

Knowledge Graph-Guided Retrieval Augmented Generation

ارائه دهنده: علی معمري

تير ماه ۱۴۰۴

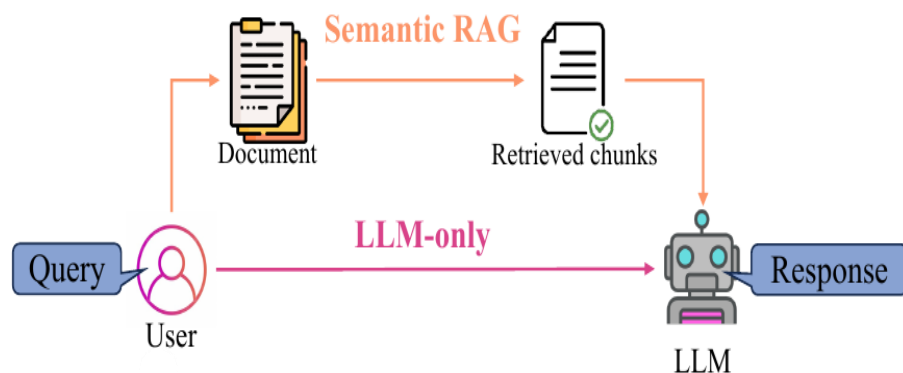
Knowledge Graph-Guided Retrieval Augmented Generation

Xiangrong Zhu[♣] Yuexiang Xie[♡] Yi Liu[♣] Yaliang Li[♡] Wei Hu[♣]
[♣] State Key Laboratory for Novel Software Technology, Nanjing University, China
[♡] Alibaba Group
{xrzhu, yiliu07}.nju@gmail.com, whu@nju.edu.cn
{yuexiang.xyx, yaliang.li}@alibaba-inc.com

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
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

DOI: <https://doi.org/10.18653/v1/2025.naacl-long.449>

۱. مقدمه



- ❑ مدل‌های زبانی بزرگ در انجام وظایف مختلف مانند پاسخ به پرسش‌ها، کمک در نوشتن، و تولید کد، به موفقیت چشمگیری دست یافته‌اند.
- ❑ با این حال، یکی از چالش‌های اساسی، وجود «توهمات» (hallucinations) در پاسخ‌های تولیدشده است.
- ❑ توهمات اغلب ناشی از اطلاعات قدیمی یا نبود دانش تخصصی در یک حوزه خاص هستند.
- ❑ «تولید تقویت‌شده با بازیابی» (RAG) به عنوان یک راه‌حل عملی برای کاهش توهمات مطرح شده است که با بازیابی دانش مرتبط از اسناد، این مشکل را کاهش می‌دهد.

- ❑ مطالعات فعلی در زمینه RAG عمدتاً از رویکردهای مبتنی بر کلیدواژه یا معنا برای بازیابی شبیه‌ترین اسناد (یا قطعه‌ها (chunk)) به پرسش کاربر استفاده می‌کنند.
- ❑ قطعه‌های بازیابی‌شده ممکن نامربوط باشند و نمی‌توانند روابط درونی را به‌خوبی ارائه دهند.
- ❑ این قطعه‌ها اغلب فاقد روابط سطح‌واقعیت (fact-level) میان یکدیگر هستند، که این امر توانایی استدلال مدل‌های زبانی بزرگ را محدود می‌کند.
- ❑ چسباندن مستقیم قطعه‌های بازیابی شده صرفاً بر اساس امتیاز شباهت، می‌تواند منجر به قطعه‌های مجزایی (isolated) از اطلاعات شود که درک جامع و پاسخ‌دهی قابل‌اعتماد را محدود می‌کند.



- ❑ گراف‌های دانش انتزاع‌های ساخت‌یافته‌ای از موجودیت‌های دنیای واقعی و روابط میان آن‌ها هستند.
- ❑ دانش موجود در یک گراف دانش به‌صورت سه‌تایی (موجودیت آغازین، رابطه، موجودیت پایانی) نمایش داده می‌شود که از طریق موجودیت‌های مشترک به هم مرتبط هستند.
- ❑ گراف‌های دانش می‌توانند با ادغام دانش ساخت‌یافته و واقعی، رویکردهای معنایی موجود در RAG را به‌طور مؤثری تکمیل کنند.
- ❑ آن‌ها روابط سطح‌واقعیت را میان بخش‌های مختلف فراهم می‌کنند و حقایق مهمی را برجسته می‌سازند که ممکن است در رویکردهای صرفاً معنایی نادیده گرفته شوند.

