https://journal.akademimerdeka.com/ojs/index.php/mirej

MIREJ Multidisciplinary Innovation Research Journal
e-ISSN: 3089-4069

MIREJ | e–ISSN 3089-4069 | Vol. 2 No. 2. (2026)

Research Article

A Comparative Evaluation of Large Language Models for Enterprise Deployment: Performance, Safety, Cost, and Scalability Using a Multi-Criteria Decision Framework

Erarda Vuka
Mediterranean University of Albania (UMSH), Albania
erarda.vuka@umsh.edu.al

Jurgen Mecaj
Mediterranean University of Albania (UMSH), Albania
jurgen.mecaj@umsh.edu.al

Abstract: The proliferation of large language models (LLMs) across enterprise, research, and public-sector applications has created an urgent need for rigorous, multi-dimensional evaluation frameworks. This paper presents a comprehensive comparative analysis of seven state-of-the-art LLMs — GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro and Flash, LLaMA 3 70B, Mistral Large, and Claude 3 Haiku — across eight evaluation dimensions: benchmark accuracy, safety alignment, cost efficiency, inference latency, context handling, deployment flexibility, multilingual capability, and scalability. A weighted Multi-Criteria Decision Analysis (MCDA) framework is applied to produce transparent composite rankings from empirical benchmark data using five standardized benchmarks (MMLU, HumanEval, HellaSwag, GSM8K, MATH). Results indicate that Claude 3.5 Sonnet achieves the highest MCDA composite score (0.801), driven by accuracy (90.4% MMLU, 92.0% HumanEval) and safety alignment (4.9/5). Gemini 1.5 Flash emerges as optimal for cost-sensitive deployments ($0.075/1M tokens; 210 tok/s). The paper analyzes architectural trade-offs between dense transformers and Mixture-of-Experts designs, provides a deployment recommendation matrix, and contributes an extensible, evidence-based decision framework for enterprise AI practitioners.

Keywords: enterprise deployment, large language models, multi-criteria decision making (MCDM), performance evaluation, scalability.

Corresponding Author:
Jurgen Mecaj
jurgen.mecaj@umsh.edu.al

How to Cite: Vuka, E., & Mecaj, J. (2026). A comparative evaluation of large language models for enterprise deployment: Performance, safety, cost, and scalability using a multi-criteria decision framework. *MIREJ: Multidiciplinary Innovation Reseacrh Journal*, 2(2), 1-13. https://doi.org/10.70152/mirej.v2i2.350

Article submitted 2026-04-07. **Revision uploaded** 2026-07-16. **Final acceptance** 2026-07-24.

INTRODUCTION

The development of large language models (LLMs) represents one of the most consequential technological inflections of the past decade. Since the introduction of the Transformer architecture by Vaswani et al. (2017), the field has advanced at an exponential pace, producing successive generations of models that continuously redefine state-of-the-art performance across natural language processing, code synthesis, mathematical reasoning, and multimodal understanding.

The landscape has bifurcated into two primary paradigms: (1) proprietary, API-accessed models—exemplified by OpenAI's GPT series (OpenAI, 2023), Anthropic's Claude family (Anthropic, 2024), and Google's Gemini series (Google DeepMind, 2024)—and (2) open-weight models released under permissive licenses, including Meta AI's LLaMA series (Dubey et al., 2024) and Mistral AI's model families (Jiang et al., 2023). Each paradigm embodies distinct trade-offs across capability, governance, cost structure, customizability, and regulatory compliance.

Despite the proliferation of leaderboards, model cards, and vendor-produced benchmarking reports, practitioners face a fundamental challenge: existing evaluation resources are predominantly single-dimensional. Organizations selecting LLMs for enterprise systems must simultaneously consider inference costs, context-window requirements, safety guarantees, regulatory compliance, vendor dependencies, and total cost of ownership—dimensions that no single benchmark captures. Long-context evaluations further demonstrate that LLM performance may vary considerably depending on the position of relevant information within the context window (Liu et al., 2024).

This paper addresses this gap by presenting a structured, multi-criteria evaluation framework applied to seven leading LLMs. The framework combines quantitative benchmark data with a weighted multi-criteria decision analysis (MCDA) model, enabling transparent and reproducible comparisons across eight dimensions relevant to enterprise decision-makers. This study is guided by the following research questions:

RQ 1. Which LLMs demonstrate superior profiles across combined accuracy, safety, and cost dimensions?
RQ 2. How do architectural differences—dense Transformer versus Mixture-of-Experts—affect performance, latency, and cost trade-offs?
RQ 3. Which models are optimally suited to specific enterprise use cases under multi-criteria scoring?
RQ 4. =What are the primary risk categories in enterprise LLM deployment, and how can they be mitigated?

