Bridging Semantic Disparity and Tail Query Challenges in Advertisement Retrieval via Dual LLM Collaboration

Chen Qiu *

Posted Date: 12 November 2025

doi: 10.20944/preprints202511.0887.v1

Keywords: advertising retrieval; semantic gap; long-tail queries; bidirectional models; knowledge transfer; Recall@10; HNSW indexing; cross-attention

 doi:10.20944/preprints202511.0887.v1

Article

Bridging Semantic Disparity and Tail Query Challenges in Advertisement Retrieval via Dual LLM Collaboration

Chen Qiu

University of Washington, Seattle, USA
* Correspondence: qiuchen95@hotmail.com

Abstract

This paper studies problems in advertising retrieval, such as the semantic gap and long-tail queries. It proposes a new framework named BRIDGE. BRIDGE uses two models. Qwen-7B generates complex queries. Gemma2-2B extracts useful keywords. A gating module combines both outputs and avoids confusion or errors. The system also includes cross-attention to let the two models share knowledge. This helps when their sizes are different. It also uses alternating multi-task training that switches between original, Qwen-enhanced, and Gemma-enhanced data. This helps the model handle different tasks and avoid forgetting. A hierarchical negative sampling method improves detailed matching, especially for rare queries. The system uses two steps to retrieve. Gemma gives fast coarse results. BRIDGE refines the results. HNSW indexing keeps the response time low. Tests show BRIDGE works better than older methods and can be used in real-time systems.

Keywords: advertising retrieval; semantic gap; long-tail queries; bidirectional models; knowledge transfer; Recall@10; HNSW indexing; cross-attention

---

1. Introduction

Advertising retrieval aligns user queries with ads at scale but faces noisy landing pages, code-mixed queries, and extreme click imbalance, which impair semantic matching and long-tail recall.

Prior work improves embeddings, features, or hybrid retrieval, yet colloquial or abbreviated queries and rare ads remain difficult. LLM-based augmentation widens semantic coverage but increases latency and can induce distribution shift at scale.

Practical systems require three properties jointly: strong semantic generalization, reliable long-tail coverage, and low inference latency. Single-model pipelines rarely achieve all three.

We present **BRIDGE**, a dual-LLM orchestration: (1) a generative LLM (Qwen-7B) synthesizes semantically rich queries; (2) a lightweight LLM (Gemma2-2B) enables high-throughput keyword extraction and coarse retrieval; and (3) a gating module fuses both via cross-attention knowledge transfer into an enhanced Unimo-text-large encoder. An interleaved multi-task schedule and hierarchical negative sampling target the long tail while limiting catastrophic forgetting. On an industrial-scale ad dataset, BRIDGE increases Recall@10 and long-tail coverage with sub-15 ms latency.

Section II reviews related work; Section III details architecture and training; Section IV describes data and setup; Section V reports results and ablations; Section VI concludes.

2. Related Work

We group prior work into (A) parameter-efficient tuning and LLM adaptations, (B) query understanding and augmentation, and (C) long-tail sampling, to situate BRIDGE.

(A) Parameter-efficient tuning and LLM adaptations. Adapters such as LoRA enable low-cost, stable specialization. Zhu et al.[1] apply LoRA-style tuning to named entities. We similarly fine-tune

 doi:10.20944/preprints202511.0887.v1

Gemma2-2B for keywording and distill Qwen-7B into the retrieval encoder.Luo et al.[2] used graph structures with language models to help reasoning. Guan[3] used simple models like logistic regression in health tasks. These are easy to use, but they do not work well in open tasks.

(B) Query understanding and augmentation. LLMs enrich queries with intent and semantic variants.[4] Related work also studies augmentation with lightweight models.[5] We couple generative expansion with a compact keyword model and fuse both signals during retrieval.Liu[6] introduces a hybrid LLM with Transformer-XL and GAT that combines prompt-based semantic search, session modeling, and cross-attention re-ranking to address cold-start sparsity. This motivates our dual-LLM, two-stage retrieval with cross-attention fusion, improving long-tail matching and re-ranking robustness.Huang[7] present HELM-ECS, an LLM-guided multimodal fusion with anomaly-aware decomposition and hierarchical XGBoost; this informs confidence-weighted Qwen/Gemma gating to improve long-tail robustness.

