GraTAG: Production AI Search via Graph-Based Query Decomposition and Triplet-Aligned Generation with Rich Multimodal Representations

Bo Tang*
AIDS and SIAR, University of Science and Technology of China
Hefei, Anhui, China
tangbo@mail.ustc.edu.cn

Junyi Zhu*
ESAT-PSI, KU Leuven
Leuven, Belgium
junyizhu.ai@gmail.com

Ang Li*
Zhejiang University
Hangzhou, Zhejiang, China
leeyon@zju.edu.cn

Yiquan Wu
Zhejiang University
Hangzhou, Zhejiang, China
wuyiquan@zju.edu.cn

Kun Kuang
Zhejiang University
Hangzhou, Zhejiang, China
kunkuang@cs.zju.edu.cn

Beihong Jin
Institute of Software, Chinese Academy of Sciences
Beijing, China
beihong@iscas.ac.cn

Jiahao Wu
The Hong Kong Polytechnic University
Hong Kong, China
notyour_mason@outlook.com

Hao Wang
MemTensor
Shanghai, China
hao.wang@my.cityu.edu.hk

Chenyang Xi
MemTensor
Shanghai, China
xicy@memtensor.cn

Yuchen Feng
MemTensor
Shanghai, China
fengyc@memtensor.cn

Wenqiang Wei
MemTensor
Shanghai, China
weiwq@memtensor.cn

Chunyu Li
MemTensor
Shanghai, China
licy@memtensor.cn

Zehao Lin
MemTensor
Shanghai, China
georgelin@zju.edu.cn

Zhiyu Li
MemTensor
Shanghai, China
lizy@memtensor.cn

Feiyu Xiong
MemTensor
Shanghai, China
xiongfy@memtensor.cn

Beibei Li
Chongqing University
Chongqing, China
libeibeics@cqu.edu.cn

Kaiwen Wei
Chongqing University
Chongqing, China
weikaiwen@cqu.edu.cn

Jingrun Chen†
School of Mathematical Sciences and SIAR, University of Science and Technology of China
Hefei, Anhui, China
jingrunchen@ustc.edu.cn

Abstract

Generative AI search engines offer a key advantage over traditional search engines through their ability to synthesize fragmented information for queries, yet they still require improvements in relevance, comprehensiveness, and presentation. To these ends, we introduce GraTAG, an AI Search framework that addresses these challenges through three core innovations. First, we introduce **graph-based query decomposition (GQD)** to dynamically break down complex or ambiguous queries into structured sub-queries with explicit dependencies, enabling more precise, stepwise retrieval. Second, we propose **triplet-aligned generation (TAG)**, which dynamically constructs relation triplets from retrieved documents to explicitly model entity relationships, factual dependencies, and logical connections, enabling the model to generate more coherent and comprehensive answers. Third, GraTAG innovates in **rich multimodal presentations** that integrates timeline visualization and textual-visual choreography to reduce cognitive load and enhance information verification. Evaluated on recent real-world queries, GraTAG

*Co-first authors.
†Corresponding author.

KDD 2026, Jeju Island, Republic of Korea.

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818426

Figure 1: Online evaluation of the GraTAG system for a complex multi-faceted query. GraTAG features built-in citation, timeline visualization (right column), and a textual-visual choreography mechanism.

outperforms eight existing systems in human expert assessments, excelling in relevance, comprehensiveness, and insightfulness. Our work publicly releases a comprehensive, industry-grade full-stack AI search engine, providing a solid reference for the community and demonstrating the effectiveness of the proposed system through rigorous evaluations and ablation studies.

CCS Concepts

• **Information systems** → **Information retrieval query processing**; • **Computing methodologies** → **Natural language generation**.

Keywords

Generative AI Search, Query Decomposition, Triplet Aligned Generation

ACM Reference Format:

Bo Tang, Junyi Zhu, Ang Li, Yiquan Wu, Kun Kuang, Beihong Jin, Jiahao Wu, Hao Wang, Chenyang Xi, Yuchen Feng, Wenqiang Wei, Chunyu Li, Zehao Lin, Zhiyu Li, Feiyu Xiong, Beibei Li, Kaiwen Wei, and Jingrun Chen. 2026. GraTAG: Production AI Search via Graph-Based Query Decomposition and Triplet-Aligned Generation with Rich Multimodal Representations. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD 2026), August 9–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3818426

1 Introduction

In the era of information, a significant volume of events, knowledge, and resources has been digitized and made accessible through the Internet. As users continually strive to quickly and accurately locate relevant information within this rapidly growing digital ecosystem, the development of search engines has emerged as a critical solution to meet their fundamental need for efficient and reliable information retrieval [8, 31].

The evolution of information retrieval has been significantly advanced by developments in language modeling. Autoregressive models based on Transformer architectures [43] enabled efficient parallel processing and large-scale unsupervised pretraining [23], leading to the emergence of intelligent behaviors in language models. Reinforcement learning from human feedback [14] further refined these models by aligning their outputs with human preferences. Modern large language models (LLMs) have demonstrated human-level performance in reading comprehension and reasoning [39], with their vast parameters enabling the encoding of extensive knowledge [37]. Building upon these advances, retrieval-augmented generation (RAG) has emerged as a framework that integrates information retrieval with LLMs [26]. Studies have shown that RAG can substantially reduce hallucinations while providing up-to-date information through in-context learning [22]. Meanwhile, LLMs enhance search quality through query rewriting, query expansion and decomposition [9]. Based on RAG, generative AI search engines such as Perplexity AI [4], Tiangong AI [5], and Metaso [34] have emerged to provide synthesized answers rather than merely returning links. While conversational LLM products like Gemini, DeepSeek and ChatGPT also employ RAG, *generative AI search engines distinguish themselves by emphasizing their dual role as both search engines and reading tools*, prioritizing answer quality and reading experiences. A survey conducted in the United States found that over a quarter of adults considered switching to AI search engines in 2023 [38].

Despite recent progress, as shown in Fig. 2, applying RAG to web-scale search still faces key challenges. **(1) Handling complex multi-faceted queries**: Web queries often encompass multiple sub-topics, temporal constraints. Treating such queries holistically results in excessive noise and irrelevant retrievals. **(2) Synthesizing knowledge across fragmented chunks**: Chunking in web RAG breaks semantic coherence and logical links, leading to missing information and hallucination in answer [26, 51]. **(3) Presenting rich and readable answers**: Web users need to quickly grasp information while also gaining comprehensive insights, which calls for clear and multi-dimensional presentations that facilitate efficient understanding. To demonstrate the flaw in existing AI search systems, we collected 300 queries across 8 domains and evaluated the answers of Perplexity AI with the help of human experts. Fig. 2 illustrates that several problems are common in the answers.

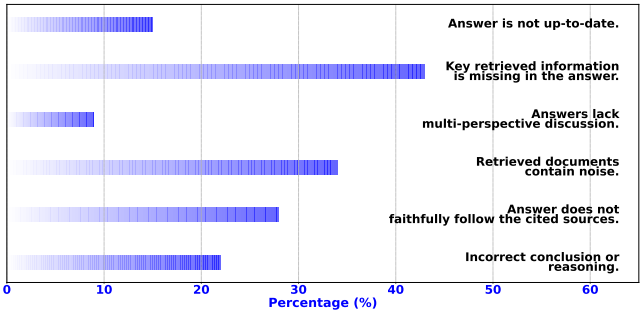

Figure 2: Common issues in generative AI search answers.

To address these challenges, we propose GraTAG, a comprehensive AI search engine framework. To tackle challenge (1), we introduce a **graph-based query decomposition (GQD)** approach that systematically breaks down ambiguous or multi-faceted queries into structured sub-queries with explicit dependencies, enabling more fine-grained retrieval with reduced noise and irrelevance. For challenge (2), we design a **triplet-aligned generation (TAG)** mechanism that dynamically constructs relation triplets from retrieved chunks to explicitly model entity relationships, factual dependencies, and logical connections, enabling the model to reason over fragmented information and generate more coherent and comprehensive answers. For challenge (3), we design a **rich multimodal presentation** that integrates timeline visualizations and textual-visual choreography in answer generation to reduce cognitive load, enhance information assimilation, and enable users to verify synthesized content.

