## 1.Fundamentals of Generative AI and LLMs

### 1.1 The Hierarchy of AI

The evolution of modern Artificial Intelligence (AI) can be understood as a layered progression of technology. At the broadest level, AI encompasses all systems designed to replicate human like cognitive functions such as learning and decision making. Within AI, Machine Learning (ML) emerged as a major paradigm by enabling models to learn patterns from data. A significant subset of ML, Deep Learning (DL), relies on multi layered artificial neural networks that emulate the structure of the human brain [1] .These deep neural networks made it possible for models to automatically extract complex features from data, laying the groundwork for generative systems. Generative AI (GenAI) marks the shift from traditional "discriminative" models, which focus on analyzing or classifying existing data, to models that can autonomously create new content text, images, or code that is realistic, coherent, and novel.

### 1.2 Transformer Architecture

A major breakthrough that accelerated this transformation was the introduction of the Transformer architecture in 2017[2]. The key innovation behind Transformers is the self attention mechanism, which allows models to evaluate relationships between all words in a sentence simultaneously, rather than processing them sequentially. This enables an understanding of long range dependencies and contextual meaning, similar to how humans emphasize certain words for comprehension. The Transformer architecture became the foundation for influential models such as BERT and the GPT family [3], enabling them to scale to unprecedented levels and perform a wide range of natural language tasks with remarkable accuracy.

### 1.3 Pre training, Scaling Las, and Efficiency

Large Language Models (LLMs) are vast probabilistic models trained to predict the next token in a sequence, using enormous amounts of text data. During pre training, these models ingest trillions of tokens from diverse sources like CommonCrawl, Wikipedia, GitHub, and scientific papers. Research such as the LLaMA series demonstrated that high quality data and effective training strategies can allow smaller models to outperform much larger ones [4], challenging the assumption that parameter count alone determines capability. Scaling laws further show that performance correlates strongly with available compute and data [5], though efficiency during inference is increasingly important for real world deployment.

### 1.4 Alignment and Reinforcement

A pre-trained LLM, however, is simply an advanced text generator and is not inherently aligned with human values or intentions. To make models safer, more cooperative, and more reliable, developers apply Reinforcement Learning from Human Feedback (RLHF). This process begins with Supervised Fine-Tuning (SFT), where human-written examples teach the model desirable behavior [6]. Next, a reward model is trained using human-ranked responses to determine which outputs are preferred. Finally, reinforcement learning algorithms such as Proximal Policy Optimization (PPO) adjust the model to maximize these reward signals, making it more helpful, honest, and harmless.

### 1.5 Prompt Engineering Techniques

How users interact with LLMs known as prompt engineering also significantly affects performance. Techniques such as zero shot and few shot prompting help guide the model by either relying entirely on prior knowledge [7] or providing examples of the desired output. The chain of thought prompting encourages the model to break down complex reasoning tasks into intermediate steps [8], greatly improving outcomes when applied explicitly. Role prompting, where the model is assigned to a professional persona or specific identity, helps shape tone, structure, and domain relevance in generated responses.

### 1.6 Limitations

Despite their power, LLMs come with inherent limitations. They may hallucinate producing confident but factually incorrect statements [9] because they predict text based on statistical patterns rather than grounding their output in verified facts. They can also replicate or amplify biases present in training data, leading to skewed or unfair responses [10]. Additionally, LLMs remain in “black box” systems, making it difficult to fully understand how specific outputs are produced. These challenges highlight the need for ongoing research in transparency, alignment, and responsible deployment of AI systems.

## 2.LLM Products

The landscape of Large Language Models has evolved into diverse product families, each optimized for specific trade-offs between speed, cost, and reasoning capability. This section outlines the major proprietary and open source models currently defining the state of the art.

### 2.1 ChatGPT (OpenAI)

OpenAI's ChatGPT has evolved across multiple generations, culminating in GPT-5. This version introduces a unified architecture with dynamic routing, allowing the model to switch between a fast lightweight mode and a deeper reasoning mode ("GPT-5 Thinking"). The GPT-5 lineup includes variants such as nano, mini, and the flagship model, each balancing cost and capability differently. GPT-5 focuses on advanced reasoning and multimodal understanding but can produce dense, complex text and requires higher computational resources [11].

