باسمه تعالی

طرح پیشنهادی برای قرارداد پژوهشی تک پروژه غیرمقیم

۱-خلاصه اطلاعات طرح:

۱-۱-عنوان طرح به فارسی: مدل‌های زبانی کوچک: تشخیص توهّم بدون دانش خارجی

۱-۲-عنوان طرح به انگلیسی: Small Language Models: Hallucination Detection without External Knowledge

۱-۳- **نوع طرح:** ☐ بنیادی ■ کاربردی ☐ توسعه ای ■ آزمایش بنیان

۲- اطلاعات مجری:

دکتر مصطفی صالحی، عضو هیات علمی و دانشیار دانشگاه تهران

موسس آزمایشگاه شبکه‌های کامپیوتری و اطلاعاتی دانشگاه تهران، https://coinlab.ut.ac.ir/

آدرس: تهران، خ کارگر شمالی، روبروی خیابان دهم، دانشکده علوم و فنون نوین، طبقه سوم، اتاق ۳۳۷

تلفن همراه: ۰۹۱۲۳۴۰۳۹۷۷، تلفن ثابت: ۰۲۱۸۶۰۹۳۲۹۸، آدرس ایمیل: mostafa_salehi@ut.ac.ir

همکار مجری: آقای علی دروگر، دانشجوی کارشناسی ارشد دانشگاه تهران

۴- چکیده طرح(حداکثر ۲۰۰ کلمه):

مدل‌های زبانی مبتنی بر معماری ترنسفورمر رمزگشا در سال‌های اخیر به نتایج پیشرو در مسائل گوناگون پردازش زبان طبیعی دست پیدا کرده‌اند. اما این مدل‌ها در برخی شرایط متن‌هایی تولید می‌کنند که ظاهری منسجم و ساختاری مطابق دستور زبان، اما محتوایی نادرست دارند. این مشکل که در ادبیات مدل های زبانی «توهّم» نامیده می‌شود، متون تولیدی مدل‌ها را نیازمند صحت‌سنجی می‌کند. این مسئله در مدل‌های زبانی کوچک که دچار توهّم بیشتری می‌شوند اهمیت دو چندان دارد. با توجه به امکان‌پذیر نبودن صحت‌سنجی در تمامی شرایط، روش‌هایی برای استفاده از عدم قطعیت مدل جهت تشخیص توهّم ارائه شده‌اند. این روش‌ها که می‌توان آن‌ها را به دو دسته‌ی کلی «جعبه‌سیاه» و «جعبه‌سفید» تقسیم کرد، عموما مستقل از یکدیگر استفاده شده و بر روی مدل‌های بزرگتر از پنج میلیارد پارامتر بررسی شده‌اند. هدف اصلی این مطالعه استفاده از مقادیر درونی مدل‌های زبانی و ترکیب آن با روش‌های مبتنی بر نمونه‌گیری، برای تشخیص خروجی قابل اعتماد از توهّم در مدل‌های زبانی کوچک است. انتظار می‌رود با در نظر گرفتن مقادیر درونی مدل در هنگام نمونه‌گیری، نسبت به روش‌های پایه بهبود حاصل شود.

Language models based on transformer architecture have achieved state-of-the-art results in various natural language processing tasks in recent years. However, these models sometimes produce texts that appear coherent and grammatically structured but contain incorrect content. This issue, referred to as "hallucination" in the literature on language models, necessitates the validation of the texts generated by the models. This problem is particularly significant in smaller language models, which tend to hallucinate more. Given the impracticality of validation in all scenarios, methods have been proposed to leverage model uncertainty to detect hallucinations. These methods can be broadly categorized into "black-box" and "white-box" approaches, and they have generally been used independently and tested on models with more than five billion parameters. The primary aim of this study is to utilize the internal values of language models and combine them with sampling-based methods to detect reliable outputs from hallucinations in smaller language models. It is expected that considering the internal values of the model during sampling will lead to improvements over baseline methods.

۴-۱- کلمات کلیدی:

تشخیص توهّم، مدل‌های زبانی، پردازش زبان طبیعی، هوش مصنوعی قابل اعتماد

۵- مشخصات موضوعی طرح:

۵-۱- تعریف مسأله، هدف و ضرورت اجرای آن:

با معرفی مدل‌های زبانی [۲] ,[۱] مبتنی بر معماری ترنسفورمر[۳] پیشرفت قابل توجهی در پردازش زبان طبیعی ایجاد شد. مدل‌های زبانی Autoregressive به دنباله‌ای از توکن‌ها[۱]مقادیر احتمال نسبت می‌دهند. احتمال دنباله‌ای از توکن‌ها همانند t که از n توکن تشکیل شده باشد را بنا بر قواعد احتمال می‌توان به شکل زیر نوشت:

$$P(t_1, t_2, \dots, t_n) = P(t_1) \prod_{i=1}^{n-1} P(t_{i+1} | t_1, \dots, t_i) \quad (۱)$$

بنابراین می‌توان یک شبکه عصبی مصنوعی را با هدف برآورد درست‌نمایی بیشینه[۲]بر روی یک پیکره‌ی متنی آموزش داد.[۴] در سال‌های اخیر مدل‌های زبانی بزرگ مبتنی بر معماری ترنسفورمر رمزگشا[۳]به نتایج پیشرو در وظایف مختلف پردازش زبان طبیعی مانند پرسش و پاسخ[۴] ترجمه و خلاصه‌سازی دست پیدا کرده‌اند. با این وجود مدل‌های زبانی همچنان ضعف‌هایی دارند که یکی از مهم‌ترین آن‌ها تولید پاسخ‌هایی مغایر با حقیقت ولی دارای ظاهری منسجم و سازگار با دستور زبان است. به این تولیدات هذیان‌گونه‌ی مدل‌های زبانی «توهّم[۵]» گفته می‌شود. برخی محققان کلمات دیگری مانند «افسانه‌بافی»[۵] را مناسبتر از توهّم می‌دانند. علی‌رغم افزایش مطالعات بر روی سازوکار درونی این مدل‌ها [۶] توهّم همچنان یک مسئله‌ی باز در حوزه‌ی مطالعاتی مدل‌های زبانی است.