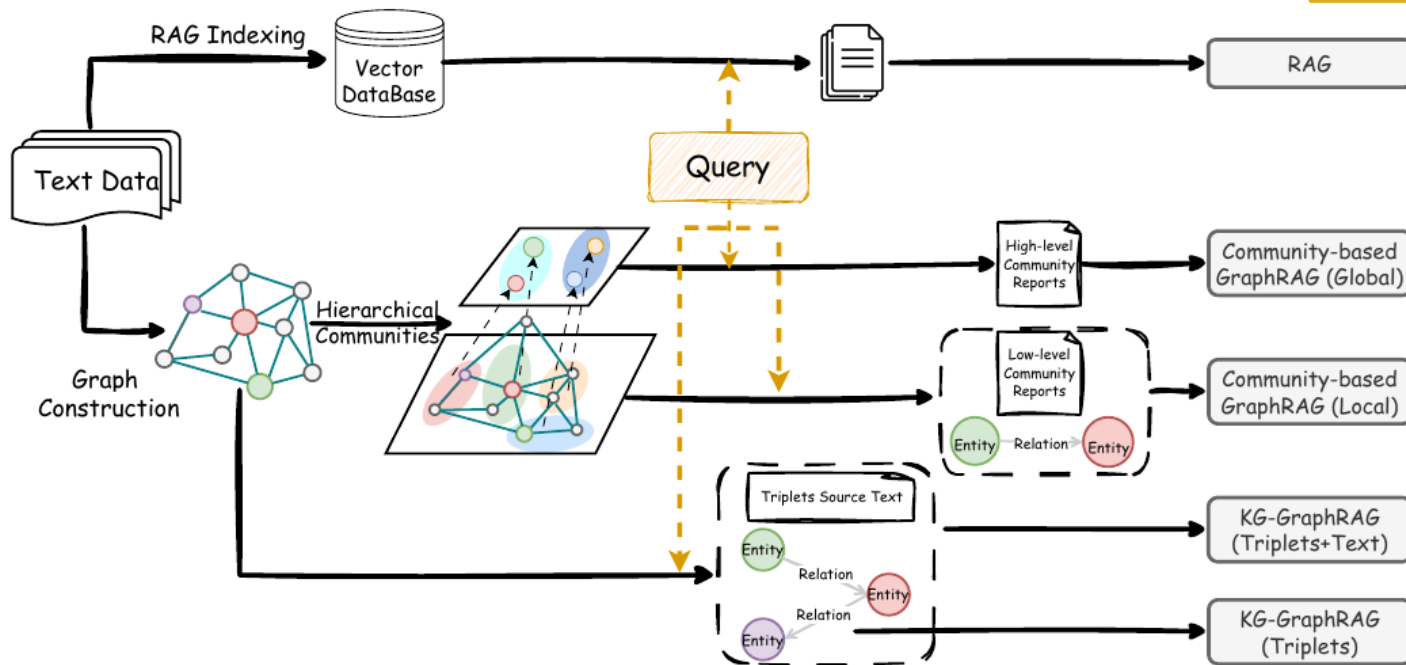


Figure 1: The illustration of RAG, KG-based GraphRAGs and Community-based GraphRAGs.

RAG vs. GraphRAG: A Systematic Evaluation and Key Insights
[arXiv:2502.11371](https://arxiv.org/abs/2502.11371)



- KG2RAG یک چارچوب نوآورانه است که از گراف‌های دانش بهره می‌برد.
- هدف: فراهم کردن روابط سطح‌واقعیت میان بخش‌ها (chunks) به‌منظور بهبود تنوع و انسجام نتایج بازیابی‌شده.
- فرآیندهای کلیدی:
 - 1. بازیابی مبتنی بر معنا: بازیابی «قطعه‌های اولیه» (seed chunks) به‌عنوان نقطه‌ی شروع.
 - 2. گسترش قطعه‌ها با کمک گراف دانش: گسترش قطعه‌های بازیابی‌شده با استفاده از ارتباطات موجود در گراف دانش.
 - 3. سازمان‌دهی بخش‌ها بر اساس گراف دانش: ارائه‌ی دانش مرتبط و مهم در قالب بندهای ساخت‌یافته و منسجم.
- مزایای اثبات‌شده: KG2RAG در معیارهای کیفیت پاسخ و کیفیت بازیابی، عملکرد بهتری نسبت به روش‌های RAG موجود در مجموعه داده‌های HotpotQA و نسخه‌های مختلف آن نشان داده است.

۲. جریان کاری KG2RAG

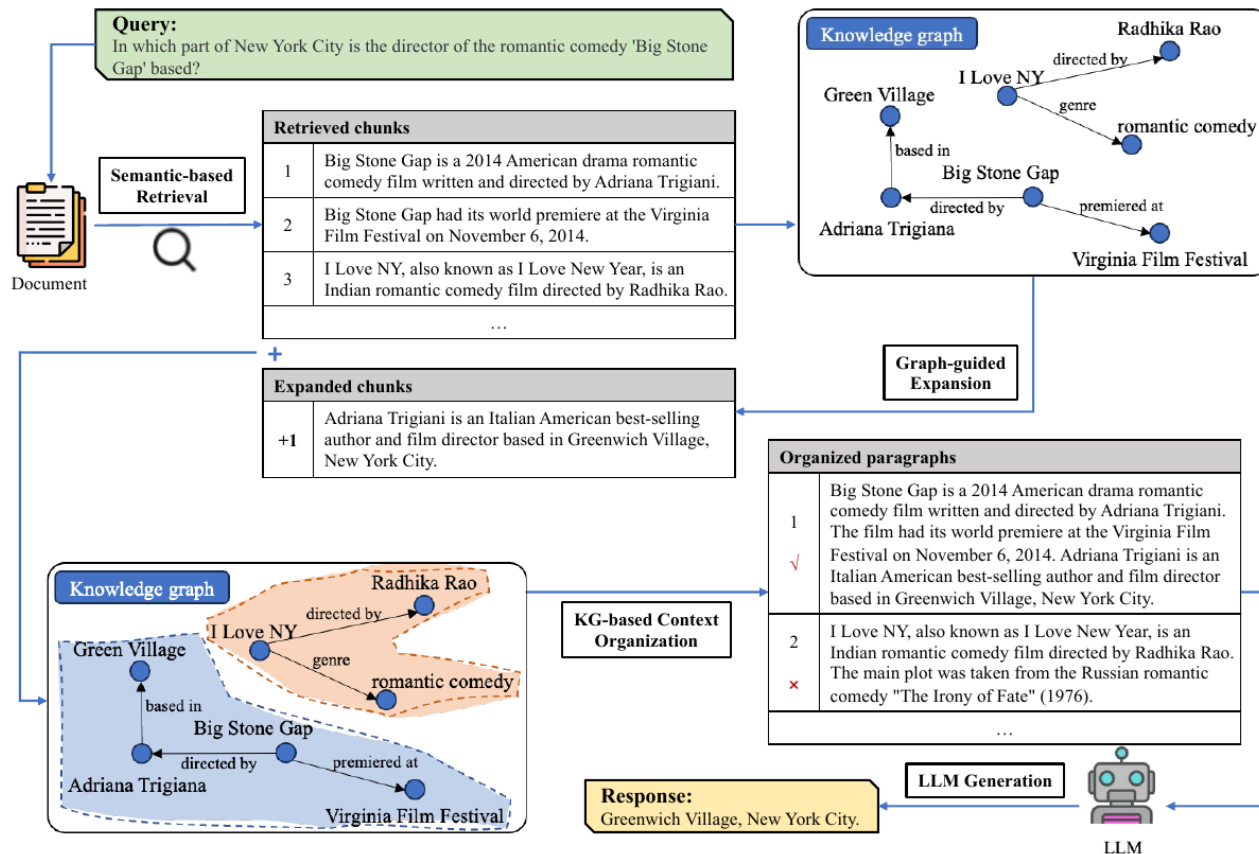


Figure 2: Workflow of the proposed KG²RAG.

□ جریان کاری KG2RAG شامل موارد زیر است:

- پردازش آفلاین اسناد
- بازیابی قطعه‌های بهبود یافته با گراف دانش
- سازمان‌دهی زمینه (context) بر اساس گراف دانش

□ این رویکرد اطمینان حاصل می‌کند که قطعه‌های بازیابی‌شده متنوع، مرتبط و سازگار هستند، و پاراگراف‌هایی منظم و مناسب برای مدل‌های زبانی بزرگ تشکیل می‌دهند.

