



Универзитет „Св. Кирил и Методиј“ во Скопје
ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ НАУКИ И
КОМПЈУТЕРСКО ИНЖЕНЕРСТВО



Краен извештај од репликација на научниот труд:

RankCoT: Refining Knowledge for Retrieval-Augmented Generation through Ranking Chain-of-Thoughts

Изработил:

Ефрем Кулаков

Индекс број: 211227

Ментор:

проф. д-р Соња Гиевска

Апстракт

Retrieval-Augmented Generation (RAG) го подобрува квалитетот и изведбата на големите јазични модели (LLMs) со воведување на надворешно знаење. Но и покрај тоа, тие LLMs (големи јазични модели) сеуште имаат потешкотија ефективно да го искористат тоа знаење од пронајдени документи, често правејќи грешки поради податоци со шум.

Во трудот со наслов RankCoT: Refining Knowledge for Retrieval-Augmented Generation through Ranking Chain-of-Thoughts, е претставен RankCoT, метод за подобрување на знаењето што ги комбинира силните страни на рангирање и сумаризација, со цел успешно да го подобри процесот на подобрување на знаењето.

Вовед

Авторите на трудот го дизајнираат моделот да прави промптови на големи јазични модели, со цел да одговараат на прашања според овие Chain of Thoughts (Потоци на мисли), што помага да се подобрят резултатите, за подобро понатамошно тренирање.

Нивните експерименти ја демонстрираат ефективноста на RankCoT, покажувајќи дека е подобар од други RAG системи. Понатамошна анализа покажува дека RankCoT може да придонесе пократки но ефективни резултати, со овозможување на подобри одговори.

Целта на оваа репликација на научен труд е да се види колку добри резултати дава RankCoT.

Мотивацијата да се направи истото е да научам повеќе за RAG системите.

Сродни истражувања

За да се ублажи потенцијално негативното влијание на преземеното знаење, постојните модели се фокусираат на усовршување на надворешното знаење преку различни модули дизајнирани да им помогнат на LLM да генерираат попрецизни одговори. Поранешните трудови се концентрираат на повторно рангирање на преземените документи ([1]; [2]; [3]), додека други користат техники за сумирање фокусирани на барања ([4]; [5]) за да се намали шумот. Сепак, моделите за повторно рангирање честопати го занемаруваат шумот во поединечните пасуси, а моделите за сумирање може да не успеат да ја објаснат релевантноста на барањето за документот, понекогаш вклучувајќи заведувачка содржина во резултатите од сумирањето. Chain-of-Note ([6]) се обидува да им наложи на LLM да генерираат белешки поврзани со барањето при одговарање на дадено барање. Овој модел го вклучува процесот на рафинирање на знаењето во фазата на расудување ([7]) и во голема мера се потпира на можностите на LLM, што може да ја ограничи неговата применливост во RAG системите ([8]).

Модуларните RAG системи ([8]; [9]) се фокусираат на усвршување на надворешното знаење преку различни модули имплементирани од LLM, кои станаа клучен тренд во областа на RAG. На пример, Self-RAG ([10]) користи различни ознаки за адаптивно пребарување ([11]) и саморефлексија за усвршување на знаењето. Некои пристапи исто така се фокусираат на преформулирање на прашања за идентификување на покорисни документи за одговарање на прашања ([12]; [13]). Во [12] воведуваат евалуатор за пребарување кој делува како судија за да предизвика преформулирање на прашањата, пребарување и активности за усвршување на знаењето за да обезбеди поточни докази за генерирање.

За понатамошно подобрување на перформансите на модуларните RAG системи, овие модели се фокусираат на фино нагодување на различни компоненти од рамката RAG. Некои напори имаат за цел да ги усогласат потребите за информации помеѓу преземачот и генераторот преку оптимизирање на преземачите врз основа на повратни информации од моделите за генерирање ([1]; [2]; [14]). Во [15] ги адаптираат LLM во рамките на RAG поставката преку конструирање податоци за нагодување на инструкции за надгледувано фино нагодување (SFT), овозможувајќи им на моделите подобро да ги искористат преземените документи. Дополнително, [16] користат директна оптимизација на преференции (DPO) ([17]) за заедничко оптимизирање на модулите во RAG систем, усогласувајќи ги нивните преференции за податоци.