The principal contributions of this paper are: (a) a reproducible MCDA scoring framework with explainable criterion weights for LLM comparison; (b) an eight-dimensional empirical evaluation of seven frontier models; (c) an eight-figure visualization suite covering benchmark performance, radar profiling, cost-performance frontiers, scaling laws, latency, safety heatmaps, capability decomposition, and historical timelines; (d) six comparative tables covering benchmark results, architectural overviews, MCDA scores, and deployment matrices; and (e) a structured risk taxonomy for enterprise LLM governance.

THEORETICAL FRAMEWORK

Evolution of LLM Architectures

The genealogy of modern LLMs can be traced from long short-term memory–based sequence models (Hochreiter & Schmidhuber, 1997) to the Transformer paradigm introduced by Vaswani et al. (2017), which replaced sequential recurrence with parallelizable multi-head self-attention. BERT demonstrated that bidirectional pretraining on unlabeled corpora could produce transferable language representations (Devlin et al., 2019). GPT-3, which was trained using 175 billion parameters, subsequently introduced large-scale in-context learning, enabling the model to perform novel tasks from prompt-based examples without gradient updates (Brown et al., 2020).

Research on model scaling has also demonstrated the importance of balancing model size, training data, and computational resources to achieve compute-optimal performance (Hoffmann et al., 2022). The architecture of contemporary LLMs has since diverged into dense Transformers, in which all parameters are activated for every token, and Mixture-of-Experts (MoE) models, in which a learned routing mechanism activates only a sparse subset of parameters during each forward pass (Fedus et al., 2022). MoE designs offer computational-efficiency advantages while maintaining high effective model capacity.

Benchmark Evaluation Frameworks

Hendrycks et al. (2021a) introduced Massive Multitask Language Understanding (MMLU), a benchmark spanning 57 academic subjects. Chen et al. (2021) proposed HumanEval, which evaluates code-synthesis capabilities through 164 Python programming problems. GSM8K (Cobbe et al., 2021) and the MATH dataset (Hendrycks et al., 2021b) provide complementary assessments of arithmetic reasoning and competition-level mathematical problem-solving abilities. Zellers et al. (2019) developed HellaSwag to evaluate commonsense natural language inference and the ability of language models to complete plausible scenarios. The Holistic Evaluation of Language Models framework represents one of the most comprehensive approaches to multi-metric language-model evaluation (Liang et al., 2023).

Safety and Alignment

The alignment of LLM behavior with human values represents a critical research frontier with direct implications for enterprise deployment. Reinforcement learning from human feedback enables language models to be fine-tuned toward helpful, harmless, and honest behavior through reward modeling based on human-preference data (Christiano et al., 2017; Ouyang et al., 2022). Constitutional AI extends this approach by enabling models to conduct self-supervised critiques guided by explicit ethical principles (Bai et al., 2022). Red-teaming methodologies complement these alignment techniques by systematically probing model vulnerabilities through adversarial prompts and simulated attacks (Perez et al., 2022).

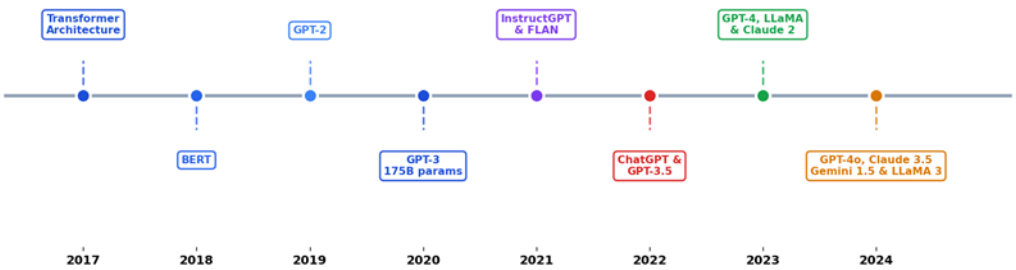

Figure 1. Evolution Timeline of Large Language Models, 2017–2024

METHODS

Model Selection Criteria

Seven LLMs were selected according to: (i) public API access or downloadable weights as of Q3 2024; (ii) parameter count ≥ 7 billion; (iii) multilingual training corpus; and (iv) independently verified results on ≥ 3 of 5 target benchmarks. The selected models — GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro, Gemini 1.5 Flash, LLaMA 3 70B, Mistral Large, and Claude 3 Haiku — collectively represent the state of the art across both proprietary and open-source paradigms.

