AED-RAG: Continuous Multi-Granular Context Fusion for Retrieval-Augmented Generation via Adaptive Ensemble Decoding

Junzhe Zhou, Fulin Lin, Tairan Cheng, Shaowen Chen, Hongwei Wang*
Zhejiang University
{junzhe1.25, fulin1.24, hongweiwang}@intl.zju.edu.cn

Abstract

Retrieval-Augmented Generation (RAG) enhances Large Language Models (LLMs) yet suffers from a mismatch between coarse retrieval granularity and fine-grained generation needs. Specifically, coarse-grained passages inherently conflate valid context with intra-passage noise that semantic retrieval often fails to filter. Existing alignment strategies, typically relying on discrete reranking, struggle to address this granularity mismatch or effectively balance external evidence with internal knowledge. To bridge this gap, we propose **AED-RAG**, a framework that synergizes discrete retrieval with continuous **A**daptive **E**nsemble **D**ecoding. Specifically, we fine-tune a utility predictor using contrastive perplexity to discern the information density differences between unstructured narrative passages and structured knowledge triplets. During inference, this predictor projects passages, triplets, and the model's parametric memory into a unified probability space, enabling a soft, token-level fusion that dynamically optimizes information gain. Extensive experiments on four open-domain QA benchmarks demonstrate that AED-RAG significantly outperforms competitive baselines, underscoring the effectiveness of integrating multi-granular contexts.

1 Introduction

Retrieval-Augmented Generation (RAG) has established itself as the de facto standard for mitigating hallucinations and enabling large language models (LLMs) to access up-to-date, domain-specific knowledge(Gao et al., 2023; Wang et al., 2025c). While early iterations of RAG predominantly relied on semantic similarity to fetch relevant contexts, the community has recently recognized a critical objective misalignment: semantic relevance does not equate to generation utility(Zhang et al., 2025a; Dai et al., 2025). Recent studies reveal that retrieved contexts, though semantically aligned with the query, often introduce distracting information that degrades reasoning performance or fail to resolve the model's specific knowledge deficits(Jiang et al., 2025). Consequently, the research frontier has shifted from merely optimizing retrieval recall to maximizing the marginal utility of information for the generator. It necessitates architectures that prioritize the actual contribution of context to the target response rather than its similarity to the input query(Zhou and Chen, 2025).

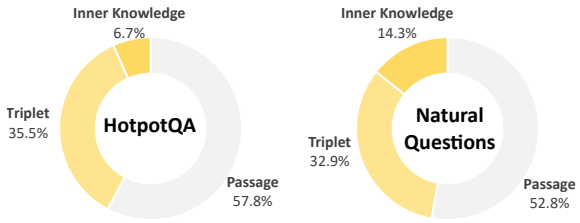

Figure 1: Distribution of Optimal Information Sources. Share of queries where passages, triplets, or parametric memory yield highest generation utility (Contrastive Perplexity) on HotpotQA and Natural Questions.

A critical barrier to high-utility generation is the inherent tension between minimizing noise for factual density and preserving narrative context for linguistic coherence(Zhong et al., 2025). Standard approaches that retrieve entire passages preserve the narrative flow and surrounding context essential for coherent generation(Gupta et al., 2024). However, they inevitably introduce "intra-passage noise"—irrelevant details that dilute the LLM's attention and induce the "lost-in-the-middle" phenomenon(Edge et al., 2025). Conversely, fine-grained retrieval units, such as atomic propositions or knowledge triplets, offer high factual density and structural clarity(Chen et al., 2024). Nonetheless, these structured representations are often too dry or fragmented, lacking the contextual nuance required

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 22885–22899

for the model to construct comprehensive and linguistically fluid responses(Wang and Han, 2025). Current methods typically force a false dichotomy between these granularities, failing to leverage their complementary strengths(Kalra et al., 2025).

Compounding these granular trade-offs, existing frameworks suffer from a discrete information bottleneck inherent in the standard "Retrieve-Rerank-Generate" pipeline(He et al., 2025). By strictly relying on top-k truncation to filter context prior to generation, these systems enforce an irreversible decision boundary that effectively severs the continuous confidence signals provided by the reranker(Gao et al., 2025). This hard-coupling mechanism forces the generator to treat all retrieved contexts equally once selected, limiting its ability to dynamically balance external evidence against the LLM's internal parametric memory—especially when conflicts arise(Choi et al., 2025). While routing-based methods like Mixture-of-Experts (MoE) offer a workaround, they typically incur prohibitive computational overhead(Liu and Yang, 2025; Zhang et al., 2025b). Consequently, there is a pressing need for a seamless, probability-aware integration strategy that operates within the efficiency constraints of standard pipelines.

We argue that the optimal RAG system should not simply select a single passage type, but rather adaptively fuse insights from diverse information granularities and memory sources in a continuous probabilistic space. Our empirical analysis underscores this need for heterogeneity. As illustrated in Figure 1, there is no single "silver bullet" for information sourcing: while context-rich passages yield the highest generation utility for approximately 55% of queries, structured knowledge triplets and the model's own internal memory prove to be superior sources in over 40% of cases. This significant variance indicates that the optimal information source is highly query-dependent. Therefore, instead of relying on static, discrete retrieval, we propose to map passages, triplets, and parametric memory into a unified continuous space, allowing the token probability distribution to dynamically determine the contribution of each source during inference.

To materialize this, we propose **AED-RAG**, a unified framework bridging discrete retrieval and continuous generation. We introduce **Triplet-Enhanced Preference Alignment**, which constructs a mixed candidate view of narrative passages and atomic triplets. By training a utility predictor via Contrastive Perplexity (CPPL) (Jiang et al., 2025) distillation, the model learns to distinguish contextually rich from factually precise units based on marginal utility. Furthermore, to overcome discrete bottlenecks, we implement **Adaptive Ensemble Decoding (AED)**. This mechanism projects the utility scores of passages, triplets, and internal memory into a continuous weight space to perform token-level fusion, enabling the model to softly balance external evidence against internal knowledge without architectural modifications.

Our contributions are summarized as follows:

- We propose **Triplet-Enhanced Preference Alignment**, a novel mechanism that leverages structured triplets to distill precise information representations, enabling the retriever to perceive the "generation utility" of information beyond coarse semantic relevance.
- We introduce **Adaptive Ensemble Decoding**, a training-free inference strategy that transforms conflict resolution from a discrete routing problem into a continuous probabilistic fusion task, effectively integrating multi-granular contexts and parametric memory at the token level.
- Extensive experiments on four open-domain QA benchmarks demonstrate that AED-RAG achieves state-of-the-art performance, significantly outperforming strong utility-aligned baselines while exhibiting robust cross-model transferability.

