EAReranker: Efficient Embedding Adequacy Assessment for Retrieval Augmented Generation

Dongyang Zeng[1] **Yaping Liu**[1*] **Wei Zhang**[1]
Shuo Zhang[1*] **Xinwang Liu**[2] **Binxing Fang**[1]
[1]Guangzhou University, Guangzhou, China
[2]National University of Defense Technology, Changsha, China
zjzdy@e.gzhu.edu.cn, {ypliu, szhang18}@gzhu.edu.cn

Abstract

With the increasing adoption of Retrieval-Augmented Generation (RAG) systems for knowledge-intensive tasks, ensuring the adequacy of retrieved documents has become critically important for generation quality. Traditional reranking approaches face three significant challenges: substantial computational overhead that scales with document length, dependency on plain text that limits application in sensitive scenarios, and insufficient assessment of document value beyond simple relevance metrics. We propose EAReranker, an efficient embedding-based adequacy assessment framework that evaluates document utility for RAG systems without requiring access to original text content. The framework quantifies document adequacy through a comprehensive scoring methodology considering verifiability, coverage, completeness and structural aspects, providing interpretable adequacy classifications for downstream applications. EAReranker employs a Decoder-Only Transformer architecture that introduces embedding dimension expansion method and bin-aware weighted loss, designed specifically to predict adequacy directly from embedding vectors. Our comprehensive evaluation across four public benchmarks demonstrates that EAReranker achieves competitive performance with state-of-the-art plaintext rerankers while maintaining constant memory usage ($\sim$550MB) regardless of input length and processing 2-3x faster than traditional approaches. The semantic bin adequacy prediction accuracy of 92.85% LACC@10 and 86.12% LACC@25 demonstrates its capability to effectively filter out inadequate documents that could potentially mislead or adversely impact RAG system performance, thereby ensuring only high-utility information serves as generation context. These results establish EAReranker as an efficient and practical solution for enhancing RAG system performance through improved context selection while addressing the computational and privacy challenges of existing methods. The source code of EAReranker is available in https://github.com/zjzdy/EAReranker.

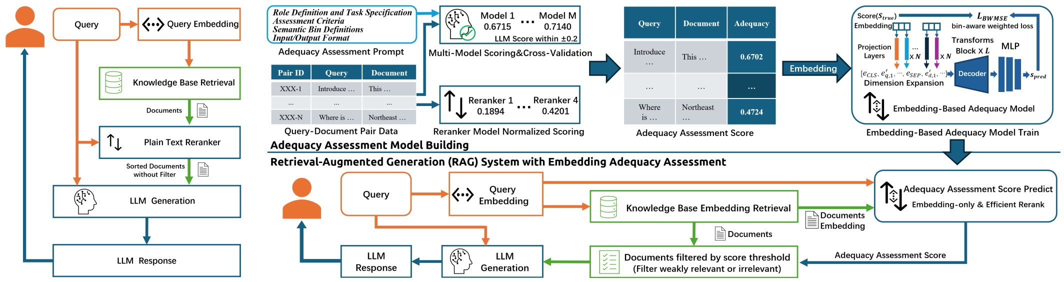
1 Introduction

Retrieval-Augmented Generation (RAG) [1] has emerged as a pivotal paradigm for enhancing Large Language Models (LLMs) in knowledge-intensive applications. By incorporating external knowledge as context, RAG systems mitigate hallucinations and enhance response accuracy. As these systems evolve, ensuring retrieved knowledge is not merely **relevant** but genuinely **adequate** in comprehensively and accurately addressing query requirements has become increasingly critical.

The canonical RAG workflow consists of retrieval and reranking stages [2, 3, 4, 5, 6]. While contemporary reranking methods effectively capture semantic relationships through cross-attention

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Traditional plantext-based Reranker RAG

(b) Overall Workflow of EAReranker: Embedding-based Document Adequacy Assessment for RAG

Figure 1: Traditional Reranker RAG vs EAReranker RAG.

mechanisms [7] or pre-trained language models [8], their direct dependency on original text introduces significant limitations.

These limitations manifest in three primary challenges. First, **substantial and unstable computational overhead** arises as computational requirements scale with text length, creating deployment barriers in resource-constrained environments and performance instability with variable document lengths. Second, **dependency on plain text** restricts deployment in scenarios where content exposure raises intellectual property or data governance concerns, impeding wider RAG adoption. Finally, a more fundamental constraint lies in the **lack of refined content value assessment**. Traditional reranking primarily evaluates "relevance," whereas effective RAG systems require assessment of "adequacy," which represents the substantive value and reliability of content in addressing the query yet remains challenging to quantify.

To address these limitations comprehensively, we propose EAReranker, an efficient embedding-based adequacy assessment model that predicts document adequacy solely from embedding vectors without accessing original text. Operating entirely in vector space, EAReranker leverages the semantic understanding capabilities of pre-trained embedding models to perform sophisticated adequacy assessment and reranking.

Our contributions include: **(1) An embedding-oriented adequacy assessment architecture:** EAReranker employs a Decoder-Only Transformer with **embedding dimension expansion** techniques and a **bin-aware weighted loss function**, eliminating dependence on plain text while effectively utilizing embedded knowledge. **(2) A principled adequacy assessment methodology:** We introduce a multi-dimensional semantic binning approach that quantifies document utility into discrete interpretable levels, validated through multiple LLMs to ensure consistent assessment quality. **(3) Empirical validation of efficiency and effectiveness:** Comprehensive experiments demonstrate that EAReranker achieves comparable performance to plaintext models while significantly reducing computational requirements and maintaining consistent efficiency regardless of document length.

2 Related Work

2.1 Language Embedding Models

Text embedding methods represent a fundamental technology in modern natural language processing, mapping textual data into continuous vector spaces where semantic relationships can be captured through geometric proximity. The evolution of these techniques has significantly enhanced the capabilities of information retrieval and ranking systems.

