International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

From Words to Intelligence: A Comprehensive Survey of Large Language Models and Their Transformative Role in Natural Language Processing

Sai Rithwik Nooguri

Abstract- — The emergence of Large Language Models (LLMs) represents one of the most consequential shifts in the history of artificial intelligence (AI) and natural language processing (NLP). Built on the Transformer architecture with self-attention mechanisms, LLMs such as BERT, GPT-3, T5, LLaMA, and GPT-4 have achieved state-of-the-art performance across a broad spectrum of linguistic tasks, fundamentally reshaping how machines comprehend and generate human language. This survey presents a systematic and comprehensive review of the evolution of NLP—from rule-based and statistical methods to the current era of foundation models—examining key architectural innovations, pre-training objectives, fine-tuning strategies including parameter-efficient methods such as Low-Rank Adaptation (LoRA), and alignment techniques including Reinforcement Learning from Human Feedback (RLHF). We critically assess performance across standard benchmarks including GLUE, SuperGLUE, and MMLU, and analyze persistent challenges such as hallucination, bias, computational cost, and explainability. Furthermore, we explore the expanding landscape of LLM applications in healthcare, education, legal reasoning, and code generation, and outline promising future directions including multimodal models, efficient inference, and AI alignment. This work aims to serve as both an accessible introduction and a scholarly reference for researchers and practitioners engaged with the rapidly evolving frontier of AI-powered language understanding.

Keywords: Large Language Models, Natural Language Processing, Transformers, BERT, GPT, Fine-tuning, RLHF, Hallucination, Benchmark Evaluation, Multimodal AI.

I. INTRODUCTION

The capacity to process, interpret, and generate natural human language has long been regarded as a hallmark of general intelligence. For decades, researchers pursued this capability through rule-based expert systems, hand-crafted grammatical parsers, and shallow statistical models. Each paradigm yielded incremental gains but ultimately fell short of broad linguistic competence. The introduction of deep learning transformed the landscape, and the subsequent development of the Transformer architecture [1] proved decisive: it enabled massively parallel training over vast textual corpora, yielding models whose scale and generalization capacity were previously unimaginable.

Today, Large Language Models (LLMs) stand at the center of AI research and commercial deployment. Models such as BERT [2], GPT-3 [3], T5 [4], and LLaMA [5] have collectively demonstrated near-human performance on tasks spanning reading comprehension, machine translation, code synthesis, and open-domain question answering. The release of ChatGPT in late 2022 marked a cultural inflection point, catalyzing public and institutional interest in AI at an unprecedented scale [6].

Despite this remarkable progress, substantial challenges remain. LLMs are known to generate factually incorrect content with high confidence—a phenomenon termed hallucination [7]—and can exhibit harmful biases absorbed from training data [8]. Their computational demands impose significant environmental and economic costs, and the opacity of their internal representations raises serious concerns about interpretability and accountability [9].

This survey contributes a structured account of the state of the field, organized as follows. Section 2 reviews the historical evolution of NLP. Section 3 details the Transformer architecture. Section 4 covers prominent LLM families. Section 5 examines pre-training and fine-tuning methods. Section 6 surveys NLP tasks and benchmark evaluation. Section 7 addresses challenges. Section 8 explores application domains. Section 9 charts future directions, and Section 10 concludes.

II. EVOLUTION OF NLP

2.1 Rule-Based Era

The earliest computational approaches to language relied on handcrafted rule systems: finite-state transducers, context-free grammars, and logic-based semantic representations. These systems, exemplified by ELIZA (1966) and SHRDLU (1971),

could produce coherent responses within narrow domains but lacked generalization. Their brittleness under paraphrase and their dependence on expert-encoded knowledge severely limited scalability.

2.2 Statistical NLP

The 1980s and 1990s saw a transition to corpus-driven statistical methods. Hidden Markov Models (HMMs) became standard for part-of-speech tagging and speech recognition. Maximum entropy and conditional random field (CRF) models improved sequence labeling. The introduction of n-gram language models enabled probabilistic text prediction. These methods were more robust than rules but remained limited by sparse data representations and the inability to capture long-range syntactic and semantic dependencies.

