# A Comprehensive Review of Large Language Models (LLMs) in Artificial Intelligence and Machine Learning

## Abstract

Large Language Models (LLMs) — transformer-based models trained at massive scale — have rapidly become a central component of modern artificial intelligence. They power conversational agents, code generation, knowledge systems, and assistive tools in science, business, and creative industries. This review surveys the foundational principles, architectural evolution, training recipes, alignment methods, and applications of LLMs. The report details a comparative analysis of major models and families, including OpenAI’s GPT, Google’s Gemini, Meta’s LLaMA, and Anthropic’s Claude, highlighting the strategic shifts in their development, from raw scaling to efficiency and open-weight distribution. A significant focus is placed on addressing pressing challenges: the engineering trade-offs of environmental cost, the mitigation of hallucinations, and the nuanced issue of bias. The analysis concludes with a discussion of promising future directions, such as multimodality, retrieval-augmented generation (RAG), and a growing emphasis on explainability and policy considerations. The review provides a principled primer for students and researchers, offering a multi-layered understanding of the complex and rapidly evolving LLM landscape.

## 1. Introduction

### 1.1 The ubiquity of LLMs in the digital age

Over the last decade, LLMs have fundamentally reshaped the field of natural language processing (NLP) and broadened the set of tasks that can be automated or augmented by AI. The success of models like GPT-3/4, PaLM, LLaMA, and Claude stems from a potent combination of the transformer architecture, unprecedented scale in parameters and data, and refined training pipelines that align their behavior with human preferences. Foundational transformer work made the architecture practicable; since then, empirical scaling laws and compute trends have guided model design and training strategies. The rapid transition of LLMs from a niche research topic to a foundational technology is reflected in recent industry data. In 2024, private AI investment reached a global total of $33.9 billion, an 18.7% increase from the previous year, while the number of organizations using AI grew from 55% in 2023 to 78%. This accelerated adoption rate signals a shift in the industry, where the focus has moved beyond the question of *if* companies will adopt AI to the more complex question of *how* they will responsibly and effectively integrate it into their operations. This transition elevates the importance of addressing the core challenges of governance, bias, environmental cost, and safety, which are no longer theoretical concerns but defining characteristics of the field.

### 1.2 Scope and objectives of this review

This report aims to provide a compact but thorough review for students and researchers who wish to obtain a deeper understanding of the LLM ecosystem. The objectives of this work are to: (1) provide a principled primer on LLMs and their foundational transformer architecture; (2) explain modern training and alignment pipelines, including instruction tuning and RLHF; (3) conduct a comparative analysis of prominent models and their strategic design choices; (4) discuss the critical engineering and societal challenges that accompany their development and deployment; and (5) outline promising directions for future research and practice.

## 2. Foundational Principles of LLMs

### 2.1 Core paradigms of machine learning for language

The journey toward modern LLMs began with early statistical NLP methods that relied on n-gram models and Hidden Markov Models (HMMs). The advent of neural sequence models, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, marked a significant step forward by introducing distributed representations and improving the capacity for capturing sequential dependencies. However, these models were inherently sequential, processing data step-by-step, which presented significant challenges for training and for modeling very long-range dependencies. The introduction of distributed word embeddings like Word2Vec and GloVe provided dense vector representations for words. A further innovation was the development of contextual embeddings, exemplified by models like ELMo and BERT, which provided word representations that were conditioned on the specific context of a sentence, leading to a substantial improvement in performance on downstream tasks.

### 2.2 The Transformer Architecture and Self-Attention

The seminal 2017 paper, “Attention Is All You Need,” introduced the Transformer architecture, which fundamentally changed the field by replacing recurrence and convolutions entirely with a self-attention mechanism. This innovation was a pivotal moment in the history of LLMs because it enabled a high degree of parallelization, allowing for efficient scaling across large compute clusters. The self-attention mechanism works by assigning a weighted importance to each token in an input sequence relative to every other token, which is essential for understanding the relationships between words in a sentence.