2 Related Works

2.1 Granularity in Retrieval Units

Recent surveys highlight a shift beyond raw passages to mitigate "intra-passage noise" (Sharma, 2025; Zhang et al., 2025c). One direction utilizes atomic units: Dense X Retrieval (Chen et al., 2024) decomposes documents into "propositions," while PropRAG (Wang and Han, 2025) extends this via beam search for multi-hop reasoning. Conversely, structural methods like GraphRAG (Edge et al., 2025) and KG2RAG (Zhu et al., 2025) leverage Knowledge Graphs to capture relational dependencies, with HippoRAG2 (Gutiérrez et al., 2025) further integrating hippocampal memory for continual learning. Intermediate strategies like Mix-of-Granularity (Zhong et al., 2025) dynamically optimize retrieval granularity. However, these methods often sacrifice narrative context or incur high

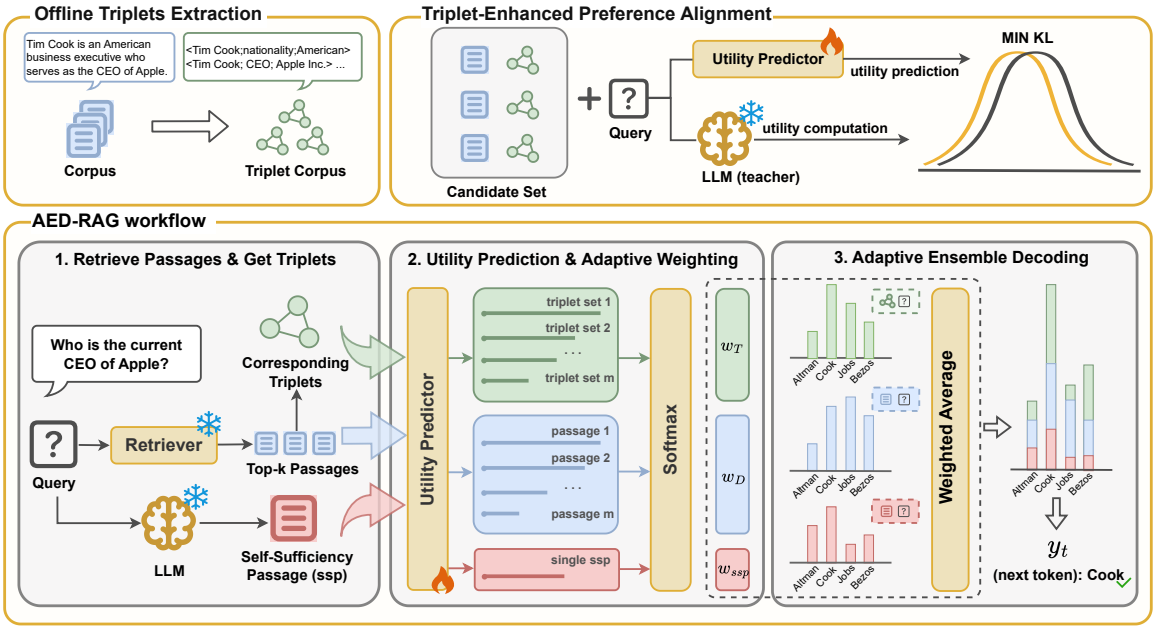

Figure 2: Overview of AED-RAG framework. Knowledge triplets are extracted from corpus passages in an offline manner and jointly employed to fine-tune a utility predictor for preference alignment. During inference, the predictor calculates adaptive weights for three distinct context representations to guide the token-level fusion within the ensemble decoding process.

construction costs. **AED-RAG** addresses this by extracting local triplets alongside passages, forming a "mixed candidate set" that balances narrative flow with structural precision without global graph overhead.

2.2 Generator-Retriever Alignment

A fundamental bottleneck in RAG is the objective mismatch between similarity-based retrieval and the generator's need for high-utility context to resolve knowledge deficits (Leung et al., 2025). GainRAG (Jiang et al., 2025) pioneers this by quantifying marginal utility via "Contrastive Perplexity" to filter candidates based on expected information gain. Similarly, Uplift-RAG (Qu et al., 2025) assesses if retrieved content offers an actual "uplift" to internal knowledge. Optimization-centric methods like RPO (Yan et al., 2025) and RankRAG (Yu et al., 2024) align ranking objectives directly with generation quality, while FaithfulRAG (Zhang et al., 2025d) employs "Self-Think" prompting to resolve fact-level conflicts. Nonetheless, these approaches typically rely on pre-generation "hard decisions" (reranking) or latency-inducing prompting. Our **AED-RAG** overcomes this by shifting alignment to inference-time decoding, employing continuous, adaptive weights to perform a "soft fusion" of information at the token level, thus avoiding premature context discarding.

2.3 Knowledge Conflict and Fusion Strategies

To navigate the volatility of external knowledge, RAG research has evolved from simple concatenation to sophisticated conflict-resolution frameworks (Li et al., 2025). Leading approaches like Astute RAG (Wang et al., 2025a) and MADAM-RAG (Wang et al., 2025b) utilize iterative source-aware consolidation and multi-agent debate, respectively, to dynamically arbitrate between conflicting parametric and retrieved information. However, CARE (Choi et al., 2025) introduces conflict-aware soft prompting to modulate context influence during inference. Despite these architectural strides, relying on model confidence remains precarious. Notably, Soudani et al. (2025) axiomatically demonstrate that traditional uncertainty metrics fail to distinguish inherent aleatoric uncertainty from retrieval-induced conflicts. Addressing this, our **AED-RAG** circumvents these pitfalls via a training-free adaptive ensemble decoding that projects multi-source evidence into a unified logit space to mathematically balance utility.

3 Methodology

3.1 Preliminary

We first formalize the RAG pipeline incorporating a reranking stage. Given an input query q and a corpus $\mathcal{C} = \{d_1, d_2, \ldots, d_M\}$, our goal is to generate a response y. The process is formulated as follows:

$$\begin{aligned} \mathcal{D}_{ret} &= \underset{d \in \mathcal{C}}{\text{Top-}k}\, f_{ret}(q, d), \\ \mathcal{D}_{ctx} &= \underset{d \in \mathcal{D}_{ret}}{\text{Top-}n}\, s_{rank}(q, d), \\ P(y|q, \mathcal{D}_{ctx}) &= \prod_{t=1}^{T} P(y_t|y_{<t}, q, \mathcal{D}_{ctx}), \end{aligned} \tag{1}$$

where $f_{\text{ret}}(\cdot)$ denotes the initial retrieval function (e.g., BM25 or dense retrieval) that retrieves k passages $\mathcal{D}_{ret}$. Subsequently, $s_{rank}(\cdot)$ serves as the reranking function to select the top-n most relevant passages $\mathcal{D}_{ctx}$ from $\mathcal{D}_{ret}$, with $n < k$. Finally, the generation probability is modeled autoregressively, where y_t represents the token predicted at step t given the history $y_{<t}$.

3.2 Method Overview

Figure 2 illustrates the AED-RAG framework, which comprises two key components designed to refine granularity and balance knowledge sources. First, we introduce a **preference-aligned utility predictor** to evaluate the generation quality of heterogeneous contexts. Facilitated by a data construction process that decomposes passages into atomic knowledge triplets, this module is supervised via Contrastive Perplexity to distinguish fine-grained information density from coarse semantic relevance. Second, leveraging these utility signals, we propose a **training-free adaptive ensemble decoding** strategy. This mechanism dynamically assigns query-specific weights to balance external retrieval against the LLM's internal parametric memory via token-level fusion.

