

Large Language Models for Natural Language Processing: Opportunities and Limitations

Asjad Ali (Corresponding Author)

Department of Computer Science, Virtual University of Pakistan

Email: asjad.ds7@gmail.com

Taib Ali

Department of Computer Science, University: University of South Asia

Email: taibali303@gmail.com

Muhammad Hadeed

Recent Graduate in Electronic Engineering, Mehran University of Engineering and

Technology Email: muhammad.hadeed2002@gmail.com

Ghulam Yasin

Department of Computer Science, University of Makran Panjgur

Email: yasinarman38@gmail.com

ABSTRACT

Large Language Models (LLMs) have fundamentally transformed natural language processing (NLP) by demonstrating emergent capabilities in reasoning, generation, multimodal synthesis, and agentic behavior through massive scaling, advanced architectures such as Mixture-of-Experts (MoE), and novel training paradigms including reinforcement learning for reasoning (e.g., DeepSeek-R1 and OpenAI o-series). This paper traces the historical evolution of computational linguistics from rule-based systems to Transformer-based foundational models and examines 2026-era advancements, including long-context processing (up to 20 million tokens), reasoning-first deliberation loops, neuro-symbolic integration, and domain-specific agentic applications in software engineering, healthcare, biomedicine, and jurisprudence. While LLMs deliver substantial operational opportunities such as 4× faster bug resolution, 50–80% reduction in legal research time, and accelerated drug discovery they continue to grapple with persistent limitations, including hallucinations (still ~6% in frontier models), adversarial vulnerabilities, sociocultural biases, position bias in long contexts, and environmental sustainability concerns amid exascale infrastructure projects like Stargate and the Genesis Mission. The analysis contrasts long-context models with Retrieval-Augmented Generation (RAG) approaches and highlights promising directions in neuro-symbolic AI to bridge reasoning gaps toward Artificial General Intelligence (AGI). Ultimately, this survey underscores the need for balanced integration of scaling, efficiency, safety, and ethical alignment to realize LLMs' full potential while mitigating risks in high-stakes real-

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

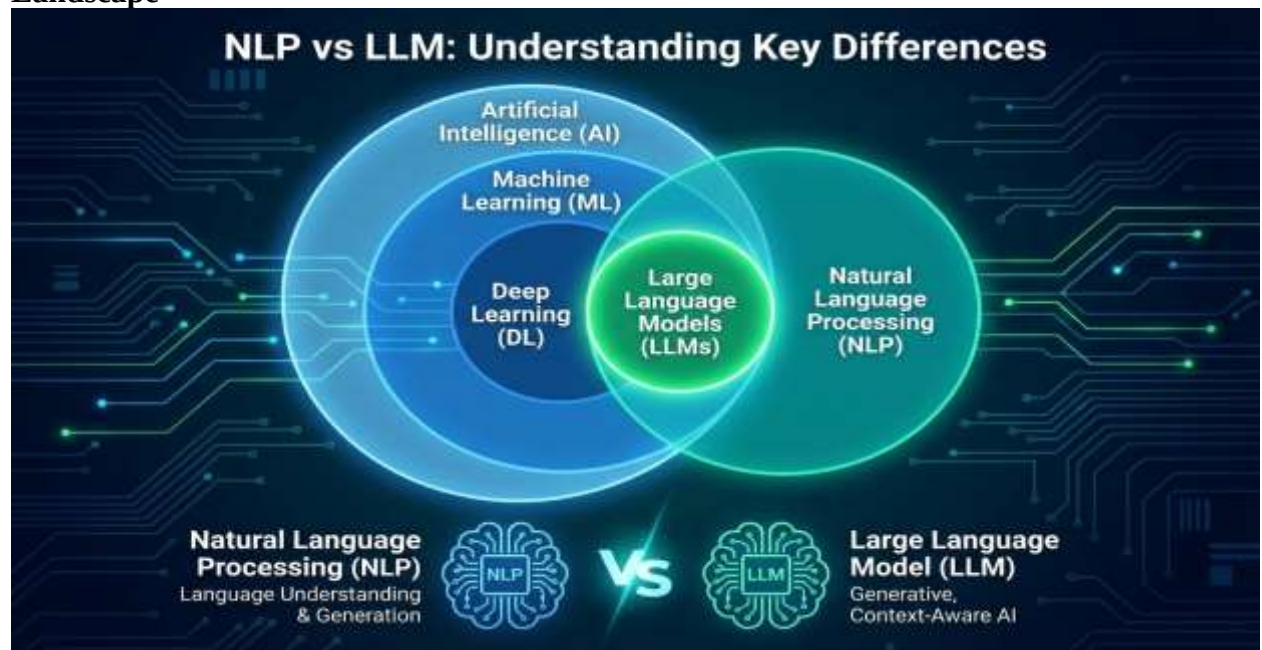
world deployment.

Keywords: Large Language Models (LLMs), Natural Language Processing (NLP), Transformer architecture, Mixture-of-Experts (MoE), Reasoning-first models, Chain-of-Thought (CoT), Agentic AI, Retrieval-Augmented Generation (RAG), Neuro-symbolic integration, Hallucinations, Bias and fairness, Long-context modeling, Artificial General Intelligence (AGI), Domain-specific applications, Environmental sustainability of AI

Introduction

The landscape of natural language processing (NLP) has undergone a series of fundamental paradigm shifts, culminating in the contemporary era of Large Language Models (LLMs) that exhibit emergent capabilities in reasoning, multimodal synthesis, and autonomous agency (Mundlamuri et al., 2025). As of early 2026, the field has transcended the initial excitement of generative text to confront the complex realities of integrating these systems into the structural fabric of global industry, science, and governance (LifeArchitect.ai, 2026). This report provides an exhaustive examination of the evolution, architectural diversification, and specialized applications of LLMs, alongside a rigorous analysis of the persistent technical and ethical limitations that define the current research frontier (Roberts, 2026).