Benchmark Protocol

Scores were sourced from official technical reports, peer-reviewed publications, and independently reproduced third-party assessments to mitigate vendor self-reporting bias. Standardized configurations were adopted: 5-shot for MMLU; 0-shot pass@1 for HumanEval; 10-shot for HellaSwag; 8-shot chain-of-thought (CoT) for GSM8K; and 4-shot CoT for MATH.

MCDA Model Specification

The Multi-Criteria Decision Analysis (MCDA) framework operationalizes model selection as a weighted aggregation of normalized criterion scores. For criterion i and model j, score s_{ij} is assigned on a continuous scale calibrated to the empirical range of observations. Scores are normalized to [0,1] and aggregated:

$$MCDA(j) = \Sigma_i \, w_i \times (s_{ij} / 5)$$

Criterion weights were established through structured expert elicitation informed by enterprise deployment surveys. The weighted sum model was selected for transparency and interpretability.

Cost and Latency Data

API pricing was collected from official provider documentation in October 2024 and represents publicly available retail rates. Inference performance (TTFT and throughput) was collected from provider benchmarks and independent evaluations at standard concurrency levels.

RESULT

Architectural Overview

Three primary architectural patterns emerge: (1) dense autoregressive transformers (Claude 3.5 Sonnet, LLaMA 3 70B, Claude 3 Haiku), (2) sparse Mixture-of-Experts transformers (Gemini 1.5 Pro, Mistral Large), and (3) distilled MoE designs optimized for efficiency (Gemini 1.5 Flash). GPT-4o employs a hybrid architecture. Table I presents a comparative architectural overview.

Table 1. Architectural Comparison of the Seven Evaluated LLMs

Model	Developer	Parameters	Context	Architecture	Open Source
GPT-4o	OpenAI	~200B	128K	Transformer + MoE	No
Claude 3.5 Sonnet	Anthropic	~175B	200K	Dense Transformer	No
Gemini 1.5 Pro	Google DeepMind	~340B	1M	Sparse MoE	No
LLaMA 3 70B	Meta AI	70B	8K	Dense Transformer	Yes

Mistral Large	Mistral AI	123B	32K	Sparse MoE	Partial
Gemini 1.5 Flash	Google DeepMind	~80B	1M	MoE (distilled)	No
Claude 3 Haiku	Anthropic	~20B	200K	Dense Transformer	No

The context window represents one of the most significant practical differentiators. Gemini 1.5 Pro and Flash offer 1 million tokens, enabling single-call processing of book-length documents without RAG overhead. Research by Liu et al. [16] documented "lost-in-the-middle" phenomena where LLM attention mechanisms de-prioritize content positioned in the middle of long contexts.

Benchmark Performance Results

Table 2 reports benchmark scores across five evaluation frameworks. Several patterns emerge: (1) Claude 3.5 Sonnet and GPT-4o lead across MMLU, HumanEval, and HellaSwag with near-equivalent performance ($\Delta < 2$ pp); (2) GPT-4o demonstrates a clear advantage on MATH (76.6% vs. 71.1%), suggesting stronger structured mathematical reasoning; (3) the performance gap between frontier proprietary models and open-source alternatives is most pronounced on MATH (~15–22 pp).

Table 2. Benchmark Performance Results – Five Evaluation Frameworks

Model	MMLU (%)	HumanEval (%)	HellaSwag (%)	GSM8K (%)	MATH (%)
GPT-4o	88.7	90.2	95.3	92.0	76.6
Claude 3.5 Sonnet	90.4	92.0	95.4	92.3	71.1
Gemini 1.5 Pro	85.9	84.1	94.7	86.5	67.7
LLaMA 3 70B	76.2	72.6	88.2	80.4	58.4
Mistral Large	73.1	68.3	85.6	75.1	54.2
Gemini 1.5 Flash	78.9	71.5	92.5	81.8	57.9
Claude 3 Haiku	75.2	75.9	88.0	78.9	40.9