We conduct large-scale, multi-faceted human and LLM evaluations, involving over 243,000 expert ratings across 9 criteria, to rigorously assess system performance. The results demonstrate that GraTAG not only achieves the highest overall quality but also excels in answer representation. Additionally, GraTAG is a commercialized (B2B) product and developed with a focus on stability and safety. An online test showcase is presented in Fig. 1. Our contributions can be summarized as follows:

- We propose GraTAG, together with a comprehensive public disclosure of the full-stack design for a commercialized generative-AI search engine. We detail an industrial-grade architecture covering the end-to-end retrieval-and-generation workflow and demonstrate how it addresses core challenges in existing search systems, including irrelevance, incompleteness, and monolithic presentation.
- We present a modular framework designed to address key challenges in AI search: (1) Graph-based Query Decomposition (GQD), which maps complex queries into structured sub-queries with explicit dependencies; (2) Triplet-Aligned Generation (TAG), which constructs semantic relation triplets to improve reasoning depth; and (3) Rich Multimodal Presentation, which integrates timelines and textual-visual choreography to boost verifiability and user experience.
- Comprehensive evaluations conducted on 1000 real-world queries and over 243,000 expert assessments validate the effectiveness of GraTAG. Compared to the strongest baseline, GraTAG improves comprehensiveness by 10.8%, insightfulness by 7.9%, and the overall average score by 4.8%. On the public benchmark BrowseComp, we outperform the best baseline by 17.3%. The relevant system code is available at https://github.com/tangbotony/GraTAG.

2 Related Work

2.1 Query Optimization for Retrieval

Effective retrieval is critical to system performance, as the quality of retrieved information significantly shapes the final output. Query rewriting techniques aim to transform user queries into more precise and retrieval-friendly formats, addressing ambiguities and enhancing alignment with indexed data [18, 29, 30, 42]. Similarly, query expansion enriches the input by generating alternative or supplementary queries, ensuring the retrieval of a broader and more contextually relevant set of documents [21]. While these techniques improve query quality, web queries remain complex, often spanning multiple sub-topics or temporal constraints. Direct retrieval thus yields noisy and incomplete documents. GraTAG mitigates this by applying graph-based query decomposition (GQD), which decomposes complex queries into parallel or sequential sub-queries to enrich retrieval while maintaining logical consistency.

2.2 Contextual Generation

After retrieving documents, the generation process synthesizes retrieved content into coherent responses. A main challenge in web-based RAG is that chunking breaks semantic coherence and logical connections, while irrelevant information distracts the model and causes hallucinations [49]. To mitigate these issues, various techniques have been proposed, including reference filtering [15], context selection [48], and reranking [53] to reduce noise and highlight salient information. Recent work has explored graph-based approaches such as GraphRAG [17], GFMRAG [28], and PathRAG [12], which construct knowledge graphs from text to model relationships among retrieved knowledge. However, these methods rely on static corpora or predefined schemas, struggling to adapt to dynamically retrieved web content. GraTAG introduces triplet-aligned generation, which dynamically constructs relation triplets from retrieved web documents to restore logical dependencies disrupted by chunking, thereby mitigating hallucinations and enhancing coherence.

2.3 Generative AI Search Engines

Recent systems such as Perplexity AI [4], Tiangong AI [5], Metaso [34], You.com [50], and Microsoft Copilot [33] integrate LLMs with web retrieval to deliver synthesized answers. However, these remain closed-source with limited architectural disclosure. We present GraTAG as a comprehensive case study, detailing our full-stack system design and targeted solutions to query complexity, knowledge fragmentation, and response structuring.

3 GraTAG AI Search

In this section, we first give an overview of our AI search system GraTAG, and then elaborate on core components: graph-based query decomposition (GQD) and triplet-aligned generation (TAG).

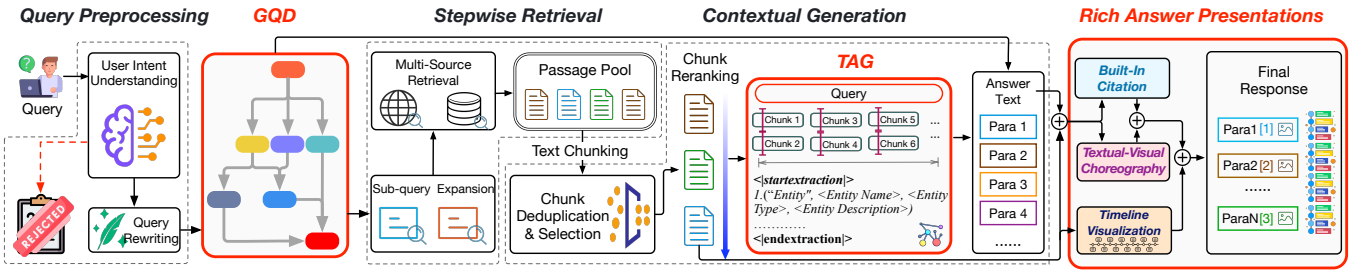

Figure 3: The workflow of GraTAG, which is an AI search system with dedicated components for query preprocessing, retrieval, and generation, featuring graph-based query decomposition, triplet-aligned generation, and rich multimodal presentations.

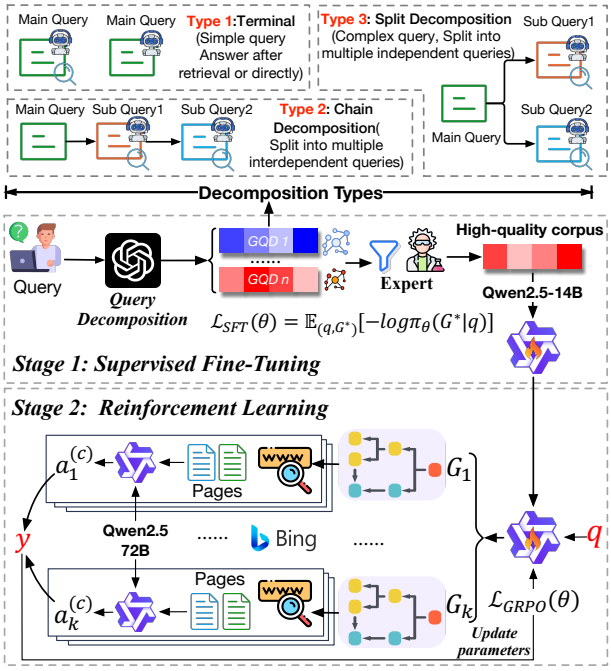

Figure 4: Graph-based query decomposition module.

3.1 Overview

GraTAG is an end-to-end production-ready RAG system comprising seven key stages. As shown in Fig. 3, the pipeline begins with Query Preprocessing, which uses a fine-tuned LLM to filter unsafe content and canonicalize spatiotemporal expressions. Next, Graph-Based Query Decomposition (GQD) (Sec. 3.2) decomposes complex queries into structured sub-queries with explicit dependencies to enable hierarchical reasoning. To improve coverage, Sub-query Expansion generates semantic variations, followed by Stepwise Retrieval, which performs multi-source retrieval guided by the GQD structure. The retrieved documents are then processed with Deduplication and Reranking to remove redundancy and suppress noise. Subsequently, Triplet-Aligned Generation (TAG) (Sec. 3.3) dynamically constructs relation triplets from evidence chunks to restore logical coherence. Finally, Rich Multimodal Presentations (Sec. 3.4) delivers structured outputs with timelines and citations to enhance user experience. Detailed implementation of those stages is provided in Appendix A.

3.2 Graph-Based Query Decomposition (GQD)

Web search queries often exhibit multiple dimensions of complexity, such as multi-intent (e.g., "compare iPhone 15 Pro battery life vs. Samsung S24 and their prices in Asia"), temporal constraints (e.g., "Tesla stock performance after 2023 Q3 earnings vs. industry trends"), and multi-hop reasoning requiring cross-document inference (e.g., "how did Fed rate changes in 2023 impact tech startup valuations"). To address these challenges, we decompose complex queries into atomic sub-queries to enable more precise retrieval and information aggregation. The decomposition is represented as a directed acyclic graph (DAG) that explicitly encodes information flow among sub-queries. In contrast to linear or tree-based approaches [52], our Graph-based Query Decomposition (GQD) captures parallel and joint dependencies, enabling finer-grained and more flexible reasoning.