Figure 1: The AI Hierarchy: Locating LLMs within the Deep Learning Landscape



The Historiography of Computational Linguistics

The progression of NLP reflects a move from manually crafted symbolic rules to statistical modeling, and ultimately to large-scale self-supervised learning on vast corpora. The earliest systems, predominant from the 1950s through the 1980s, were rule-based architectures such as ELIZA and SHRDLU, which relied on hardcoded

grammars and pattern matching (Mundlamuri et al., 2025). These models simulated dialogue by transforming user inputs based on predefined templates but possessed no genuine comprehension or capacity for generalization beyond their limited domains (Valgimigli, 2024).

The 1990s marked the transition to statistical language models (SLMs), where language was modeled probabilistically (MDPI, 2025a). Techniques such as n-gram models, Hidden Markov Models (HMMs), and Conditional Random Fields (CRFs) enabled more robust, data-driven approaches (Deekshith, 2024). For example, HMMs utilized probabilities to identify word roles, such as nouns or verbs, which significantly improved speech recognition and part of speech tagging (Smarts et al., 2024). However, these models were constrained by narrow context windows and an inherent inability to capture long-range semantic dependencies (Naseem et al., 2021). The advent of sequential neural language models in the 2010s addressed some of these limitations. The introduction of Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs), allowed for better context modeling (Janghorbani, 2023). Simultaneously, embedding-based models like Word2Vec and GloVe introduced distributed word representations that captured semantic similarities in high-dimensional vector spaces (HatchWorks, 2025). Despite these breakthroughs, early neural models struggled with vanishing gradients during the processing of long sequences and were difficult to parallelize (Chen et al., 2024).

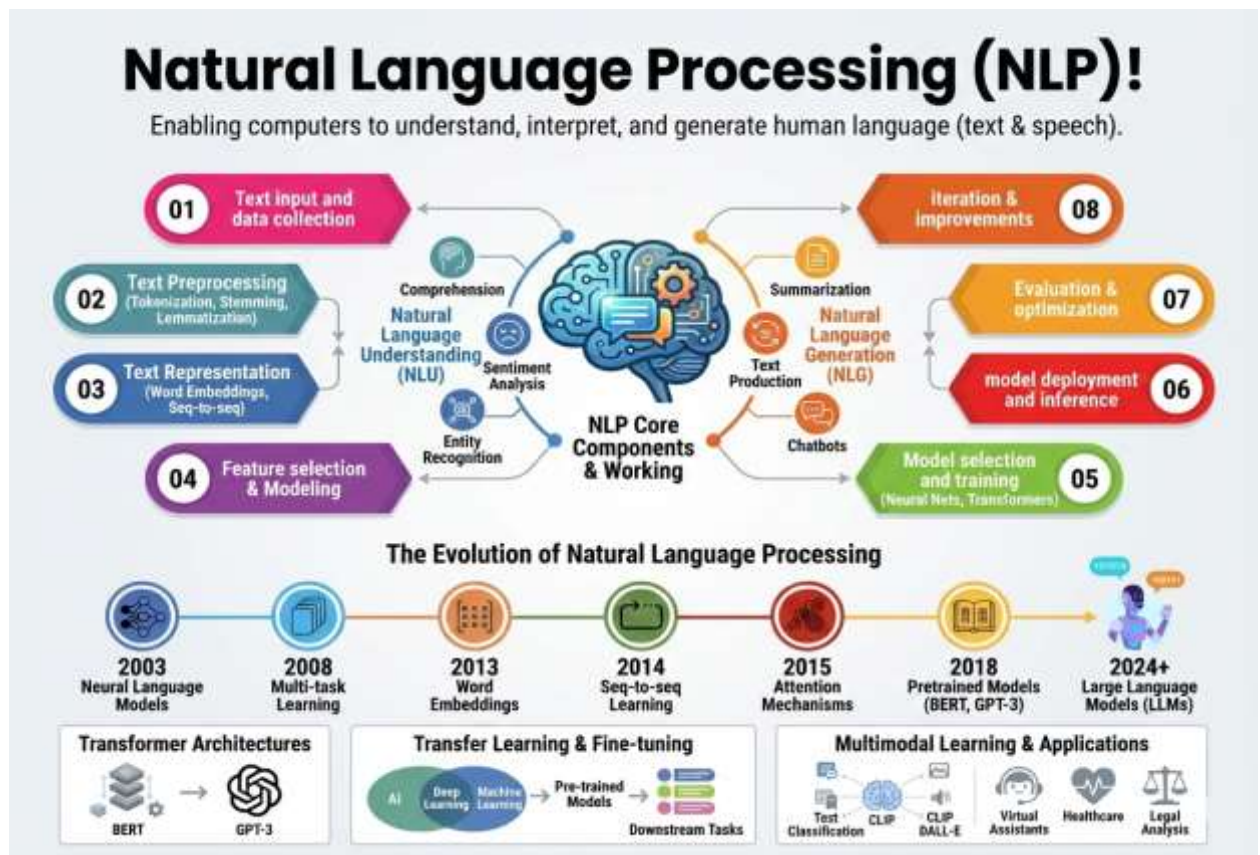
Figure 2: Evolutionary Timeline and Core Functional Components of NLP

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

The current era began in 2017 with the introduction of the Transformer architecture and its self-attention mechanism, which revolutionized the field by enabling the efficient, parallelizable modeling of long-range dependencies (Hassanpour et al., 2024). This architectural shift facilitated the training of foundational models such as BERT, GPT, T5, and later, the Llama and DeepSeek families, which pushed the boundaries of natural language understanding (NLU), generation (NLG), and reasoning (NLR) (Ramdani, 2025).