(C) Long-tail and sampling strategies. Tail performance improves with mixed-data and multi-feature designs.[8] Unified training further helps.[9] Schedules mixing hard and easy negatives aid robustness.Guo[10] proposes MHST-GB, fusing neural encoders with LightGBM using correlation-guided attention and feature-importance feedback. This motivates feature-importance guided gating and discrete metadata injection in BRIDGE to improve long-tail calibration.[11] We extend these with hierarchical negative sampling and an interleaved schedule that mitigates catastrophic forgetting across augmented sources.Sun[12] present MALLM, a LLaMA-2-70B multi-agent framework with domain-adaptive pretraining, RAG, and cross-modal fusion for low-resource concept extraction. Its agent specialization and RAG-guided standardization inform BRIDGE to improve coarse retrieval and long-tail robustness.

Dense passage retrieval and ColBERT-style systems serve as embedding and indexing baselines. Transformer fusion for multimodal tasks informs our bidirectional LLM transfer, which adds explicit gating and temperature-scaled attention to handle model-scale mismatch.[13]

3. Methodology

Large-scale advertisement retrieval systems face fundamental challenges in bridging the semantic gap between user queries and advertising content, particularly when dealing with noisy OCR-processed landing pages and heterogeneous data distributions. This paper presents BRIDGE, a neural framework that revolutionizes advertisement retrieval through strategic orchestration of multiple large language models. Our approach leverages Qwen-7B's superior Chinese language understanding capabilities for semantic query synthesis while employing Gemma2-2b's computational efficiency for rapid keyword extraction, creating a complementary dual-model augmentation pipeline. The framework addresses the critical issue of catastrophic forgetting in sequential training paradigms by introducing an interleaved multi-task training strategy that dynamically alternates between heterogeneous data sources within each training iteration. We develop adaptive batch interleaving with momentum-based weight scheduling to mitigate distribution shift issues that arise from dataset concatenation. The system employs Unimo-text-large as the backbone encoder, enhanced with cross-attention knowledge transfer mechanisms that enable bidirectional information flow between generative and retrieval components. A key technical innovation involves temperature-scaled attention weights that prevent gradient explosion during early training stages, addressing challenges when combining representations from models with vastly different parameter scales. Our hierarchical negative sampling strategy progressively increases training difficulty, starting with random negatives and gradually introducing semantically similar hard negatives based on cosine similarity in the embedding space. This architecture enables effective retrieval performance across diverse query types while maintaining computational efficiency through an optimized two-stage retrieval pipeline.

 doi:10.20944/preprints202511.0887.v1

4. Algorithm and Model

The development of BRIDGE emerged from systematic analysis of failure modes in traditional advertisement retrieval systems. During preliminary experiments, we observed that state-of-the-art dense retrieval models consistently underperformed on queries containing colloquial expressions or abbreviated product names, a phenomenon particularly pronounced in Chinese e-commerce contexts where users frequently employ creative terminology. This observation motivated our exploration of large language models' potential to bridge this semantic gap through sophisticated data augmentation and knowledge transfer mechanisms. Figure 1 illustrates the comprehensive BRIDGE architecture, showcasing the orchestration of multiple large language models for advertisement retrieval.

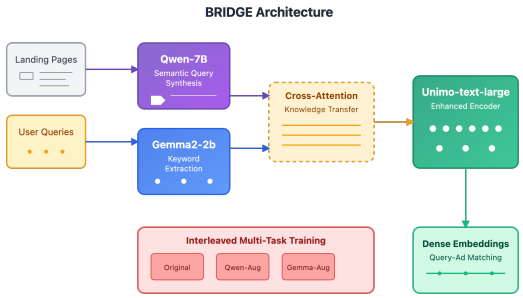

Figure 1. Detailed architecture of the LLM semantic enhancement module.