Embedding approaches have advanced from context-independent word representations to sophisticated contextual models. Early techniques like Word2Vec [9] and GloVe [10] established foundational vector space modeling principles, while Transformer-based architectures including BERT [11] and RoBERTa [12] later introduced dynamic contextual understanding. The development of sentence-level embedding frameworks such as Sentence-BERT [13] further refined semantic representation by optimizing vector spaces specifically for similarity assessment.

Recent architectural innovations have substantially expanded embedding capabilities. Models based on XLM-RoBERTa [14], such as bge-m3 [15] and jina-embeddings-v3 [16], demonstrate robust multilingual and multi-domain processing abilities. Advanced architectures including Qwen2-based [17] KaLM-embedding [18] and Gemma2-based [19] bge-multilingual-gemma2 [15] leverage increased parameter scales to achieve enhanced semantic modeling and extended context support [20].

The demonstrated capability of modern embeddings to effectively encode semantic content suggests that embedding vectors alone can sufficiently represent semantic adequacy information. Our EAReranker framework builds upon this foundation, conducting document adequacy evaluation exclusively within the embedding space while maintaining semantic fidelity.

2.2 Language Reranking Models

Document reranking serves as a critical component in RAG for refining initially retrieved candidates [2, 3, 4, 5]. Traditional approaches utilized lexical heuristics such as BM25 [21] and TF-IDF [22], which offer computational efficiency but exhibit limited semantic understanding.

Neural reranking models have introduced sophisticated semantic modeling through various architectural paradigms. Encoder-based approaches, exemplified by bge-reranker-v2-m3 [15] and jina-reranker-v2-base-multilingual [23], process query-document pairs using token-level cross-attention mechanisms. Decoder-Only architectures, including bge-reranker-v2-gemma [15] and lb-reranker [8], leverage autoregressive modeling over combined sequences to capture complex dependencies.

An alternative approach involves directly prompting large language models for ranking tasks, as demonstrated by RankGPT [6] and UPR [24]. While this method offers flexibility, it requires substantial computational resources and full text access.

Current reranking approaches face two significant limitations. First, they require direct access to query and document text, resulting in computational costs that scale with input length. Second, they primarily focus on relevance assessment rather than comprehensive adequacy evaluation for RAG.

Our proposed EAReranker addresses these limitations by operating solely on embedding vectors, enabling efficient computation and privacy preservation while maintaining competitive performance. This approach represents a significant advancement in developing practical and scalable solutions for document adequacy assessment in RAG.

3 Embedding Adequacy Assessment Framework

3.1 Framework Overview

Figure 1 contrasts traditional reranker-based RAG systems with our proposed EAReranker approach. While conventional rerankers perform detailed query-document interaction requiring computational resources proportional to text length, EAReranker operates exclusively on fixed-dimension embedding vectors, enabling efficient adequacy assessment without accessing original text content. This design strategically addresses three critical limitations of existing approaches: **(1) Computational invariance:** Processing fixed-dimension embeddings decouples runtime and memory requirements. **(2) Avoidance of dependence on original text:** By operating solely on embeddings, our method mitigates issues arising from direct text exposure or handling. **(3)Adequacy-centric evaluation:** The model learns to assess multi-dimensional adequacy factors rather than relying solely on relevance.

A critical distinction underlying our work is that between *relevance* and *adequacy*. Traditional retrieval systems prioritize relevance—measuring topical alignment between queries and documents through lexical or semantic similarity. However, for RAG systems performing complex generation tasks, relevance alone is insufficient. Adequacy measures a document's *functional utility* in enabling the generation model to produce complete, accurate, and well-structured responses. For instance, a document containing only the query keywords may achieve high relevance scores but provide minimal generation value, while a comprehensive, well-structured explanation may score lower on similarity metrics yet offer substantially higher utility for answer generation. Our adequacy framework evaluates documents across four complementary dimensions to capture this multi-faceted utility.

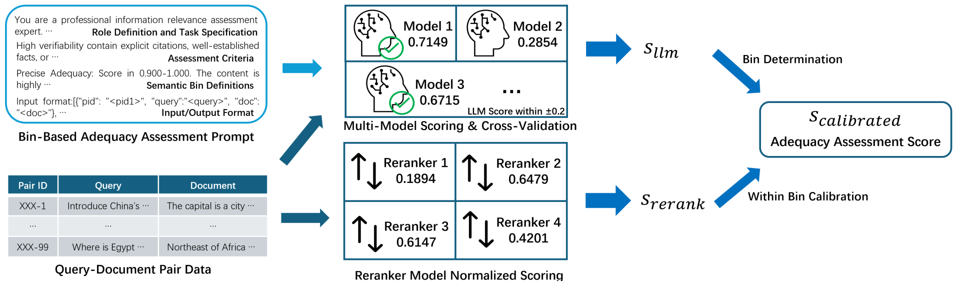


Figure 2: Pipeline of Adequacy Assessment Annotation.

3.2 Problem Formal Characterization

Traditional document reranking is typically defined as learning a function $f(q, d)$ that evaluates the relevance between the original query text q and document text d, and ranks documents accordingly. We redefine this task as learning a model $\mathcal{M}$ that predicts a scalar score $s' \in [0, 1]$ reflecting the "adequacy" of document d for query q, using only their embedding vectors $\mathbf{e}_q \in \mathbb{R}^d$ and $\mathbf{e}_d \in \mathbb{R}^d$ generated by pre-trained embedding models:

$$f(q, d) \to s \quad \text{(Traditional)} \quad \Rightarrow \quad \mathcal{M}(\mathbf{e}_q, \mathbf{e}_d) \to s' \quad \text{(Ours)}, \tag{1}$$

where documents $\{d_i\}$ are then reranked based on these predicted adequacy scores s'_i.

