Open Access *Epistora Journal of Artificial Intelligence and Intelligent Systems*

EPISTORA *Review Article*

The Evolution, Capabilities, Limitations and Future of Large Language Models: A Comprehensive Review

Divyansh Shukla
Independent Researcher, India

Corresponding author: Divyansh Shukla, Independent Researcher, India.

Submitted: July 05, 2026 **Accepted: July 15, 2026** **Published: August 16, 2026**

Citation: Divyansh Shukla (2026) The Evolution, Capabilities, Limitations, and Future of Large Language Models: A Comprehensive Review. Epistora J. Artif. Intell. & Intell. Syst. 1(1), 1-17. Article EJAIS-105

Abstract

Large language models (LLMs) have emerged as one of the most transformative technological developments of the twenty-first century. Originating from statistical language modelling and neural network research, these systems have grown from modest n-gram models into massively parameterised, multi-trillion-token-trained architectures capable of complex reasoning, code synthesis, scientific hypothesis generation, and open-ended conversation. This review surveys the trajectory of LLM development from foundational language-modelling research through the seminal Transformer architecture (Vaswani et al., 2017) to contemporary frontier models released in 2025–2026, including the GPT, Claude, Gemini, Llama, Qwen, and DeepSeek families.

We systematically examine the core technical innovations that have driven progress: the self-attention mechanism, Reinforcement Learning from Human Feedback (RLHF), Constitutional AI, Chain-of-Thought prompting, Mixture-of-Experts (MoE) routing, Retrieval-Augmented Generation (RAG), and emerging agentic paradigms. Benchmark comparisons across MMLU, HumanEval, GSM8K, MATH, and GPQA are presented to contextualise the relative capabilities of major model families. We further analyse domain-specific applications in education, enterprise software, and scientific research, before turning to the critical limitations that temper optimism: hallucination, factual inconsistency, social bias, adversarial fragility, privacy risks, intellectual property concerns, and the substantial environmental footprint of large-scale training.

The paper closes with an examination of ethical and societal implications, emerging governance frameworks at national and supranational levels, and a forward-looking synthesis of research directions likely to define the next generation of AI systems, including reasoning-specialised models, embodied agents, neuromorphic integration, and interpretability science. The review is intended as a comprehensive reference for researchers, practitioners, and policy-makers navigating the rapidly evolving LLM landscape in 2026.

Keywords: Large language models, Transformer architecture, Scaling laws, RLHF, Chain-of-thought, Mixture of experts, Retrieval-augmented generation, AI agents, Hallucination, AI safety, AI governance, GPT, Claude, Gemini, Llama, DeepSeek, Qwen, Multimodal AI

Introduction

The history of artificial intelligence contains many inflection points, but few compare in breadth of societal impact to the emergence of large language models capable of fluent, contextually rich natural-language generation.

Within a span of less than a decade, architectures built upon the self-attention mechanism introduced by Vaswani et al. [1] have progressed from producing grammatically plausible but semantically shallow text to passing professional examinations [2], generating functional software [3], assisting in protein structure annotation [4], and engaging in multi-step scientific reasoning [5]. As of 2026, LLMs are embedded in consumer applications used by hundreds of millions of people daily, yet the field continues to evolve at a pace that challenges systematic understanding.

This review aims to provide that systematic understanding. We trace the intellectual lineage of LLMs from Shannon's information-theoretic framing of language [6] and Elman's recurrent network language models [7] through the neural language model of Bengio et al. [8], the word embedding revolution of Mikolov et al. [9], and the sequence-to-sequence breakthrough of Sutskever et al. [10], arriving at the Transformer [1] and its derivatives. We then examine how scaling-measured in parameters, data, and compute-has unlocked qualitatively new capabilities [11, 12], and how post-training alignment techniques have shaped the behavioural profile of deployed systems [13, 14].

A particular focus is placed on the diversity of the current frontier. Unlike the earlier period dominated by a single laboratory's releases, 2024–2026 has been characterised by vigorous international competition. OpenAI's GPT-4o and GPT-4.1 [15], Anthropic's Claude 3.5 and 3.7 [16], Google DeepMind's Gemini 1.5 and 2.0 [17], Meta's Llama 3 and 4 [18], DeepSeek's V3 and R1 [19, 20], and Alibaba's Qwen2.5 and Qwen3 [21, 22] represent distinct engineering philosophies, training pipelines, and alignment approaches that collectively expand the frontier while offering researchers a rich comparative dataset.

The structure of this paper is as follows. Section 2 traces the historical evolution of language models. Section 3 details the Transformer architecture and its key technical components. Section 4 discusses scaling laws and emergent abilities. Sections 5–7 profile major model families. Section 8 surveys benchmarking methodology. Sections 9–11 examine reasoning, coding, and application domains. Sections 12–14 address multimodality and agentic systems. Section 15 catalogues known limitations. Sections 16–18 treat ethical, societal, and regulatory dimensions. Section 19 outlines future research directions, and Section 20 offers a synthesising discussion before the conclusion.

Historical Evolution of Language Models

Language modelling-the task of assigning probabilities to sequences of words-predates neural networks by several decades. Shannon [6] formalised the entropic properties of English text in 1948, and subsequent work on n-gram models (Jelinek, 1990; Kneser & Ney, 1995) provided the statistical backbone of speech recognition and machine translation systems through the 2000s. These models were powerful for their time yet fundamentally limited by their inability to generalise beyond their training n-grams.

The neural language model introduced by Bengio et al. [8] in 2003 demonstrated that distributed word representations learned jointly with a feedforward language model substantially outperformed n-gram baselines on held-out perplexity. This representation-learning paradigm was later operationalised at scale by Mikolov et al. [9] through the Word2Vec skip-gram and CBOW objectives, and by Pennington et al. through GloVe [23], establishing pre-trained word vectors as a standard NLP ingredient.

Recurrent architectures-particularly Long Short-Term Memory networks [24] and Gated Recurrent Units [25]-enabled the processing of variable-length sequences, and the sequence-to-sequence framework of Sutskever et al. [10] applied them to machine translation with notable success. The introduction of additive attention by Bahdanau et al. [26] in 2015 alleviated the encoder bottleneck by allowing the decoder to selectively attend to source tokens, a mechanism that would prove foundational for the Transformer.