Specifically, we employ an LLM $\pi_\theta(G \mid q)$ (based on Qwen2.5-72B) to conduct GQD by defining nodes and their pairwise relationships. This GQD model π_θ is post-trained via two stages: supervised fine-tuning on curated query-GQD pairs, followed by reinforcement learning alignment with RAG task performance.

3.2.1 Decomposition Types (Fig. 4 top row). Our GQD model is instructed to follow three decomposition strategies (a combination of them is allowed to fully break down an input query): 1) *Terminal:* The input query is atomic, and decomposition is unnecessary. 2) *Chain decomposition:* A query is sequentially decomposed into a series of sub-queries, where each parent node provides preliminary information for its child node and the descendant nodes, e.g., least-to-most prompting [52]. 3) *Split decomposition:* The query is divided into independent sub-queries that can be processed in parallel.

3.2.2 Stage 1: Supervised Fine-Tuning (SFT) of the GQD Model (Fig. 4 middle row). To improve the performance of GQD model, we first conduct SFT using a set of query-GQD pairs (q, G^*) collected through a systematic curation process. Multiple foundation models are instructed to generate candidate GQDs following the decomposition strategies described above. The generated GQDs are then programmatically validated to ensure that the parent-child relationships meet the requirements and remove duplicate GQDs. Finally, human experts examine the data and select correctly generated high-quality samples. SFT of π_θ is performed via minimizing the loss:

$$\mathcal{L}_{\text{SFT}}(\theta) = \mathbb{E}_{(q,G^*)}\left[-\log \pi_\theta(G^* \mid q)\right]. \tag{1}$$

This stage produces a reference model π_{ref} and initializes the policy model for subsequent reinforcement learning.

3.2.3 Stage 2: Reinforcement Learning (RL) of the GQD Model (Fig. 4 bottom row). To further align GQD generation with downstream RAG performance while stabilizing the reward signal, we adopt a reinforcement learning strategy based on GRPO [16] with environment semi-determinism, variance control, and minimal shaping. For each query q, we sample K candidate GQDs $\{g_k\}_{k=1}^{K}$ from the current policy π_θ. We then evaluate the quality of each g_k based on the likelihood of the pipeline producing the ground-truth response y. Specifically, for each g_k, we perform C independent retrievals to obtain evidence sets $\{a_k^{(c)}\}_{c=1}^{C}$. Top-retrieved URLs/chunks for each sub-query are pre-cached to ensure consistency and reduce web fluctuation noise. For high-similarity queries (similarity > 0.95, measured via BGE-M3 embeddings), cached evidence is automatically reused. Variance is further controlled by averaging over multiple retrievals and clipping extreme rewards. Formally, the reward for candidate g_k is computed as:

$$r_k = \frac{1}{C}\sum_{c=1}^{C} \log \mathcal{M}(y \mid a_k^{(c)}), \tag{2}$$

where $\mathcal{M}(\cdot \mid \cdot)$ denotes the response-generation model (Qwen2.5-72B). Next, we compute group-normalized rewards and use them as advantages:

$$\tilde{r}_k = \frac{r_k - \mu_r}{\sigma_r}, \mu_r = \frac{1}{K}\sum_{j=1}^{K} r_j, \sigma_r = \sqrt{\frac{1}{K}\sum_{j=1}^{K}(r_j - \mu_r)^2}, \hat{A}_{k,t} = \tilde{r}_k. \tag{3}$$

Finally, we optimize the GQD model π_θ via the GRPO objective:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{\substack{(q,y)\sim\mathcal{D},\\ \{g_k\}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}} \Bigg[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|g_k|}\sum_{t=1}^{|g_k|}\min\Big(\rho_{k,t}(\theta)\,\hat{A}_{k,t}, \text{clip}\big(\rho_{k,t}(\theta),\,1-\varepsilon,\,1+\varepsilon\big)\,\hat{A}_{k,t}\Big) - \beta\,\mathbb{D}_{\text{KL}}\big(\pi_\theta \| \pi_{\text{ref}}\big)\Bigg], \tag{4}$$

where

$$\rho_{k,t}(\theta) = \frac{\pi_\theta(g_{k,t} \mid q, g_{k,<t})}{\pi_{\theta_{\text{old}}}(g_{k,t} \mid q, g_{k,<t})}. \tag{5}$$

3.3 Triplet-Aligned Generation (TAG)

Since retrieved documents are split into chunks for information selection and ranking, supporting evidence may span multiple chunks. This fragmentation can break long-range dependencies, leading to information gaps and logical discontinuities that amplify hallucinations. To address this, we propose Triplet-Aligned Generation (TAG), which extracts triplets from the retrieved documents and aligns them with the answer generation process to explicitly bridge missing logic and relations across chunks. By injecting these structural cues, TAG enhances cross-chunk reasoning coherence and mitigates hallucination.

3.3.1 Triplet Extraction Cold Start (Fig. 5 top row). In this paper, triplets encapsulate entity details, inter-entity relations, and implicit factual and logical dependencies, enabling structured cross-chunk reasoning rather than isolated snippet reading. To train triplet extraction stably, we adopt a cold-start stage: For each sub-query, a strong teacher model (e.g., GPT-4o) extracts p parallel triplets

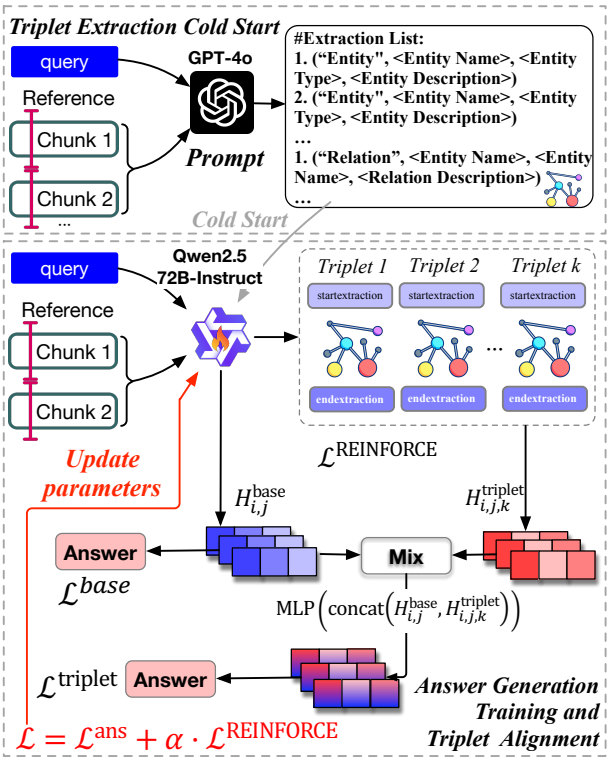

Figure 5: Triplet-aligned generation module.

$t = \{t^1, \ldots, t^p\}$ for each (sub-query, chunks) pair and then we verify their quality manually. We then SFT the target LLM using sub-queries and their corresponding chunks as input, training the model to produce concise and effective triplets. Skipping this stage typically yields lengthy, noisy triplets that fail to guide response generation and harm training stability. During fine-tuning, dedicated extraction tokens ⟨!startextraction!⟩ mark spans, helping the model internalize graph-like structure.

