Available at
www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

A Comprehensive Review of Large Language Model (LLM) in Artificial Intelligence and Machine Learning

Kishan Shingarakhiya[1], Ms Ashvini vaidya[2]
1. (B Tech in Computer Engineering, Atmiya University, Rajkot, India
Email: kishanshingrakhiya3805@gmail.com)
2. (Faculty of Engineering and Technology (CE), Atmiya University, Rajkot, India
Email: ashvini.vaidya@atmiyauni.ac.in)

Abstract- Large Language Models (LLMs) emerge today powering, amongst other things, chatbots, code generation, knowledge systems, and creative tools. This review traces their foundations, evolution, training, alignment, and use. It contrasts the main families of models-GPT, Gemini, LLaMA, and Claude-tracing recent trends of scaling and open-weight distribution. In this backdrop, go-to key challenges stand to be trade-offs in efficiency, environmental cost, hallucinations, and biases. The future unfolds into multimodality, retrieval-augmented generation, explainability, and policy framework.

1. Introduction

1.1 The ubiquity of LLMs in the digital age

Throughout the last ten years, LLMs have managed to take NLP from an esoteric area of research and turned it into a foundational AI technology. Their success originates from coupling transformers, immense barbells, and fine-tuned training aimed at human preference. Prompted by scaling laws and compute trends, now they have become the mainstream mass adoption trends, garnering investments of $33.9 billion in 2024 and 78% usage. From this point onwards, instead of questioning whether companies will adopt AI, it will be questioned how they will do so responsibly. This makes governance, bias, environmental cost, and safety the central issues of the field.

1.2 Scope and objectives of this review

This report offers a comprehensive yet succinct survey of the LLM ecosystem. It covers their transformer-based foundations, methods guided by alignment- and training-based objectives, including instruction tuning and RLHF, and presents a comparative analysis of major models with respect to their design choices and engineering, as well as societal challenges, with an outlook into the eventual research directions and practical applications.

2. Foundational Principles of LLMs

2.1 Core paradigms of machine learning for language

We speak of standard computational linguists: A first generation-say statistical language methods, established upon simple concepts of n-gram-models and HMMs-and so forth, gave way to the second generation of neural modelers, empowering RNNs and LSTMs,

distributing representations yet suffering from long-range dependency. Of course, the turn of the century brought Word2Vec and GloVe, capable of representing words in high-dimensional vector spaces with meaning, but it was the advent of contextualized word embeddings derived from ELMo and BERT that truly liberated the field of NLP into a new horizon, where downstream tasks earn unprecedented performance gains.

2.2 The Transformer Architecture and Self-Attention

Introduced initially were the Transformer and its self-attention alongside the concepts of replacing recurrence and convolutions, positing this as a simpler base that allowed parallelization of large-scale training. Weights in self-attention go between tokens, and such a weighting phenomenon helps models track relationships necessary for sentence understanding: each token stands for Q (query), K (key), and V (value) vectors. The degree of match between queries and keys ends up being used to weigh the combination of values into a contextual representation. So unlike RNNs, which work step-by-step through a sentence, Transformers treat all tokens at once, scaling up to several hundred billion parameters. It is often the downside of the high parallelization: computation is great, and so is the environment cost.

2.3 The evolution of language models

Two large-scale pretraining objectives were enabled by the Transformer: Masked Language Modeling and Autoregressive Modeling. Under the MLM setting, BERT masks some tokens in the input and tries to predict them, doing well on many NLP tasks when fine-tuned. The GPT models use the autoregressive setting to predict the next token, enabling programs for text generation and few-shot prompting, as in GPT-3. Chinchilla scaling laws came in later to dictate more precise best practices for maximizing efficiency, indicating that the bigger model and more data paradigm was not the most efficient way to go after although there was improvement in performance.

3. The Modern LLM Pipeline

3.1 Self-supervised pretraining at scale

The LLM pipeline begins with self-supervised training on massive datasets of text, code, and dialogue to build broad language abilities. Raw text is tokenized into numerical form, often using Byte-Pair Encoding to compress data efficiently, though English-optimized tokenizers can perform poorly for some languages. Pretraining also involves extensive data cleaning to remove low-quality, toxic, or duplicate content, sometimes using LLMs themselves to refine future training datasets.

3.2 Fine-tuning and alignment techniques

Sometimes, the static knowledge in an LLM's parameters can wear out of date or even become erroneous. RAG that combine an LLM with external knowledge bases to improve factuality and ease the updating of knowledge.

1. **Data Collection:** The bevy of diverse prompts is created, the model is then rendered to generate any number of responses for each prompt, and then humans rank those responses according to

whichever criteria-they consider useful, honest, safe, etc.

