Algorithms and Intelligence

Vol 2 Issue 1 (2026)

www.algointel.org

Preference-Aligned Passage Reranking Without Decoding for Efficient Retrieval-Augmented Generation

Alison Charlene, Jongmin Ko, Keisuke Endo

Business School, University of Auckland, Auckland, New Zealand

Abstract

Retrieval-augmented generation has emerged as a cornerstone paradigm for mitigating hallucinations and expanding the factual boundaries of large language models. However, the efficacy of the generation stage remains fundamentally bounded by the quality of the retrieved context. Consequently, passage reranking systems are routinely employed as an intermediate step to refine the output of first-stage retrievers. Modern high-performance rerankers increasingly rely on large language models utilizing autoregressive decoding to score passage relevance, a process that introduces severe latency bottlenecks and prohibitive computational overhead. To address this critical inefficiency, this paper introduces a novel methodology for preference-aligned passage reranking that entirely circumvents the decoding phase. By leveraging human and artificial intelligence preference alignment techniques directly on the internal hidden states of the model, our approach projects high-dimensional contextual representations into a calibrated preference space using a single forward pass. This decoding-free architecture eliminates the sequential token generation bottleneck while preserving the sophisticated reasoning capabilities of large-scale models. Through comprehensive empirical evaluations across multiple benchmark datasets, we demonstrate that the proposed framework achieves retrieval performance commensurate with state-of-the-art generative rerankers but operates with an order-of-magnitude reduction in latency. The findings suggest that direct preference optimization of encoder representations offers a highly scalable solution for enterprise-grade retrieval-augmented generation systems, fundamentally altering the trade-off between retrieval accuracy and computational efficiency.

Keywords

Retrieval-Augmented Generation, Passage Reranking, Preference Alignment, Efficiency, Natural Language Processing..

INTRODUCTION

1.1 Background on Retrieval-Augmented Generation

The rapid evolution of natural language processing has been largely driven by the scaling of foundational models, which exhibit remarkable capabilities in comprehension, reasoning, and

generation. Despite these advancements, parametric language models remain constrained by their static training corpora, leading to temporal degradation of knowledge and a persistent propensity to generate plausible but factually incorrect statements, phenomena widely documented in recent literature [1]. Retrieval-augmented generation addresses these fundamental limitations by decoupling the storage of factual knowledge from the parametric weights of the model. In a standard retrieval-augmented generation pipeline, a user query is first processed by a retriever, which searches an external knowledge corpus to identify a set of potentially relevant documents. These documents are subsequently prepended to the query and fed into a generative language model, grounding the generation process in external, verifiable facts [2]. The success of this paradigm hinges almost entirely on the relevance and information density of the retrieved context [3]. If the provided documents lack the necessary information or contain distracting noise, the generation quality degrades significantly, regardless of the sophistication of the generative model. Consequently, ensuring the highest possible quality of retrieved context has become a central objective in the design of modern artificial intelligence systems.

1.2 The Bottleneck of Decoding in Reranking

First-stage retrievers, predominantly utilizing dual-encoder architectures for dense vector similarity search, prioritize computational efficiency over deep contextual understanding to scan millions of documents in milliseconds. Because these models process queries and documents independently, they inherently lack the capacity to model the complex cross-attentional dynamics required for nuanced relevance judgment [4]. To bridge the gap between fast but coarse retrieval and the stringent contextual requirements of the generation phase, passage reranking is employed. Historically, this intermediate stage utilized cross-encoder models that concatenate the query and document, allowing for rich, bidirectional self-attention. More recently, the field has gravitated toward repurposing massive, decoder-only large language models as rerankers, prompting them to output a relevance score or a categorical label through autoregressive decoding [5]. While these generative rerankers achieve state-of-the-art accuracy, they impose a crippling computational burden. Generating tokens autoregressively requires sequential matrix multiplications where each step depends on the previous output, preventing parallelization and resulting in latency that scales linearly with the number of generated tokens. For real-world systems attempting to rerank dozens or hundreds of passages per query, this decoding bottleneck frequently violates strict latency budgets [6]. The operational costs associated with hosting such computationally intensive models further limit their accessibility and scalability for widespread deployment [7].

