Novelty Assessment Report

Paper: Supervised Fine-Tuning or Contrastive Learning? Towards Better Multimodal LLM Reranking
PDF URL: https://openreview.net/pdf?id=1Mh2q7L0eY
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

In information retrieval, training reranking models mainly focuses on two types of objectives: metric learning (e.g. contrastive loss to increase the predicted scores on relevant query-document pairs) and classification (binary label prediction of relevance vs. irrelevance). For BERT-style encoders, various studies have shown that contrastive learning (CL) can be more effective than discriminative (classification) learning. However, for large language models (LLMs), classification via supervised fine-tuning (SFT), which predicts ''yes'' (resp. ''no'') token for relevant (resp. irrelevant) pairs, appears more promising as it aligns well with the generative nature of LLMs. This divergence raises a central question: which objective is intrinsically better suited to LLM-based reranking, and what mechanism underlies the difference? In this work, we conduct a comprehensive comparison and analysis between CL and SFT for reranking, taking the universal multimodal retrieval (UMR) as the experimental playground. We first decompose the objectives into two components: weight, which controls the magnitude of those updates, and direction, which guides the model updates, then present a unified framework for understanding their interactions. Through probing experiments, we find that SFT provides a substantially stronger weighting scheme than CL, whereas the preferred scoring direction shows no clear winner. Taken together, these results point to a consistent advantage of SFT over CL for LLM reranking. To further validate our findings, we conduct large-scale training with SFT and present new state-of-the-art rerankers on the MRB benchmark. We also provide ablations on SFT settings and expect our findings to benefit future research and applications in this area.

Core Task Landscape

This paper addresses: **Comparing Training Objectives for LLM-Based Reranking**

A total of **8 papers** were analyzed and organized into a taxonomy with **6 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Training Objective Design and Comparison**
- **Training Data and Supervision Sources**
- **Bias Mitigation and Robustness**

Complete Taxonomy Tree

- Comparing Training Objectives for LLM-Based Reranking Survey Taxonomy
- Training Objective Design and Comparison
  - Contrastive vs. Supervised Fine-Tuning Objectives ★ (2 papers)
  - [0] Supervised Fine-Tuning or Contrastive Learning? Towards Better Multimodal LLM Reranking (Anon et al., 2026) View paper
  - [2] Rethink Training of BERT Rerankers in Multi-Stage Retrieval Pipeline (Luyu Gao, 2021) View paper
  - Reinforcement Learning and Hybrid Training Approaches (2 papers)
  - [4] ERank: Fusing Supervised Fine-Tuning and Reinforcement Learning for Effective and Efficient Text Reranking (Cai, 2025) View paper
  - [5] Large Language Models for Reranking: A Survey (Y Zhou, 2025) View paper
  - Specialized Loss Functions for Reranking (2 papers)
  - [3] FIRST: Faster Improved Listwise Reranking with Single Token Decoding (Revanth Gangi Reddy, 2024) View paper
  - [8] EARerranker: Efficient Embedding Adequacy Assessment for Retrieval Augmented Generation (D Zeng, n.d.) View paper
- Training Data and Supervision Sources
  - Alternative Supervision Signals (1 papers)
  - [6] Can LLM Annotations Replace User Clicks for Learning to Rank? (Lulu Yu, 2025) View paper
  - Cross-Domain and Generalization Studies (1 papers)
  - [7] Make Large Language Model a Better Ranker (Zhi Zheng, n.d.) View paper
- Bias Mitigation and Robustness
  - Multi-View and Position Bias Reduction (1 papers)
  - [1] Multi-view-guided Passage Reranking with Large Language Models (Kwon Jun, 2025) View paper

Narrative

Core task: Comparing training objectives for large language model based reranking. The field structure suggested by the taxonomy reveals three main branches that organize research around how LLM-based rerankers are trained and refined. The first branch, Training Objective Design and Comparison, examines the choice and formulation of loss functions—particularly contrasting contrastive learning approaches with supervised fine-tuning strategies—and explores how different objectives shape model behavior. Works such as Rethink BERT Rerankers[2] and FIRST[3] illustrate early efforts to understand which training signals best capture relevance. The second branch, Training Data and Supervision Sources, addresses where supervision comes from, including human annotations, click logs, and synthetic labels generated by LLMs themselves, as seen in LLM Annotations Replace Clicks[6]. The third branch, Bias Mitigation and Robustness, focuses on ensuring that rerankers generalize fairly across diverse queries and resist spurious correlations or position biases.