3.3.2 Answer Generation Training and Triplet Alignment (Fig. 5 bottom row). First, we enhance the model's ability to generate answers both with and without triplet augmentation. Specifically, given an instruction-answer pair (x_i, y_i) and its k-th triplet t_i^k. We first obtain the hidden state $H_{i,j}^{\text{base}} = \text{hs}[x_i; y_i^{(j-1)}]$ of the j-th token in the answer under the original RAG method. In our proposed method, when incorporating the triplet t_i^k, the corresponding hidden state is $H_{i,j,k}^{\text{triplet}} = \text{hs}[x_i; y_i^{(j-1)}; t_i^k]$. We concatenate the two obtained hidden states and input them into a three-layer MLP with ReLU activation to compute the weight for each token in the answer:

$$\omega_{i,j,k} = \text{MLP}(\text{concat}(H_{i,j}^{\text{base}}, H_{i,j,k}^{\text{triplet}})). \tag{6}$$

Meanwhile, we compute the log-likelihoods under the conditions of without and with the triplet:

$$\mathcal{L}^{\text{base}} = \log p_\theta(y_i^j \mid x_i, y_i^{(j-1)}) = f(H_{i,j}^{\text{base}}), \tag{7}$$

$$\mathcal{L}^{\text{triplet}} = \log p_\theta(y_i^j \mid x_i, y_i^{(j-1)}, t_i^k) = f(H_{i,j,k}^{\text{triplet}}), \tag{8}$$

where $f(\cdot) = \text{softmax}(\text{lmhead}(\cdot))$. Finally, we compute the weighted sum of the j-th token to obtain the log-likelihood incorporating

the triplet for training answer generation:

$$\mathcal{L}^{\text{ans}} = \omega_{i,j,k} \cdot \mathcal{L}^{\text{base}} + (1 - \omega_{i,j,k}) \cdot \mathcal{L}^{\text{triplet}}. \tag{9}$$

Given that the extracted triplets vary in length and content, we aim to determine the most beneficial triplet for optimizing the answer. Specifically, for i-th instruction-answer pair (x_i, y_i), our objective is to find an optimal triplet $t^{(i)}$ that maximizes $\pi_\theta(y_i \mid x_i, t^{(i)})$. To achieve this, we design a reward function R based on the REINFORCE algorithm to quantify the impact of introducing k-th triplet on answer generation as follows:

$$R_{i,k} = \max(0, \mathcal{L}_i^{\text{base}} - \mathcal{L}_{i,k}^{\text{triplet}} - \bar{R}_i). \tag{10}$$

This reward function increases the likelihood of selecting triplets that perform better than the average reward $\bar{R}_i$. Finally, the REINFORCE loss term is defined as:

$$\mathcal{L}^{\text{REINFORCE}} = -R_{i,k} \cdot \log \pi_\theta(g_k^i \mid x_i). \tag{11}$$

To further encourage succinct triplets, we introduce a length-aware bonus. Let $\ell_{i,k}$ be the tokenized length of t_k, and define the candidate-wise mean $\bar{\ell}_i = 1/K_i \sum_{k=1}^{K_i} \ell_{i,k}$, for the pair (x_i, y_i). We then set $r_{i,k}^{\text{len}} = \gamma\left(\bar{\ell}_i - \ell_{i,k}\right)$, which increases the reward for shorter-than-average triplets and decreases it otherwise ($\gamma > 0$ controls the strength). The augmented reward and the resulting REINFORCE loss are $\tilde{R}_{i,k} = R_{i,k} + r_{i,k}^{\text{len}}$. Finally, the REINFORCE loss is:

$$\mathcal{L}^{\text{REINFORCE}} = -\tilde{R}_{i,k} \cdot \log \pi_\theta(t_k \mid x_i), \tag{12}$$

which encourages the model to select triplets that improve answer generation, while ignoring unhelpful ones. This formulation directly aligns with our GQD-style preference modeling, where log-probability differences capture the relative advantage of candidate knowledge structures. Thus, the total loss function for model training is defined as:

$$\mathcal{L} = \mathcal{L}^{\text{ans}} + \alpha \cdot \mathcal{L}^{\text{REINFORCE}}, \tag{13}$$

where α is a hyperparameter controlling the trade-off between different training objectives.

3.3.3 Answer Generation Inference. For answer generation, the trained model processes sub-queries either sequentially or in parallel based on the dependency structure in the GQD. The generator takes as input both the extracted triplets and the reranked chunks. For nodes with ancestors, the cumulative Q&A pairs of the ancestor chain are prepended to provide contextual information, ensuring that the generation process respects the hierarchical reasoning flow defined by the GQD. When all leaf nodes are processed, their Q&A pairs are concatenated with the main query to form the final comprehensive answer. This triplet-aligned generation approach significantly improves answer coherence and reasoning quality.

3.4 Rich Multimodal Presentations

Traditional chatbots often rely on linear text stacking, which can impose a high cognitive load on users. Given that AI-powered search engines facilitate extensive knowledge transmission, integrating cognitive scaffolding is essential to support user comprehension. Cognitive science research has shown that structured information and multimodal presentations enhance the efficiency of information assimilation [32, 36, 40]. To address these challenges, we incorporate timeline visualizations and textual-visual choreography to optimize the reading experience.

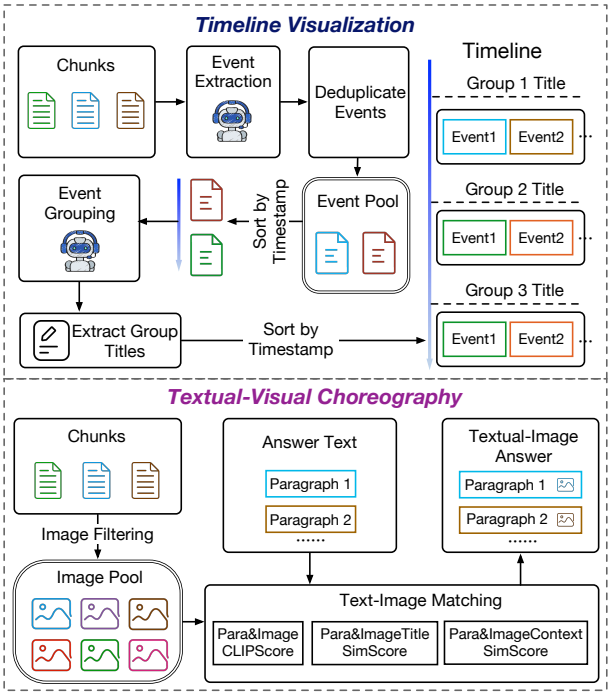

Figure 6: Rich multimodal presentations module.

3.4.1 Timeline Visualization (Fig. 6 top row). In online search scenarios, integrating timeline visualizations enables users to better understand the evolution and context of events. We propose a novel timeline visualization scheme. First, we collect all retrieved chunks following the chunk selection (see Fig. 3). Then, we instruct an LLM (Qwen2.5-14B-Instruct) to extract any event time mentioned in each chunk and to generate a corresponding title and summary. If a chunk does not explicitly mention a time, we resort to using the document's report time as extracted by the same LLM. Chunks lacking temporal information in both the chunk and the document are discarded. Because the retained chunks may describe the same content, we then employ `bge-large` to compute text embedding of the concatenated title and summary and calculate pairwise cosine similarities across chunks. Chunks with a similarity score exceeding 0.9 are merged by discarding the one with the later timestamp, resulting in a list of distinct events with timestamps. To make timeline visualization more structured, we further instruct the LLM to group these events and derive relevant keywords based on their summaries. Finally, we present the event titles for each group, sorted according to their timestamps. Our timeline visualization scheme is integrated within the framework, where upstream pipeline components (GQD, retrieval, and chunk ranking) enhance the relevance and precision of the displayed events. This integration distinguishes our solution from other methods [20, 45], as GQD decomposes complex queries into simpler sub-queries, enabling targeted retrieval,

Table 1: Dataset statistics.

Category	Sources	Number
Event Timeline	TopHub, Google Trends	350
Multi-Constraint Factual	Zhihu, Wikipedia	300
Decision-Making	Stack Exchange, Hacker News	200
Technical Troubleshooting	StackOverflow, GitHub	150

Table 2: Pearson correlation coefficients between human and LLM scores under different evaluation criteria.

Metric	Value	Metric	Value
Conciseness	0.802	Numerical Precision	0.715
Relevance	0.821	Factuality	0.729
Timeliness	0.808	Comprehensiveness	0.642
Clarity	0.646	Coherence	0.727
Insightfulness	0.659	**Average**	**0.891**

while chunk ranking removes noise, together yielding more precise and context-aware results.

