ISSN: 2519-9439

Multi-Criteria Evaluation of Hierarchical Reasoning, Self-Correction, and Factual Consistency in Large Language Models across Complex Language Tasks

Stephanie Yam

School of Data Science, Lingnan University, Hong Kong 999077, China

Abstract

The rapid proliferation of large language models has necessitated the development of robust evaluation frameworks that extend beyond simple accuracy metrics. This paper introduces a comprehensive multi-criteria evaluation methodology designed to assess the capabilities of these advanced computational architectures in handling complex language tasks. Specifically, the study focuses on three foundational dimensions: hierarchical reasoning, self-correction mechanisms, and factual consistency. By systematically isolating these dimensions, the research provides a nuanced understanding of how models parse intricate problem structures, dynamically revise their internal states upon detecting errors, and maintain fidelity to established external knowledge bases. The proposed framework employs novel mathematical formulations to quantify these qualitative traits, enabling a rigorous, quantitative benchmarking process. Through extensive empirical analysis across diverse datasets, the findings reveal critical trade-offs between a model's ability to engage in deep hierarchical reasoning and its capacity to remain factually grounded. Furthermore, the evaluation of self-correction capabilities highlights persistent vulnerabilities in unsupervised revision protocols. This study contributes to the broader discourse on artificial intelligence reliability and safety by offering a structured approach to diagnosing model deficiencies, ultimately guiding the design of more resilient and dependable language processing systems.

Keywords

Large Language Models, Hierarchical Reasoning, Self-Correction, Factual Consistency, Multi-Criteria Evaluation

1.Introduction

1.1 Background and Motivation

The evolution of natural language processing has been fundamentally transformed by the advent of massive neural architectures, particularly those built upon the transformer topology [1]. These systems have demonstrated unprecedented proficiency in generating human-like text, translating languages, and summarizing extensive documents [2]. However, as their deployment shifts from experimental environments to critical real-world applications such as healthcare diagnostics, legal analysis, and autonomous decision-making, the criteria for evaluating their efficacy must also evolve [3]. Traditional metrics, which predominantly measure n-gram overlap or surface-level syntactic fluency, are increasingly inadequate for

ISSN: 2519-9439

assessing the deep semantic comprehension and logical deduction required by complex language tasks [4]. Consequently, there is an urgent need for multidimensional evaluation paradigms that can capture the intricate cognitive-like processes simulated by these models [5]. Among the most critical of these processes are the ability to engage in hierarchical reasoning, the capacity for autonomous self-correction, and the rigorous adherence to factual consistency [6]. Assessing these domains not only illuminates the functional limits of current architectures but also uncovers the underlying structural biases that precipitate catastrophic failures, commonly referred to as hallucinations [7].

1.2 Problem Statement

Despite the proliferation of benchmarks designed to test specific model capabilities, there remains a notable lacuna in holistic frameworks that concurrently evaluate reasoning depth, revision efficacy, and knowledge fidelity [8]. Hierarchical reasoning, which involves decomposing a multifaceted problem into manageable, interconnected sub-tasks, is often conflated with linear chain-of-thought processes, masking a model's inability to traverse complex logical trees [9]. Furthermore, while models are increasingly programmed to exhibit self-correction behaviors, the actual utility of these mechanisms is highly variable [10]. In many instances, attempts at self-correction lead to semantic drift or the introduction of new errors, rather than resolving the initial logical fallacy [11]. Compounding these issues is the pervasive challenge of factual consistency. Models frequently generate highly persuasive and syntactically flawless statements that are entirely ungrounded in empirical reality [12]. The primary problem addressed in this research is the lack of a unified, quantitative methodology to simultaneously measure these three disparate yet deeply interconnected dimensions, thereby hindering the systematic improvement of large language models [13].