1.3 Proposed Preference-Aligned Approach

To overcome the severe inefficiencies associated with generative reranking, we propose a paradigm-shifting approach that maintains the robust reasoning capabilities of large models while entirely eliminating the autoregressive decoding step. Our framework, termed preference-aligned passage reranking without decoding, introduces a methodology for extracting relevance signals directly from the final hidden states of a language model following a single forward pass. By bypassing the language modeling head and the subsequent token-by-token generation, the computational complexity is drastically reduced [8]. However, raw hidden states are not inherently calibrated to reflect nuanced human judgments of passage relevance. To bridge this semantic gap, we adapt preference alignment methodologies traditionally used to align generative behaviors with human intent. By utilizing datasets comprised of chosen and rejected query-document pairs, we fine-tune the hidden state projections using a margin-based contrastive loss, explicitly aligning the continuous representation space with human relevance preferences. This alignment process ensures that the scalar values derived from the hidden states accurately reflect the comparative utility of different passages for a given query. The resulting system offers a seamless integration path for existing retrieval-augmented generation pipelines, providing the accuracy of a massive

language model reranker at a fraction of the computational latency, thereby enabling high-throughput, enterprise-grade deployments without sacrificing factual grounding or generation quality.

2. Literature Review

2.1 Evolution of Retrieval-Augmented Generation Systems

The conceptual foundations of augmenting neural networks with external memory can be traced back to early architectures designed for question answering and memory networks. However, the modern instantiation of retrieval-augmented generation gained critical momentum with the introduction of end-to-end architectures that jointly optimized a dense retriever and a sequence-to-sequence generator [9]. Early dense retrieval models revolutionized the field by demonstrating that deep contextual representations could outperform traditional sparse lexical matching algorithms like term frequency-inverse document frequency [10]. As generative language models scaled in parameter count and capability, the necessity for joint training diminished, giving rise to modular retrieval-augmented generation frameworks where off-the-shelf retrievers and generators could be dynamically combined [11]. The flexibility of this modular approach accelerated adoption, but it also exposed the fragility of large language models when presented with suboptimal context. Researchers observed that the presence of irrelevant information could actively distract the generator, leading to a phenomenon where the model ignores its parametric knowledge in favor of erroneous retrieved context [12]. Subsequent studies focused on optimizing the context window size and exploring mechanisms to filter noise prior to generation [13]. Despite improvements in long-context modeling, the quadratic complexity of self-attention mechanisms ensures that feeding excessively large numbers of retrieved documents directly into the generator remains computationally prohibitive, reaffirming the indispensable role of highly accurate retrieval and reranking stages [14].

2.2 Passage Reranking Methodologies

Passage reranking has historically served as a vital component in multi-stage search architectures. Initially dominated by feature-based machine learning models utilizing tree ensembles over lexical signals, the paradigm shifted dramatically with the advent of transformer-based architectures. The introduction of cross-encoders established a new benchmark for relevance scoring by allowing exhaustive, token-level attention between the query and the document [15]. While highly accurate, the computational cost of processing every query-document pair through a deep transformer limits cross-encoders to reranking only a small subset of candidates. To mitigate this, late interaction models were developed, which precompute token-level representations and perform a lightweight interaction step during the query phase [16]. These models strike a balance between the efficiency of dual-encoders and the accuracy of cross-encoders [17]. More recently, the field has witnessed a surge in zero-shot and few-shot reranking methodologies leveraging large language models. By formulating relevance assessment as a prompt-based generation task, researchers have achieved unprecedented accuracy on zero-shot benchmarks [18]. However, as previously highlighted, these generation-based rerankers suffer from severe latency issues, prompting an urgent search for methods that decouple the reasoning power of large language models from the bottleneck of autoregressive generation.