محققان از زوایای گوناگونی به تعریف و دسته‌بندی توهّمات مدل‌های زبانی پرداخته‌اند. به عنوان نمونه [۷] با ارجاع به [۸] به تولید محتوایی که با حقایق انحراف داشته و خروجی مدل را غیرقابل اعتماد کند، توهّم می‌گوید. [۷] توهّم را بر اساس درستی متن ورودی به دو دسته کلی و بر اساس محتوای توهّمات تولید شده به شش نوع متفاوت تقسیم می‌کند. در مقابل [۹] توهّم را به یکی از دو دسته‌ی «توهّم غیر وفادار» و «توهّم غیر حقیقی» تقسیم می‌کند. محتوایی که با حقایق جهان در تعارض یا دارای تناقض باشد، توهّم غیر حقیقی است. محتوایی که با دستورات کاربر یا زمینه‌ی فراهم شده توسط او همخوانی ندارد توهّم غیر وفادار است. به شکل مشابه، [۱۰] نیز توهّم را در سه گروه «تعارض با ورودی»، «تعارض با زمینه» و «تعارض با حقایق» دسته‌بندی می‌کند. ما در این تحقیق توهّم را فارغ از نوع آن مورد بررسی قرار داده و تنها بر تشخیص توهّم تمرکز داریم. برخی محققان توهّم را در یک وظیفه‌ی خاص پردازش زبان طبیعی مانند ترجمه[۱۱] یا خلاصه‌سازی[۸] بررسی می‌کنند. هدف ما بررسی توهّم در پرسش و پاسخ در مدل‌های زبانی کوچک است.

مدل‌های زبانی کوچک (Small Language Models) به مدل‌هایی گفته می‌شود که در مقایسه با مدل‌های زبانی بزرگ، تعداد پارامتر کمتری دارند. اگرچه در مورد محدوده‌ی تعریف SLMها توافق مشخصی وجود ندارد اما ما مشابه [۱۲] مدل‌های دارای یکصد میلیون تا پنج میلیارد پارامتر را به عنوان مدل زبانی کوچک در نظر می‌گیریم. در جدول ۱ لیستی از برخی مدل‌های زبانی کوچک ارائه شده است.

جدول ۱: مدل‌های زبانی کوچک

مدل زبانی	تعداد پارامتر (میلیارد)	تاریخ انتشار	منتشر کننده	مرجع
Bloom	۰٫۵۶۰ و ۱	۲۰۲۲/۱۱	BigScience	[۱۳]
Pythia	۰٫۴۱۰ ، ۰٫۱۶۰	۲۰۲۳/۰۳	EleutherAI	[۱۴]

[1] Tokens
[2] Maximum likelihood estimation
[3] Decoder-only Transformer
[4] Question Answering
[5] Hallucination
[۶] Confabulation – بر گرفته از یک اصطلاح روانپزشکی که در آن شخص به دلیل اختلال حافظه، به طور ناخودآگاه در ذهن خود رویدادهای غیر واقعی می‌سازد.
[7] Faithfulness hallucination
[8] Factfulness hallucination

			۱ ، ۱٫۴ و ۲٫۸	
MobileLLM	۰٫۱۲۵ و ۰٫۳۵	۲۰۲۴/۰۲	Meta	[۱۵]
Phi-۳-Mini	۳٫۸	۲۰۲۴/۰۴	Microsoft	[۱۶]
Gemma ۲	۲	۲۰۲۴/۰۷	Google	[۱۷]
LLaMa ۳٫۲	۱ و ۳	۲۰۲۴/۰۹	Meta	[۱۸]

SLMها نسبت به LLMها دو مزیت اساسی دارند. مدل‌های کوچکتر به توان پردازشی کمتری برای آموزش و استنتاج نسبت به مدل‌های بزرگتر نیاز دارند؛ در نتیجه در شرایط محدودیت منابع پردازشی بر مدل‌های بزرگ ارجحیت دارند. همچنین مدل‌های کوچکتر برای مطالعه از نظر تفسیرپذیری گزینه‌های مناسبتری هستند. در مقابل مدل‌های زبانی بزرگ عملکرد بسیار بهتری در وظایف گوناگون پردازش زبان طبیعی دارند.[۱۹] این نکته مدل‌های کوچک را به گزینه‌های جذابتری برای مسئله‌ی مورد بحث تبدیل می‌کند چرا که در این مدل‌ها توهّمات بیشتری مشاهده می‌شود. با این وجود، همان گونه که در بررسی پیشینه‌ی تحقیق نشان داده خواهد شد، تقریبا تمامی مطالعات اصلی در مسئله‌ی مورد بحث از مدل‌هایی با بیش از هفت میلیارد پارامتر استفاده کرده‌اند.

