موضوع:

سیستم های پرسش و پاسخ با استفاده از مدل های زبانی بزرگ

Question answering with LLm

استاد:

دکتر بهشید بهکمال

تهیه و تنظیم:

فاطمه باغخانی

بهمن 1402

فهرست

1 مقدمه

1-1 مدل های بزرگ زبانی[1] :

ظهور مدل‌های زبان بزرگ (LLM) را می‌توان به پیشرفت در روش‌های یادگیری عمیق (DL)، در دسترس بودن منابع محاسباتی عظیم، و در دسترس بودن مقادیر زیادی از داده‌های آموزشی نسبت داد. این مدل‌ها که اغلب بر روی مجموعه‌های گسترده از وب از قبل آموزش داده شده‌اند، توانایی یادگیری الگوهای پیچیده، تفاوت‌های زبانی و روابط معنایی را دارند. تنظیم دقیق این مدل‌ها در کارهای پایین دستی خاص، نتایج امیدوارکننده‌ای را نشان داده است و به عملکرد پیشرفته‌ای در معیارهای مختلف دست یافتند .و اینکه از دیرباز هدف دانشمندان دستیابی به خواندن، نوشتن و ارتباط انسان گونه بوده اند و این مساله از دیرباز یک چالش تحقیقاتی طولانی مدت بوده است.[1]

[1] Summary of ChatGPT-Related research and perspective towards the future of large language models

مدل‌سازی زبان، یک وظیفه که در آن یک چارچوب محاسباتی برای درک و مدل‌سازی نحوه زبان، معنا و تسهیل تولید مصنوعی زبان ابداع شده است، به‌عنوان چالش تحقیقاتی مهمی در پردازش زبان طبیعی در چند دهه اخیر ظاهر شده است. تکامل این حوزه شاهد اولین وعده روش‌های مبتنی بر قوانین بود که دیکشنری‌ها را در بر می‌گرفت. پس از آن، روش‌های آماری مانند مدلسازی n-gramجذب کرد. با این حال، ظهور مدل‌های زبان احتمالی عصبی بیشترین تغییرات را به‌همراه داشت. در این زمینه، با داده شدن یک دنباله متنی x، احتمال آن توسط احتمال شرطی تولید هر نشانه xi با در نظر گرفتن نشانه‌های قبلی تعیین می‌شود؛ یک فرآیند که به طور موثر به یک ضرب احتمالات با استفاده از قانون زنجیره‌ای تجزیه می‌شود.

ظهور معماری ترنسفورمر نقطه عطفی در توسعه مدل‌های زبان احتمالی عصبی را نشان داد، که پردازش کارآمد داده‌های متوالی و امکان پردازش موازی را فراهم کرد و در عین حال وابستگی‌های دوربرد متنی را دربرمی‌گرفت. این نوآوری راه را برای توسعه مدل‌هایی همچون GPT-3، GPT-4 و PaLM با اندازه پارامترهای

[1] llm

گسترده و ظرفیت یادگیری بی‌نظیر آنها باز کرد. این مدل‌ها قادر به تولید متن با پیوستگی شبیه به انسان و درک زمینه‌های پیچیده هستند. در سال‌های بعدی، پیشرفت‌های در الگوریتم‌های زبان بزرگ توسط بهبودهای دقیق مشخص شده‌اند. تکنیک‌هایی همچون تنظیم دستورالعمل و یادگیری تقویتی از بازخورد انسانی (RLHF) توانایی‌های مکالمه و استدلال مدل‌ها را در محیط‌های مختلف افزایش داده‌اند. این روش‌ها عملکرد خود را به عنوان مدل‌های جهانی بهینه‌سازی کرده و به تولید توانایی‌های نوظهور کمک کرده‌اند.[2][1]

[1] OpenAI. 2023b. Gpt-4 technical report

[2]Garima Agrawal, Tharindu Kumarage, Zeyad Alghami, Huan Liu. "Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey." arXiv preprint .arXiv:2311.07914, November 14, 2023.

2-1 تاریخچه

درواقع مدل سازی زبان (LM) یک رویکرد حیاتی برای تقویت هوش زبانی ماشین ها است وبه طوری که از شکل ۱ قابل مشاهده است، تحقیقات مربوط به مدل زبان (LM) توجه گسترده‌ای را به خود جلب کرده و چهار مرحله توسعه مهم را طی کرده است، به شرح زیر:

اولین مرحله در توسعه lmها statistical language models مانند n-gram-models بوده است این مدل‌ها احتمال وقوع کلمه بعدی در یک دنباله را بر اساس فراوانی تکرار n-گرام‌های قبلی کلمات تخمین می‌زنند.

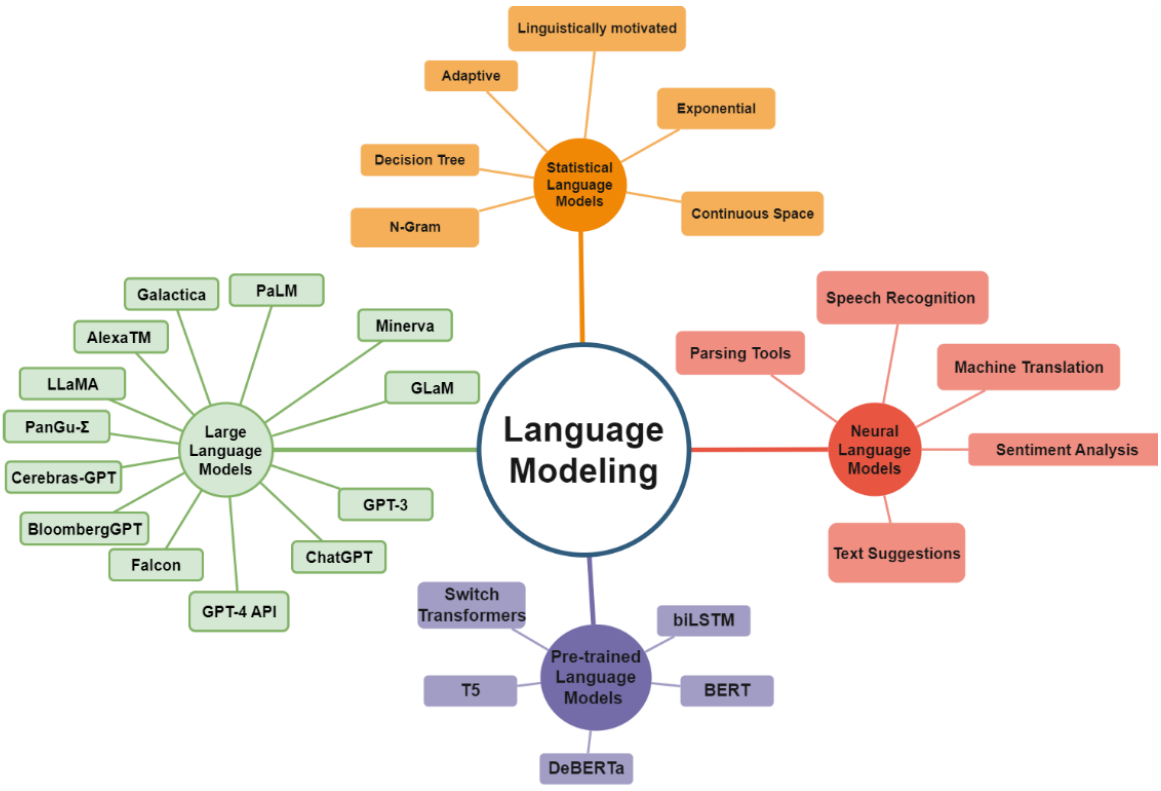
دومین مرحله توسعه lmها مرحله دوم توسعه مدل زبان (LM) شامل معرفی مدل‌های زبانی مبتنی بر شبکه عصبی بود که به آن‌ها مدل‌های زبانی عصبی (NLMs) نیز گفته می‌شود. این رویکرد که به عنوان مدل‌سازی زبانی عصبی نیز شناخته می‌شود، از شبکه‌های عصبی برای پیش‌بینی توزیع احتمال کلمه بعدی در یک دنباله با توجه به کلمات قبلی در دنباله استفاده می‌کند. (RNNs)

مرحله سوم توسعه مدل زبان (LM) شامل ایجاد تعبیرهای کلمات متنی است که هدف آن ضبط معنا و سازوکار کلمات در جمله یا متن است. این تعبیرها به عنوان مدل‌های زبان پیش‌آموزش دیده (PLMs) نامگذاری می‌شوند.این مدل‌ها از شبکه‌های عصبی استفاده می‌کنند تا یک نمایش برداری(embedding) از کلمات را یاد بگیرند که با در نظر گرفتن متناسب با متن کلمه ظاهر می‌شود.