Several active lines of work highlight key trade-offs and open questions. One central theme is whether contrastive objectives—which encourage models to distinguish relevant from irrelevant passages—offer advantages over pointwise or listwise supervised fine-tuning, especially when training data is noisy or limited. Another theme concerns the integration of multimodal signals and the use of LLM-generated annotations to reduce reliance on expensive human labels. Within this landscape, Multimodal LLM Reranking[0] sits naturally alongside efforts like Multi-view Passage Reranking[1] and ERank[4], which also explore richer input representations and alternative training regimes. Compared to Rethink BERT Rerankers[2], which revisited foundational BERT-based objectives, and LLM Reranking Survey[5], which provides a broader overview, the original paper emphasizes the comparative evaluation of objectives in a multimodal setting, bridging objective design with emerging data modalities.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Rethink Training of BERT Rerankers in Multi-Stage Retrieval Pipeline

Authors: Luyu Gao, Zhuyun Dai, Jamie Callan | **Year/Venue**: 2021 | **URL**: View paper

Abstract

Pre-trained deep language models~(LM) have advanced the state-of-the-art of text retrieval. Rerankers fine-tuned from deep LM estimates candidate relevance based on rich contextualized matching signals. Meanwhile, deep LMs can also be leveraged to improve search index, building retrievers with better recall. One would expect a straightforward combination of both in a pipeline to have additive performance gain. In this paper, we discover otherwise and that popular reranker cannot fully exploit th...

Relationship Analysis

Both papers belong to the same taxonomy category examining contrastive learning versus supervised fine-tuning objectives for reranking, though they target different model architectures and eras. The original paper focuses on multimodal LLM-based rerankers, comparing SFT (predicting 'yes'/'no' tokens) against contrastive InfoNCE loss, and finds SFT superior due to stronger weighting schemes. The candidate paper addresses BERT-style encoder rerankers for text retrieval, proposing Localized Contrastive Estimation (LCE) that samples negatives from target retriever results, demonstrating that contrastive learning with localized negatives outperforms vanilla binary classification for BERT models—a contrasting conclusion that highlights how optimal training objectives differ between encoder-based and generative LLM architectures.

Contributions Analysis

Overall novelty summary. The paper contributes a unified framework for analyzing contrastive learning versus supervised fine-tuning in LLM-based reranking, alongside a multimodal reranking benchmark (MRB) and state-of-the-art GMR models. It resides in the 'Contrastive vs. Supervised Fine-Tuning Objectives' leaf, which contains only two papers total (including this one). This leaf sits within the broader 'Training Objective Design and Comparison' branch, indicating a relatively sparse research direction focused specifically on direct objective comparisons for LLM rerankers. The taxonomy reveals that while training objective design is an active area, head-to-head comparisons of CL versus SFT remain underexplored.

The taxonomy shows neighboring leaves addressing reinforcement learning hybrids, specialized loss functions, and alternative supervision sources. The 'Reinforcement Learning and Hybrid Training Approaches' leaf explores multi-objective optimization, while 'Specialized Loss Functions for Reranking' examines novel loss designs for ranking errors. The 'Alternative Supervision Signals' leaf investigates LLM annotations versus click data. The original paper diverges from these by focusing on foundational objective comparison rather than hybrid methods or supervision sources, and by extending the analysis to multimodal retrieval contexts where text and vision signals interact.