2.3 Preference Alignment in Large Language Models

The alignment of artificial intelligence systems with human values and intentions has become a central focus of contemporary machine learning research. The most prominent technique in this domain is reinforcement learning from human feedback, which trains a reward model on human preference data and subsequently optimizes the language model policy to maximize this reward [19]. While highly effective at shaping the tone, safety, and helpfulness of generative models, reinforcement learning from human feedback is notoriously complex and unstable to train. In

response, direct preference optimization emerged as a mathematically elegant alternative that bypasses the need for a separate reward model by directly optimizing the language model policy using a classification loss over preference pairs [20]. This breakthrough has democratized the alignment process, allowing researchers to efficiently fine-tune models to exhibit desired behaviors [21]. While these alignment techniques have been extensively applied to open-ended text generation, their application to information retrieval and passage ranking remains largely underexplored. A limited number of recent studies have attempted to use preference data to improve the ranking capabilities of language models, but they have primarily maintained the autoregressive generation paradigm [22]. By failing to rethink the architectural mechanism of scoring, these early attempts yield highly accurate models that are nonetheless too slow for practical retrieval-augmented generation pipelines [23].

2.4 Decoding-Free Inference Techniques

The computational expense of large language model inference has catalyzed a distinct subfield of research dedicated to accelerating model execution without sacrificing performance. Techniques such as quantization, distillation, and speculative decoding have shown significant promise in reducing memory footprints and increasing token generation throughput [24]. However, these methods still fundamentally operate within the sequential generation paradigm. A parallel track of research has investigated entirely decoding-free inference for specific classification and regression tasks. By appending specific pooling layers or linear classifiers to the final hidden states of encoder-decoder or decoder-only models, researchers have demonstrated that it is possible to extract rich semantic signals without generating text [25]. In the context of embedding generation, contrastive learning techniques have been successfully adapted to train models that output high-quality dense vectors directly from the final layer representations [26]. Despite these advancements, the integration of rigorous preference alignment protocols with decoding-free inference architectures specifically for the task of passage reranking represents a significant gap in the literature. This paper addresses this gap by synthesizing these disparate research threads into a cohesive methodology that leverages the best aspects of large language model reasoning, preference alignment, and efficient inference.

3. Methodology

3.1 System Architecture Overview

The core objective of the proposed architecture is to evaluate the relevance of a candidate passage given a specific user query using the deep contextual understanding of a large language model, while entirely avoiding the latency associated with autoregressive token generation. Standard generative rerankers format the query and passage into a specific prompt template, append an instruction requesting a relevance assessment, and execute a forward pass through the transformer blocks to predict the probabilities of subsequent tokens. The model then generates tokens sequentially, which are subsequently parsed to extract a numerical or categorical relevance score [27]. Our architecture completely excises this generative phase. The model receives the concatenated query and passage, formatted using a specialized instruction template to activate the model's zero-shot or few-shot reasoning capabilities. We process this combined sequence through the stacked transformer layers in a single, parallelized forward pass. Instead of passing the final hidden states through the language modeling head to project into the vocabulary space, we isolate the hidden state vector associated with the final token of the input sequence. This vector serves as a dense, high-dimensional representation of the semantic interaction between the query and the passage. By applying a trained linear projection layer to this specific hidden state, we map the complex multidimensional representation down to a single scalar value. This scalar value represents the unnormalized relevance score of the passage, allowing the system to process large

batches of candidate documents simultaneously and sort them based on this directly computed metric.