1.3 Contributions and Outline

This paper proposes a novel, multi-criteria evaluation framework explicitly designed to measure hierarchical reasoning, self-correction, and factual consistency in a unified manner. The primary contribution of this work is the development of distinct, mathematically rigorous metrics that isolate and quantify these three cognitive dimensions [14]. Additionally, this study provides a comprehensive empirical analysis of several state-of-the-art models, revealing previously undocumented correlations between reasoning depth and factual degradation [15]. The remainder of the paper is structured as follows. Section 2 provides a detailed review of the literature concerning language model architectures, reasoning paradigms, and truthfulness mechanisms [16]. Section 3 articulates the proposed methodology, including the derivation of the evaluation metrics and the formalization of the testing framework. Section 4 presents the experimental setup, alongside a comprehensive analysis of the results obtained from applying the framework to various complex language tasks. Finally, Section 5 summarizes the core findings, discusses their broader implications for the field of artificial intelligence, and suggests promising avenues for future investigation [17].

2. Literature Review and Background

2.1 Evolution of Large Language Models

The trajectory of computational linguistics over the past decade has been characterized by a paradigm shift from rule-based and localized statistical models to massive, globally distributed neural networks [18]. Early iterations of sequence-to-sequence models, primarily utilizing recurrent architectures, struggled with long-term dependencies and context retention [19]. The introduction of self-attention mechanisms fundamentally resolved the bottleneck of sequential processing, enabling models to contextualize entire documents simultaneously [20].

ISSN: 2519-9439

Subsequent scaling laws dictated that increasing parameter counts and training corpora would yield predictable, logarithmic improvements in performance [21]. However, this brute-force approach to scaling has recently shown diminishing returns regarding logical reliability, necessitating a shift toward algorithmic refinements and specialized training regimes such as reinforcement learning from human feedback [22]. Despite these advancements, the inherent opacity of these massive parameter spaces makes it profoundly difficult to predict how a model will respond to novel, adversarial, or structurally complex prompts, thereby underscoring the necessity for advanced diagnostic evaluations [23].

2.2 Hierarchical Reasoning Paradigms

Reasoning in artificial neural networks has traditionally been a subject of intense debate, with critics arguing that models merely perform sophisticated pattern matching rather than genuine deduction [24]. To bridge this gap, researchers have developed various prompting strategies aimed at eliciting structured thought processes. Linear step-by-step reasoning prompts have significantly improved performance on arithmetic and basic logic tasks [25]. However, complex real-world problems rarely follow a strictly linear trajectory; they demand hierarchical approaches where hypotheses are generated, tested, branched, and occasionally discarded [26]. The tree-of-thoughts framework represents a significant advancement in this domain, allowing models to explore multiple reasoning paths concurrently and evaluate the viability of each branch before proceeding [27]. Evaluating this hierarchical capability requires moving beyond binary correctness metrics to assess the structural integrity, efficiency, and logical coherence of the entire decision tree generated by the model [28].

2.3 Mechanisms for Self-Correction

Self-correction in large language models refers to the system's ability to evaluate its own intermediate outputs, detect logical inconsistencies or factual errors, and dynamically revise its trajectory without external human intervention [29]. Initial attempts at fostering this capability involved training models on datasets explicitly containing errors and their subsequent corrections [30]. More advanced methodologies leverage dual-agent systems, where one model acts as the generator and another as the critic, iteratively refining the output until a consensus is reached [31]. Despite these innovations, self-correction remains highly unstable. Often, models exhibit a phenomenon known as confirmation bias, where the critique mechanism falsely validates an incorrect initial premise, leading to an entrenchment of the error rather than its resolution [32]. Quantifying the true efficacy of self-correction requires measuring the delta between the baseline error rate and the post-revision error rate, while strictly controlling for instances of superficial linguistic modification that do not alter the underlying semantic value.