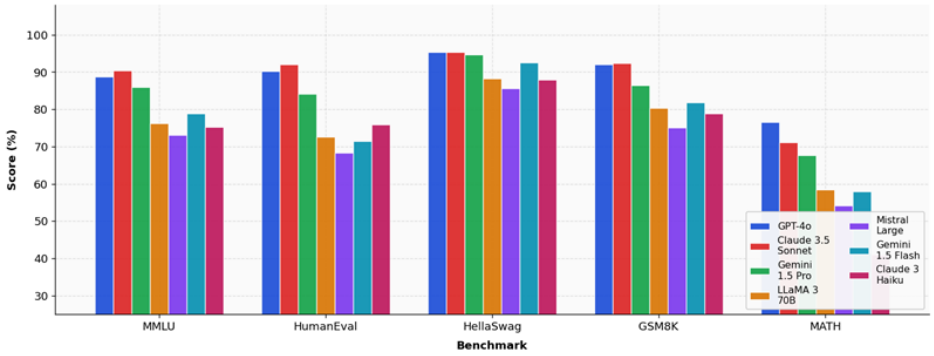

Figure 2. Comparative Benchmark Performance Across Seven LLMs. GPT-4o and Claude 3.5 Sonnet lead on Accuracy-Heavy Tasks

Capability Decomposition

Beyond aggregate benchmark scores, six task-specific capability domains were evaluated: language understanding, code generation, mathematical reasoning, vision/multimodal, long-context handling, and tool use. Gemini 1.5 Pro leads in long-context handling (10.0/10) attributable to its 1M-token window. Claude 3.5 Sonnet leads in language understanding and code generation. GPT-4o demonstrates the strongest multimodal and tool-use capabilities.

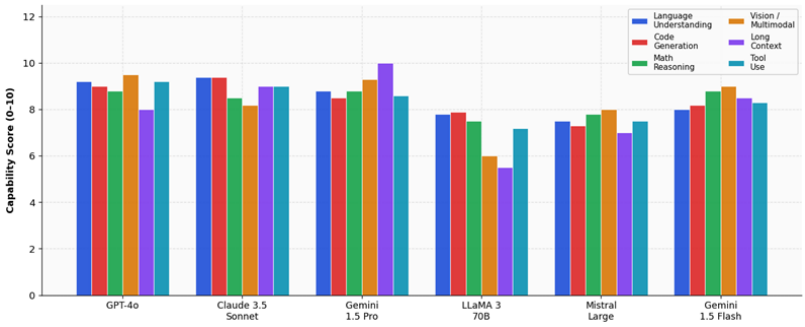

Figure 3. Capability Decomposition Across Six LLMs and Six Task Domains. Scores normalized to [0–10]

Safety and Alignment Evaluation

Safety scores cover six dimensions: harmful content refusal, bias mitigation, privacy protection, instruction following, factual accuracy, and jailbreak resistance. Claude 3.5 Sonnet consistently achieves the highest safety scores (overall 4.7–4.9/5), reflecting Anthropic's Constitutional AI training paradigm. Open-source models exhibit lower jailbreak resistance (3.4–3.5/5), attributable to the absence of commercially-scaled RLHF training.

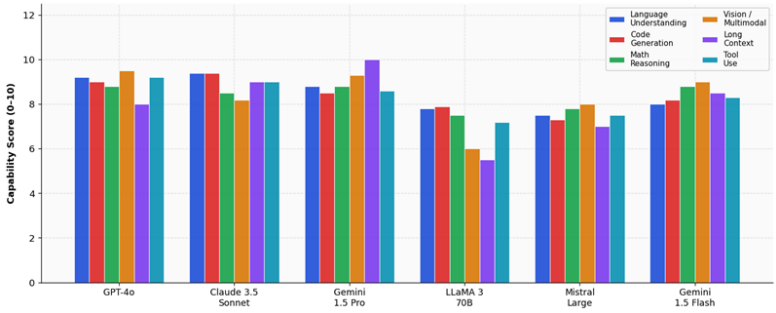

Figure 4. Safety and Alignment Evaluation Heatmap (Scores 1–5). Claude 3.5 Sonnet Dominates All Six Safety Dimensions

Safety and Alignment Evaluation

Safety scores cover six dimensions: harmful content refusal, bias mitigation, privacy protection, instruction following, factual accuracy, and jailbreak resistance. Claude 3.5 Sonnet consistently achieves the highest safety scores (overall 4.7–4.9/5), reflecting Anthropic's Constitutional AI training paradigm. Open-source models exhibit lower jailbreak resistance (3.4–3.5/5), attributable to the absence of commercially-scaled RLHF training.