۱- پردازش آفلاین اسناد

- **قطعه‌بندی (chunking):** اسناد به n قطعه تقسیم می‌شوند، معمولاً بر اساس ساختار جمله/پاراگراف و یک اندازه قطعه از پیش تعیین شده.

$$\mathcal{D} = \{c_1, \dots, c_n\}.$$

- **مرتبط کردن قطعه‌ها و گراف دانش:** برای درک روابط غنی در سطح حقایق بین قطعه‌ها، هر قطعه به یک گراف دانش مرتبط می‌شود.
- از مدل‌های زبانی بزرگ و مهندسی اعلان برای استخراج موجودیت‌ها و روابط از قطعه‌ها برای ساخت زیرگراف‌هایی که در نهایت با هم ترکیب شده و یک گراف کامل را تشکیل می‌دهند استفاده می‌شود.

$$\mathcal{G} = \{(h, r, t, c) \mid c \in \mathcal{D}\}$$

- این فرایند مستقل از پرسش (query-independent) بوده و به صورت آفلاین انجام می‌شود؛ فقط یک بار ساخته می‌شود و از به روزرسانی‌های تدریجی پشتیبانی می‌کند.

Prompt for Triplet Extraction

Instruction:

Extract informative triplets directly from the text following the examples. Do not add any extra words, line breaks, or spaces.

Example 1:

Text: Scott Derrickson (born July 16, 1966) is an American director, screenwriter and producer.

Triplets:

<Scott Derrickson, born in, 1966>,
<Scott Derrickson, nationality, America>,
<Scott Derrickson, occupation, director>,
<Scott Derrickson, occupation, screenwriter>,
<Scott Derrickson, occupation, producer>

Example 2:

Text: A Kiss for Corliss is a 1949 American comedy film directed by Richard Wallace and written by Howard Dimsdale.

Triplets:

<A Kiss for Corliss, year, 1949>,
<A Kiss for Corliss, country, America>,
<A Kiss for Corliss, genre, comedy film>,
<A Kiss for Corliss, director, Richard Wallace>,
<A Kiss for Corliss, writer, Howard Dimsdale>

Target Text: <target text>

Triplets:

Figure 3: The prompt for triplet extraction.

۲- بازیابی قطعه‌های بهبودیافته با گراف دانش

□ فرآیند بازیابی دو مرحله‌ای:

۱. بازیابی مبتنی بر معنا (Semantic-based Retrieval):

■ شباهت معنایی بین پرسش کاربر (q) و تمام قطعه‌ها (c) اندازه‌گیری می‌شود.

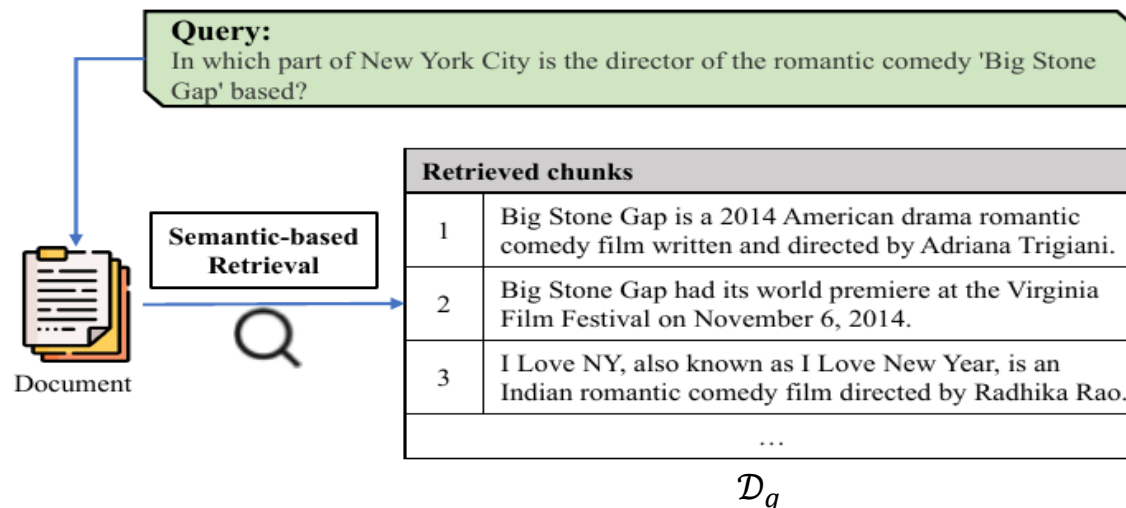
$$\mathcal{S} = \{s(q, c) \mid c \in \mathcal{D}\}$$

- تابع s شباهت کسینوسی بردار تعبیه (embedding) q و c را محاسبه می‌کند.
- شبیه‌ترین k قطعه با بالاترین میزان شباهت انتخاب می‌شوند Dq. این قطعه‌ها به عنوان «قطعه‌های اولیه (seed chunks)» استفاده می‌شوند.
- محدودیت تکیه صرف بر بازیابی معنایی: تکیه صرف بر این روش ممکن است منجر به انتخاب قطعه‌های پراکنده و بدون ارتباط شود که در نتیجه، اطلاعات واقعی مهمی از دست می‌رود.

۲- بازیابی قطعه‌های بهبودیافته با گراف دانش

□ فرآیند بازیابی دو مرحله‌ای:

۱. بازیابی مبتنی بر معنا (Semantic-based Retrieval):



۲- بازیابی قطعه‌های بهبودیافته با گراف دانش

□ فرآیند بازیابی دو مرحله‌ای:

۲. گسترش قطعه‌ها با کمک گراف دانش (Graph-guided Expansion):

- انگیزه: این روش یک قطعه را از طریق موجودیت‌های مشترک به قطعه‌های دیگر متصل می‌کند.
- فرآیند:

- یک زیرگراف مرتبط G_q^0 از روی قطعه‌های اولیه (seed chunks) استخراج می‌شود.

$$\mathcal{G}_q^0 = \{(h, r, t, c) \mid c \in \mathcal{D}_q\} \subseteq \mathcal{G}$$

- با استفاده از الگوریتم جستجوی اول-سطح (BFS)، ناحیه m-همسایگی گراف $\mathcal{G}_q^0$ پیمایش می‌شود تا موجودیت‌ها و یال‌های مرتبط شناسایی شوند.

$$\mathcal{G}_q^m = \text{traverse}(\mathcal{G}, \mathcal{G}_q^0, m)$$

- تمام قطعه‌هایی که با این زیرگراف گسترش‌یافته مرتبط هستند، به عنوان قطعه‌های گسترش‌یافته