The decisive architectural breakthrough came in 2017 when Vaswani et al. [1] eliminated recurrence entirely, replacing it with a stack of multi-head self-attention layers and position-wise feedforward networks. The resulting Transformer trained more efficiently on parallel hardware and achieved new state-of-the-art results on translation.

BERT [27] demonstrated that bidirectional pre-training via masked language modelling yielded powerful contextual representations transferable to downstream tasks. GPT [28] and GPT-2 [29] pursued the complementary autoregressive pre-training objective, showing that sufficiently large causal language models could generate coherent long-form text.

GPT-3 [30], with 175 billion parameters trained on hundreds of billions of tokens, marked the arrival of few-shot in-context learning as an emergent property of scale. Subsequent years witnessed rapid iteration: Codex [3] applied the paradigm to code; InstructGPT [13] introduced RLHF to align model behaviour with human preferences; and ChatGPT's public release in November 2022 brought LLMs to mainstream awareness. The period 2023–2026 has been characterised by the proliferation of model families, the open-sourcing of capable weights (Llama, Qwen, DeepSeek), and the development of specialised reasoning models.

Table 1: Timeline of Key Language Model Milestones

Year	Milestone	Key Model / Paper
1950	Turing Test proposed	Turing (1950)
1966	First chatbot	ELIZA (Weizenbaum)
1986	Backpropagation popularised	Rumelhart et al.
2001	Neural language models	Bengio et al. (2003)
2013	Word embeddings	Word2Vec (Mikolov et al.)
2014	Sequence-to-sequence models	Sutskever et al.
2015	Attention mechanism formalised	Bahdanau et al.
2017	Transformer architecture	Vaswani et al. — "Attention Is All You Need"
2018	BERT bidirectional pre-training	Devlin et al. (Google)
2018	GPT-1 autoregressive pre-training	Radford et al. (OpenAI)
2019	GPT-2 / T5 / XL Net	OpenAI / Google
2020	GPT-3 (175 B parameters)	Brown et al. (OpenAI)
2021	Codex / GitHub Copilot	Chen et al. (OpenAI)
2022	Instruct GPT / ChatGPT / BLOOM	OpenAI / Big Science
2023	GPT-4 / LLaMA / Claude-1 / PaLM 2	OpenAI / Meta / Anthropic / Google
2024	Gemini 1.5 / Llama 3 / Claude 3 / Qwen2	Google / Meta / Anthropic / Alibaba
2025	GPT-4o / Claude 3.5 / DeepSeek-V3 / Qwen3	OpenAI / Anthropic / DeepSeek / Alibaba
2026	Multimodal agents; reasoning specialisation	Ongoing frontier research

Transformer Architecture and Technical Foundations

Self-Attention and Multi-Head Attention

The core innovation of the Transformer [1] is the scaled dot-product attention function. Given query matrix Q, key matrix K, and value matrix V - each derived from the input by learned linear projections - attention is computed as:

Attention(Q, K, V) = softmax(QK^T / sqrt(d_k)) V where d_k is the key dimensionality. Scaling by 1/sqrt(d_k) prevents the dot products from growing large enough to push softmax into saturation regions. Multi-head attention runs h parallel attention heads with separate projections, concatenates their outputs, and applies a final linear transformation, allowing the model to jointly attend to information from different representational subspaces at different positions [1].

In autoregressive (decoder-only) models such as GPT and Llama, a causal mask is applied so that position i attends only to positions <= i, preserving the left-to-right generative property. Bidirectional models such as BERT use full attention across all positions.

Positional Encoding

Because attention is permutation-equivariant, positional information must be injected explicitly. Vaswani et al. [1] used sinusoidal encodings; later work introduced learned absolute embeddings (BERT), relative position encodings (Shaw et al., 2018), ALiBi (Press et al., 2022) [31], and Rotary Position Embedding (RoPE) (Su et al., 2022) [32]. RoPE has become the dominant choice for modern LLMs owing to its length extrapolation properties and compatibility with efficient attention kernels.

Feedforward Layers and Activation Functions

Each Transformer layer contains a position-wise feedforward network (FFN) comprising two linear transformations with a non-linearity. SwiGLU [33], a gated variant of the Swish activation, has replaced ReLU in most contemporary LLMs (Llama 2/3, Qwen, Mistral) owing to its improved training stability and downstream performance [34].

Reinforcement Learning from Human Feedback (RLHF)

RLHF [13] is the dominant post-training alignment paradigm. The pipeline proceeds in three stages: (i) supervised fine-tuning (SFT) on curated demonstration data to initialise the policy; (ii) reward model training, where annotators rank model completions and a Bradley-Terry model is fitted; and (iii) proximal policy optimisation (PPO) [35] to maximise the reward signal while penalising KL divergence from the SFT reference policy to prevent reward hacking. Anthropic's Constitutional AI (CAI) [36] augments RLHF with a self-critique loop guided by an explicit set of principles, reducing dependence on human labour in the preference phase. Direct Preference Optimisation (DPO) [37] later offered a simpler, reference-model-free alternative that has been widely adopted for instruction tuning.

Chain-of-Thought Reasoning

Wei et al. [38] demonstrated that prompting LLMs to produce intermediate reasoning steps-'chain-of-thought' (CoT) prompting-dramatically improved performance on arithmetic, commonsense, and symbolic reasoning tasks. Kojima et al. [39] showed that appending 'Let's think step by step' to zero-shot prompts elicited similar improvements without few-shot exemplars. Self-consistency [40] further boosts CoT by sampling multiple reasoning paths and marginalising over the majority answer. Process reward models (PRMs) [41] provide fine-grained supervision over individual reasoning steps, and have been central to training o1-style reinforcement-learning-based reasoning models.

Mixture of Experts (MoE)

MoE architectures [42] replace the dense FFN in each Transformer layer with a set of expert sub-networks and a learned routing function. The routing function-typically a top-k softmax gate-activates only k of E experts per token, so that total compute scales with k/E rather than E. Shazeer et al.'s sparse MoE [42] was subsequently refined by Switch Transformer [43] (top-1 routing for simplicity), Mixtral 8x7B [44] (top-2 routing), and DeepSeek-V2/V3 [19] (fine-grained expert decomposition with auxiliary load-balancing losses). MoE enables large total parameter counts-beneficial for model capacity-while maintaining manageable per-token FLOP budgets.