The mechanism's operation can be intuitively understood through the creation of three primary vectors for each token in an input sequence: a query (Q), a key (K), and a value (V). The query vector represents the current token's search for relevance within the sequence. The key vector acts as a marker or "tag" for all tokens, and the value vector represents the content of the token itself. To determine the importance of one token to another, a similarity score is calculated between the query of the current token and the keys of all other tokens. These scores are then normalized and used to compute a weighted sum of the value vectors, which forms the new, contextually-aware representation of the token. This process can be executed simultaneously for all tokens in a sequence, a stark contrast to the sequential processing of RNNs. This parallelization is the foundational reason that the transformer architecture has enabled the training of models with hundreds of billions of parameters. The capacity for parallel computation, while a necessary innovation for training large models, is also the primary driver of their high computational and environmental costs, which are discussed in a later section.

### 2.3 The evolution of language models

The Transformer architecture paved the way for pretraining at an unprecedented scale. Two common pretraining objectives emerged: masked language modeling (MLM) and autoregressive causal language modeling. BERT, a model that uses the MLM objective, was designed as a bidirectional encoder that learns a deep representation of a text by predicting randomly masked tokens. This approach, combined with a fine-tuning step, proved highly effective for a wide range of NLP tasks. In contrast, models in the GPT family use a causal autoregressive objective, predicting the next token in a sequence. This approach naturally lends itself to text generation and has been the basis for models that popularized few-shot prompting, as demonstrated by GPT-3. These innovations in pretraining and prompting established an empirical observation that "more data + larger models" often yielded better performance. However, subsequent research, such as the work on Chinchilla scaling laws, refined this view by establishing a more nuanced trade-off between model size and data volume to achieve compute-optimal performance.

## 3. The Modern LLM Pipeline

### 3.1 Self-supervised pretraining at scale

The modern LLM pipeline begins with self-supervised training on massive corpora, which typically consist of web text, books, code, and dialogue data. This pretraining phase is the most computationally intensive step and is designed to impart general linguistic knowledge and broad capabilities to the model. To prepare the data, a process called tokenization is used to convert the raw text into numerical representations that the model can process. A common method is Byte-Pair Encoding (BPE), which identifies frequent pairs of characters and merges them into new tokens, creating a vocabulary of tokens that can compress large datasets. However, tokenizers optimized for English can be suboptimal for other languages, sometimes using up to 15 times more tokens per word for languages like Shan or having a 50% premium for languages like Portuguese and German. The pretraining phase also involves extensive data cleaning to remove low-quality, toxic, or duplicated content, a process that can be further enhanced by using a trained LLM to clean datasets for future models.

### 3.2 Fine-tuning and alignment techniques

Following pretraining, models undergo a multi-stage refinement process to align their behavior with human expectations. The first stage is often Supervised Fine-Tuning (SFT), where the pre-trained model is fine-tuned on a smaller dataset of high-quality, human-curated examples of prompts and desired responses. This step primes the model to follow instructions and generate outputs in a useful format, effectively teaching it to transition from being a simple text completer to a helpful assistant.

The next and most critical stage for creating safe and useful models is Reinforcement Learning from Human Feedback (RLHF). This process involves a feedback loop where human preferences guide the model's behavior. The process typically involves three phases :

1. **Data Collection:** A dataset of prompts is created, and the model generates several different responses for each prompt. Humans then rank these responses based on criteria like helpfulness, honesty, and safety.
2. **Reward Model Training:** A separate, smaller model, known as the "reward model," is trained on this human preference data. The goal of this model is to learn to predict the human preference score for any given response.
3. **Reinforcement Learning (RL) Optimization:** The original LLM, now called the "policy model," is fine-tuned using a reinforcement learning algorithm (e.g., Proximal Policy Optimization). The reward model acts as the "reward function" for the RL algorithm, providing a score for each response generated by the policy model. The policy model then iteratively updates its weights to maximize this reward, aligning its outputs with the human judgments encoded in the reward model.

RLHF represents a profound strategic shift. It transforms a model from a mere statistical pattern predictor into an agent that is explicitly optimized to embody human values of helpfulness and safety, effectively addressing the complex challenge of aligning AI behavior with human goals in a practical, scalable way.

### 3.3 Retrieval-Augmented Generation (RAG) and non-parametric memory

While LLMs store vast amounts of knowledge implicitly in their parameters, this knowledge is static and prone to becoming outdated or factually incorrect. Retrieval-Augmented Generation (RAG) is an architectural framework designed to address this by coupling LLMs with an external knowledge base or retrieval system. This approach improves factual accuracy and makes knowledge maintenance tractable.