### 2.2 Gemini (Google)

Google's Gemini 3 family, including variants like Gemini 3 Pro, emphasizes advanced multimodal capabilities across text, audio, images, video, and code, with efficient deployment for complex tasks. Gemini 3 Pro is optimized for high-accuracy reasoning, agentic performance, and long-context processing up to 1 million tokens, while supporting reinforcement learning for multi-step problem-solving. The family prioritizes efficiency through a sparse mixture-of-experts architecture, making it suitable for developer tools, AI agents, and integrated systems, with notable strengths in benchmark-leading output quality and safety [12].

### 2.3 Claude (Anthropic)

Anthropic's Claude models, including Claude Opus 4.1 and Claude Sonnet-4, are built around the principles of Constitutional AI, prioritizing safety, reliability, and structured reasoning. They often generate well-organized and coherent text, though their style can sometimes become overly complex. Claude models are typically favored in scenarios requiring careful, rule-guided behavior and strong alignment with safety constraints [13].

### 2.4 Grok (xAI)

xAI's Grok-4 integrates real-time information from the X platform, allowing it to provide up-to-date responses. It is designed to be conversational, dynamic, and highly responsive to current events. Grok also emphasizes reliability in tasks requiring grounded information, especially in areas like reference retrieval. Its main stylistic characteristic is a tendency toward high-density, complex language.

### 2.5 Copilot (Microsoft)

Microsoft's Copilot is designed as an assistant integrated across the Windows ecosystem and productivity tools. It excels at general tasks such as summarizing content, drafting text, generating code, and interacting with Microsoft services. However, it requires external validation when used for academic or research heavy tasks due to limitations in handling detailed citation or scholarly material [14] .

### 2.6 DeepSeek

DeepSeek-v3.1 focuses on cost efficient deployment while maintaining high accuracy in structured tasks like reference generation. It performs well in theoretical and knowledge-based operations and is positioned as a highly competitive low-cost alternative to major proprietary LLMs. Its main limitation is inconsistency in tasks requiring detailed numerical or calculation-heavy reasoning [15].

### 2.7 LLaMA (Meta)

Meta's LLaMA family ranging from 7B to 65B parameters is fully open source and trained exclusively on publicly available datasets such as CommonCrawl, GitHub, Wikipedia, and ArXiv. LLaMA emphasizes accessibility, transparency, and customizability. Because of its open-source nature, it has become a widely adopted foundation model for both enterprise solutions and community-built LLM derivatives [16] .

## 3. LLM Performance Overview

This section provides an integrated evaluation of how modern AI models perform across domains such as medicine, dentistry, supply-chain decision-making, and academic reliability. Overall, frontier models like GPT-5, Gemini 3, Claude 4.x, and Grok-4.1 demonstrate near "expert-level" knowledge, but their strengths and weaknesses vary significantly by task type and domain.

### 3.1 Performance in Medical & Dental Domains

To evaluate how well the evaluators agreed with each other, Intraclass Correlation Coefficients (ICC) were calculated separately for accuracy and completeness scores. The normality of continuous variables was checked using the Shapiro–Wilk test. Normally distributed variables are presented as mean ± standard deviation (SD), while non-normal variables are presented as median (minimum–maximum).

Based on the normality results, group comparisons were done using ANOVA when there were more than two groups, and the data were normally distributed. When normality was not met, the Kruskal–Wallis test was used. If ANOVA showed significant differences, Bonferroni-adjusted post-hoc tests were applied. If Kruskal–Wallis showed significance, Dunn–Bonferroni post-hoc tests were used.

Correlations between scores were examined with Pearson or Spearman correlation coefficients, depending on data distribution.

Effect sizes were calculated using eta-squared ($\eta^2$) for both parametric and non-parametric tests to assess the practical importance of group differences.

All statistical analyses were performed using SPSS for Windows (version 27.0; IBM, Armonk, NY, USA). 5% types I error level was accepted, and statistical significance was set at $p < 0.05$ [17].

### 3.2 Supply Chain, Decision Making

This domain evaluates models on progressively harder tasks: theoretical multi-choice items, reasoning-based questions, and open numerical problems.

#### 3.2.1 Numerical vs. Theoretical Tasks

GPT-5 dominates numerical tasks, achieving 92% accuracy when no answer options are provided. By contrast, smaller models like Claude-Haiku 3.5 perform poorly on these tasks (under 20%). For theoretical, single-choice questions, Gemini-3 ranks highest.

#### 3.2.2 Impact of Chain of Thought Prompting

Studies show that implicit CoT (“think step by step”) often fails to improve accuracy and can even decrease performance in complex tasks, particularly for Claude

models.
However, explicit CoT, where the model is forced to write out its reasoning, leads to major jumps in accuracy.
DeepSeek, for example, rises from 0% → 80% on medium-difficulty numerical problems when explicit reasoning is required [18].