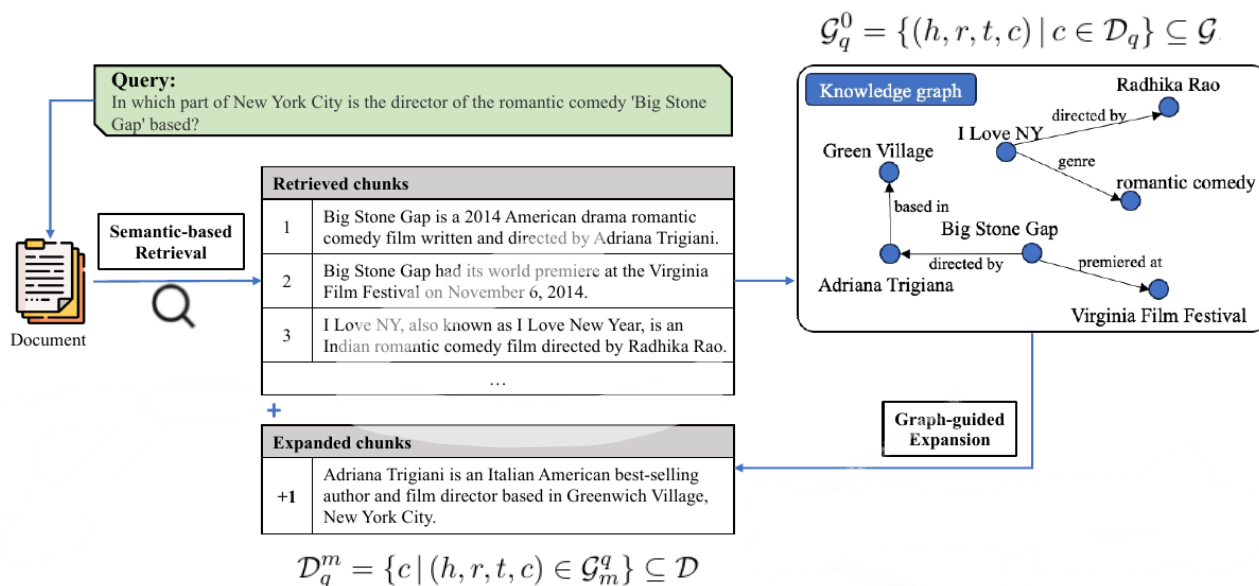
$$\mathcal{D}_q^m = \{c \mid (h, r, t, c) \in \mathcal{G}_q^m\} \subseteq \mathcal{D} \quad \mathcal{D}_q^m \text{ خوانده می‌شوند.}$$

- مزیت: بخش‌هایی را که شامل موجودیت‌ها یا سه‌تایی‌های مشابه یا مرتبط هستند گردآوری می‌کند، تنوع را افزایش داده و یک شبکه دانشی جامع ایجاد می‌کند.

۲- بازیابی قطعه‌های بهبودیافته با گراف دانش

□ فرآیند بازیابی دو مرحله‌ای:

۲. گسترش قطعه‌ها با کمک گراف دانش (Graph-guided Expansion):



۳- سازماندهی زمینه بر پایه‌ی گراف دانش

□ یک مرحله‌ی پس‌پردازش قبل از تولید پاسخ توسط مدل زبانی بزرگ.

□ انگیزه‌ها:

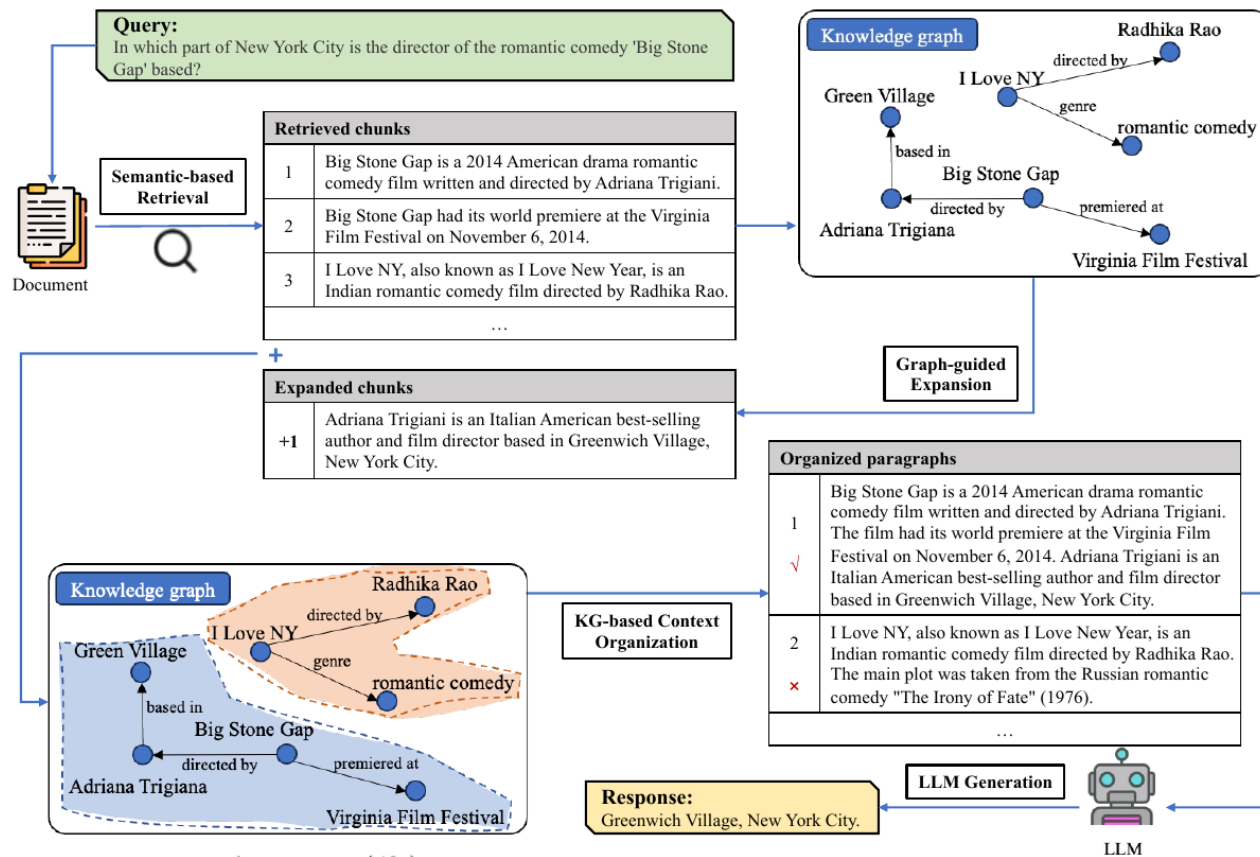
- بخش‌های گسترش‌یافته ممکن است بیش از حد زیاد باشند، که می‌تواند از طول مجاز زمینه فراتر رود یا نویز ایجاد کند.
- زمینه‌ای (context) که از نظر معنایی منسجم و خوب سازمان‌دهی شده باشد، تأثیر مثبتی بر درک و تولید مدل زبانی دارد.
- این مؤلفه هم به‌عنوان یک پالایش‌گر و هم به‌عنوان یک سازمان‌دهنده عمل می‌کند.