یکی از روش‌های افزایش کیفیت خروجی مدل‌های زبانی و مقابله با توهّم، استفاده از Retrieval-Augmented Generation است.[۲۰] در این روش از یک پایگاه داده برداری[۹] برای بازیابی اطلاعات مرتبط با موضوع استفاده می‌شود. سپس با غنی کردن پرامپت[۱۰] ورودی با اطلاعات بازیابی شده، مدل خروجی باکیفیت‌تری تولید می‌کند. یک روش دیگر برای مقابله با توهّم، بررسی درستی محتوای تولید شده پس از استنتاج است. در این صورت تشخیص توهّم به یک مسئله‌ی صحت‌سنجی[۱۱] تبدیل می‌شود که در آن می‌توان از پایگاه داده، یا جستجوی اینترنتی استفاده کرد. همچنین روش‌های متعددی مبتنی بر گراف دانش برای استنتاج متکی بر دانش[۱۲] یا صحت‌سنجی خروجی معرفی شده‌اند.[۲۱] واضح است که استفاده از RAG یا گراف دانش در فرآیند استنتاج یا صحت‌سنجی اطلاعات پس از تولید خروجی، در همه‌ی شرایط امکان‌پذیر نمی‌باشد. برای استفاده از RAG به پایگاه داده شامل اطلاعات مرتبط پرسش نیاز است. صحت‌سنجی نیز فرآیندی پیچیده است. لذا نیازی اساسی به روش‌هایی برای تشخیص توهّم در مدل‌های زبانی بدون دانش خارجی وجود دارد. این امر تنها با درک علت توهّم در مدل‌های زبانی امکان‌پذیر است.

فرضیه‌های مختلفی برای عوامل ایجاد توهّم در مدل‌های زبانی ارائه شده است. می‌دانیم که کیفیت متون خروجی مدل‌های زبانی با حجم و کیفیت داده‌های آموزشی ارتباط مستقیم دارد. برخی از محققان نادرستی داده‌های آموزشی را عامل این مسئله می‌دانند.[۲۲] وابستگی درستی برخی از داده‌ها به زمان نیز در این مسئله تاثیر دارد زیرا دانش مدل از نظر زمانی توسط مجموعه‌داده‌ی آموزشی محدود می‌شود.

یکی دیگر از فرضیه‌های ارائه شده، روش پیش‌بینی کلمه به کلمه‌ی مدل‌های زبانی را عامل توهّم می‌داند. در [۲۳] محققان نشان می‌دهند مدل‌های زبانی در متون تولیدی خود گزاره‌های نادرستی را بیان می‌کنند که اگر به شکل جداگانه مورد پرسش قرار گیرند، مدل نادرستی آن‌ها را تشخیص می‌دهد. در مدل‌های زبانی کلمات ابتدایی تولید شده توسط مدل بر روی انتخاب کلمات بعدی تاثیر گذاشته و بدین ترتیب، زمینه[۱۳] می‌تواند عامل اشتباهی پس از اشتباه دیگر شود. در نتیجه زنجیره‌ای از توهّمات تولید می‌شود که [۲۳] آن را «بهمن توهّم»[۱۴] نامگذاری کرده است.

از دیگر فرضیه‌های مطرح شده، ارتباط میان تردید و عدم قطعیت مدل و توهّم است. مدل‌های زبانی با استفاده از حجم زیادی از متون آموزش می‌بینند تا کلمه‌ی بعدی را پیش‌بینی کنند. در این فرآیند مدل با تمامی اسامی و مفاهیم به شکل یکسان برخورد نخواهد کرد. به عنوان نمونه مدل در فرآیند آموزش بارها در متون مختلف با نام «سورن کی‌یرکگور»، به عنوان یک فیلسوف شهیر، برخورد داشته است. اگر ورودی مدل عبارت «سورن کی‌یرکگور _ » باشد، مدل به کلماتی مانند «فیلسوف»، «متفکر» و «الهی‌دان» مقادیر احتمال بالایی نسبت می‌دهد در حالی که کلماتی مانند «کارگردان» یا «بسکتبالیست» مقادیر احتمال اندکی خواهند داشت. با انتخاب کلمه‌ی فیلسوف و تبدیل عبارت به «سورن کی‌یرکگور فیلسوف _»، مدل گزینه‌های جدیدی با مقادیر احتمال بالا مانند «دانمارکی»، «پروتستان» و «اگزیستانسیالیست» برای تکمیل جمله خواهد داشت. لذا مدل با تردید اندک می‌تواند جمله را به پایان برساند. در مقابل اگر نام شخصی نه چندان شناخته شده از مدل پرسیده شود، مدل تعداد بسیار زیادی

[9] Vector Database
[10] Prompt
[11] Fact-checking
[12] Knowledge-Aware Inference
[13] Context
[14] Hallucination Snowballing

گزینه پیشنهادی با مقادیر احتمال اندک برای تکمیل جمله خواهد داشت. لذا عدم قطعیت مدل در پاسخ‌گویی سبب تولید خروجی متوهّمانه می‌شود. روش‌هایی مانند [۲۴] و [۲۵] با نمونه‌گیری و مقایسه‌ی معنایی جملات بر پایه‌ی همین فرضیه بنا شده‌اند. اگر مدل در پاسخ به سوال قطعیت بالایی داشته باشد، باید متونی با معنای نزدیک تولید کند.

