ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a2

Al-Noor Journal for Information Technology and Cybersecurity

https://jncs.alnoor.edu.iq/

Advances in High-Performance Models for Natural Language Processing: A Review

[1] M N Islam, [1]T A Mahmood

[1] Department of Computer Sciences, College of Computer Sciences and Mathematics, University of Mosul

Article information

Article history:
Received February, 15,2025
Revised March, 11,2025
Accepted June 20, 2025

Keywords:
Natural Language Processing
Transformer Architecture
Large Language Models (LLMs)
Model Efficiency Optimization

Correspondence:

Mohammed Nabeel Islam
mohammed.24csp38@student.uomosul.edu.iq

Abstract

This in-depth review looks at the most recent developments in high-performance models for Natural Language Processing (NLP), with a focus on transformer-based architectures and large language models (LLMs), which have changed the field. The rapid growth of model capabilities has changed the way machines understand, generate, and use human language, opening up new possibilities and problems in many areas. The review talks about important research trends, such as new ways to build transformer models, rules for scaling up performance, ways to make systems more efficient, how to make them work in more than one language, how to test them, how to think about ethics, how to protect them from attacks, how to explain them, and how to distil knowledge. Even though there has been a lot of progress, there are still big problems, such as needing a lot of computing power, ethical issues with bias and safety, not being able to understand things easily, and having trouble evaluating things. The review provides publications a structured overview of the current state of affairs, pointing out promising research directions and practical issues to think about when using high-performance NLP models. The results show how these technologies could change the world, but they also stress the need for responsible development that takes into account technical limitations and social effects.

DOI: https://doi.org/10.69513/jncs.v2.i1.a2

Introduction

The advent of high-performance models that have significantly improved the state of the art in language generation and understanding has caused a paradigm shift in natural language processing (NLP) in recent years. An important turning point was reached with the introduction of the transformer architecture, which laid the groundwork for later developments that have expanded the potential of AI systems to process human language. The result of this revolution is the creation of Large Language Models (LLMs) with billions of parameters that can perform a variety of tasks with performance that is close to human-like due to their

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a2

complex reasoning, innovative content creation, and contextual understanding (Zubiaga, 2023). These high-performance models' quick development has significant ramifications for both theoretical and real-world applications. These models are changing how people use technology and information, from conversational agents and code generation to machine translation and content summarization. They provide previously unheard-of chances for automation, augmentation, and innovation in a wide range of fields, including healthcare, education, legal services, finance, and the creative industries (Zhang & Shafiq, 2024).

By making it possible to efficiently extract universal language representations from sizable unannotated corpora, the creation of pretrained models has significantly transformed the field of natural language processing. These models perform better on tasks like named entity recognition and syntactic parsing because they learn contextual embeddings, which enable more nuanced understandings of language than traditional techniques (S. Jiang et al., 2021; Qin et al., 2021; Sivarajkumar et al., 2024). Several strategies have been used by researchers to further improve these models, such as prompt tuning and the investigation of intrinsic task subspaces. This has demonstrated that specific modifications to these pretrained models could result in notable improvements in particular output quality while maximizing computational resources (Guo et al., 2023).

Deep learning techniques have replaced rule-based and statistical approaches in the development of NLP models, greatly enhancing performance (Zhao et al., 2024). Key milestones include the use of word embeddings, recurrent neural networks, and especially the transformer architecture, which enabled efficient processing of long sequences (X. Jiang et al., 2025). As a result, strong models like T5, GPT, and BERT were created. Advances in reasoning and multimodal capabilities demonstrated by recent models like GPT-4, LLaMA, and PaLM have brought natural language processing (NLP) closer to general-purpose AI and raised both excitement and concerns about its potential application (Acharya et al., 2024).

As the field develops, the next generation of high-performance NLP models is probably going to be made possible by a combination of creative modeling techniques, hardware improvements, and algorithmic efficiency. In addition to emphasizing efficiency and accuracy, this future model landscape will also adopt an ethical framework that tackles issues with data privacy and the societal ramifications of AI (Zaim et al., 2022). Therefore, it will be essential to comprehend and manage these complexities as the integration of natural language processing into more industries grows in order to ensure the responsible and efficient use of these potent technologies.