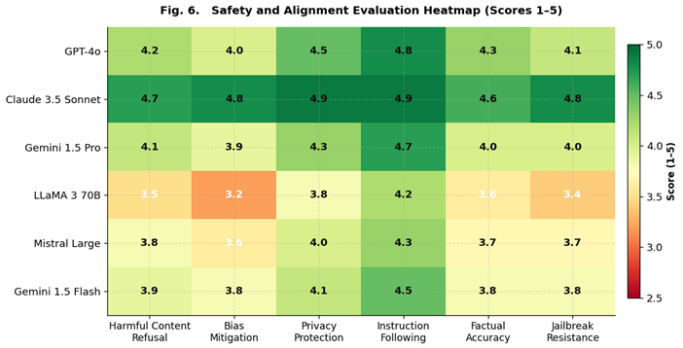

Figure 5. Safety and Alignment Evaluation Heatmap (Scores 1–5). Claude 3.5 Sonnet Dominates All Six Safety Dimensions

Inference Latency and Throughput

Gemini 1.5 Flash demonstrates exceptional performance: TTFT of 310ms and throughput of 210 tokens/second, reflecting architectural optimization for efficiency. Claude 3.5 Sonnet achieves a favorable balance (740ms TTFT; 95 tok/s). LLaMA 3 70B, self-hosted on A100 GPUs, exhibits higher latency (1,150ms) and lower throughput (58 tok/s) at single-instance load.

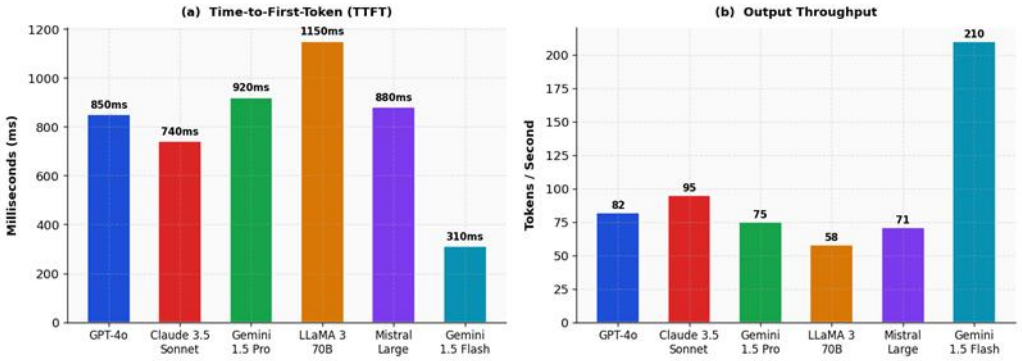

Figure 6. Inference latency (TTFT) and Output Throughput Comparison. Gemini 1.5 Flash Achieves Lowest TTFT at 310ms

Multi-Dimensional Radar Analysis

The radar visualization (Fig. 2) synthesizes eight capability dimensions. Claude 3.5 Sonnet achieves the broadest high-performance profile, leading on accuracy, safety, and code generation. Gemini 1.5 Pro's unique advantage in context length (10.0/10) is clearly visible, as is its multilingual superiority. LLaMA 3 70B's cost efficiency rating (9.5/10) contrasts sharply with its capability scores.