More precisely, given a query embedding $\mathbf{e}_q$ and a collection of document embeddings $\{\mathbf{e}_{d_i}\}_{i=1}^N$ obtained during the initial retrieval phase, we aim to train a model $\mathcal{M}(\mathbf{e}_q, \mathbf{e}_{d_i}) \to s'_i$ that approximates the true adequacy score s_i with high fidelity. The final document ranking is determined by sorting the documents in descending order of their predicted adequacy scores, such that:

$$s'_{i_1} \geq s'_{i_2} \geq ... \geq s'_{i_N}, \tag{2}$$

where $(i_1, i_2, ..., i_N)$ represents the indices of the documents after ranking. A critical property of our framework is score interpretability—documents with similar predicted scores should demonstrate comparable utility across query-document pairs, enabling meaningful threshold-based filtering.

To effectively assess adequacy from embeddings, our model must exhibit four essential capabilities: extraction of fine-grained semantic information from fixed-dimension vectors; exclusive operation on embeddings without requiring original text; evaluation across multiple adequacy dimensions beyond simple relevance; and computational efficiency invariant to document length. These requirements guide our architectural design and implementation strategies detailed in subsequent sections, with Section 4 presenting our annotation methodology and Section 5 describing our model architecture.

4 Adequacy Assessment Annotation

This section establishes a formal framework for evaluating document adequacy in RAG systems, transcending traditional relevance metrics to capture the multidimensional utility of retrieved content as generation context. We introduce a rigorous methodology for quantifying document adequacy through interpretable value classification, providing both theoretical foundations and high-quality labeled datasets for training our proposed EAReranker model. As illustrated in Figure 2, our annotation pipeline integrates semantic bin-based standards, multi-model evaluation, and cross-model validation to ensure consistency across diverse query-document domains.

4.1 Semantic Bin-Based Adequacy Scoring

Adequacy assessment represents a fundamental advancement beyond conventional relevance evaluation. While relevance primarily quantifies semantic proximity between queries and documents, adequacy measures the substantive utility of documents as generation context. This distinction is critical: relevance indicates topical alignment, whereas adequacy evaluates whether a document provides reliable information that comprehensively addresses query requirements.

Table 1: Semantic Binning Scheme for Adequacy Scoring.

Semantic Bin	Score Range	Verifiability	Need Coverage	Evidence Completeness	Structure Suitability	Description
Precise Adequacy	$[0.90, 1.00]$	High	High	High	High	Optimal context
High Adequacy	$[0.75, 0.90)$	High	High	High	Medium	High-quality context
Middle Adequacy	$[0.50, 0.75)$	Medium	Medium	Medium	Medium	Usable context with supplementation
Marginal Relevance	$[0.25, 0.50)$	Low	Low	Low	Low	Marginally usable context
Weak Relevance	$[0.10, 0.25)$	Very Low	Low	Very Low	-	Negligible value context
Irrelevance	$[0.00, 0.10)$	-	Very Low	-	-	Unusable context

We formalize adequacy assessment through four complementary dimensions: **(1) Verifiability**: Quantifies information reliability and factual foundation. High verifiability documents present explicitly cited or well-established facts, while low verifiability documents contain predominantly unsubstantiated claims. **(2) Need Coverage**: Evaluates comprehensive addressing of query requirements. High coverage documents provide complete responses to information needs, whereas low coverage documents may contain superficially relevant terminology without addressing query intent. **(3) Evidence Completeness**: Assesses logical coherence and supporting evidence. Documents with high completeness present well-structured arguments with sufficient substantiation, while those with low completeness exhibit logical discontinuities or insufficient support. **(4) Structure Suitability**: Measures alignment between document presentation and required content format. High suitability documents present information in readily utilizable formats, while low suitability documents require significant transformation.

We discretize the adequacy spectrum into six distinct bins spanning [0,1], as detailed in Table 1. Documents in the **Precise Adequacy** bin ([0.90, 1.00]) represent optimal context with high performance across all dimensions. **High Adequacy** documents ([0.75, 0.90)) contain all required information with minimal redundancy. **Middle Adequacy** documents ([0.50, 0.75)) address main issues but require supplementation. **Marginal Relevance** documents ([0.25, 0.50)) cover only peripheral aspects of the query. **Weak Relevance** documents ([0.10, 0.25)) provide minimal substantive information. **Irrelevance** documents ([0.00, 0.10)) have no discernible relationship to query requirements.

This structured binning provides clear operational thresholds for RAG systems. Documents scoring above 0.75 can be directly incorporated, while those below 0.25 should generally be excluded to prevent hallucination or inaccuracy in generated content. The granular discrimination between bins enables automated optimization of context selection, significantly enhancing generation quality.

4.2 Multi-Model Semantic Bin Scoring

Constructing high-quality training data necessitates a robust annotation methodology. We employ a multi-model scoring framework that leverages diverse large language models under a principled cross-validation scheme to ensure reliable and consistent scoring across heterogeneous query-document pairs. Our framework employs a cross-validation architecture that aggregates assessments from multiple LLMs with distinct architectures and training paradigms. This approach mitigates individual model biases and captures diverse evaluation perspectives, enhancing annotation reliability particularly for ambiguous cases.

Effective LLM-based scoring relies on carefully engineered prompts encompassing: (1) Professional role definition with explicit evaluation criteria, (2) Four-dimensional adequacy framework articulation, (3) Precise semantic bin definitions with boundary cases, and (4) Structured input/output formats. We incorporate exemplar cases demonstrating assessment criteria across diverse scenarios, enhancing model understanding of nuanced adequacy distinctions.

The scoring process follows Algorithm 1, which implements a hierarchical validation mechanism to identify consistent model assessments. The algorithm's innovation lies in its combinatorial approach, systematically evaluating all possible combinations of four model scores to find consistent subsets. Specifically, it begins with four LLM scores and checks consistency (tolerance 0.2); if unsuccessful, additional models are consulted and all four-score combinations are evaluated to find self-consistent subsets. This approach offers several advantages: **flexibility in finding reliable assessments** without requiring all models to agree, **computational efficiency** through progressive validation, and **effective handling of outlier scores** by identifying alternative combinations that provide consistent assessments.

Algorithm 1: Multi-Model Semantic Bin Scoring.