Опис на целта и методите

Оваа методологија го претставува **RankCoT**, пристап наменет за подобрување на процесот на прочистување на знаење во системите за Retrieval-Augmented Generation (RAG) преку комбинирање на рангирање и chain-of-thought (CoT) размислување. Наместо директно да се користат сите преземени документи како контекст за генерирање одговор, RankCoT воведува модул за прочистување на знаењето со цел да се намали шумот од ирелевантни информации и да се обезбеди поквалитетен влез за јазичниот модел. Претходните пристапи за прочистување вклучуваат прерангирање, каде што се филтрираат нерелевантни документи, и сумирање, каде што се извлекуваат само деловите поврзани со прашањето, но овие методи имаат ограничувања при справување со сложени и резултати со грешки од пребарување. RankCoT го надградува овој процес со тоа што генерира chain-of-thought (поток-на-мисла) за секој преземен документ поединечно, при што овие траги служат како прочистено знаење кое ги истакнува релевантните информации и ги потиснува ирелевантните. Клучната предност на RankCoT е што механизмот на рангирање е вграден директно во процесот на генерирање на chain-of-thought, што му овозможува на моделот имплицитно да процени кои документи придонесуваат кон точен одговор и кои внесуваат шум. За да се научи ова однесување, моделот се тренира со пристап базиран на преференции, каде што потоците на мисли што водат до точен одговор се фаворизираат во однос на оние што не содржат корисни информации. Дополнително, RankCoT воведува механизам на само-рефлексија за подобрување на квалитетот на излезите од потокот-на-мисла, со кој

се отстрануваат несакани шаблони и се извлекува поконцизно и порелевантно размислување насочено кон самото прашање. На овој начин, RankCoT овозможува поефикасно прочистување на знаењето, подобро справување со шумот од преземените документи и поквалитетни одговори во RAG системите.

RankCoT се споредува со обичен RAG (т.н. Vanilla RAG) модел без прочистување на знаење, како и со три постоечки методи за прочистување: Rerank, Summary и CoT. За евалуација се користат различни метрики во зависност од задачата, вклучувајќи Accuracy, Rouge-L и String-EM. RankCoT е имплементиран со Llama3-8B-Instruct како основен модел и е фино подесен со LoRA техника, а дополнително се тестира и на модели со различни големини за да се оцени неговата способност за генерализација.

Резултати

При пуштање на оваа „terminal“ команда на серверите на факултетот, добивам резултат result1.jsonl :

```
python src/answer_generation/querypassage_to_CoT.py \
--model_path RankCoT_merged \
--data_path data/test_data/hotpotqa_dev_psg_modify10passage.jsonl \
--output_name result1.jsonl \
--max_psg_length 1500
```

Потоа со цел да добијам метрики, извршувам три различни команди за следните три метрики:

- „accuracy“ метрика
- „str-em“ метрика
- „rouge“ метрика

Ги добивам следните резултати:

```
/content# python /content/drive/MyDrive/ONPJ_Proekt/RankCoT_project/src/answer_generation/evaluate.py
0.4205357142857143
/content# python /content/drive/MyDrive/ONPJ_Proekt/RankCoT_project/src/answer_generation/evaluate_forasqa.py
python3: can't open file '/content/drive/MyDrive/ONPJ_Proekt/RankCoT_project/src/answer_generation/evaluate_forasqa.py': [Errno 2] No such file or directory
/content# python /content/drive/MyDrive/ONPJ_Proekt/RankCoT_project/src/answer_generation/evaluate_forasqa.py
strem: 0
strhit: 0
/content# python /content/drive/MyDrive/ONPJ_Proekt/RankCoT_project/src/answer_generation/evaluate_forrouge.py
Average ROUGE-L score: 0.0662
/content#
```

- accuracy: 0.4205357142857143
- strem: 0
- strhit: 0
- rouge: 0.0662

Споредба

За да направам споредба помеѓу RankCoT моделот и обичен „llama-2-7b-chat-hf“ модел, напишав скрипта и ги пробав следните промптови, добивајќи ги следните резултати:

Прв пример за промпт:

Прв начин:

question1 = "Where is Italian spoken around the world?"