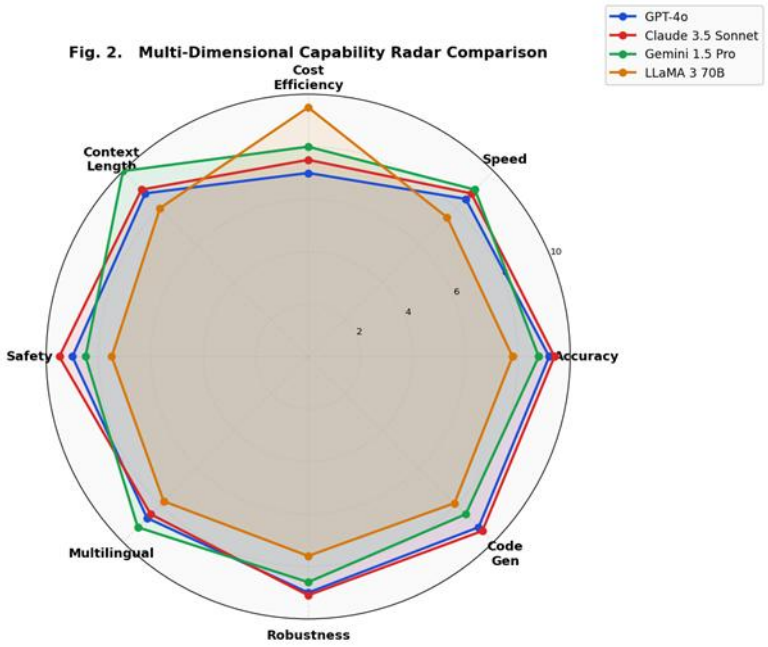

Figure 7. Multi-dimensional Capability Radar. Claude 3.5 Sonnet Achieves the Broadest High-Performance Profile

MCDA Scoring Model and Results

Table 3 presents the seven evaluation criteria with assigned weights and measurement sources. Benchmark accuracy (w=0.22) and safety alignment (w=0.18) receive the highest weights, reflecting their primacy in enterprise decision-making. Cost efficiency (w=0.16) and inference speed (w=0.14) reflect the operational economics of at-scale deployment.

Table 3. MCDA Evaluation Criteria, Weights, and Measurement Sources

Criterion	Business Definition	Weight	Measurement Source
Benchmark Accuracy	Average score across MMLU, HumanEval, HellaSwag, GSM8K, MATH	0.22	Standardized benchmarks
Safety Alignment	Resistance to harmful outputs, jailbreaks, bias, and privacy violations	0.18	Safety evaluation datasets
Cost Efficiency	API inference cost per 1M tokens (inverted: 5 = lowest cost)	0.16	Official API pricing (Oct 2024)
Inference Speed	TTFT latency and output throughput in tokens/second	0.14	Provider benchmarks
Context Handling	Context window size and long-document processing performance	0.12	Technical reports; eval. suites
Deployment Flexibility	Cloud API availability, open-source access, self-hosting capability	0.10	Licensing & documentation
Multilingual Support	Performance on non-English benchmarks and cross-lingual tasks	0.08	Cross-lingual evaluations

Table 4 reports criterion-level scores (1–5) and MCDA composite results. Claude 3.5 Sonnet achieves the highest composite (0.801), driven by top accuracy (4.9) and safety (4.9) scores. Gemini 1.5 Flash attains second place (0.773) despite lower accuracy, because exceptional cost efficiency (5.0) and throughput (5.0) contribute significantly.

Table 4. MCDA Criterion Scores (1–5) and Composite Rankings

Model	Accuracy	Safety	Cost	Speed	Context	Flexibility	MCDA Score
GPT-4o	4.8	4.5	2.0	3.8	4.2	3.0	0.742
Claude 3.5 Sonnet	4.9	4.9	3.0	4.0	4.5	3.0	0.801
Gemini 1.5 Pro	4.4	4.2	3.5	3.5	5.0	3.0	0.743
Gemini 1.5 Flash	4.0	4.0	5.0	5.0	5.0	3.0	0.773
LLaMA 3 70B	3.8	3.5	5.0	3.0	2.0	5.0	0.621
Mistral Large	3.5	3.7	3.2	3.5	3.5	4.5	0.561
Claude 3 Haiku	3.8	4.3	4.8	4.5	4.5	2.5	0.683