Input: Query-document pair (q, d), LLM models $\mathcal{M} = \{M_1, ..., M_m\}$
Output: LLM adequacy score s_{llm}

$\mathcal{S} \leftarrow \{\text{Score}(M_i, q, d) \mid i \in \{1, 2, 3, 4\}\}$; `/* Initialize with first 4 scores */`
$\mu \leftarrow \text{Mean}(\mathcal{S})$;
if $\max_{s \in \mathcal{S}} |s - \mu| \leq 0.2$ **then**
  return μ;
for $i \leftarrow 5$ **to** m **do**
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{\text{Score}(M_i, q, d)\}$;
  foreach $\mathcal{S}_{subset} \in$ *Combinations*$(\mathcal{S}, 4)$ **do**
    $\mu_{subset} \leftarrow \text{Mean}(\mathcal{S}_{subset})$;
    if $\max_{s \in \mathcal{S}_{subset}} |s - \mu_{subset}| \leq 0.2$ **then**
      return μ_{subset};
return $\text{Mean}(\mathcal{S})$;

4.3 Within-Bin Score Calibration Methodology

While multi-model validation provides reliable bin assignments, we observed that LLMs tend to produce clustered scores within each bin, limiting intra-bin differentiation. To enhance score granularity while preserving semantic boundaries, we developed a within-bin calibration methodology that incorporates auxiliary ranking signals from diverse reranking models.

The calibration process begins with normalization of reranker scores across the dataset:

$$s_{R_j}^{norm} = \frac{s_{R_j} - \min(s_{R_j})}{\max(s_{R_j}) - \min(s_{R_j})}. \tag{3}$$

These normalized scores are aggregated to produce a composite reranking signal:

$$s_{rerank} = \langle s_{R_j}^{norm} \rangle_{j=1}^{|\mathcal{R}|}. \tag{4}$$

For documents with LLM-assigned scores s_{llm} within bin $[l_i, h_i]$, the calibration formula applies:

$$s_{calibrated} = 0.5 \times ((s_{llm} - l_i)/(h_i - l_i) + s_{rerank}) \times (h_i - l_i) + l_i. \tag{5}$$

This calibration methodology preserves semantic bin boundaries while enhancing intra-bin differentiation, transforming discrete score clusters into continuous distributions and maintaining semantic consistency with original assessments.

Through this comprehensive annotation framework, we construct large-scale training datasets with high-quality adequacy scores. The methodology's ability to produce consistent, fine-grained adequacy annotations while preserving semantic interpretability represents a significant advancement in automated document assessment for RAG.

5 Embedding-Based Adequacy Assessment Model

We propose EAReranker, an efficient embedding-based model for document adequacy assessment that operates exclusively in the vector space, as illustrated in Figure 3. The model leverages a specialized Decoder-Only Transformer architecture to evaluate adequacy using only query and document embedding, eliminating the need for access to original text while maintaining competitive performance with traditional text-based approaches. The architecture consists of two primary components:

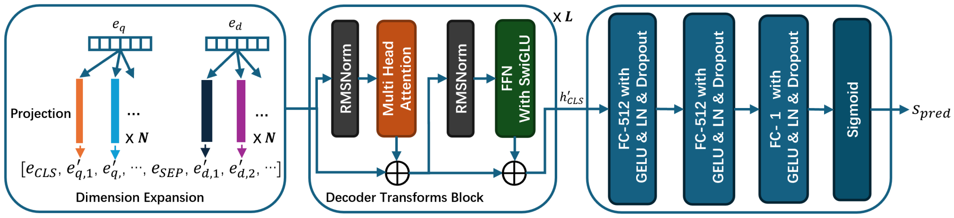


Figure 3: Overall of Adequacy Assessment Model.

Embedding Dimension Expansion: To enable rich semantic feature extraction from fixed-dimensional embeddings, we transform single embedding vectors into sequential representations through independent projection layers. Given query embedding $\mathbf{e}_q$ and document embedding $\mathbf{e}_d$, the expansion is performed as:

$$\mathbf{e}'_{q,i} = L_{q,i}(\mathbf{e}_q) = \mathbf{W}_{q,i}\mathbf{e}_q + \mathbf{b}_{q,i}, \quad \mathbf{e}'_{d,i} = L_{d,i}(\mathbf{e}_d) = \mathbf{W}_{d,i}\mathbf{e}_d + \mathbf{b}_{d,i}, \tag{6}$$

where each projection layer $L_{q,i}$ and $L_{d,i}$ is initialized independently to capture distinct semantic aspects. The expanded sequence combines these transformed embeddings with trainable special tokens embedding vectors $\mathbf{e}_{CLS}$ and $\mathbf{e}_{SEP}$: $[\mathbf{e}_{CLS}, \mathbf{e}'_{q,1}, \ldots, \mathbf{e}'_{q,N}, \mathbf{e}_{SEP}, \mathbf{e}'_{d,1}, \ldots, \mathbf{e}'_{d,N}]$.

Transformer Processing: The expanded sequence is processed through L stacked Decoder layers with Multi-Head Self-Attention and Feed-Forward Networks. The [CLS] token output is passed through a three-layer MLP head to predict the final adequacy score.

The model is trained using a bin-aware weighted loss function that emphasizes accurate prediction near semantic bin boundaries:

$$\mathcal{L}_{BWMSE}(S_{true}, S_{pred}) = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot (s_{true,i} - s_{pred,i})^2, \tag{7}$$

where weights w_i are computed based on prediction deviation from bin boundaries.

This architecture effectively addresses the key challenges of embedding-based adequacy assessment while maintaining high computational efficiency and privacy preservation.

6 Experiments

We present a comprehensive evaluation of EAReranker's effectiveness in embedding-based adequacy assessment and reranking. Our experiments assess the model's ranking performance across public benchmarks, adequacy classification accuracy, and computational efficiency, demonstrating its ability to achieve competitive results while operating exclusively on embedding vectors.