3.3 Triplet-Enhanced Preference Alignment

Contrastive Perplexity We employ Contrastive Perplexity (CPPL) (Jiang et al., 2025) to quantify the marginal utility of the retrieved context c in generating the ground truth answer a. This metric effectively decouples the contribution of external retrieval from the LLM's internal parametric knowledge. Formally, CPPL is defined as:

$$\begin{aligned} &\tilde{p}(a_j|q, c, a_{<j}) = \\ &\text{Softmax}\big((1+\alpha) \cdot \text{logit}(a_j|q, c, a_{<j}) \\ &-\alpha \cdot \text{logit}(a_j|q, a_{<j})\big), \end{aligned} \tag{2}$$

$$\text{CPPL}(c, a|q) = \exp\left(-\frac{1}{N}\sum_{j=1}^{N} \log \tilde{p}(a_j|q, c, a_{<j})\right), \tag{3}$$

where N denotes the answer length and $a_{<j}$ represents the history tokens. By penalizing logits derived solely from the query q, the adjusted distribution $\tilde{p}$ highlights the information gain specifically attributable to the context c. This allows the model to differentiate between contexts that are merely semantically relevant and those that provide necessary factual support.

Data Construction In traditional RAG, retrieving raw passages often introduces background noise, potentially hampering the inference of LLMs. While prior works have explored ensemble decoding to utilize multiple contexts (Shi et al., 2024b), they often struggle to accurately fuse disparate information sources due to the lack of fine-grained utility calibration. To address this, we construct a hybrid candidate set that combines the original passages with structured knowledge triplets, enabling the utility predictor to assess information granularity more effectively.

Specifically, we employ a triplet extractor $\Psi(\cdot)$ (Fang et al., 2024) to decompose unstructured text into atomic facts. For efficiency, we **pre-compute** the extraction process for the entire corpus offline. Consequently, for a given query q and a set of retrieved passages $\mathcal{D}_{ret} = \{d_1, \ldots, d_k\}$, we can directly map each passage d_i to its corresponding structured set:

$$\mathcal{T}_i = \Psi(d_i) = \{(h_1, r_1, t_1), \ldots, (h_s, r_s, t_s)\}, \tag{4}$$

where (h, r, t) denotes the head entity, relation, and tail entity, and s is the number of triplets in passage d_i. This distillation process isolates core facts with relatively low noise, though it may occasionally omit context. The mapped $\mathcal{T}_{ret}$ is $\{\mathcal{T}_1, \ldots, \mathcal{T}_k\}$.

To mitigate potential blind spots in external retrieval and leverage parametric memory, we follow Jiang et al. (2025) to incorporate a self-generated passage $d_{gen} = \mathcal{G}(\rho(q))$, synthesized by the LLM $\mathcal{G}$ via instruction $\rho(\cdot)$. The final mixed candidate

set $\mathcal{U}_{mix}$ is formally defined as:

$$\mathcal{U}_{mix} = \mathcal{D}_{ret} \cup \mathcal{T}_{ret} \cup \{d_{gen}\}. \tag{5}$$

Utility Predictor To identify the optimal context within the heterogeneous set $\mathcal{U}_{mix}$, we employ a lightweight utility predictor based on a pre-trained Cross-Encoder (Xiao et al., 2024). This module quantifies the semantic relevance between the query q and a candidate unit u via a scalar score:

$$s_u = S_\phi(q, u), \tag{6}$$

where S_ϕ is the scoring network parameterized by ϕ. We distill knowledge from the LLM reader into the predictor using CPPL (Eq. 3) as a silver-standard signal. Specifically, for a training pair (q, a), we first compute the utility score $\mathcal{C}_j = \text{CPPL}(u_j, a|q)$ to construct a teacher distribution $\mathcal{P}_{teacher}$ based on the inverse log-scores:

$$\mathcal{P}_{teacher}(u_j) = \frac{\exp\left(-\log(\mathcal{C}_j + 1)\right)}{\sum_{u' \in \mathcal{U}_{mix}} \exp\left(-\log(\mathcal{C}_{u'} + 1)\right)}. \tag{7}$$

Simultaneously, the predictor's output is normalized into a student distribution $\mathcal{P}_{student}$:

$$\mathcal{P}_{student}(u_j) = \frac{\exp(S_\phi(q, u_j))}{\sum_{u' \in \mathcal{U}_{mix}} \exp(S_\phi(q, u'))}. \tag{8}$$

Following Lin et al. (2023); Shi et al. (2024b), we optimize ϕ by minimizing the KL divergence, thereby aligning the predictor's ranking mechanism with the generator's inherent preferences:

$$\mathcal{L} = \sum_{u_j \in \mathcal{U}_{mix}} \mathcal{P}_{teacher}(u_j) \cdot \log\left(\frac{\mathcal{P}_{teacher}(u_j)}{\mathcal{P}_{student}(u_j)}\right). \tag{9}$$

This objective enables the predictor to effectively arbitrate among raw passages, structured triplets, and internal parametric memory. The pseudocode details in Algorithm 1.

3.4 Adaptive Ensemble Decoding

The conventional information fusion strategy (Shi et al., 2024b) fails to aggregate information of the retrieval set, limiting the potential gains from hybrid contexts. To address this, we introduce an Adaptive Ensemble Decoding strategy. This approach moves beyond simple selection, aiming to dynamically weigh the collective contribution of passages, triplets, and internal parametric memory based on their utility.

Sequential Utilities Calculation Specifically, given a query q, retrieved passages $\mathcal{D}_{ret}$, and corresponding pre-extracted triplets $\mathcal{T}_{ret}$, we first evaluate all units using the fine-tuned predictor $S_{\tilde{\phi}}$. We then isolate the top-m scores to construct ordered utility sequences:

$$\begin{aligned} \mathcal{S}^k &= \text{Top}_m\left(\{S_{\tilde{\phi}}(q, x) \mid x \in \zeta\}\right) \\ &= [s^\zeta_{(1)}, \ldots, s^\zeta_{(m)}], \quad \zeta \in \{\mathcal{D}_{ret}, \mathcal{T}_{ret}\} \end{aligned} \tag{10}$$

where $\text{Top}_m(\cdot)$ sorts and retains the m highest values, $s^\cdot_{(j)}$ denotes the utility score of the candidate at rank j. Consequently, we derive the passage set $\mathcal{D}$ and the triplet set $\mathcal{T}$ including the top-m candidates respectively. To quantify the holistic quality of each information source, we define the aggregate utilities $U_\mathcal{D}$ and $U_\mathcal{T}$ as:

$$U_\zeta = \sum_{j=1}^{m} \beta^{j-1} s^\zeta_{(j)}, \quad \zeta \in \{\mathcal{D}, \mathcal{T}\} \tag{11}$$

where $\beta > 0$ is a "visibility" hyperparameter controlling the weight distribution. A smaller β concentrates weight on the top-1 candidate, while a larger β emphasizes the collective utility of the sequence. Notably, setting $\beta > 1$ amplifies the impact of lower-ranked context on the overall score.