Table 1: Historiography of Computational Linguistics Paradigms

Era	Paradigm	Primary Mechanism	Representative Systems
1950s–1980s	Rule-Based	Handcrafted rules & templates	ELIZA, SHRDLU (MDPI, 2025a; HatchWorks, 2025)
1990s–2000s	Statistical	N-grams, HMMs, CRFs	Early IBM Translation, PageRank (MDPI, 2025a; Cabello, 2025)
2010s–2017	Sequential Neural	RNNs, LSTMs, Word2Vec	Seq2seq, early Google Translate (MDPI, 2025a; HatchWorks, 2025)
2017–Present	Transformer-Based	Self-attention, Scale, RLHF	GPT-4, Llama 3, DeepSeek-R1 (MDPI, 2025a; SiliconFlow, 2026)

Architectural Taxonomy and Engineering Advancements

Modern LLMs are predominantly built on the Transformer architecture, which effectively captures long-range dependencies through the self-attention mechanism (SiliconFlow, 2026). These models are typically categorized into three main groups based on their configuration: encoder-only models like BERT, which excel at understanding and classification; decoder-only models like the GPT family, which are optimized for generative tasks; and encoder-decoder models like T5 and BART, which are frequently utilized for sequence-to-sequence tasks such as translation and summarization.

By 2026, the industry has pivoted toward Mixture-of-Experts (MoE) architectures to manage the "curse of scale". MoE models, such as DeepSeek V3 and Qwen 3, employ sparse activation where only a subset of parameters is utilized for any given token (Huang et al., 2023). This allows for trillion-parameter capacity without the prohibitive computational costs associated with dense models. For instance, Qwen3-Next-80B activates only 3 billion parameters via 512 experts, achieving high-level reasoning with significantly lower per-token compute requirements (Raiaan et al., 2024).

Training Paradigms and Thinking Modes

The capabilities of LLMs are refined through a multi-stage training process. Pre-training on massive datasets sometimes including nearly the entire web establishes the model's foundational linguistic properties (Han et al., 2024). This is followed by task-specific fine-tuning and instruction tuning, which enhance the model's ability to follow natural language commands. Seminal efforts like FLAN and T0 have demonstrated that instruction tuning significantly improves zero-shot performance across diverse tasks, making models more usable and steerable (Bhattacharya et al., 2024).

A critical development in early 2026 is the emergence of reasoning-first LLMs that utilize internal deliberation loops to improve correctness. These models, exemplified by OpenAI's o1/o3 series and DeepSeek-R1, generate an internal chain-of-thought (CoT) before outputting final answers (Zapier, 2025). This "thinking mode" allows models to tackle highly complex academic problems, mathematical proofs, and coding challenges by breaking them into manageable steps and verifying the logic at each interval (Shakudo, 2026).

Figure 3: Comparative Analysis: Supervised NLP vs. Self-Supervised LLM Paradigms



DeepSeek-R1-Zero, for instance, demonstrated that reasoning capabilities could be incentivized through pure reinforcement learning (RL) without human-labeled reasoning trajectories (Kumar et al., 2025). This training pipeline allows the model to explore novel reasoning patterns that may be missed by human-defined heuristics, although it occasionally struggles with readability and language mixing. The subsequent DeepSeek-R1 model addressed these issues by aligning model behavior with human preferences through additional non-reasoning data (DeepSeek-AI, 2025).

Operational Opportunities and Domain-Specific Integration

The opportunities presented by LLMs in 2026 extend beyond general conversation to the specialized automation of high-value professional sectors. Organizations are moving away from general-purpose models toward domain-specific LLMs fine-tuned with proprietary data to improve accuracy, compliance, and efficiency (Prajna AI Wisdom, 2025).

The Agentic Revolution in Software Engineering

The role of LLMs in software engineering has shifted from static assistants to autonomous agents capable of managing end-to-end workflows. These agents plan tasks, run tests, and manage continuous integration/continuous delivery (CI/CD) pipelines with minimal oversight (Osmani, 2025). Telemetry from early 2026 indicates that pairing engineers with agentic debugging assistants has led to a four-fold acceleration in resolving bugs. Unlike traditional assistants, agentic models pursue a hypothesis iteratively until the bug is resolved (DigitalDefynd, 2026).

The AgentBench evaluation suite in 2025 highlighted that purpose-built agentic systems outperform baseline LLMs by up to 12 times in completing complex multi-step tasks (Sun et al., 2025). Tools like OpenDevin and Cursor now explicitly support chunked workflows, where an agent implements one function or fixes one bug at a time, preventing the "drift" often associated with monolithic task prompts (Lendman,

2025).