3.2 Constructing Preference-Aligned Representations

The transition from a generative scoring mechanism to a direct hidden state projection necessitates a sophisticated training regimen to ensure the output scalars accurately reflect nuanced relevance judgments. We achieve this through the careful construction of preference-aligned representations. The foundational concept relies on assembling a vast dataset of queries, each associated with at least one highly relevant passage and one highly irrelevant or distracting passage. These sets form the basis of our preference pairs, simulating a scenario where a human evaluator explicitly prefers one document over another for answering a specific query [28]. The challenge in modern retrieval-augmented generation systems is that documents are rarely purely relevant or entirely irrelevant; they often contain varying degrees of partial information or conflicting facts. Therefore, our dataset construction process involves not only leveraging traditional search benchmarks but also utilizing large teacher models to generate synthetic preference labels based on detailed grading rubrics [29]. We present the model with a pair of documents for a single query. The model processes each query-document concatenation independently through its transformer blocks. By isolating the final hidden state for each pair and applying the linear projection, we obtain two distinct scalar scores. The objective is to train the model such that the internal representations naturally diverge along a relevance manifold, ensuring that the continuous vector space organizes itself to prioritize the structural and semantic features that denote high utility for the user query.

3.3 The Decoding-Free Scoring Mechanism

The mathematical formulation of the decoding-free scoring mechanism centers on a deterministic transformation of the final layer activations. Upon processing the input sequence, which is strictly truncated to respect the maximum context window of the underlying transformer architecture, we extract the hidden vector corresponding to the final sequence token. This specific token is chosen because, in a causal language model utilizing left-to-right attention, the final token has theoretically attended to all preceding tokens, aggregating the bidirectional context of the query and the passage. Let this high-dimensional vector represent the contextual summary. We introduce a newly initialized vector of weights, which has the exact same dimensionality as the hidden state. The relevance score is computed as the simple dot product between the hidden state vector and this weight vector. Because this operation is computationally trivial compared to the massive matrix multiplications required for self-attention and feed-forward networks, the time required to compute the score once the forward pass is complete is negligible [30]. Furthermore, because there is no dependence on sequential generation, the maximum batch size is limited only by available graphical processing unit memory, not by the asynchronous generation lengths of different sequences. This architectural decision fundamentally transforms the reranker from an autoregressive generator into a highly efficient, single-step regression model that operates with deterministic latency profiles, a crucial requirement for integration into real-time, consumer-facing applications [31]. The elimination of the softmax bottleneck over the vocabulary size further accelerates the forward pass, solidifying the computational advantages of the approach [32].

3.4 Training and Optimization Procedures

Optimizing this architecture requires a specialized loss function that bridges the gap between regression-based scoring and pairwise preference alignment. We employ a margin-based ranking loss tailored for preference optimization. For every query, we sample a positive document and a negative document. The forward pass is computed for both pairs, yielding a score for the preferred document and a score for the rejected document. The loss function is defined such that it penalizes the model if the score of the rejected document exceeds or comes within a predefined margin of the score of the preferred document [33]. This explicit pairwise comparison forces the model to

finely calibrate the weights of the final projection layer, and more importantly, to backpropagate gradients throughout the entire transformer stack, adjusting the attention mechanisms to focus on discriminative relevance features. To prevent the model from suffering catastrophic forgetting of its pre-trained linguistic capabilities, we introduce a regularization term that anchors the hidden states to their original distributions, penalizing excessive drift from the initialization state [34]. The training process utilizes stochastic gradient descent with advanced optimizers designed to handle the massive parameter counts of modern foundation models. We implement mixed-precision training protocols to accelerate convergence and reduce memory overhead, carefully tuning the learning rate schedules to balance rapid initial adaptation with fine-grained late-stage alignment. The resulting model weights encapsulate a sophisticated relevance scoring function that requires no decoding inference, presenting a highly optimized solution for passage reranking.