2.3 Neural NLP

Deep learning fundamentally altered NLP beginning around 2013. Word2Vec introduced distributed word representations, embedding semantic similarity into continuous vector spaces. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, brought sequential modeling that captured context across time steps. Sequence-to-sequence models with attention mechanisms enabled neural machine translation at scale. These innovations set the stage for the Transformer revolution.

2.4 The Foundation Model Era

The convergence of large-scale unlabeled text corpora, distributed GPU computing, and the Transformer architecture gave rise to what is now termed foundation models—massive pretrained systems adaptable to thousands of downstream tasks. This paradigm shift, formalized by Bommasani et al. (2021), decouples general-purpose representation learning from task-specific adaptation, dramatically reducing the data requirements for individual applications. LLMs represent the linguistic instantiation of this broader trend.

III. THE TRANSFORMER ARCHITECTURE

3.1 Self-Attention Mechanism

Vaswani et al. [1] introduced the Transformer in 2017, proposing an architecture built entirely on self-attention—eliminating recurrence and convolution in favor of direct pairwise token interactions. Given an input sequence of token embeddings, self-attention computes query (Q), key (K), and value (V) projections, producing contextually weighted output representations: $\text{Attention}(Q, K, V) = \text{softmax}(QK^T / dk) \cdot V$ where dk is the dimensionality of the key vectors. This formulation allows every token to attend to every other token simultaneously, enabling parallelization and capturing long-range dependencies that sequential architectures struggle with.

3.2 Multi-Head Attention

Rather than computing a single attention function, the Transformer applies h parallel attention heads, each operating in a projected subspace of reduced dimensionality. The outputs are concatenated and linearly projected. Multi-head attention allows the model to jointly attend to information from different representational subspaces at different positions—capturing syntactic, semantic, and positional patterns simultaneously.

3.3 Positional Encoding and Feed-Forward Layers

Since self-attention is permutation-invariant, positional information is injected via sinusoidal positional encodings added to input embeddings. Each Transformer block is completed by a position-wise feed-forward sub-layer, layer normalization, and residual connections. The encoder processes input sequences bidirectionally; the decoder generates outputs autoregressively, attending to encoder outputs via cross-attention. This encoder-decoder design was central to the original machine translation task [1] and subsequently adapted into a diverse family of architectures.

3.4 Scaling Laws

A pivotal finding in LLM research concerns the predictable relationship between model size, dataset size, and performance. Kaplan et al. (2020) demonstrated that loss scales as a power law with respect to both parameters and training tokens. Hoffmann et al. (2022) refined this insight in the Chinchilla study, showing that many existing large models were undertrained relative to their parameter count and advocating for compute-optimal training. These scaling laws guide architectural decisions and resource allocation across the field.

IV. PROMINENT LLM FAMILIES

4.1 Encoder-Only: BERT

Devlin et al. [2] introduced BERT (Bidirectional Encoder Representations from Transformers) in 2018, pre-training a deep bidirectional Transformer using two objectives: Masked Language Modeling (MLM), which predicts randomly masked tokens from both left and right context, and Next Sentence Prediction (NSP). BERT established new state-of-the-art results on eleven NLP benchmarks, including pushing the GLUE score to 80.5 (a 7.7-point absolute improvement). Its bidirectional context modeling proved especially effective for understanding tasks. Subsequent variants—RoBERTa, ALBERT, DeBERTa—refined BERT's pre-training through dynamic masking, parameter sharing, and disentangled attention.