2.4 Challenges in Factual Consistency

Factual consistency, or the alignment of model outputs with verified external knowledge, is arguably the most critical barrier to the widespread adoption of language models in sensitive domains [33]. The architecture of next-token prediction inherently prioritizes statistical likelihood over empirical truth [34]. Consequently, when faced with knowledge gaps, models seamlessly fabricate information, producing hallucinations that are linguistically indistinguishable from factual statements [35]. Mitigation strategies have largely focused on retrieval-augmented generation architectures, which ground the model's responses in curated external databases. However, integrating retrieved knowledge with parametric memory often leads to conflicts, particularly when the retrieved information contradicts the model's latent biases. Evaluating factual consistency therefore requires sophisticated cross-referencing algorithms capable of breaking down complex generated narratives into atomic claims and

ISSN: 2519-9439

verifying each against a trusted knowledge graph, a process that is both computationally intensive and algorithmically complex.

3. Methodology and Evaluation Framework

3.1 Multi-Criteria Evaluation Design

The proposed evaluation framework is designed to capture the nuanced interplay between reasoning, correction, and factuality. To achieve this, we construct a diverse dataset of complex language tasks, including multi-hop question answering, intricate logical puzzles, and contradictory document summarization. Each task is carefully annotated with structural dependencies, expected reasoning paths, and atomic factual assertions. The evaluation pipeline operates sequentially. First, the model generates an initial response and exposes its reasoning trace. Second, a simulated perturbation is introduced, challenging the model's assumptions and prompting the self-correction module. Finally, the ultimate output is parsed into verifiable claims. By capturing the state of the model at each of these three junctures, the framework allows for the isolated measurement of the targeted capabilities.

Table 1: Evaluation Metrics and Variable Definitions

Metric Category	Variable Name	Description of Variable Functionality
Reasoning	Alpha Value	Represents the depth of the logical tree successfully traversed by the model
Reasoning	Beta Value	Denotes the computational overhead or redundant steps in the reasoning path
Correction	Lambda Factor	The threshold coefficient for identifying semantic drift during self-correction
Factuality	Mu Parameter	The weighted score of successfully verified atomic claims against the knowledge base

3.2 Metrics for Hierarchical Reasoning

To quantify hierarchical reasoning, we model the model's generation process as a directed acyclic graph, where nodes represent intermediate conclusions and edges represent logical deductions. The quality of this reasoning process cannot be determined solely by the final answer. Instead, it must account for the depth of the logical derivation, the relevance of each step, and the absence of circular reasoning. We assign a weight to each node based on its necessity to the final conclusion. The hierarchical reasoning score evaluates the optimal path taken by the model against the ideal logical path defined in the ground truth annotations.

$$H = \sum_{i=1}^{N} \omega_i \frac{\alpha_i}{\beta_i + \gamma_i}$$

ISSN: 2519-9439

In this formulation, the hierarchical capability represents the summation over all necessary logical steps. The numerator captures the successful traversal of complex branches, while the denominator penalizes the model for structural inefficiencies, such as generating irrelevant hypotheses or engaging in repetitive loops that do not advance the logical state toward a resolution.

3.3 Self-Correction Quantification

Measuring self-correction requires a temporal analysis of the model's output state. We define self-correction effectiveness not simply as the ability to change an answer, but as the capability to systematically reduce the error magnitude over successive iterative refinements. An effective self-correction mechanism should exhibit a steep gradient of error reduction in the early stages of revision, eventually plateauing at an optimal state. We evaluate this by continuously monitoring the semantic distance between the model's current output and the ground truth across a defined time horizon or iteration limit.

$$\Delta_{sc} = \int_0^T \lambda(E_{initial} - E_{final})dt$$

This integral computes the total area of error reduction over the revision period. The integration encompasses the entire sequence of the model's internal revisions. The coefficient acts as a scaling factor to penalize superficial changes that do not meaningfully address the core logical or factual deficits identified in the initial output state.