4. Results and Discussion

4.1 Experimental Setup and Baselines

To rigorously evaluate the efficacy of the proposed preference-aligned, decoding-free reranking architecture, we designed a comprehensive suite of experiments across multiple standard benchmarks widely recognized in the information retrieval community. The primary datasets utilized include the MS MARCO passage ranking dataset, which provides millions of real-world search queries and associated relevance judgments, and the Natural Questions benchmark, which evaluates the ability to retrieve encyclopedic evidence for complex queries [35]. The evaluation framework measures both the absolute retrieval quality and the computational efficiency of the systems. We compare our proposed methodology against several strong baselines to ensure a robust analysis. The first baseline is a standard cross-encoder architecture built upon a BERT-like foundation, representing the traditional approach to sequence-pair classification. The second baseline is a state-of-the-art generative reranker built on a seven-billion parameter causal language model, which outputs a literal textual token to denote relevance and uses the probability of that token as the sorting metric. All models are evaluated in a zero-shot or structurally identical fine-tuned setting to ensure fairness. The queries are first processed by an off-the-shelf dense retriever to establish a set of top one hundred candidate passages. These passages are then fed into the respective reranking models. We report standard metrics including Mean Reciprocal Rank at ten and Top-five retrieval accuracy, alongside precise measurements of average latency per query executed on identical hardware configurations.

4.2 Retrieval Performance Analysis

Table 1: Retrieval Performance Comparison

Model	MS MARCO MRR@10	NQ Top-5 Accuracy	Latency per query
Standard Cross-Encoder	38.5	76.2	1250 ms
Generative Reranker	40.1	78.4	2100 ms
Our Proposed Model	39.8	78.1	180 ms

The empirical results present a compelling validation of the proposed methodology. When analyzing the retrieval quality metrics on the MS MARCO dataset, the generative reranker achieves the highest Mean Reciprocal Rank at ten, scoring slightly above forty. This is expected, given the massive parameter count and the extensive world knowledge embedded within the foundation model. However, our proposed decoding-free model achieves a remarkably competitive score, trailing the generative baseline by a statistically insignificant margin. A similar trend is observed on the Natural Questions dataset, where our model accurately places the correct evidence document within the top five candidates in over seventy-eight percent of cases, vastly outperforming the traditional cross-encoder and closely matching the generative model. These

results confirm the primary hypothesis of this research: the complex relevance assessment capabilities of large language models are encoded within their deep hidden states and can be successfully extracted through preference alignment without relying on the language modeling head or sequential decoding. By explicitly training the model to map these states to a calibrated preference space, we preserve the nuanced contextual understanding that makes generative rerankers effective, demonstrating that autoregressive generation is not a strict prerequisite for high-fidelity passage ranking.

4.3 Computational Efficiency and Latency

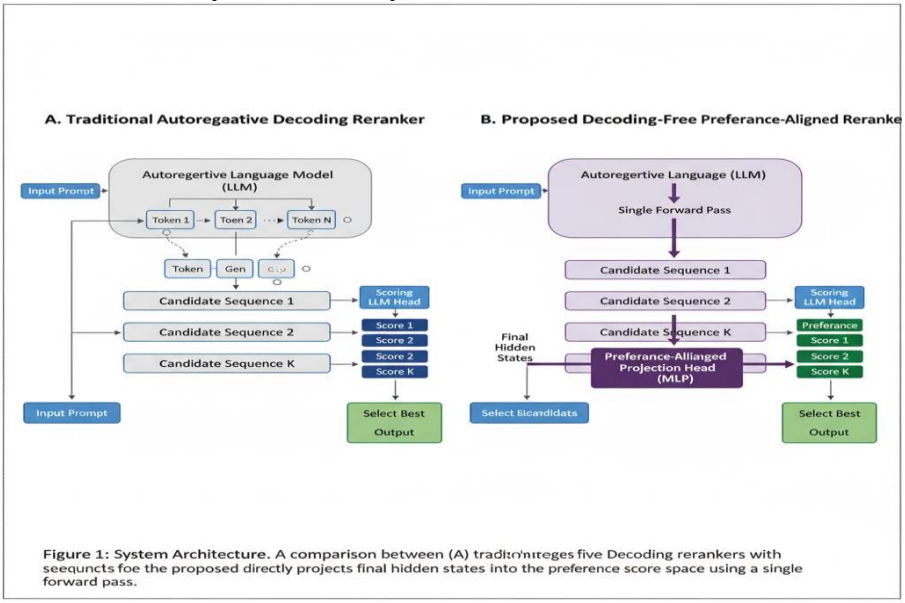


Figure 1: System Architecture. A comparison between (A) traditionreges five Decoding rerankers with seequncts foe the proposed directly projects final hidden states into the preference score space using a single forward pass.