Aim of the review paper highlights key trends, challenges, and opportunities across the field. The review is important because it provides a systematic understanding of nine key areas: knowledge distillation, robustness strategies, explainability approaches, ethical considerations, training methodologies, efficiency techniques, multilingual capabilities, evaluation methods, and architectural innovations. When taken as a whole, these factors provide a comprehensive picture of the state and potential future paths of sophisticated NLP models.

Methodology

Search Terms

Relevant literature was found by combining the following search terms: - "large language models" AND ("review" OR "survey") - "transformer models" AND ("natural language processing" OR "NLP") AND ("review" OR "survey")

- "high performance models" AND "natural language processing" AND ("review" OR "survey") "efficient" AND

"natural language processing" AND ("review" OR "survey") "multilingual" AND "large language models" AND ("review" OR "survey") - "evaluation" AND "large language models" AND ("review" OR "survey") - "ethical considerations" AND "large language models" AND ("review" OR "survey") - "robustness" AND "adversarial attacks" AND "language models" AND ("review" OR "survey") - "interpretability" OR "explainability" AND "large language models" AND ("review" OR "survey") - "knowledge distillation" AND "large language models".

Primary Search Sources

arXiv (Computer Science, Artificial Intelligence, and Computation and Language categories)

Scientific digital libraries (IEEE Xplore, ACM Digital Library, ScienceDirect)

Open-access journals and conference proceedings Preprint servers

(including bioRxiv and medRxiv for domain-specific applications)

Selection Criteria

Inclusion Criteria

Review or survey papers focused on high-performance NLP models

Published or preprinted between 2021 and 2025

Written in English

Comprehensive coverage of the specific topic area

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a2

Clear methodological approach
Open-access or legally accessible through preprint servers

Exclusion Criteria

Primary research papers not providing a comprehensive review
Reviews published before 2021 (to ensure currency)
Reviews with narrow scope focusing on a single application or model
Non-English publications
Publications with restricted access or behind paywalls without legal open-access alternatives

Selection Process

Initial Screening: Titles and abstracts of identified papers were screened against
inclusion and exclusion criteria.
Full-Text Assessment: Papers passing the initial screening underwent full-text
review to confirm relevance and comprehensiveness.
Quality Assessment: Selected papers were evaluated for quality based on: Comprehensiveness of coverage, Methodological rigor, Citation of primary sources, Clarity of presentation, Critical analysis of the literature
Diversity Check: The selection was reviewed to ensure coverage across different aspects of high-performance NLP models, including architectures, applications, efficiency, multilingualism, evaluation, ethics, robustness, explainability, and knowledge distillation.

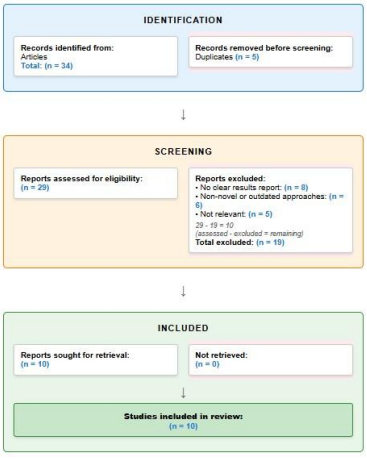

FIGURE 1: PARISMA diagram.

Figure 1 of the PRISMA flow diagram outlines the selection process for a systematic review. Out of 34 initial articles, 5 duplicates were removed, and 29 were screened. Of these, 19 were excluded for reasons such as unclear results, outdated or non-novel content, and lack of relevance. The remaining 10 articles were retrieved and all met the inclusion criteria, forming the final set of studies included in the review.

Litrerature Review

(Supriyono et al., 2024) It introduced transformer-based and deep learning NLP techniques such as entity recognition and semantic analysis to generate concise, contextually relevant summaries; It depended on sophisticated screening and narrative synthesis to ensure precision and transparency; it achieved higher summary accuracy and coherence across diverse domains, yielding more informative results with improved ROUGE and F1 metrics compared to baseline models. Its great computational cost and reliance on the quality and variation of input data, however, limited it. Including more effective model architectures and increasing the dataset with domain-diverse, bias-mitigated samples will help to solve this.