مرحله چهارم توسعه مدل زبان (LM) شامل ایجاد مدل‌های زبانی پیش‌آموزش دیده در مقیاس بزرگ است که به آن‌ها مدل‌های زبانی بزرگ (LLMs) نیز گفته می‌شود.که قادر به انجام از تسک های مختلف پردازش زبان

طبیعی (NLP) با عملکرد بسیار عالی هستند. مانند GPT3 که روی تعداد زیادی داده آموزش دیده اند و میتوانند برای یک تسک خاص fine-tuned شوند[2][1]

[1] M. Fraiwan and N. Khasawneh, “A review of chatgpt applications in education, marketing, software engineering, and healthcare: Benefits, drawbacks, and research directions,” arXiv preprint arXiv:2305.00237, 2023.
[2] Summary of ChatGPT-Related research and perspective towards the future of large language models



شکل 1-دسته بندی انواع مدل های زبانی

2 بیان مساله

به طور کلی امروزه مدل های زبانی با توجه به دامنه گسترده‌ی کاربردهای مدل‌های زبانی بزرگ(LLMs) در حوزه‌های پزشکی، آموزش، مالی و پرسش و پاسخ کاربردهای متنوعی را دارد .در حوزه پزشکی، مدل‌های زبانی بزرگ (LLMs) مانند ChatGPT می‌توانند در تشخیص بیماری‌ها، پیش‌بینی نتایج آزمایش‌ها، تحلیل اسناد پزشکی، و ارائه راهنمایی‌های درمانی مورد استفاده قرار بگیرند.

در حوزه آموزش، ChatGPT می‌تواند به دانش آموزان در تمرین و تکالیف و یادگیری شان کمک کند. همچنین می‌تواند به معلمان در ارزیابی خودکار پاسخ‌ها و کارنامه‌های دانش‌آموزان کمک کند و بار کاری آن‌ها را کاهش دهد.[2][1]

چت بات‌ها در برنامه‌های خدمات مشتریان و در حوزه آموزش مورد استفاده قرار می‌گیرند، که می‌توانند به صورت خودکار و فوری به سوالات مشتریان پاسخ دهند و در رفع مشکلات و ارائه راهنمایی به آن‌ها کمک کنند.

ChatGPT همین طور به عنوان یک سیستم پرسش و پاسخ در کاربردهای مختلف استفاده می‌شود مانند حوزه آموزش قابلیت پاسخگویی به سوالات و بررسی امتحانات را فراهم می‌کند. همچنین می‌تواند در موضوعات علمی و مفهومی مختلف مانند ریاضیات، فیزیک، و فلسفه به کار رود، با این حال، عملکرد آن ممکن است در برخی زمینه‌ها نسبت به دانشجویان فارغ‌التحصیل متفاوت باشد. [2]

ChatGPT قادر است به یادگیری، مقایسه و تأیید پاسخ‌ها در موضوعات مختلف علمی مانند فیزیک، ریاضیات و شیمی، و یا موضوعات مفهومی مانند فلسفه و دین باشد. به طور خاص، آنها بیان می‌کنند که توانایی‌های ریاضی چت‌جی‌پی‌تی کمتر از دانشجویان فارغ‌التحصیل ریاضی معمولی است. با این حال، عملکرد چت‌جی‌پی‌تی می‌تواند به طور قابل توجهی بسته به نیازهای شغلی خاص متفاوت باشد.[2]

در حل مسائل ریاضی، چت‌جی‌پی‌تی به طور کلی مسائل را درک می‌کند اما قادر به ارائه پاسخ صحیح نمی‌باشد. عملکرد آن با استفاده از مجموعه‌داده‌های مختلف، از جمله مجموعه‌داده Grad Text که در مسائل ساده تئوری مجموعه و منطق عملکرد بهتری داشت، مورد ارزیابی قرار گرفت. با این حال، در مجموعه‌داده‌هایی مانند مسائل حل‌مسابقه‌های المپیادی و مجموعه‌داده‌های Holes-in-Proofs، چت‌جی‌پی‌تی نمرات پایین‌تری دریافت کرد که این نکته نشان‌دهنده محدودیت‌های آن در حل مسائل ریاضی پیچیده است.[3][2]
مطالعات انجام شده در حوزه پزشکی نشان می‌دهد که چت‌جی‌پی‌تی قادر است به سوالات پزشکی بیماران پاسخ دهد و در تشخیص بیماری‌ها به پزشکان کمک کند.[4]

[1] J. S. () and W. Y. (), “Unlocking the power of chatgpt: A framework for applying generative ai in education,” ECNU Review of Education, vol. 0, no. 0, p. 20965311231168423, 0

[2] M. Fraiwan and N. Khasawneh, “A review of chatgpt applications in education, marketing, software engineering, and healthcare:Benefits,drawbacks, and research directions,” arXivpreprint arXiv:2305.00237,
2023.

[3] S. Frieder, L. Pinchetti, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, A. Chevalier, and J. Berner, “Mathematical capabilities of chatgpt,” arXiv preprint arXiv:2301.13867 , year=2023.

[4] mbzuai oryx, “Xraygpt: Chest radiographs summarization using med-ical vision-language models,” 2023.

1-2 چالش های مساله

- Outdated knowledge

مدل‌های فعلی بر روی داده‌های تاریخی (تا سال 2021) آموزش داده شده‌اند و به همین دلیل توانایی درک زمان واقعی از رویدادهای کنونی را ندارند. این یک نگرانی جدی در دوران انفجار اطلاعات امروزی است، زیرا قابلیت اعتماد به پایگاه دانش‌های قبلی به تدریج کاهش می‌یابد و این موضوع ممکن است منجر به پاسخ‌های نادرست، به ویژه در زمینه‌هایی که به سرعت در حال تحول هستند مانند قانون و فناوری، شود. علاوه بر این، این مدل‌ها قادر به بررسی صحت ادعاها نیستند در حالی که داده‌های آموزشی شامل محتواهایی از منابع مختلف است که برخی از آنها ممکن است غیرقابل اعتماد باشند و این موضوع ممکن است منجر به پاسخ‌های به نظر معقول اما بی‌معنی شود.

- Insufficient understanding

هنگام پرداختن به سؤالات مبهم یا پیچیده‌ی متنی، مدل‌ها ممکن است با چالش‌های درکی روبرو شوند. علاوه بر این، در برخی حوزه‌های تخصصی، وجود فراوانی از اختصارات منحصر به فرد، چالش‌های درک مدل‌ها را تشدید می‌کند و باعث پاسخ‌های نادرست و بی‌معنی می‌شود.[2][1]

- Energy consumption

در طول مراحل آموزش و استنتاج، این مدل‌های بزرگ نیاز به منابع محاسباتی و منابع برق قابل توجهی دارند که منجر به مصرف انرژی بالا و انتشار گازهای گلخانه‌ای قابل توجه می‌شود. این موضوع باعث محدود شدن استقرار و کاربردهای عملی این مدل‌ها می‌شود.[2][1]

- Malicious usage

اگرچه OpenAI مجموعه‌ای از محدودیت‌ها را برای کاهش سمیت مدل اجرا کرده است، اما مواردی از استفاده‌کنندگان وجود دارد که با استفاده از دستوراتی با دقت طراحی شده، تلاش می‌کنند تا از این محدودیت‌ها خارج شوند و مدل را مجبور به تولید محتوای ناسالم یا حتی استفاده غیرقانونی تجاری کنند.[2][1]

- Bias and discrimination

به دلیل تأثیر داده‌های پیش‌آموزش، مدل‌ها در زمینه‌های سیاسی، ایدئولوژیکی و سایر حوزه‌ها تعصباتی نشان می‌دهند. به همین دلیل، استفاده از مدل‌های زبانی در حوزه‌های عمومی، مانند آموزش و تبلیغات، باید با احتیاط بسیار بیشتری انجام شود. این موضوع به این معنی است که نیاز است با دقت و هوشمندی فراوان به استفاده از این مدل‌ها در این حوزه‌ها نگرانی شود و تأثیرات تعصباتی آنها در نظر گرفته شود.[2][1]

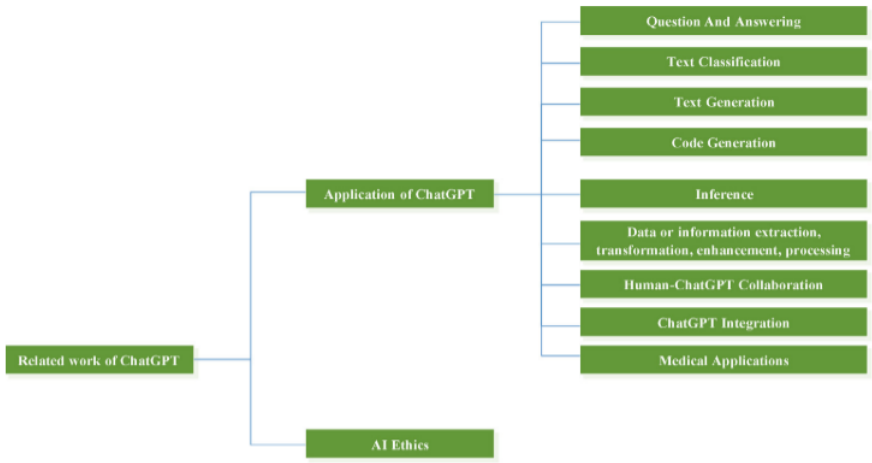
- Privacy and data security

همزمان با افزایش تعداد کاربران، حفاظت از حریم خصوصی کاربران و امنیت داده‌ها به مراتب مهم‌تر می‌شود. در واقع، در اوایل آوریل، به دلیل نگرانی‌های حریم خصوصی، ChatGPT در ایتالیا ممنوع شد. این موضوع به ویژه مهم است زیرا مدل‌ها در طول تعاملات، اطلاعات شخصی و ترجیحات شخصی را جمع‌آوری می‌کنند و در آینده، مدل‌های چندحالته مانند GPT-4 ممکن است از کاربران خواسته شود تا عکس‌های خصوصی خود را بارگذاری کنند.[2][1]

[1] M. Fraiwan and N. Khasawneh, “A review of chatgpt applications in education, marketing, software engineering, and healthcare: Benefits, drawbacks, and research directions,” arXiv preprint arXiv:2305.00237, 2023.