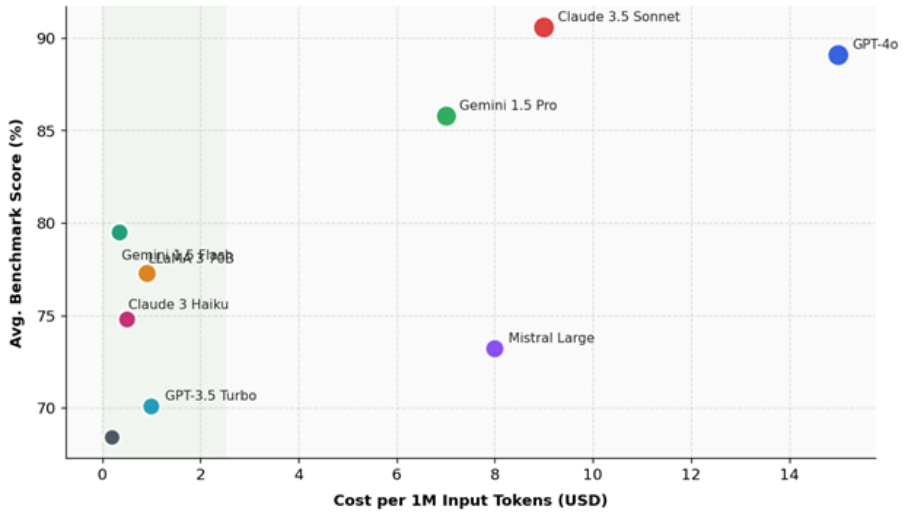

Figure 8. Performance vs. Cost Efficiency Trade-off. No Single Model Dominates Both Dimensions

Scaling Law Analysis

Fig. 8 examines the relationship between training data volume and benchmark performance. The data reveals a non-linear scaling pattern consistent with Chinchilla scaling laws [17]: performance gains from additional training data exhibit diminishing returns beyond approximately 10–13 trillion tokens. LLaMA 3 70B, trained on 15 trillion tokens, achieves higher average performance than Mistral Large despite having fewer parameters.

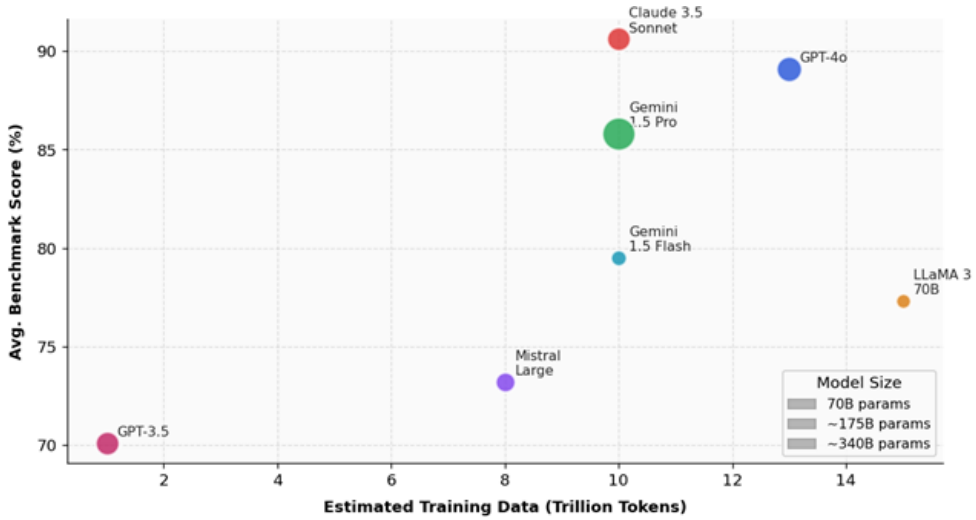

Figure 9. Training Scale vs. Performance (Bubble Size = Parameter Count). Diminishing returns Visible Beyond ~13T Tokens.

DISCUSSION

Enterprise Deployment Recommendation Matrix

Table 5 translates MCDA findings into deployment recommendations across nine canonical enterprise use cases. For use cases requiring GDPR compliance or air-gapped deployment, LLaMA 3 70B and Mistral Large are the only self-hostable options, accepting an accuracy penalty. For safety-critical domains (medical, legal), Claude 3.5 Sonnet provides the optimal risk-adjusted profile.

Table 5. Enterprise Deployment Recommendation Matrix

Use Case	Recommended Model	Alternative	Rationale
Medical / Legal Analysis	Claude 3.5 Sonnet	GPT-4o	Highest safety score (4.9) + accuracy; regulated domain
Code Generation	Claude 3.5 Sonnet	GPT-4o	92.0% HumanEval; strong instruction following; 200K context
Long-Document Processing	Gemini 1.5 Pro	Claude 3.5 Sonnet	1M-token context; book-length docs; no RAG needed
High-Volume Chatbot	Gemini 1.5 Flash	Claude 3 Haiku	$0.075/1M tokens; 210 tok/s throughput; optimal cost/volume
Air-Gapped / Data-Sovereign	LLaMA 3 70B	Mistral Large	Open-source; fully self-hostable; no third-party data egress
Multilingual Applications	Gemini 1.5 Pro	GPT-4o	Trained on broadest multilingual corpus; leads on mT5 evals
Mathematical Reasoning	GPT-4o	Claude 3.5 Sonnet	Leads on MATH benchmark (76.6%); structured reasoning tasks
Startup / Low-Budget	Gemini 1.5 Flash	Claude 3 Haiku	Combined cost-performance-speed advantage; MCDA = 0.773

Enterprise Deployment Risk Analysis

Table 6 classifies the primary risk categories associated with enterprise LLM deployment, identifies the most affected model categories, and provides mitigation strategies.