4.2 Encoder-Decoder: T5

Raffel et al. [4] proposed T5 (Text-to-Text Transfer Transformer), reformulating all NLP tasks as text-to-text problems within a unified encoder-decoder framework. By treating translation, summarization, classification, and question answering as variants of the same conditional generation problem, T5 enabled a systematic exploration of pre-training objectives, model scales, and dataset compositions using the Colossal Clean Crawled Corpus (C4). The 11B-parameter T5-XXL achieved state-of-the-art across multiple benchmarks, demonstrating that architectural simplicity combined with scale could be highly effective.

4.3 Decoder-Only: GPT Family

OpenAI's GPT series pioneered autoregressive language modeling at scale. GPT-3 [3], with 175 billion parameters, demonstrated startling few-shot and zero-shot competence: given only a handful of input-output demonstrations in the prompt, the model generalized to translation, arithmetic, code generation, and more, without gradient updates. This in-context learning capability—absent in smaller models—revealed emergent properties of scale. InstructGPT [6], trained via RLHF, improved alignment with human intent, producing outputs rated as more helpful and less harmful than the base GPT-3 model despite having far fewer parameters.

4.4 Open-Source Models: LLaMA Family

Meta AI's LLaMA [5] demonstrated that high performance is achievable with publicly available training data. LLaMA-13B surpassed GPT-3 (175B) on most benchmarks by training on more tokens relative to model size. LLaMA 2 [10] extended this with models from 7B to 70B parameters, fine-tuned with RLHF for chat applications. Llama 3 [11] further advanced the series with improved multilingual, coding, and reasoning capabilities and extended context windows up to 128K tokens. These open-weight releases catalyzed a community of derivative fine-tuned models and research extensions.

V. PRE-TRAINING AND FINE-TUNING

5.1 Pre-Training Objectives

Two principal pre-training paradigms have emerged. Masked Language Modeling (MLM), used in BERT and its variants, trains bidirectional encoders by masking 15% of input tokens and predicting them from surrounding context. Causal Language Modeling (CLM), used in GPT-style models, trains autoregressive decoders to predict each token given all preceding tokens. T5 introduced span corruption, masking contiguous spans and predicting them in the decoder.

Each objective shapes the model's inductive biases: MLM favors understanding; CLM favors generation; span corruption balances both.

5.2 Supervised Fine-Tuning

Full fine-tuning updates all model parameters on labeled downstream data. While effective, this approach is computationally prohibitive at scale: storing and serving a fully fine-tuned 70B model for each downstream task is impractical. Moreover, catastrophic forgetting—wherein fine-tuning on a specific task degrades performance on others—remains a persistent challenge.

5.3 Parameter-Efficient Fine-Tuning (PEFT)

PEFT methods address these limitations by updating only a small fraction of model parameters. Adapter modules insert small trainable bottleneck layers between Transformer blocks while freezing the backbone. Prefix tuning prepends learnable continuous tokens to the input. LoRA [12], introduced by Hu et al., is perhaps the most widely adopted PEFT method: it freezes pre-trained weight matrices and injects trainable low-rank decomposition matrices A R^{d×r} and B R^{r×k} at each attention layer. LoRA can reduce trainable parameters by 10,000-fold compared to full fine-tuning on GPT-3, with negligible performance degradation and no additional inference latency.

5.4 RLHF and Alignment

Ouyang et al. [6] demonstrated that scaling alone does not produce models aligned with human intent. Reinforcement Learning from Human Feedback

(RLHF) addresses this through a three-stage pipeline:

(1) supervised fine-tuning on curated demonstrations;
(2) training a reward model on human preference comparisons; and
(3) optimizing the language model against the reward model using Proximal Policy Optimization (PPO). InstructGPT models trained with RLHF were consistently preferred by human evaluators over the larger GPT-3 base, demonstrating that alignment is distinct from raw capability. Constitutional AI and Direct Preference Optimization (DPO) have since extended RLHF principles with improved stability and efficiency.