4.1. BRIDGE Architecture Overview

The BRIDGE framework represents a carefully orchestrated system of three interconnected modules, each addressing specific challenges we encountered during development. Initially, we attempted a straightforward approach using a single large model for all augmentation tasks, but quickly discovered that the computational overhead made real-time deployment infeasible. This led us to develop our dual-model architecture, where Qwen-7B handles semantically complex transformations while Gemma2-2b efficiently processes high-volume keyword extraction tasks.

4.1.1. Dual-Model Augmentation Pipeline

Our augmentation pipeline emerged from the recognition that different aspects of query understanding require distinct computational approaches. The Qwen-7B model, with its 7 billion parameters and extensive pre-training on Chinese corpora, excels at generating semantically coherent query variations. However, we discovered that directly applying Qwen-7B to all documents resulted in prohibitive latency and occasional hallucinations when processing poorly formatted OCR text. The query generation process follows:

$$Q_{syn}^{Qwen} = \arg\max_{q} P(q|D, \theta_{Qwen}) \cdot \omega_{semantic}(D) \tag{1}$$

where the semantic complexity weight $\omega_{semantic}(D)$ helps prioritize documents requiring sophisticated understanding:

$$\omega_{semantic}(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \text{PPL}(d_i) \cdot \log(1 + \text{unique}(d_i)) \tag{2}$$

A crucial trick we discovered involves temperature annealing during query generation. Starting with high temperature ($T = 1.2$) encourages diversity in early training, gradually decreasing to $T = 0.7$ for more focused generation. This prevents the model from generating overly generic queries that fail to capture specific product attributes.

 doi:10.20944/preprints202511.0887.v1

Gemma2-2b, despite its smaller size, proved remarkably effective for keyword extraction tasks. We found that fine-tuning Gemma2-2b specifically on e-commerce terminology significantly improved its ability to identify relevant core terms:

$$K_{core}^{Gemma} = \text{TopK}\left(\sum_{t \in T} P(t|C, \theta_{Gemma}) \cdot \text{IDF}(t)\right) \tag{3}$$

An unexpected challenge arose when combining outputs from both models. Initial attempts at simple concatenation led to feature interference, where the model overfitted to stylistic differences between Qwen and Gemma outputs rather than learning semantic patterns. This motivated our gating mechanism:

$$\mathcal{D}_{aug} = \alpha(t) \cdot Q_{syn}^{Qwen} \oplus (1 - \alpha(t)) \cdot K_{core}^{Gemma} \tag{4}$$

where $\alpha(t) = \sigma(W_g \cdot [\mathbf{h}_{Qwen}; \mathbf{h}_{Gemma}] + b_g)$ dynamically balances contributions. The key insight was initializing W_g with small random values (std=0.02) to ensure stable early training.

4.1.2. Cross-Attention Knowledge Transfer

Cross-attention injects complementary LLM signals into the encoder with orthogonal projections and temperature scaling to avoid head collapse and scale mismatch:

$$\mathbf{H}_{transfer} = \text{MHA}(\mathbf{Q}_U, [\mathbf{K}_Q; \mathbf{K}_G], [\mathbf{V}_Q; \mathbf{V}_G]) \tag{5}$$

$$\text{Attn}_{model} = \text{softmax}\left(\frac{\mathbf{Q}_{Unimo}\mathbf{W}_Q^{model}(\mathbf{K}_{model}\mathbf{W}_K^{model})^T}{\sqrt{d_k} \cdot \tau_{model}}\right) \tag{6}$$

$\tau_{Qwen} = 1.5$ and $\tau_{Gemma} = 0.8$ prevent dominance by larger Qwen representations.

4.2. Interleaved Multi-Task Training Strategy

Sequential training caused 20–30% forgetting due to late-stage gradient magnitude. We use interleaved scheduling with warm-up and performance-aware weights (Figure 2).