Table 2: Domain-Specific LLM Integration and Benchmarks

Domain	Application	Primary Benefit	Key Benchmark
Software Engineering	Agentic Debugging	4x faster bug resolution	SWE-bench Verified (DigitalDefynd, 2026; Digital Applied, 2025)
Healthcare	Clinical Decision Support	Reduced task load & dropout rates	PubMed/USMLE (JMIR, 2025; PMC, 2025b)
Law	Contract Analysis	50-80% reduction in research time	JUDGEBERT (Lendman, 2025; Beauchemin & Khoury, 2025)
Science	Drug Discovery	Accelerated de novo design	SMILES accuracy (PMC, 2025b; World Economic Forum, 2025)

Healthcare and Biomedical Discovery

In healthcare, LLMs are being integrated into 5-stage clinical workflows, covering everything from patient registration and pre-diagnosis to treatment planning and discharge summaries (Li et al., 2025). Research published in 2025 demonstrates that GPT-4, when used as a therapy support tool, significantly reduced dropout rates and improved clinical outcomes in group interventions. However, widespread adoption remains constrained by regulatory barriers, such as classification as software-as-a-medical-device (SaMD), which requires lengthy approval processes (Evangelista et al., 2026).

In the biomedical realm, LLMs like DrugLLM and MolGPT are utilized for structure-based drug molecule design. These models create novel molecules with required scaffolds by generating SMILES (Simplified Molecular Input Line Entry System) strings from molecular graphs (Holmes, 2025). This capability facilitates high-quality and rapid drug design, particularly for rare medical conditions where traditional pharmaceutical investment is limited (World Economic Forum, 2025).

Jurisprudence and the Legal Domain

LLMs have transformed legal NLP applications by ensuring accuracy and accessibility for text simplification. Specialized evaluation metrics like JUDGEBERT have been introduced to assess legal meaning preservation in French legal text simplification, outperforming existing metrics by demonstrating a superior correlation with human judgment (Siino et al., 2025). Law firms utilize specialized agents for contract analysis, case research, and precedent identification, which has been documented to reduce research time by 50-80% (Lendman, 2025).

Technical Limitations and the Persistence of Risk

Despite their transformative potential, LLMs in 2026 face significant technical, social, and ethical challenges. These models remain "black box" systems, particularly proprietary ones like GPT-5 and Gemini 3, where training data and internal

implementations are kept confidential. This lack of transparency hinders reproducibility and raises concerns among researchers about the validity of findings derived from these models (Ariai et al., 2025). The Mystery-Zebra benchmark further underscores these limitations, demonstrating that LLM reasoning is often restricted to a few consecutive steps and remains heavily dependent on lexical clues present in the training data (Resimić et al., 2025).

Hallucinations and Factual Accuracy

The tendency of LLMs to "hallucinate" producing factually incorrect or nonsensical outputs remains a major bottleneck. While newer models like GPT-5.2 have reduced hallucination rates by approximately 40% compared to previous generations, the rate still stands at a non-negligible 6.2% (Shakudo, 2026). This is particularly dangerous in high-stakes environments like medicine or finance, where incorrect information can lead to catastrophic outcomes (Banerjee et al., 2025). Retrieval-Augmented Generation (RAG) has emerged as an effective mitigation strategy, grounding model outputs in externally retrieved, factual information to improve accuracy (Dataiku, 2025).

Vulnerabilities to Adversarial and Natural Distribution Shifts

Safety mechanisms in aligned LLMs can be bypassed through sophisticated adversarial attacks or "natural distribution shifts". The ActorBreaker method, introduced in late 2025, identifies actors related to toxic prompts within the pre-training distribution to craft multi-turn queries that gradually guide a model to reveal unsafe content (Ren et al., 2025). This highlights a critical failure mode: models may be robust to explicit toxic prompts but remain brittle to benign queries that are semantically related to harmful content (Wei et al., 2025). Experimental results indicate that these prompts can bypass detection by systems like Llama-guard 2, demonstrating the need for safety training to encompass broader semantic relationships (Wang et al., 2025).

Sociocultural and Gender Bias

LLMs continue to exhibit striking disparities in their creative and instructive outputs (Beauchemin & Khoury, 2025). In stories generated for children, protagonists described as girls receive 55.26% more appearance-related attributes than boys (Rooein et al., 2025). Furthermore, narratives for non-Western children disproportionately emphasize tradition and heritage compared to Western counterparts, raising concerns about the equity of AI in educational settings. These biases are often rooted in the pre-training data and are difficult to eliminate entirely through current alignment techniques (Appen, 2025).

The Long-Context vs. RAG Performance Spectrum

The expansion of context windows reaching up to 20 million tokens in Gemini 3 has sparked a debate on the obsolescence of RAG. Long-context models are ideal for tasks where understanding the entirety of a document archive is beneficial, such as in holistic summarization or maintaining context across extended multi-stage tasks (Ali, 2025). However, long-context models suffer from "position bias," performing better

when relevant insights are at the beginning or end of a document rather than the middle (Phan et al., 2024).

RAG remains the preferred approach for scenarios where precise retrieval of specific details and high citation accuracy are crucial. RAG reduces computational load, cost, and latency by only including the most relevant sections of a document in the prompt (Dataiku, 2025). Furthermore, research indicates that thermal throttling on mobile hardware remains a major bottleneck for long-context models, as the intense compute demands can cause device temperatures to rise, triggering performance degradation or app crashes (Google AI Developers, 2025).