۵-۲- پیشینه‌ی تحقیق:

مطالعات را می‌توان بسته به دسترسی به مقادیر درونی مدل به دو دسته‌ی جعبه‌سیاه[۲۶] ,[۲۵] ,[۲۴] و جعبه‌سفید[۲۸] ,[۲۷] تقسیم کرد. در مطالعات جعبه‌سیاه با نمونه‌گیری از مدل، متون تولیدی از نظر یکپارچگی معنایی مورد بررسی قرار می‌گیرند. در صورتی که پاسخ‌های مدل به یک سوال دارای تعارضات یا تناقضاتی با یکدیگر باشند، می‌توان نتیجه گرفت که محتوای تولیدی مدل برای آن پرامپت قابل اعتماد نمی‌باشد. در واقع تفاوت معنایی قابل توجه در پاسخ‌های تولیدی نشانه‌ای از عدم قطعیت مدل در پاسخگویی است. در دسته‌ی دوم مطالعات، هدف روش‌های جعبه سفید این است که با آموزش یک دسته‌بند بر روی مقادیر درونی، مانند مقادیر فعالسازی و توجه، الگوهای مربوط به قاطعیت مدل را بیابند. مطالعات جعبه‌خاکستری دسته‌ی دیگری از مطالعات هستند که بعضا به عنوان حالتی خاص از شرایط جعبه‌سیاه در نظر گرفته می‌شوند. در شرایط جعبه‌خاکستری همانند جعبه‌سیاه دسترسی به مقادیر درونی مدل زبانی امکان‌پذیر نمی‌باشد اما مقادیر احتمال توکن‌های تولید شده در دسترس هستند.

یکی از مقالات اصلی در مسئله‌ی تشخیص توهّم، SelfCheckGPT [۲۴] است. این مطالعه با معرفی یک مجموعه‌داده‌ی اختصاصی مبتنی بر مجموعه‌داده‌ی WikiBio و استفاده از پنج روش مختلف برای تشخیص توهّم به روش جعبه‌سیاه، به مقایسه‌ی روش‌ها با یکدیگر می‌پردازد. مجموعه‌داده‌ی معرفی شده شامل دویست‌وسی‌وهشت مورد پاسخ برچسب خورده‌ی GPT-۳ و بیست مورد نمونه برای هر پاسخ است. هر پاسخ جمله به جمله در سه کلاس «صحیح»، «غلط جزئی» و «غلط جدی» برچسب زده شده و بیست مورد نمونه دیگر نیز توسط GPT-۳ برای مقایسه تولید شده است. سپس از پنج روش متفاوت مبتنی بر NLI، n-gram ، Question Answering [۲۹]،BERTScore [۳۰] [8] و پرامپت برای تشخیص توهّم استفاده شده است. از میان آن‌ها دو روش پرامپت‌نویسی با استفاده از GPT-۳ و NLI با استفاده از یک مدل DeBERTa-v3-large [۳۱] تنظیم دقیق شده[۱۵]بهترین نتیجه را داشتند. علاوه بر پنج روش ذکر شده، SelfCheckGPT از محاسبه‌ی آنتروپی با استفاده از مقادیر احتمال توکن‌های خروجی نیز استفاده می‌کند اما نتیجه‌ی آن ضعیفتر از پنج روش دیگر است.

[۳۲] از یک مدل زبانی به عنوان پراکسی[۱۶]استفاده می‌کند تا مقادیر احتمال توکن‌های خروجی را محاسبه کند. در این مقاله با الهام از شیوه‌ی برخورد انسان با متن در هنگام صحت‌سنجی سه ایده مطرح شده است؛ ۱- تمرکز بر کلمات کلیدی حاوی اطلاعات مفید ۲- تمرکز بر کلمات قبلی با انتشار عدم قطعیت به وسیله‌ی مقادیر توجه و ۳- مشخص کردن نوع هر اسم خاص در قبل آن. در [۳۳] نیز بر اساس مشابهت ژاکارد و NLI روشی برای محاسبه عدم قطعیت مدل در شرایط جعبه‌سیاه معرفی شده است. در HaLoCHECK [۳۴] از معیار SummaC [۳۵] برای مقایسه‌ی یکپارچگی میان هر جفت نمونه متن تولیدی و از یک گراف دو بخشی[۱۷]برای مدلسازی و محاسبه‌ی امتیاز استفاده شده است.

در برخی مقالات از روش‌های مبتنی بر پرامپت‌نویسی استفاده شده است. [۲۶] با در نظر گرفتن یک مدل به عنوان تولید کننده‌ی محتوا و یک مدل دیگر به عنوان تحلیلگر، چارچوبی مبتنی بر پرامپت برای تشخیص تناقض در نمونه‌های متون معرفی می‌کند. به طور مشابه [۳۶] گزاره‌های نادرست را به وسیله‌ی مباحثه‌ی دو مدل زبانی کشف می‌کند. [۳۷] با تولید گزاره‌هایی مرتبط با متن خروجی، یک مدل گرافی احتمالاتی[۱۸]درخت مارکوف مخفی) می‌سازد و سپس با رویکردی بیزی[۱۹]نمره‌ی اعتماد برای گزاره محاسبه می‌کند. برای تولید گزاره‌های جدید از پرامپت‌نویسی بر اساس سه استراتژی استفاده می‌شود؛ شکستن به گزاره‌های ساده‌تر، تولید فرض‌های متناقض و سازگار، و پرسش از مدل برای تصحیح گزاره.