VI. NLP TASKS AND BENCHMARK EVALUATION

6.1 Core NLP Tasks

LLMs have achieved transformative performance across the full spectrum of NLP tasks:

- Text Classification: Sentiment analysis, topic labeling, and natural language inference are standard benchmarks where fine-tuned BERT-family models routinely exceed 90% accuracy.
- Named Entity Recognition (NER): Sequence labeling for person, organization, and location entities. Fine-tuned models approach human-level F1 scores on CoNLL-2003.
- Machine Translation: Neural sequence-to-sequence models have largely displaced phrase-based systems. Transformer-based models set BLEU scores of 41.8 on WMT 2014 En-Fr [1].
- Question Answering: Reading comprehension benchmarks such as SQuAD have been surpassed by LLMs. BERT achieved 93.2 F1 on SQuAD v1.1 [2].
- Summarization: Abstractive summarization with models like BART and PEGASUS achieves state-of-the-art ROUGE scores on CNN/DailyMail.
- Code Generation: Models such as Code LLaMA [13] and Codex achieve up to 67% on the HumanEval benchmark for Python code synthesis.

6.2 Benchmark Suites

The General Language Understanding Evaluation (GLUE) benchmark aggregates nine NLU tasks including sentiment analysis, textual entailment, and co-reference resolution. SuperGLUE introduced harder tasks requiring multi-sentence reasoning. Both benchmarks were rapidly saturated by LLMs, prompting the development of more challenging evaluations.

The Massive Multitask Language Understanding (MMLU) benchmark spans 57 academic subjects from elementary mathematics to law and medicine, probing world knowledge and reasoning. GPT-4 achieves approximately 86% on MMLU, surpassing expert human performance on many subjects. BIG-Bench (Beyond the Imitation Game Benchmark) and HellaSwag evaluate commonsense reasoning and future text prediction, revealing persistent gaps between model capability and human performance in complex inferential tasks.

VII. CHALLENGES AND LIMITATIONS

7.1 Hallucination

Hallucination—the generation of fluent but factually incorrect content—is among the most consequential limitations of LLMs [7]. LLMs produce text by predicting high-probability continuations, not by retrieving verified facts. As a result, they may confidently state fabricated statistics, invent citations, or misrepresent historical events. Zhang et al. [14] provide a comprehensive taxonomy of LLM hallucination phenomena and evaluate mitigation strategies including retrieval augmentation, chain-of-thought prompting, and factuality training objectives. Xu et al. [15] demonstrate theoretically that hallucination cannot be fully eliminated—it is an innate limitation of the probabilistic generation paradigm.

7.2 Bias and Fairness

LLMs absorb and amplify biases present in their training corpora, which are predominantly composed of internet text reflecting historical inequalities and cultural prejudices. Models exhibit gender stereotyping in professional role associations, racial bias in sentiment generation, and geographical disparities in world knowledge. Debiasing techniques include data filtering, counterfactual data augmentation, and adversarial training, but none fully resolves embedded societal biases. Evaluation frameworks such as WinoBias and StereoSet provide systematic measurements of these effects.

7.3 Computational Cost and Efficiency

Training a 175B-parameter model requires thousands of GPU-hours at substantial financial and carbon cost. Inference at scale introduces latency that limits real-time applications. Addressing this, researchers have developed quantization (reducing weight precision to 4- or 8-bit), knowledge distillation (training smaller student models from larger teachers), and sparse architectures.

Parameter-efficient methods such as LoRA [12] dramatically reduce adaptation cost. Efficient LLM research [16] surveys these techniques systematically, documenting progress toward accessible deployment.

7.4 Explainability and Interpretability

The internal representations of LLMs remain opaque, limiting accountability in high-stakes applications. Attention visualization, probing classifiers, and mechanistic interpretability have shed light on specific phenomena: attention heads specialize in syntactic roles, factual associations are localized to specific layers, and arithmetic is computed through identifiable circuits. Zhao et al. [9] survey

explainability techniques for LLMs, categorizing approaches by training paradigm and explanation type. Despite progress, comprehensive mechanistic understanding of large-scale language models remains a major open problem.