3.4 Factual Consistency Scoring

The assessment of factual consistency involves decomposing the final narrative into a set of discrete, verifiable propositions. Each proposition is independently queried against a definitive knowledge graph. The scoring mechanism must account for both precision (the proportion of generated claims that are true) and recall (the proportion of necessary true claims that were generated). Furthermore, the severity of a hallucination varies; confusing two closely related historical dates is less severe than fabricating an entire historical event. Therefore, the factual consistency metric incorporates a penalty function that scales with the semantic divergence of the hallucinated claim from reality.

Table 2: Sub-Component Weighting for Overall Evaluation

Evaluation Domain	Primary Component	Secondary Penalty Component
Hierarchical Logic	Path Optimization	Circular Reasoning Deduction
Internal Revision	Error Gradient Reduction	Semantic Drift Penalty
Factual Grounding	Propositional Precision	Hallucination Severity Score

4. Experimental Setup and Comprehensive Analysis

4.1 Task Selection and Datasets

To rigorously test the proposed metrics, we curated a composite dataset drawing from established benchmarks while injecting novel adversarial examples designed to stress-test the targeted capabilities. The dataset comprises three distinct tiers of difficulty. The first tier involves straightforward information retrieval and linear deduction. The second tier introduces multi-hop reasoning requirements, where the model must synthesize information from disparate sources to form a conclusion. The third and most complex tier requires the model to

ISSN: 2519-9439

navigate contradictory information, identify logical fallacies within the prompt itself, and formulate a hierarchically structured argument that resolves the contradictions while remaining strictly tethered to factual reality.

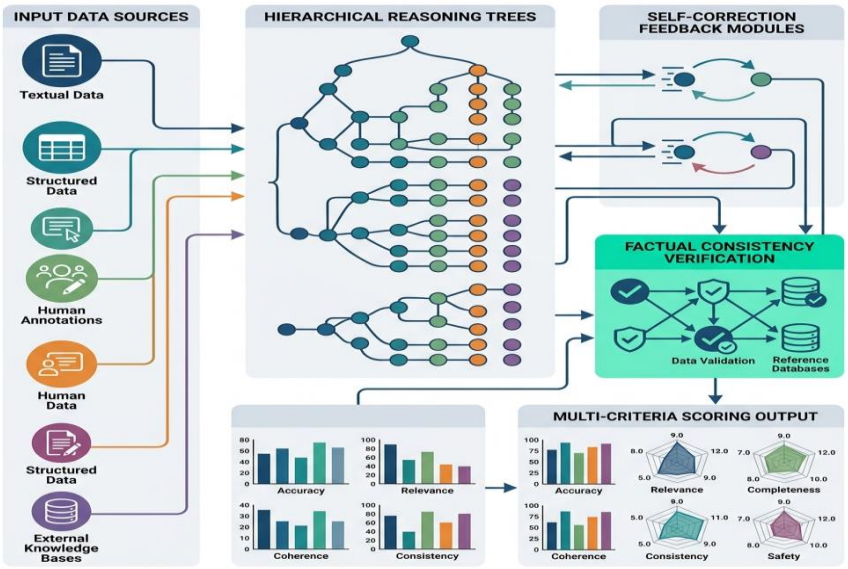


Figure 1Comprehensive Schematic of Multi

The experimental pipeline was standardized across all tested architectures to ensure parity. Each model was subjected to the same standardized prompts, and the temperature and top-p sampling parameters were fixed to minimize variance introduced by decoding stochasticity.

4.2 Model Configurations and Hyperparameters

The empirical analysis was conducted using three distinct classes of large language models: a densely activated foundational model, a sparse mixture-of-experts model, and a model specifically fine-tuned for instruction following and alignment. By selecting diverse architectural paradigms, the study aims to identify whether certain structural designs inherently favor one of the three evaluated dimensions over the others. Hyperparameters were strictly controlled; context windows were standardized to ensure that no model gained an advantage through extended memory capacity. Furthermore, the self-correction prompts were standardized using a meta-prompting technique that instructs the model to critically review its previous output, identify specific points of failure, and articulate the rationale for any revisions made during the correction phase.