Among 24 candidates examined, the unified framework contribution (4 candidates, 0 refutable) appears relatively novel, with no clear prior work decomposing objectives into weight and direction components for LLM reranking. The MRB benchmark contribution (10 candidates, 1 refutable) shows more overlap, suggesting existing multimodal evaluation resources may partially cover this ground. The GMR models contribution (10 candidates, 0 refutable) appears novel in achieving state-of-the-art multimodal reranking performance. The limited search scope means these assessments reflect top-30 semantic matches and immediate citations, not exhaustive field coverage.

Given the sparse taxonomy leaf and limited refutation signals, the work appears to occupy a relatively underexplored niche at the intersection of objective comparison and multimodal reranking. The analysis is constrained by the 24-candidate search scope and may not capture all relevant prior work in adjacent areas like BERT-based objective studies or broader multimodal retrieval benchmarks. The framework and model contributions show stronger novelty signals than the benchmark component within this limited examination.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Unified framework for analyzing SFT and CL in LLM reranking

Description: The authors develop a unified framework (URL) that decomposes reranking loss functions into weight and direction components, enabling systematic comparison between supervised fine-tuning and contrastive learning. Through this decomposition, they demonstrate that SFT's superior performance stems primarily from its weight component, which provides stronger optimization signals than CL.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Large Language Models for Reranking: A Survey

URL: View paper

Brief Assessment

LLM Reranking Survey[5] appears to be a survey paper that mentions reranking frameworks in passing. The provided context contains only fragmentary references to reranking methods and does not present a unified analytical framework decomposing loss functions into weight and direction components as the original paper does.

---

2. Improving Fine-tuning of Language Models with an Emphasis on Isotropy and Rank

URL: View paper

Brief Assessment

Isotropy and Rank[11] focuses on improving text embeddings through isotropy enhancement and rank reduction for dense retrieval and sentence embedding tasks, not on comparing SFT versus CL for LLM reranking optimization.

---

3. Self-supervised scientific document recommendation based on contrastive learning

URL: View paper

Brief Assessment

Self-supervised Document Recommendation[9] focuses on contrastive learning for scientific document recommendation, not on comparing supervised fine-tuning versus contrastive learning for LLM-based reranking or developing a unified framework with weight/direction decomposition.

---

4. HMCL: Task-Optimal Text Representation Adaptation through Hierarchical Contrastive Learning

URL: View paper

Brief Assessment

HMCL[10] focuses on hierarchical contrastive learning for text representation adaptation across similarity, retrieval, and reranking tasks. It does not develop a unified framework decomposing reranking loss functions into weight and direction components to compare SFT versus CL.

Contribution 2: MRB benchmark for multimodal reranking evaluation

Description: The authors construct MRB (multimodal reranking benchmark), a comprehensive evaluation benchmark containing 40 test datasets spanning diverse modalities including single-modal, cross-modal, and fused-modal retrieval tasks. This benchmark enables rigorous assessment of universal multimodal reranking models across different domains and task types.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Wikido: A new benchmark evaluating cross-modal retrieval for vision-language models

URL: View paper

Brief Assessment

Wikido[20] focuses on cross-modal retrieval (image-to-text and text-to-image) for vision-language models, not multimodal reranking. The original paper's MRB benchmark evaluates reranking models across single-modal, cross-modal, and fused-modal tasks, which is a different problem domain.

---

2. Multibench: Multiscale benchmarks for multimodal representation learning

URL: View paper

Brief Assessment

Multibench[17] focuses on multimodal representation learning across diverse tasks (affective computing, robotics, healthcare) with 15 datasets and 10 modalities, while MRB specifically targets multimodal reranking evaluation with 40 test datasets for retrieval tasks. These serve fundamentally different purposes in the multimodal learning pipeline.

---

3. A survey on multimodal benchmarks: In the era of large ai models

URL: View paper

Brief Assessment

Multimodal Benchmarks Survey[14] focuses on benchmarks for evaluating multimodal large language models (MLLMs) across understanding, reasoning, generation, and application tasks. The original paper's MRB is specifically designed for multimodal reranking evaluation with 40 datasets spanning single-modal, cross-modal, and fused-modal retrieval tasks, which is a distinct focus not covered in the survey's taxonomy.