7.5 Privacy and Data Concerns

LLMs trained on internet-scale data may memorize and reproduce personally identifiable information, posing serious privacy risks. Membership inference attacks can determine whether a specific text appeared in training data. Differential privacy provides theoretical guarantees but introduces significant utility trade-offs at scale. Additionally, copyright questions around training on proprietary content remain legally unresolved in most jurisdictions, creating uncertainty for commercial deployment.

VIII. APPLICATIONS OF LLMS

8.1 Healthcare and Biomedical NLP

Clinical NLP involves processing unstructured text from medical records, discharge summaries, and research literature. Domain-adapted models such as BioBERT and ClinicalBERT achieve strong performance on biomedical named entity recognition, clinical relation extraction, and medical question answering. LLMs have demonstrated the ability to pass the US Medical Licensing Examination (USMLE), raising both promise and caution regarding clinical deployment. Hallucination in medical contexts—where fabricated treatment recommendations could harm patients—demands robust mitigation [7] before clinical use.

8.2 Education and Intelligent Tutoring

LLMs enable adaptive educational systems capable of generating personalized explanations, grading open-ended responses, and providing immediate feedback on student writing. Automated short answer grading using BERT, T5, and GPT-3 has been benchmarked across multiple datasets, with transformer-based systems approaching inter-rater human agreement. Instruction tuning with GPT-4 has been explored for generating high-quality teaching hints for programming exercises. Ethical considerations around academic integrity and over-reliance on AI-generated content require careful institutional governance.

8.3 Legal Reasoning and Contract Analysis

The legal domain presents unique challenges: long documents, specialized vocabulary, jurisdiction-specific knowledge, and high consequences for error. LegalBERT and similar domain-adapted encoders improve clause classification and contract review. GPT-4 has been evaluated on bar examination questions and demonstrates near-passing performance. Applications include contract due diligence, regulatory compliance checking, and case law summarization, though hallucination of legal citations remains a critical risk.

8.4 Code Generation and Software Engineering

Codex, Code LLaMA [13], and similar code-specialized models represent a significant advance in AI-assisted software development. These models support code completion, bug detection, test generation, and documentation synthesis. GitHub Copilot, powered by Codex, has been adopted by millions of developers. Evaluations on HumanEval and MBPP benchmarks show continuous improvement, with the best open models now exceeding 65% pass rates. Security implications—including the generation of vulnerable code patterns—require ongoing attention.

8.5 Information Retrieval and Search

Dense retrieval using LLM-derived embeddings has improved ad hoc search, semantic document matching, and question answering.

Retrieval-Augmented Generation (RAG) architectures combine retrieval with generation, grounding LLM responses in retrieved documents to reduce hallucination [17]. Models such as WikiChat [17] demonstrate 97.3% factual accuracy in simulated conversations by grounding responses in Wikipedia. RAG has become a standard component of production LLM systems, enabling knowledge currency without full model retraining.

8.6 Multimodal Applications

The success of LLMs in text has motivated extension to multimodal settings. Vision-language models such as GPT-4V and LLaVA process images alongside text, enabling visual question answering, image captioning, and optical character recognition. Audio-language models such as Whisper perform robust speech recognition, and multimodal foundation models like Gemini natively process text, images, audio, and video within a single architecture. Yin et al. [18] survey the landscape of multimodal LLMs, documenting architecture strategies, evaluation protocols, and emerging capabilities.

IX. FUTURE RESEARCH DIRECTIONS

9.1 Efficient and Green AI

The environmental cost of training frontier models has spurred research into compute-optimal training, sparse mixture-of-experts architectures, and hardware-software co-design. State-space models such as Mamba offer linear-time alternatives to quadratic self-attention for long sequences. Flash Attention and

continuous batching address GPU memory bottlenecks in inference. The democratization of LLMs depends on reducing their resource footprint without proportional capability loss.

