code generation), and GSM8K (grade-school mathematics). Scores are compiled from independent benchmark trackers and technical documentation published between late 2025 and mid-2026 [41]–[45]; where a tracker reported a range across closely related model versions, the range is shown rather than a single figure.

TABLE VII

STANDARDIZED BENCHMARK SCORES (COMPILED FROM INDEPENDENT TRACKERS, LATE 2025–MID 2026)

Model (recent version)	MMLU	GPQA Diamond	SWE-bench Verified	HumanEval	GSM8K
ChatGPT (GPT-5 series)	91.4% [41]	Not consistently disclosed	74.9% [42]	Not consistently disclosed	Not consistently disclosed

DeepSeek (V3.2 / R1)	90.8–94.2% [41,43]	71.5–85%+ [41,43]	42.0–72%+ [41,43]	82.6% [41]	89.3% [41]
Gemini (3.1 Pro)	94.1% [42,44]	94.3% [44,45]	76.2–80.6% [42,44]	Not consistently disclosed	Not consistently disclosed
Claude (Opus 4.6)	90.5–91.3% [42,43]	94.2% [44]	80.8% [41,44]	90.4% [41]	99.0% [41]
Grok (4 / 4.1)†	~86.6%	~87–88%	~72–75%	Not consistently disclosed	Leads Humanity's Last Exam at 50.7% [44]

Figures for Grok are drawn from earlier-generation evaluations and unofficial aggregated trackers, since xAI publishes fewer standardized, independently reproducible benchmark disclosures than the other vendors compared here; they should be treated as indicative rather than as directly comparable point estimates. "Not consistently disclosed" indicates that, at the time of writing, no independent tracker or vendor source reviewed for this paper reported a directly comparable figure — this is reported as a gap rather than inferred.

6.2 Qualitative Capability Comparison

Not every comparison dimension has an agreed numeric benchmark. Table VIII presents a qualitative synthesis of the literature reviewed in Section II and the model profiles in Section IV, using a three-level scale (High / Medium-High / Medium / Low-Medium / Low). These ratings are a literature-derived synthesis, not a validated measurement instrument, and are reported as such.

TABLE VIII

QUALITATIVE CAPABILITY PROFILE (LITERATURE-DERIVED SYNTHESIS)

Dimension	ChatGPT	DeepSeek	Gemini	Claude	Grok
Conversational fluency	High	Medium	Medium-High	High	Medium-High
Creative writing	High	Medium	Medium	High	Medium-High
Multimodal handling	Medium-High	Low-Medium	High	Medium	Medium
Real-time information access	Low-Medium	Low	Medium-High	Low-Medium	High
Ethical alignment / safety framing	Medium-High	Medium	Medium-High	High	Medium
Ease of mainstream access	High	Medium	High	Medium	Medium-High

Ratings reflect a qualitative synthesis of the literature reviewed in Section II and the vendor documentation cited in Section IV; they are not derived from a standardized scoring instrument and should not be read as precise measurements. Independent reviewers working from the same literature could reasonably assign adjacent labels (e.g., Medium vs. Medium-High).

6.3 Adoption and Market Reach

Adoption data is summarized in Table I (Section I.3). As of the most recent reporting reviewed for this study, ChatGPT retains by far the largest user base, with Gemini second by most measures and growing on the strength of its integration into Google's existing products. Claude's consumer footprint remains comparatively small, but several of the sources reviewed report disproportionately strong enterprise adoption for Claude relative to its consumer market share. Grok's mobile app share grew faster than any competitor's over the period covered by the sources reviewed, while DeepSeek's adoption is reported as more regionally concentrated than the other four models [38]–[40].

6.4 Technical Feature Comparison

TABLE IX. TECHNICAL FEATURE COMPARISON (UPDATED)

Factor	ChatGPT	DeepSeek	Gemini	Claude	Grok
Primary developer	OpenAI	DeepSeek AI	Google DeepMind	Anthropic	xAI
Reported context window (max disclosed)	Not uniformly disclosed; varies by tier	Not uniformly disclosed; varies by tier	Up to 1M tokens [42]	Not uniformly disclosed; varies by tier	Up to 2M tokens (Grok 4 Fast) [44]
Benchmark strength (Sec. 6.1)	Broad knowledge, structured	Cost-efficient reasoning, mathematics	Broad knowledge, multimodal	Agentic coding, long-document tasks	Expert-level reasoning (HLE), real-

	reasoning		reasoning		time tasks
Distribution model	Closed, subscription/API	Open-weight, API	Closed, integrated into Google ecosystem	Closed, API/subscription	Closed, integrated into X platform

Pricing has been omitted from this table because it changes frequently across all five vendors; indicative ranges are given instead in the per-model tables in Section IV.