Self-Sufficiency Passage Furthermore, to simultaneously consider the model's internal parametric memory, we introduce a self-sufficiency passage d_{ss}. While it shares the same prompt template as d_{gen}, d_{ss} is required to participate in the LLM generation process, serving as a representation of the model's internal parametric memory to counterbalance external knowledge. Given that d_{ss} functions as a standalone unit, we define its overall utility U_{ss} directly as its predicted score.

$$U_{ss} = [S_{\tilde{\phi}}(q, d_{ss})]. \tag{12}$$

Adaptive Ensemble Decoding To quantify the contribution of each knowledge source, we normalize the utility scores $U_\mathcal{D}, U_\mathcal{T}$, and U_{ss} into an adaptive weight vector $\mathbf{w}$ via the Softmax function:

$$\mathbf{w} = [w_\mathcal{D}, w_\mathcal{T}, w_{ss}] = \text{Softmax}\left([U_\mathcal{D}, U_\mathcal{T}, U_{ss}]\right). \tag{13}$$

These weights reflect the relative importance of context-rich information, low-noise information, and the model's internal parametric knowledge, respectively. During generation, we compute the output probability distributions conditioned on each

Method	Settings	HotpotQA			NQ			TriviaQA			WebQ		
		EM	F1	Avg	EM	F1	Avg	EM	F1	Avg	EM	F1	Avg
Naive	-	20.96	24.72	22.84	20.77	24.60	22.69	48.00	49.17	48.59	35.16	32.82	33.99
Standard RAG	$k=1$	30.94	34.98	32.96	26.79	28.98	27.89	52.89	52.40	52.65	35.99	32.51	34.25
	$k=5$	37.07	39.06	38.07	31.45	31.01	31.23	57.98	55.91	56.95	**40.82**	33.60	37.21
RECOMP	$k=5$	33.45	39.23	36.34	28.16	29.58	28.87	49.70	51.14	50.42	31.30	32.35	31.83
REPLUG	$k=5$	25.63	28.17	26.90	28.26	24.96	26.61	54.61	52.46	53.54	35.04	31.87	33.46
BGE-reranker	$n=1$	35.98	40.85	38.42	30.54	33.19	31.87	57.19	56.95	57.07	36.98	34.39	35.69
	$n=5$	39.15	42.06	40.61	32.97	33.25	33.11	59.58	57.48	58.53	39.70	34.22	36.96
GainRAG	$n=1$	37.69	42.65	40.17	30.82	33.81	32.32	58.43	58.90	58.67	38.26	35.28	36.77
	$n=5$	36.38	40.69	38.54	31.52	32.60	32.06	59.73	58.56	59.15	40.23	34.19	37.21
AED-RAG (Ours)	$m=1$	38.01	43.39	40.70	31.46	34.77	33.12	60.13	**60.20**	**60.17**	39.39	**36.91**	**38.15**
	$m=2$	**40.04**	**45.05**	**42.55**	**33.78**	**36.35**	**35.07**	**60.14**	60.06	60.10	39.88	36.27	38.08

Table 1: Main results (EM, F1, and Avg) on four datasets. k denotes the number of initially retrieved passages. n represents the number of selected passages for reranking baselines. For AED-RAG, m indicates selecting the top-m passages and top-m triplets. We also report the minimal setting ($n=1$ or $m=1$), where the initial retrieval is fixed at $k=5$. The best results are highlighted in **bold**, and the second-best results are underlined.

context source independently. The subsequent token y_t is selected by maximizing the weighted sum of log-probabilities:

$$y_t = \arg\max_{v \in V} \sum_{\zeta \in \mathcal{Z}} w_\zeta \log P(v|q, \zeta, y_{<t}), \quad (14)$$

where $\mathcal{Z} = \{\mathcal{D}, \mathcal{T}, d_{ss}\}$, V denotes the vocabulary and $y_{<t}$ represents the generation history. This strategy projects external evidence (passages, triplets) and internal memory into a **unified continuous space** for dynamic token-level fusion. Notably, our approach generalizes the standard discrete retrieve-rerank-generate paradigm (reducing to it when $w_\mathcal{T} = w_{ss} = 0$), thereby facilitating optimal integration of latent information without the bottlenecks of hard selection. The pseudocode is detailed in Algorithm 2.

4 Experiment

4.1 Experimental Setup

Training Details For each sample in the HotpotQA (Yang et al., 2018) training set, we retrieve the top-20 passages using Contriever (Lei et al., 2023) and map their corresponding pre-extracted triplets. This results in a candidate set of 41 items per sample: 20 external passages, 20 triplet sets, and 1 reader-generated passage. We employ Qwen3-8B (Yang et al., 2025) to calculate the contrastive perplexity for these candidates given the query q and ground truth a, setting the decoding parameter α to 0.5 following CAD (Shi et al., 2024a). The utility predictor is initialized with BGE-reranker-base (Xiao et al., 2024) and trained on two RTX Pro 6000 GPUs. Further details are provided in Appendix C.

Datasets We evaluate AED-RAG on four widely-used open-domain question answering datasets: HotpotQA (Yang et al., 2018), Natural Questions (NQ) (Kwiatkowski et al., 2019), WebQuestions (Berant et al., 2013), and TriviaQA (Joshi et al., 2017). Detailed in Appendix A.

Evaluation Metrics We evaluate performance using Exact Match (EM) and F1 score. Following Asai et al. (2023) and Jiang et al. (2025), we adopt a relaxed EM criterion, where a prediction is deemed correct if it contains the ground truth. F1 measures the token-level overlap between the prediction and reference. We observe a trade-off where longer generations tend to inflate EM (via higher answer coverage) but penalize F1 (due to redundancy). To counterbalance this length bias, we additionally report the average of EM and F1 to provide a more holistic assessment.

Baselines We compare our proposed AED-RAG against five baselines: 1) Naive Generation, 2) Standard RAG, 3) RECOMP (Xu et al., 2023), 4) REPLUG (Shi et al., 2024b), 5) BGE-Reranker (Xiao et al., 2024), 6) GainRAG (Jiang et al., 2025). Detailed descriptions of these methods are provided in Appendix B.

Implementation Details To ensure a fair comparison, we employ Contriever as the unified re-

Method	HotpotQA			NQ		
	EM	F1	Avg	EM	F1	Avg
AED-RAG	**40.04**	**45.05**	**42.55**	**33.78**	**36.35**	**35.07**
w/o AED (triplets)	39.93	44.57	42.25	32.99	35.25	34.12
w/o AED (ssp)	39.66	44.30	41.98	32.57	35.18	33.88
w/o AED (full)	38.11	43.16	40.64	31.92	34.26	33.09
w/o AED (full) & triplets in reranking	35.87	40.31	38.09	31.10	32.61	31.86

Table 2: Ablation study of different components on HotpotQA and Natural Questions (NQ) datasets. We report the results formatted to two decimal places. The best results are **bolded**, and the second-best results are underlined.

triever and BGE-reranker-base (Xiao et al., 2024) for reranking tasks. Qwen3-8B serves as the backbone LLM. By default, the model receives $k = 5$ contexts for generation. Specifically, standard RAG directly retrieves the top-5 passages. For reranking baselines, we first retrieve the top-10 passages and select the top-5 after reranking. Similarly, our AED-RAG initially retrieves the top-10 passages, then selects the top-m passages and top-m triplets. In our experiments, we set $m = 2$ and the visibility parameter $\beta = 0.5$. Additionally, we discuss the computational costs including input token and runtime overheads in Appendix F.