Retrieval-Augmented Generation (RAG)

RAG [45] addresses LLM knowledge staleness and hallucination by conditioning generation on documents retrieved from an external corpus at inference time. The original formulation by Lewis et al. [45] used a pre-trained bi-encoder to retrieve top-k passages and concatenated them with the query before passing to the generator. Contemporary RAG systems employ dense and sparse hybrid retrieval, re-ranking with cross-encoders, chunk-level embedding with vector databases (Pinecone, Weaviate, Chroma), and multi-hop retrieval chains [46]. RAG complements parametric LLM knowledge with verifiable, up-to-date information, and is foundational to enterprise knowledge-base applications.

Table 2: Comparison of Major LLM Architectures

Architecture	Attention	Positional Encoding	Training Objective	Examples
Encoder-only	Full bidirectional	Absolute / learned	MLM, NSP	BERT, RoBERTa, ALBERT
Decoder-only	Causal (autoregressive)	RoPE / ALiBi	Next-token prediction	GPT, Llama, DeepSeek
Encoder-Decoder	Cross-attention	Relative	Seq2Seq (span prediction)	T5, BART, mT5
Mixture of Experts	Sparse gated	RoPE	Next-token (top-k routing)	Mixtral, DeepSeek-V3, Qwen3
Multimodal	Cross-modal (vision+text)	2D RoPE / patch	Contrastive + LM	Gemini, GPT-4o, LLaVA

Scaling Laws and Emergent Abilities

Major Model Families

Table 3: Comparison of Major LLM Families (2024–2026)

Model Family	Developer	Release	Params (B)	Context	Multimodal	OpenWeight	Notable Feature
GPT-4o	OpenAI	2024	~1,000+	128 K	Yes	No	Omni-modal I/O
GPT-4.1	OpenAI	2025	Undisclosed	1 M	Yes	No	1 M-token context
Claude 3.5 Sonnet	Anthropic	2024	Undisclosed	200 K	Yes	No	Top coding/reasoning
Claude 3.7	Anthropic	2025	Undisclosed	200 K	Yes	No	Extended thinking
Gemini 1.5 Pro	Google DeepMind	2024	Undisclosed	1 M	Yes	No	MoE long context
Gemini 2.0 Flash	Google DeepMind	2025	Undi Major Reviewscl osed	1 M	Yes	No	Agentic speed
Llama 3.3 70B	Meta AI	2024	70	128 K	No	Yes	Open-weight SOTA
Llama 4 Scout	Meta AI	2025	109	10 M	Yes	Yes	MoE open-source
DeepSeekV3	DeepSeek	2024	685 (MoE)	128 K	No	Yes	Cost-efficient MoE
DeepSeekR1	DeepSeek	2025	671 (MoE)	128 K	No	Yes	RL-driven reasoning
Qwen2.5-72B	Alibaba	2024	72	128 K	No	Yes	Multilingual dense
Qwen3-235B	Alibaba	2025	235 (MoE)	128 K	No	Yes	Thinking/non-thinking

Kaplan et al. [11] established that language model loss decreases as a smooth power law in compute, data, and parameter count, and that for a fixed compute budget the optimal trade-off allocates roughly equal resources to model size and data volume. Hoffmann et al. [12] ('Chinchilla') revised this ratio substantially, finding that Chinchilla-optimal training requires approximately 20 tokens per parameter - a result that prompted the research community to train smaller, more data-rich models.

Beyond smooth scaling, Wei et al. [47] identified 'emergent abilities': capabilities that appear abruptly as model scale crosses certain thresholds and are effectively absent in smaller models. Examples include multi-step arithmetic, analogical reasoning, and multi-language translation. The mechanism underlying emergence remains debated: Schaeffer et al. [48] argue that many claimed emergencies are artefacts of discontinuous evaluation metrics rather than genuine phase transitions, while others maintain that some qualitative shifts are real.

Scaling has also been applied to post-training compute. The 'inference-time scaling' paradigm - exemplified by OpenAI's o1 [5] and DeepSeek-R1 [20]-allocates additional compute at inference by generating extended internal monologues before producing a final answer, achieving marked gains on competition-level mathematics and coding without expanding model parameters. This represents a shift from the 'scale the model' to the 'scale the thinking' philosophy.

GPT Series (OpenAI)

OpenAI's Generative Pre-trained Transformer series represents the most commercially influential line of LLMs to date. GPT-1 [28] established the viability of unsupervised pre-training followed by fine-tuning. GPT-2 [29] demonstrated emergent text generation quality that prompted concerns about misuse. GPT-3 [30] introduced few-shot in-context learning at scale. GPT-3.5 (InstructGPT [13]) applied RLHF and became the backbone of ChatGPT. GPT-4 [2] achieved human-level performance on a suite of professional exams and introduced vision capabilities. GPT-4o ('omni') [15] unified text, image, and audio modalities in a single end-to-end model with real-time voice capabilities, while GPT-4.1 extended the context window to one million tokens for document-level processing tasks. OpenAI's o1 [5] and o3 series introduced dedicated reasoning models trained with extended chain-of-thought reinforcement learning.

Llama Series (Meta AI)

Meta's Llama (Large Language Model Meta AI) series has been the dominant open-weight LLM family. Llama 1 [49] demonstrated that carefully curated open-source data could produce competitive models. Llama 2 [50] added supervised fine-tuning and RLHF variants and was released under a permissive research licence. Llama 3 [18] significantly expanded training data and improved multilingual capabilities. Llama 3.3 (70B) achieved state-of-the-art open-weight performance on standard benchmarks. Llama 4, announced in 2025, introduced a Scout variant with a 10-million-token context window using a mixture-of-experts architecture, and a Maverick variant targeting top-tier quality. The open-weight availability of Llama models has catalysed an ecosystem of fine-tuned derivatives, quantised deployments, and research studies.

Claude Series (Anthropic)

Anthropic's Claude models are distinguished by their grounding in Constitutional AI [36]-a self-supervised alignment approach in which the model critiques and revises its own outputs according to a written constitution of principles before human labellers provide preference judgements. Claude 1 was released in early 2023. Claude 2 extended the context window to 100K tokens. Claude 3 (launched March 2024) introduced three tiers - Haiku, Sonnet, and Opus - spanning the quality-cost frontier; Claude 3 Opus achieved the highest benchmark scores at launch across MMLU, HumanEval, and GPQA. Claude 3.5 Sonnet (2024) surpassed Opus on coding tasks while maintaining Sonnet-tier latency. Claude 3.7 Sonnet (2025) introduced 'extended thinking' - a visible chain-of-thought reasoning mode - and demonstrated industry-leading performance on agentic software-engineering tasks such as SWE-bench.