The RAG process typically involves three main steps :

1. **Retrieval:** When a user submits a query, an information retrieval component searches an external data source (e.g., a document repository, a vector database, or the web) for relevant information. Vector databases, in particular, are used to store documents as numerical embeddings, enabling efficient retrieval based on semantic similarity.
2. **Augmentation:** The retrieved documents or snippets of information are then used to augment the user's original prompt. This newly-composed, augmented prompt, which includes both the user's query and the relevant context, is then passed to the LLM.
3. **Generation:** The LLM generates a response based on the augmented prompt. By grounding its generation in the retrieved, up-to-date information, the model can produce a more accurate and verifiable output.

RAG is a critical architectural solution because it makes the model's knowledge dynamic and auditable, effectively shifting the paradigm from "train on everything" to "reason with relevant information." This is a significant step toward building trustworthy and up-to-date AI systems, and it mitigates the critical problem of hallucinations and data staleness in a scalable, cost-effective manner.

### 3.4 Evaluation and benchmarks

LLMs are evaluated using a combination of methods. Standardized task benchmarks, such as GLUE, SuperGLUE, and MMLU, measure performance on a range of linguistic and reasoning tasks. However, these benchmarks are often complemented by human evaluations to rate the model's helpfulness and safety. A growing emphasis is placed on emergent properties, such as chain-of-thought reasoning, which can be elicited through sophisticated prompting techniques. It is also critical that benchmarking accounts for training-data contamination and includes robust, held-out, and adversarial tests to ensure the reliability and generalizability of the results.

## 4. Comparative Analysis of Leading LLM Architectures

The LLM landscape is dominated by a few major players, each with a distinct architectural and business strategy. While early models competed on raw size, the field is now marked by strategic differentiation on factors such as multimodality, efficiency, and open-source accessibility.

### 4.1 OpenAI’s GPT family

The GPT family, developed by OpenAI, has been a central force in the popularization of LLMs. GPT-3, with its 175 billion parameters, revolutionized the field by demonstrating impressive zero- and few-shot learning abilities, showcasing that a model could perform new tasks with only a handful of examples in the prompt, without any further fine-tuning. The subsequent release of GPT-4 extended these capabilities, notably with multimodality, allowing it to process and reason over both text and images. The GPT family is known for its proprietary, closed-source nature, which has raised concerns about distribution control, customization, and cost. However, OpenAI has made a strategic move to address this by releasing "open-weight" models under a permissive license. This provides strong real-world performance at a lower cost, enabling more efficient deployment and use in agentic workflows.

### 4.2 Google’s PaLM and Gemini

Google’s PaLM model scaled to hundreds of billions of parameters, demonstrating strong performance on few-shot and reasoning benchmarks. A key technical contribution of PaLM was its efficient training on TPU clouds using pathways systems. Google’s subsequent Gemini effort emphasizes a more holistic, natively multimodal approach. Gemini’s core innovation lies in its ability to process and reason over text, images, audio, and video from the outset, rather than as a secondary add-on. Additionally, Gemini models include advanced agentic features like "Deep Research" and "Deep Think". Deep Research is an agentic feature that can autonomously browse the web, think through its findings, and synthesize them into multi-page reports. It addresses the technical challenge of long-running inference with an asynchronous task manager, which allows for graceful error recovery and enables multi-step, iterative research without restarting from scratch. This approach enables Gemini to act as a part of a larger, more complex decision system, moving beyond simple generation to more sophisticated, goal-oriented reasoning.

### 4.3 Meta’s LLaMA family: The power of open and efficient models

Meta’s LLaMA family of models represents a strategic counterpoint to the closed-source giants. LLaMA focuses on efficiency and open research, with smaller parameter models (7B-65B) trained on high-quality, publicly available corpora. This strategy has democratized access to powerful foundation models and accelerated the ecosystem of open-weight models and community fine-tuning. The release of LLaMA 3.1 further cemented this position, with the introduction of a new 405 billion-parameter model with a greatly increased context length of 128,000 tokens. This model’s performance is comparable to leading closed-source models like GPT-4, and its open-weight nature allows organizations to fine-tune it for specialized applications without licensing restrictions. This approach signals a critical market split, where Meta is cultivating a massive developer community that rapidly prototypes specialized models, presenting a clear contrast to the proprietary model of OpenAI and Anthropic. This fosters flexibility and innovation, especially for companies that prioritize data security and customization by running models locally within a secure environment.