[2] Summary of ChatGPT-Related research and perspective towards the future of large language models

حال با توجه به اینکه بیشتر کاربردهایی که این مدل های زبانی در زمینه پرسش و پاسخ است. طراحی یک سیستم پرسش و پاسخ یکپارچه و دقیق مبتنی بر llm که در حوزه های مختلف مانند پزشکی ، آموزش و.. کاربردی است حائز اهمیت است .شکل 2 نمونه هایی از کاربردهای یکی از مدل های زبانی بزرگ مانند chatgpt بیان کرده است.



شکل2:کاربردهای چت جی پی تی

3 پیشینه تحقیق

3-1 کارهای پیشین

مدل‌های زبانی مدرن امروزی مستعد تولید توهم[۲] هستند که عمدتاً ناشی از خلأهای دانشی درون مدل‌هاست. برای رفع این محدودیت بنیادی، محققان از راهبردهای متنوعی برای تقویت مدل‌های زبانی از طریق افزودن دانش خارجی استفاده می‌کنند تا توهم‌زایی را کاهش دهند و دقت استدلال را افزایش دهند. در میان این راهبردها، بهره‌گیری از گراف‌های دانش[۳] به عنوان منبع اطلاعات خارجی نتایج امیدوارکننده‌ای نشان داده است.

روش ها برای تلفیق گراف‌های دانش با LLM ها را می‌توان به ۳ دسته اصلی تقسیم کرد:

1. استنتاج آگاه از دانش[۴]: استفاده از گراف دانش برای بهبود فرایند استنتاج و استدلال مدل. این به مدل کمک می‌کند پیش‌بینی‌های آگاهانه‌تری داشته باشد.[1]

[2] hallucination
[3] Knowledge graphs
[4] KnowledgeAware Inference

2. یادگیری آگاه از دانش[5]: بهینه‌سازی فرایند آموزش مدل با استفاده از سیگنال‌هایی از گراف دانش. این به مدل کمک می‌کند موثرتر یاد بگیرد.[1]

3. اعتبارسنجی آگاه از دانش[6]: استفاده از گراف دانش برای اعتبارسنجی خروجی‌های مدل. این به شناسایی خطاها و ارزیابی قابلیت اطمینان کمک می‌کند.[1]

در مجموع، گراف‌های دانش راه مفیدی برای تزریق دانش ساختارمند دنیای واقعی به مدل‌های زبانی بزرگ هستند. این مدل‌ها را دقیق‌تر، منطقی‌تر و قادر به تولید خروجی‌های باکیفیت‌تر می‌کند. راهبردهای اصلی شامل استفاده از دانش برای هدایت استنتاج، بهینه‌سازی یادگیری و اعتبارسنجی پیش‌بینی‌ها است.

با توجه به توانایی گراف‌های دانش در نمایش روابط پیچیده بین موجودیت‌ها، کاربردهای متنوعی در حوزه‌های گوناگون دارند (Fensel و همکاران)[1] .آن‌ها در جستجوی معنایی برای افزایش درک معنایی موتورهای جستجو (Singhal)[1] ، مدیریت دانش سازمانی (Deng و همکاران)،[1] بهینه‌سازی زنجیره تأمین (Deng و همکاران)[1] ، ادغام داده‌ها برای تحلیل جامع، تشخیص تقلب مالی (Mao و همکاران)،[1] سیستم‌های توصیه‌گر (Guo و همکاران)[1] ، آموزش (Agrawal و همکاران)[1]و سیستم‌های پاسخگویی به پرسش با استفاده از چت‌بات‌ها و دستیاران مجازی استفاده می‌شوند.[2][3]

[1] Garima Agrawal, Tharindu Kumarage, Zeyad Alghami, Huan Liu. "Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey." arXiv preprint arXiv:2311.07914, November 14, 2023.

[2]Agarwal, A., Gawade, S., Channabasavarajendra, S., & Bhattacharyya, P. (2023). There is No Big Brother or Small Brother: Knowledge Infusion in Language Models for Link Prediction and Question Answering. arXiv preprint arXiv:2301.04013.

[3] Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023.Knowledge-augmented language model prompting for zero-shot knowledge graph question answering.

arXiv preprint arXiv:2306.04136.

[5] Knowledge-Aware Learning
[6] Knowledge-Aware Validation

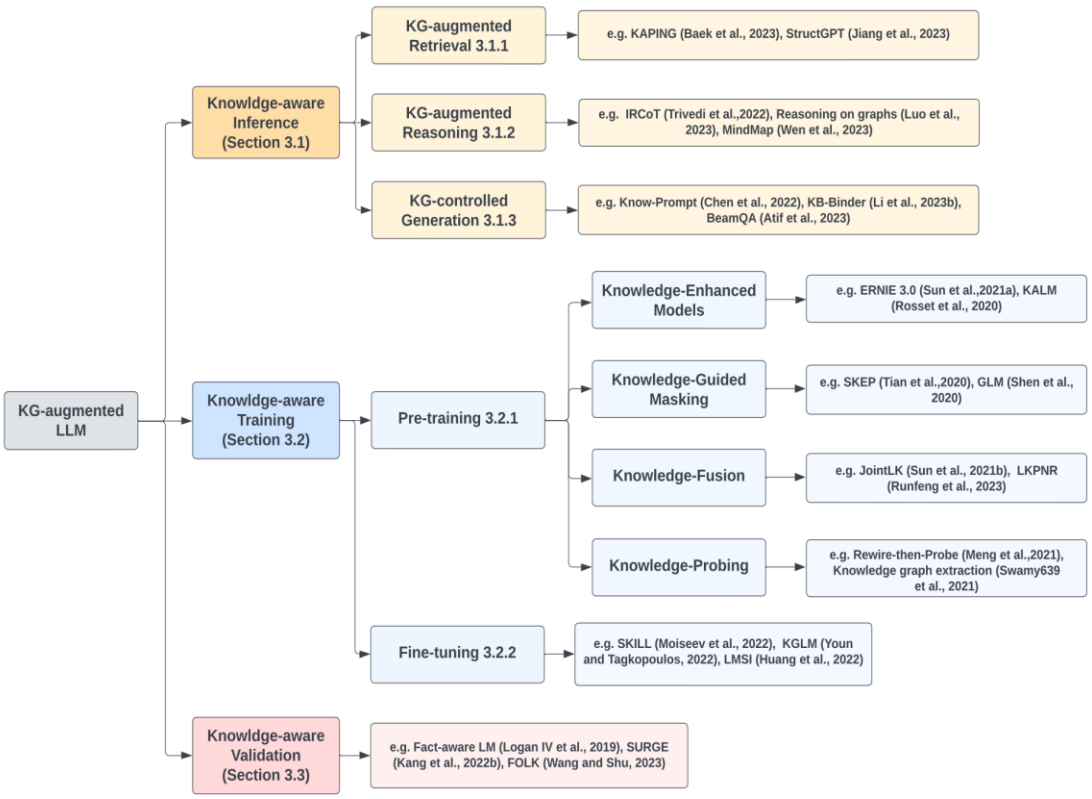
مدل‌های زبانی بزرگ عمدتاً دارای سه نقطه ضعف هستند :

1.عدم درک سؤال به دلیل فقدان بافت

2.عدم دانش کافی برای پاسخگویی دقیق

3.عدم توانایی بازیابی حقایق خاص

بهبود توانایی‌های شناختی این مدل‌ها نیازمند پالایش فرایند استنتاج‌گری، بهینه‌سازی مکانیزم‌های یادگیری و ایجاد مکانیسمی برای اعتبارسنجی نتایج است .این روش‌ها به سه دسته استنتاج مبتنی بر دانش، یادگیری مبتنی بر دانش و اعتبارسنجی مبتنی بر دانش طبقه‌بندی می‌شوند که در شکل 3 بیشتر تفکیک شده‌اند.



شکل3-انواع روش های پرسش پاسخ مبتنی برllm

Knowledge-Aware Inference 1-1-3

در زمینه مدل های زبانی بزرگ (LLMs)، "استنتاج" به فرایند استفاده از مدل پیش‌آموزش دیده برای تولید متن یا پیش‌بینی‌های خاص بر اساس ورودی یا بافت داده شده اشاره دارد.

LLMs با چالش‌هایی در استنتاج مواجه هستند و نمی‌توانند خروجی صحیح یا نتایج بهینه ارائه دهند.

عوامل مختلفی مانند ورودی ابهام‌آلود یا فقدان بافت واضح ممکن است منجر به این شکست‌ها شود. دلیل دیگر می‌تواند خلأ دانش، سوگیری داده‌های آموزشی یا عدم توانایی تعمیم‌پذیری به سناریوهای جدید و ناشناخته باشد.