4.2 Main Results

The experimental results are presented in Table 1, yielding the following observations:

1) AED-RAG achieves SOTA performance across all four datasets. We attribute this success to two key factors: the adaptive weights effectively balance the interplay between external knowledge and internal memory, and the ensemble decoding strategy provides fine-grained control at the token level. Notably, despite being trained solely on HotpotQA, AED-RAG demonstrates robust generalization capabilities across diverse benchmarks.

2) Our method consistently outperforms GainRAG, which similarly utilizes CPPL for preference alignment. This comparison underscores the superiority of adaptive ensemble decoding over coarse-grained reranking strategies in precisely managing contextual information.

3) On WebQuestions, reranking-based methods underperform standard RAG when utilizing five passages. As WebQuestions is arguably the simplest dataset evaluated, this indicates that multiple passages may introduce noise that degrades performance on simple questions. Nevertheless, AED-RAG mitigates this limitation, securing the best overall performance.

4.3 Ablation Study

We conduct ablation studies on HotpotQA and NQ datasets to assess the contribution of each component to the generation performance. The results are clarified in Table 2.

1) **w/o AED (triplets):** This setting excludes triplet data from the Adaptive Ensemble Decoding stage, forcing the model to generate tokens based on external passages and the self-sufficiency paragraph. The observed performance decline across all datasets confirms that structured triplets provide distinct information representations that effectively mitigate background noise inherent in unstructured passages.

2) **w/o AED (ssp):** Here, we omit the self-sufficiency paragraph during decoding. The resulting drop in metrics underscores the critical role of the model's internal parametric memory in complementing retrieved contexts for accurate question answering.

3) **w/o AED (full):** Removing the AED component entirely reduces the framework to a standard reranking approach. While significant performance drops occur, our method still outperforms GainRAG (Jiang et al., 2025), a SOTA preference-aligned passage reranker. This advantage stems from our Utility Predictor, which dynamically selects the higher-utility option between passages and their corresponding triplet sets, rather than relying on passages alone.

4) **w/o AED (full) & triplets in reranking:** Building on the removal of AED, this setting further excludes triplet data during the reranking phase. The subsequent performance degradation demonstrates that structured knowledge remains beneficial even within a standard reranking paradigm.

4.4 Impact of the Visibility Coefficient

We introduce a non-negative visibility coefficient, β, to modulate the contribution of long-tail can-

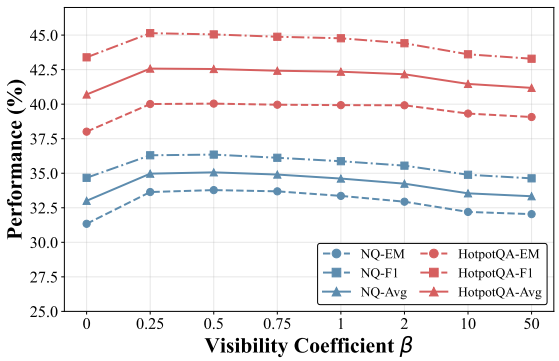

Figure 3: Impact of the visibility coefficient β on generation performance on HotpotQA and NQ datasets.

didates to the sequential utility. For $\beta < 1$, the coefficient acts as a decay factor, attenuating the influence of lower-ranked passages or triplets. Notably, at $\beta = 0$, the global utility collapses to that of the top-ranked candidate, reducing the process to greedy selection ($m = 1$). Conversely, $\beta \to 1$ yields a more uniform contribution distribution, while $\beta > 1$ amplifies the influence of sequence quantity and minimum utility values.

Figure 3 illustrates the generation performance across varying β. Results indicate that AED-RAG is robust when $\beta < 1$ but exhibits a noticeable decline at extremes (e.g., $\beta = 0$ or $\beta = 50$). This suggests that collective utility estimation should prioritize top-ranked candidates. Furthermore, since excessive β values suppress the weight of internal parametric memory, the observed performance drop underscores the necessity of the model's intrinsic knowledge for accurate generation.

4.5 Alignment is a Prerequisite

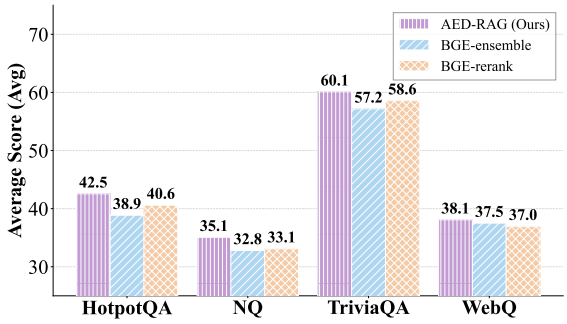

Figure 4: Comparison validating the necessity of alignment. BGE-ensemble denotes using raw scores from the pre-trained BGE-reranker for ensemble decoding.

To validate the necessity of preference alignment, we evaluated AED-RAG by replacing the utility predictor with a pre-trained BGE-reranker, directly using its raw output scores as weights for ensemble decoding. As shown in Figure 4, this strategy failed to yield performance gains. Furthermore, compared to standard reranking baselines, we observed **performance degradation** across all datasets, with the exception of WebQuestions. These results indicate that while pre-trained rerankers excel at discriminative tasks, their uncalibrated scores are ill-suited for ensemble decoding. Without alignment, these scores fail to effectively balance the interplay between external context and the model's internal parametric knowledge.

Backbone	Method	HotpotQA (Avg)	NQ (Avg)
Qwen3-1.7B	BGE-reranker	31.89	24.87
	AED-RAG	**34.08** (↑2.19)	**26.31** (↑1.44)
Qwen2.5-7B	BGE-reranker	40.86	32.43
	AED-RAG	**42.76** (↑1.90)	**33.61** (↑1.18)
Qwen3-14B	BGE-reranker	42.51	33.46
	AED-RAG	**44.88** (↑2.37)	**35.08** (↑1.62)

Table 3: Evaluation of cross-model transferability. Values in parentheses denote the absolute performance gain over the BGE-Rerank baseline.

4.6 Cross-Model Transferability

We evaluated the transferability of AED-RAG across different models within the same series. Specifically, we deployed the utility predictor—originally fine-tuned using data annotated by **Qwen3-8B**—directly into AED-RAG pipelines instantiated with **Qwen3-1.7B**, **Qwen2.5-7B** and **Qwen3-14B** as reader modules. As detailed in Table 3, empirical results demonstrate that AED-RAG achieves overall performance improvements, even though the utility predictor was not specifically fine-tuned for these target models. This outcome reflects the inherent consistency in preferences among models of the same lineage, suggesting that AED-RAG possesses a certain degree of transferability across same-series models without the necessity for separate optimization.