### 4.4 Anthropic’s Claude and Constitutional AI

Anthropic, founded by former OpenAI employees, introduced the Claude family of models with a core focus on safety-by-design. Claude models are developed using a training method called "Constitutional AI," where a set of human-readable principles or a "constitution" guides the model's training and self-supervision to produce safer and more harmless outputs. This method offers an alternative to relying solely on human feedback for alignment, enhancing transparency in the rules that shape the model’s behavior. Claude is particularly well-suited for applications that require extended contextual understanding, such as technical support and complex problem-solving. Its safety characteristics and ability to be honest about its limitations make it a strong choice for high-stakes applications where ethical considerations are paramount.

### 4.5 The trend towards efficient and specialized architectures

Beyond the major players, a significant trend has emerged toward efficiency and compute-optimal strategies. Research into scaling laws, notably the Chinchilla paper, established that training on more tokens with moderately smaller parameter counts can be more effective than simply scaling parameters blindly. This lesson has motivated the development of specialized and distilled variants that maintain capability while lowering inference costs. A key architectural innovation supporting this trend is the Mixture-of-Experts (MoE) architecture, used in models like Mixtral and GPT-4. MoE models contain multiple smaller neural networks, or "experts". For a given input, a "gating network" or "router" dynamically activates and routes the input to only a small subset of these experts, combining their outputs to form the final response. This sparse activation allows models to achieve high capacity with significantly lower computational overhead during inference, a critical step toward reducing the environmental and economic costs of deployment. The shift toward MoE and other efficiency-focused architectures is a direct response to the scalability challenges of traditional dense models.

<br>

| **Model Family** | **Typical Parameter Counts** | **Architecture & Key Feature** | **Open/Closed-Source** | **Primary Use Cases** | **Differentiating Factor** |
| --- | --- | --- | --- | --- | --- |
| **OpenAI GPT** | GPT-3 (175B), GPT-4 (MoE, 1.7T est.) | Decoder-only Transformer, Few-shot/Zero-shot learning, Multimodality (GPT-4) | Closed-Source, but with open-weight models released | Creative writing, code generation, complex reasoning | High versatility and state-of-the-art performance, but at a high cost |
| **Google Gemini** | Nano, Pro, Ultra | Multimodal Transformer, Native multimodality | Closed-Source | Complex reasoning, agentic tasks, scientific discovery, iterative design | Native multimodal understanding and advanced agentic features like Deep Research/Think |
| **Meta LLaMA** | 7B to 405B (LLaMA 3.1) | Decoder-only Transformer | Open-Weight | Research, on-device deployment, customized fine-tuning | Open-source nature democratizes AI and enables flexibility/customization |
| **Anthropic Claude** | Not publicly disclosed | Constitutional AI-trained Transformer | Closed-Source | High-context tasks, technical support, safety-critical applications | Emphasis on safety and ethical considerations through Constitutional AI |
| **Mistral AI** | 7B to 141B (MoE) | Mixture-of-Experts (MoE) | Open-Source/Open-Weight | Efficient deployment, compliance-focused applications | High performance-to-cost ratio due to MoE architecture; open and aligned with EU regulations |

<br>

## 5. Critical Challenges and Engineering Considerations

### 5.1 Data quality, provenance, and societal bias

LLMs learn from massive, web-scale corpora that inevitably contain societal biases, inaccuracies, and toxic content. These biases can be amplified during generation and require complex mitigation strategies. The problem of bias is not static; it is a highly contextual issue that shifts depending on the specific application. For example, a bias mitigation strategy that works for financial decision-making may not be effective for commercial transactions or hiring decisions. This means that the sources of bias do not reside in a single, fixed location within the model's architecture.