---

4. Polysemous visual-semantic embedding for cross-modal retrieval

URL: View paper

Brief Assessment

Polysemous Visual-Semantic Embedding[18] focuses on cross-modal retrieval using polysemous embeddings for image-text and video-text pairs, not on constructing comprehensive multimodal reranking benchmarks spanning single-modal, cross-modal, and fused-modal tasks.

---

5. A language-guided cross-modal semantic fusion retrieval method

URL: View paper

Brief Assessment

Cross-modal Semantic Fusion[13] focuses on retrieval methods for the WebQA dataset, not on constructing comprehensive benchmarks spanning diverse modalities and task types for reranking evaluation.

---

6. Mmdocir: Benchmarking multi-modal retrieval for long documents

URL: View paper

Brief Assessment

Mmdocir[16] focuses on multimodal document retrieval (page-level and layout-level retrieval from long documents), not reranking. The original paper's MRB benchmark evaluates reranking models that reorder retrieved candidates, while Mmdocir[16] evaluates retrieval systems that identify relevant pages/layouts from scratch.

---

7. SMIL: Multimodal Learning with Severely Missing Modality

URL: View paper

Brief Assessment

SMIL[19] focuses on multimodal learning with severely missing modalities during training/testing, not on multimodal retrieval benchmarks. The paper addresses a different problem domain (missing modality learning) rather than retrieval evaluation.

---

8. Mm-embed: Universal multimodal retrieval with multimodal llms

URL: View paper

Prior Art Analysis

Mm-embed[15] demonstrates prior work on comprehensive multimodal retrieval benchmarks. The candidate paper presents M-BEIR, a benchmark comprising 10 datasets with 16 diverse multimodal retrieval tasks across 4 domains, evaluated on 190k test queries with a unified candidate pool of 5.6m multimodal documents. This benchmark spans single-modal, cross-modal, and fused-modal retrieval scenarios similar to the original paper's MRB. The M-BEIR benchmark was published and used for training and evaluation before the original paper's submission, establishing prior art in comprehensive multimodal retrieval benchmarking that encompasses the same diversity of modalities and task types claimed as novel by the original authors.

Evidence

Evidence 1 - **Rationale**: This pair shows that M-BEIR (wei et al., 2023) already provided a comprehensive multimodal retrieval benchmark with diverse tasks before the original paper's MRB, challenging the novelty claim. - **Original**: we compile a new unified

benchmark called mrb (multimodal reranking benchmark, §5). Through analysis and comparison, we find that sft consistently outperforms cl for llm-based rerankers - **Candidate**: we evaluate models' universal multimodal retrieval capabilities using the m-beir dataset (wei et al., 2023), which is constructed from 10 datasets with 16 diverse multimodal retrieval tasks across 4 domains, as listed in appendix table 10. we train our models on the m-beir 1.1m training queries and ...

Evidence 2 - **Rationale**: This demonstrates that the framework for evaluating single-modal, cross-modal, and fused-modal (interleaved text-image) retrieval was already established by wei et al. (2023) in M-BEIR, predating the original paper's MRB benchmark. - **Original**: we introduce the mrb benchmark, comprising 40 datasets across single-, cross-, and fusedmodal retrieval, offering a comprehensive evaluation for universal multimodal reranking. - **Candidate**: in this work, we borrow the setting of universal multimodal retrieval from wei et al. (2023), where user queries and candidates may consist of text, image or interleaved text-image; i.e., q ∈ {qtxt, qimg, (qtxt, qimg)}; c ∈ {ctxt, cimg, (ctxt, cimg)}.