2. **Reward Model Training:** A smaller separate model, called the reward model, is then trained on this data of human preferences so that it can predict the human preference score given to any response.
3. **Reinforcement Learning Optimization:** Here, the original LLM is renamed the policy model and fine-tuned using an RL algorithm (e.g., PPO). The reward model serves as the reward function in the RL algorithm, providing a reward for each response generated by the policy model. The policy model updates its weights iteratively to maximize the associated reward and thereby align its outputs with human judgments encoded in the reward model.

RLHF turns LLMs from pattern predictors into agents optimized along human values such as helpfulness and safety, therefore providing a practical and scalable way to bring AI behavior into alignment with human goals.

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

While LLMs possess the implicit knowledge of a corpus endowed in their parameters, such knowledge is static and prone to becoming out of date or incorrect in respect of facts. RAG is an architecture created for addressing such an issue by joining the LLMs with an external knowledge base or a retrieval system, thus improving factual accuracy while making knowledge maintenance tractable.

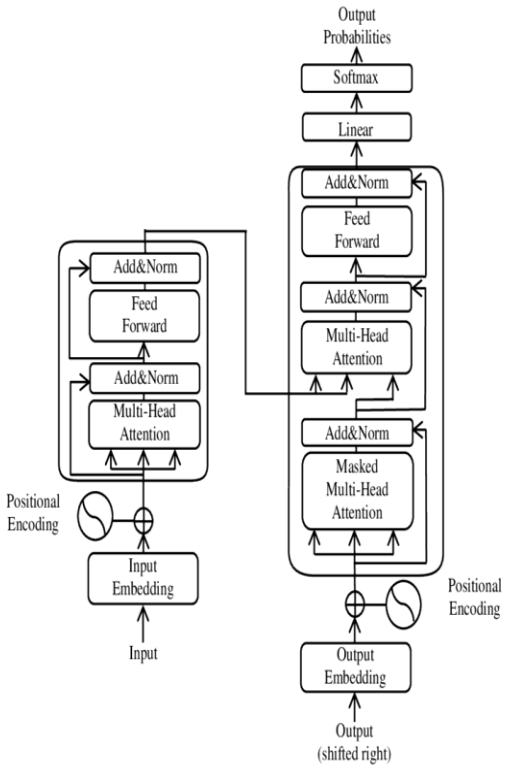
The three main steps in general are:

1. **Retrieval:** A user poses a query to an information retrieval component that subsequently fetches relevant information out of an external data source (e.g., a document repository, a vector DB, or the web). Vector DBs specifically consider storing documents as numerical embeddings to efficiently retrieve them based on semantic similarity.
2. **Augmentation:** The documents or snippets of information are used to amplify the original user prompt. This composed prompt is an augmented one, consisting of the user's query and relevant context, and is therefore passed to the LLM.
3. **Generation:** The LLM generates a response based on the augmented prompt. By grounding the generation in retrieved, up-to-date information, the model produces an accurate and checkable output.

Ran is the transformation of LLMs: it makes knowledge dynamic and auditable, from "train on everything" to "reason with relevant information," hence decreasing hallucinations and data staleness, while supporting a trustworthy, up-to-date, and scalable AI.

3.4 Evaluation and benchmarks

Benchmarks like GLUE, SuperGLUE, and MMLU are used for evaluating language and reasoning tasks, while human assessments measure helpfulness and safety. Emphasis is increasingly being placed on emergent abilities such as chain-of-thought-type reasoning, with robust evaluation against contamination, held-out tests, and adversarial scenarios to ensure credibility and generalizability of the results.



4. Comparative Analysis of Leading LLM Architectures

There are only a few major players in the LLM industry, each applying its own architectural thought and business strategy. While early models competed on sheer size, the field today is strategically differentiated on factors such as multimodality, efficiency, and open-source accessibility.

4.1 OpenAI's GPT family

LLMs have been made popular by OpenAI's GPT family of models, which range in power, the smaller GPT-3 (at 175B parameters) allowing powerful few-shot learning while the bigger GPT-4 added multimodal reasoning across text and images. The closed-source nature keeps it from being tailor-made for a specific purpose and adds to its final cost, with OpenAI doing a bit of mitigation by releasing some open-weight models for cheaper deployment.

4.2 Google's PaLM and Gemini

PaLM scaled into the domain of hundreds of billions of parameters, making it exceptional in few-shot learning and reasoning, trained efficiently on TPU clouds via pathways. Google's Gemini has built upon this with true multimodality-from scratch; that is, it works with text, images, audio, and video-and is agentic, including Deep Research, where the agent can autonomously browse, reason, and synthesize multi-step reports, allowing Gemini to be integrated into advanced decision systems rather than being a mere text generator.