### 3.2.3 Cost Speed Accuracy Balance (AHP Analysis)

When accuracy, latency, and cost are weighted together, GPT-5 mini emerges as the most balanced model—offering high reasoning performance with significantly lower resource usage than the full GPT-5 model.

### 3.3 Academic Reliability & Hallucination Rates

A key finding of this study is the persistent and structurally embedded problem of hallucination in the bibliographic outputs of most chatbots. Hallucination occurs when a model, unable to retrieve a real source from its training distribution, generates a reference that is grammatically coherent and academically plausible but factually nonexistent. This behavior is not random; it reflects the probabilistic nature of LLMs, which prioritize producing linguistically well-formed output even in the absence of verified factual information. As a result, many fabricated references adopt the appearance of legitimate citations, blending real-sounding author names, common journal titles, and syntactically correct DOIs into entries that can be difficult for non-experts to detect false.

The results of this study show considerable differences across models. While Grok and DeepSeek did not fabricate any references, other systems most notably Copilot, Perplexity, and Claude displayed extremely high hallucination rates, sometimes generating entire lists of references that do not exist in any academic database. A particularly concerning behavior is the production of “structured” hallucinations, in which elements from real works (such as well-known authors or publishers) are mixed with incorrect details like fabricated article titles or invalid publication years. These references appear legitimate at a glance and therefore pose significant risks for students who might accept them uncritically.

Hallucination is also closely tied to the type of document requested. Models were far more likely to fabricate journal articles than books, likely because widely used textbooks appear frequently in their training data, whereas journal content—especially specialized or recent articles is less consistently represented. This helps explain why hallucination was more common in fields dominated by article based scholarship, such as Engineering, Experimental Sciences, and Health Sciences. In some cases, chatbots attempted to satisfy the user’s implied preference for recent sources by assigning

publication years in the 2020–2025 range to entirely fabricated works, reinforcing the idea that the system's goal is to produce plausible output rather than accurate information.

Overall, hallucination emerges as a fundamental limitation of current LLM based chatbots in academic tasks requiring citation-level precision. Even when models provide real references, they frequently contain subtle inaccuracies that stem from the same generative mechanisms responsible for full fabrication. These patterns underscore the importance of thorough manual verification and highlight the ongoing need for improved training data, better integration with authoritative bibliographic sources, and stronger educational strategies to ensure students use AI tools critically and responsibly [19].

### 3.3.1 Fabrication Rates

While Grok and DeepSeek demonstrated high integrity with zero fabricated references, Copilot exhibited significant issues, fabricating all of its journal citations. ChatGPT and Gemini presented a different challenge: they often generated 'hallucinated' citations that appeared plausible by pairing legitimate authors with non-existent article titles [20].

### 3.3.2 Document Type Differences

Models are far more reliable with book references (only 12.9% fabricated) compared to journal articles, where fabricated entries jump to 78% [21]. This significant discrepancy highlights that LLMs still struggle with the granularity of academic literature. While books are often treated as distinct, major entities within the model's training data, appearing frequently across multiple sources and editions, journal articles are far more numerous and specific. Consequently, instead of retrieving an exact article, models often resort to a probabilistic "reconstruction" strategy: they hallucinate citations by pairing real, established authors with plausible sounding but non-existent article titles that fit the context. This pattern-matching behavior explains why the error rate for specific journal articles remains critically high, rendering most models unreliable for deep literature reviews without rigorous human verification.

## 3.4 General NLP & Reasoning Benchmarks

Beyond domain-specific tasks, evaluating LLMs on general Natural Language Processing (NLP) and reasoning benchmarks provides a clearer picture of their fundamental capabilities. While earlier research focused heavily on scaling laws assuming that larger models inherently yield better results recent evidence suggests a paradigm shift. Current benchmarks indicate that architectural efficiency and

advanced training methodologies, such as instruction tuning, are often more critical for determinants of performance than raw parameter count alone.