LLMs همچنین ممکن است در انجام تسک هایی که نیاز به استدلال پیچیده و چند مرحله‌ای دارند، با مشکل مواجه شوند. برخلاف انسان‌ها، LLMs اغلب نمی‌توانند برای روشن سازی سوالات مبهم، اطلاعات بیشتری کسب کنند و درک خود را اصلاح نمایند.تکنیک‌های مختلفی برای بهبود توانایی‌های استنتاج و استدلال مدل طراحی شده‌اند. مدل نه تنها باید ظرایف سوال را درک کند بلکه باید از بافت[7] مرتبط مورد نیاز برای استدلال دقیق آگاه باشد. هدایت یا راهنمایی برای تکالیف استدلالی خاص نیز می‌تواند به مدل کمک کند. این ممکن است نیاز به دانستن حقایق و هنجارهای دنیای واقعی داشته باشد.گراف‌های دانش (KGs) منبع عالی نمایش ساختاریافته دانش نمادین حقایق دنیای واقعی هستند. محققان به طور فعال در حال کار روی استفاده از گراف‌های دانش موجود و تقویت دانش خارجی در سطح ورودی (یا سرنخ) هستند تا مدل بتواند به بافت مرتبط دسترسی پیدا کند و توانایی‌های استدلالی خود را بهبود بخشد.

ما همه این تکنیک‌ها را روش‌های استنتاج آگاه از دانش می‌نامیم. این‌ها را به طور خاص به بازیابی تقویت شده با KG[8]، استدلال تقویت شده با KG[9] و تولید کنترل شده توسط KG[10] طبقه‌بندی می‌کنیم.

[7] context
[8] KG-Augmented Retrieval
[9] KG-Augmented Reasoning
[10] KG-Controlled Generation

مدل‌های تولید متن تقویت‌شده با بازیابی مانند RAG[1] و RALM [2]برای افزایش آگاهی متنی مدل‌های زبانی بزرگ (LLMs) در انجام تکالیف مبتنی بر دانش، محبوبیت یافته‌اند.آن‌ها در طول فرایند تولید متن، اسناد مرتبط را در اختیار LLMs قرار می‌دهند و بدون تغییر معماری LLM، موضوع توهم‌زایی را به طور مؤثری کاهش داده‌اند.این روش‌ها از قدرت حافظه پارامتریک مدل پیش‌آموزش‌دیده و حافظه غیرپارامتریک اسناد بازیابی‌شده برای تولید متن استفاده می‌کنند. آن‌ها می‌توانند برخی از مشکلات توهم‌زایی در LLMs را مرتفع کنند، زیرا دانش مدل گسترش یافته و از طریق ماژول بازیابی به طور مستقیم پالایش می‌شود.

LLMs می‌توانند بدون آموزش اضافی، دانش دسترسی‌یافته را تفسیر و بررسی کنند. این روش‌ها به ویژه برای تکالیف نیازمند دانش خارجی که انسان‌ها بدون منابع اضافی دشواری در انجام آن دارند، باارزش هستند.

در مدل RALM، k سند برتر توسط بازیاب‌های LLM انتخاب و به ورودی الحاق می‌شوند تا بر اساس اسناد مرتبط از منبع داده‌های داخلی، پاسخ تولید شود. اما اگر دانش خارجی به خوبی سازماندهی و از منابع داده‌های ساختاریافته، پایگاه‌داده‌ها یا گراف‌های دانش فراهم شود، همراستایی بیشتری با دقت واقع‌گرایانه خواهد داشت.

Baek و همکاران مدل KAPING [3]را پیشنهاد کردند که دانش مرتبط را با مطابقت موجودیت مورد پرسش بازیابی و سپس سه‌تایی‌های مرتبط را از گراف دانش استخراج می‌کند. این سه‌تایی‌ها به سرنخها الحاق شده و برای پاسخ‌دهی به پرسش بدون نمونه استفاده می‌شوند.

Knowledge-Augmented language model PromptING (KAPING) را برای بهبود پاسخگویی به سوالات گراف دانش بدون نیاز به دانش قبلی معرفی می‌کند. نویسندگان تاکید می‌کنند که مدل‌های زبان بزرگ (LLMs) قادرند به انجام وظایف پاسخگویی به سوالات بدون آموزش خاص در دامنه مورد نظر بپردازند. این مدل‌ها بر اساس دانش داخلی که در زمان پیش‌آموزش در پارامترهای خود ذخیره کرده‌اند، عملکرد خود را انجام می‌دهند. با این حال، این دانش داخلی ممکن است ناکافی یا نادقیق باشد که موجب تولید پاسخ‌های اشتباه شود. علاوه بر این، آموزش مجدد این مدل‌ها برای به‌روزرسانی دانش آنها هزینه‌بر است.

برای حل این محدودیت‌ها، نویسندگان پیشنهاد می‌دهند که دانش را مستقیماً در ورودی مدل‌های زبانی بزرگ افزود کنند. آنها ابتدا حقایق مرتبط با سوال ورودی را از گراف دانش با استفاده از شباهتهای معنایی بین سوال و حقایق مربوطه استخراج می‌کنند. سپس این حقایق به صورت پیشنهاد در ابتدای سوال ورودی قرار می‌گیرند و

به عنوان پراکنده به مدل‌های زبانی بزرگ ارسال می‌شوند تا پاسخ را تولید کنند. این چارچوب KAPING نیازی به آموزش مدل ندارد و به همین دلیل به‌طور کامل بدون نیاز به دانش قبلی عمل می‌کند.

عملکرد چارچوب KAPING در وظیفه پاسخگویی به سوالات گراف دانش، که هدف آن پاسخ به سوال کاربر بر اساس حقایق گراف دانش است، ارزیابی شده است. نتایج نشان می‌دهد که KAPING در مقایسه با روش‌های مرجع بدون نیاز به دانش قبلی، در میان چندین LLM با اندازه‌های مختلف، به طور میانگین تا 48٪ عملکرد بهتری داشته است.[1][3]

Wu و همکاران پیشنهاد کردند که بازنویسی سه‌تایی‌های استخراج‌شده به عبارت‌های متنوع، عملکرد مدل‌های زبانی بزرگ را بیشتر بهبود می‌بخشد.[4]

یکی از محدودیت‌های مدل‌های زبان بزرگ در حل مسائل مرتبط با دانش، عدم توانایی آن‌ها در حفظ دانش جهانی، به ویژه دانش‌هایی کمتر شناخته شده است در این روش پیشنهاد شده Wu ، تلاش می‌شود تا با استفاده از دانش موجود در گراف دانش، عملکرد مدل‌های زبان بزرگ در حل مسئله پرسش و پاسخ در گراف دانش بهبود یابد. پژوهش‌های قبلی نشان داده‌اند که استفاده از دانش گرافی برای افزایش تحریک مدل‌های زبان بزرگ می‌تواند عملکرد آن‌ها را به طور قابل توجهی در مسئله پرسش و پاسخ در گراف دانش بهبود بخشد.

با این حال، روش‌های قبلی در این حوزه ناکارآمدی در تعبیر متنی دانش گراف را دارند، به این معنی که فاصله بین نمایش گرافی و نمایش متنی را نادیده می‌گیرند. به منظور رفع این مشکل، در این مقاله روشی به نام "KG-to-Text" پیشنهاد شده است که قادر است دانش موجود در گراف را به جملات متنی با اطلاعات دقیق و کاربردی برای مسئله پرسش و پاسخ تبدیل کند. بر اساس این روش، یک چارچوب تقویت شده برای مدل‌های زبان بزرگ با استفاده از گراف دانش پیشنهاد شده است. نتایج آزمایشات انجام شده بر روی چندین مجموعه داده نشان می‌دهد که روش پیشنهادی KG-to-Text در مقایسه با روش‌های قبلی، در دقت پاسخ و کاربردی بودن جملات دانش بهبود قابل توجهی دارد.[5][4]

[1] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. arXiv preprint arXiv:2212.10509.

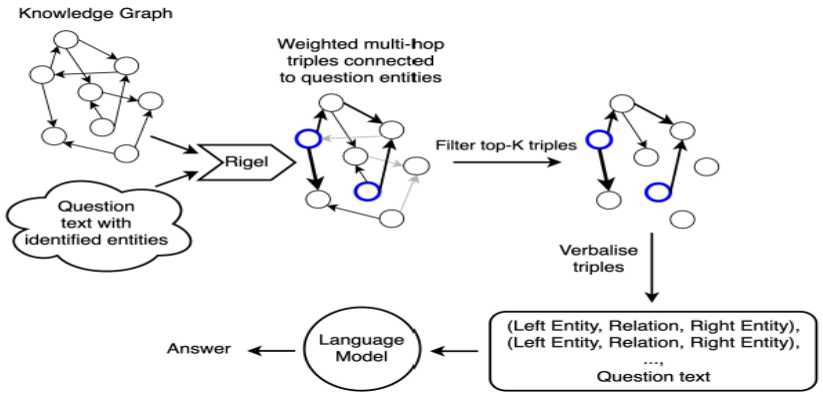
[2] Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM Tonmoy, Aman Chadha, Amit P Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models–an extensive definition, quantification, and prescriptive remediations. arXiv preprint arXiv:2310.04988

[3] Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. arXiv preprint arXiv:2306.04136

[4] Yike Wu, Nan Hu, Guilin Qi, Sheng Bi, Jie Ren, Anhuan Xie, and Wei Song. 2023. Retrieve-rewriteanswer: A kg-to-text enhanced llms framework for knowledge graph question answering. arXiv preprint arXiv:2309.11206.