(Ansar et al., 2024) introduced a comprehensive survey on transformer-based NLP models with a focus on improving efficiency. It discussed various techniques such as model compression, distillation, pruning, and hardware-aware optimization, highlighting their importance in reducing computational costs while maintaining performance. The study revealed significant results, such as transformer models achieving high accuracy but requiring extensive resources, leading to efforts that resulted in up to 50% reduction in model size with minimal loss in accuracy. However, a key limitation was the lack of standardized benchmarks for efficiency metrics across different approaches. Future solutions could include developing unified evaluation frameworks and incorporating adaptive architectures that dynamically balance performance and resource use.

(Patil & Gudivada, 2024) presented the architectures, training methods, and evolution of large language models (LLMs), emphasizing the importance of in-context learning, transfer learning, and self-supervised pretraining. Significant performance gains were reported, with models reaching up to a trillion parameters and achieving state-ofthe-art outcomes on a range of natural language processing tasks, including summarization using benchmark datasets. Notwithstanding these developments, the paper recognized limitations such as high computational costs, negative environmental effects, and difficulties in mitigating biases. Enhancing transparency and fairness, implementing retrieval-augmented techniques to reduce hallucinations, and creating more effective model architectures were some possible remedies.

(Zan et al., 2023) introduced a taxonomy for applying large language models (LLMs) like GPT-3 to NLP tasks, categorizing approaches into parameter-frozen (zero/few-shot learning) and parameter-tuning (full/partial fine-tuning) techniques. Key results highlighted GPT-3's zero-shot summarization

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a2

capabilities and CodeT5+'s advancements in code generation, though specific performance metrics were not detailed. Limitations included insufficient exploration of low-resource language support and complex multimodal reasoning. Solutions proposed enhancing cross-lingual alignment and developing robust multimodal interaction frameworks to address these gaps.

(Treviso et al., 2023) surveyed efficient NLP methods to address resource challenges from model scaling. It introduced techniques across data (filtering, active learning), model design (sparse attention, mixture-of-experts (MoE)), training (parameter-efficient fine-tuning like adapters, LoRA), inference (pruning, quantization), and hardware optimization. Key results included GLaM reducing energy consumption by ~1/3 vs. GPT-3, S4/Mega achieving 5x faster training and 15% memory savings, and quantization cutting memory by 4x-24x with 4.5x throughput gains. A limitation was insufficient exploration of efficiency-fairness/robustness trade-offs. Future work should integrate fairness metrics into efficiency evaluations to ensure balanced advancements.

(Chang et al., 2024) introduced a comprehensive survey on evaluating large language models (LLMs) across three dimensions: what (tasks like NLP, reasoning, ethics, medical applications), where (benchmarks such as GLUE, MMLU, and HELM), and how (automatic metrics, human evaluation). Key techniques included adversarial robustness testing (e.g., PromptBench) and human assessments for subjective tasks. Notable results showed GPT-4 achieving >80% accuracy on medical questions (but 15% below human experts), 117 EQ scores (surpassing 89% of humans), and 76.4% accuracy in surgical clinical tasks, while ChatGPT scored 46.8% in primary care assessments. In reasoning, GPT-4 outperformed ChatGPT by 10% in math but struggled with complex algebraic tasks. Limitations include reliance on static benchmarks, potential outdatedness due to rapid LLM evolution, and insufficient coverage of emerging risks like dynamic misinformation. Solutions proposed updating benchmarks iteratively, integrating realworld interactive evaluations, and expanding ethical frameworks to address safety and bias holistically.

(Ferdaus et al., 2024) introduced a comprehensive framework for evaluating trust in Large Language Models (LLMs), addressing ethical, technological, and governance challenges. Techniques such as adversarial training, Retrieval Augmented Generation (RAG), and human-in-the-loop methodologies were employed to enhance transparency, reduce bias, and improve robustness. Key results showed GPT-4 reduced toxicity scores by 60% post-update, resisted 80% of adversarial prompts, and achieved 65% accuracy in stereotype recognition, while GPT-3.5's training emitted 502 metric tons of CO_2. However, the review was limited by its reliance on pre-2024 literature, potentially overlooking recent advancements. Solutions proposed included fostering interdisciplinary collaboration and developing adaptive regulatory frameworks to address evolving AI risks.