6.1 Experimental Settings

Our evaluation utilized a dataset comprising 1 million query-document pairs, partitioned into training (80%) and testing (20%) sets. Vector representations were generated using established embedding models: bge-m3 (1024-dimensional)[15], jina-embeddings-v3 (jina-v3, 1024-dimensional) [16], gte-multilingual-base (gte-base, 768-dimensional) [18], and KaLM-embedding-multilingual-mini-instruct-v1.5 (KaLM, 896-dimensional) [20]. We deliberately excluded higher-dimensional models like NV-Embed-v2 (4096-dimensional) [25] to focus on commonly dimensions in RAG.

The evaluation framework compares EAReranker against multiple baselines: traditional lexical models (BM25 [21]), embedding cosine similarity methods, and plaintext-based reranking models including gte-multilingual-reranker-base [18], bge-reranker-v2-m3 [15], jina-reranker-v2-base-multilingual [23], and lb-reranker-v1.0 [8].

EAReranker employs a stacked Transformer architecture with 4 layers and an embedding dimension expansion factor of 4, balancing representational capacity with computational efficiency. Training utilized the AdamW optimizer (batch size 256, learning rate 1e-5) for 50 epochs with early stopping.

Table 2: Comparative Case Study: Traditional Retrieval Scores Versus Adequacy Assessment.

Query	Document	bge-m3	gte-base	bge-r2	gte-rb	Adequacy
Introduce China's capital.	Beijing City (Beijing), abbreviated as Jing, historically known as Yanjing and Beiping ...	0.7305	0.7983	2.4641	0.3608	0.9261
Introduce China's capital.	The capital is a nation's most important political center. Throughout Chinese history ...	0.6684	0.7723	2.4453	0.9229	0.8749
Introduce China's capital.	Beijing is the capital of the People's Republic of China, categorized as a super first-tier city.	0.6717	0.7634	0.7334	0.2303	0.5763
Introduce China's capital.	Beijing is the capital	0.7858	0.8427	2.4805	0.3511	0.3102
Introduce China's capital.	Introduce China's capital	1.0000	1.0000	10.1562	3.2539	0.1740
The formula for multiplying two numbers to 10.	1*5=5; 2*5=10; 3*5=15; 4*5=20 ...	0.6137	0.6300	0.1333	0.3533	0.8137
Where is Egypt located in Africa? A. Northeast B. Southwest	A	0.3278	0.5939	-4.7656	0.6436	0.1523
What is China's capital city?	Beijing	0.6024	0.7688	2.8633	0.6348	0.1893
What is China's capital city?	The railway runs east to west	0.3977	0.4263	-10.7734	-0.4365	0.0324

6.2 Metrics

Our evaluation framework incorporates both ranking effectiveness and classification accuracy metrics:

Ranking Metrics: Normalized Discounted Cumulative Gain (NDCG@10) evaluates ranking quality considering both relevance and position. Mean Reciprocal Rank (MRR) quantifies the position of the first relevant document.

Adequacy Assessment Metrics: ACC25 measures the proportion of samples where the absolute error between predicted and true scores is within 0.25, directly reflecting bin classification accuracy. LACC@25 and LACC@10 evaluate binary classification accuracy at thresholds of 0.25 and 0.10 respectively, indicating the model's ability to distinguish relevant from irrelevant documents. These metrics are particularly significant for RAG systems that require effective filtering mechanisms to prevent noise contamination.

6.3 Dataset

We curated a comprehensive adequacy assessment dataset by augmenting the `bge-m3-data` [15] comprising diverse query-document pairs with additional samples sourced from multiple heterogeneous collections [26, 27, 28, 29, 30, 31, 32, 33, 34, 35]. This aggregated dataset spans a wide spectrum of domains and query intents, encompassing fact verification, specialized knowledge retrieval, etc.

To ensure annotation quality, we adopted a multi-model scoring framework employing seven LLMs [36, 37, 38, 39, 17, 40, 41] with diverse architectures and training regimes. Each query-document pair was scored multiple times, with cross-validation and iterative selection to reduce model-specific bias and improve consistency, particularly for ambiguous boundary samples.

Table 2 contrasts traditional retrieval scores with our bin-based adequacy assessment across representative examples. Here, bge-r2 and gte-rb denote reranking scores from bge-reranker-v2-m3 [15] and gte-multilingual-reranker-base [18], respectively; bge-m3 and gte-base represent cosine similarity scores. The comparison illustrates how traditional methods often conflate lexical or semantic similarity with true informational value. For instance, exact query repetition achieves maximal similarity scores but receives low adequacy ratings, reflecting its lack of substantive content. Conversely, documents offering structured, detailed information receive high adequacy despite lower similarity scores, highlighting traditional models' inability to fully capture content utility for generation tasks.

6.4 Performance Experiments

Ranking Effectiveness. Evaluation across four public benchmarks (FEVER [30], NFCorpus [28], DuRetrieval [42], and T2Ranking [43]) reveals consistent performance patterns, as shown in Table 3. Traditional lexical methods demonstrate the weakest performance, with embedding similarity methods showing significant improvements in NDCG and MRR. Plaintext-based reranking models achieve superior performance. However, EAReranker maintains competitive results within 0.54%-1.30% of the best plaintext model. This performance is particularly noteworthy given that EAReranker

Table 3: Ranking Performance on Public Benchmarks (%).

Type	Model	FEVER		NFCorpus		DuRetrieval		T2Ranking	
		NDCG@10	MRR	NDCG@10	MRR	NDCG@10	MRR	NDCG@10	MRR
Lexical	BM25	48.09	32.97	32.08	43.96	19.47	21.03	46.34	35.89
Embedding Similarity	Cosine (gte-base)	92.11	94.19	36.66	56.60	87.54	93.38	84.71	92.60
	Cosine (bge-m3)	81.37	85.64	31.41	52.17	83.96	90.77	81.37	90.30
	Cosine (jina-v3)	89.05	91.28	36.44	55.30	83.17	89.72	83.16	91.53
	Cosine (KaLM)	86.54	89.65	25.92	39.31	80.23	86.49	79.78	87.61
Plaintext Rerankers	gte-reranker-base	94.83	96.12	38.79	59.25	89.37	94.81	87.06	94.37
	bge-reranker-v2-m3	95.26	96.78	39.41	60.83	90.16	95.24	87.85	95.41
	jina-reranker-v2	93.74	95.82	38.04	58.96	88.65	93.94	86.34	93.85
	lb-reranker-v1.0	95.11	96.44	39.28	60.71	89.82	94.76	87.52	94.93
Embedding Reranker	EAReranker (gte-base)	94.36	95.87	38.14	58.93	88.96	94.23	86.47	93.91
	EAReranker (bge-m3)	94.72	96.21	38.76	59.32	89.57	94.82	87.11	94.68
	EAReranker (jina-v3)	94.51	96.08	38.51	59.14	89.23	94.56	86.89	94.32
	EAReranker (KaLM)	93.84	95.43	37.65	58.46	88.54	93.95	86.12	93.57

Table 4: EAReranker Adequacy Performance(%).