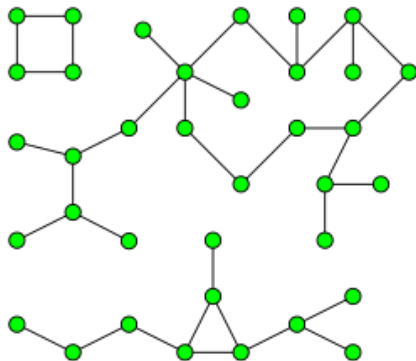
۳- سازماندهی زمینه بر پایه‌ی گراف دانش

□ عملکرد به‌عنوان یک پالایش‌گر:

- شباهت معنایی بین بخش‌های گسترش‌یافته و پرسش را محاسبه می‌کند.
- زیرگراف گسترش‌یافته ($\mathcal{G}_q^m$) را به یک گراف بدون جهت و وزن‌دار ($\mathcal{U}_q^m$) تبدیل می‌کند که در آن وزن یال‌ها میزان شباهت قطعه‌ی مربوطه با پرسش است.
$$\mathcal{U}_q^m = \{(h \leftrightarrow t, \text{rel} : r, \text{src} : c, \text{weight} : s(q, c)) \mid (h, r, t, c) \in \mathcal{G}_q^m\},$$
- مولفه‌های هم‌بند (connected components) را در $\mathcal{U}_q^m$ شناسایی می‌کند ($\mathcal{B}_i$) (هر مؤلفه هم‌بند، زیرمجموعه‌ای از موجودیت‌هاست که از نظر معنایی با هم ارتباط دارند).
- برای هر مولفه‌ی هم‌بند، یک درخت پوشای بیشینه (Maximum Spanning Tree) ایجاد می‌کند تا فقط اطلاعات پیوندی مرتبط‌تر را حفظ کرده و یال‌های تکراری یا غیرضروری را حذف کند.

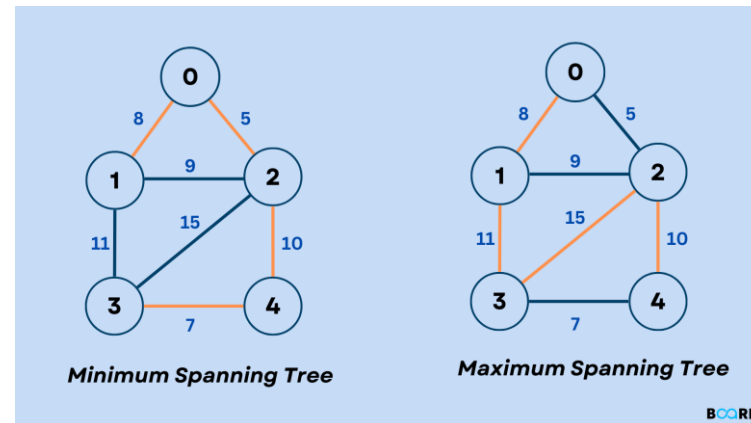
$$\mathcal{T}_i = \text{MST}(\mathcal{B}_i)$$





گرافی با ۳ مؤلفه همبند

[https://en.wikipedia.org/wiki/Component_\(graph_theory\)](https://en.wikipedia.org/wiki/Component_(graph_theory))



<https://www.boardinfinity.com/blog/prims-and-kruskal-algorithm-for/>

۳- سازماندهی زمینه بر پایه گراف دانش

□ عملکرد به عنوان یک سازمان دهنده:

□ بخش های بازیابی شده را در قالب پاراگراف هایی که مرتبط و از نظر معنایی منسجم هستند، یکپارچه می کند.

□ هر درخت پوشای بیشینه (T_i) که تولید می شود، دو نمایش دارد:

▪ **نمایش متنی:** با استفاده از پیمایش عمقی (DFS) و شروع از یال با بیشترین وزن، قطعه هایی که توسط

یال های درخت به هم متصل شده اند را به هم متصل کرده و یک پاراگراف منسجم ایجاد می کند.

▪ **نمایش سه تایی:** تمامی یال ها را در قالب سه گانه های $\langle h, r, t \rangle$ به هم متصل می کند.

□ امتیاز ارتباط بین درخت های پوشای بیشینه (MST) و پرسش کاربر، بر اساس نمایش سه تایی آن ها محاسبه می شود.

□ این کار با استفاده از یک تابع بازرتبه بندی مبتنی بر رمزگذار دوجهته (cross-encoder) انجام می گیرد.

$$R(q, T_i) = C(q, \text{conc}(T_i)),$$

□ سه تایی ها یک بازنمایی فشرده و ساخت یافته از اطلاعات کلیدی ارائه می دهند.

□ MST ها بر اساس میزان ارتباط، به صورت نزولی مرتب می شوند.

□ سپس نمایش های متنی MST های انتخاب شده (تا سقف k مورد برتر) به مدل های زبانی بزرگ داده می شوند تا پاسخ نهایی تولید شود.