Table 3: Performance Comparison: Long-Context LLMs vs. RAG-Based Systems

Feature	Long-Context LLMs	RAG-Based Systems
Max Context Size	20M tokens (Gemini 3)	Theoretically unlimited via external DB
Latency	Higher (linearly with context)	Lower (targeted retrieval)
Citation Accuracy	Variable (Position bias)	High (Direct attribution)
Best For	Repository analysis, context retention	Precise fact retrieval, cost control (Digital Applied, 2025; Dataiku, 2025)

Foundational Research: Neuro-Symbolic Integration

To overcome the reasoning bottlenecks of pure neural architectures, researchers have turned to neuro-symbolic AI (NeSy), which integrates the learning capabilities of neural networks with the formal reasoning of symbolic logic (Yang et al., 2025). This integration aligns with the "Dual Process Theory" in cognitive science, where System 1 is fast and intuitive (neural) and System 2 is slow and deliberate (symbolic) (IJCAI, 2025).

NeSy methods improve reasoning across three primary paradigms:

Symbolic to LLM: Rigorous symbolic methods generate high-quality data to "distill" logic into the model (Bhuyan et al., 2024).

LLM to Symbolic: The LLM acts as a controller that translates queries into formal languages like Python or PDDL for execution by external solvers (Beurer-Kellner, 2025).

LLM + Symbolic: Hybrid architectures where symbolic knowledge is used as a regularization constraint during training to ensure model predictions align with domain-specific knowledge (Acharya et al., 2024).

The ENVISIONS framework, introduced in 2025, represents a significant advance in neural-symbolic self-training. By leveraging the interactive nature of an embodied environment, LLMs generate trajectories and learn symbolic language processing without the need for human annotations or a stronger teacher model (Xu et al., 2025).

In evaluations, ENVISIONS consistently converted existing LLMs into stronger variants, boosting performance by an average of 30% for LLaMA2-Chat models (Škrlj et al., 2025).

Formalizing Reasoning in the 2026 Landscape

Reasoning in NeSy frameworks is defined as a recursive process where an answer A is derived from a problem Q and background knowledge K through a sequence of intermediate steps $Z = \{z_1, z_2, \dots, z_n\}$ (Ozarde et al., 2026). Researchers distinguish between three primary reasoning types:

Deductive Reasoning: Deriving steps by applying formal logical rules from K .

Inductive Reasoning: Generalizing broader rules from specific examples.

Abductive Reasoning: Generating the most plausible hypothesis to explain an observed phenomenon (sun et al., 2026).

The goal of this research is to achieve Artificial General Intelligence (AGI) by aligning LLMs with these rigorous logical structures, allowing them to handle ambiguity without losing consistency (Deckker et al., 2026).

Toward AGI: The Infrastructure of Scale and the Stargate Project

As of 2026, the pursuit of AGI has transitioned from a theoretical objective to a massive industrial and geopolitical mission. OpenAI's five-level framework classifies progress toward AGI, with current models reaching Level 2 (Reasoners) and Level 3 (Agents) (Negele et al., 2025). Achieving Level 5 would mark the realization of AGI, where systems can perform the work of entire organizations (Quinteft, 2025).

A critical component of this trajectory is the "Stargate" project, an AI data center initiative utilizing over 2 million Nvidia GB200 chips with a capacity exceeding 5 gigawatts. This infrastructure is designed to train GPT-7 and beyond, models expected to have trillion-plus parameters and datasets reaching quadrillions of tokens (LifeArchitect.ai, 2026).

The Genesis Mission and Nature's Language

Parallel to hardware scaling, the "Genesis Mission" represents a coordinated national effort by the U.S. Department of Energy to harness the world's largest collection of federal scientific data. This dataset, estimated at 3.6 exabytes, includes materials imaging, nuclear physics collision data, and viral genomics (Smith et al., 2022). The objective is to allow models to learn the "rules of nature" directly from raw sensor data, discovering near-infinite energy sources and new materials to replace steel and concrete (Tithi et al., 2025). Moreover, the integration of AI frameworks is being explored for large-scale social benefits, such as enhancing climate resilience in urban infrastructure (Mahadeva & Sindhushree, 2025).

Table 4: Scientific Data Repositories of the Genesis Mission

Dataset Component	Size	Scientific Purpose
-------------------	------	--------------------

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

APS-U X-Ray Data	500PB	Materials imaging and discovery loops (LifeArchitect.ai, 2026)
Stockpile Stewardship	1,000PB	Nuclear physics and classified archives (LifeArchitect.ai, 2026)
JGI Genomics	80PB	DNA sequence as "nature's language" (LifeArchitect.ai, 2026)
ESGF Climate Data	40PB	Climate AI twins and modeling (LifeArchitect.ai, 2026)

Expert Critiques and the AGI Timeline Debate

Despite the massive investments, the timeline for AGI remains controversial among leading experts. Sam Altman and Aleksander Madry suggest AGI may be achieved by the end of 2026 or shortly thereafter, arguing that the necessary scientific breakthroughs are complete and only scaling remains (Altman, 2024). Conversely, Yann LeCun and Gary Marcus argue that AGI remains decades away, as current architectures lack consciousness, common sense, and the ability to interact adaptively with the physical world (Madry, 2026).

Marcus emphasizes that we lack adequate safety measures for superintelligent systems and proposes a global treaty to verify compliance with bans on recursive self-improvement (ZombieCodeKill, 2025). Meanwhile, Apple researchers published findings in June 2025 casting doubt on whether current model reasoning abilities can scale to truly "generalizable" reasoning, suggesting that models may only be replicating patterns found in training data rather than performing genuine logical synthesis (IBM Think, 2025).

Economic and Environmental Sustainability

The environmental impact of LLMs has become a primary constraint on deployment (Roberts, 2026). Goldman Sachs predicts data center power demand will rise by 160% by 2030, necessitating a shift toward "Green AI". This has spurred research into smaller, more efficient models like Microsoft's Phi-4 and Google's Gemma 3, which deliver high performance at a fraction of the energy cost (Prajna AI Wisdom, 2025).