Embedding	ACC25	LACC@25	LACC@10
gte-base	81.98	84.17	91.32
bge-m3	84.28	86.12	92.85
jina-v3	83.52	85.42	92.27
KaLM	83.09	85.06	91.98

Table 5: Ablation Performance on bge-m3 (%).

Configuration	ACC25	LACC@10
Complete model	84.28	92.85
w/o Dimension Expansion	80.58 (-3.70)	89.62 (-3.23)
w/o Bin-Aware Loss	81.61 (-2.67)	90.65 (-2.20)
w/o Score Calibration	82.95 (-1.33)	91.98 (-0.87)

Table 6: Inference Computational Efficiency Comparison.

Model (Max Length)	VRAM(MB)	Inference Time(s)
gte-reranker-base (8K)	1209-2225	0.1697-0.3506
bge-reranker-v2-m3 (8K)	2176-2527	0.1312-0.5468
jina-reranker-v2 (1K)	547-603	0.1846-0.2128
lb-reranker-v1.0 (128K)	967-8441	0.3479-5.9219
EAReranker (gte-base)	544	0.1111
EAReranker (bge-m3)	550	0.1128
EAReranker (jina-v3)	550	0.1127
EAReranker (KaLM)	547	0.1124

processes only fixed-dimensional embedding vectors rather than full text. Additional experimental information is provided in Appendix B.

Adequacy Assessment Capability. Table 4 presents the model's performance on adequacy classification metrics. EAReranker achieves 84.28% ACC25 using bge-m3 embeddings, with LACC@10, reaching 92.85% and LACC@25 achieving 86.12%. These results demonstrate effective discrimination between adequacy levels, particularly in identifying irrelevant or marginal documents that should be filtered from RAG contexts.

Computational Efficiency. Table 6 compares the computational efficiency of EAReranker with traditional reranking models. EAReranker demonstrates significant advantages in resource utilization, maintaining consistent memory usage (∼550MB) regardless of input length, while traditional models exhibit substantial variation (up to 8441MB for lb-reranker-v1.0).

In terms of inference speed, EAReranker processes queries in 0.11s-0.13s, consistently faster than traditional models, especially for longer documents. Unlike traditional approaches with fixed maximum sequence lengths, EAReranker's processing capability depends only on the underlying embedding model's input capacity, providing greater flexibility for varied document lengths.

Component Contribution Analysis. Ablation studies confirm the importance of EAReranker's architectural components, as shown in Table 5. Removing the embedding dimension expansion strategy results in significant performance degradation (3.70% in ACC25, 3.23% in LACC@10), demonstrating its essential role in extracting fine-grained semantic features from compressed embeddings. Similarly, the bin-aware weighted loss proves crucial for learning precise adequacy boundaries, with its removal causing notable performance reduction (2.67% in ACC25, 2.20% in LACC@10). Our score calibration methodology also contributes 1.33% improvement in ACC25, confirming its effectiveness in enhancing within-bin score granularity.

The experimental results collectively validate EAReranker's effectiveness as an embedding-based adequacy assessment model for RAG systems. The model maintains competitive ranking performance while operating exclusively on embedding vectors, effectively distinguishes between documents, and significantly reduces computational requirements compared to traditional approaches.

7 Conclusion

This paper introduces a novel framework for document adequacy assessment in RAG systems, shifting from traditional relevance-centric metrics to a comprehensive multi-dimensional evaluation framework that encompasses *verifiability*, *need coverage*, *evidence completeness*, and *structure suitability*. Through our semantic binning methodology and multi-model scoring approach, we provide a quantification of retrieved documents' intrinsic utility as generative context.

Our proposed EAReranker operates exclusively on embedding vectors without accessing original text, achieving comparable ranking performance to state-of-the-art plaintext models while significantly reducing computational overhead. This demonstrates the viability of embedding-based methods for efficient and plaintext-preserving RAG deployments in resource-constrained environments.

Despite promising empirical results, several challenges remain. The current semantic bin segmentation and integration of adequacy dimensions rely on empirically-informed heuristic criteria that require further theoretical research and discussion. Additionally, our reliance on multi-model LLM annotations introduces potential biases that warrant systematic investigation through more robust validation methodologies. Future research directions include refining calibration techniques for enhanced scoring granularity, and adapting the framework for domain-specific applications to improve generation factuality. Exploring alternative embedding expansion mechanisms and model architectures may yield further efficiency gains and performance improvements.

In conclusion, this work establishes fundamental theoretical and practical contributions toward embedding-based adequacy assessment, advancing the development of semantically nuanced, computationally efficient, and trustworthy retrieval-augmented generation systems.

8 Acknowledgements

This work is supported by National Natural Science Foundation of China (62372124).

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020, pp. 9459–9474.

[2] N. Hossain, M. Ghazvininejad, and L. Zettlemoyer, "Simple and effective retrieve-edit-rerank text generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2532–2538.

[3] G. Geigle, J. Pfeiffer, N. Reimers, I. Vulić, and I. Gurevych, "Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 503–521, 2022.

[4] M. Glass, G. Rossiello, M. F. M. Chowdhury, A. Naik, P. Cai, and A. Gliozzo, "Re2g: Retrieve, rerank, generate," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, 2022, pp. 2701–2715.