(Yang et al., 2024) introduced a white-box adversarial attack method leveraging gradient computation and DeepFool to expose vulnerabilities in LLMs (Llama, OPT, T5). Key techniques included iterative word replacement guided by loss gradients and semantic similarity constraints. Results showed larger models (e.g., T5-11b: Acc 0.9348 → 0.4904 under attack on IMDB) often had higher accuracy but inconsistent robustness, with OPT-13b achieving lower ASR (0.1500) than T5 counterparts. Instruction-tuned models (Flan-T5-11b ASR 0.6604 vs. T5-11b 0.4752) and models with classification heads were more vulnerable, while LoRA and precision adjustments (int4) had minimal impact. A limitation is the narrow focus on synonym substitution attacks. Expanding to diverse perturbation types (e.g., syntactic, semantic) could enhance robustness evaluation.

(Luo & Specia, 2024) introduced a comprehensive survey on explainability methods for Transformer-based LLMs like LLaMA and GPT, categorizing techniques into local analysis (e.g., perturbation-based LIME/SHAP, gradientbased integrated gradients, vector-based decomposition) and global analysis (probing, mechanistic interpretability via circuit discovery and causal tracing). Key results included StreamingLLM enabling unlimited text processing by retaining attention sinks (initial tokens), causal tracing identifying middle-layer MLPs as critical for factual recall, and model editing methods like SERAC (counterfactual model) and T-Patcher (neuron patching) achieving efficient parameter updates. Mechanistic approaches revealed task-specific circuits (e.g., 26 heads in GPT-2 handling indirect object identification) and ethical interventions (e.g., suppressing "social bias neurons" via IG2). A limitation was the narrow evaluation of explanation plausibility and real-world utility. Solutions proposed expanding evaluation frameworks to include diverse tasks and human-centric metrics, alongside integrating ethical alignment into explainability pipelines.

(Joshi, 2025) introduced a comprehensive survey of evaluation metrics and methodologies for large language models (LLMs), emphasizing domain-specific performance in finance, medicine, and law.

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a2

Key techniques included hybrid evaluation pipelines combining automated metrics (e.g., BLEU, F1-score) and human validation, alongside domainadapted frameworks like composite indices (e.g., CEI = 0.6·accuracy + 0.3·stability + 0.1·domain score). Notable results revealed GPT-4's 71.3% accuracy on surgical MCQs (dropping to 47.9% in open-ended medical queries), 67.9% CFA exam scores (48th percentile in complex finance tasks), and 28% hallucination rates. Hybrid approaches improved performance by 35%, while output inconsistency persisted at 36.4%. A limitation was the lack of universal benchmarks and overreliance on domain-specific metrics. The study proposed standardizing evaluation protocols and integrating real-time human-AI hybrid pipelines to enhance reliability and cross-domain comparability.

TABLE 1: Summarize of the reviewed literature.