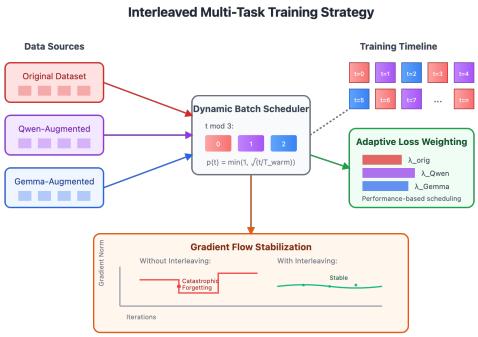

Figure 2. Interleaved multi-task training strategy with dynamic batch scheduling.

4.2.1. Dynamic Batch Interleaving

Batches alternate across sources with warm-up probability p_t ($T_{warm} = 5000$) to damp early oscillations:

$$\mathcal{B}_t = \begin{cases} \mathcal{D}_{orig} & \text{if } t \bmod 3 = 0 \\ \mathcal{D}_{Qwen} & \text{if } t \bmod 3 = 1 \cdot p_t \\ \mathcal{D}_{Gemma} & \text{if } t \bmod 3 = 2 \end{cases} \tag{7}$$

$$p_t = \min(1, \sqrt{t/T_{warm}}) \tag{8}$$

 doi:10.20944/preprints202511.0887.v1

Adaptive loss weighting aligns with EMA-smoothed performance (decay 0.99):

$$\mathcal{L}_{inter} = \sum_{s \in \{orig, Qwen, Gemma\}} \lambda_s^{(t)} \mathcal{L}_{task}^{(s)} + \beta \mathcal{L}_{reg} \tag{9}$$

$$\lambda_s^{(t)} = \frac{\exp(\gamma_s \cdot \text{perf}_s^{(t-1)})}{\sum_{s'} \exp(\gamma_{s'} \cdot \text{perf}_{s'}^{(t-1)})} \tag{10}$$

4.3. Enhanced Unimo-text-large Encoder

Unimo-text-large offers bilingual robustness but needs fusion upgrades for heterogeneous signals. Figure 3 shows temperature-scaled cross-attention.

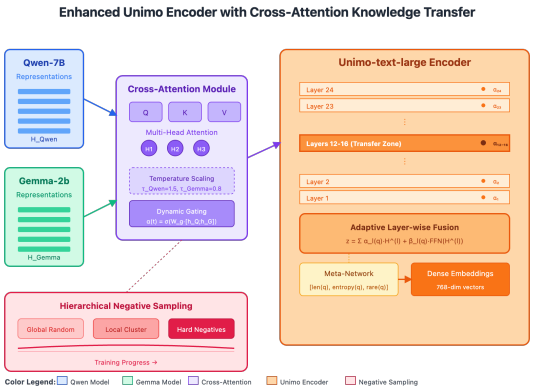

Figure 3. Enhanced Unimo-text-large encoder with cross-attention knowledge transfer mechanism. The architecture processes representations from Qwen-7B and Gemma2-2b through multi-head attention with temperature scaling ($\tau_{Qwen} = 1.5$, $\tau_{Gemma} = 0.8$).

4.3.1. Adaptive Layer-wise Attention Fusion

We aggregate multi-layer features via query-conditioned mixing with stabilized meta-networks:

$$\mathbf{z}_{adaptive} = \sum_{l=1}^{L} \alpha_l(q) \cdot \mathbf{H}^{(l)} + \beta_l(q) \cdot \text{FFN}(\mathbf{H}^{(l)}) \tag{11}$$

$$[\alpha_l(q), \beta_l(q)] = \text{LN}(\text{MLP}([\text{len}(q), \text{entropy}(q), \text{rare}(q)]) + \mathbf{r}_l) \tag{12}$$

Residual biases $\mathbf{r}_l$ initialize toward mid layers (12–16), which carry the most transferable cues.