Evidence 3 - **Rationale**: This shows M-BEIR already implemented a unified evaluation approach across diverse multimodal datasets with a global candidate pool, similar to what the original paper claims as novel in MRB. - **Original**: To make comprehensive evaluations of multimodal reranking, we compile a new unified benchmark called mrb (multimodal reranking benchmark, §5). Through analysis and comparison, we find that sft consistently outperforms cl for llm-based rerankers - **Candidate**: following the global evaluation setting of the m-beir dataset, for each query, candidates are retrieved from a unified candidate pool of 5.6m multimodal documents spanning all 10 datasets. we report the average recall@5 (r@5) as the retrieval accuracy across all test queries in each dataset

---

9. Uniir: Training and benchmarking universal multimodal information retrievers

URL: View paper

Brief Assessment

Uniir[12] focuses on multimodal retrieval (not reranking) and introduces M-BEIR benchmark for retrieval tasks. The original paper's MRB is specifically designed for reranking evaluation with 40 test datasets, while Uniir[12]'s M-BEIR addresses retrieval with different task formulations and evaluation protocols.

---

10. Deep supervised cross-modal retrieval

URL: View paper

Brief Assessment

Deep Supervised Cross-modal[21] focuses on cross-modal retrieval (finding relevant data across modalities like image-to-text), not on reranking tasks. It does not present a benchmark for evaluating reranking models.

Contribution 3: GMR models achieving state-of-the-art multimodal reranking

Description: The authors develop GMR-3B and GMR-7B, instruction-aware multimodal LLM rerankers trained using supervised fine-tuning on approximately 1.5 million diverse query-document pairs. These models establish new state-of-the-art performance on the MRB benchmark, demonstrating the practical effectiveness of their SFT-based approach for universal multimodal reranking.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. E2LVLM:Evidence-Enhanced Large Vision-Language Model for Multimodal Out-of-Context Misinformation Detection

URL: View paper

Brief Assessment

E2LVLM[28] focuses on multimodal out-of-context misinformation detection using evidence-enhanced vision-language models, not on universal multimodal reranking across diverse query-document pairs as in the original paper's GMR models.

---

2. EarthGPT: A Universal Multimodal Large Language Model for Multisensor Image Comprehension in Remote Sensing Domain

URL: View paper

Brief Assessment

EarthGPT[22] focuses on remote sensing image comprehension tasks (scene classification, image captioning, visual grounding, object detection) rather than multimodal reranking for information retrieval. The technical approaches and application domains are fundamentally different.

---

3. Visual in-context learning for large vision-language models

URL: View paper

Brief Assessment

Visual In-context Learning[23] focuses on enhancing in-context learning capabilities for large vision-language models through visual demonstration retrieval and image summarization, not on developing instruction-aware multimodal rerankers for universal reranking tasks.

---

4. CAT: Enhancing Multimodal Large Language Model to Answer Questions in Dynamic Audio-Visual Scenarios

URL: View paper

Brief Assessment

CAT[24] focuses on audio-visual question answering using multimodal LLMs with clue aggregation, not on reranking tasks or training rerankers with supervised fine-tuning for information retrieval.

---

5. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training

URL: View paper

Brief Assessment

MLLM Strong Reranker[25] focuses on multimodal RAG with noise-injected training for generation tasks, while the original paper develops universal multimodal rerankers trained via supervised fine-tuning on diverse query-document pairs for ranking tasks.

---

6. LamRA: Large Multimodal Model as Your Advanced Retrieval Assistant

URL: View paper

Brief Assessment

LamRA[26] focuses on re-purposing generative LMMs for retrieval tasks with a two-stage training strategy, while the original paper develops instruction-aware rerankers using supervised fine-tuning on 1.5M query-document pairs specifically for reranking.

---

7. Conditional Sentence Generation and Cross-Modal Reranking for Sign Language Translation

URL: View paper

Brief Assessment

Sign Language Translation[30] focuses on sign language video translation using conditional sentence generation and cross-modal re-ranking for a specific domain task, not on developing universal multimodal reranking models trained on diverse query-document pairs across multiple modalities.