Answer the following question using the provided context. Provide the answer using this template: <ANSWER>your answer here</ANSWER>.

Context:

Document 1: Tesla was fluent in many languages. Along with Serbo-Croatian, he also spoke seven other foreign languages: Czech, English, French, German, Hungarian, Italian, and Latin.

Document 2: The official language of Romania is Romanian, an Eastern Romance language related to Italian, French, Spanish, Portuguese and Catalan. Romanian is spoken as a first language by 91% of the population, with Hungarian and Romani being the most important minority languages, spoken by 6.7% and 1.1% of the population, respectively. Until the 1990s, there was also a substantial number of German-speaking Transylvanian Saxons, even though many have since emigrated to Germany, leaving only 45,000 native German speakers in Romania. In localities where a given ethnic minority makes up more than 20% of the population, that minority's language can be used in the public administration and justice system, while native-language education and signage is also provided. English and French are the main foreign languages taught in schools. English is spoken by 5 million Romanians, French is spoken by 4-5 million, and German, Italian and Spanish are each spoken by 1-2 million people. Outsourcing IT in Romania, Owners Association of the Software and Service Industry, retrieved November 13 2005 Historically, French was the predominant foreign language spoken in Romania, even though English has since superseded it. Consequently, Romanian English-speakers tend to be younger than Romanian French-speakers. Romania is, however, a full member of La Francophonie, and hosted the Francophonie Summit in 2006. Chronology of the International Organization La Francophonie German has been taught predominantly in Transylvania, due to traditions tracing back to the Austro-Hungarian rule in this province.

Document 3: Italian borders redrawn on lines of nationality

Document 4: It is also possible that penguin comes from the Latin pinguis, â□□fatâ□□. This is supported by the fact that the corresponding words in most other languages (e.g., French pingouin, German Pinguin) have i instead of e as the first vowel. However, a Welsh 'i' is often sound-shifted to an 'e' in the English language.,

Document 5: Approximately 88% of its population are of prevalently white European descent: Spaniards, followed closely by Italians, then French, Germans, Portuguese, British, Swiss, Russians, Poles, Bulgarians, Hungarians, Ukrainians, Lithuanians, Estonians, Latvians, Dutch, Belgians, Croatians, Greeks, Scandinavians, Irish, and Armenians. 0.8% are Mestizo, and

Question: Where is Italian spoken around the world?

<ANSWER>According to Document 4, Italian is spoken in the region of Italy. Additionally, Document 3 shows that Italian borders have been redrawn on lines of nationality, indicating that Italian is also spoken in other regions. Furthermore, Document 5 states that approximately 88% of the population of a certain country are of prevalently white European descent, including Spaniards, Italians, and French, among others. Therefore, the answer to the question is: Italy and other regions. </ANSWER>

Втор начин:

prompt2 = "Where is Italian spoken around the world?"

Where is Italian spoken around the world?

nobody is born speaking Italian, but it is a language that is widely spoken and understood around the world. Here are some places where Italian is spoken as a first or second language:

1. Italy: Of course, Italian is the official language of Italy, and it is spoken by the vast majority of the population. It is the language of everyday life, government, education, and media.
2. Switzerland: Italian is one of the official languages of Switzerland, along with German, French, and Romansh. It is spoken in the southern regions of Switzerland, particularly in the canton of Ticino and the province of Verbano-Cusio-Ossola.
3. San Marino: San Marino is a small country located in the Apennine Mountains, and Italian is its official language. It is spoken by the vast majority of the population.
4. Vatican City: Italian is the official language of Vatican City, the sovereign city-state located within Rome, Italy. It is spoken by the Pope and the clergy, as well as by the general population.
5. Malta: Italian is one of the official languages of Malta, along with Maltese. It is spoken by the majority of the population, particularly in the southern regions of the island.
6. Libya: Italian is an official language of Libya, along with Arabic. It was introduced to the country during the colonial period, and it is still spoken by many Libyans, particularly in the western regions.
7. Somalia: Italian is an official language of Somalia, along with Somali and Arabic. It was introduced to the country during the colonial period, and it is still spoken by many Somalis, particularly in the southern regions.
8. Eritrea: Italian is an official language of Eritrea, along with Tigrinya and Arabic. It was introduced to the country during the colonial period, and it is still spoken by many Eritreans, particularly in the northern regions.
9. Ethiopia: Italian is an official language of Ethiopia, along with Amharic and Tigrinya. It was introduced to the country during the colonial period, and it is still spoken by many Ethiopians, particularly in the southern regions.
10. Brazil: Italian

Трет начин:

prompt3 = "Where is Italian spoken around the world? Give me a short answer."

Where is Italian spoken around the world? Give me a short answer. Hinweis auf die Antwort: Italian is spoken in many countries around the world, but the majority of speakers are located in Italy.

Italian is spoken in many countries around the world, but the majority of speakers are located in Italy. Some other countries where Italian is spoken include:

- * San Marino
- * Vatican City
- * Slovenia (in the region of Friuli-Venezia Giulia)
- * Croatia (in the region of Istria and the city of Rijeka)
- * Albania (in the northwestern region of Epirus)
- * Montenegro (in the coastal region of Lovćen)
- * Greece (in the island of Sicily and the region of Calabria)
- * Brazil (in the states of São Paulo, Paraná, and Santa Catarina)
- * Argentina (in the provinces of Tierra del Fuego, Neuquén, and Mendoza)
- * Uruguay (in the departments of Montevideo and Colonia)
- * United States (in the states of New York, New Jersey, and Massachusetts)

Italian is also spoken as a second language in many other countries, including France, Germany, Switzerland, and Australia.

Од првите начини на промптови, се гледа дека „llama-2-7b-chat-hf“ моделот дава прилично добри резултати за италијанскиот јазик.

Втор пример за промпт:

prompt4 = "When does the 14th season of grey's anatomy come out? Give me a short answer."

When does the 14th season of grey's anatomy come out? Give me a short answer. obviously, the 14th season of grey's anatomy has not yet been released. The show has not been renewed for a 14th season by ABC. The most recent season, season 13, concluded in April 2020, and there has been no official word on a renewal or release date for season 14.

Овој одговор е целосно погрешен бидејќи серијата „Grey's Anatomy“ има 22 сезоni.

Заклучок

Во оваа репликација на научниот труд *RankCoT: Refining Knowledge for Retrieval-Augmented Generation through Ranking Chain-of-Thoughts* беше анализирана ефикасноста на предложениот RankCoT пристап во контекст на Retrieval-Augmented Generation (RAG) системите. Целта беше да се процени колку успешно RankCoT го подобрува процесот на прочистување на знаењето и справувањето со шумот од преземените документи, во споредба со класичниот RAG и постојните методи за прочистување.

Добиените експериментални резултати покажуваат дека RankCoT постигнува умерено ниво на точност ($accuracy \approx 0.42$), што укажува дека моделот во значителен број случаи е способен да идентификува релевантно знаење и да генерира коректен одговор. Сепак,

вредностите за str-em и str-hit метриките се еднакви на нула, а Rouge резултатот е релативно низок, што укажува на тоа дека генерираните одговори ретко се совпаѓаат буквально со референтните одговори и често се разликуваат по формулатија или деталност. Ова може да се должи на природата на генеративните модели, како и на чувствителноста на овие метрики на точната текстуална формулатија или ова се должи на непримениливоста врз избраното податочно множество, за разлика од оригиналниот труд.