[5] Garima Agrawal, Tharindu Kumarage, Zeyad Alghami, Huan Liu. "Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey." arXiv preprint arXiv:2311.07914, November 14, 2023.

در مقاله Sen et al. (Sen et al، 2023) ادعا شده است که شباهت برای یافتن حقایق مرتبط با سوالات پیچیده کافی نیست. آنها پیشنهاد دادند از یک ماژول بازیابی بر اساس مدل پرسش و پاسخ گراف دانش به صورت دنباله به دنباله استفاده شود تا توزیع روی روابط چندگانه در یک گراف دانش برای پاسخ به سوالات پیش‌بینی شود. سه‌تایی‌های برتر که به دست آمده‌اند به عنوان متن زمینه به سوال معرفی شده به مدل زبان طبیعی (LLM) اضافه می‌شوند. آنها نشان می‌دهند که در مقایسه با طرح مستقیم سوالات به LLM بدون دانش خارجی، میانگین بهبود 47٪ در مجموعه داده‌های مختلف پرسش و پاسخ حاصل می‌شود.[1](شکل4)



شکل4

مدل Jiang et al. (StructGPT، 2023) از سه منبع داده ساختاری استفاده می‌کند: گراف‌های دانش، جداول داده و پایگاه‌های داده ساختاری. استفاده از کوئری‌های ساختاری برای استخراج اطلاعات استفاده شد که

با ادغام آنها به عنوان جملات بلند وارد LLM می‌شوند تا روابط کاندیدای مرتبط‌ترین را ارائه دهند. سه‌تایی‌ها با رابطه پیشنهاد شده به همراه متن اصلی افزوده می‌شوند و به LLM برای پاسخ نهایی معرفی می‌شوند.[2]

الگوریتمی که در مقاله "StructGPT: چارچوبی کلی برای مدل زبان بزرگ برای استدلال بر روی داده‌های ساختاری" معرفی شده است، هدف آن بهبود توانایی استدلال مدل‌های زبان بزرگ بر روی داده‌های ساختاری است. این الگوریتم یک چارچوب خواندن-استدلال تکراری (IRR) به نام StructGPT را معرفی می‌کند. در ادامه، خلاصه‌ای از این الگوریتم آورده شده است:

۱. هدف اصلی StructGPT توانایی استدلال مدل‌های زبان بزرگ بر روی داده‌های ساختاری است. داده‌های ساختاری به داده‌هایی اطلاق می‌شود که در یک فرمت استاندارد سازماندهی شده‌اند، مانند نمودارهای دانش و پایگاه‌های داده.

۲. برای پشتیبانی از استدلال بر روی داده‌های ساختاری، الگوریتم از رابط‌های ویژه استفاده می‌کند که دسترسی و فیلترینگ داده را برای مدل‌های زبان بزرگ فراهم می‌کنند. این رابط‌ها بر اساس ویژگی‌های داده‌های ساختاری طراحی شده‌اند و دسترسی دقیق و کارآمد به شواهد مورد نیاز را فراهم می‌کنند.

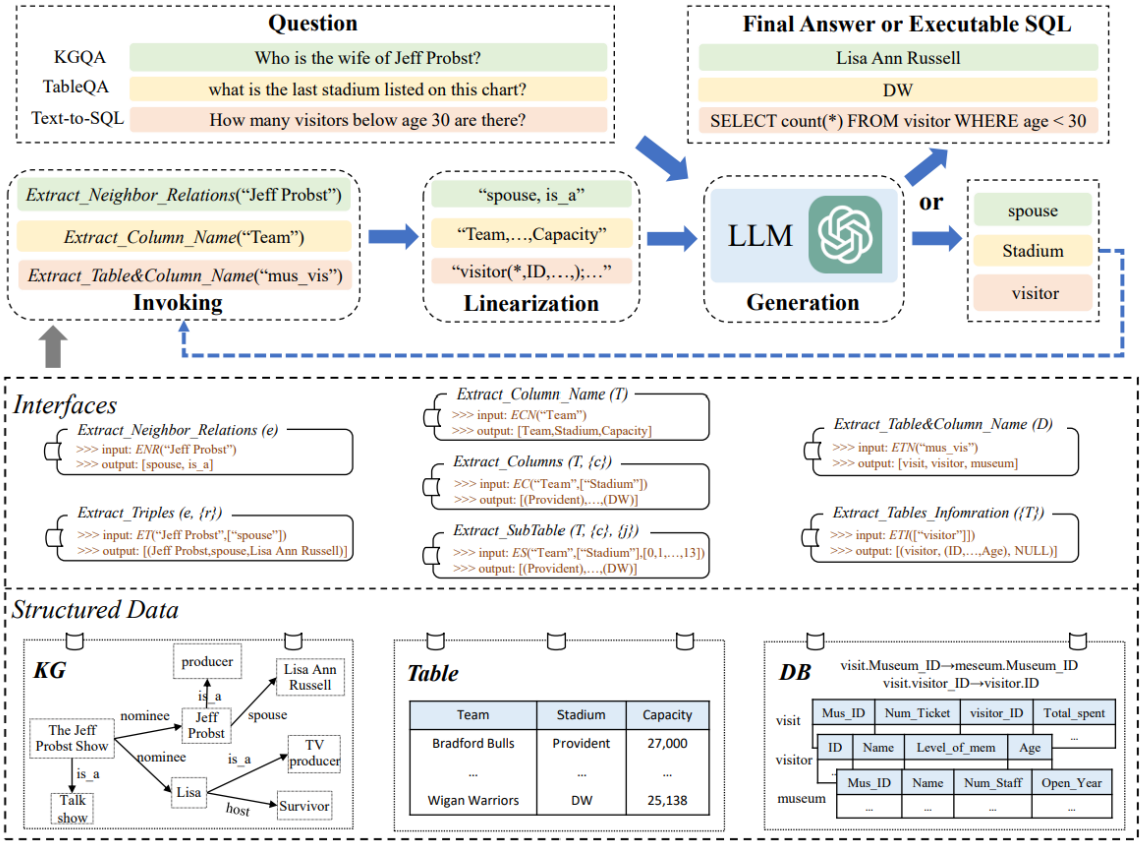
۳. الگوریتم StructGPT از یک فرآیند فراخوانی-خطی‌سازی-تولید برای حمایت از استدلال مدل‌های زبان بر روی داده‌های ساختاری با استفاده از رابط‌های خارجی استفاده می‌کند. با تکرار این فرآیند با استفاده از رابط‌های ارائه شده، می‌توان به آرامی به پاسخ هدف برای یک پرسش داده شده نزدیک شد.[2]

۴. این الگوریتم اولین کاری است که بررسی می‌کند چگونه می‌توان مدل‌های زبان بزرگ را در استدلال بر روی انواع مختلفی از داده‌های ساختاری (شامل جداول، نمودارهای دانش و پایگاه‌های داده) در یک الگومشابه واحد حمایت کرد.[2]

که توضیحاتش را میتوان در شکل5 دید.

[1] Priyanka Sen, Sandeep Mavadia, and Amir Saffari. 2023. Knowledge graph-augmented language models for complex question answering

[2] Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Structgpt: A general framework for large language model to reason over structured data. arXiv preprint arXiv:2305.09645



شکل5-structGPT

درباره‌ی حفظ دانش واقعی توسط LLMs (مدل‌های زبان بزرگ) و تأثیر حافظه غیرپارامتریک تحلیلی وجوددارد. این تحلیل می‌کند تا چه اندازه دانش واقعی به بهبود عملکرد کمک می‌کند و آیا امکان ساخت یک سیستمی که به طور تطبیقی حافظه‌های غیرپارامتریک و پارامتریک را ترکیب کند، وجود دارد یا خیر. پژوهشگران دریافتند که این مدل‌ها در موجودیت‌ها و روابط محبوب تر عملکرد بهتری دارند. با این حال، LLMs اغلب با موضوعات کمتر محبوب یا روابط خاص مشکل دارند و افزایش اندازه مدل در این موارد عملکرد را بهبود نمی‌بخشد. با این حال، تقویت LLMs با داده‌های بازیابی شده، بهبودهای قابل توجهی را ایجاد می‌کند.

KG-Augmented Reasoning 1-2-3

در این روش‌ها، از مراحل استدلال میانی متوالی استفاده می‌شود تا توانایی استدلال پیچیده مدل‌های زبان بزرگ را بهبود بخشید. روش‌های مورد استفاده شامل Chain of Thought (CoT)[1]، Chain of Thought with Self-Consistency (CoT-SC)[2]، Program-Aided Language Model (PAL)[3]، Reason and Act (ReAct)[4]، و Reflexion[5] می‌باشند.