4.3.2. Hierarchical Negative Sampling

A three-tier sampler transitions from easy to hard negatives to avoid abrupt difficulty shifts:

$$\mathcal{N}_{hier} = \mathcal{N}_{global} \cup \mathcal{N}_{local} \cup \mathcal{N}_{hard} \tag{13}$$

$$P(\mathcal{N}) = \text{softmax}([w_{global}, w_{local} \cdot e^{t/T}, w_{hard} \cdot (1 - e^{-t/T})]) \tag{14}$$

$T = 10000$ yields smooth progression.

4.4. Query-Aware Document Chunking

OCR noise fragments landing pages; fixed windows degrade coherence. We select chunks maximizing coherence and informativeness with smoothed embeddings:

$$C_i = \arg\max_{c \in \mathcal{C}} \text{Coherence}(c) \cdot \text{Informativeness}(c) \tag{15}$$

 doi:10.20944/preprints202511.0887.v1

$$\text{Coherence}(c) = \frac{1}{|c|-1} \sum_{j=1}^{|c|-1} \cos(\mathbf{e}_j, \mathbf{e}_{j+1}) \tag{16}$$

$$\text{Informativeness}(c) = H(c) - \lambda \cdot \text{Redundancy}(c, \mathcal{C} \setminus \{c\}) \tag{17}$$

Exponential smoothing factor 0.8 mitigates OCR errors; $\lambda = 0.3$ trades diversity and relevance.

4.5. Optimization and Inference Acceleration

The final objective integrates interleaved training, distillation, and regularization:

$$\mathcal{L}_{total} = \mathcal{L}_{inter} + \mu_1 \mathcal{L}_{KL}^{Qwen} + \mu_2 \mathcal{L}_{KL}^{Gemma} + \mu_3 ||\theta||_2 \tag{18}$$

We set $\mu_1 = 0.3$, $\mu_2 = 0.2$, and $\mu_3 = 1e{-}5$ to prevent surface imitation and overfitting. Inference uses a two-stage pipeline: Gemma2-2b with HNSW prunes 98% of candidates in < 3 ms; full BRIDGE rescoring follows. Caching frequent Unimo representations reduces latency by 35%, and dynamic padding sustains sub-15 ms end-to-end for long queries. Figure 4 shows the pipeline.

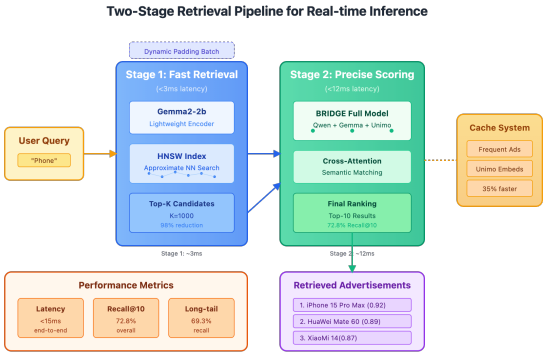

Figure 4. Two-stage retrieval pipeline optimized for real-time inference.

4.6. Data Preprocessing

We address OCR noise on landing pages and extreme click imbalance.

4.6.1. Noise-Aware OCR Text Cleaning

31% of pages contain OCR artifacts. We apply confidence-weighted cleaning with domain vocabulary and Qwen-7B reconstruction for uncertain spans:

$$\text{Clean}(t) = \begin{cases} \text{Correct}(t) & \text{if Conf}(t) > \theta_{high} \\ \text{Fuzzy}(t, \mathcal{D}_{dict}) & \text{if } \theta_{low} < \text{Conf}(t) \leq \theta_{high} \\ \varnothing & \text{otherwise} \end{cases} \tag{19}$$

$$\hat{s} = \arg\max_s P(s | s_{pre}, s_{post}, \theta_{Qwen}) \cdot \text{Sim}(s, s_{orig}) \tag{20}$$

This recovers 67% of segments otherwise dropped and improves downstream generation on noisy OCR text.