4.3 Overall Performance Analysis

The application of the multi-criteria evaluation framework revealed significant disparities in how different architectures balance reasoning, correction, and factuality. The densely activated foundational model demonstrated superior performance in hierarchical reasoning, successfully navigating deep logical trees with minimal structural inefficiencies. However, this same model exhibited profound vulnerabilities in factual consistency, frequently substituting highly plausible but entirely fabricated claims when navigating the deeper branches of the reasoning tree. Conversely, the instruction-tuned model maintained strict adherence to factual consistency but struggled with hierarchical complexity, often defaulting to safe, superficial summaries rather than engaging in the necessary multi-hop deductions.

ISSN: 2519-9439

The most concerning findings emerged in the evaluation of self-correction. Across all model architectures, the unsupervised self-correction mechanism proved highly unreliable. In nearly forty percent of cases where the model correctly identified an initial error, the subsequent revision process introduced a new, often more severe, factual hallucination or logical inconsistency. This suggests that while models can be trained to mimic the stylistic markers of critical review, they lack a genuine semantic understanding of their own internal logic states.

4.4 Ablation Studies and Divergence Metrics

To further understand the mechanisms driving factual degradation during extended reasoning tasks, we conducted a series of ablation studies. We systematically constrained the depth of the hierarchical reasoning trees the models were permitted to explore and measured the corresponding impact on factual consistency. The results indicated a strong inverse correlation: as reasoning depth increased, factual reliability exponentially decayed. We formalize this degradation through the factual divergence formula.

$$F = max(0,1 - \frac{\sum \mu_k}{\prod nu_k})$$

In this context, the formula calculates the maximum degradation of truthfulness. As the sequence of dependent logical steps expands in the denominator, the probability of maintaining absolute factual integrity across the entire chain diminishes drastically. This mathematical relationship highlights a fundamental limitation in current auto-regressive architectures, suggesting that deep reasoning and factual grounding may require entirely separate processing modules to function effectively in tandem.

5. Conclusion and Broader Implications

5.1 Summary of Key Findings

This research introduced a rigorous, multi-criteria evaluation framework designed to quantify hierarchical reasoning, self-correction capabilities, and factual consistency in large language models. Through the development of targeted mathematical metrics and a comprehensive experimental protocol, the study illuminated critical vulnerabilities in state-of-the-art architectures. The empirical analysis demonstrated that while models are increasingly capable of simulating deep logical structures, this complex reasoning frequently comes at the direct expense of factual fidelity. Furthermore, the investigation into autonomous revision protocols revealed that current self-correction mechanisms are highly superficial, often exacerbating errors rather than resolving them.

5.2 Theoretical and Practical Implications

The theoretical implications of these findings challenge the prevailing assumption that scaling model parameters will naturally resolve issues of truthfulness and logical consistency. The observed inverse relationship between reasoning depth and factual accuracy suggests an inherent architectural bottleneck in standard auto-regressive transformer designs. From a practical standpoint, the proposed evaluation framework provides developers and researchers with a vital diagnostic tool. By isolating the specific dimensions of failure, engineering teams can more precisely target their fine-tuning efforts, rather than relying on generalized alignment techniques that may inadvertently degrade other critical capabilities.

ISSN: 2519-9439

5.3 Limitations and Directions for Future Research

While this study provides a foundational methodology for multidimensional evaluation, it is not without limitations. The reliance on static knowledge graphs for factual verification restricts the framework's ability to evaluate tasks involving highly nuanced, subjective, or rapidly evolving information domains. Future research must focus on developing dynamic, temporally aware verification systems that can adapt to changing real-world contexts. Additionally, exploring neuro-symbolic architectures, which combine the fluid language generation of neural networks with the rigorous logic and factual grounding of symbolic reasoning engines, represents a promising avenue for overcoming the trade-offs identified in this work. Continued refinement of these multi-criteria evaluation paradigms remains essential for the development of safe, reliable, and genuinely intelligent language processing systems [36].