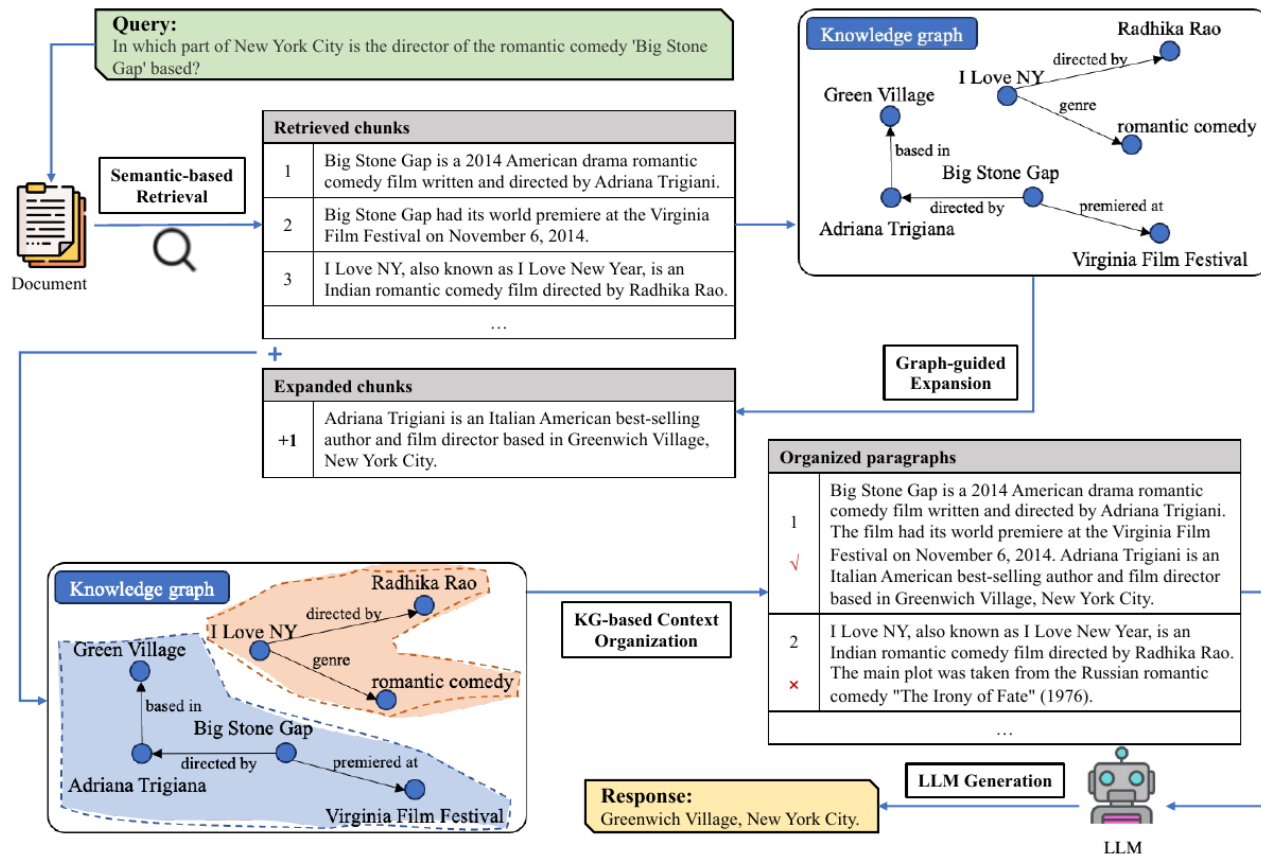


Figure 2: Workflow of the proposed KG²RAG.

۳. آزمایش‌ها

❑ **HotpotQA:** معیار پرکاربرد برای پاسخ‌دهی به پرسش‌های چند-مرحله‌ای (multi-hop).

❑ حالت Distractor: شامل ۱۰ سند پشتیبان است که برخی از آن‌ها بی‌ربط هستند.

❑ حالت Fullwiki: استخراج دانش از میان ۶۶,۵۸۱ سند ویکی‌پدیا.

❑ **نسخه‌های Shuffle-HotpotQA:** طراحی‌شده برای کاهش اتکای مدل‌های زبانی به دانش پیش‌زمینه (مانند محتوای ویکی‌پدیا در داده‌های آموزشی).

❑ در این نسخه‌ها، موجودیت‌ها به صورت تصادفی با موجودیت‌های دیگر از همان دسته جایگزین می‌شوند، و پرسش‌ها، سه‌تایی‌ها، و اسناد متناظر نیز به‌روزرسانی می‌شوند.

Paragraph A, Return to Olympus:

[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:

[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?

A: Malfunkshun

Supporting facts: 1, 2, 4, 6, 7

Figure 1: An example of the multi-hop questions in HOTPOTQA. We also highlight the supporting facts in *blue italics*, which are also part of the dataset.

HOTPOTQA: A Dataset for Diverse, Explainable Multi-hop Question Answering
<https://aclanthology.org/D18-1259/>

❑ **HotpotQA**: معیار پرکاربرد برای پاسخ‌دهی به پرسش‌های چند-مرحله‌ای (multi-hop).

- ❑ حالت Distractor: شامل ۱۰ سند پشتیبان است که برخی از آن‌ها بی‌ربط هستند.
- ❑ حالت Fullwiki: استخراج دانش از میان ۶۶,۵۸۱ سند ویکی‌پدیا.

❑ **نسخه‌های Shuffle-HotpotQA**: طراحی‌شده برای کاهش اتکای مدل‌های زبانی به دانش پیش‌زمینه (مانند محتوای ویکی‌پدیا در داده‌های آموزشی).

- ❑ در این نسخه‌ها، موجودیت‌ها به صورت تصادفی با موجودیت‌های دیگر از همان دسته جایگزین می‌شوند، و پرسش‌ها، سه‌تایی‌ها، و اسناد متناظر نیز به روزرسانی می‌شوند.

❑ **ساخت گراف دانش (KG)**: برای استخراج موجودیت‌ها و روابط از میان ۶۶,۵۸۱ سند مجموعه داده‌ی HotpotQA، از مدل **Llama-3-8b** استفاده شد.

- ❑ در نهایت، ۲۱۱,۳۵۶ سه‌گانه تولید شد که هر یک به قطعه (chunk) منبع خود پیوند داده شده‌اند.

❑ معیارهای ارزیابی

- کیفیت پاسخ: امتیاز F1، دقت و فراخوانی
- کیفیت بازیابی: امتیاز F1، دقت و فراخوانی

❑ مدل‌های استفاده شده

- مدل زبانی بزرگ: LLaMA3-8B
- مدل تعبیه‌سازی: mx-bai-embed-large
- مدل بازرتبه‌بند (reranker): bge-reranker-large

❑ روش‌های پایه برای مقایسه

- LLM-only، Semantic RAG، Hybrid RAG، GraphRAG، LightRAG

Methods	Hotpot-Dist			Hotpot-Full			Shuffle-Hotpot-Dist			Shuffle-Hotpot-Full		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
LLM-only	0.237	0.259	0.234	0.237	0.259	0.234	0.158	0.175	0.158	0.158	0.175	0.158
Semantic RAG	0.617	0.646	0.643	0.528	0.558	0.535	0.508	0.533	0.524	0.422	0.449	0.433
+ Rerank	0.652	0.685	0.665	0.587	0.613	0.603	0.532	0.560	0.546	0.447	0.476	0.456
Hybrid RAG	0.653	0.676	0.655	0.551	0.582	0.558	0.520	0.548	0.534	0.443	0.473	0.446
LightRAG	0.293	0.288	0.480	0.261	0.259	0.364	0.285	0.284	0.404	0.202	0.199	0.293
GraphRAG	0.400	0.408	0.491	0.169	0.157	0.429	0.351	0.365	0.401	0.163	0.155	0.362
KG ² RAG	0.663	0.690	0.683	0.631	0.665	0.643	0.545	0.572	0.566	0.507	0.539	0.512

Table 1: Comparisons in terms of response quality between KG²RAG and baselines.