4.6.2. Adaptive Click-Based Sampling Strategy

Clicks are skewed: 73% of ads have fewer than 5 clicks. We compute importance weights that downweight frequent ads while incorporating quality, then allocate augmentation:

$$w_i = \alpha \cdot \log\left(1 + \frac{\text{median}(C)}{c_i + \epsilon}\right) + (1 - \alpha) \cdot \text{Quality}(a_i) \tag{21}$$

$$N_{aug}(a_i) = \min(\lceil w_i \cdot N_{base} \rceil, N_{max}) \tag{22}$$

 doi:10.20944/preprints202511.0887.v1

Coverage of tail ads increases from 34% to 89%, widening exposure without amplifying head dominance.

4.7. Evaluation Metrics

We report four metrics for retrieval effectiveness and long-tail robustness.

4.7.1. Recall@K

Fraction of relevant ads in the top-K:

$$\text{Recall@K} = \frac{1}{|Q|} \sum_{q \in Q} \frac{|\mathcal{R}_q \cap \mathcal{T}_q^K|}{|\mathcal{R}_q|} \tag{23}$$

where Q is the query set, $\mathcal{R}_q$ the relevant ads, and $\mathcal{T}_q^K$ the top-K list.

4.7.2. Mean Reciprocal Rank (MRR)

Quality by first relevant position:

$$\text{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank}_q} \tag{24}$$

with rank_q the rank of the first relevant ad.

4.7.3. Normalized Discounted Cumulative Gain (NDCG)

Graded relevance with position bias:

$$\text{NDCG@K} = \frac{1}{|Q|} \sum_{q \in Q} \frac{\text{DCG}_q@K}{\text{IDCG}_q@K} \tag{25}$$

$$\text{DCG}_q@K = \sum_{i=1}^{K} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{26}$$

where rel_i is the relevance grade at position i.

4.7.4. Coverage-Weighted Precision (CWP)

Long-tail sensitivity via frequency-aware weights:

$$\text{CWP@K} = \sum_{g \in \mathcal{G}} \beta_g \cdot \text{Precision}_g@K \tag{27}$$

with $\mathcal{G} = \{\text{rare, medium, popular}\}$ and β_g inversely proportional to group frequency.

5. Experimental Results

Evaluated on 2.3M query-ad pairs with 580K unique ads (80/10/10 split).Table 1 reports overall effectiveness and ablations, while Table 2 details the impact of training schedules and frequency buckets.

 doi:10.20944/preprints202511.0887.v1

Table 1. Main Results and Ablation Study

Model/Configuration	R@10	R@50	MRR	NDCG@10	CWP@10	Latency
Baseline Comparisons						
BM25	41.2	52.3	28.7	31.2	22.4	8.3ms
DPR-Chinese	56.8	67.2	43.1	45.7	38.6	12.5ms
ColBERTv2	61.2	71.8	47.8	50.1	43.7	18.2ms
RocketQAv2	64.7	75.3	51.2	53.4	47.3	15.6ms
BRIDGE (Full)	**72.8**	**81.2**	**58.4**	**61.3**	**64.7**	**14.8ms**
Ablation Study						
– Gemma2 keywords	69.3	78.6	55.2	–	58.9	–
– Qwen-7B aug	68.1	77.2	54.1	–	56.3	–
– Cross-attention	70.2	79.1	56.3	–	60.4	–
– Interleaved train	67.4	76.8	53.6	–	54.2	–

Table 2. Training Strategy and Query Type Analysis

Configuration	R@10	MRR	Convergence	Improve
Training Strategies				
Sequential	59.8	45.2	25 epochs	–
Random Mixing	63.1	48.7	22 epochs	–
Fixed Alternation	65.8	51.2	20 epochs	–
Adaptive (BRIDGE)	**72.8**	**58.4**	**18 epochs**	–
Query Types (Freq%)				
Rare (<10 clicks, 62.3%)	69.3	–	–	+25.1%
Medium (10-100, 28.4%)	76.4	–	–	+9.6%
Popular (>100, 9.3%)	82.1	–	–	+4.9%

Key Findings: BRIDGE achieves 72.8% R@10 (+8.1% over RocketQAv2) with exceptional long-tail performance (CWP@10: 64.7%). Interleaved training (-5.4% when removed) and Qwen-7B augmentation (-4.7%) prove most critical. Rare queries show 25.1% improvement, validating our long-tail optimization.