از دیگر روش‌های جعبه‌سیاه برای محاسبه عدم قطعیت، Semantic Entropy [۲۵] است. در این روش پاسخ به همراه نمونه‌های تولید شده توسط مدل از نظر معنایی خوشه‌بندی می‌شوند و سپس برای پاسخ‌ها آنتروپی محاسبه می‌گردد. اگر متون تولیدی طولانی باشند، مثلا زندگینامه یک شخص، ابتدا گزاره‌های درون متن استخراج شده و برای هر گزاره سوالی طرح می‌شود. سپس برای هر سوال نمونه‌گیری شده و نمونه‌ها خوشه‌بندی می‌شوند. در [۲۸] یک روش جعبه‌سفید به نام Semantic Entropy Probe بر اساس Semantic Entropy معرفی می‌شود که هدف آن تخمین مقدار

[1] Fine-tuned
[1] Proxy
[1] Bipartite graph
[1] Directed Graphical Model
[1] Bayesian

Semantic Entropy با استفاده از مقادیر لایه‌های پنهان است. برای این کار یک دسته‌بند بر روی مقادیر پنهان مدل آموزش داده می‌شود.
یکی دیگر از مهم‌ترین مقالات منتشر شده SAPLMA [۲۷] است که رویکردی جعبه‌سفید دارد. این مطالعه بر این ادعا استوار است که مقادیر درونی مدل زبانی می‌توانند باور مدل درباره‌ی درستی یک گزاره را آشکار کنند. این گزاره هم می‌تواند گزاره‌ی ورودی باشد و هم گزاره‌ی در حال تولید. بر این اساس محققان ابتدا یک مجموعه‌داده‌ی صحیح-غلط در شش موضوع متفاوت ساخته‌اند. سپس با استفاده از پنج موضوع مقادیر فعالسازی مدل در لایه‌های مختلف مدل، مانند شانزدهم، بیستم، بیست‌وچهارم و بیست‌وهشتم، را در هنگام استنتاج ثبت کرده‌اند. در ادامه بر اساس برچسب گزاره ورودی، روی مقادیر مربوط به هر لایه یک دسته‌بند آموزش داده‌اند. برای آزمون از موضوع باقی‌مانده استفاده شده است. دسته‌بند دارای یک معماری MLP با سه لایه مخفی و ابعاد کاهشی ۲۵۶، ۱۲۸ و ۶۴ بوده و ورودی آن مقادیر فعالسازی خروجی یک لایه از مدل زبانی است. بدین ترتیب می‌توان تقریبا همزمان با فرآیند استنتاج، به وسیله‌ی مقادیر درونی مدل، در مورد درستی متن تصمیم‌گیری کرد.
[۳۸] به روشی نزدیک به SAPLMA از تعبیه‌ی متنی زمینه‌سازی شده[۲۰] یا همان مقادیر درونی آخرین توکن در آخرین لایه برای آموزش یک دسته‌بند MLP به نام MIND استفاده کرده است. تفاوت اصلی MIND با SAPLMA در مجموعه‌داده‌ی مورد استفاده است. MIND از مجموعه‌داده‌ی اختصاصی خود به نام HELM برای ثبت مقادیر و آموزش دسته‌بند بهره برده است. در HELM به جای استفاده از گزاره‌های صحیح-غلط، پاسخ‌های مدل پس از تولید برچسب‌خورده‌اند و از آن‌ها برای آموزش دسته‌بند استفاده شده است.
در ITI [۳۹] نیز همانند [۲۷] یک مدل بر روی مقادیر درونی مدل آموزش داده می‌شود تا در صورت لزوم در متن تولیدی مدل مداخله صورت گیرد. تفاوت اصلی این مطالعه با SAPLMA در استفاده از رگرسیون لجستیک به جای معماری MLP است. مقاله‌ی ITI متعلق به یک مسیر تحقیقاتی موازی و نزدیک به مسئله‌ی ما است که هدف اصلی آن مداخله در فرآیند استنتاج جهت افزایش درستی[۲۱] خروجی است. از دیگر کارهای این مسیر می‌توان به [۴۰] اشاره کرد.
در [۴۱] مسئله‌ی تشخیص توهّم به عنوان یک مسئله‌ی ارضای محدودیت صورتبندی شده است. در این مطالعه نشان داده شده است که میان توجه[۲۲] به توکن‌های عامل محدودیت و درستی گزاره تولید شده، همبستگی وجود دارد. بر این اساس یک دسته‌بند به نام SAT Probe جهت کشف توهّم با استفاده از الگوهای توجه آموزش داده می‌شود.
[۴۲] روشی به نام PoLLMgraph را معرفی می‌کند. هر توکن در دنباله‌ی توکن‌های خروجی مدل با عبور از تعداد محدودی وضعیت درونی تولید می‌شود که [42] آن را «رَد» می‌نامد. PoLLMgraph با استفاده از یک مجموعه‌داده‌ی مرجع و رد انتقالات درونی مدل، پس از اعمال تبدیل PCA، یک مدل مارکوف مخفی و یک مدل GMM[۲۳] برای تشخیص توهّم آموزش می‌دهد.

جدول شماره‌ی ۲: پیشینه‌ی تحقیق

تکنیک کلی	نام	تاریخ انتشار	مجموعه‌داده	مدل زبانی*	مرجع
جعبه‌سیاه و جعبه‌خاکستری	SelfCheckGPT	۲۰۲۳/۱۱	Wikibio-GPT-۳	GPT-۳	[۲۴]
	EXAMINER	۲۰۲۳/۱۲	LAMA، TriviaQA، PopQA، Natural Questions (NQ)	GPT–۳، GPT–۳٫۵، LLaMa – (۷)	[۳۶]
	Stronger Focus	۲۰۲۳/۱۲	WikiBio-GPT-۳، XSumFaith، FRANK	۲۲ مدل متفاوت تا ۷۰ میلیارد پارامتر	[۳۲]
	HaLoCHECK	۲۰۲۳ (*preprint*)	اختصاصی	BLOOM – (۷)	[۳۴]

[0] Contextualized Embedding
[1] Truthfulness
[2] Attention
[3] Gaussian Mixture Model