Efficiency is also driven by cost optimization strategies, such as prompt caching which can reduce repetitive workflow costs by 90% and model routing. Organizations achieving the best ROI implement "routing" where simple tasks like classification are handled by DeepSeek or small open-weight models, while complex reasoning is reserved for frontier models like GPT-5.2 or Claude Opus 4.5 (Digital Applied, 2025).

Conclusion

As of early 2026, Large Language Models stand at the forefront of a profound shift in natural language processing and broader artificial intelligence, having evolved from statistical and sequential precursors into scalable, reasoning-capable systems that exhibit emergent agency, long-range comprehension, and domain-specialized automation. Breakthroughs in Mixture-of-Experts architectures, reinforcement-learned reasoning pathways, extended context windows, and neuro-symbolic hybridization have enabled unprecedented performance in complex tasks ranging

from autonomous software engineering agents and clinical decision support to structure-based drug design and legal contract analysis.

Nevertheless, the field confronts enduring technical and societal hurdles. Hallucinations, though reduced, remain non-negligible in high-stakes domains; adversarial robustness is fragile against semantically subtle attacks; sociocultural and gender biases persist in generated outputs; and the environmental footprint of frontier-scale training exemplified by multi-gigawatt data centers demands urgent progress toward greener, more efficient models. The debate over long-context versus RAG paradigms, coupled with critiques of scaling-alone paths to genuine generalization and common sense, highlights that current architectures may excel at sophisticated pattern replication rather than true understanding or adaptive world modeling.

Looking forward, the pursuit of AGI through initiatives like the Stargate project and Genesis Mission signals an era of massive industrial and geopolitical investment, yet expert opinions remain divided on feasible timelines, with some forecasting arrival by late 2026–2027 and others emphasizing fundamental architectural shortcomings. Sustainable advancement will require prioritizing neuro-symbolic reasoning, robust safety alignment, bias mitigation, energy-efficient inference (via prompt caching, model routing, and smaller specialized models), and interdisciplinary governance frameworks to ensure equitable benefits across global societies. In summary, while LLMs have transitioned from experimental curiosities to indispensable tools reshaping industry, science, and governance, their responsible maturation depends on addressing transparency deficits, ethical risks, and resource constraints. Only through concerted efforts to combine scale with symbolic rigor, verifiable reasoning, and societal safeguards can the field move beyond impressive imitation toward trustworthy, generalizable intelligence that truly augments human capability.

References

- Ahrabian, K., Jandaghi, P., Mokhberian, N., Karimireddy, S. P., & Pujara, J. (2025). A systematic analysis of base model choice for reward modeling. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 146–164. <https://doi.org/10.48550/arXiv.2505.10775>
- Altman, S. (2024). The path to AGI: Scaling and scientific breakthroughs. OpenAI.
- Appen. (2025). ACL 2025 trends: Fairness, bias, and multimodal capabilities. <https://www.appen.com/blog/acl-2025>
- Beauchemin, D., & Khoury, R. (2025). QFrCoLA: A Quebec-French corpus of linguistic acceptability judgments. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.
- Beyer, H., & Reed, C. (2025). Lexical recall or logical reasoning: Probing the limits of reasoning abilities in large language models. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13532–13557. <https://doi.org/10.18653/v1/2025.acl-long.664>
- Cabello, A. (2025, May). The evolution of language models: A journey through time. <https://medium.com/@adria.cabello/the-evolution-of-language-models-a-journey-through-time-3179f72ae7eb>
- Dataiku. (2025, December). Is RAG obsolete? Evaluating long-context models vs. retrieval-augmented generation. <https://www.dataiku.com/stories/blog/is-rag->

obsolete

- DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing reasoning capabilities through reinforcement learning. <https://arxiv.org/abs/2501.12948>
- Digital Applied. (2025, December 19). LLM comparison guide: Performance, pricing, and specialization. <https://www.digitalapplied.com/blog/llm-comparison-guide-december-2025>
- DigitalDefynd. (2026). Top 100 agentic AI facts & statistics. <https://digitaldefynd.com/IQ/agentic-ai-statistics/>
- Google AI Developers. (2025, December). Gemini 3 significantly worse than 2.5 Pro at long context. Google AI Developers Forum. <https://discuss.ai.google.dev/t/gemini-3-significantly-worse-thant-2-5-pro-at-long-context-temperature-likely-to-blame/110888>
- HatchWorks. (2025). Large language models guide: Evolution, architecture, and milestones. <https://hatchworks.com/blog/gen-ai/large-language-models-guide/>
- IBM Think. (2025, June). Reasoning LLMs and generalizable reasoning: Insights from Apple research. <https://www.ibm.com/think/topics/reasoning-model>
- JMIR. (2025). Selecting suitable LLM architectures for clinical tasks: A systematic review. *Journal of Medical Internet Research*, 23(1), e70315. <https://www.jmir.org/2025/1/e70315>
- Lendman, T. (2025). Agentic AI workflows: Transformative case studies revealed. <https://troylendman.com/2025-agentic-ai-workflows-transformative-case-studies-revealed/>
- LifeArchitect.ai. (2026, January). AI timeline: 2025-2026 milestones and breakthroughs. <https://lifearchitect.ai/timeline/>
- Madry, A. (2026). The transition from ANI to AGI: Engineering and scaling. OpenAI.
- Mahadeva, M., & Sindhushree. (2025). Artificial intelligence-driven framework for enhancing climate resilience in urban infrastructure. *MAT Journals*, 10(3).
- MDPI. (2025a). Large language models for natural language processing: A leap in capabilities and ethical concerns. *Applied Sciences*, 16(1), 382. <https://www.mdpi.com/2076-3417/16/1/382>
- MDPI. (2025b). Challenges and issues surrounding large language models in real-world applications. *Applied Sciences*, 15(14), 8103. <https://www.mdpi.com/2076-3417/15/14/8103>
- Osmani, A. (2025). My LLM coding workflow going into 2026. <https://medium.com/@addyosmani/my-llm-coding-workflow-going-into-2026-52fe1681325e>
- PMC. (2025a). Ethical aspects of deploying large language models in healthcare: A systematic review. *PubMed Central*. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12460403/>
- PMC. (2025b). Large language models in nephrology and structure-based drug design. *PubMed Central*. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12461145/>
- Prajna AI Wisdom. (2025, April). LLM trends 2025: A deep dive into the future of large language models. <https://prajnaaiwisdom.medium.com/llm-trends-2025-a-deep-dive-into-the-future-of-large-language-models-bff23aa7cdbe>
- Quinteft. (2025). Understanding OpenAI's five levels of AI progress towards AGI. <https://quinteft.com/understanding-openais-five-levels-of-ai-progress-towards->