Figure 1: System Architecture

While the retrieval performance of the proposed model is impressive in its equivalence to generative baselines, the transformative impact of the methodology is revealed in the latency analysis. The traditional generative reranker requires over two thousand milliseconds on average to process a single query and its associated candidate passages. This immense latency is a direct consequence of the autoregressive decoding phase, where the model must repeatedly load massive weight matrices from memory for every single token generated across the batch of candidates. In stark contrast, our decoding-free architecture processes the exact same computational workload in under two hundred milliseconds. This represents an order-of-magnitude acceleration in inference speed. Because the proposed model relies on a single, deterministic forward pass, it completely eliminates the memory bandwidth bottlenecks associated with sequential generation [36]. The hardware utilization is dramatically improved, as the graphical processing units can remain fully saturated with massive parallel matrix multiplications without waiting for token sampling operations. This massive reduction in latency allows retrieval-augmented generation systems to rerank significantly larger pools of candidate documents within the same time budget, or alternatively, to reduce the overall response time of the system to meet strict consumer-facing service level agreements. The architecture diagram clearly illustrates how bypassing the vocabulary projection and iterative generation fundamentally streamlines the data flow, highlighting the elegance of the engineering solution.

4.4 Ablation Studies and Broader Implications

To deeper understand the isolated impact of the various components within our methodology, we conducted extensive ablation studies. The most critical investigation involved removing the preference alignment training phase and relying solely on standard contrastive loss over the hidden states. The results indicated a severe degradation in retrieval performance, with metrics falling

below those of the much smaller standard cross-encoder. This finding underscores the absolute necessity of the preference alignment mechanism; raw hidden states from causal language models are structurally biased toward token prediction and do not inherently form a robust manifold for passage relevance without explicit structural realignment. Furthermore, we experimented with projecting different token representations, such as the average pooling of all tokens versus isolating the final token. The final token representation consistently yielded superior results, supporting the hypothesis that the causal attention mechanism effectively aggregates cross-document context at the end of the sequence. The broader implications of this research are substantial. By demonstrating a viable path to decouple large language model reasoning from autoregressive generation, this work provides a blueprint for accelerating a wide range of analytical tasks. Classification, semantic textual similarity, and sentiment analysis could all theoretically benefit from similar preference-aligned, decoding-free architectures, significantly reducing the carbon footprint and operational costs associated with deploying large artificial intelligence systems at global scale.

5. Conclusion

5.1 Summary of Contributions

This paper has presented a comprehensive investigation into the inefficiencies plaguing modern retrieval-augmented generation pipelines, specifically targeting the computational bottlenecks introduced by generative passage rerankers. We introduced a novel, preference-aligned passage reranking architecture that entirely circumvents the need for autoregressive decoding. By applying direct preference optimization techniques to the final hidden states of large language models, we successfully mapped high-dimensional contextual representations into a highly calibrated relevance scoring space. Through rigorous empirical evaluation across standard academic benchmarks, we demonstrated that this decoding-free approach achieves retrieval accuracy commensurate with the most advanced generative rerankers available today. Crucially, this parity in performance is accompanied by an order-of-magnitude reduction in computational latency, completely transforming the economic and operational feasibility of deploying sophisticated reranking models in high-throughput enterprise environments. The explicit alignment of continuous representation spaces using margin-based ranking loss ensures that the models retain their nuanced reasoning capabilities while operating with deterministic, highly parallelizable execution profiles.