Gemini Series (Google DeepMind)

Google DeepMind's Gemini series represents the first natively multimodal LLM family from a major laboratory, trained from inception on interleaved sequences of text, image, audio, and rather than adapting a -

-text-only base model [51]. Gemini 1.0 Ultra achieved state-of-the-art scores on MMLU. Gemini 1.5 Pro [52] introduced a one-million-token context window using a hybrid sparse-attention mechanism, enabling document-level understanding, full-codebase analysis, and hour-long video comprehension in a single pass. Gemini 2.0 Flash (2025) introduced real-time multimodal streaming and tool-use capabilities for agentic pipelines, while Gemini 2.5 Pro further refined reasoning performance.

Qwen Series (Alibaba Cloud)

Alibaba Cloud's Qwen (Tongyi Qianwen) series [21] has established itself as the leading multilingual open-weight model family, with particular strength in Chinese-English bilingual tasks. Qwen2 [53] demonstrated competitive performance with Llama 3 across standard English benchmarks while significantly outperforming it on Chinese-language tasks. Qwen2.5 introduced specialised variants for coding (Qwen2.5-Coder), mathematics (Qwen2.5-Math), and vision-language tasks (Qwen-VL). Qwen3 [22], released in 2025, is a mixture-of-experts architecture available in sizes up to 235 billion parameters that introduced a 'thinking/non-thinking' mode toggle, allowing users to trade reasoning depth against latency.

DeepSeek Series

DeepSeek, a Chinese AI research company, has produced a series of models that have attracted broad attention for their combination of competitive quality and cost-efficient training. DeepSeek-V2 introduced a Multi-Head Latent Attention (MLA) mechanism that compresses the KV-cache to reduce memory overhead. DeepSeek-V3 [19] is a 685-billion-parameter MoE model trained on 14.8 trillion tokens; its training cost of approximately 5.6 million GPU-hours was notably below industry norms for comparable-quality dense models, attributed to FP8 mixed-precision training and architectural efficiency. DeepSeek-R1 [20] applied group relative policy optimisation (GRPO) reinforcement learning - without supervised chain-of-thought bootstrapping - to produce a reasoning model that matched o1-preview on competition mathematics while releasing its weights openly.

Benchmarking and Evaluation of LLMs

Systematic evaluation of LLMs requires benchmarks that assess both breadth of knowledge and depth of reasoning. MMLU (Massive Multitask Language Understanding) [54] evaluates 57-subject multiple-choice knowledge from elementary to professional level and has become the canonical measure of general knowledge. HumanEval [3] and MBPP assess functional code correctness through unit tests. GSM8K [55] tests grade-school arithmetic problem-solving. MATH [56] extends this to competition-level mathematics. GPQA (Graduate-Level Google-Proof Q&A) [57] targets expert-level science questions designed to be difficult to answer by searching the web, serving as a proxy for genuine reasoning.

MT-Bench and Chatbot Arena [58] provide human preference-based evaluation of open-ended conversation quality, capturing dimensions - style, tone, helpfulness - that automated benchmarks miss. MMMU evaluates multimodal reasoning across diverse visual-question-answering tasks. SWE-bench [59] measures the ability to resolve real GitHub issues in Python repositories, serving as an increasingly important real-world coding evaluation.

Benchmark saturation is an acknowledged concern: frontier models now exceed 90% on MMLU, diminishing its discriminative power. The field has responded with harder benchmarks (GPQA Diamond, FrontierMath, LiveCodeBench) and process-level evaluations that assess reasoning correctness rather than answer accuracy alone. Contamination - overlap between training data and benchmark test sets - remains a persistent confound [60], motivating the development of live, continuously updated evaluation suites.

Reasoning Capabilities

Mathematical and logical reasoning has emerged as a primary differentiator among frontier LLMs. The chain-of-thought prompting paradigm [38] catalysed a wave of research into how intermediate computation can be externalised and leveraged. The introduction of process reward models (PRMs) [41] - which score each intermediate reasoning step - allows training signals to target the quality of the reasoning process rather than merely its output, providing denser learning signal for multi-step tasks.

Table 4: Benchmark Performance of Major LLMs (scores as of mid-2025; - = not reported)

Model	MMLU	Human Eval	GSM8K	MATH	GPQA	MMMU	MT-Bench
GPT-4o (2024)	88.7	90.2	97.0	76.6	53.6	69.1	9.0
Claude 3.5 Sonnet	88.3	92.0	96.8	78.2	59.4	70.2	9.0
Gemini 1.5 Pro	85.9	84.1	94.4	67.7	46.2	65.8	8.9
Llama 3.3 70B	86.0	88.4	95.1	68.0	46.7	64.9	8.7
DeepSeek-V3	87.5	89.3	95.5	75.3	57.4	68.4	8.9
DeepSeek-R1	90.8	96.3	97.2	92.3	71.5	-	-
Qwen2.5-72B	86.1	87.2	95.5	73.2	49.3	64.6	8.8
Qwen3-235B (think)	90.1	95.7	97.4	93.1	72.0	-	-

OpenAI's o1 and o3 models represent the most prominent application of test-time compute scaling: the models are trained to produce extended internal reasoning traces through reinforcement learning, with inference cost scaling with the number of reasoning tokens generated [5]. DeepSeek-R1 [20] demonstrated that this approach could be implemented without supervised CoT bootstrapping, using pure RL from a base model - an approach termed 'cold start' training. These reasoning-specialised models substantially outperform standard instruction-tuned models on competition-level mathematics (AIME), physics, and advanced coding tasks.

Remaining challenges include multi-hop reasoning over long contexts, robust spatial and temporal reasoning, causal inference from observational data, and the avoidance of sycophantic reasoning - the tendency to adjust conclusions to match perceived user preferences rather than evidence. Systematic errors on seemingly simple counting and ordering tasks ('reversal curse') [61] suggest that LLM reasoning is not yet fully general.