روش CoT شامل سه بخش <input, chainofthought, output> است. در این روش، chainofthought مجموعه‌ای از مراحل استدلال زبان طبیعی است که به نتیجه نهایی منجر می‌شود. این روش به تعداد مثال‌های یادگیری کم در زمینه‌ی کار بیشتر از روش‌های دیگر شبیه است، تنها تفاوت آن این است که از راهنمایی به جای تنظیم مدل برای هر وظیفه استفاده می‌شود. این رویکردها به روند استدلال مرحله به مرحله برای به دست آوردن پاسخ شباهت دارند و تفسیری از اینکه مدل چگونه به یک پاسخ خاص رسیده است، ارائه می‌دهند. همچنین، این روش به مدل فرصتی می‌دهد تا مشکلات در مسیر استدلال را رفع کند.

این روش‌ها اغلب برای وظایفی مانند مسائل ریاضی با کلمات، استدلال مشترک و محاسبات نمادین مانند ادغام آخرین حرف، پرتاب سکه و یا هر وظیفه دیگری که انسان‌ها با توضیح مراحل به زبان قادر به حل آن هستند، استفاده می‌شوند. دلیل اینکه روش CoT موثر است، این است که به مدل امکان می‌دهد برای مسائل سخت‌تر، محاسبات یا توکن‌های میانی بیشتری را صرف کند و دوم، این راهنماها به مدل اجازه می‌دهند تا به دانش مربوطه که در طول آموزش قبلی کسب کرده است، دسترسی پیدا کند. به عنوان مثال، در یک مسئله ریاضی، برای مدل سخت است تا تمام معناشناختی را به یک معادله ترجمه کند. اما یک زنجیره از افکار به او این امکان را می‌دهد تا با استفاده از مراحل میانی در زبان طبیعی به هر بخش از سوبهتر استدلال کند. روش CoT دامنه‌ی وظایفی را که مدل‌های زبانی می‌توانند پتانسیل حل کنند، گسترش می‌دهد. اما افزودن دستی زنجیره‌ی افکار به منظور تعمیم برای حالت بدون نمونه، هزینه‌بر است.

روش "درخت افکار" (ToT)[6] یک تعمیم از روش "زنجیره افکار" است. این روش امکان بررسی واحدهای هماهنگ متن یا "افکار" را که به عنوان مراحل میانی در حل مسئله عمل می‌کنند، فراهم می‌کند. این روش به مدل‌های زبان بزرگ امکان می‌دهد تصمیم‌گیری هدفمندتری را با در نظر گرفتن مسیرهای استدلال متعدد و ارزیابی خود برای انتخاب مسیر بعدی و به جلو نگاه کردن یا بازگشت به عقب در صورت نیاز به انتخاب‌های

سراسری داشته باشند. این روش توانایی حل مسائل LLMs را بهبود می‌بخشد. آنها وظایفی را در نظر گرفتند که نیاز به برنامه‌ریزی یا جستجوی غیرتافل نظیر نوشتن خلاقانه یا انجام کراسورد دارند. تکنیک ToT الهام گرفته شده از رویکرد ذهن انسان به حل و فصل واژه‌های استدلال پیچیده از طریق آزمون و خطا است.

[1] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35:24824–24837

[2] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

[3] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In International Conference on Machine Learning, pages 10764–10799. PMLR

[4] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629.

[5] Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. arXiv preprint arXiv:2303.11366.

[6] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. arXiv preprint arXiv:2305.10601.

بر اساس راهبردهای "زنجیره افکار" و "درخت افکار"، روش‌های مختلفی برای افزایش توانایی استدلال در استنتاج عمومی و خاص در حوزه‌های مختلف و پاسخ‌دهی چند مرحله‌ای ارائه شده است.

در مقاله He و همکاران (He et al.، 2022)[1]، روش بازنگری با بازیابی (RR) ارائه شده است که در آن مراحل استدلال تجزیه‌شده از راهنمایی زنجیره افکار برای بازیابی دانش خارجی مرتبط با LLMs استفاده می‌شوند تا توضیحات دقیق و بهبود دقت پیش‌بینی پاسخ را فراهم کنند. روش IRCoT (Trivedi و همکاران، 2022) [2]زنجیره تولید افکار را با بازیابی دانش از گراف دانش ترکیب می‌کند تا با توجه به مراحل بازیابی قبلی، بازیابی و استدلال را به صورت تکراری برای سوالات استدلال چند مرحله‌ای هدایت کند.

روش MindMap (Wen و همکاران، 2023) [3]یک رویکرد قابل پیاده‌سازی است که توانایی استدلال گرافی را در LLMs استخراج می‌کند. این روش با ایجاد ارتباط بین موجودیت‌های کلیدی در سوال و موجودیت‌های همسایه آنها، یک زیرگراف را از گراف دانش استخراج می‌کند. سپس با استفاده از این زیرگراف، مدرک‌های مبتنی بر مسیر را ایجاد می‌کند که به LLMs در استدلال کمک می‌کند و آنها را قادر می‌سازد تا ورودی‌های

گرافی را درک کرده و نقشه ذهنی خود را بر اساس تولید مبتنی بر مدرک ایجاد کنند که پشتوانه تولید پاسخ است.

روش استدلال بر روی گراف (RoG) (Luo et al. ،2023[4]) ابتدا تمام مسیرهای رابطه‌ای مورد نیاز برای یک سوال داده شده را تولید می‌کند که به آنها "مسیرهای برنامه‌ریزی دقیق" می‌گویند. این برنامه‌ها به گراف دانش داده می‌شوند تا مسیرهای استدلال دقیقی را تولید کنند. این مسیرها به نوبه خود به LLMs امکان می‌دهند تا استدلال دقیق و قابل تفسیر را انجام دهند.

. با این حال، سؤال اساسی این است که آیا شبکه‌های عصبی واقعاً در "استدلال" مشغول هستند و نیست و نمی‌توان قطعیت داشت که دنبال کردن مسیر استدلال صحیح همیشه به پاسخ‌های دقیق منجر می‌شود (Qiao et al.، 2022؛ Jiang et al.، 2020).[5] این جهت تحقیق، تأکید بر رابطه پیچیده بین راهنماها، مدل‌های زبانی و استدلال، پتانسیل قابل توجهی برای بررسی‌های بیشتر دارد.

[1] Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. arXiv preprint arXiv:2301.00303.

[2] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. arXiv preprint arXiv:2212.10509.

[3] Yilin Wen, Zifeng Wang, and Jimeng Sun. 2023. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. arXiv preprint arXiv:2308.09729.

[4] Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2023. Reasoning on graphs: Faithful and interpretable large language model reasoning. arXiv preprint arXiv:2310.01061.

[5] Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. arXiv preprint arXiv:2212.09597.

Knowledge-Controlled Generation 1-3-3

در واقع روش‌های دیگری وجود دارد که از مدل‌های زبانی برای تولید دانش و اجرای وظایف مختلف استفاده می‌کنند. در ادامه تکنیک‌های دیگری شرح داده میشود

1. نمونه‌های کم: در این روش، ابتدا اظهارنظرهای مرتبط با سؤال با استفاده از مدل زبانی و تعداد کمی نمونه تولید می‌شود. سپس یک مدل جداگانه از اظهارنظرهای تولید شده استنتاج می‌برد و پیش‌بینی با بالاترین اطمینان را به عنوان پاسخ نهایی انتخاب می‌کند.[1]

2. مدل‌های زبانی مبتنی بر برنامه نویسی (PLMs): در این روش، با استفاده از متن محیط، یک PLM مبتنی بر برنامه نویسی به نام Codex فراخوانی می‌شود. Codex تماس‌های API مربوطه را برای اجرای وظیفه مورد نیاز تولید می‌کند.[2]

3. اتصال به گراف دانش: در روش KB-Binder، Codex با یک گراف دانش ترکیب می‌شود. Codex شکل منطقی یک پیش‌نویس برای سؤال خاصی تولید می‌کند و گراف دانش موجودیت‌ها را به آن متصل کرده و پاسخ کامل را فراهم می‌کند.[3]

4. ارائه پراکنده مبتنی بر کلمات کلیدی: در این روش، جملاتی با قالب پراکنده برای موجودیت‌ها در گراف دانش تولید می‌شود. این جملات با استفاده از پرس‌وجوهای SPARQL روی گراف دانش، بازدهی بیشتری و دقت پیش‌بینی بهتری دارند.[4]

5. تنظیم پراکنده: در روش KnowPrompt، پراکنده‌هایی از یک مدل زبانی پیش‌تر آموزش دیده تولید می‌شوند و برای استخراج روابط در وظایف پراکنده، تنظیم پراکنده انجام می‌شود.

این روش‌ها نشان می‌دهند که چگونه می‌توان از مدل‌های زبانی برای تولید دانش و بهبود وظایف مختلف استفاده کرد. هر روش مزایا و محدودیت‌های خود را دارد و کارایی آن‌ها بستگی به مورد استفاده خاص دارد.