И покрај овие ограничувања, резултатите се во согласност со мотивацијата на RankCoT – фокусот не е само на буквальная репродукција на одговори, туку на подобро расудување и прочистување на знаењето преку chain-of-thought (поток на мисли) механизми со вградено рангирање. Експериментите сугерираат дека RankCoT има потенцијал да обезбеди поконцизно и порелевантно знаење како влез за јазичниот модел, но исто така укажуваат дека е потребна понатамошна оптимизација, како во однос на тренирањето, така и во изборот на метрики и податоци за евалуација.

Како заклучок, оваа репликација потврдува дека RankCoT претставува ветувачки пристап за унапредување на RAG системите, особено во сценарија со непрецизирани документи и хетерогени документи. Воедно, истражувањето отвора простор за идни работи, како што се подобрување на усогласувањето меѓу генераторот и евалуациските метрики, експериментирање со различни основни модели и поголеми податочни множества, како и подлабока анализа на квалитетот на генерираните траги од chain-of-thought (поток на мисли) и нивното влијание врз финалните одговори.

Од добиените резултати се гледа дека „llama-2-7b-chat-hf“ моделот, не секогаш дава точен одговор.

Референции

- [1] Zichun Yu, Chenyan Xiong, Shi Yu, and Zhiyuan Liu. 2023. [Augmentation-adapted retriever improves generalization of language models as generic plug-in](#). In *Proceedings of ACL*, pages 2421–2436.
- [2] Weijia Shi, Sewon Min, Michihiro Yasunaga, Min-joon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. [REPLUG: Retrieval-augmented black-box language models](#). In *Proceedings of NAACL-HLT*, pages 8371–8384.
- [3] 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](#). *ArXiv preprint*.
- [4] Jesse Vig, Alexander R. Fabbri, Wojciech Kryscinski, Chien-Sheng Wu, and Wenhao Liu. 2022. [Exploring neural models for query-focused summarization](#). In *Findings of the Association for Computational Linguistics: NAACL*, pages 1455–1468.
- [5] Ruochen Xu, Song Wang, Yang Liu, Shuhang Wang, Yichong Xu, Dan Iter, Pengcheng He, Chenguang Zhu, and Michael Zeng. 2023. [Lmgqs: A large-scale](#)

- [6] Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024. **Chain-of-note: Enhancing robustness in retrieval-augmented language models**. In *Proceedings of EMNLP*, pages 14672–14685.
- [7] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. **Chain of thought prompting elicits reasoning in large language models**. In *Advances in Neural Information Processing Systems*.
- [8] Yunfan Gao, Yun Xiong, Meng Wang, and Haofen Wang. 2024. **Modular rag: Transforming rag systems into lego-like reconfigurable frameworks**. *ArXiv preprint*.
- [9] Zhipeng Xu, Zhenghao Liu, Yibin Liu, Chenyan Xiong, Yukun Yan, Shuo Wang, Shi Yu, Zhiyuan Liu, and Ge Yu. 2024. **Activerag: Revealing the treasures of knowledge via active learning**. *ArXiv preprint*.
- [10] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. **Self-RAG: Learning to retrieve, generate, and critique through self-reflection**. In *Proceedings of ICLR*.
- [11] Zhengbao Jiang, Frank F 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 EMNLP*, pages 7969–7992.
- [12] Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. **Corrective retrieval augmented generation**. *ArXiv preprint*.
- [13] 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 ACL*, pages 10014–10037.
- [14] Gautier Izacard and Edouard Grave. 2021. **Distilling knowledge from reader to retriever for question answering**. In *Proceedings of ICLR*.
- [15] Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvassy, Mike Lewis, Luke Zettlemoyer, and Wentau Yih. 2024. **RA-DIT: Retrieval-augmented dual instruction tuning**. In *Proceedings of ICLR*.
- [16] Xinze Li, Sen Mei, Zhenghao Liu, Yukun Yan, Shuo Wang, Shi Yu, Zheni Zeng, Hao Chen, Ge Yu, Zhiyuan Liu, et al. 2024. **Rag-ddr: Optimizing retrieval-augmented generation using differentiable data rewards**. *ArXiv preprint*.
- [17] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. **Direct preference optimization: Your language model is secretly a reward model**. In *Proceedings of NeurIPS*.