Table 6. Enterprise LLM Deployment Risk Categories and Mitigations

Risk Category	Description	Affected Models	Mitigation Strategy
Hallucination	Confident generation of factually incorrect or fabricated content	All models	RAG pipelines; confidence thresholds; human-in-the-loop verification
Prompt Injection	Adversarial inputs overriding system instructions or jailbreaking safety filters	All models	Input sanitization; output filtering; sandboxed execution environments
Data Privacy	PII leakage through memorization or context window exposure	Proprietary APIs	On-premise deployment; data minimization; DLP controls; GDPR compliance
Vendor Lock-in	Excessive dependency on a single API provider for mission-critical workflows	GPT-4o, Claude, Gemini	Multi-model architecture; open-source fallback; API abstraction layers

Algorithmic Bias	Systematic discriminatory outputs affecting protected groups	All models	Red-teaming; diverse evaluation datasets; bias metrics; human oversight
Training Data Poisoning	Supply chain contamination of training corpora affecting model behavior	All models	Model cards; transparency reports; third-party audits; anomaly detection

Framework Limitations

This study has several limitations. First, benchmark scores are static snapshots; rapid model updates may alter comparative rankings within weeks of publication. Second, MCDA weights reflect a generalized enterprise perspective; domain-specific deployments should recalibrate weights. Third, all primary benchmarks are conducted in English; multilingual capability differences may be substantially more pronounced in low-resource language evaluation. Fourth, the weighted sum MCDA model assumes linear independence among criteria; interaction-aware extensions (AHP, ANP) are warranted when criteria are strongly correlated. Fifth, API pricing is subject to change; organizations should verify current pricing at decision-making time.

CONCLUSION

This paper has presented a comprehensive comparative evaluation of seven state-of-the-art large language models using a structured, multi-dimensional MCDA framework integrating benchmark performance, safety alignment, cost efficiency, inference performance, and deployment flexibility. The analysis produces five principal findings: Claude 3.5 Sonnet achieves the highest MCDA composite score (0.801) driven by combined superiority in accuracy and safety alignment. Gemini 1.5 Flash represents the optimal choice for cost-sensitive, high-volume deployments at $0.075/1M input tokens and 210 tok/s throughput. GPT-4o leads specifically in mathematical reasoning (76.6% MATH) and multimodal capability. LLaMA 3 70B provides competitive value for data-sovereign contexts despite a 10–20 pp accuracy penalty versus frontier models. No single model dominates across all evaluation dimensions, confirming the necessity of multi-criteria analysis over single-metric leaderboard comparisons. Future research directions include: (1) empirical validation of MCDA rankings through controlled deployment pilots; (2) development of domain-specific frameworks for regulated industries; (3) longitudinal tracking of model performance evolution; (4) investigation of multi-model ensemble architectures; and (5) rigorous cross-linguistic evaluation extending to low-resource language contexts.

REFERENCES

Anthropic. (2024). Claude 3 model card [Technical report].

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., . . . Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv*:2212.08073.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language

models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., . . . Zaremba, W. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv*:2107.03374.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv*:2110.14168.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186).

Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120), 1–39.

Gemini Team. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., . . . Ma, Z. (2024). The Llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Hendrycks, D., Burns, C., Arora, S., Basart, S., Tang, E., Song, D., & Steinhardt, J. (2021a). Measuring mathematical problem solving with the MATH dataset. *Advances in Neural Information Processing Systems*, 34.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021b). Measuring massive multitask language understanding. In International Conference on Learning Representations.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., . . . Sifre, L. (2022). Training compute-optimal large language models. *Advances in Neural Information Processing Systems*, 35, 30016–30030.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Renard Lavaud, L., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., . . . Koreeda, Y. (2023). Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157–173.

OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022). Red teaming language models with language models. arXiv preprint arXiv:2202.03286.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). HellaSwag: Can a machine really finish your sentence? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 4791–4800).