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

Meta has developed LLaMA models focused on open, efficient research, ranging from smaller models (7B–65B) trained on public datasets, providing open-weight access for fine-tuning by the community. LLaMA 3.1 scaled up to 405B parameters with 128,000-token context, having performance comparable to GPT-4 and the like. However, LLaMA 3.1 also provides the ability to be outfitted and locally deployed in a secure manner. This open environment encourages the quick building of specialized models, setting it apart from the closed environments of OpenAI and Anthropic.

4.4 Anthropic's Claude and Constitutional AI

Claude by Anthropic, developed by former OpenAI staff, focuses on safety-by-design through Constitutional AI processes whereby human-readable principles dictate training and self-supervision for outputs that are safe and transparent. Claude functions best in long-

context scenarios such as technical support and complicated problem-solving, rendering it suitable for high-stakes, ethics-sensitive applications.

4.5 The trend towards efficient and specialized architectures

There has been an efficiency-oriented trend in line with scaling laws like those set out by Chinchilla, preferring to train more tokens but with fewer parameters. This engendered specialized and distilled models, of which Mixture-of-Experts (MoE) varieties would be found in architectures behind Mixtral and GPT-4. An MoE routes an input to a very small subset of expert networks and combines their outputs, thereby providing a very high-capacity response at lower computational cost and hence with lesser environmental and economic impact relative to the dense models.

A decoder-only transformer is used for the OpenAI GPT family, including GPT-3 (175B) and GPT-4 (1.7T estimated with MoE), promoting few-shot and zero-shot learning, with GPT-4 beyond into multimodality. Closed-source but highly versatile to handle creative writing and coding/reasoning-type tasks, though very expensive. Google Gemini (Nano, Pro, Ultra) is a multimodal transformer with native multimodality, intending to handle complex multimodal reasoning, agentic tasks, and scientific discovery. Deep Research-type advanced agentic features set it apart, although it is closed source. Meta's LLaMA models are in the range of 7B to 405B and are open-weight, widely used in research, fine-tuning, or independent deployment-large-scale and hard-to-get LLMs being put within the reach of many. Anthropic Claude, with its training using Constitutional AI, describes a product focused on safety and ethics and fit for high-context, safety-critical uses. Lastly, Mistral AI (7B–141B, MoE) gives efficient open-weight models aligning well with the EU regulatory framework with a great value-for-money proposition.

5. Critical Challenges and Engineering Considerations

5.1 Data quality, provenance, and societal bias

Large language models trained on massive web-scale data inherit societal biases, inaccuracies, and toxic content that differ depending upon the application and require a fairly complicated kind of mitigation. Recent approaches aim to mitigate bias by fine pruning of neurons or attention heads responsible for the harmful behavior while keeping most of the model useful. It would be unpractical to impose liability on developers; hence, large companies developing and deploying AI systems step in to be in charge of managing bias, ensuring transparency, and abiding by anti-discrimination legislation.

5.2 Hallucinations and factuality

The so-called hallucinations happen when false statements are confidently generated by LLMs-prediction of the statistically likely token than retrieval of facts. RAG attempts to reduce the falsehood by applying external knowledge to the outputs. Other methods include generating options, seeking attribution, employing a separate model for verification, or developing confidence scores which can be used to flag potentially questionable output so that they can be examined by humans.

A decoder-only transformer is used for the OpenAI GPT family, including GPT-3 (175B) and GPT-4 (1.7T estimated with MoE), promoting few-shot and zero-shot learning, with GPT-4 beyond into multimodality.

Closed-source but highly versatile to handle creative writing and coding/reasoning-type tasks, though very expensive. Google Gemini (Nano, Pro, Ultra) is a multimodal transformer with native multimodality, intending to handle complex multimodal reasoning, agentic tasks, and scientific discovery. Deep Research-type advanced agentic features set it apart, although it is closed source. Meta's LLaMA models are in the range of 7B to 405B and are open-weight, widely used in research, fine-tuning, or independent deployment-large-scale and hard-to-get LLMs being put within the reach of many. Anthropic Claude, with its training using Constitutional AI, describes a product focused on safety and ethics and fit for high-context, safety-critical uses. Lastly, Mistral AI (7B–141B, MoE) gives efficient open-weight models aligning well with the EU regulatory framework with a great value-for-money proposition.

5.3 Scalability, compute, and environmental cost

Massive computation is required for training LLMs, making them highly environmentally unfriendly. Consider that the training operation of GPT-3 generated CO_2 emissions equivalent to that of five cars in an average lifetime, while conducting inference could be an annual expense 25 to 1400 times more. Water co-usage represents another environmental concern, with 700,000 liters being used on the training of ChatGPT. In turn, this justifies improvement in inference efficiency backed by MoE architectures, green energy, cooling techniques, hardware-software co-design, and the like, toward minimizing the environmental burden with a computational one.