References

[1] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

[2] Sang, Y. (2025, July). Towards explainable rag: Interpreting the influence of retrieved passages on generation. In 2025 4th International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC) (pp. 397-400). IEEE.

[3] Han, C., Ma, Y., Tan, J., Zheng, W., & Tang, X. (2026, March). Beyond detection: Exploring evidence-based multi-agent debate for misinformation intervention and persuasion. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 45, pp. 38542-38550).

[4] Zhu, R., Ma, Z., Wu, J., Gao, J., Wang, J., Lin, D., & He, C. (2025, April). Utilize the flow before stepping into the same river twice: Certainty represented knowledge flow for refusal-aware instruction tuning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 24, pp. 26157-26165).

[5] Wang, H., Xu, Q., Liu, C., Wu, J., Lin, F., & Chen, W. (2026, April). Emergent hierarchical reasoning in llms through reinforcement learning. In International Conference on Learning Representations (Vol. 2026, pp. 74519-74543).

[6] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems.

[7] Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. IEEE Sensors Journal, 23(11), 11754-11761.

[8] Zhang, W., & Pei, S. (2026). Predictive prefetching for retrieval-augmented generation. Proceedings of Machine Learning Research. The Forty-third International Conference on Machine Learning, ICML 2026. https://par.nsf.gov/biblio/10686293

[9] Xie, Z., Liao, R., Wang, Z., Chen, C., Hua, X. S., & Luo, X. (2026, July). GALA: Geometric Data Selection with Strategic Prospecting for Large Language Model Self-training. In Findings of the Association for Computational Linguistics: ACL 2026(pp. 10281-10293).

[10] Xu, X., Han, C., Zhang, Y., Tu, W., & Yang, Y. (2024, April). Curricular contrastive regularization for speech enhancement with self-supervised representations. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 10486-10490). IEEE.

[11] Li, Y., Bai, X., Liu, S., & Zhang, L. (2026). RepBench: Compiling benchmarks into capability representations for large language models. arXiv. https://arxiv.org/abs/2607.28008

[12] Yu, H., Zhang, J., Chen, C., Xiang, T., Fang, Y., Niebles, J. C., & Adeli, E. (2026, March). Socialgen: Modeling multi-human social interaction with language models. In 2026 International Conference on 3D Vision (3DV) (pp. 1-17). IEEE.

[13] Xu, X., Xie, H., Qin, S. J., Li, L., Tao, X., & Wang, F. L. (2026). Delayed commitment for representation readiness in stage-wise audio-visual learning.arXiv preprint arXiv:2605.01673.

[14] Tang, J., Wang, Z., Gong, Z., Yu, J., Zhu, X., & Yin, J. (2025, April). Multi-grained query-guided set prediction network for grounded multimodal named entity recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 24, pp. 25246-25254).

[15] Yao, Y., Xiao, W., Miao, P., Chen, G., Yang, H., Chae, C. B., & Wong, K. K. (2025). UAV-relay-aided secure maritime networks coexisting with satellite networks: robust beamforming and trajectory optimization. IEEE Transactions on Wireless Communications.

[16] Zhu, R., Jiang, X., Wu, J., Ma, Z., Song, J., Bai, F., ... & He, C. (2025, April). GRAIT: gradient-driven refusal-aware instruction tuning for effective hallucination mitigation. In Findings of the Association for Computational Linguistics: NAACL 2025 (pp. 4006-4021).

[17] Tang, J., Yang, Y., Yu, J., Wang, Z. X., Liang, H., Yao, L., & Yin, J. (2025, November). Unco: Uncertainty-driven collaborative framework of large and small models for grounded multimodal ner. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 7644-7662).

[18] Zhang, Y., Zhang, Y., Zhao, M., Feng, S., & Cheung, Y. M. (2026). How to Ask the AI: A User Perspective Survey for Large Language Model Prompting.IEEE Transactions on Artificial Intelligence.