3.4.2 Textual-Visual Choreography (Fig. 6 bottom row). A picture is worth a thousand words. GraTAG integrates relevant images into textual answers to enhance information assimilation. These images are extracted from retrieved documents. To ensure quality and relevance, we first filter out noisy images, retaining only those of high quality. Specifically, a rule-based filtering algorithm eliminates logos, icons, and low-resolution images. Subsequently, we compute the similarity between the textual description associated with the image and the main query using bge-reranker-v2-m3 and remove those smaller than 0.3. To determine the optimal placement of images, we compute the pairwise similarity between generated answer paragraphs and candidate images. This computation involves a weighted average of three measures: (1) the embedding cosine similarity between the generated text paragraph and the image, computed using the clip-vit-huge-patch14; (2) the estimated similarity between the synthesized paragraph and the retrieved document's title, obtained via bge-reranker-v2-m3; and (3) the embedding cosine similarity between the synthesized paragraph and the retrieved document's text using bge-large. Pairwise similarities are assembled into a matrix. Then we determine the optimal image-to-text alignment using the Hungarian algorithm [24].

4 Experiments

4.1 Experimental Setup

Multi-faceted Evaluation Criteria. Due to the open-ended nature of answers in real scenarios, we adopted rating criteria for evaluating generated answers rather than computing their match to a fixed reference answer. We invited experts with master's degrees to develop a multi-faceted rating criteria, including: (1) Conciseness, (2) Numerical Precision, (3) Relevance, (4) Factuality, (5) Timeliness, (6) Comprehensiveness, (7) Clarity, (8) Coherence, and (9) Insightfulness. Our nine evaluation metrics were chosen based on two principles: (1) each metric is concise, easily interpretable, and independent; and (2) collectively, they capture the majority of observed issues.

Test Set. We construct a test set of 1,000 real-world complex queries, which is organized into four categories: (1) Event Timeline, which requires reconstructing evolving events or comparing entities across time. (2) Multi-Constraint Factual, which involves multiple explicit constraints such as time, location, scope, or causality. (3) Decision-Making, which specifies a target objective together with feasibility or preference constraints. (4) Technical Troubleshooting, which combines procedural intents with concrete environment constraints (e.g., version or platform). The dataset statistics are summarized in Tab. 1. In addition, we evaluate our method on the public benchmark to examine generalization: BrowseComp [44], an English benchmark where answering requires sustained, multi-step retrieval and yields short, verifiable outputs. We required three ratings for each combination of system, query, and rating criterion. Just Tab. 3 alone contains 243,000 human ratings, underscoring the sheer scale of our evaluation.

Annotation Protocol. GQD and TAG training data each consist of 10,756 samples. Annotation was conducted by 27 master's-level annotators using a 3-way independent evaluation scheme with 2-of-3 majority voting. For TAG triplets, each sample was assessed along four criteria: Factual Correctness, Faithfulness, Coverage, and Conciseness. The direct pass rates were 82.13% for GQD (completed in 12 working days) and 73.62% for TAG (25 working days); all remaining samples were manually corrected before training. The 1,000 test queries are drawn from 8 public sources (Tab. 1) and contain no overlap with the training data used for GQD and TAG.

LLM Evaluation. Evaluating the generated answers for all experiments using multi-faceted criteria by human experts is prohibitively expensive. Therefore, we considered employing an LLM to evaluate the generated text for a subset of experiments. To validate the effectiveness of the LLM as an evaluator, we also conducted a comparison experiment involving both human and LLM (GPT-4o-2024-08-06) evaluators and calculated the Pearson correlation between their final scores. As shown in Tab. 2, human and LLM scores exhibit a high correlation coefficient, indicating that it is feasible to use LLM as an evaluator. Hence, experiments reported in Tabs. 7 to 9, 11 and 12 were evaluated with GPT-4o.

Baselines. We compare GraTAG with eight existing systems, including: **Generative AI search engines**: Perplexity AI [4], Metaso [34], Tiangong AI [5], ChatGLM [11], and Tongyi [6]; and **Conversational LLMs with RAG**: KIMI [3], Ernie Bot [2], and Baichuan [1]. For Perplexity AI, we selected GPT-4o [35] as the backend model. For the GPT-4o automatic evaluation, we also include GINGER [25], a nugget-based RAG baseline. We further conduct ablations on the two core components, GQD and TAG. Additional extended experiments, such as latency analysis and hyperparameter exploration, are provided in the appendices.

4.2 Main Results

4.2.1 Overall Performance. We first compare GraTAG with existing systems under multi-faceted human evaluation in a model-agnostic setting. As shown in Tab. 3, GraTAG achieves the highest average score (9.235 vs. 8.810), outperforming strong baselines, with notable gains in comprehensiveness (9.143 vs. 8.252, $p < 0.001$) and insightfulness (7.333 vs. 6.796, $p < 0.001$). Each response was evaluated by at least three experts. Pairwise Pearson correlations yield

Table 3: Multi-faceted comparison of different systems via human expert evaluation (Higher is better, 10-point scale).

Model	Conciseness	Numerical Precision	Relevance	Factuality	Timeliness	Comprehensiveness	Clarity	Coherence	Insightfulness	Average
Perplexity AI [4]	**9.851**	9.631	9.436	8.525	8.552	7.284	9.612	9.853	6.546	8.810
Tiangong AI [5]	9.840	**9.722**	7.812	8.924	8.103	8.020	9.604	9.802	6.535	8.707
Ernie Bot [2]	9.772	9.328	8.887	8.028	8.404	7.798	9.524	**9.900**	5.969	8.623
KIMI [3]	9.840	9.515	8.529	8.227	8.966	8.155	9.220	9.709	6.796	8.773
Metaso [34]	9.766	8.941	8.515	7.408	8.408	5.689	9.388	9.681	4.755	8.061
ChatGLM [11]	9.817	9.428	8.949	**9.124**	8.347	6.168	9.538	9.726	5.047	8.460
Baichuan [1]	9.661	9.596	6.486	7.611	8.220	8.252	9.223	9.612	6.117	8.309
Tongyi [6]	9.806	9.009	7.585	7.212	8.194	7.677	9.291	9.899	5.859	8.281
GraTAG (Ours)	9.813	9.714	**9.533**	8.932	**9.205**	**9.143**	**9.633**	9.810	**7.333**	**9.235**

Table 4: Performance comparison on the BrowseComp.

Model	Correct Number	Acc (%)
Perplexity AI	298	23.54
Tiangong AI	29	2.29
Ernie Bot	255	20.14
KIMI	330	26.07
Metaso	77	6.08
ChatGLM	152	12.01
Baichuan	80	6.32
Tongyi	26	2.05
GraTAG (Ours)	**387**	**30.57**

Table 5: Comparison of timeline visualization.

Approach	Average Event Count	Precision	Wall Time (s) ↓
CHRONOS [45]	5.12	79%	67.44
GraTAG (Ours)	**10.14**	**84**%	**33.27**

Table 6: Comparison of textual-visual choreography.

Approach	Inclusion (%) ↑	Precision (%) ↑
Metaso [34]	3.0	72.2
GraTAG (Ours)	**80.0**	**90.0**

an average agreement of 0.890, indicating high inter-rater consistency. As a complementary automatic evaluation, Tab. 10 shows that GraTAG achieves a higher average score than GINGER under GPT-4o evaluation (9.312 vs. 9.086). We further evaluate on the BrowseComp benchmark, which contains 1,266 test samples. As shown in Tab. 4, GraTAG achieves the best accuracy (30.57%), outperforming the strongest baseline by 17.3% and demonstrating strong generalization under standardized evaluation.

4.2.2 Performance on Answer Representation. We evaluate the effectiveness of our method in improving user experience and information accessibility. For **Timeline**, we compare our method with CHRONOS [45] using event count (the number of event presented in the timeline), precision (the percentage of relevant events), and wall time (latency for generating the timeline) for online deployment. As shown in Tab. 5, GraTAG outperforms CHRONOS across all dimensions. For **Textual-Visual**, Metaso [34] also implements textual-visual choreography. We compare it with GraTAG by: inclusion (the rate at which images are incorporated into the generated answers) and precision (the percentage of included images that are contextually relevant). As shown in Tab. 6, GraTAG outperforms Metaso significantly on both metrics.