در روش BeamQA (Atif et al., 2023)، مدل زبانی برای تولید مسیرهای استدلالی استفاده می‌شود که برای پیش‌بینی پیوند در گراف دانش، از جستجوی مبتنی بر تعبیه گراف دانش استفاده می‌کند. اندازه شعاع بیم برگشتی کنترل کننده‌ی انتخاب چندین رابطه در هر مرحله است. مسیرها و پیوندهایی که تولید می‌شوند، برای استخراج سه‌تایی‌های گراف دانش استفاده می‌شوند که در ادامه برای وظایف پرسش و پاسخ درباره‌ی گراف دانش (KGQA) استفاده می‌شوند.[5]

استفاده از guardrails (قوانین و محدودیت‌ها) نیز در هوش مصنوعی نمونه‌برداری برای تعیین محدوده‌هایی که مدل می‌تواند در آن عمل کند، پیشنهاد می‌شود. guardrails به عنوان محدودیت‌ها و قوانینی عمل

می‌کنند که به کنترل فرایند تولید خروجی مدل زبانی کمک می‌کنند و اطمینان از استفاده ایمن و امن از هوش مصنوعی نمونه‌برداری را فراهم می‌کنند.

روش‌های تولید کنترل‌شده دانش می‌توانند به مدل‌های زبانی کمک کنند تا اطمینان حاصل کنند که اطلاعات با واقعیت‌ها هماهنگ است و جلوی گسترش اطلاعات نادرست را بگیرند. انتولوژی گراف دانش، دستورالعمل‌های دقیقی برای محدودیت‌های خاصی که به دامنه مشخصی سفارشی شده است، توصیف می‌کند. توسعه مدل‌های زبانی با گراف‌های دانش می‌تواند برای تعیین محدوده‌های تولید خروجی برای مدل‌های زبانی آسانتر و صمیمی‌تر شود.

[1] Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, et al. 2022. Binding language models in symbolic languages. arXiv preprint arXiv:2210.02875.

[2] Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, et al. 2022. Binding language models in symbolic languages. arXiv preprint arXiv:2210.02875

[3] Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhu Chen. 2023b. Few-shot in-context learning for knowledge base question answering. arXiv preprint arXiv:2305.01750.

[4] Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledgeaware prompt-tuning with synergistic optimization for relation extraction. In Proceedings of the ACM Web conference 2022, pages 2778–2788.

[5] Farah Atif, Ola El Khatib, and Djellel Difallah. 2023. Beamqa: Multi-hop knowledge graph question answering with sequence-to-sequence prediction and beam search. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 781–790.

3-1-2 ارزیابی و نتیجه گیری:

استفاده از گراف‌های دانش با استفاده از تکنیک‌های متنوع مشروح در بخش قبل، یک روش ارزشمند برای بهبود عملکرد مدل‌های زبانی (LLMs) در وظایف شناختی مختلف است. هر یک از این روش‌ها نقاط قوت و محدودیت‌های خود را دارند.

افزایش بازیابی، که در آن مدل در هنگام استنتاج اطلاعاتی را از گراف دانش یا پایگاه داده خارجی بازیابی می‌کند، برای برنامه‌های زمان واقعی کارآمد است. همچنین، به مدل امکان دسترسی به اطلاعات به‌روز را

می‌دهد. KAPING (Baek و همکاران، 2023) عملکرد را بر اساس این اندازه‌گیری می‌کند که آیا توکن‌های تولید شده از پرسمان‌ها در موجودیت‌های پاسخ گنجانده شده‌اند و چقدر سه‌تایی‌های بازیابی شده از گراف دانش در تولید پاسخ مفید هستند. KAPING به طور قابل توجهی دقت وظایف نیازمند دانش واقعی را بهبود داد. به‌ویژه، بهبود عملکرد در مدل‌های کوچکتر بود که نشان می‌دهد افزایش دانش به جای افزایش اندازه مدل مفید بود. با این حال، عملکرد روش‌های افزایش بازیابی به کارایی ماژول‌های بازیابی وابسته است. آنها ممکن است به پرسمان‌های پیچیده یا نیواندیشی به خوبی پاسخ ندهند زیرا محدود به اطلاعات موجود در گراف دانش هستند.

استدلال از طریق پرسمان با استفاده از "زنجیره افکار (CoT)" شامل ارائه دستورالعمل‌ها یا پرسمان‌های صریح به مدل برای هدایت آن در فرآیند استدلال در حالی که از اطلاعات موجود در گراف دانش استفاده می‌کند، ارزان و عملی است و کنترل مستقیم بر تمرکز مدل و هدایت برای تنظیم دقیق پرسمان‌ها برای وظایف خاص را فراهم می‌کند. دقت ChatGPT از 66.8٪ با استفاده از پرسمان استاندارد به 85.7٪ با افزایش گراف‌های دانش با استفاده از روش‌های استدلال بر روی گراف (RoG) (Luo و همکاران، 2023) در وظایف پاسخ‌دهی به سوال‌ها افزایش یافت.

با افرض کردن، نیاز به ساخت دقیق پرسمان‌ها و تطبیق بهتر با پرسمان‌های متنوع یا غیرمنتظره را دارد. عملکرد به طور قابل توجهی بهبود نمی‌یابد وقتی روش‌های CoT برای حل وظایف استدلال حسابی در مدل‌های کوچک (حدود 100 میلیارد پارامتر) اعمال می‌شوند. این مدل‌ها توانایی‌های حسابی ضعیفی دارند و درک معنایی نسبتاً محدودی دارند، بنابراین قادر به تعمیم به وظایف جدید نیستند. با این حال، برای مسائل پیچیده‌تر در مدل‌های بزرگ مانند بزرگترین مدل GPT و PALM (Wei و همکاران، b2022)، عملکرد دو برابر شد.

یک روش دیگر افزودن اطلاعات به گراف دانش برای تولید متناظر یا اطلاعات بیشتر در طول آموزش یا استنتاج است. آنها اطلاعات مرتبط بازمانده را تولید می‌کنند، بنابراین انعطاف‌پذیری در برخورد با پرسمان[11]های متنوع را

[11] prompt

دارند. با این حال، کیفیت تولید ممکن است متنوع باشد و ممکن است باعث ایجاد اطلاعات نادرست یا غیرمرتبط شود.

روش‌های آموزش مدل با استفاده از آموزش اولیه یا تنظیم دقیق مدل را قادر می‌سازد تا مدل از دانش مشخصی که از داده گراف دانش در مدل وارد شده است یاد بگیرد. این روش‌ها می‌توانند عملکرد وظیفه‌ای را بهبود بخشند، اما منابع و هزینه‌های مرتبط با آنها زیاد است. ترکیب استراتژیک دانش خارجی با مدل پیش‌آموزشی، نیاز به قدرت محاسباتی قابل توجه، مجموعه داده‌های گسترده و تلاش‌های دقیق برای تنظیم دقیق دارد. یک چالش دیگر این است که تنظیم دقیق وابسته به داده است، بنابراین وظیفه-محدود و قابل کلیت نیست.

روش‌های اعتبارسنجی دانش با استفاده از بررسی حقایق اطمینان حاصل می‌کنند که محتوای تولید شده توسط مدل قابل اعتماد است[2][1].

[1] Garima Agrawal, Tharindu Kumarage, Zeyad Alghami, Huan Liu. "Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey." arXiv preprint arXiv:2311.07914, November 14, 2023.

[2] Yasumasa Onoe, Michael JQ Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. 2023. Can lms learn new entities from descriptions? challenges in propagating injected knowledge. arXiv preprint arXiv:2305.01651.

ارزیابی چت جی پی تی در مقایسه با سایر llm:

در دوره فعلی، تعداد زیادی از مدل‌های زبانی برای پاسخ به پرسش‌های کاربران ظاهر شده‌اند. به ویژه، مدل زبانی GPT-3.5 Turbo به عنوان تکنولوژی پایه ChatGPT توجه قابل توجهی جلب کرده است. این مدل با بهره‌گیری از پارامترهای گسترده، به بهترین نحو به سوالات مختلف پاسخ می‌دهد. با این حال، به دلیل وابستگی به دانش داخلی و outdatedبودن دانش آن، دقت پاسخ‌ها ممکن است زیاد نباشد. در اینجا ChatGPT را به عنوان یک سامانه پرسش و پاسخ (QAS) مورد بررسی قرار می‌دهد و عملکرد آن را با سایر سامانه‌های QAS مقایسه می‌کند. متمرکز اصلی این بررسی بر ارزیابی توانایی ChatGPT در استخراج پاسخ از پاراگراف‌های ارائه شده است، که یک قابلیت اصلی در QAS است. به علاوه، مقایسه عملکرد در حالاتی بدون استخراج دانش از نیز انجام شده است. ارزیابی از مجموعه داده‌های شناخته‌شده پرسش و پاسخ (QA)، از جمله SQuAD، NewsQA و PersianQuAD بر روی زبان‌های انگلیسی و فارسی استفاده شد. در این ارزیابی معیارهایی از قبیل امتیاز F-score، exact match،accuracy به‌کار رفت.