Recent research has explored a novel approach to addressing this by "pruning" specific computational units within the model, such as artificial neurons or attention heads, that are identified as contributing to biased behavior. The findings indicate that neuron-level pruning is more effective at reducing bias while maintaining the model’s overall utility. The contextual nature of bias has profound implications for legal and regulatory frameworks. It suggests that holding AI model developers liable for all harmful outputs may be impractical. Instead, a more effective legal and policy approach is to hold accountable the companies that are deploying the models in a particular use case. This encourages regulators to require companies that deploy AI models to conduct rigorous bias audits, maintain transparency about their AI usage, and ensure compliance with anti-discrimination laws.

### 5.2 Hallucinations and factuality

LLMs are prone to hallucinations, which are confidently produced but incorrect statements. These are not simply "mistakes" but are a natural consequence of a model’s probabilistic generation process, where it predicts the most statistically likely next token rather than retrieving a verified fact from a knowledge base. The primary technical mitigation strategy is Retrieval-Augmented Generation (RAG), which grounds the model’s output in external knowledge. By augmenting the prompt with verified information, RAG significantly reduces the risk of factual errors. Other strategies include providing the LLM with possible options, asking it to provide attributions, and using an auxiliary LLM to evaluate the generated output for trustworthiness. A promising area of research is the use of "confidence scores" to identify potential hallucinations and flag untrustworthy outputs for human review.

<br>

| **Strategy** | **Description** | **Key Benefit** | **Drawbacks/Trade-offs** |
| --- | --- | --- | --- |
| **Retrieval-Augmented Generation (RAG)** | Augmenting the user's prompt with information from an external, verified knowledge base before generation. | Improves factuality, provides up-to-date information, and makes outputs verifiable. | Requires maintaining and updating an external knowledge base, can be complex to implement, and requires effective search algorithms. |
| **Fine-tuning** | Fine-tuning the model on a dataset of factually correct and verifiable examples. | Can improve the model's general factual grounding and response style. | Costly and time-consuming to create and maintain high-quality fine-tuning datasets. |
| **Prompt Engineering** | Instructing the model to be cautious or to state its lack of knowledge when a query is ambiguous or lacks relevant data. | Inexpensive and simple to implement. | Can be brittle and less reliable than architectural solutions, and depends heavily on the user's ability to craft effective prompts. |
| **Confidence Scoring** | Using an external or auxiliary model to generate a score that estimates the trustworthiness of a generated response. | Can automatically flag potentially incorrect responses for review, reducing the need for extensive human oversight. | Requires a separate system or model to be effective; not a fix for the underlying generation process. |

<br>

### 5.3 Scalability, compute, and environmental cost

The training and operation of state-of-the-art LLMs require massive computational resources and specialized hardware. The environmental footprint of a single model's training can be significant; for example, a 2020 paper estimated that training GPT-3 could produce as much carbon dioxide as five cars over their lifetimes. Beyond the one-time training cost, the ongoing inference phase—where the model is deployed and used for daily queries—can be even more problematic. For an application like ChatGPT, the computational cost of inference can be as much as 25 times greater per year than the initial training cost. Similarly, for a system like the Google search engine, which is shifting to use LLMs, the cost could be up to 1,400 times the annual training cost.

The environmental impact also extends to water consumption. The training of a model like ChatGPT is estimated to have consumed 700,000 liters of water, equivalent to a household's consumption over five years. The ongoing inference phase is estimated to use 500 milliliters of fresh water for every 20-50 user requests. The fact that inference costs can be orders of magnitude greater than training costs reveals a critical challenge: the industry’s focus must shift from simply optimizing the training of new models to making every single user query as efficient as possible. This highlights the importance of architectures like MoE, which can reduce computations and data transfer by a factor of 10 to 100 times during inference. Efforts to mitigate these environmental impacts include improving data center sustainability through the use of green energy and more efficient cooling methods, as well as hardware and software co-design to improve computational efficiency.