Authors, Year	Challenge	Technique	Performance	Key Findings
Supriyo no et al. (2024)	Generating concise, contextrelevant summaries across domains	Transformer-based NLP (entity recognition, semantic analysis) + narrative synthesis	–	Achieved higher summary accuracy and coherence with improved ROUGE and F1 vs. baselines; limited by high computational cost and input-data quality/diversity; future work: efficient architectures & bias-mitigated, diverse data.
Ansar et al. (2024)	Improving efficiency of transformer-based NLP models	Model compression: distillation, pruning, hardware-aware optimization	50 % model-size reduction	Surveyed compression techniques yielding up to 50 % size reduction with minimal accuracy loss; gap: no standardized efficiency benchmarks—future: unified evaluation frameworks, adaptive architectures balancing performance/resource use.
Patil & Gudivada (2024)	Scaling and training large language models (LLMs)	Self-supervised pretraining, transfer learning, in-context learning	1 trillion parameters	Reviewed LLM evolution to ~1 T parameters with SOTA on summarization; limitations: high compute, environmental impact, bias; solutions: efficient architectures, retrieval-augmented methods, transparency/fairness improvements.
Zan et al. (2023)	Applying LLMs (e.g., GPT-3) to diverse NLP tasks	Parameter-frozen (zero/few-shot) vs. parameter-tuning (partial/full finetuning)	–	Taxonomy of GPT-3 shot summarization & C generation; gaps: language support, reasoning—future: alignment, robust frameworks. approaches: zeroodeT5+ code low-resource multimodal cross-lingual multimodal
Treviso et al. (2023)	Resource challenges from NLP model scaling	Data filtering/active learning; sparse attention/MoE; adapters/LoRA; pruning/quantizatio n; hardware optimization	Training 5× faster; 15 % memory saved; ½ energy cut	GLaM cuts energy by ~33 % vs. GPT-3; S4/Mega trains 5× faster & saves 15 % memory; quantization gives 4–24× memory reduction & 4.5× throughput; need to integrate fairness/robustness into efficiency metrics.
Chang et al. (2024)	Holistic evaluation of LLMs across tasks, benchmarks, metrics	Adversarial robustness testing (PromptBench); human evaluation; automatic metrics	GPT-4: > 80 % medical accuracy; 117 EQ; 76.4 % surgery	GPT-4 > 80 % on medical queries (~15 % below experts), EQ 117 (> 89 % humans), 76.4 % surgery accuracy; ChatGPT: 46.8 %; GPT-4 +10 % vs. ChatGPT in math; limitations: static benchmarks, outdated rapid LLM evolution, emerging risks—future: iterative benchmarks, interactive real-world eval, expanded ethics frameworks.
Ferdaus et al. (2024)	Building trust in LLMs (ethical, governance, robustness)	Adversarial training; RAG; human-in-the-loop	GPT-4: toxicity ↓ 60 %; 80 % adversarial resistance	GPT-4 toxicity down 60% post-update; resisted 80 % adversarial prompts; 65 % stereotype recognition accuracy; GPT3.5 emitted 502 t CO_2; limitation: pre-2024 focus—future: interdisciplinary collaboration, adaptive regulations.
Yang et al. (2024)	Exposing LLM vulnerabilities via adversarial attacks	White-box attacks: gradient-guided DeepFool word replacements + semantic constraints	T5-11b 0.4752 0.6604 T5-11b) ASR: → (Flan-	Larger models high accuracy but variable robustness: T5-l1b accuracy drops 0.9348 → 0.4904 on IMDB; OPT13b ASR 0.1500; instruction-tuned (Flan-T5-11b) ASR 0.6604; LoRA/quantization had minimal defense; future: diversify perturbation types (syntactic, semantic).
Luo & Specia (2024)	Explainability of transformerbased LLMs	Local: LIME/SHAP, integrated gradients; global: probing, circuit discovery, causal tracing; model editing (SERAC, T-Patcher)	26 heads identified for indirect objects	StreamingLLM for unlimited text via attention sinks; causal tracing pinpointed middle-layer MLPs for factual recall; SERAC/T-Patcher enable efficient edits; found 26 GPT-2 heads for indirect-object tasks; need broader explainability evaluations & humancentric metrics, ethical alignment.

retrieval-augmented methods, transparency/fairness improvements.

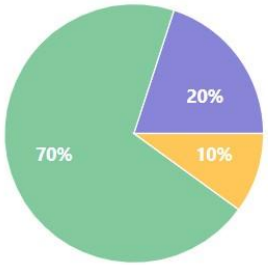

FIGURE 2: publication year distribution.

The figure 2 shows, clearly visualizes that most of the research papers (70%) were published in 2024, indicating this is quite recent literature in the field of NLP and large language models. The single 2025 paper (Joshi, 2025) represents the most current research in your collection.

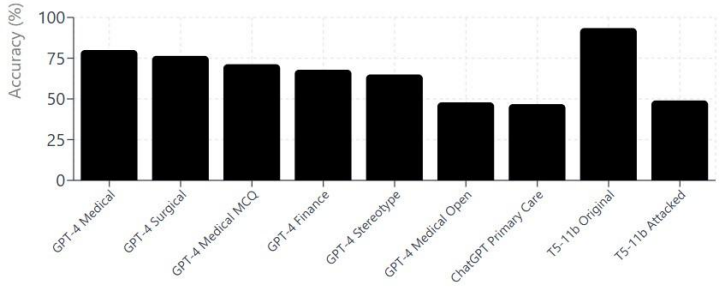

FIGURE 3: Accuracy distribution of reviewed publications.