9.2 Improved Factuality and Grounding

Reducing hallucination requires progress on multiple fronts: calibrated uncertainty estimation (so models express doubt when appropriate), retrieval augmentation for dynamic knowledge integration, and training objectives that reward factual accuracy rather than fluency alone. Self-consistency decoding and chain-of-thought prompting partially address multi-step reasoning errors. Benchmark development for factuality evaluation [7,14] will be essential to measure progress.

9.3 AI Alignment and Safety

As LLMs are deployed in consequential settings, alignment with human values becomes paramount. Beyond RLHF [6], Constitutional AI trains models using AI-generated feedback grounded in explicit principles. Scalable oversight—using AI assistance to evaluate AI outputs—addresses the challenge of supervising superhuman capabilities. Mechanistic interpretability aims to reverse-engineer model computations to enable reliable safety verification. Wang et al. [19] survey alignment methodologies comprehensively, documenting data collection strategies, training methods, and evaluation frameworks.

9.4 Long-Context and Memory Architectures

Standard Transformer attention scales quadratically with sequence length, limiting context windows to tens of thousands of tokens at reasonable cost. Extended context models (Llama 3 supports 128K tokens [11]) use positional encoding improvements such as RoPE and ALiBi. External memory architectures and retrieval mechanisms supplement parametric knowledge with dynamic, updatable information stores. These advances are critical for document-level understanding, legal contract analysis, and scientific literature synthesis.

9.5 Specialized and Domain-Adapted LLMs

General-purpose LLMs often underperform on highly specialized domains due to distributional mismatch with training data. Continued investment in domain-specific pre-training—analogous to BioBERT for biomedicine or K2 for geoscience—will produce models better suited to vertical applications. Instruction tuning on curated domain-specific demonstrations, as demonstrated by InvestLM for finance, offers a cost-effective alternative to full pre-training for domain adaptation. Ling et al. [20] provide a comprehensive survey of domain specialization techniques for LLMs.

X. CONCLUSION

This survey has traced the evolution of natural language processing from rule-based expert systems through statistical methods and deep learning to the current era of large language models. The Transformer architecture [1], with its self-attention mechanism and capacity for parallelized training at scale, proved to be the enabling innovation for this transformation. Models such as BERT [2], GPT-3 [3], T5 [4], InstructGPT [6], and LLaMA [5,10,11] have collectively advanced the state of the art across every major NLP task, demonstrated emergent capabilities absent in smaller systems, and enabled a new generation of AI-powered applications.

Yet the field faces substantial unresolved challenges. Hallucination [7,14,15] threatens deployment in high-stakes domains. Bias and fairness concerns [8] require ongoing vigilance and structural mitigation. Computational costs and carbon footprint demand continued research into efficiency [12,16].

Explainability and interpretability [9] remain foundational for trust and accountability. Alignment with human values [6,19] grows more pressing as model capabilities expand. The trajectory of the field—toward multimodal, efficient, aligned, and domain-specialized foundation models—offers substantial promise. Progress on factual grounding, long-context reasoning, and safe deployment will determine whether LLMs fulfill their potential as genuinely transformative tools for human knowledge work. We hope this survey provides a useful foundation for researchers and practitioners navigating this rapidly evolving landscape.

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
   a. Processing Systems (NeurIPS), 30.
   https://arxiv.org/abs/1706.03762
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423
3. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.,Dhariwal, P., … Amodei, D. (2020). Language models