<br>

|  | **Training** | **Inference** |
| --- | --- | --- |
| **Carbon Footprint** | An estimated 284 tons of CO2eq for BERT. Training a model like GPT-3 could produce as much CO2 as five cars over their lifetimes. | Can be orders of magnitude greater than the training cost over time. For an application like ChatGPT, the annual computational cost could be 25x the training cost. |
| **Water Consumption** | An estimated 700,000 liters for ChatGPT's training, equivalent to a household's consumption over five years. | Estimated to be 500 ml of fresh water for every 20-50 user requests. This is a continuous, compounding cost with millions of users. |
| **Mitigation Strategies** | Compute-optimal training strategies (e.g., Chinchilla), energy-efficient architectures (e.g., MoE) , and using data centers powered by renewable energy. | Efficient architectures (e.g., MoE) , quantization, and distillation to reduce the compute needed per query. |

<br>

## 6. Emerging Directions and Promising Research Areas

The LLM landscape is constantly evolving, with several key trends shaping the next generation of models:

**Multimodal Foundation Models:** The integration of text, image, audio, and video inputs is expanding the applications of LLMs into new domains such as robotics, medical imaging, and creative tools. Models like GPT-4 and Google's Gemini are at the forefront of this trend, demonstrating richer understanding and generation capabilities by processing multiple modalities natively.

**Retrieval / Tool-Augmented Agents and Modular Systems:** The evolution beyond simple RAG is toward more sophisticated agentic systems. By combining generation with deterministic external logic, such as search indices, APIs, and calculators, LLMs can act as agents that can perform multi-step, iterative tasks. Examples include Google's Deep Research and Deep Think, which enable the model to formulate a multi-step plan, autonomously search for information, and synthesize its findings into a comprehensive report. This trend increases reliability and makes models a functional part of larger, more complex decision systems.

**Efficient Models and Deployment Strategies:** The focus on efficiency, driven by both economic and environmental concerns, will continue to grow. Techniques like distillation, where a smaller model is trained to mimic the behavior of a larger one, as well as sparsely activated models (MoE) and quantization, will reduce inference costs and enable on-device or low-cost cloud deployment.

**Responsible AI, Standards, and Regulation:** As LLMs become more integrated into high-stakes domains, the need for robust standards and regulatory frameworks has become more urgent. Initiatives such as the EU AI Act and research into a risk-based approach to governance highlight the growing importance of establishing clear policies for safe and ethical deployment. Standards for model documentation, such as "model cards" or "datasheets," will also become essential for ensuring transparency and accountability.

## 7. Applications & Societal Impact

The widespread adoption of LLMs is having a profound effect across various sectors. In productivity and creativity, they are transforming workflows by augmenting human capabilities in writing, coding, and design. LLMs are also poised to benefit education and healthcare through applications like personalized tutoring and medical summarization, though these require careful evaluation of accuracy and liability. However, the technology also presents significant challenges. The ability of LLMs to generate plausible but false narratives at scale increases the risk of misinformation and requires the development of robust detection tools and responsible usage policies. Furthermore, the automation of tasks may shift labor demand and require strategic reskilling and policy measures to manage workforce transitions.

## 8. Conclusion

LLMs represent a major milestone in the history of artificial intelligence. The field has transitioned from foundational algorithmic breakthroughs, such as the Transformer architecture, to a complex ecosystem defined by scaling laws and sophisticated training and alignment pipelines. While the raw power of these models is now well established, significant challenges remain, including ensuring factuality, mitigating societal bias, managing environmental costs, and addressing privacy concerns. The analysis indicates that the future of LLMs is moving away from a singular focus on raw scale toward a more nuanced landscape defined by efficiency, multimodality, and verifiable, agentic behavior. The critical problems of environmental cost and bias, in particular, are driving a fundamental shift toward more efficient architectures and a re-evaluation of legal and policy frameworks. To realize the full benefits of this technology responsibly, continued cross-disciplinary work—combining machine learning research, systems engineering, ethics, and public policy—will be essential.