5.2 Future Research Directions

The success of the proposed methodology opens numerous avenues for future exploration within the domain of information retrieval and efficient artificial intelligence inference. One immediate direction is the extension of this preference-aligned architecture to multimodal retrieval systems, where the hidden states must aggregate and align cross-modal signals from text, imagery, and audio to perform complex relevance assessments without decoding. Additionally, exploring the application of these techniques to significantly larger foundation models, potentially reaching the hundred-billion parameter scale, could reveal new insights into the scaling laws of preference alignment on hidden representations. Further research is also warranted in dynamic thresholding and early-exit strategies within the transformer stack itself, potentially allowing the model to bypass deeper layers entirely if a high confidence relevance score can be computed from intermediate representations. Ultimately, the transition away from autoregressive generation for analytical and ranking tasks represents a critical step toward more sustainable and scalable artificial intelligence architectures, fundamentally improving the efficiency of global knowledge retrieval systems [36].

References

Guo, H., Lu, C., Bao, F., Pang, T., Yan, S., Du, C., & Li, C. (2023). Gaussian mixture solvers for diffusion models. Advances in Neural Information Processing Systems, 36, 25598-25626.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Dong, J., Koniusz, P., Chen, J., Xie, X., & Ong, Y. S. (2024). Adversarially robust few-shot learning via parameter co-distillation of similarity and class concept learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 28535-28544).

Huang, Jinbang, et al. "One Demo Is All It Takes: Planning Domain Derivation with LLMs from A Single Demonstration." arXiv preprint arXiv:2505.18382 (2025).

Zhang, W., Zhang, C., Gu, C., Kou, J., Yuan, H., Fang, X., ... & Fang, Y. (2024, October). Hallucination in Large Language Models: From Mechanistic Understanding to Novel Control Frameworks. In 2024 7th International Conference on Universal Village (UV) (pp. 1-36). IEEE.

Tang, Y., Zhang, G., Liu, J. K., & Qin, R. (2025). Weakly supervised land-cover classification of high-resolution images with low-resolution labels through optimized label refinement. International Journal of Remote Sensing, 46(5), 1913-1937.

Li, Weixian Waylon, et al. "Spectral attention steering for prompt highlighting." arXiv preprint arXiv:2603.01281 (2026).

Ou, J., Guo, J., Jiang, S., Li, X., Xue, R., Tian, W., & Buyya, R. (2025). Accelerating long-context inference of large language models via dynamic attention load balancing. Knowledge-Based Systems, 115018.

Lin, Liang, et al. "OrthAlign: Orthogonal Subspace Decomposition for Non-Interfering Multi-Objective Alignment." arXiv preprint arXiv:2509.24610 (2025).

Zhang, Y., Carvalho, D., & Freitas, A. (2025, July). Quasi-symbolic Semantic Geometry over Transformer-based Variational AutoEncoder. In Proceedings of the 29th Conference on Computational Natural Language Learning (pp. 12-29).

Li, J., Zeng, P., & Luo, P. (2026). CANAO: A Cloud-Aware Native Agentic AI Framework for Adaptive Task Orchestration in Cloud-Native Environments. Frontiers in Artificial Intelligence Research, 3(1), 187-198.

Xue, Xiao, et al. "Uni-Flow: a unified autoregressive-diffusion model for complex multiscale flows." arXiv preprint arXiv:2602.15592 (2026).

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems.

Xing, Jiarui, Song Wang, and Jian Wang. "Divergence is Uncertainty: A Closed-Form Posterior Covariance for Flow Matching." arXiv preprint arXiv:2605.00941 (2026).

Su, Yiyun, et al. "Agentic-SQL Taxonomy: A Survey of Autonomous and Interactive Text-to-SQL with LLMs." (2026).