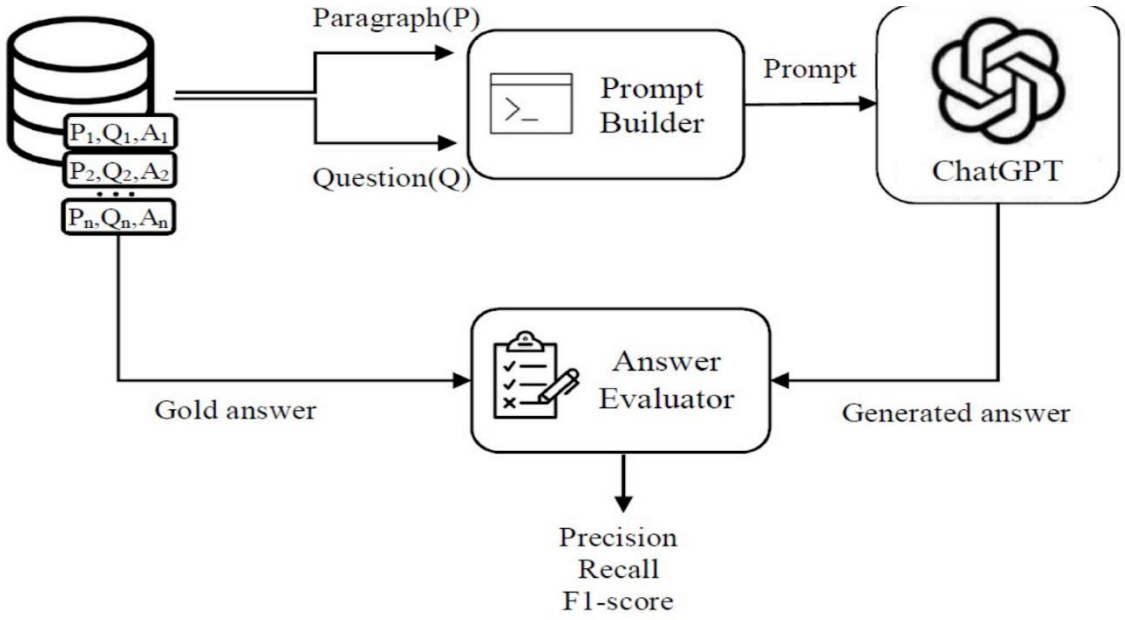
این مطالعه نشان می‌دهد که در حالی که ChatGPT به عنوان یک مدل تولیدی کارآمد عمل می‌کند، در پاسخ به سوالات نسبت به مدل‌های خاص وظیفه کمتر مؤثر است. ارائه متن ,context کمک می‌کند تا عملکرد آن بهبود یابد و prompt engineering دقت را به ویژه برای سوالاتی که پاسخ صریحی در پاراگراف‌های ارائه شده وجود ندارد، افزایش می‌دهد. ChatGPT در سوالات حقیقی ساده نسبت به انواع سوالات "چگونه" و "چرا" عملکرد برتری دارد.
پاسخ دقیق و موثر به سوالات، یک جزء بحرانی در فهم زبان طبیعی و سیستم‌های ارتباطی است. وظایف پرسش و پاسخ (QA) به عنوان یک چالش اساسی در حوزه هوش مصنوعی (AI) توجه زیادی جلب کرده‌اند، زیرا نمایانگر یک چالش اساسی برای توسعه سیستم‌های هوشمند است که بتوانند به سوالات کاربران به شیوه‌ای مشابه انسانی فهمیده و پاسخ دهند. با ظهور مدل‌های زبان بزرگ مقیاس مانند ChatGPT که بر پایه تکنیک‌های پیشرفته یادگیری عمیق قرار دارند، علاقه به ارزیابی عملکرد آنها در تسک های QA در حال افزایش است.

ما یک سری آزمایشات و ارزیابی‌ها با استفاده از دستگاه‌های مختلف پرسش و پاسخ برای ارزیابی کارایی و دقت ChatGPT در ارائه پاسخ‌های دقیق به سوالات کاربران انجام دادیم. با بهره‌گیری از مجموعه داده‌های بنچمارک خوب شناخته‌شده مانند Stanford Question Answering Dataset (SQuAD)، ما می‌توانیم عملکرد ChatGPT را با سایر مدل‌های برتر و ارزیابی توانایی‌های آن در مقابل انواع سوالات و پیچیدگی‌های زبانی مختلف مقایسه کنیم. انتخاب مجموعه‌های داده برای ارزیابی جامع عملکرد ChatGPT در پرسش و پاسخ بسیار حیاتی است. علاوه بر SQuAD، ما از NewsQA به عنوان یک مجموعه داده QA در حوزه اخبار و PersianQuAD که گستره وسیعی از موضوعات و انواع سوالات را در زبان فارسی پوشش می‌دهد، استفاده کردیم.
نتیجه ارزیابی:
در این مطالعه، یک تجزیه و تحلیل جامع از عملکرد ChatGPT به عنوان یک سامانه پرسش و پاسخ (QAS) انجام داده و عملکرد آن را با سایر مدل‌های موجود مقایسه کردیم. ارزیابی ما بر روی توانایی مدل برای استخراج پاسخ از پاراگراف‌های ارائه شده و همچنین عملکرد آن در حالات بدون متن اطراف تمرکز داشت. ما جنبه‌های مختلف را بررسی کردیم که شامل هالوسیناسیون پاسخ، پیچیدگی سوال و تأثیر متن روی عملکرد بوده است.

در اختتامیه، ارزیابی جامع ChatGPT به عنوان یک سامانه پرسش و پاسخ نقاط قوت، محدودیت‌ها و زمینه‌هایی که می‌تواند بهبود یابد، را ارائه داده است. با بهره‌گیری از مجموعه داده‌های شناخته‌شده پرسش و

پاسخ (QA) به هر دو زبان انگلیسی و فارسی، از معیارهایی نظیر امتیاز اف-اسکور، تطابق دقیق و بازیابی برای ارزیابی عملکرد ChatGPT استفاده کردیم.

مدلی که برای ارزیابی chatgptدر سیستم های question answering ارائه شده است:



شکل 6

برای ارزیابی دقت و کارآیی پاسخ‌های حاصل از سامانه ChatGPT، یک چارچوب ارزیابی به کار گرفته شده است (شکل 6 را ببینید). این چارچوب شامل سه مؤلفه اصلی است که شامل سازنده پرسمان، ChatGPT و ارزیابی پاسخ است. ساختار کلی مجموعه داده پرسش و پاسخ به شکل مجموعه‌های سه‌تایی (P، Q، A) است، که در آن P به یک پاراگراف متنی اشاره دارد، Q سوالی درباره پاراگراف را نشان می‌دهد و A به پاسخ متناظر اشاره دارد. در برخی از مجموعه‌های داده، پاسخ به سوال ممکن است در پاراگراف وجود نداشته باشد و یک فیلد اضافی مشخص می‌کند که پاسخ در پاراگراف متناظر وجود ندارد. در ادامه، به این دو نوع مجموعه داده به ترتیب با نام‌های نوع1 و نوع2 اشاره می‌شود. برای هر سه‌تایی از مجموعه داده، Pi و Qi وارد بخش سازنده پرسمان می‌شوند و یک prompt آماده و به ChatGPT ارسال می‌شود تا از P سوال شود. سپس پاسخ تولید شده توسط ChatGPT همراه با پاسخ طلایی A به بخش ارزیابی پاسخ منتقل می‌شود تا عملکرد با استفاده از معیارهای ارزیابی پیش‌تعریف شده تجزیه و تحلیل شود.

Model Name	Exact Match	Recall
RAG-original Siriwardhana et al. [2022]	4.33	7.92
BERT+ASGen Back et al. [2021]	54.7	64.5
AMANDA Kundu and Ng [2020]	48.4	63.7
DecaProp Tay et al. [2020](Yi Tay et al. 2019)	53.1	66.3
ChatGpt for Questions with all validator-confirmed answers	41.07	46.70

Table 4: Assessing the Effectiveness of Various Language Models on NewsQA

Model Name	Exact Match	Recall
PersianQA Kazemi et al. [2022]	78.8	82.97
ChatGpt	41	55

Table 5: Results on PersianQuad

شکل 7

یافته‌ها نشان می‌دهد که ChatGPT، هرچند که به عنوان یک مدل تولیدی عملکرد نشان می‌دهد، با چالش‌ها در پاسخ به سوالات نسبت به مدل‌های خاص وظیفه روبه‌رو است. متن پیرامون به عنوان یک عامل حیاتی ثابت می‌شود و عملکرد مدل در استخراج پاسخ با فراهم کردن پاراگراف‌های اطراف بهبود می‌یابد.
Prompt engineering به ویژه به شکل prompt های دو مرحله‌ای، دقت را افزایش می‌دهد، به ویژه برای سوالاتی که در پاراگراف‌های ارائه شده پاسخ صریحی ندارند.
ما توانایی ChatGPT در پاسخ به سوالات ساده و واقعی را مشاهده کردیم که نقاط قوت آن را پدیدار میکند . با این حال، با سوالات پیچیده "چگونه" و "چرا" چالش‌ها به وجود می‌آید اما با استفاده از knowledge graphها توانستند این مشکل را حل کنند.[1]

[1]Bahak, H., Taheri, F., Zojaji, Z., & Kazemi, A. (2023). Evaluating ChatGPT as a Question Answering System: A Comprehensive Analysis and Comparison with Existing Models. arXiv preprint arXiv:2312.07592.