4.7 Impact of Triplets

In AED-RAG, we utilize knowledge triplets as condensed textual evidence to enhance factual precision. By directly concatenating these triplets, we provide a low-noise information source that complements the contextual richness of narrative passages. This straightforward approach bypasses the overhead of knowledge graph maintenance while effectively bridging the gap between narrative flow

Method	Settings	HotpotQA	NQ
standard RAG	k=5	38.07	31.23
standard RAG	triplet, k=5	36.31	30.93
standard RAG	both, k=5+5	37.54	31.48
BGE-reranker	k=10, n=5	40.61	33.11
BGE-reranker	triplet, k=10, n=5	38.66	32.62
BGE-reranker	both, k=10, n=5	40.08	33.21
AED-RAG	m=2	42.55	35.07

Table 4: Performance (avg. scores) comparison between triplet corpus and original paragraph corpus. "Triplet" refers to using only the triplet corpus as retrieval context. "Both" denotes the simultaneous use of both passages and their corresponding triplets, concatenated within a single context window.

and factual exactness. As shown in Table 2, incorporating these distinct information representations yields significant performance gains. Furthermore, the results in Table 4 indicate that using only the triplet corpus (or obtaining both triplets and paragraphs) in standard RAG or BGE-Reranker yields negligible performance gains, and may even result in deterioration. It demonstrates that the performance gains of AED-RAG are not solely attributable to the informational advantages of the triplets.

5 Conclusion

In this work, we presented AED-RAG, a framework designed to mitigate the granularity mismatch and discrete information bottlenecks in RAG. Instead of relying solely on the hard selection, our approach seeks to harmonize the contextual richness of passages with the factual precision of triplets via *Triplet-Enhanced Preference Alignment*. Furthermore, we explored a continuous conflict resolution strategy through *Adaptive Ensemble Decoding*, which enables a token-level fusion of multi-granular information. Empirical results suggest that AED-RAG offers a robust alternative to discrete baselines, highlighting the potential of softly balancing external retrieval and internal parametric memory for high-utility generation.

Limitations

While AED-RAG effectively controls generation via adaptive weights, we acknowledge a few limitations in our current work. First, our approach focuses on individual passages; future research could investigate the potential synergies of combinatorial contexts to further enhance preference alignment. Second, the retrieval count m is currently fixed; exploring adaptive retrieval quantities might further optimize performance. Finally, as our model was trained on standard benchmarks, extending experiments to large-scale datasets would be valuable to verify the model's scalability and robustness.

Ethics Statement

This work strictly adheres to the ACL Ethics Policy. All datasets and models utilized in the experiments are publicly available. We believe that the presented work does not directly involve potential risks.

Acknowledgment

This work is supported by the National Key Research and Development Program of China (Grant No. 2024YFF0907802) and the National Natural Science Foundation of China (Grant Nos. 62276230 and 62576308).

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024. Dense X retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15159–15177, Miami, Florida, USA.

Eunseong Choi, June Park, Hyeri Lee, and Jongwuk Lee. 2025. Conflict-aware soft prompting for retrieval-augmented generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26969–26983, Suzhou, China.

Lu Dai, Yijie Xu, Jinhui Ye, Hao Liu, and Hui Xiong. 2025. Seper: Measure retrieval utility through the lens of semantic perplexity reduction. In *The Thirteenth International Conference on Learning Representations*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and

Jonathan Larson. 2025. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

Jinyuan Fang, Zaiqiao Meng, and Craig MacDonald. 2024. TRACE the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8472–8494, Miami, Florida, USA.

Guangze Gao, Zixuan Li, Chunfeng Yuan, Jiawei Li, Wu Jianzhuo, Yuehao Zhang, Xiaolong Jin, Bing Li, and Weiming Hu. 2025. D-RAG: Differentiable retrieval-augmented generation for knowledge graph question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35386–35405, Suzhou, China.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. *arXiv preprint arXiv:2410.12837*.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From RAG to memory: Non-parametric continual learning for large language models. In *Forty-second International Conference on Machine Learning*.

Jie He, Richard He Bai, Sinead Williamson, Jeff Z. Pan, Navdeep Jaitly, and Yizhe Zhang. 2025. Clara: Bridging retrieval and generation with continuous latent reasoning. *Preprint*, arXiv:2511.18659.

Yi Jiang, Sendong Zhao, Jianbo Li, Haochun Wang, and Bing Qin. 2025. GainRAG: Preference alignment in retrieval-augmented generation through gain signal synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10746–10757, Vienna, Austria.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. 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)*, pages 1601–1611, Vancouver, Canada.

Rishi Kalra, Zekun Wu, Ayesha Gulley, Airlie Hilliard, Xin Guan, Adriano Koshiyama, and Philip Colin Treleaven. 2025. HyPA-RAG: A hybrid parameter adaptive retrieval-augmented generation system for AI legal and policy applications. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 1036–1054, Albuquerque, New Mexico.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Yibin Lei, Liang Ding, Yu Cao, Changtong Zan, Andrew Yates, and Dacheng Tao. 2023. Unsupervised dense retrieval with relevance-aware contrastive pretraining. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10932–10940, Toronto, Canada.

Kin Kwan Leung, Mouloud Belbahri, Yi Sui, Alex Labach, Xueying Zhang, Stephen Rose, and Jesse C Cresswell. 2025. Classifying and addressing the diversity of errors in retrieval-augmented generation systems. *arXiv preprint arXiv:2510.13975*.

Yangning Li, Weizhi Zhang, Yuyao Yang, Wei-Chieh Huang, Yaozu Wu, Junyu Luo, Yuanchen Bei, Henry Peng Zou, Xiao Luo, Yusheng Zhao, Chunkit Chan, Yankai Chen, Zhongfen Deng, Yinghui Li, Hai-Tao Zheng, Dongyuan Li, Renhe Jiang, Ming Zhang, Yangqiu Song, and Philip S. Yu. 2025. A survey of RAG-reasoning systems in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12120–12145, Suzhou, China.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, and 1 others. 2023. Ra-dit: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Lihui Liu and Carl J Yang. 2025. Mixrag: Mixture-of-experts retrieval-augmented generation for textual graph understanding and question answering. *arXiv preprint arXiv:2509.21391*.

Changle Qu, Sunhao Dai, Hengyi Cai, Yiyang Cheng, Jun Xu, Shuaiqiang Wang, and Dawei Yin. 2025. Uplift-RAG: Uplift-driven knowledge preference alignment for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9632–9644, Suzhou, China.

Chaitanya Sharma. 2025. Retrieval-augmented generation: A comprehensive survey of architectures, enhancements, and robustness frontiers. *arXiv preprint arXiv:2506.00054*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024a.

Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024b. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico City, Mexico.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2025. Why uncertainty estimation methods fall short in RAG: An axiomatic analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16596–16616, Vienna, Austria.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. 2025a. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30553–30571, Vienna, Austria.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025b. Retrieval-augmented generation with conflicting evidence. In *Second Conference on Language Modeling*.