Coding Capabilities

Code generation was among the earliest demonstrations of LLM utility beyond natural language. Codex [3], fine-tuned from GPT-3 on GitHub repositories, achieved 28.8% pass@1 on HumanEval and powered GitHub Copilot. Subsequent model generations have dramatically raised this ceiling: Claude 3.5 Sonnet, DeepSeek-R1, and Qwen3 all exceed 90% on HumanEval, and frontier models approach 50–60% on the harder LiveCodeBench.

SWE-bench [59] provides a more ecologically valid coding evaluation by requiring models to produce patches that resolve real GitHub issues, including test suite verification. Claude 3.7 Sonnet achieved 70.3% on SWE-bench Verified as of early 2025, significantly ahead of comparable models, attributed to training on extensive agentic software-engineering trajectories. Agentic coding frameworks - wherein the LLM iteratively writes code, executes it, reads error output, and revises - have further extended effective coding capability beyond single-shot completion.

Low-code and no-code applications are transforming software development workflows. Studies indicate that developers using AI coding assistants complete tasks 55% faster and report higher satisfaction [62], though concerns about over-reliance, security vulnerabilities in AI-generated code [63], and intellectual property questions around training data sourced from copyleft repositories remain active debates.

Educational Applications

The application of LLMs to education encompasses intelligent tutoring systems, automated essay scoring, curriculum generation, and accessibility tools. Khanmigo, powered by GPT-4 and deployed by Khan Academy, delivers personalised Socratic tutoring in mathematics and science, adapting question difficulty and explanation style in real time. Studies in randomised controlled settings have documented measurable learning gains attributable to AI tutoring [64], though effect sizes vary substantially with subject, student age, and implementation quality.

LLMs demonstrate considerable promise for language learning, providing immediate corrective feedback, contextualised vocabulary instruction, and immersive conversational practice at scale - capabilities previously feasible only with costly human tutors. Automated essay scoring systems powered by LLMs correlate well with expert human raters for holistic quality and grammar, but show reduced validity for argumentation quality and disciplinary knowledge application [65].

Critical concerns in educational deployment include academic integrity (student use of LLMs for assignment completion), the potential for LLM factual errors to propagate into learner knowledge, and equity - models trained predominantly on English text may provide inferior support to speakers of lower-resource languages. Guidelines from UNESCO and national education ministries are beginning to codify acceptable use, though institutional adoption varies widely.

Enterprise Applications

Enterprise adoption of LLMs has accelerated dramatically, with a McKinsey Global Survey (2024) estimating that over 65% of organisations regularly employ generative AI in at least one business function. Dominant use cases include customer support automation, document summarisation, contract analysis, internal knowledge retrieval, and software development assistance. LLM-powered customer service systems reduce average handling time while maintaining satisfaction scores comparable to human agents at a fraction of the operational cost.

RAG architectures [45] are central to enterprise deployment, enabling LLMs to answer questions grounded in proprietary document repositories without requiring fine-tuning or the risks associated with exposing sensitive data in training pipelines. Fine-tuning remains important for domain-specific stylistic adaptation, entity recognition, and task-format specialisation, and is increasingly accessible through efficient methods such as LoRA [66] and QLoRA [67].

Governance challenges in enterprise settings include prompt injection attacks, in which malicious content in retrieved documents hijacks model instructions [68]; hallucination-induced errors in high-stakes processes (legal, medical, financial); and model drift - gradual performance degradation as domain data distributions shift. Structured evaluation, human-in-the-loop escalation protocols, and retrieval confidence scoring are standard mitigations in mature enterprise deployments.

Scientific Research Applications

The intersection of LLMs and scientific research has produced a cluster of high-impact demonstrations. AlphaFold 2 [69], though not an LLM per se, demonstrated that attention-based architectures trained on protein sequence databases could predict protein structure with experimental accuracy - a breakthrough in structural biology with direct therapeutic implications. Subsequent protein language models (ESMFold, ProtTrans) have extended this paradigm to sequence-function prediction.

In chemistry, LLMs have been applied to retrosynthetic planning, reaction prediction, and molecular property prediction. Transformer models pre-trained on SMILES strings and reaction databases outperform prior graph neural network approaches on multiple benchmarks [70]. In materials science, LLM-assisted discovery pipelines have accelerated the identification of battery electrolytes and photovoltaic materials by coupling generative models with high-throughput density functional theory calculations.

For literature-intensive disciplines - medicine, law, social science - LLMs enable systematic review drafting, hypothesis generation from large corpora, and evidence synthesis. Clinical LLMs fine-tuned on electronic health records have demonstrated performance comparable to generalist physicians on diagnostic reasoning tasks [71], though deployment in clinical workflows raises substantial safety and liability concerns. The emerging paradigm of 'AI scientists' - fully autonomous systems that design, execute, and interpret experiments - has been demonstrated in limited domains [72] and represents a transformative but still nascent direction.

Multimodal AI Systems

The fusion of language modelling with visual perception has produced a generation of vision-language models (VLMs) that can interpret, reason about, and generate across text and image modalities. Early approaches such as CLIP [73] established contrastive image-text alignment as an effective pre-training objective. LLaVA [74] and InstructBLIP demonstrated that connecting a pre-trained vision encoder to an LLM via a learned adapter yielded strong performance on visual question answering and image captioning.

Natively multimodal models such as Gemini [51] and GPT-4o [15] represent a more ambitious approach: training a single model end-to-end on interleaved multimodal sequences from inception. Gemini 1.5 Pro's one-million-token context enables video understanding at the level of individual frames over hour-long footage. GPT-4o processes audio, image, and text in real time, enabling natural conversational interaction with visual content. Claude 3 introduced image understanding capabilities across the model range.

Beyond image and text, multimodal research is expanding into video understanding (sequential temporal reasoning over thousands of frames), audio-language models (speech recognition, speaker diarisation, music generation), and 3D scene understanding. Unified tokenisation schemes that represent diverse modalities as discrete tokens within a single vocabulary are an active research direction, promising to further unify the architecture landscape.

AI Agents and Autonomous Systems

The concept of an LLM-based agent refers to a system in which the language model serves not merely as a one-shot text generator but as the cognitive core of an iterative perceive-reason-act loop. Foundational work by ReAct [75] demonstrated that interleaving reasoning traces with discrete actions (web search, calculator, code execution) significantly improved performance on interactive tasks. LangChain and LlamaIndex subsequently provided engineering frameworks that operationalised this pattern, enabling chaining of LLM calls, tool integrations, and memory management.