[۲۶]	GPT-۴، GPT-۳٫۵، LLaMa-۲ – Chat – (۷۰)، Vicuna – (۱۳)	اختصاصی - بر پایه‌ی Wikipedia	۲۰۲۴/۰۱	ChatProtect		
[۳۳]	GPT-۳٫۵، OPT – (۱۳)، LLaMa – (۱۳)، LLaMa-۲ – (۱۳)	TriviaQA، CoQA، Natural Questions (QA)	۲۰۲۴/۰۵	UQ-NLG		
[۲۵]	LLaMa-۲ – Chat – (۷، ۱۳، ۷۰)، Falcon – Instruct – (۷۴۰)، Mistral – Instruct – (۷)	SQuAD، TriviaQA، FactualBio، BioASQ، SVAMP، Natural Questions (NQ) - Open	۲۰۲۴/۰۶	Semantic Entropy		
[۳۷]	GPT-۳٫۵، LLaMa-۳ – Instruct – (۸)	Wikibio-GPT-۳، FELM–Science، FactCheckGPT	۲۰۲۴ (*preprint*)	BTProp		
[۳۹]	LLaMa – (۷)	TruthfulQA	۲۰۲۳/۰۹	ITI	جعبه‌سفید	
[۲۷]	OPT – (۶٫۷)، LLaMa-2 – (۷)	اختصاصی	۲۰۲۳/۱۲	SAPLMA		
[۴۱]	LLaMa-۲ – (۷، ۱۳، ۷۰)	اختصاصی - بر پایه‌ی WikiData و CounterFact	۲۰۲۴/۰۱	SAT Probe		
[۴۲]	Alpaca – (۱۳)، Vicuna – (۱۳)، T۵ – (۱۱)، LLaMa – (۱۳)، LLaMa-۲ – (۱۳)	HaluEval، TruthfulQA	۲۰۲۴/۰۶	PoLLMgraph		
[۳۸]	LLaMa-۲ – Chat – (۷، ۱۳)، LLaMa-۲ – Base – (۷)، Falcon – (۴۰)، OPT – (۶٫۷)، GPT-J – (۶)	HELM	۲۰۲۴/۰۸	MIND		
[۲۸]	GPT-۳٫۵، LLaMa-۳ – (۷۰)، LLaMa-۲ – (۷، ۷۰)، Mistral – (۷)	BioASQ، TriviaQA، Natural Questions (NQ) – Open، SQuAD	۲۰۲۴ (*preprint*)	Semantic Entropy Probes		

				Phi-۳-Mini – (۳.۸)	

*مدل‌های زبانی از چپ به راست به فرم «نام مدل – نوع – (تعداد پارامتر به میلیارد)» نوشته شده‌اند.

۵-۳- روش ها و فنون اجرایی طرح:

اصول و فنون پردازش زبان طبیعی:

مطالعه بر روی مسئله‌ی مطرح شده بدون تسلط کافی بر روی اصول و فنون پردازش زبان طبیعی و به خصوص روش‌های مبتنی بر یادگیری ماشین امکان‌پذیر نمی‌باشد.

معماری ترنسفورمر و مکانیزم توجه:

تقریبا تمامی مدل‌های زبانی فعلی بر پایه‌ی معماری ترنسفورمر هستند. درک این معماری یادگیری عمیق و پایه اساسی آن، مکانیزم توجه، دارای اهمیت است. مکانیزم توجه به مدل امکان می‌دهد بر روی کلمات کلیدی تمرکز کرده و با طولانی شدن متن، از اهمیت اطلاعات مهم کاسته نشود.

تفسیرپذیری مکانیستیک:

تفسیرپذیری مکانیستیک یک حوزه‌ی تحقیقاتی است که بر درک سازوکارهای درونی مدل‌های زبانی تمرکز دارد. محققان تفسیرپذیری با استفاده از مدل‌های زبانی کوچک، به دنبال مهندسی معکوس و یافتن دلایل رفتارهای این مدل‌ها هستند. درک شفاف و تفسیرپذیر عملکرد مدل‌های زبانی می‌تواند در حل مشکلات آن‌ها، از جمله توهّم، مفید باشد.

چارچوب‌های نرم‌افزاری یادگیری عمیق:

در این مطالعه از چارچوب‌های نرم‌افزاری مانند Pytorch جهت آموزش دسته‌بند و ارزیابی مدل‌هایی زبانی استفاده می‌شود.

روش‌های ارزیابی:

برای ارزیابی روش ارائه شده از مجموعه‌داده‌های TruthfulQA [۴۳] و HaluEval [۴۴] استفاده خواهد شد. TruthfulQA یک مجموعه‌داده شامل هشتصدوهفده سوال است که برای ارزیابی عمومی مدل‌های زبانی مورد استفاده قرار می‌گیرد. HaluEval یک بنچمارک مختص ارزیابی توهّم در مدل‌های زبانی است. این مجموعه‌داده دارای سی‌وپنج هزار نمونه در چهار بخش QA، Summarization، Dialogue و General بوده که پنج هزار نمونه‌ی بخش General برچسب‌گذاری Yes/No شده‌اند و در سه بخش دیگر از یک مدل زبانی پیشرفته‌تر، مانند ChatGPT، برای داوری استفاده می‌شود. با توجه به اینکه تمرکز ما در این مطالعه بر روی پرسش و پاسخ است، ما از بخش QA شامل ده هزار نمونه استفاده خواهیم کرد.
همان‌گونه که گفته شد، اغلب روش‌های پایه از مدل‌های زبانی با بیش از هفت میلیارد پارامتر استفاده کرده‌اند. لذا لازم است برای مقایسه‌ی عادلانه، نتایج روش‌های پایه را بر روی مدل زبانی کوچک مورد نظر بازتولید کرد. شاخص مورد استفاده AUROC خواهد بود.