کیفیت پاسخ

- ❑ KG2RAG عملکرد بهتری نسبت به روش‌های پایه دارد.
- ❑ تمام روش‌های RAG به‌طور قابل‌توجهی نسبت به حالت تنها LLM، بهبود نشان می‌دهند.
- ❑ مثلاً بیش از ۲۹.۱٪ افزایش در F1 روی مجموعه داده HotpotQA.
- ❑ این روش به‌ویژه در شرایط چالش‌برانگیز عملکرد قوی‌تری دارد:
- ❑ در حالت Fullwiki: KG2RAG حداقل ۸٪ افزایش امتیاز F1 نسبت به روش‌های پایه دارد، که نشان‌دهنده‌ی بازیابی بهتر بخش‌های مرتبط است.
- ❑ در مجموعه داده‌ی Shuffle-HotpotQA: KG2RAG بهبود حداقل ۲.۵٪ در حالت Distractor و ۶.۴٪ در حالت Fullwiki را در F1 نشان می‌دهد، که بیانگر کارایی آن زمانی است که مدل زبانی بیشتر به RAG متکی است تا دانش ذاتی خود.

Methods	Hotpot-Dist			Hotpot-Full			Shuffle-Hotpot-Dist			Shuffle-Hotpot-Full		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Semantic RAG	0.343	0.206	0.894	0.300	0.178	0.790	0.321	0.201	0.837	0.268	0.167	0.708
+ Rerank	0.357	0.224	0.932	0.306	0.197	0.833	0.339	0.213	0.886	0.286	0.179	0.754
Hybrid RAG	0.354	0.222	0.921	0.302	0.189	0.795	0.334	0.210	0.837	0.279	0.174	0.739
LightRAG	0.234	0.150	0.638	0.132	0.083	0.340	0.227	0.148	0.535	0.116	0.073	0.295
GraphRAG	0.255	0.167	0.594	0.180	0.113	0.470	0.210	0.138	0.482	0.199	0.126	0.510
KG ² RAG	0.436	0.301	0.908	0.310	0.203	0.838	0.405	0.279	0.840	0.305	0.193	0.790

Table 2: Comparisons in terms of retrieval quality between KG²RAG and baselines.

کیفیت بازیابی

- ❑ KG2RAG تعادل مطلوبی میان دقت و فراخوانی در بازیابی برقرار می‌کند.
- ❑ این نتایج نشان‌دهنده کارایی گسترش تکه‌ها با گراف دانش و سازمان‌دهی زمینه است.
- ❑ در حالت Distractor: فراخوانی مشابه با دیگر روش‌ها، اما دقت به‌طور چشم‌گیری بالاتر است (بیش از ۷.۹٪ در HotpotQA و ۶.۹٪ در Shuffle-HotpotQA).
- ❑ در حالت Fullwiki: بهبودهای پیوسته و قابل‌توجه در هر دو معیار دقت و فراخوانی نسبت به سایر روش‌های مبتنی بر RAG مشاهده می‌شود.
- ❑ این نتایج توانایی KG2RAG را در ارائه نتایج بازیابی با کیفیت بالا، با بهره‌گیری از گراف‌های دانش، تأیید می‌کند.

۴. بحث‌های بیشتر

	Response Quality			Retrieval Quality			
	F1	Precision	Recall	F1	Precision	Recall	#Avg.
KG ² RAG	0.663	0.690	0.683	0.436	0.301	0.908	8.11
w/o organization	0.660	0.678	0.679	0.259	0.153	0.963	16.76
w/o expansion	0.626	0.653	0.645	0.473	0.341	0.842	4.41

Table 3: Experimental results of an ablation study conducted on HotpotQA in the distractor setting.

بدون سازماندهی زمینه (فقط گسترش تکه‌ها با گراف دانش)

- ❑ کیفیت پاسخ‌ها (معیار F1) تقریباً مشابه باقی می‌ماند، اما کیفیت بازیابی اطلاعات به‌طور قابل‌توجهی کاهش می‌یابد.
- ❑ تعداد قطعه‌های بازیابی‌شده به‌طور محسوسی بیشتر است، که ممکن است شامل قطعه‌های نامرتب نیز باشد.
- ❑ این نتایج نشان می‌دهد که سازماندهی زمینه مبتنی بر گراف دانش در انتخاب و سازماندهی مؤثر قطعه‌های بازیابی‌شده نقش مهمی دارد تا اطلاعات مرتبط حفظ شوند.

	Response Quality			Retrieval Quality			#Avg.
	F1	Precision	Recall	F1	Precision	Recall	
KG ² RAG	0.663	0.690	0.683	0.436	0.301	0.908	8.11
w/o organization	0.660	0.678	0.679	0.259	0.153	0.963	16.76
w/o expansion	0.626	0.653	0.645	0.473	0.341	0.842	4.41

Table 3: Experimental results of an ablation study conducted on HotpotQA in the distractor setting.

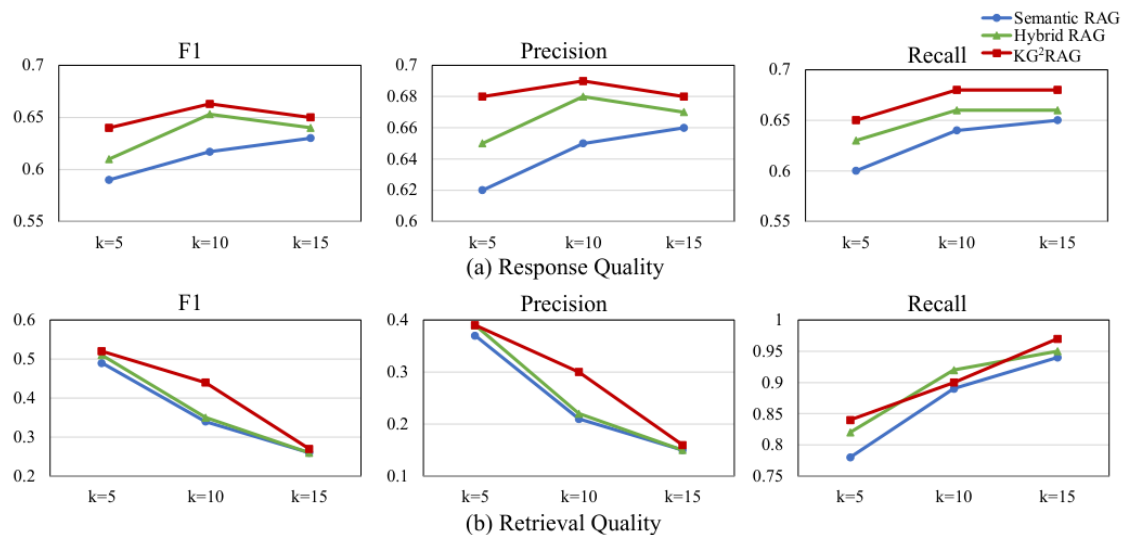
بدون گسترش (فقط سازماندهی محتوا بر اساس گراف دانش)