are few-shot learners. Advances in Neural Information Processing Systems (NeurIPS), 33, 1877–1901. https://arxiv.org/abs/2005.14165
4. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S.,Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67. https://arxiv.org/abs/1910.10683
5. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux,M.-A., Lacroix, T., … Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint. https://arxiv.org/abs/2302.13971
6. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L.,Mishkin, P., … Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems (NeurIPS), 35. https://arxiv.org/abs/2203.02155
7. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., …Liu, T. (2024). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems. DOI: 10.1145/3703155
8. Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., …Liu, Q. (2023). Aligning large language models with human: A survey. arXiv preprint. https://arxiv.org/abs/2307.12966
9. Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., … Du,M. (2024). Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology. DOI: 10.1145/3639372
10. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi,A., Babaei, Y., … Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint. https://arxiv.org/abs/2307.09288
11. Grattafiori, A., Dubey, A., Jauhri, A., … Touvron, H.(2024). The Llama 3 herd of models. arXiv preprint. https://arxiv.org/abs/2407.21783
12. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S.,& Chen, W. (2022). LoRA: Low-rank adaptation of large language models. Proceedings of ICLR 2022. https://arxiv.org/abs/2106.09685
13. Rozière, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan,X., … Synnaeve, G. (2023). Code Llama: Open foundation models for code. arXiv preprint. https://arxiv.org/abs/2308.12950
14. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., … Shi, S.(2023). Siren's song in the AI ocean: A survey on hallucination in large language models. Computational Linguistics. DOI: 10.1162/coli.a.16
15. Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination isinevitable: An innate limitation of large language models. arXiv preprint. https://arxiv.org/abs/2401.11817
16. Wan, Z., Wang, X., Liu, C., Alam, S., Zheng, Y., Liu, J., …Zhang, M. (2024). Efficient large language models: A survey. Transactions on Machine Learning Research. https://arxiv.org/abs/2312.03863
17. Semnani, S. J., Yao, V. Z., Zhang, H., & Lam, M. (2023). WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia.
    a. Findings of EMNLP 2023. DOI: 10.18653/v1/2023.findings-emnlp.157
18. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E.(2024). A survey on multimodal large language models. National Science Review. DOI: 10.1093/nsr/nwae403
19. Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., …Liu, Q. (2023). Aligning large language models with human: A survey. arXiv preprint arXiv:2307.12966. https://arxiv.org/abs/2307.12966
20. [20] Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., …Zhao, L. (2024). Domain specialization as the key to make large language models disruptive: A comprehensive survey. ACM Computing Surveys. DOI: 10.1145/3764579
21. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., …Wen, J.-R. (2023). A survey of large language models. arXiv preprint. https://arxiv.org/abs/2303.18223
22. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., …Xie, X. (2023). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology. https://arxiv.org/abs/2307.03109
23. Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-task deepneural networks for natural language understanding. Proceedings of ACL 2019. DOI: 10.18653/v1/P19-1441
24. Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester,B., … Le, Q. V. (2022). Finetuned language models are zero-shot learners. Proceedings of ICLR 2022. https://arxiv.org/abs/2109.01652
25. Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A.,Khashabi, D., & Hajishirzi, H. (2023). Self-instruct: Aligning language models with self-generated instructions. Proceedings of ACL 2023. DOI: arXiv:2212.10560
26. Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y., & Wen, J.-R. (2023). HaluEval: A large-scale hallucination evaluation benchmark for large language models. arXiv preprint. https://arxiv.org/abs/2305.11747
27. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., …Wen, J.-R. (2024). A survey on large language model

based autonomous agents. Frontiers of Computer Science. DOI: 10.1007/s11704-024-40231-1

28. Patil, R. & Gudivada, V. (2024). A review of currenttrends, techniques, and challenges in large language models (LLMs). Applied Sciences, 14(5), 2074. DOI: 10.3390/app14052074
29. Huang, X., Ruan, W., Huang, W., Jin, G., Dong, Y., Wu, C.,… Mustafa, M. A. (2024). A survey of safety and trustworthiness of large language models through the lens of verification and validation. Artificial Intelligence Review. DOI: 10.1007/s10462-024-10824-0
30. Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., …Chen, E. (2024). Large language models for generative information extraction: A survey. Frontiers of Computer Science. DOI: 10.1007/s11704-024-40555-y
31. Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self-attention with relative position representations. arXiv preprint. https://arxiv.org/abs/1803.02155
32. Gao, T., Fisch, A., & Chen, D. (2021). Making pre-trainedlanguage models better few-shot learners. Proceedings of ACL-IJCNLP 2021. DOI: 10.18653/v1/2021.acl-long.295