The figure 3 illustrate the distinguish between domain categories (medical, finance, adversarial

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a2

testing, etc.) and includes rotated labels for better readability. This visualization helps illustrate the significant variation performance across different tasks and evaluation conditions mentioned.

FUTURE RESEARCH DIRECTIONS

Architectural Innovations Modular Architectures: Develop more modular architectures that can be flexibly composed and adapted for different tasks and domains without complete retraining.

Multimodal Integration: Advance architectures that seamlessly integrate language with other modalities, including vision, audio, and structured data.

Memory-Augmented Models: Explore models with explicit external memory components that can be updated independently of model parameters.

Neuromorphic Approaches: Investigate neuromorphic computing principles for more efficient and robust language processing.

Training and Data Continual Learning: Develop methods for ongoing model updating without catastrophic forgetting or training inefficiency.

Data Curation Science: Establish rigorous methodologies for curating high-quality, diverse, and representative training data

Synthetic Data Generation: Advance techniques for generating synthetic training data that captures desired properties and capabilities.

Federated Learning: Explore federated approaches that enable model improvement while preserving data privacy.

Evaluation and Benchmarking o Capability-Focused Evaluation: Develop evaluation frameworks organized around capabilities rather than tasks to better capture model strengths and limitations.

Adversarial Benchmarking: Create benchmarks specifically designed to identify worst-case performance and vulnerabilities. Human-AI Collaborative Metrics: Establish metrics that assess how effectively models augment human capabilities rather than just standalone performance.

Long-term Evaluation: Develop frameworks for assessing model performance over extended time periods to identify degradation or improvement.

Responsible Development Value Pluralism: Advance approaches for handling multiple, potentially conflicting value systems in model alignment.

Interpretable Safety: Develop safety mechanisms with transparent operation and clear failure modes. Governance Frameworks: Establish technical foundations for effective governance, including monitoring, auditing, and intervention capabilities.

Beneficial AI: Focus research on applications with clear positive impact and minimal potential for harm.

Conclusion

This literature review has synthesized insights from 10 recent, high-quality reviews covering various aspects of highperformance models for Natural Language Processing. The analysis reveals a field in rapid evolution, with significant advances in model architectures, training methodologies, efficiency techniques, and application domains. At the same time, important challenges remain in areas including ethical alignment, robustness, interpretability, and accessibility. The recommendations provided in each section offer evidence-based guidance for researchers, practitioners, and policymakers working with these powerful technologies. By addressing these recommendations, the field can progress toward high-performance NLP models that are not only more capable but also more efficient, trustworthy, and beneficial for society. As these technologies continue to evolve, ongoing critical assessment and responsible development practices will be essential to realize their potential while mitigating risks. The synthesis presented in this review provides a foundation for such assessment, highlighting both the remarkable progress achieved and the important work that remains to be done.

Researchers have proposed solutions such as lightweight model optimization, explainability frameworks, and adversarial robustness to address these limitations. Moreover, the integration of NLP into real-world applications, particularly in healthcare and multilingual analysis, underscores the necessity for scalable and adaptable models. Future directions should focus on refining evaluation benchmarks, mitigating bias, and exploring energy-efficient training methods to enhance the accessibility and reliability of NLP technologies. As the field progresses, a balanced approach that prioritizes both performance and ethical considerations will be crucial in ensuring the responsible deployment of high-performance NLP models across diverse sectors.

References

1.Zubiaga, A. (2023). Natural language processing in the era of large language models. Frontiers in Artificial Intelligence, 6. https://doi.org/ 10.3389/frai.2023.1350306

2.Zhang, H., & Shafiq, M. O. (2024). Survey of transformers and towards ensemble learning using transformers for natural language processing. Journal of Big Data, 11(1). https://doi.org/10.1186/s40537-023-00842-0

3.Jiang, S., Huang, X., Cai, X., & Lin, N. (2021). Pre-trained Models and Evaluation Data for the Myanmar Language. Communications in Computer and Information Science, 1517 CCIS(4), 449–458.