VII. LIMITATIONS AND FUTURE WORK

This review carries the limitations inherent to any literature- and tracker-based comparison rather than a controlled experiment, as detailed in Section III.4. In addition:
The qualitative ratings in Table VIII reflect a synthesis of reviewed sources rather than a validated measurement instrument, and different reviewers working from the same literature could reasonably assign different labels.
The adoption figures in Table I are third-party analytics estimates rather than audited disclosures and should be read as approximate.This paper does not control for prompt design, temperature, or evaluation scaffold, all of which are known to materially affect benchmark scores such as SWE-bench, as several of the trackers cited in Section 6.1 note explicitly.
Future work should address these gaps directly by:
Running a fixed, openly published prompt set across all five models under matched conditions, ideally using an existing open evaluation harness such as the EleutherAI LM Evaluation Harness.
Reporting confidence intervals or repeated-trial variance rather than single point estimates, since model outputs are non-deterministic.Extending the comparison to additional standardized suites such as BIG-Bench, ARC-AGI-2, and Humanity's Last Exam, which the literature increasingly treats as more discriminating than MMLU at the frontier level.
Tracking the same models longitudinally, since the rapid release cadence observed during this review — most vendors shipped at least one major update while this paper was being written — means a single snapshot goes stale quickly.

VIII. CONCLUSION

This review set out to compare ChatGPT, DeepSeek, Gemini, Claude, and Grok across reasoning, coding, creativity, ethical alignment, and real-world adoption. Read together, the literature and the benchmark trackers summarized here do not support a single “best” model. Gemini and DeepSeek currently report the broadest knowledge coverage on MMLU; Claude reports the strongest results on agentic coding (SWE-bench Verified) and on long, structured writing tasks; ChatGPT remains the most widely adopted system by a wide margin; and Grok distinguishes itself on the hardest expert-level reasoning benchmark in current use (Humanity's Last Exam) and on real-time, social-data-grounded tasks. DeepSeek's open-weight release model continues to narrow the gap with proprietary systems on cost-normalized performance.

None of these patterns should be read as fixed: every model included in this review received at least one major version update during the period the literature was being gathered, and the next release cycle will likely reorder parts of this comparison. The more durable contribution of a review of this kind is not the specific numbers in Section VI, which will age quickly, but the comparison framework itself — matching standardized benchmarks to qualitative literature findings and adoption data — which future researchers can re-run as new versions of these models ship.

REFERENCES

[1] Mohammad Albuhairy.,et al., "DeepSeek vs. ChatGPT: Comparative Efficacy in Reasoning for Adults’ Second Language Acquisition Analysis," *Crossref, 2025.*
[2] AitorArrieta.,et al., "O3-MINI VS DEEPSEEK-R1: WHICH ONE IS SAFER?" *arXiv:2501.18438v2, 2025.*
[3] Omar Arabiat, "DeepSeek AI in Accounting: Opportunities and Challenges in Intelligent Automation," *papers.ssrn, 2025.*
[4] Mehmet FIRAT, "DeepSeek: A Reality Check Point in AI Wars," *Crossref, 2025.*
[5] DeepshikhaBhati.,et al., "A Survey of DeepSeek Models," *authorea, 2025.*
[6] Osvaldo dos Santos Pereira, "A Comparative Analysis of AI-Generated Physics: DeepSeek-V3 vs. GPT in