There are such great environmental footprints from training and inference of large language models. The carbon footprint is inordinate: the training period of BERT was estimated to be around 284 tons of CO_2 equivalent, and GPT-3 produced carbon emissions equatable to the lifetime emissions of five average cars. Training is more costly; inference is even more so in many cases: applications like ChatGPT generate up to 25 times the training cost in a year. Water consumption is a crucial factor; fifty training runs of GPT-scale models (e.g., ChatGPT) are estimated at consuming 700,000 liters, which a household would consume over five years; inference, on the other hand, uses approximately 500 ml per 20 to 50 user queries. These costs become massive with millions of requests each day. To lessen such effects, mitigating approaches are put forward: compute-optimal training-and-inference paradigm (e.g., Chinchilla), renewable-powered data centers; architectural efficiency, improvement through inference to reduce per-query computation (e.g., via MoE, quantization, and distillation).

6. Emerging Directions and Promising Research Areas

Based on ChatGPT's output, some of the main trends shaping the future of the LLM landscape are:

- **Multimodal Models**: When the system integrates text, image, audio, and video modalities, it offers a much richer understanding and generation capability in cross-domain areas of robotics, medical imaging, and creative tools (e.g., GPT-4, Gemini).
- **Retrieval/Tool-Augmented Agents**: The model executes multi-stage tasks by combining generation and external logic, such as in Google's Deep Research, thereby enhancing reliability in complex decision-making.
- **Efficient Models and Deployment**: Distillation, MoE, and quantization bring

about a decrease in the cost of inference, easing the process of deployment on the edge and low-cost cloud.

Responsible AI and Regulation: Guidelines, documentation, and regulatory frameworks (e.g., EU AI Act) foster the usage of LLMs in high-stake applications that are safe, ethical, and transparent.

7. Applications & Societal Impact

A shift is observed due to LLMs in sectors such as productivity, creativity, education, and health, wherein they assist in writing, coding, designing, tutoring, and ultimately patient summarization. Problems posed by LLMs are: cases of producing reasonable false content leading to misinformation; how automation alters the nature and demand for labor so that detection tools would need to be built, a responsible-use policy crafted, and retraining pursued strategically.

8. Conclusion

LLMs stand as greatest milestones in AI, moving from foundational breakthroughs such as Transformers toward scaling, training, and alignment pipelines. Despite being powerful, problems remain in factuality, biases, environmental impact, and privacy. Future emphasis is placed on efficiency, multimodality, and verifiable behavior, with environmental and ethical concerns giving rise to new architectures and policies. Responsible adoption therefore requires urgent interdisciplinary collaboration between ML research, systems engineering, ethics, and public policies.

Works cited

Here's a condensed list of 15 references covering your content:

[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] Attention Is All You Need - NIPS, https://papers.neurips.cc/paper/7181-attention-is-all-you-need.pdf

[5] BERT: Pre-training of Deep Bidirectional Transformers, https://blog.paperspace.com/bert-pre-training-of-deep-bidirectional-transformers-for-language-understanding/

[6] What Is Reinforcement Learning From Human Feedback (RLHF)? | IBM, https://www.ibm.com/think/topics/rlhf

[7] What is Retrieval-Augmented Generation (RAG)? | Google Cloud, https://cloud.google.com/use-cases/retrieval-augmented-generation

[8] What are AI hallucinations & how to mitigate them | Medium, https://medium.com/low-code-for-advanced-

data-science/what-are-ai-hallucinations-how-to-mitigate-them-in-llms-4abf0cd54a7a

[9] Comparing GPT, Claude, Llama, and Mistral | Epista, https://www.epista.com/knowledge/break-down-four-of-the-biggest-players-in-ai-gpt-claude-llama-and-mistral

[10] Gemini Deep Research — your personal research assistant, https://gemini.google/overview/deep-research/

[11] The Llama 3 Herd of Models | Meta AI, https://ai.meta.com/research/publications/the-llama-3-herd-of-models/

[12] Mixture of Experts (MoE) Architecture in Modern Machine Learning | Medium, https://medium.com/@prabhuss73/mixture-of-experts-moe-architecture-in-modern-machine-learning-1-introduction-b9f5930d860f

[13] 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

[14] Bias and Fairness in Large Language Models: A Survey | MIT Press, https://direct.mit.edu/coli/article/50/3/1097/121961/Bias-and-Fairness-in-Large-Language-Models-A

[15] 18 Artificial Intelligence LLM Trends in 2025 | AI Bistrot, https://medium.com/data-bistrot/15-artificial-intelligence-llm-trends-in-2024-618a058c9fdf