- ☐ دقت بازیابی و معیار F1 بالا باقی می‌ماند، در حالی که تعداد قطعه‌های بازیابی‌شده به‌طور قابل‌توجهی کمتر است.
- ☐ با این حال، کیفیت پاسخ‌ها بهتر نمی‌شود چون ممکن است قطعه‌های ضروری با روش‌های صرفاً مبتنی بر معنا (semantic-based) نادیده گرفته شوند.
- ☐ این نتایج اهمیت گسترش هدایت‌شده توسط گراف دانش را نشان می‌دهد که با استفاده از گراف دانش می‌تواند روابط سطح-حقیقت (fact-level) را درک کرده و اطلاعات کلیدی را که روش‌های معنایی از قلم می‌اندازند بازیابی کند.
- ☐ نتیجه‌گیری: هر دو مؤلفه به‌صورت مکمل با هم کار می‌کنند تا اثربخشی مدل KG2RAG را افزایش دهند.

عملکرد نسب به مقادیر مختلف k (تعداد قطعه‌های بازایی شده)

□ KG2RAG عملکرد برتری را در مقادیر مختلف k حفظ می‌کند.

□ این مدل کمترین حساسیت را نسبت به k نشان می‌دهد، که فرآیند RAG را پایدارتر می‌سازد.



عملکرد نسب به مقادیر مختلف m (تعداد hop برای گسترش)

□ تنظیم مقدار $m=1$ برای آزمایش‌ها مناسب است.

□ KG2RAG حساسیت کمی نسبت به m دارد.

	Response Quality			Retrieval Quality			#Avg.
	F1	Precision	Recall	F1	Precision	Recall	
$m = 1$	0.663	0.690	0.683	0.436	0.301	0.908	8.11
$m = 2$	0.656	0.681	0.674	0.420	0.291	0.917	8.53
$m = 3$	0.658	0.678	0.675	0.421	0.284	0.924	8.19

Table 5: Experimental results on HotpotQA in the distractor setting with varying m .

تحلیل پایداری در شرایط کاهش کیفیت گراف دانش

□ حتی زمانی که ۵٪ یا ۱۰٪ از سه‌گانه‌ها از گراف حذف شوند، مدل KG2RAG همچنان عملکرد پایداری دارد و بهتر از مدل‌های پایه عمل می‌کند.

	Response Quality			Retrieval Quality		
	F1	Precision	Recall	F1	Precision	Recall
Hybrid RAG	0.653	0.676	0.655	0.354	0.222	0.921
KG ² RAG	0.663	0.690	0.683	0.436	0.301	0.908
–5%	0.662	0.681	0.676	0.434	0.306	0.898
–10%	0.654	0.688	0.682	0.432	0.305	0.890

Table 6: Experimental results on HotpotQA in the distractor setting with triplets dropped.

تحلیل کارایی

- ساخت گراف دانش: در KG2RAG از نظر هزینه‌ی توکن، تعداد فراخوانی‌های مدل زبانی بزرگ و زمان، نسبت به LightRAG و GraphRAG کارآمدتر است.
- زمان بازیابی و تولید پاسخ: زمان موردنیاز برای بازیابی و تولید در KG2RAG کمتر از LightRAG و GraphRAG است و بسیار نزدیک به Semantic RAG می‌باشد.

Method	Avg. retrieval time	Avg. generation time
Semantic RAG	21ms	2,500ms
LightRAG	40ms	5,600ms
GraphRAG	42ms	5,500ms
KG ² RAG	25ms	2,300ms

Table 10: Comparison of average retrieval and generation time per query.

	#Input tokens	#Output tokens	#LLM calls	Extraction time
LightRAG	1,269	381	1	3s
GraphRAG	2,791	629	5	6s
KG ² RAG	561	22	1	1s

Table 9: Comparison of average LLM and time cost per chunk during KG construction.

تحلیل کارایی

- ساخت گراف دانش: در KG2RAG از نظر هزینه‌ی توکن، تعداد فراخوانی‌های مدل زبانی بزرگ و زمان، نسبت به LightRAG و GraphRAG کارآمدتر است.
- زمان بازیابی و تولید پاسخ: زمان موردنیاز برای بازیابی و تولید در KG2RAG کمتر از LightRAG و GraphRAG است و بسیار نزدیک به Semantic RAG می‌باشد.

Method	Avg. retrieval time	Avg. generation time
Semantic RAG	21ms	2,500ms
LightRAG	40ms	5,600ms
GraphRAG	42ms	5,500ms
KG ² RAG	25ms	2,300ms

Table 10: Comparison of average retrieval and generation time per query.

	#Input tokens	#Output tokens	#LLM calls	Extraction time
LightRAG	1,269	381	1	3s
GraphRAG	2,791	629	5	6s
KG ² RAG	561	22	1	1s

Table 9: Comparison of average LLM and time cost per chunk during KG construction.

۴. محدودیت، نتیجه‌گیری و کارهای آینده

محدودیت‌ها

- ❑ مدل KG2RAG عمدتاً بر بهینه‌سازی بازیابی و گسترش/سازماندهی محتوا با هدایت گراف دانش تمرکز دارد.
- ❑ این مدل ماژول‌های دیگر RAG مانند بازنویسی پرس‌وجو یا گفت‌وگوی چندمرحله‌ای (multi-turn) را بهینه نمی‌کند.

کارهای آینده

- ❑ مدل پیشنهادی KG2RAG به صورت مستقل (orthogonal) و سازگار با سایر ماژول‌های RAG طراحی شده است.
- ❑ در آینده، تمرکز بر توسعه‌ی KG2RAG به عنوان یک ابزار Plug-and-Play خواهد بود تا ادغام آن ساده‌تر شود و پژوهش در این زمینه بیشتر گسترش یابد.

nju-websoft/ KG2RAG



Knowledge Graph-Guided Retrieval Augmented
Generation, NAACL 2025



2

Contributors



2

Issues



64

Stars



9

Forks



<https://github.com/nju-websoft/KG2RAG/>

سپاس