#### Works cited

1. Large language model - Wikipedia, https://en.wikipedia.org/wiki/Large\_language\_model 2. The 2025 AI Index Report | Stanford HAI, https://hai.stanford.edu/ai-index/2025-ai-index-report 3. What is self-attention? | IBM, https://www.ibm.com/think/topics/self-attention 4. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, https://blog.paperspace.com/bert-pre-training-of-deep-bidirectional-transformers-for-language-understanding/ 5. Attention is All you Need - NIPS, https://papers.neurips.cc/paper/7181-attention-is-all-you-need.pdf 6. brightmart/bert\_language\_understanding: Pre-training of Deep Bidirectional Transformers for Language Understanding: pre-train TextCNN - GitHub, https://github.com/brightmart/bert\_language\_understanding 7. Popular LLM Models: GPT, Claude, Llama, and Others Compared ..., https://medium.com/@rizqimulkisrc/popular-llm-models-gpt-claude-llama-and-others-compared-eaed7c5da1b3 8. What Is Reinforcement Learning From Human Feedback (RLHF)? - IBM, https://www.ibm.com/think/topics/rlhf 9. What is RLHF? - Reinforcement Learning from Human Feedback Explained - AWS, https://aws.amazon.com/what-is/reinforcement-learning-from-human-feedback/ 10. What is Retrieval-Augmented Generation (RAG)? - Google Cloud, https://cloud.google.com/use-cases/retrieval-augmented-generation 11. What is RAG? - Retrieval-Augmented Generation AI Explained - AWS - Updated 2025, https://aws.amazon.com/what-is/retrieval-augmented-generation/ 12. What are AI hallucinations & how to mitigate them in LLMs | by Roberto Cadili - Medium, https://medium.com/low-code-for-advanced-data-science/what-are-ai-hallucinations-how-to-mitigate-them-in-llms-4abf0cd54a7a 13. Comparing GPT, Claude, Llama, and Mistral: Which Large ..., https://www.epista.com/knowledge/break-down-four-of-the-biggest-players-in-ai-gpt-claude-llama-and-mistral 14. Top 9 Large Language Models as of September 2025 - Shakudo, https://www.shakudo.io/blog/top-9-large-language-models 15. Gemini Deep Research — your personal research assistant, https://gemini.google/overview/deep-research/ 16. Gemini 2.5 Pro - Google DeepMind, https://deepmind.google/models/gemini/pro/ 17. Gemini Deep Research – your personal research assistant, https://gemini.google/overview/deep-research/?hl=en-GB 18. 9 Top Open-Source LLMs for 2024 and Their Uses - DataCamp, https://www.datacamp.com/blog/top-open-source-llms 19. The Llama 3 Herd of Models | Research - AI at Meta, https://ai.meta.com/research/publications/the-llama-3-herd-of-models/ 20. en.wikipedia.org, https://en.wikipedia.org/wiki/Mistral\_AI#:~:text=The%20model%20uses%20an%20architecture,experts%2C%20but%20offering%20higher%20performance. 21. Mixture of Experts (MoE) Architecture in Modern Machine Learning 1. Introduction - Medium, https://medium.com/@prabhuss73/mixture-of-experts-moe-architecture-in-modern-machine-learning-1-introduction-b9f5930d860f 22. Measuring the environmental impact of delivering AI at Google Scale, https://services.google.com/fh/files/misc/measuring\_the\_environmental\_impact\_of\_delivering\_ai\_at\_google\_scale.pdf 23. Mistral AI: Models, Capabilities and Latest Developments | Built In, https://builtin.com/articles/mistral-ai 24. Measuring the environmental impact of AI inference | Google Cloud Blog, https://cloud.google.com/blog/products/infrastructure/measuring-the-environmental-impact-of-ai-inference 25. Bias and Fairness in Large Language Models: A Survey - MIT Press Direct, https://direct.mit.edu/coli/article/50/3/1097/121961/Bias-and-Fairness-in-Large-Language-Models-A 26. Bias in Large Language Models—and Who Should Be Held ..., https://law.stanford.edu/press/bias-in-large-language-models-and-who-should-be-held-accountable/ 27. Tips to measure confidence and mitigate LLM hallucinations : r/OpenAI - Reddit, https://www.reddit.com/r/OpenAI/comments/1g6fiaa/tips\_to\_measure\_confidence\_and\_mitigate\_llm/ 28. Managing the environmental impact of Large Language ... - Eviden, https://eviden.com/insights/blogs/llms-and-the-effect-on-the-environment/ 29. 18 Artificial Intelligence LLM Trends in 2025 | by Gianpiero Andrenacci | AI Bistrot | Medium, https://medium.com/data-bistrot/15-artificial-intelligence-llm-trends-in-2024-618a058c9fdf