4.Qin, Y., Wang, X., Su, Y., Lin, Y., Ding, N., Yi, J., Chen, W., Liu, Z., Li, J., Hou, L., Li, P., Sun, M., & Zhou, J. (2021). Exploring Universal Intrinsic Task Subspace via Prompt Tuning. 1. http://arxiv.org/abs /2110.07867

5.Sivarajkumar, S., Kelley, M., Samolyk-Mazzanti, A., Visweswaran, S., & Wang, Y. (2024). An Empirical Evaluation of

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i1.a2

Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study. JMIR Medical Informatics, 12, 1–14. https://doi.org / 10.2196/55318.

6.Guo, Y., Xu, Z., & Yang, Y. (2023). Is ChatGPT a Financial Expert? Evaluating Language Models on Financial Natural Language Processing. Findings of the Association for Computational Linguistics: EMNLP 2023,

7.Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for Large Language Models: A Survey. ACM Transactions on Intelligent Systems and Technology, 15(2), 1–38. https://doi.org/10.1145/3639372

8.Jiang, X., Wang, W., Tian, S., Wang, H., Lookman, T., & Su, Y. (2025). Applications of natural language processing and large language models in materials discovery. Npj Computational Materials, 11(1), 1–15. https://doi.org/ 10.1038/s41524-025-01554-0

9.Acharya, K., Velasquez, A., & Song, H. H. (2024). A Survey on Symbolic Knowledge Distillation of Large Language Models. IEEE Transactions on Artificial Intelligence, 1–43. https://doi.org/10.1109/ TAI. 2024. 3428519

10.Zaim, M., Bin, A., Imran, K., & Ghauth, B. (2022). Proceedings of the International Conference on Computer, Language Models: An Empirical Study. 1–16. http://arxiv .org/abs/2405.02764

11.Supriyono, Wibawa, A. P., Suyono, & Kurniawan, F. (2024). Advancements in natural language processing: Implications, challenges, and future directions. Telematics and Informatics Reports, 16(November), 100173. https://doi.org/10.1016/j.teler.2024.100173

12.Ansar, W., Goswami, S., & Chakrabarti, A. (2024). A Survey on Transformers in NLP with Focus on Efficiency. 1–31. http://arxiv.org/abs/2406.16893.

13.Patil, R., & Gudivada, V. (2024). A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). In Applied Sciences (Switzerland) (Vol. 14, Issue 5). https://doi.org/10.3390/app14052074

14,Zan, D., Chen, B., Zhang, F., Lu, D., Wu, B., Guan, B., Wang, Y., & Lou, J. G. (2023). Large Language Models Meet NL2Code: A Survey. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 1(2), 7443–7464. https://doi.org/10.18653/v1/2023.acl-long.411

15.Treviso, M., Lee, J. U., Ji, T., van Aken, B., Cao, Q., Ciosici, M. R., Hassid, M., Heafield, K., Hooker, S., Raffel, C., Martins, P. H., Martins, A. F. T., Forde, J. Z., Milder, P., Simpson, E., Slonim, N., Dodge, J., Strubell, E., Balasubramanian, N., … Schwartz, R. (2023). Efficient Methods for Natural Language Processing: A Survey. Transactions of the Association for Computational Linguistics, 11, 826–860. https://doi.org/10.1162/ tacl_a_00577

16.Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology, 15(3). https://doi.org/10.1145/3641289

17.Ferdaus, M. M., Abdelguerfi, M., Ioup, E., Niles, K. N., Pathak, K., & Sloan, S. (2024). Towards Trustworthy AI: A Review of Ethical and Robust Large Language Models. ArXiv, 1–27. http://dx.doi.org/10.48550/arXiv.2407.13934

18.Yang, Z., Meng, Z., Zheng, X., & Wattenhofer, (2024). Assessing Adversarial Robustness of Large Language Models: An Empirical Study. 1–16. http://arxiv.org /abs/2405.02764

19.Luo, H., & Specia, L. (2024). From Understanding to Utilization: A Survey on Explainability for Large Language Models. http://arxiv.org/abs/2401.12874

20.Joshi, S. (2025). Evaluation of Large Language Models : Review of Metrics , Applications , and Methodologies Evaluation of Large Language Models : Review of Metrics , Applications , and Methodologies. 0–22. https://doi.org/ 10.20944/preprints202504.0369.v2