۶- نوآوری تحقیق:

ترکیب روش‌های جعبه‌سیاه و جعبه‌سفید به وسیله‌ی تابع امتیاز:

همان گونه که اشاره شد روش‌های ارائه شده برای تشخیص توهّم، عموما یا متکی بر نمونه‌گیری هستند و یا متکی بر تشخیص تردید مدل با تکیه بر مقادیر درونی. یکی از مشکلات اساسی در روش‌های جعبه‌سیاه که بر پایه‌ی نمونه‌گیری کار می‌کنند، عدم در نظر گرفتن میزان اعتمادپذیری نمونه‌ها هستند. می‌توان با ادغام تکنیک‌های جعبه‌سفید در هنگام نمونه‌گیری، میزان قطعیت مدل در تولید این نمونه‌ها را محاسبه کرد. سپس با استفاده از

یک تابع امتیازدهی، نتیجه‌ی نهایی را به دست آورد. انتظار می‌رود با ترکیب روش‌های جعبه‌سفید و جعبه‌سیاه نتایج بهتری حاصل شود.

کاهش تعداد استنتاج در روش NLI:

در روش‌های مبتنی بر مدل NLI مانند SelfCheckGPT – NLI [۲۴] به تعداد نمونه‌ها استنتاج انجام می‌شود. می‌توان با محاسبه‌ی میزان عدم قطعیت مدل در هنگام تولید نمونه‌ها، تعدادی از نمونه‌ها که با بیشترین قطعیت تولید شده‌اند را برای استنتاج انتخاب کرد.

کاهش ابعاد دادگان:

با توجه به اینکه مقادیر لایه‌های درونی مدل‌های زبانی ابعاد بالایی دارند، کاهش ابعاد به وسیله‌ی معماری خودرمزنگار[۴]ممکن است منجر به بهبود نتایج شود.

ادغام مقادیر درونی تمام توکن‌های خروجی:

در روش‌هایی مانند [۳۸] ,[۲۷] تنها از مقادیر درونی مدل مربوط به آخرین توکن استفاده می‌شود، در حالی که ممکن است مقادیر درونی توکن‌های قبلی نیز حامل اطلاعات مفیدی باشند. می‌توان مقادیر درونی برای تمام یا تعدادی از توکن‌های تولیدی را به وسیله‌ی یک خودرمزنگار بازگشتی یا شبکه عصبی بازگشتی[۵] با یکدیگر ترکیب کرد.

استفاده از ماشین بردار پشتیبان:

در مطالعاتی مانند [۳۸] ,[۲۷] از یک پرسپترون چند لایه (MLP) برای معماری مدل دسته‌بند استفاده می‌شود. در شرایطی که ابعاد داده‌ها بالا و تعدادشان کم است، انتظار می‌رود ماشین بردار پشتیبان عملکرد بهتری داشته باشد.

۷- جدول زمانبندی انجام طرح:

شروع از فروردین ۱۴۰۳

زمان (ماه) / فعالیت	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲
بازتولید نتایج روش‌های پایه	■	■										
آماده‌سازی دادگان		■										
پیاده‌سازی خودرمزنگار جهت کاهش ابعاد دادگان و ماشین بردار پشتیبان جهت دسته‌بندی			■									
تغییر هایپرپارامترها، لایه‌های مورد استفاده و ابعاد دادگان			■	■	■							
ارزیابی و مستندسازی نتایج اولیه و مقایسه با روش‌های پایه				■	■							
پیاده‌سازی ترکیب روش جعبه‌سفید با جعبه‌سیاه (تاثیر دادن مقایسه‌ی معنایی نمونه‌ها)						■	■	■	■			
ارزیابی نهایی نتایج										■		
مستندسازی نتایج و تهیه مقاله											■	■

[2] Autoencoder
[2] Recurrent neural network
[4]
[5]

مراجع

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners".

[3] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Oct. 06, 2024. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[4] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A Neural Probabilistic Language Model," *Journal of Machine Learning Research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[5] B. Millidge, "LLMs confabulate not hallucinate." Accessed: Oct. 27, 2024. [Online]. Available: http://www.beren.io/2023-03-19-LLMs-confabulate-not-hallucinate/

[6] J. Ferrando, G. Sarti, A. Bisazza, and M. R. Costa-jussà, "A Primer on the Inner Workings of Transformer-based Language Models," Oct. 13, 2024, *arXiv*: arXiv:2405.00208. doi: 10.48550/arXiv.2405.00208.

[7] V. Rawte *et al.*, "The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2541–2573. doi: 10.18653/v1/2023.emnlp-main.155.

[8] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On Faithfulness and Factuality in Abstractive Summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 1906–1919. doi: 10.18653/v1/2020.acl-main.173.

[9] L. Huang *et al.*, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," Nov. 09, 2023, *arXiv*: arXiv:2311.05232. doi: 10.48550/arXiv.2311.05232.

[10] Y. Zhang *et al.*, "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," Sep. 24, 2023, *arXiv*: arXiv:2309.01219. doi: 10.48550/arXiv.2309.01219.

[11] K. Benkirane *et al.*, "Machine Translation Hallucination Detection for Low and High Resource Languages using Large Language Models," Jul. 25, 2024, *arXiv*: arXiv:2407.16470. doi: 10.48550/arXiv.2407.16470.

[12] Z. Lu *et al.*, "Small Language Models: Survey, Measurements, and Insights," Sep. 24, 2024, *arXiv*: arXiv:2409.15790. doi: 10.48550/arXiv.2409.15790.

[13] N. Muennighoff *et al.*, "Crosslingual Generalization through Multitask Finetuning," May 29, 2023, *arXiv*: arXiv:2211.01786. doi: 10.48550/arXiv.2211.01786.