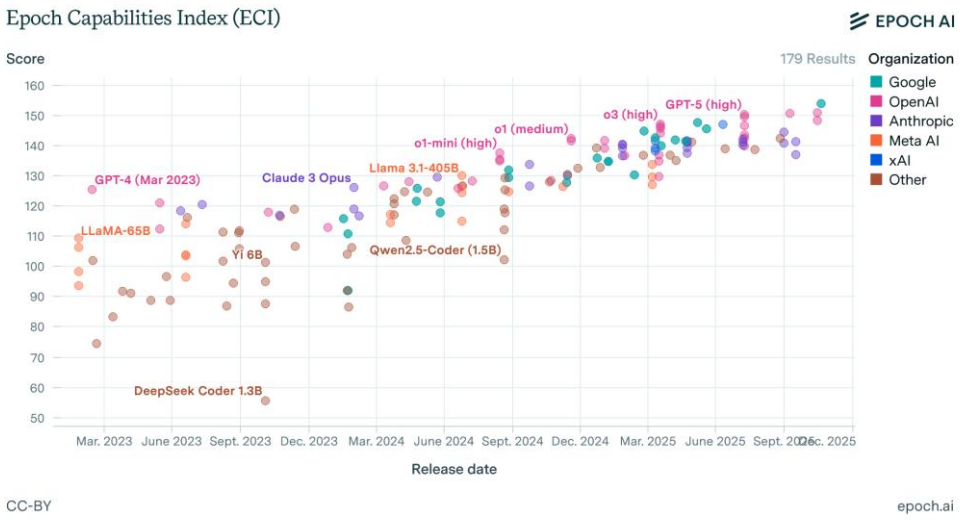


*Figure 1: The Epoch Capabilities Index (ECI) showing the progression of general model capabilities over time, highlighting the rapid rise of reasoning-focused models like the GPT-5 and o1 series.*

### 3.4.1 Zero-Shot Reasoning

The results challenge the idea that parameter count alone determines performance. LLaMA-65B outperforms GPT-3 (175B) on major reasoning benchmarks such as PIQA, SIQA, and OpenBookQA. On TriviaQA, it also surpasses Gopher-280B [16].

### 3.4.2 Instruction Following & RLHF

InstructGPT (1.3B) is preferred by human evaluators over GPT-3 (175B), demonstrating that RLHF is more influential than model size for alignment and helpfulness. RLHF fine-tuning also. Furthermore, the application of RLHF has been shown to increase model truthfulness by minimizing hallucinations, while simultaneously reducing the generation of toxic outputs by approximately 25%.

These findings show that training strategy and alignment techniques matter as much as, if not more than, raw scale [22].

## 4. References


[1] "Untangling deep learning from artificial intelligence and machine learning."

[2] A. Vaswani *et al.*, "Attention Is All You Need."

[3] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." [Online]. Available: https://github.com/tensorflow/tensor2tensor

[4] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.13971

[5] J. Kaplan *et al.*, "Scaling Laws for Neural Language Models," Jan. 2020, [Online]. Available: http://arxiv.org/abs/2001.08361

[6] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback."

[7] T. B. Brown *et al.*, "Language Models are Few-Shot Learners." [Online]. Available: https://commoncrawl.org/the-data/

[8] J. Wei *et al.*, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models Chain-of-Thought Prompting."

[9] Z. Ji *et al.*, "Survey of Hallucination in Natural Language Generation," Jul. 2024, doi: 10.1145/3571730.

[10] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, Inc, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[11] OpenAI, "GPT-4 Technical Report."

[12] "Gemini 3 Pro-Model Card Model Information."

[13] "The Claude 3 Model Family: Opus, Sonnet, Haiku Anthropic." [Online]. Available: https://docs.anthropic.com/

[14] S. Bubeck *et al.*, "Sparks of Artificial General Intelligence: Early experiments with GPT-4," Apr. 2023, [Online]. Available: http://arxiv.org/abs/2303.12712

[15] DeepSeek-AI *et al.*, "DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model," Jun. 2024, [Online]. Available: http://arxiv.org/abs/2405.04434

[16] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.13971

[17] M. Liu *et al.*, "Textbook-Level Medical Knowledge in Large Language Models: A Comparative Evaluation Using the Japanese National Medical Examination," Sep. 12, 2025. doi: 10.1101/2025.09.10.25335398.

[18] M. Yasunaga *et al.*, "LARGE LANGUAGE MODELS AS ANALOGICAL REASONERS."

[19] Á. Cabezas-Clavijo and P. Sidorenko-Bautista, “Assessing the performance of 8 AI chatbots in bibliographic reference retrieval: Grok and DeepSeek outperform ChatGPT, but none are fully accurate.”

[20] W. H. Walters and E. I. Wilder, “Fabrication and errors in the bibliographic citations generated by ChatGPT,” *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-41032-5.

[21] Z. Ji *et al.*, “Survey of Hallucination in Natural Language Generation,” Jul. 2024, doi: 10.1145/3571730.

[22] L. Ouyang *et al.*, “Training language models to follow instructions with human feedback.”