agi/

- Ren, Q., Li, H., Liu, D., Xie, Z., Lu, X., Qiao, Y., Sha, L., Yan, J., Ma, L., & Shao, J. (2025). LLMs know their vulnerabilities: Uncover safety gaps through natural distribution shifts. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing.
- Roberts, M. (2026). 10 AI predictions for 2026: The year of assimilation. Capgemini Future Labs. <https://aibusiness.com/generative-ai/10-ai-predictions-2026>
- Rooein, D., Zouhar, V., Nozza, D., & Hovy, D. (2025). Biased tales: Cultural and topic bias in generating children's stories. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing.
- Shakudo. (2026, January 6). Top 9 large language models: A comparison of proprietary and open-source architectures. <https://www.shakudo.io/blog/top-9-large-language-models>
- SiliconFlow. (2026). The best large language models for academic research in 2026. <https://www.siliconflow.com/articles/en/best-LLMs-for-academic-research>
- World Economic Forum. (2025, December). Neurosymbolic AI: Real-world outcomes and drug discovery. <https://www.weforum.org/stories/2025/12/neurosymbolic-ai-real-world-outcomes/>
- Xu, Y., et al. (2025). ENVISIONS: ENV-guided self-training framework for neural symbolic scenarios. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025). <https://aclanthology.org/2025.acl-long.635.pdf>
- Yang, X. W., Shao, J. J., Guo, L. Z., Zhang, B. W., Zhou, Z., Jia, L. H., Dai, W. Z., & Li, Y. F. (2025). Neuro-symbolic artificial intelligence: Towards improving the reasoning abilities of large language models. Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI-25). <https://doi.org/10.48550/arXiv.2508.13678>
- Zapier. (2025, October). Best large language models in 2026. <https://zapier.com/blog/best-llm/>
- ZombieCodeKill. (2025). Explosive AI timeline predictions: AGI and global treaties. <https://medium.com/@ZombieCodeKill/explosive-ai-timeline-predictions-b29744f839a9>
- Mundlamuri, R., Gunnam, G. R., Mysari, N. K., & Pujuri, J. (2025). The Evolution of AI: From Classical Machine Learning to Modern Large Language Models. Ieee Access.
- Valgimigli, L. (2024). Novel language models and methods for semantic representation learning, self-supervised retrieval, and summarization.
- Deekshith, A. (2024). Advances in natural language processing: A survey of techniques. International Journal of Innovations in Engineering Research and Technology, 8, 74-83.
- Smarts, N., Selvaraj, R., & Kuthadi, V. (2024, December). Natural Language Processing Revolution From Humble Beginnings to Deep Learning and Beyond. In 2024 IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG) (pp. 1-13). IEEE.
- Naseem, U., Razzak, I., Khan, S. K., & Prasad, M. (2021). A comprehensive survey