4.3 Ablation Study

4.3.1 System Components. We conduct a comprehensive ablation study to evaluate the contribution of each component in our pipeline. As shown in Tab. 7, removing any module leads to a measurable drop in overall performance, with the full system achieving the highest average score. Notably, omitting GQD results in the most significant degradation, particularly in Relevance and Comprehensiveness, underscoring its critical role in generating focused and thorough answers. Removing TAG also causes a notable decline, especially in Factuality, Insightfulness and Numerical Precision, while also reducing Coherence. Overall, the results demonstrate that the integration of all modules is essential for achieving balanced, high-quality responses across all evaluation criteria.

4.3.2 Graph-based Query Decomposition (GQD). We conduct detailed ablation experiments to evaluate different implementations of GQD. As shown in Tab. 8, the choice of implementation significantly impacts the final answer quality. Using prompt engineering with GPT-4o achieves strong average performance (9.204), demonstrating the effectiveness of the GQD approach itself. Fine-tuning a smaller model (Qwen-2.5-14B-Instruct) with SFT achieves competitive results (8.962), showing that the capability can be distilled into more efficient models. Incorporating reinforcement learning further improves performance, with GRPO achieving the best overall score (9.312) and Reinforce++ [19] also showing strong results (9.181). Notably, GQD with RL shows consistent improvements across all nine evaluation metrics, with particularly substantial gains in Relevance and Comprehensiveness.

4.3.3 Triplet-Aligned Generation (TAG). We evaluate the effectiveness of integrating triplet information into the generation process. As shown in Tab. 9, we compare several TAG variants: entity-only TAG that extracts entities without constructing triplets, Oracle TAG with manually verified relation–entity consistency, prompt engineering with GPT-4o, fine-tuning with Qwen-2.5-14B-Instruct, and the Graph-RAG baseline. Our lightweight TAG design with RL nears Oracle TAG while surpassing both entity-only TAG and Graph-RAG, demonstrating the effectiveness of automatic triplet construction. TAG consistently improves answer quality across implementations, with improvements particularly pronounced in Factuality and Coherence, showing that structured knowledge from triplets enhances both the accuracy and organization of the generated content while maintaining temporal awareness.

Table 7: Ablation study of sub-modules in our approach, "−" indicates skipping the sub-module.

Model	Conciseness	Numerical Precision	Relevance	Factuality	Timeliness	Comprehensiveness	Clarity	Coherence	Insightfulness	Average
Full Approach	**9.797**	**9.747**	**9.559**	**9.222**	9.185	**9.129**	**9.599**	**9.784**	**7.790**	**9.312**
– Query Preprocessing	9.704	9.287	9.352	9.105	9.051	8.810	9.477	9.493	6.890	9.019
– Query Expansion	9.661	9.169	9.491	9.085	**9.200**	8.372	9.363	9.670	6.722	8.970
– GQD	9.615	9.129	7.54	8.695	8.956	7.445	9.032	9.607	6.183	8.467
– Chunk Selection	9.730	9.334	9.392	9.083	9.100	8.582	9.546	9.647	6.924	9.038
– Chunk Rerank	9.711	9.483	9.462	9.031	8.989	8.773	9.465	9.656	6.849	9.047
– TAG	9.564	9.163	9.283	8.397	8.669	8.955	9.103	9.391	7.119	8.849

Table 8: Ablation study on Graph-based Query Decomposition (GQD).

Model	Conciseness	Numerical Precision	Relevance	Factuality	Timeliness	Comprehensiveness	Clarity	Coherence	Insightfulness	Average
– GQD	9.615	9.129	7.540	8.695	8.956	7.445	9.032	9.607	6.183	8.467
Prompt+GPT-4o	**9.798**	9.561	9.419	9.096	9.165	8.825	9.477	**9.807**	7.688	9.204
SFT+Qwen-2.5-14B	9.734	9.088	9.025	8.826	8.959	8.604	9.347	9.636	7.438	8.962
GQD w/ RL (GRPO)	9.797	**9.747**	**9.559**	**9.222**	**9.185**	**9.129**	**9.599**	9.784	**7.790**	**9.312**
GQD w/ RL (Reinforce++)	9.784	9.577	9.436	9.101	8.956	8.987	9.597	9.777	7.415	9.181

Table 9: Ablation study on Triplet-Aligned Generation (TAG).

Model	Conciseness	Numerical Precision	Relevance	Factuality	Timeliness	Comprehensiveness	Clarity	Coherence	Insightfulness	Average
Oracle TAG	9.774	9.786	9.554	9.415	9.175	9.053	9.731	9.811	7.864	9.351
– TAG	9.564	9.163	9.283	8.397	8.669	8.955	9.103	9.391	**7.119**	8.849
Prompt+GPT-4o	9.762	9.682	**9.652**	9.148	8.994	8.638	9.496	**9.869**	7.371	9.179
SFT+Qwen-2.5-14B	9.593	9.292	9.367	8.702	8.832	8.455	9.355	9.488	7.127	8.912
entity-only TAG	9.545	9.175	9.293	8.413	8.664	8.862	9.200	9.363	7.082	8.844
Graph-RAG	9.442	9.207	9.383	8.789	9.053	8.954	9.226	9.510	7.588	9.017
TAG w/ RL (GRPO)	**9.797**	**9.747**	9.559	**9.222**	**9.185**	**9.129**	**9.599**	9.784	**7.790**	**9.312**

4.4 Production Deployment Analysis

4.4.1 Test-Time Latency Analysis. In GraTAG, we maximize parallelism to reduce test-time latency, e.g., executing lateral sub-queries in GQD concurrently. We further fine-tune and quantize candidate models, selecting the fastest one that satisfies our performance requirements. In deployment, GraTAG achieves a response latency of 14.2 seconds. For comparison, the latencies of Perplexity, Tiangong, ErnieBot, KIMI, Metaso, ChatGLM, Baichuan, and Tongyi are 13.9, 4.0, 6.0, 2.8, 10.4, 7.9, 6.1, and 10.4 seconds, respectively. The latency of GraTAG is measured on a cluster of 16 Muxi MXC500 GPUs (each providing approximately 70% of the computing power of an NVIDIA A800), while baselines are evaluated via their publicly available interfaces. In production deployment, latency is further reduced via two caching strategies: (1) query-level caching, where identical or highly similar queries within a 10-minute window return cached results directly; and (2) retrieval result caching, which reuses previously retrieved URLs and chunks.

4.4.2 Performance on Simple Queries. To examine GraTAG's behavior on non-complex inputs, we isolate a low-complexity subset of 121 queries (decomposition nodes ≤ 2) from the 1,000-query benchmark. All systems score higher on this easier subset, and the overall ranking remains consistent with the full benchmark: GraTAG (9.44) > Perplexity (9.29) > KIMI (9.24) > Tiangong (9.16) > Ernie Bot (9.08) > ChatGLM (9.03) > Baichuan (8.93) > Tongyi (8.90) > Metaso (8.69), confirming that GraTAG's advantage generalizes beyond complex queries.

4.4.3 Online Deployment and A/B Test. GraTAG has been commercially deployed as the product Xinyu at Xinhua News Agency (21,765 users) and China Telecom (85,439 users). Prior to full deployment, a blind evaluation by 10 expert journalists on 150 real business queries showed Xinyu outperforming KIMI [3] on 85 queries versus 44 (56.67% vs. 29.33%, with 14.00% ties). Following deployment, a 14-day online A/B test (8,752 FlexRAG users vs. 8,754 Xinyu users, bucketed by user ID) showed that Xinyu increased active users' average daily Q&A count from 2.73 to 5.56 and average usage days within the 14-day window from 2.46 to 5.28, indicating clear improvements in online engagement and user stickiness. All blind-evaluation queries and scoring results are available on our GitHub repository.