Jingjin Wang and Jiawei Han. 2025. PropRAG: Guiding retrieval with beam search over proposition paths. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6223–6238, Suzhou, China.

Zilong Wang, Zifeng Wang, Long Le, Steven Zheng, Swaroop Mishra, Vincent Perot, Yuwei Zhang, Anush Mattapalli, Ankur Taly, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2025c. Speculative RAG: Enhancing retrieval augmented generation through drafting. In *The Thirteenth International Conference on Learning Representations*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. *arXiv preprint arXiv:2310.04408*.

Shi-Qi Yan, Quan Liu, and Zhen-Hua Ling. 2025. RPO: Retrieval preference optimization for robust retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5228–5240, Vienna, Austria.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *Advances in Neural Information Processing Systems*, 37:121156–121184.

Hengran Zhang, Minghao Tang, Keping Bi, Jiafeng Guo, Shihao Liu, Daiting Shi, Dawei Yin, and Xueqi Cheng. 2025a. Utility-focused LLM annotation for retrieval and retrieval-augmented generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1683–1702, Suzhou, China.

Jiarui Zhang, Xiangyu Liu, Yong Hu, Chaoyue Niu, Fan Wu, and Guihai Chen. 2025b. RAGRouter: Learning to route queries to multiple retrieval-augmented language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Junnan Dong, Yi Chang, and Xiao Huang. 2025c. A survey of graph retrieval-augmented generation for customized large language models. *Preprint*, arXiv:2501.13958.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025d. FaithfulRAG: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21863–21882, Vienna, Austria.

Zijie Zhong, Hanwen Liu, Xiaoya Cui, Xiaofan Zhang, and Zengchang Qin. 2025. Mix-of-granularity: Optimize the chunking granularity for retrieval-augmented generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5756–5774, Abu Dhabi, UAE. Association for Computational Linguistics.

Jiawei Zhou and Lei Chen. 2025. OpenRAG: Optimizing RAG end-to-end via in-context retrieval learning. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. 2025. Knowledge graph-guided retrieval augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8912–8924, Albuquerque, New Mexico.

A Dataset Details

We provide a brief overview of the four datasets used in our experiments. The statistics of these datasets are shown in Table 5

HotpotQA (Yang et al., 2018) is a dataset designed for multi-hop reasoning, requiring systems to retrieve and reason over multiple supporting documents to answer questions. It consists of question-answer pairs accompanied by supporting facts from English Wikipedia.

Natural Questions (NQ) (Kwiatkowski et al., 2019) comprises real user queries issued to the Google search engine. The questions are paired with entire Wikipedia pages, challenging the model to handle open-domain questions that may be ambiguous or require identifying long and short answers.

WebQuestions (Berant et al., 2013) focuses on entity-centric questions primarily derived from the Freebase knowledge base. It is widely used to evaluate the capability of QA systems in retrieving specific factual entities.

TriviaQA (Joshi et al., 2017) includes complex question-answer pairs authored by trivia enthusiasts. The evidence documents are collected from Wikipedia and the Web, requiring the model to handle long context and diverse lexical variations.

We evaluate on the development sets for HotpotQA and NQ, and the test sets for WebQuestions and TriviaQA. This selection relies on the availability of public splits and prioritizes larger sample sizes to ensure statistically robust evaluation.

Dataset	# Train	# Dev	# Test
HotpotQA	90,447	7,405	/
Natural Questions (NQ)	79,168	8,757	3,610
TriviaQA	78,785	8,837	11,313
WebQuestions(WebQ)	3,778	/	2,032

Table 5: Statistics of the experimental datasets. The values indicate the number of question-answer pairs in each split. The symbol "/" denotes that the specific dataset split is unavailable or official labels are not provided.

B Baseline Details

We benchmark our approach against the following common or state-of-the-art methods:

Naive Generation: This method relies solely on the internal parametric knowledge of the LLM to generate responses, serving as a raw performance baseline without external retrieval.

Standard RAG: Following the classic "Retrieve-then-Read" paradigm, it retrieves relevant passages and directly concatenates them with the query to form the input context.

RECOMP (Xu et al., 2023): This approach compresses retrieved documents into concise summaries prior to in-context integration, employing selective augmentation to omit the context entirely if it is deemed unhelpful.

REPLUG (Shi et al., 2024b): A framework that treats the LLM as a black box using ensemble decoding. Rather than concatenating all passages, it inputs each passage-query pair independently and ensembles the output probabilities.

BGE-Reranker (Xiao et al., 2024): This method incorporates an intermediate reranking module between retrieval and generation to refine the quality of retrieved documents.

GainRAG (Jiang et al., 2025): A recent approach that fine-tunes the reranker based on contrastive perplexity, aiming to align the reranker's preference with the LLM.

C Training Details

In the utility annotation phase, we filter out training samples where the maximum utility score among candidates is negligible (specifically, lower than $1+10^{-10}$). This filtering process ensures the model focuses on solvable instances, resulting in a final training set of 19,291 unique queries from the original HotpotQA training split. Although we construct an initial pool of 41 candidates (20 passages, 20 triplet sets, and 1 self-generated passage) for each query, using all candidates during training is computationally redundant. Therefore, for each training step, we select the top-32 candidates with the highest annotated utility scores to form the input context list. The model is trained with a batch

Task	Prompt Template
Generation	*{context}* \n ### Instruction: \n Answer the question below concisely in a few words. \n\n ### Input: \n *{query}*
Self-sufficiency Passage	Please provide background for the question below in 100 words. Do not respond with anything other than background. If you do not know or are unsure, please generate "N/A" directly. Question: *{query}*

Table 6: Full list of prompt instructions employed for zero-shot evaluation and pseudo-passage generation. The placeholders {query} and {context} represent the input question and the retrieval context, respectively.

size of 8. The learning rate is set to 6e-5, with a warmup ratio of 0.1. We employed a fixed random seed during the training process to ensure reproducibility. Consequently, all results reported in this paper are obtained from single runs.

D Prompt Templates

We adopt the prompt settings from Jiang et al. (2025). To ensure a fair comparison, the prompts used for generation are identical across AED-RAG and all baselines, as detailed in Table 6.

Algorithm 1 Triplet-Enhanced Preference Alignment (Training)

Require: LLM Reader $\mathcal{G}$, Utility Predictor S_ϕ, Triplet Extractor Ψ.
Input: Training Set $\mathcal{D}_{train} = \{(q, a)\}$, Corpus $\mathcal{C}$, Retrieval Limit k.
Output: Optimized parameters $\tilde{\phi}$.