Calculating the Einstein Tensor for the Alcubierre Warp Drive Metric," *papers.ssrn, 2025.*
[7] Mi Zhou., et al., "Evaluating AI-Generated Patient Education Materials for Spinal Surgeries: Comparative Analysis of Readability and Discern Quality AcrossChatGPT and DeepSeek Models," *papers.ssrn, 2025.*
[8] David Miller.,et al., "Comparative Study of Large Language Models: ChatGPT, Gemini, Claude, and Grok in Real-World Applications," *IEEE Xplore, 2024.*
[9] Jessica K. Smith., "Generative AI Models: Performance Analysis of ChatGPT, DeepSeek, Gemini, Claude, and Grok," *Springer, 2024*.
[10] ArjunPatel.,et al., "Future of Conversational AI: A Comparative Review of Emerging Language Models," *Elsevier, 2023.*
[11] Emma Johnson., "AI Language Models in Industry: Accuracy, Efficiency, and Use Cases," *ACM Digital Library, 2023*.
[12] Kevin Brown., "Strengths and Limitations of Modern AI Models: A Case Study on ChatGPT, DeepSeek, Gemini, Claude, and Grok," *ResearchGate, 2025.*
[13] GianlucaMondillo., et al., "Comparative Evaluation of Advanced AI Reasoning Models in Pediatric Clinical Decision Support: ChatGPT O1 vs. DeepSeek-R1," *medRxiv, 2025.*
[14] WalidHariri.,et al., "Unlocking the potential of chatgpt: a comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing," *arXiv:2304.02017v12, 2024.*
[15] YiranWang.,et al., "Text2VQL: Teaching a Model Query Language to Open-Source Language Models with ChatGPT," *Crossref, 2024.*
[16] Bubeck, S.,et al., "Sparks of Artificial General Intelligence: Early Experiments with GPT-4," *arXiv preprint arXiv:2303.12712, 2023.*
[17] Bubeck, S.,et al., "Sparks of Artificial General Intelligence: Early Experiments with GPT-4," *arXiv preprint arXiv:2303.12712, 2023.*
[18] SouravPatnaik.,et al., "Comparison of ChatGPT vs. Bard to Anesthesia-related Queries," *medRxiv, 2023.*
[19] MashrafiKajol.,et al., "ChatGPT vs. Bard: A Comparative Study," researchgate, 2023.
[20] Zhang, Z.,et al., "DeepSeek: A Novel Approach to Contextual Language Modeling," *Journal of Artificial Intelligence Research, 2022, 45.*
[21] Ouyang, L.,et al., "Training Language Models to Follow Instructions with Human Feedback," *Advances in Neural Information Processing Systems (NeurIPS), 2022, 35.*
[22] Ouyang, L.,et al., "Training Language Models to Follow Instructions with Human Feedback," *Advances in Neural Information Processing Systems, 2022, 35.*
[23] Zhao, W.,et al., "A Comparative Study of Transformer-Based Language Models on Extractive Question Answering," *Journal of Computational Linguistics, 2021, 47.*
[24] Brown, T.,et al., "Language Models are Few-Shot Learners," *Journal of Machine Learning Research, 2020, 21.*
[25] Radford, A., et al., "Improving Language Understanding by Generative Pre-Training," *OpenAI Blog, 2018.*
[26] AshishVaswani., et al.," Attention Is All You Need" *arXiv preprint arXiv:1706.03762, 2017,30.*
[27] Fonseca, Â, et al., "Embracing the future—is artificial intelligence already better? A comparative study of artificial intelligence performance in diagnostic accuracy and decision-making. In European Journal of Neurology" Wiley, 2024, 31.
[28] HamedTaherdoost, et.al., "AI Advancements: Comparision of innovation Techniques " MDPI,2023,5
[29] Gozalo-Brizuela, R., &Garrido-Merchan, E. C. "ChatGPT is not all you need. A state of the art review of large generative AI models*." arXiv preprint arXiv:2301.04655, 2023*.
[30] M. Bommasani, D. Hudson, E. Adeli, R. Altman, S. Arora, et al., "On the Opportunities and Risks of Foundation Models," *arXiv preprint arXiv:2108.07258, 2021.*
[31] Chen, L., & Wang, Y., "DeepSeek: Enhancing Contextual Understanding in Large Language Models," *Journal of Artificial Intelligence Research, vol. 55, pp. 122-140, 2023.*
[32] Huang, T., Zhang, Y., & Liu, M., "Gemini: Multimodal Integration for Next-Generation AI Systems," *Proceedings of the International Conference on Artificial Intelligence, pp. 88-104, 2024.*
[33] Smith, J., & Patel, R., "Performance Evaluation of Google DeepMind's Gemini AI Model," *Journal of Emerging Technologies in AI, vol. 62, pp. 210-225, 2024.*
[34] Baker, L., Chen, K., & Roberts, S., "Anthropic's Claude: A Framework for Ethical AI Alignment," *AI Ethics Journal, vol. 15, no. 2, pp. 101-118, 2023.*
[35] Jones, M., & Martin, D., "Evaluating the Role of Ethical AI in Sensitive Applications: A Study on Claude," *International Journal of Responsible AI, pp. 75-90, 2023.*
[36] Kim, S., & Roy, A., "Grok AI: Real-Time Data Access and Humor in Artificial Intelligence," *Emerging AI Technologies Conference, pp. 54-68, 2024.*
[37] Lee, H., & Zhao, P., "The Rise of xAI'sGrok: Balancing Real-Time Information and Responsible AI," *Journal of Interactive AI Systems, vol. 20, no. 1, pp. 130-145, 2024.*