- on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5), 1-35.
- Janghorbani, S. (2023). Self-Supervised Object-Centric Representations Learning of Computer Vision and Natural Language Understanding Models (Doctoral dissertation, Rutgers The State University of New Jersey, School of Graduate Studies).
- Chen, Z., Xu, L., Zheng, H., Chen, L., Tolba, A., Zhao, L., ... & Feng, H. (2024). Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models. *Computers, Materials & Continua*, 80(2).
- Hassanpour, H., & Majidi, M. (2024). From Statistical Models to LLMs: A Comprehensive Survey of Language Model Evolution. *Journal of Artificial Intelligence, Applications and Innovations*, 1(4), 55-75.
- Ramdani, F. (2025). Natural Language Processing and Sentiment Analysis. In *Data Science: Foundations and Hands-on Experience: Handling Economic, Spatial, and Multidimensional Data with R* (pp. 359-394). Singapore: Springer Nature Singapore.
- Huang, Y., Xu, J., Lai, J., Jiang, Z., Chen, T., Li, Z., ... & Zhao, P. (2023). Advancing transformer architecture in long-context large language models: A comprehensive survey. *arXiv preprint arXiv:2311.12351*.
- Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., ... & Azam, S. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access*, 12, 26839-26874.
- Han, S., Wang, M., Zhang, J., Li, D., & Duan, J. (2024). A review of large language models: Fundamental architectures, key technological evolutions, interdisciplinary technologies integration, optimization and compression techniques, applications, and challenges. *Electronics*, 13(24), 5040.
- Bhattacharya, P., Prasad, V. K., Verma, A., Gupta, D., Sapsomboon, A., Viriyasitavat, W., & Dhiman, G. (2024). Demystifying chatgpt: An in-depth survey of openai's robust large language models. *Archives of Computational Methods in Engineering*, 31(8), 4557-4600.
- Kumar, K., Ashraf, T., Thawakar, O., Anwer, R. M., Cholakkal, H., Shah, M., ... & Khan, S. (2025). Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*.
- Sun, J., Zheng, C., Xie, E., Liu, Z., Chu, R., Qiu, J., ... & Li, Z. (2025). A survey of reasoning with foundation models: Concepts, methodologies, and outlook. *ACM Computing Surveys*, 57(11), 1-43.
- Li, H., Fu, J. F., & Python, A. (2025). Implementing large language models in health care: clinician-focused review with interactive guideline. *Journal of medical Internet research*, 27, e71916.
- Evangelista, E., Ruba, F., Bukhari, S., Nazir, A., & Sharma, R. (2026). GraphRAG-Enabled Local Large Language Model for Gestational Diabetes Mellitus: Development of a Proof-of-Concept. *JMIR diabetes*, 11, e76454.
- Holmes, B. A. (2025). Patient Subset Classification Using Encoded Embeddings And Knowledge Graph Retrieval-Augmented Generation (Doctoral dissertation,

- Wright State University).
- Kocaballi, A. B., Sezgin, E., Clark, L., Carroll, J. M., Huang, Y., Huh-Yoo, J., ... & Zubatiy, T. (2022). Design and evaluation challenges of conversational agents in health care and well-being: selective review study. *Journal of medical Internet research*, 24(11), e38525.
- Siino, M., Falco, M., Croce, D., & Rosso, P. (2025). Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*.
- Ariai, F., Mackenzie, J., & Demartini, G. (2025). Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *ACM Computing Surveys*, 58(6), 1-37.
- Resimić, M., Freigang, V., Jenkins, M., Vrushni, J., & TI, T. W. (2025). Harnessing artificial intelligence (AI) for anti-corruption.
- Banerjee, S., Zhu, Y., Freeman, I., Machado, J. V., Ahmed, A., Sarker, A., & Al-Garadi, M. (2025). Agentic AI in Healthcare: A Comprehensive Survey of Foundations, Taxonomy, and Applications. *Authorea Preprints*.
- Wei, R., Niu, P., Shen, X., Tu, T., Li, Y., Wu, R., ... & Li, P. (2025). The Trojan Knowledge: Bypassing Commercial LLM Guardrails via Harmless Prompt Weaving and Adaptive Tree Search. *arXiv preprint arXiv:2512.01353*.
- Wang, J., Li, H., Peng, H., Zeng, Z., Wang, Z., Du, H., & Yu, Z. (2025). Activation-Guided Local Editing for Jailbreaking Attacks. *arXiv preprint arXiv:2508.00555*.
- Ali, O. (2025). Retrieval Augmented Generation for Intelligent Querying of Databases and Documents.
- Phan, H., Acharya, A., Chaturvedi, S., Sharma, S., Parker, M., Nally, D., ... & Horawalavithana, S. (2024). Rag vs. long context: Examining frontier large language models for environmental review document comprehension. *arXiv preprint arXiv:2407.07321*.
- Bhuyan, B. P., Ramdane-Cherif, A., Tomar, R., & Singh, T. P. (2024). Neuro-symbolic artificial intelligence: a survey. *Neural Computing and Applications*, 36(21), 12809-12844.
- Beurer-Kellner, L. (2025). Language Model Programming (Doctoral dissertation, ETH Zurich).
- Acharya, K., Velasquez, A., & Song, H. H. (2024). A survey on symbolic knowledge distillation of large language models. *IEEE Transactions on Artificial Intelligence*.
- Škrlić, B., Koloski, B., Pollak, S., & Lavrač, N. (2025). From Symbolic to Neural and Back: Exploring Knowledge Graph–Large Language Model Synergies. *Challenges and Algorithms for Knowledge Discovery from Data: Essays Dedicated to Arno Siebes on the Occasion of His 67th Birthday*, 181-197.
- Ozarde, S., & Carta, S. (2026). Cartesian Methodical Doubt: A Cognitive Framework for Reasoning and Explainability. In *HCI International 2025–Late Breaking Papers: 27th International Conference on Human-Computer Interaction, HCII 2025, Gothenburg, Sweden, June 22–27, 2025, Proceedings, Part XV* (p. 266). Springer Nature.
- Deckker, D., & Sumanasekara, S. (2026). Scaling Laws, Foundation Models, and the AI Singularity: A Critical Appraisal of 2023–2025 Evidence.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

- Sun, B., You, H., Wu, J., Wang, Q., & Du, J. (2026). Belief Update Modeling for AI Agents by Human-Derived Dual-Thresholds Evidence Accumulation Process. *Journal of Computing in Civil Engineering*, 40(1), 04025116.
- Negele, M., Juijn, D., Shamir, A., Janku, D., Özcan, B., Soder, L., ... & Andriushchenko, M. (2025). Europe and the geopolitics of AGI.
- Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., ... & Catanzaro, B. (2022). Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.
- Tithi, J. J., Wu, H., Abuhatzera, A., & Petrini, F. (2025). Scaling Intelligence: Designing Data Centers for Next-Gen Language Models. *arXiv preprint arXiv:2506.15006*.