Offline Phase:
$\mathcal{C}_\mathcal{T} \leftarrow \{\Psi(d) \mid d \in \mathcal{C}\}$ ▷ extract triplet sets
Online Fine-tuning:
for each batch $(q, a) \in \mathcal{D}_{train}$ **do**
  $\mathcal{D}_{ret} \leftarrow \text{Retrieve}(q, \mathcal{C}, k)$
  $\mathcal{T}_{ret} \leftarrow \{MAP(d, \mathcal{C}_\mathcal{T}) \mid d \in \mathcal{D}_{ret}\}$ ▷ Get k triplet sets, one per passage
  $d_{gen} \leftarrow \mathcal{G}(q)$
  // Candidate Set Construction (Treating each set $\mathcal{T}_i \in \mathcal{T}_{ret}$ as a unit)
  $\mathcal{U}_{mix} \leftarrow \mathcal{D}_{ret} \cup \mathcal{T}_{ret} \cup \{d_{gen}\}$
  // Teacher Distribution Calculation
  for unit $u_j \in \mathcal{U}_{mix}$ **do**
    $\mathcal{C}_j \leftarrow \text{CPPL}(u_j, a, q)$ ▷ Eq. 3
  end for
  Calculate $\mathcal{P}_{teacher}(u_j)$ ▷ Eq. 7
  // Student Distribution Calculation
  for unit $u_j \in \mathcal{U}_{mix}$ **do**
    $s_{u_j} \leftarrow S_\phi(q, \tilde{u}_j)$ ▷ Eq. 6
  end for
  Calculate $\mathcal{P}_{student}(u_j)$ ▷ Eq. 8
  // Optimization
  $\mathcal{L} \leftarrow \sum \mathcal{P}_{teacher} \cdot \log(\frac{\mathcal{P}_{teacher}}{\mathcal{P}_{student}})$ ▷ Distillation loss, Eq. 9
  Update ϕ to minimize $\mathcal{L}$.
end for

E Algorithm for AED-RAG

We provide the detailed pseudocode for the two core components of AED-RAG: the Triplet-Enhanced Preference Alignment (Algorithm 1) and the Adaptive Ensemble Decoding strategy (Algorithm 2). Specifically, **Algorithm 1** outlines the training phase, detailing the construction of heterogeneous candidate sets and the utility alignment objective. **Algorithm 2** describes the inference process, illustrating how adaptive weights are computed to dynamically fuse passages, triplets, and internal memory at the token level.

Algorithm 2 Adaptive Ensemble Decoding (Inference)

Require: Utility Predictor $S_{\tilde{\phi}}$, LLM $\mathcal{G}$.
Input: Query q, Hyperparameters m, β.
Output: Response y.

$\mathcal{D}_{ret} \leftarrow \text{Retrieve}(q, \mathcal{C}, k)$
$\mathcal{T}_{ret} \leftarrow \{MAP(d, \mathcal{C}_\mathcal{T}) \mid d \in \mathcal{D}_{ret}\}$
$d^{ss} \leftarrow \mathcal{G}(q)$ ▷ Generate self-sufficiency passage
// Utility Prediction & Selection
$s^{\mathcal{D}r} \leftarrow \{S_\phi(q, d) \mid d \in \mathcal{D}_{ret}\}$
$s^{\mathcal{T}r} \leftarrow \{S_\phi(q, \mathcal{T}_u) \mid \mathcal{T}_u \in \mathcal{T}_{ret}\}$
$\mathcal{D} \leftarrow \text{Top}_m(\mathcal{D}r)$ ▷ Select top-m passages
$\mathcal{T} \leftarrow \text{Top}_m(\mathcal{T}r)$ ▷ Select top-m triplet sets
$U_{ss} \leftarrow S_\phi(q, d^{ss})$ ▷ Eq. 12
// Sequential Utilities Calculation
$U_\mathcal{D} \leftarrow \sum_{j=1}^{m} \beta^{j-1} \cdot s^{\mathcal{D}}_{(j)}$
$U_\mathcal{T} \leftarrow \sum_{j=1}^{m} \beta^{j-1} \cdot s^{\mathcal{T}}_{(j)}$ ▷ Eq. 11
// Adaptive Weighting
$w \leftarrow \text{Softmax}([U_\mathcal{D}, U_\mathcal{T}, U_{ss}])$ ▷ Eq. 13
// Token-level Fusion Decoding
$y \leftarrow$ empty sequence
for step $t = 1, 2, \ldots$ **do**
  $\mathcal{L}_\mathcal{D} \leftarrow \log P(v|q, \mathcal{D}, y_{<t})$
  $\mathcal{L}_\mathcal{T} \leftarrow \log P(v|q, \mathcal{T}, y_{<t})$
  $\mathcal{L}_{ss} \leftarrow \log P(v|q, d^{ss}, y_{<t})$
  $\text{Score}(v) \leftarrow w_\mathcal{D}\mathcal{L}_\mathcal{D} + w_\mathcal{T}\mathcal{L}_\mathcal{T} + w_{ss}\mathcal{L}_{ss}$ ▷ Eq. 14
  $y_t \leftarrow \arg\max_v \text{Score}(v)$
  $y \leftarrow y \oplus y_t$
  if y_t is EOS **then break**
  end if
end for
return y

F Analysis of Inference Overhead

Since autoregressive generation is memory-bound rather than compute-bound, AED-RAG can batch the three data streams simultaneously, keeping its decoding latency consistent with standard single-stream generation. To demonstrate that the infer-

Method	Settings	HotpotQA		NQ	
		$\text{latency}_{\text{total}}$	$\text{latency}_{\text{inference}}$	$\text{latency}_{\text{total}}$	$\text{latency}_{\text{inference}}$
standard RAG	k=5	0.49	0.23	0.54	0.28
standard RAG	k=10	0.83	0.49	0.69	0.32
BGE-reranker	k=10, n=5	0.48	0.22	0.55	0.29
BGE-reranker (sentence-level)	k=24, n=12	0.80	0.14	0.93	0.15
GainRAG	k=10, n=5	0.52	0.20	0.58	0.22
RECOMP	k=5	0.83	0.06	1.16	0.08
AED-RAG(en1)	k=10, m=1	0.54	0.17	0.66	0.25
AED-RAG(en2)	k=10, m=2	0.57	0.20	0.71	0.30

Table 7: Comparison of latency overhead (in seconds) per query on HotpotQA and NQ across different methods. Specifically, $\text{latency}_{\text{total}}$ refers to the end-to-end latency, and $\text{latency}_{\text{inference}}$ denotes the time overhead for the model to perform the final question answering once all contexts are retrieved.

Method	Settings	HotpotQA	NQ
standard RAG	k=5	777.54	735.42
standard RAG	k=10	1459.41	1367.21
standard RAG	k=20	2795.95	2624.17
BGE-reranker	n=5	752.59	783.44
AED-RAG	m=1	585.12	580.05
AED-RAG	m=2	765.34	782.99

Table 8: Average input token consumption per query on HotpotQA and NQ datasets.

ence overhead of AED-RAG is highly controllable, we empirically compare the time overhead of our method with several baselines within the same codebase. As shown in Table 7, under equivalent settings, the time overhead of AED-RAG (including ensemble decoding) aligns closely with standard RAG and reranking pipelines, introducing no significant increase in end-to-end latency. Notably, simpler sentence-level retrieval or lightweight compression models (RECOMP) actually incur greater retrieval latency than AED-RAG. Furthermore, Table 8 shows that AED-RAG maintains a token overhead commensurate with standard RAG and the BGE-reranker, confirming its efficiency alongside its performance gains.