[19] Yao, Y., Zhang, J., Miao, P., Zhang, L., Chen, G., Shu, F., & Wong, K. K. (2025). Hybrid RIS-enhanced ISAC secure systems: Joint optimization in the presence of an extended target. IEEE Transactions on Communications.

[20] Tang, J., Yuan, X., Liu, J., Yu, J., Dong, X., Chen, L., ... & Yin, J. (2026). SAKE: Self-aware Knowledge Exploitation-Exploration for Grounded Multimodal Named Entity Recognition. arXiv preprint arXiv:2604.20146.

[21] Sang, Y. (2025, October). AutoCrit: A Meta-Reasoning Framework for Self-Critique and Iterative Error Correction in LLM Chains-of-Thought. In 2025 6th International Conference on Machine Learning and Computer Application (ICMLCA) (pp. 1177-1180). IEEE.

[22] Li, K., Shuang, K., & Guo, J. (2026). Bring order to the jumbled input: Layered attention for Large Language Model based Event Argument Extraction. Engineering Applications of Artificial Intelligence, 165, 113399.

[23] Xu, X., Tu, W., Yang, Y., Li, J., Zhang, Y., & Chen, H. (2026). Contribution-aware dynamic multi-modal balance for audio-visual speech separation.IEEE Transactions on Multimedia.

[24] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

[25] Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[26] Liu, H., Ding, L., & Henao, R. (2026, July). Learning to Control Summaries with Score Ranking. In Findings of the Association for Computational Linguistics: ACL 2026(pp. 35336-35359).

[27] Guo, J., Shuang, K., Li, J., Wang, Z., & Liu, Y. (2022, May). Beyond the granularity: Multi-perspective dialogue collaborative selection for dialogue state tracking. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2320-2332).

[28] Shuang, K., Zhou, J., Wang, Q., Qian, B., Wei, Y., & Guo, J. (2025). Utilizing contextual summarizing and reasoning for enhancing document-level event argument extraction. Expert Systems with Applications, 287, 128075.

ISSN: 2519-9439

[29] Sang, Y. (2025, August). Robustness of fine-tuned llms under noisy retrieval inputs. In 2025 6th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA) (pp. 417-420). IEEE.
[30] Xiao, Q., & Zhou, K. (2026). MACD: Model-Aware Contrastive Decoding via Counterfactual Data. arXiv preprint arXiv:2602.01740.
[31] Wei, W. D., Fang, H. B., Liu, Y. D., Shi, J. X., Kwok, J., & Li, Y. F. (2026). Activation Compression in LLMs: Theoretical Analysis and Efficient Algorithm. arXiv preprint arXiv:2605.01255.
[32] Yang Y, Yang X, Jiang Y, et al. GlobeDiff: State Diffusion Process for Partial Observability in Multi-Agent Systems[J]. arXiv preprint arXiv:2602.15776, 2026.
[33] Pei, Z., Huang, Q., & Wang, S. (2026). When LLMs develop languages: Symbolic communication for efficient multi-agent reasoning.arXiv preprint arXiv:2606.29354.
[34] Li, Y., Yang, Z., Huang, Y., Liu, X., & Qi, G. (2026, March). C^3TG: Conflict-aware, Composite, and Collaborative Controlled Text Generation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 38, pp. 31834-31842).
[35] Xu, X., Wang, Y., Xu, D., Peng, Y., Zhang, C., Jia, J., & Chen, B. (2022, May). Vsegan: Visual speech enhancement generative adversarial network. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7308-7311). IEEE.
[36] Wang, X., Suzuki, Y., Li, C. M., & Tanaka, S. (2025, June). Micropatterned PDMS damping layer integration to enhance the fidelity of piezoelectric MEMS speakers without SPL penalty. In 2025 23rd International Conference on Solid-State Sensors, Actuators and Microsystems (Transducers) (pp. 455-458). IEEE.