AutoGPT, BabyAGI, and similar early agentic systems demonstrated the potential - and fragility - of fully autonomous LLM agents operating over extended planning horizons.

Contemporary agentic systems are more carefully constrained: Anthropic's Claude with computer use [76], Microsoft Copilot Studio agents, and OpenAI Operator enable browser and desktop manipulation within sandboxed environments with human oversight checkpoints.

Multi-agent systems [77] - in which multiple specialised LLM instances communicate and coordinate - have shown promise for complex tasks exceeding the context capacity of a single model. An orchestrator agent decomposes a task into subtasks, dispatches them to specialist sub-agents (researcher, coder, critic, executor), and synthesises results. Debate and reflection between agents improves factual accuracy over single-agent chains.

Key open challenges in agentic AI include long-horizon planning consistency, robust tool use and error recovery, memory architectures that scale beyond context windows, and safety in open-ended action spaces. The risk of cascading errors - where an early mistake compounds across subsequent steps - remains a limiting factor in fully autonomous deployment, motivating research into agent introspection, uncertainty quantification, and graceful failure modes.

Limitations and Challenges

Hallucinations

Hallucination - the generation of plausible-sounding but factually incorrect content - is the most widely cited limitation of LLMs in high-stakes applications [78]. Hallucinations arise from the fundamental mismatch between the autoregressive training objective (predict the next token) and the requirement for factual accuracy. Mitigation strategies include RAG (grounding generation in retrieved documents), factuality-targeted training rewards, uncertainty estimation, and citation-generation with subsequent verification. Despite progress, no current method eliminates hallucination entirely, and detection remains an open problem: LLMs themselves are unreliable detectors of their own errors.

Bias and Fairness

LLMs trained on large web corpora inherit and can amplify societal biases present in that data [79]. Demographic biases manifesting as differential performance, stereotype reinforcement, and offensive content generation have been documented across models, with documented disparities by race, gender, nationality, religion, and socioeconomic status. Debiasing approaches include curated data curation, targeted fine-tuning, and constitutional principles prohibiting discriminatory outputs, but trade-offs between bias reduction and general capability remain an active tension.

Safety and Alignment

The alignment problem - ensuring LLM behaviour is consistent with human values and intentions - encompasses both near-term concerns (jailbreaking, harmful content generation, deceptive alignment) and longer-horizon existential risks [80]. Jailbreaking - adversarial prompting that bypasses safety guardrails - remains an ongoing arms race between attack and defence. Sleeper agent experiments [81] demonstrated that models could be trained to behave safely under normal conditions while maintaining hidden malicious behaviours triggered by specific inputs, motivating investment in interpretability research.

Privacy

LLMs trained on web data unavoidably ingest personal information, and can reproduce memorised training data including email addresses, phone numbers, and in rare cases medical records [82]. Membership inference attacks can in some cases determine whether a specific document was included in training data. Differential privacy training mitigates memorisation at the cost of model quality. GDPR's right to erasure creates legal tensions with the difficulty of 'unlearning' specific training data from large models.

Copyright and Intellectual Property

The legal status of training LLMs on copyrighted text remains contested across jurisdictions, with active litigation against OpenAI, Meta, and other developers. LLMs can reproduce substantial portions of memorised copyrighted text verbatim, raising infringement concerns. The output side is equally contested: courts in the United States and European Union are actively adjudicating whether LLM-generated content can receive copyright protection and under what conditions. Regulatory clarity is emerging slowly, with the EU AI Act among the first frameworks to address training data transparency.

Computational Cost and Environmental Impact

The training of frontier LLMs requires massive computational resources with correspondingly significant energy consumption. GPT-4's training has been estimated to require tens of millions of GPU-hours; Gemini Ultra and Llama 3 405B are similarly resource-intensive. Data centre electricity consumption attributable to AI workloads is growing rapidly, raising concerns about carbon emissions in the context of climate commitments [83]. Efficiency research - MoE architectures, quantisation, sparse activation, knowledge distillation, and hardware-software co-design - is active and has produced meaningful gains, but the overall compute envelope of frontier training continues to expand.

Ethical and Societal Implications

The deployment of LLMs at societal scale raises questions that extend beyond technical performance metrics. In the labour market, LLMs automate cognitive tasks previously requiring professional training, creating -

-both efficiency gains and displacement risks. Studies by Acemoglu (2024) and Brynjolfsson et al. project significant disruption in writing, legal, accounting, and customer-service professions, while also creating new roles in AI development, prompt engineering, and AI oversight.

Misinformation represents a systemic risk: LLMs lower the cost of producing convincing synthetic text, audio, and video, enabling disinformation campaigns at previously infeasible scale. Synthetic media detection research [84] has not kept pace with generation quality, and policy responses - watermarking mandates, provenance standards such as C2PA - are in early stages of adoption.

Concentration of AI capability in a small number of well-resourced organisations raises concerns about the democratisation of access. Open-weight models (Llama, Qwen, DeepSeek) partially counterbalance this dynamic by enabling independent researchers and smaller organisations to deploy competitive models, though the compute required for training remains inaccessible to most. Discussions of 'compute governance' as a lever for AI oversight have entered policy discourse.

Power asymmetries between AI developers and users - with training objectives, data composition, and alignment decisions made by developers with limited public accountability - motivate calls for participatory design, diverse annotator pools, and public audit mechanisms. The philosophy of AI ethics has been criticised for focusing on individual harms at the expense of structural and political dimensions of AI deployment.

Regulatory and Governance Considerations

The regulatory landscape for LLMs is evolving rapidly and diverges significantly across jurisdictions. The European Union's AI Act [85], passed in 2024 and entering phased implementation, establishes a risk-tiered framework in which 'general-purpose AI' systems with significant societal impact are subject to transparency, documentation, and testing obligations. Foundation model providers above a compute threshold (10^25 FLOP) face systemic risk classification with additional requirements.

In the United States, President Biden's Executive Order on AI (October 2023) required developers of dual-use foundation models to share safety test results with the government and established the AI Safety Institute within NIST as a coordination body. The subsequent administration has recalibrated emphasis from safety mandates toward innovation promotion, though sector-specific regulators (FDA for medical AI, SEC for financial AI) continue to develop domain-specific guidance.