---

8. MM-R5: MultiModal Reasoning-Enhanced ReRanker via Reinforcement Learning for Document Retrieval

URL: View paper

Brief Assessment

MM-R5[29] focuses on reasoning-enhanced reranking via reinforcement learning with a two-stage training approach (SFT + RL), while the original paper develops GMR models using only supervised fine-tuning on diverse query-document pairs without explicit reasoning chains or RL components.

---

9. Mm-embed: Universal multimodal retrieval with multimodal llms

URL: View paper

Brief Assessment

Mm-embed[15] focuses on universal multimodal retrieval (bi-encoder retrievers), not reranking. While it explores zero-shot reranking with MLLMs, it does not present instruction-aware supervised fine-tuned reranking models like GMR.

---

10. V2Xum-LLM: Cross-Modal Video Summarization with Temporal Prompt Instruction Tuning

URL: View paper

Brief Assessment

V2Xum-LLM[27] focuses on video summarization tasks (video-to-video, video-to-text) using temporal prompts, not on multimodal reranking for information retrieval across diverse query-document pairs.

Appendix: Text Similarity Detection

Textual similarity detection checked 24 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Rethink Training of BERT Rerankers in Multi-Stage Retrieval Pipeline

Detected in: Core Task (sibling)

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Supervised Fine-Tuning or Contrastive Learning? Towards Better Multimodal LLM Reranking View paper
- [1] Multi-view-guided Passage Reranking with Large Language Models View paper
- [2] Rethink Training of BERT Rerankers in Multi-Stage Retrieval Pipeline View paper
- [3] FIRST: Faster Improved Listwise Reranking with Single Token Decoding View paper
- [4] ERank: Fusing Supervised Fine-Tuning and Reinforcement Learning for Effective and Efficient Text Reranking View paper
- [5] Large Language Models for Reranking: A Survey View paper
- [6] Can LLM Annotations Replace User Clicks for Learning to Rank? View paper
- [7] Make Large Language Model a Better Ranker View paper
- [8] EAReranker: Efficient Embedding Adequacy Assessment for Retrieval Augmented Generation View paper
- [9] Self-supervised scientific document recommendation based on contrastive learning View paper
- [10] HMCL: Task-Optimal Text Representation Adaptation through Hierarchical Contrastive Learning View paper
- [11] Improving Fine-tuning of Language Models with an Emphasis on Isotropy and Rank View paper
- [12] Uniir: Training and benchmarking universal multimodal information retrievers View paper
- [13] A language-guided cross-modal semantic fusion retrieval method View paper
- [14] A survey on multimodal benchmarks: In the era of large ai models View paper
- [15] Mm-embed: Universal multimodal retrieval with multimodal llms View paper
- [16] Mmdocir: Benchmarking multi-modal retrieval for long documents View paper
- [17] Multibench: Multiscale benchmarks for multimodal representation learning View paper
- [18] Polysemous visual-semantic embedding for cross-modal retrieval View paper
- [19] SMIL: Multimodal Learning with Severely Missing Modality View paper
- [20] Wikido: A new benchmark evaluating cross-modal retrieval for vision-language models View paper
- [21] Deep supervised cross-modal retrieval View paper
- [22] EarthGPT: A Universal Multimodal Large Language Model for Multisensor Image Comprehension in Remote Sensing Domain View paper
- [23] Visual in-context learning for large vision-language models View paper
- [24] CAT: Enhancing Multimodal Large Language Model to Answer Questions in Dynamic Audio-Visual Scenarios View paper
- [25] Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training View paper
- [26] LamRA: Large Multimodal Model as Your Advanced Retrieval Assistant View paper
- [27] V2Xum-LLM: Cross-Modal Video Summarization with Temporal Prompt Instruction Tuning View paper
- [28] E2LVLM:Evidence-Enhanced Large Vision-Language Model for Multimodal Out-of-Context Misinformation Detection View paper
- [29] MM-R5: MultiModal Reasoning-Enhanced ReRanker via Reinforcement Learning for Document Retrieval View paper
- [30] Conditional Sentence Generation and Cross-Modal Reranking for Sign Language Translation View paper