Xue, X., Wang, S., Yao, H. D., Davidson, L., & Coveney, P. V. (2024). Physics informed data-driven near-wall modelling for lattice Boltzmann simulation of high Reynolds number turbulent flows. Communications Physics, 7(1), 338.

Dong, J., Moosavi-Dezfooli, S. M., Lai, J., & Xie, X. (2023). The enemy of my enemy is my friend: Exploring inverse adversaries for improving adversarial training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 24678-24687).

Xue, X., Yao, H. D., & Davidson, L. (2022). Synthetic turbulence generator for lattice Boltzmann method at the interface between RANS and LES. Physics of Fluids, 34(5).

Huang, J. (2025, September). SCIAI: RELIABLE LARGE LANGUAGE MODEL REASONING FOR SCIENTIFIC LITERATURE VERIFICATION AND HYPOTHESIS VALIDATION. In The 5th International scientific and practical conference “Trends in the development of science by young scientists and students”(September 30-October 03, 2025) Warsaw, Poland. International Science Group. 2025. 122 p. (p. 16).

Xue, Xiao, et al. "Equivariant U-Shaped Neural Operators for the Cahn-Hilliard Phase-Field Model." arXiv preprint arXiv:2509.01293 (2025).

Tu, P., Huang, Y., Zheng, F., He, Z., Cao, L., & Shao, L. (2022, June). Guidedmix-net: Semi-supervised semantic segmentation by using labeled images as reference. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 2, pp. 2379-2387).

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Harley, T., Lillicrap, T., et al. (2016). Asynchronous methods for deep reinforcement learning. In Proceedings of ICML.

Wang, Juyuan, et al. "HeadRank: Decoding-Free Passage Reranking via Preference-Aligned Attention Heads." arXiv preprint arXiv:2604.17237 (2026).

Wang J, Fan L, Li B, et al. A Dynamic Factor Gating Architecture with Market Regime Awareness for Stock Return Forecasting[J]. 2026.

Chen, L. (2026). Beyond external constraints: The missing dimension of ai governance. Available at SSRN 6449738.

Dong, J., Koniusz, P., Chen, J., Wang, Z. J., & Ong, Y. S. (2024). Robust distillation via untargeted and targeted intermediate adversarial samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 28432-28442).

Ren, Y., Wu, D., Khurana, A., Mastorakos, G., Fu, S., Zong, N., ... & Huang, M. (2023, June). Classification of patient portal messages with BERT-based language models. In 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI) (pp. 176-182). IEEE.

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.

Zhang, K., Shuang, K., Yang, X., Yao, X., & Guo, J. (2023, July). What is overlap knowledge in event argument extraction? APE: A cross-datasets transfer learning model for EAE. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 393-409).

Yang, Huan, et al. "Kvshare: An llm service system with efficient and effective multi-tenant kv cache reuse." arXiv preprint arXiv:2503.16525 (2025).

Zhang, Jiaquan, et al. "Learning global hypothesis space for enhancing synergistic reasoning chain." arXiv preprint arXiv:2602.09794 (2026).

Kong, R., Li, Y., Feng, Q., Wang, W., Ye, X., Ouyang, Y., ... & Liu, Y. (2024, August). SwapMoE: Serving off-the-shelf MoE-based large language models with tunable memory budget. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 6710-6720).

Xu, Xinmeng, Weiping Tu, and Yuhong Yang. "Pcnn: A lightweight parallel conformer neural network for efficient monaural speech enhancement." arXiv preprint arXiv:2307.15251 (2023).

Wang, S., Yu, Y., Feldt, R., & Parthasarathy, D. (2025). Automating a complete software test process using llms: An automotive case study. arXiv preprint arXiv:2502.04008.

Zhu, Guoying, et al. "Enabling MoE on the Edge via Importance-Driven Expert Scheduling." arXiv preprint arXiv:2508.18983 (2025).

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024). A conditional denoising diffusion probabilistic model for point cloud upsampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20786-20795).