The People's Republic of China has enacted regulations on generative AI services (effective August 2023) requiring content moderation, real-name registration for users of public-facing services, and security assessments for models with substantial societal influence. International coordination on AI governance is nascent: the Hiroshima AI Process (G7) and the Global Partnership on AI have produced voluntary principles, while the UN's High-Level Advisory Body on AI has called for a more binding international framework analogous to nuclear non-proliferation treaties.

Technical governance tools include model evaluations for dangerous capabilities (CBRN, cyberoffence), structured access regimes, red-teaming requirements, and incident reporting. Anthropic [86], OpenAI, and DeepMind have published 'responsible scaling policies' or 'preparedness frameworks' committing to internal capability thresholds above which deployment will be paused or additional mitigations required, though critics note these are self-imposed and lack independent verification.

Future Directions of Artificial Intelligence

The near-term trajectory of LLM research is shaped by several converging trends. Inference-time compute scaling - allocating additional reasoning tokens at inference rather than training larger models - is likely to continue delivering capability gains for structured reasoning tasks, with research into more efficient search algorithms (tree of thought, Monte Carlo tree search) over reasoning traces. Multimodal unification will proceed toward models that process and generate across all human communication modalities - text, image, audio, video, and eventually robotic proprioception - within a single architecture.

Embodied AI systems that ground language in physical environment perception and action represent the frontier at which LLMs meet robotics; projects such as Google's RT-2 and Figure's humanoid robot demonstrations preview this direction.

Long-context and memory architectures are an active research front. Current approaches - sliding window attention, retrieval augmentation, memory-augmented transformers - partially address the limitation of fixed context windows but introduce latency and complexity trade-offs. Neurologically inspired persistent memory mechanisms that selectively encode and retrieve episodic experience across arbitrarily long horizons remain an open challenge.

Interpretability and mechanistic understanding of LLM internals have advanced from circuit-level analysis [87] of toy models to sparse autoencoder decomposition [88] of frontier model residual streams. Understanding the computational mechanisms underlying factual recall, reasoning, and value learning is foundational to both safety and capability research, and is likely to yield training and alignment insights over the next decade.

The development of smaller, more efficient models deployable on consumer hardware ('on-device AI') is democratising access and enabling privacy-preserving applications. Quantisation to 4-bit and 2-bit precision, speculative decoding, and neural architecture search have produced models that run adequately on mobile devices, with further efficiency expected from next-generation neural processing unit designs.

Finally, the integration of symbolic reasoning systems with neural LLMs - neurosymbolic AI - may provide a path to more robust logical consistency and verifiable reasoning. Hybrid systems that use LLMs for natural language parsing and generation while delegating formal reasoning steps to theorem provers, constraint solvers, or knowledge graphs have shown promise on mathematical and planning tasks [89].

Discussion

The preceding review reveals a field characterised by extraordinary momentum, genuine transformative capability, and equally genuine unresolved challenges. The dominant narrative of continuous improvement on benchmark leaderboards must be contextualised by the equally consistent finding that benchmark performance does not reliably predict deployment performance in high-stakes, open-ended real-world tasks. The gap between what frontier LLMs can do in controlled evaluation and what they do reliably and safely at scale remains a central practical and scientific problem.

The bifurcation between open-weight and proprietary models has emerged as a defining structural feature of the current landscape. Proponents of openness cite accelerated research, democratic access, and the practical security benefits of transparent systems. Proponents of controlled access cite the difficulty of deploying safety measures on weights that cannot be updated post-release and the risk of adversarial fine-tuning to remove alignment. The DeepSeek R1 release crystallised this debate, demonstrating that open-weight models at frontier quality are technically achievable at lower cost than previously assumed, intensifying the urgency of policy discussions.

Anthropic's Constitutional AI [36] and Google's RLHF-based alignment [13] represent two ends of a spectrum from rule-based to preference-based alignment. The practical convergence of these approaches - most deployed models use hybrids - suggests that alignment is not a binary property but a multidimensional engineering problem requiring ongoing iteration. The long-term question of whether systems trained by next-token prediction can acquire the kind of stable value representations required for trustworthy autonomous agency is unsettled, and should not be assumed affirmative without substantially more interpretability evidence.

From an applications perspective, the evidence base for educational and scientific LLM applications is growing but still immature relative to the scale of deployment. Randomised controlled trials are rare; most evidence consists of observational studies with significant confounders. As LLMs are embedded in consequential workflows, the demand for rigorous evaluation methodologies analogous to clinical trial standards will intensify - particularly in medicine, law, and public policy.

Conclusion

This review has surveyed the evolution, technical foundations, capabilities, limitations, and future trajectory of large language models as of 2026. From the statistical language models of the 1990s to the transformer-based systems capable of graduate-level scientific reasoning and autonomous software engineering, the intellectual journey represents one of the most rapid capability expansions in the history of technology.

The transformative promise of LLMs - accelerating scientific discovery, democratising education, augmenting human cognitive work, and enabling new forms of human-machine collaboration - is credible and substantiated by a growing body of empirical evidence. Yet this promise is inseparable from risks that demand serious, sustained attention: the systemic production of plausible misinformation, the amplification of societal biases, the concentration of transformative power in a small number of organisations, and the long-horizon alignment challenge of ensuring that increasingly capable autonomous systems remain reliably beneficial.

The field's greatest near-term scientific priorities include: reliable factual grounding without retrieval degrading reasoning quality; interpretability methods scalable to frontier models; alignment techniques robust to distributional shift and adversarial pressure; and evaluation frameworks that measure real-world performance rather than saturated benchmark metrics. Progress on these fronts will determine whether the extraordinary capabilities demonstrated in 2026 translate into durable, equitable, and trustworthy societal benefit.

The governance infrastructure to manage these risks and opportunities is nascent and inconsistent across jurisdictions. The coming years will require close collaboration among researchers, developers, policymakers, civil society, and affected communities to craft frameworks that neither stifle beneficial innovation nor permit the unchecked deployment of systems whose failure modes remain incompletely understood. This review is offered as a contribution to the shared knowledge base on which such collaboration must rest.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. arXiv:1706.03762.
2. OpenAI. (2023). GPT-4 technical report. arXiv:2303.08774.
3. Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., et al. (2021). Evaluating large language models trained on code. arXiv:2107.03374.
4. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589.
5. OpenAI. (2024). Learning to reason with LLMs. OpenAI Technical Blog. https://openai.com/research/learning-to-reason-with-llms
6. Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379–423.
7. Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14(2), 179–211.
8. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. Journal of Machine Learning Research, 3, 1137–1155.
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26. arXiv:1310.4546.
10. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems, 27. arXiv:1409.3215.
11. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., et al. (2020). Scaling laws for neural language models. arXiv:2001.08361.
12. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., et al. (2022). Training compute-optimal large language models. arXiv:2203.15556.

13. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35. arXiv:2203.02155.
14. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862.
15. OpenAI. (2024). Hello GPT-4o. OpenAI Technical Blog. https://openai.com/research/gpt-4o
16. Anthropic. (2024). The Claude 3 model family: Opus, Sonnet, Haiku. Anthropic Technical Report. https://www.anthropic.com/news/claude-3-family
17. Google DeepMind. (2023). Gemini: A family of highly capable multimodal models. arXiv:2312.11805.
18. Meta AI. (2024). Meta Llama 3. arXiv:2407.21783.
19. DeepSeek-AI. (2024). DeepSeek-V3 technical report. arXiv:2412.19437.
20. DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv:2501.12948.
21. Qwen Team, Alibaba Cloud. (2024). Qwen2 technical report. arXiv:2407.10671.
22. Qwen Team, Alibaba Cloud. (2025). Qwen3 technical report. arXiv:2505.09388.
23. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. Proceedings of EMNLP, 1532–1543.
24. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
25. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078.
26. Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. ICLR 2015. arXiv:1409.0473.
27. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT 2019. arXiv:1810.04805.
28. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI Blog.
29. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8).
30. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33. arXiv:2005.14165.
31. Press, O., Smith, N. A., & Lewis, M. (2022). Train short, test long: Attention with linear biases enables input length extrapolation. ICLR 2022. arXiv:2108.12409.
32. Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., & Liu, Y. (2022). RoFormer: Enhanced transformer with rotary position embedding. Neurocomputing, 568. arXiv:2104.09864.
33. Shazeer, N. (2020). GLU variants improve transformer. arXiv:2002.05202.
34. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.
35. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv:1707.06347.
36. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073.
37. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36. arXiv:2305.18290.
38. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35. arXiv:2201.11903.
39. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems, 35. arXiv:2205.11916.
40. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., et al. (2023). Self-consistency improves chain of thought reasoning in language models. ICLR 2023. arXiv:2203.11171.

41. Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., et al. (2023). Let's verify step by step. arXiv:2305.20050.
42. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. ICLR 2017. arXiv:1701.06538.
43. Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research, 23(120), 1–39. arXiv:2101.03961.
44. Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., et al. (2024). Mixtral of experts. arXiv:2401.04088.
45. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33. arXiv:2005.11401.
46. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., et al. (2024). Retrieval-augmented generation for large language models: A survey. arXiv:2312.10997.
47. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., et al. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research. arXiv:2206.07682.
48. Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage? Advances in Neural Information Processing Systems, 36. arXiv:2304.15004.
49. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv:2302.13971.
50. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.
51. Anil, R., Chowdhery, A., Roberts, A., Askell, A., Amodei, A., Chen, A., et al. (2023). Gemini: A family of highly capable multimodal models. arXiv:2312.11805.
52. Google DeepMind. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530.
53. Hui, B., Yang, J., Cui, Z., Yang, J., Liu, D., Zhang, L., et al. (2024). Qwen2.5-Coder technical report. arXiv:2409.12186.
54. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. ICLR 2021. arXiv:2009.03300.
55. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., et al. (2021). Training verifiers to solve math word problems. arXiv:2110.14168.
56. Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., et al. (2021). Measuring mathematical problem solving with the MATH dataset. Advances in Neural Information Processing Systems, 34. arXiv:2103.03874.
57. Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., et al. (2023). GPQA: A graduate-level Google-proof Q&A benchmark. arXiv:2311.12022.
58. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., et al. (2024). Judging LLM-as-a-judge with MT-bench and Chatbot Arena. Advances in Neural Information Processing Systems, 36. arXiv:2306.05685.
59. Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2024). SWE-bench: Can language models resolve real-world GitHub issues? ICLR 2024. arXiv:2310.06770.
60. Jacovi, A., Goldberg, Y., & Goldberger, J. (2023). Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. arXiv:2305.10160.
61. Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A., Korbak, T., & Evans, O. (2023). The reversal curse: LLMs trained on “A is B” fail to learn “B is A”. arXiv:2309.12288.
62. Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot. arXiv:2302.06590.
63. Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., & Karri, R. (2022). Asleep at the keyboard? Assessing the security of GitHub Copilot's code contributions. IEEE Symposium on Security and Privacy.
64. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2024). Generative AI can harm learning. The Wharton School, University of Pennsylvania Working Paper.
65. Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2(2), 100050.

66. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al. (2022). LoRA: Low-rank adaptation of large language models. ICLR 2022. arXiv:2106.09685.
67. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems, 36. arXiv:2305.14314.
68. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. arXiv:2302.12173.
69. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589.
70. Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., & Lee, A. A. (2019). Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. ACS Central Science, 5(9), 1572–1583.
71. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., et al. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180.
72. Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. Nature, 624(7992), 570–578.
73. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. ICML 2021. arXiv:2103.00020.
74. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. Advances in Neural Information Processing Systems, 36. arXiv:2304.08485.
75. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. ICLR 2023. arXiv:2210.03629.
76. Anthropic. (2024). Developing a computer use model. Anthropic Technical Blog.
77. Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. ACM UIST 2023. arXiv:2304.03442.
78. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.
79. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? ACM FAccT 2021.
80. Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.
81. Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., et al. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. arXiv:2401.05566.
82. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., et al. (2021). Extracting training data from large language models. USENIX Security 2021. arXiv:2012.07805.
83. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., et al. (2021). Carbon considerations for large language model training. arXiv:2104.10350.
84. Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. PNAS, 119(1), e2110013119.
85. European Parliament. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council: Artificial Intelligence Act. Official Journal of the European Union.
86. Anthropic. (2023). Anthropic's responsible scaling policy. https://www.anthropic.com/news/anthropics-responsible-scaling-policy
87. Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., et al. (2021). A mathematical framework for transformer circuits. Transformer Circuits Thread. https://transformer-circuits.pub
88. Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. arXiv:2309.08600.
89. Pan, L., Saxon, M., Xu, W., Nathani, D., Wang, X., & Wang, W. Y. (2023). Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. arXiv:2308.03188.