5 Conclusion

In this work, we present GraTAG, a generative AI search engine designed to tackle multi-faceted challenges in answer generation and user experience through a fully integrated pipeline. Our approaches not only build on state-of-the-art research but also introduce novel solutions to specific challenges. Extensive experiments demonstrate the superiority of GraTAG over existing technologies. The system has been successfully deployed in real-world applications, with finalized industry contracts now exceeding USD 2 million and expansion to English-language scenarios (e.g., Suanova), validating its practical value and scalability.

Acknowledgements

This work was supported by NSFC grant 12425113 and the Basic Research Program of Jiangsu under Grant BK20253039, the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2025C02037), the Research Project of the Philosophy and Social Sciences Planning Research Platform (Laboratory) of Zhejiang Province, and the Key R&D Program of Hangzhou (2025SZDA0254).

References

[1] Baichuan AI. 2024. Baichuan. https://ying.baichuan-ai.com/ Accessed: 2025-02-05.

[2] Baidu AI. 2024. Yiyan. https://yiyan.baidu.com/ Accessed: 2025-02-05.

[3] Moonshot AI. 2024. KIMI. https://kimi.moonshot.cn/ Accessed: 2025-02-05.

[4] Perplexity AI. 2024. Perplexity AI. https://www.perplexity.ai/ Accessed: 2025-02-05.

[5] Tiangong AI. 2024. Tiangong AI. http://tiangong.cn/ Accessed: 2025-02-05.

[6] Tongyi AI. 2024. Tongyi. https://tongyi.ai/ Accessed: 2025-02-05.

[7] Alec Berntson. 2023. Azure AI Search: Outperforming Vector Search with Hybrid Retrieval and Reranking. https://techcommunity.microsoft.com/blog/azure-ai-services-blog/azure-ai-search-outperforming-vector-search-with-hybrid-retrieval-and-reranking/3929167 Accessed: 2025-02-01.

[8] Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks* 30 (1998), 107–117.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[10] Harrison Chase. 2022. *LangChain.* https://github.com/langchain-ai/langchain

[11] ChatGLM. 2024. ChatGLM. https://chatglm.cn/ Accessed: 2025-02-05.

[12] Boyu Chen, Zirui Guo, Zidan Yang, Yuluo Chen, Junze Chen, Zhenghao Liu, Chuan Shi, and Cheng Yang. 2025. PathRAG: Pruning Graph-based Retrieval Augmented Generation with Relational Paths. *arXiv preprint arXiv:2502.14902* (2025).

[13] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv:2402.03216 [cs.CL]

[14] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[15] Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR* (2023).

[16] DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437 [cs.CL] https://arxiv.org/abs/2412.19437

[17] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130* (2024).

[18] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise Zero-Shot Dense Retrieval without Relevance Labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 1762–1777.

[19] Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. 2025. Reinforce++: An efficient rlhf algorithm with robustness to both prompt and reward models. *arXiv preprint arXiv:2501.03262* (2025).

[20] Qisheng Hu, Geonsik Moon, and Hwee Tou Ng. 2024. From Moments to Milestones: Incremental Timeline Summarization Leveraging Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7232–7246. doi:10.18653/v1/2024.acl-long.390

[21] Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *arXiv preprint arXiv:2305.03653* (2023).

[22] Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 7969–7992. doi:10.18653/v1/2023.emnlp-main.495

[23] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[24] H. W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* 2, 1-2 (1955), 83–97. doi:10.1002/nav.3800020109 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/nav.3800020109

[25] Weronika Łajewska and Krisztian Balog. 2025. GINGER: Grounded Information Nugget-Based Generation of Responses. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2723–2727. doi:10.1145/3726302.3730166

[26] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[27] Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023. Making Large Language Models A Better Foundation For Dense Retrieval. arXiv:2312.15503 [cs.CL]

[28] Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Dinh Phung, Chen Gong, and Shirui Pan. 2025. GFM-RAG: Graph Foundation Model for Retrieval Augmented Generation. *arXiv preprint arXiv:2502.01113* (2025).

[29] Xinbei Ma, Yeyun Gong, Pengcheng He, hai zhao, and Nan Duan. 2023. Query Rewriting in Retrieval-Augmented Large Language Models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. https://openreview.net/forum?id=gXq1cwkUZc

[30] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query Rewriting in Retrieval-Augmented Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5303–5315.

[31] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval.* Cambridge University Press.

[32] Richard E. Mayer. 2014. Cognitive Theory of Multimedia Learning. In *The Cambridge Handbook of Multimedia Learning*, Richard E. Mayer (Ed.). Cambridge University Press, Cambridge, 43–71.

[33] Yusuf Mehdi. 2023. Reinventing Search with a New AI-powered Microsoft Bing and Edge, Your Copilot for the Web. Microsoft Official Blog. https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

[34] Metaso. 2024. Metaso. https://metaso.cn/ Accessed: 2025-02-05.

[35] OpenAI. 2024. ChatGPT. https://chatgpt.com/ Accessed: 2025-02-05.

[36] Luis P. Prieto, Kshitij Sharma, Łukasz Kidzinski, María Jesús Rodríguez-Triana, and Pierre Dillenbourg. 2018. Multimodal Teaching Analytics: Automated Extraction of Orchestration Graphs from Wearable Sensor Data. *Journal of Computer Assisted Learning* 34, 2 (April 2018), 193–203. doi:10.1111/jcal.12232

[37] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[38] Statista. 2023. https://www.statista.com/statistics/1377993/us-adults-ai-powered-search-engines-usage-choice/ Accessed: 2025-01-21.

[39] Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Robin IM Dunbar, et al. 2024. LLMs achieve adult human performance on higher-order theory of mind tasks. *arXiv preprint arXiv:2405.18870* (2024).

[40] John Sweller, Paul Ayres, and Slava Kalyuga. 2020. Cognitive load theory and educational technology. *Educational Technology Research and Development* 68, 1 (2020), 1–16. doi:10.1007/s11423-019-09701-3

[41] Robert Endre Tarjan and Anthony E Trojanowski. 1977. Finding a maximum independent set. *SIAM J. Comput.* 6, 3 (1977), 537–546.

[42] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*. 10014–10037.

[43] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[44] Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. arXiv:2504.12516 [cs.CL] https://arxiv.org/abs/2504.12516

[45] Weiqi Wu, Shen Huang, Yong Jiang, Pengjun Xie, Fei Huang, and Hai Zhao. 2025. Unfolding the Headline: Iterative Self-Questioning for News Retrieval and Timeline Summarization. *arXiv preprint arXiv:2501.00888* (2025).

[46] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. arXiv:2309.07597 [cs.CL]

[47] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115* (2024).

[48] Haoyan Yang, Zhitao Li, Yong Zhang, Jianzong Wang, Ning Cheng, Ming Li, and Jing Xiao. 2023. PRCA: Fitting Black-Box Large Language Models for Retrieval Question Answering via Pluggable Reward-Driven Contextual Adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5364–5375. doi:10.18653/v1/2023.emnlp-main.326

[49] Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. In *The Twelfth International Conference on Learning Representations*.

[50] You.com. 2024. You.com: The AI Search Engine You Control. https://you.com/. Accessed: 2024-10-04.

[51] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv preprint arXiv:2309.01219* (2023).

[52] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=WZH7099tgfM

[53] Shengyao Zhuang, Bing Liu, Bevan Koopman, and Guido Zuccon. 2023. Open-source Large Language Models are Strong Zero-shot Query Likelihood Models for Document Ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 8807–8817.

A Other Implementations in the Workflow

A.1 Query Preprocessing

To ensure robustness and practical deployment in real-world scenarios, GraTAG incorporates a comprehensive query preprocessing pipeline that handles edge cases and user intent clarification before GQD generation. Upon receiving a query, a fine-tuned LLM (currently based on Qwen-2.5-14B) is employed to filter out unsafe or harmful queries and clarifies ambiguous ones by prompting the user with options, such as suggesting a specific region for a general query like "the current state of the economy." If the query is ambiguous or lacks specificity, the LLM prompts the user with clarifying questions or options to refine his/her intent. The query might be rejected if it is regarded as one that warrants refusal.