[5] S. Kongyoung, C. Macdonald, and I. Ounis, "monoqa: Multi-task learning of reranking and answer extraction for open-retrieval conversational question answering," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 7207–7218.

[6] W. Sun, L. Yan, X. Ma, P. Ren, D. Chen, H. Ren, Z. Ren, and M. de Rijke, "Is chatgpt good at search? investigating large language models as re-ranking agents," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 14 918–14 937.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017, pp. 5998–6008.

[8] lightblue. (2025) lb-reranker-0.5b-v1.0. [Online]. Available: https://huggingface.co/lightblue/lb-reranker-0.5B-v1.0

[9] K. W. Church, "Word2vec," *Natural Language Engineering*, vol. 23, no. 1, pp. 155–162, 2017.

[10] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[13] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[14] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451.

[15] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Liu, C. J. Li, and Z. Yang, "Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," 2024.

[16] S. Sturua, I. Mohr, M. K. Akram, A. Bhardwaj, J. Kimmig, N. Fiedler, A. El-Kishky, M. Seo, J. Chen, M. Sclar, S. Chandrasekar, M. Josifoski, and G. Vassilev, "jina-embeddings-v3: Multilingual embeddings with task lora," 2024.

[17] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, "Qwen2.5 technical report," 2024.

[18] Z. Li, X. Zhang, Y. Zhang *et al.*, "Towards general text embeddings with multi-stage contrastive learning," 2023.

[19] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, J. Ferret, P. Liu, P. Tafti, A. Friesen, M. Casbon, S. Ramos, R. Kumar, C. L. Lan, S. Jerome, A. Tsitsulin, N. Vieillard, P. Stanczyk, S. Girgin, N. Momchev, M. Hoffman, S. Thakoor, J.-B. Grill, B. Neyshabur, O. Bachem, A. Walton, A. Severyn, A. Parrish, A. Ahmad, A. Hutchison, A. Abdagic, A. Carl, A. Shen, A. Brock, A. Coenen, A. Laforge, A. Paterson, B. Bastian, B. Piot, B. Wu, B. Royal, C. Chen, C. Kumar, C. Perry, C. Welty, C. A. Choquette-Choo, D. Sinopalnikov, D. Weinberger, D. Vijaykumar, D. Rogozińska, D. Herbison, E. Bandy, E. Wang, E. Noland, E. Moreira, E. Senter, E. Eltyshev, F. Visin, G. Rasskin, G. Wei, G. Cameron, G. Martins, H. Hashemi, H. Klimczak-Plucińska, H. Batra, H. Dhand, I. Nardini, J. Mein, J. Zhou, J. Svensson, J. Stanway, J. Chan, J. P. Zhou, J. Carrasqueira, J. Iljazi, J. Becker, J. Fernandez, J. van Amersfoort, J. Gordon, J. Lipschultz, J. Newlan, J. yeong Ji, K. Mohamed, K. Badola, K. Black, K. Millican, K. McDonell, K. Nguyen, K. Sodhia, K. Greene, L. L. Sjoesund, L. Usui, L. Sifre, L. Heuermann, L. Lago, L. McNealus, L. B. Soares, L. Kilpatrick, L. Dixon, L. Martins, M. Reid, M. Singh, M. Iverson, M. Görner, M. Velloso, M. Wirth, M. Davidow, M. Miller, M. Rahtz, M. Watson, M. Risdal, M. Kazemi, M. Moynihan, M. Zhang, M. Kahng, M. Park, M. Rahman, M. Khatwani, N. Dao, N. Bardoliwalla, N. Devanathan, N. Dumai, N. Chauhan, O. Wahltinez, P. Botarda, P. Barnes, P. Barham, P. Michel, P. Jin, P. Georgiev, P. Culliton, P. Kuppala, R. Comanescu, R. Merhej, R. Jana, R. A. Rokni, R. Agarwal, R. Mullins, S. Saadat, S. M. Carthy, S. Cogan, S. Perrin,

S. M. R. Arnold, S. Krause, S. Dai, S. Garg, S. Sheth, S. Ronstrom, S. Chan, T. Jordan, T. Yu, T. Eccles, T. Hennigan, T. Kocisky, T. Doshi, V. Jain, V. Yadav, V. Meshram, V. Dharmadhikari, W. Barkley, W. Wei, W. Ye, W. Han, W. Kwon, X. Xu, Z. Shen, Z. Gong, Z. Wei, V. Cotruta, P. Kirk, A. Rao, M. Giang, L. Peran, T. Warkentin, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, D. Sculley, J. Banks, A. Dragan, S. Petrov, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, S. Borgeaud, N. Fiedel, A. Joulin, K. Kenealy, R. Dadashi, and A. Andreev, "Gemma 2: Improving open language models at a practical size," 2024. [Online]. Available: https://arxiv.org/abs/2408.00118

[20] X. Hu, Z. Shan, X. Zhao *et al.*, "Kalm-embedding: Superior training data brings a stronger embedding model," 2025.

[21] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *Proceedings of the Third Text REtrieval Conference (TREC-3)*. NIST, 1995, pp. 109–126.

[22] D. A. Grossman and O. Frieder, *Information Retrieval: Algorithms and Heuristics*, 2nd ed., ser. The Information Retrieval Series. Dordrecht: Springer, 2004, vol. 15.

[23] jinaai. (2024) jina-reranker-v2-base-multilingual. [Online]. Available: https://huggingface.co/jinaai/jina-reranker-v2-base-multilingual

[24] D. Sachan, M. Lewis, M. Joshi, A. Aghajanyan, W.-t. Yih, J. Pineau, and L. Zettlemoyer, "Improving passage retrieval with zero-shot question generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022, pp. 3781–3797.

[25] C. Lee, R. Roy, M. Xu, J. Raiman, M. Shoeybi, B. Catanzaro, and W. Ping, "Nv-embed: Improved techniques for training llms as generalist embedding models," *arXiv preprint arXiv:2405.17428*, 2024.