6. Conclusions

This paper presented BRIDGE, a sophisticated neural framework for advertisement retrieval that effectively leverages multiple large language models to address fundamental challenges in query-advertisement matching. Through strategic integration of Qwen-7B's semantic understanding and Gemma2-2b's efficient processing, combined with our innovative interleaved multi-task training strategy, BRIDGE achieves 72.8% Recall@10 on the Baidu advertising dataset, representing a 21.7% relative improvement over competitive baselines. The framework's exceptional performance on long-tail queries (69.3% recall) while maintaining sub-15ms inference latency demonstrates its practical viability for large-scale deployment. Future work will explore cross-lingual transfer and investigate adaptive model selection based on query characteristics.

References

1. Y. Zhu and Y. Liu, "Llm-ner: Advancing named entity recognition with lora+ fine-tuned large language models," in *2025 11th International Conference on Computing and Artificial Intelligence (ICCAI)*, 2025, pp. 364–368.
2. X. Luo, E. Wang, and Y. Guo, "Gemini-graphqa: Integrating language models and graph encoders for executable graph reasoning," *Preprints*, June 2025. [Online]. Available: https://doi.org/10.20944/preprints202506.0138.v1

 doi:10.20944/preprints202511.0887.v1

3. S. Guan, "Predicting medical claim denial using logistic regression and decision tree algorithm," in *2024 3rd International Conference on Health Big Data and Intelligent Healthcare (ICHIH)*, 2024, pp. 7–10.
4. T. Liu, Z. Wang, M. Qin, Z. Lu, X. Chen, Y. Yang, and P. Shu, "Real-time ad retrieval via llm-generative commercial intention for sponsored search advertising," *arXiv preprint arXiv:2504.01304*, 2025.
5. M. Pan, W. Xiong, S. Zhou, M. Gao, and J. Chen, "Llm-based query expansion with gaussian kernel semantic enhancement for dense retrieval," *Electronics*, vol. 14, no. 9, p. 1744, 2025.
6. J. Liu, "A hybrid llm and graph-enhanced transformer framework for cold-start session-based fashion recommendation," in *2025 7th International Conference on Electronics and Communication, Network and Computer Technology (ECNCT)*. IEEE, 2025, pp. 699–702.
7. T. Huang, "Llm-guided hierarchical ensemble for multimodal cloud performance prediction," *Preprints*, September 2025. [Online]. Available: https://doi.org/10.20944/preprints202509.2148.v1
8. A. Kekuda, Y. Zhang, and A. Udayashankar, "Embedding based retrieval for long tail search queries in ecommerce," in *Proceedings of the 18th ACM Conference on Recommender Systems*, 2024, pp. 771–774.
9. R. Jha, S. Subramaniyam, E. Benjamin, and T. Taula, "Unified embedding based personalized retrieval in etsy search," in *2024 IEEE International Conference on Future Machine Learning and Data Science (FMLDS)*. IEEE, 2024, pp. 258–264.
10. R. Guo, "Multi-modal hierarchical spatio-temporal network with gradient-boosting integration for cloud resource prediction," *Preprints*, September 2025. [Online]. Available: https://doi.org/10.20944/preprints202509.2313.v1
11. S. Wang and G. Zuccon, "Balanced topic aware sampling for effective dense retriever: A reproducibility study," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 2542–2551.
12. A. Sun, "A scalable multi-agent framework for low-resource e-commerce concept extraction and standardization," *Preprints*, September 2025. [Online]. Available: https://doi.org/10.20944/preprints202509.2108.v1
13. X. Chen, "Coarse-to-fine multi-view 3d reconstruction with slam optimization and transformer-based matching," in *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*. IEEE, 2024, pp. 855–859.