Following this, any open-source LLM (currently Qwen-2.5-14B) is used to rewrite the query by translating relative spatiotemporal terms (e.g., “last week”) into precise timestamp ranges, resolving vague locations (“nearby”) to specific places, and canonicalizing incomplete entity names.

A.2 Sub-query Expansion

Given a sub-query in the graph after GQD, we ask an LLM to generate multiple relevant queries. These queries are called retrieval queries since they are subsequently entered into a search engine to retrieve more information. Specifically, the LLM is instructed to act as a subject matter expert in a university, expanding the given query to create related questions that assess students’ comprehensive understanding of the topic across multiple dimensions: 1) content mastery, 2) understanding of key elements, 3) contextual analysis, and 4) extended thinking.

A.3 Retrieval

To ensure the recency and relevance of retrieved information, we submit each retrieval query to multiple search engines simultaneously, leveraging the ranking algorithms in search engines to obtain more comprehensive content and mitigate biases. Additionally, we incorporate `Milvus` and `Elasticsearch` for our B2B clients to build local retrieval systems.

Directly feeding raw retrieved documents (e.g., a web page) into an LLM can lead to high perplexity and degraded response quality. To make the input content LLM-friendly and safe, we implement a robust content filtering pipeline. This process involves removing disruptive elements, filtering sensitive or extraneous information, and standardizing formatting.

After filtering, we segment documents into chunks using the `RecursiveCharacterTextSplitter` method [10]. We adopt a small chunk size (i.e., 350) with a relatively large overlap (i.e., 25%) to optimize the performance of the subsequent text embedding model, following the research findings by Azure AI Search [7].

A.4 Chunk Deduplication and Reranking

Notably, retrieved passages vary in relevance and often contain duplicate information, which can distract the model. To mitigate this problem, we use a fine-tuned text embedding model (bge-large [46]) to compute embeddings for the chunks and then calculate pairwise cosine similarities. We aim to identify the largest subset of chunks where the similarity score between any two chunks is no greater than 0.8. Finding the optimal solution to this problem corresponds to solving the maximum independent set problem [41], which is NP-hard. To improve computational efficiency, we adopt a greedy strategy that processes each chunk sequentially, retaining it only if its similarity to all previously retained chunks remains below 0.8. For re-ranking, we finetune a reranker model (bge-reranker-v2-m3 [13, 27]) to sort the chunks based on their similarity to the sub-query.

B Further Analysis

B.1 Impact of Fine-Tuning on Answer Generation

In GraTAG, we fine-tune LLMs for entity extraction, GQD generation, and answer generation. We validate their effectiveness through ablation studies. As shown in Tab. 11, replacing these modules with GPT-4o and Qwen-2.5-72B confirms that improvements in query rewriting, entity extraction, and GQD generation positively impact final answer quality. Tab. 12 demonstrates that fine-tuning the answer generation model yields consistent performance gains across eight evaluation metrics, including Conciseness and Numerical Precision.

B.2 Effect of Reranking Model Selection

We conduct an ablation study on the reranking component by comparing our fine-tuned reranker model against its base model (bge-reranker-v2-m3) and two LLMs instructed to generate a ranking order based on relevance. As shown in Tab. 13, our fine-tuned method preserves the efficiency of the base model while significantly outperforming it and achieving performance comparable to GPT-4o. The wall time for GPT-4o reflects API response time, whereas the wall time for other models is measured on a local cluster. This demonstrates that fine-tuning allows us to achieve strong performance with more efficient models, maintaining low latency while approaching the quality of much larger language models.

B.3 Impact of Number of Sampled GQDs

We further investigate the optimal number of sampled GQDs. As shown in Tab. 14, performance improves as the number of GQDs increases from 1 to 8. Beyond this point, the performance remains relatively stable with diminishing returns, suggesting that 8 GQDs provide an effective balance between query coverage and computational efficiency.

Table 10: Multi-faceted comparison of different systems via GPT-4o automatic evaluation. Higher is better, and 10 is the maximum. Best scores are in bold and second-best scores are underlined.

Model	Conciseness	Numerical Precision	Relevance	Factuality	Timeliness	Comprehensiveness	Clarity	Coherence	Insightfulness	Average
Perplexity AI	**9.844**	9.528	**9.670**	9.659	8.057	8.194	9.811	9.800	6.547	9.012
Tiangong AI	9.528	8.934	9.315	9.479	7.473	7.239	9.543	9.540	5.895	8.550
Ernie Bot	9.626	8.928	9.334	9.415	7.836	7.692	9.518	9.601	6.362	8.701
KIMI	9.574	9.259	9.449	9.580	7.941	8.176	9.578	9.614	6.312	8.831
Metaso	9.215	8.346	8.920	9.013	7.005	6.248	9.143	8.993	5.203	8.010
ChatGLM	9.723	9.275	9.562	**9.715**	7.947	7.917	**9.826**	**9.809**	6.599	8.930
Baichuan	8.962	8.575	8.812	8.936	7.348	7.357	8.888	8.753	6.171	8.200
Tongyi	9.321	8.467	8.904	9.101	7.288	7.534	9.394	9.741	6.063	8.424
GINGER [25]	9.629	9.358	9.432	8.807	9.011	8.901	9.374	9.597	7.665	9.086
GraTAG (Ours)	9.797	**9.747**	9.559	9.222	**9.185**	**9.129**	9.599	9.784	**7.790**	**9.312**

Table 11: Ablation study of replacing our fine-tuned LLMs with proprietary and open-source models.

Model	Conciseness	Numerical Precision	Relevance	Factuality	Timeliness	Comprehensiveness	Clarity	Coherence	Insightfulness	Average
GPT-4o	9.745	9.622	**9.633**	8.940	8.823	9.065	9.456	9.754	7.147	9.132
Qwen 2.5-72B	9.698	9.485	9.475	8.517	8.912	8.709	9.592	9.670	7.330	9.043
GraTAG (Ours)	**9.797**	**9.747**	9.559	**9.222**	**9.185**	**9.129**	**9.599**	**9.784**	**7.790**	**9.312**

Table 12: Ablation study of replacing our fine-tuned answer generation model with proprietary and open-source models.

Model	Conciseness	Numerical Precision	Relevance	Factuality	Timeliness	Comprehensiveness	Clarity	Coherence	Insightfulness	Average
GPT-4o	9.771	9.681	**9.600**	9.095	8.971	9.110	9.550	9.771	7.661	9.246
Qwen 2.5-72B	9.741	9.577	9.563	8.849	8.794	9.005	9.549	9.72	7.546	9.149
GraTAG (ours)	**9.797**	**9.747**	9.559	**9.222**	**9.185**	**9.129**	**9.599**	**9.784**	**7.790**	**9.312**

Table 13: Ablation study of fine-tuning the rerank model.

Model	Precision	Recall	F1 Score	Wall Time (s)
GPT-4o [35]	**0.717**	0.719	**0.718**	3.6
Qwen 2.5-72B [47]	0.541	**0.894**	0.674	2.4
bge-reranker-v2-m3 [13]	0.568	0.671	0.615	0.1
GraTAG (ours)	0.627	0.735	0.677	0.1

Table 14: Impact of the number of sampled GQDs.

Number	Average Score
1	9.119
4	9.290
8	9.312
16	9.304
32	**9.314**

Table 15: Text-Image Matching Performance with bge-reranker-v2-m3 at Different Thresholds

Threshold	Inclusion (%)	Precision (%)
0.1	93	45
0.2	85	80
0.3	80	90
0.4	68	93
0.5	43	95

B.4 Threshold Analysis for Text-Image Choreography

We analyze the matching threshold for Text-Image Choreography using bge-reranker-v2-m3. As shown in Table 15, the threshold controls the trade-off between inclusion (recall) and precision: lower thresholds (0.1) achieve high inclusion (93%) but low precision (45%), while higher thresholds (0.5) improve precision (95%) at the cost of reduced inclusion (43%). We select threshold 0.3 as it provides the optimal balance with 80% inclusion and 90% precision, ensuring comprehensive coverage while maintaining answer quality.