[26] H. Wachsmuth, M. Trenkmann, B. Stein, and G. Engels, "A review corpus for argumentation analysis," in *Computational Linguistics and Intelligent Text Processing*, ser. Lecture Notes in Computer Science, vol. 8404. Springer Berlin Heidelberg, 2014, pp. 115–127.

[27] D. Hoogeveen, K. M. Verspoor, and T. Baldwin, "Cqadupstack: A benchmark data set for community question-answering research," in *Proceedings of the 20th Australasian Document Computing Symposium*, ser. ADCS '15. Association for Computing Machinery, 2015.

[28] V. Boteva, D. Gholipour, A. Sokolov, and S. Riezler, "A full-text learning to rank dataset for medical information retrieval," in *Advances in Information Retrieval*, ser. Lecture Notes in Computer Science, vol. 9626. Springer International Publishing, 2016, pp. 716–722.

[29] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2017, pp. 1601–1611.

[30] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "Fever: a large-scale dataset for fact extraction and verification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018, pp. 809–819.

[31] M. Maia, S. Handschuh, A. Freitas, B. Davis, R. McDermott, M. Zarrouk *et al.*, "Financial opinion mining and question answering," 2018. [Online]. Available: https://sites.googlecom/view/fiqa/home

[32] T. Diggelmann, J. Boyd-Graber, J. Bulian, M. Ciaramita, and M. Leippold, "Climate-fever: A dataset for verification of real-world climate claims," 2020.

[33] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, and H. Hajishirzi, "Fact or fiction: Verifying scientific claims," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 7534–7550.

[34] E. Voorhees, T. Alam, S. Bedrick, D. Demner-Fushman, W. R. Hersh, E. Kanoulas, K. Lo, J. Mott, L. Scibek, and M. Thornblade, "Trec-covid: Constructing a pandemic information retrieval test collection," *ACM SIGIR Forum*, vol. 54, no. 1, pp. 1–12, 2021.

[35] Q. Works, "Quiz works." [Online]. Available: https://quiz-works.com/about

[36] OpenAI, "Gpt-4o system card," 2024.

[37] DeepSeek-AI, "Deepseek-v3 technical report," 2024.

[38] THUDM, "Chatglm: A family of large language models from glm-130b to glm-4 all tools," 2024.

[39] G. DeepMind, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," 2024.

[40] Meta, "Llama 3.3-70b-instruct," 2024. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

[41] Anthropic, "Introducing the next generation of claude," 2024. [Online]. Available: https://www.anthropic.com/news/claude-3-family

[42] Y. Qiu, H. Li, Y. Qu, Y. Chen, Q. She, J. Liu, H. Wu, and H. Wang, "DuReader-retrieval: A large-scale Chinese benchmark for passage retrieval from web search engine," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 5326–5338.

[43] X. Xie, Q. Dong, B. Wang, F. Lv, T. Yao, W. Gan, Z. Wu, X. Li, H. Li, Y. Liu, and J. Ma, "T2ranking: A large-scale chinese benchmark for passage ranking," 2023.

[44] J. Chen, N. Wang, C. Li, B. Wang, S. Xiao, H. Xiao, H. Liao, D. Lian, and Z. Liu, "AIR-bench: Automated heterogeneous information retrieval benchmark," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 19 991–20 022. [Online]. Available: https://aclanthology.org/2025.acl-long.982/

[45] T. s Koˇ ciský, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, and E. Grefenstette, "The NarrativeQA reading comprehension challenge," *Transactions of the Association for Computational Linguistics*, vol. TBD, p. TBD, 2018. [Online]. Available: https://TBD

[46] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 2790–2799.

[47] R. Zhang, J. Han, C. Liu, P. Gao, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, and Y. Qiao, "Llama-adapter: Efficient fine-tuning of language models with zero-init attention," *arXiv preprint arXiv:2303.16199*, 2023.

[48] S. MacAvaney, F. M. Nardini, R. Perego, N. Tonellotto, N. Goharian, and O. Frieder, "Efficient document re-ranking for transformers by precomputing term representations," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 49–58.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction (L45-58) describe the proposed EAReranker, its benefits (efficiency, no reliance on plaintext) and the key contributions (embedding-oriented architecture, adequacy methodology).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated discussion of limitations and challenges in the Conclusion section (Section 7, L310-317), outlining areas for future research such as refining empirical heuristics in annotation and investigating LLM annotation biases.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not center on formal theoretical results or new theorems but rather on the design and empirical evaluation of a neural architecture and annotation methodology.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 6.1 and Appendix B provide detailed information on experimental settings, datasets used, model architecture, hyperparameters, optimizer details, training epochs, and metrics, sufficient for reproducing the described experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: This research utilizes established public datasets in conjunction with open-source embedding models. The complete implementation code has been made available in the supplementary materials.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 6.1 and Appendix B.1 provide details on data splits, hyperparameters, optimizer, training epochs, and hardware/software configuration.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report standard deviations over 5 runs with different random seeds for the main adequacy assessment metrics in Appendix B.6 (Table 10).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B.1 specifies the technical hardware (GPU, CPU, memory, OS) used for experiments (lines 521-523), and Table 6 details VRAM usage and inference time for different models, including EAReranker (lines 282-289).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is conducted on public datasets, adheres to responsible data handling (no private text is exposed). No ethical violations are apparent.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses both the benefit of improved efficiency and privacy in RAG systems as well as the potential risks (related to mislabeling or bias in adequacy assessment, incorrect filtration) in several sections (introduction, conclusion, and limitations).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not state that high-risk assets such as powerful generative models or sensitive datasets.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper provides references to all used datasets and models (see References [15-41]), credits their sources, and discusses licensing and terms of use in each case.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces a new, annotated adequacy dataset. Collection and annotation procedures, binning schemes, score distributions, and annotation protocols are given in Sections 4 and B.2, with codes in supplemental material.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The application of multiple LLMs for semantic bin annotation is front-and-center in the methodology (see Section 4.2, Appendix A.1–A.3). Prompt engineering, model aggregation, and LLM selection are described in detail.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.