[14] S. Biderman *et al.*, "Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling," in *Proceedings of the 40th International Conference on Machine Learning*, PMLR, Jul. 2023, pp. 2397–2430. Accessed: Dec. 11, 2024. [Online]. Available: https://proceedings.mlr.press/v202/biderman23a.html

[15] Z. Liu *et al.*, "MobileLLM: Optimizing Sub-billion Parameter Language Models for On-Device Use Cases," in *Proceedings of the 41st International Conference on Machine Learning*, PMLR, Jul. 2024, pp. 32431–32454. Accessed: Dec. 11, 2024. [Online]. Available: https://proceedings.mlr.press/v235/liu24ce.html

[16] M. Abdin *et al.*, "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone," Aug. 30, 2024, *arXiv*: arXiv:2404.14219. doi: 10.48550/arXiv.2404.14219.

[17] G. Team *et al.*, "Gemma 2: Improving Open Language Models at a Practical Size," Oct. 02, 2024, *arXiv*: arXiv:2408.00118. doi: 10.48550/arXiv.2408.00118.

[18] "Llama 3.2: Revolutionizing edge AI and vision with open, customizable models," Meta AI. Accessed:

Dec. 11, 2024. [Online]. Available: https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/
[19] L. Chen and G. Varoquaux, “What is the Role of Small Models in the LLM Era: A Survey,” Sep. 30, 2024, *arXiv*: arXiv:2409.06857. doi: 10.48550/arXiv.2409.06857.
[20] P. Lewis *et al.*, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, in NIPS ’20. Red Hook, NY, USA: Curran Associates Inc., Dec. 2020, pp. 9459–9474.
[21] G. Agrawal, T. Kumarage, Z. Alghamdi, and H. Liu, “Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey,” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds., Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 3947–3960. doi: 10.18653/v1/2024.naacl-long.219.
[22] Z. Ji *et al.*, “Survey of Hallucination in Natural Language Generation,” *ACM Comput. Surv.*, vol. 55, no. 12, p. 248:1-248:38, Mar. 2023, doi: 10.1145/3571730.
[23] M. Zhang, O. Press, W. Merrill, A. Liu, and N. A. Smith, “How Language Model Hallucinations Can Snowball,” May 22, 2023, *arXiv*: arXiv:2305.13534. doi: 10.48550/arXiv.2305.13534.
[24] P. Manakul, A. Liusie, and M. Gales, “SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 9004–9017. doi: 10.18653/v1/2023.emnlp-main.557.
[25] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, “Detecting hallucinations in large language models using semantic entropy,” *Nature*, vol. 630, no. 8017, pp. 625–630, Jun. 2024, doi: 10.1038/s41586-024-07421-0.
[26] N. Mündler, J. He, S. Jenko, and M. Vechev, “Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation,” presented at the The Twelfth International Conference on Learning Representations, Oct. 2023. Accessed: Oct. 20, 2024. [Online]. Available: https://openreview.net/forum?id=EmQSOi1X2f
[27] A. Azaria and T. Mitchell, “The Internal State of an LLM Knows When It’s Lying,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 967–976. doi: 10.18653/v1/2023.findings-emnlp.68.
[28] J. Kossen, J. Han, M. Razzak, L. Schut, S. Malik, and Y. Gal, “Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs,” Jun. 22, 2024, *arXiv*: arXiv:2406.15927. doi: 10.48550/arXiv.2406.15927.
[29] P. Manakul, A. Liusie, and M. Gales, “MQAG: Multiple-choice Question Answering and Generation for Assessing Information Consistency in Summarization,” in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, and A. A. Krisnadhi, Eds., Nusa Dua, Bali: Association for Computational Linguistics, Nov. 2023, pp. 39–53. doi: 10.18653/v1/2023.ijcnlp-main.4.
[30] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT,” presented at the Eighth International Conference on Learning Representations, Apr. 2020. Accessed: Dec. 11, 2024. [Online]. Available: https://iclr.cc/virtual_2020/poster_SkeHuCVFDr.html
[31] P. He, J. Gao, and W. Chen, “DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing,” presented at the The Eleventh International Conference on Learning Representations, Sep. 2022. Accessed: Dec. 11, 2024. [Online]. Available: https://openreview.net/forum?id=sE7-XhLxHA
[32] T. Zhang *et al.*, “Enhancing Uncertainty-Based Hallucination Detection with Stronger Focus,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 915–932. doi: 10.18653/v1/2023.emnlp-main.58.
[33] Z. Lin, S. Trivedi, and J. Sun, “Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models,” *Transactions on Machine Learning Research*, Feb. 2024, Accessed: Oct. 24, 2024.

[Online]. Available: https://openreview.net/forum?id=DWkJCSxKU5
[34] M. Elaraby *et al.*, "Halo: Estimation and Reduction of Hallucinations in Open-Source Weak Large Language Models," Sep. 13, 2023, *arXiv*: arXiv:2308.11764. doi: 10.48550/arXiv.2308.11764.
[35] P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, "SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 163–177, Feb. 2022, doi: 10.1162/tacl_a_00453.
[36] R. Cohen, M. Hamri, M. Geva, and A. Globerson, "LM vs LM: Detecting Factual Errors via Cross Examination," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 12621–12640. doi: 10.18653/v1/2023.emnlp-main.778.
[37] B. Hou, Y. Zhang, J. Andreas, and S. Chang, "A Probabilistic Framework for LLM Hallucination Detection via Belief Tree Propagation," Jun. 11, 2024, *arXiv*: arXiv:2406.06950. doi: 10.48550/arXiv.2406.06950.
[38] W. Su *et al.*, "Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models," in *Findings of the Association for Computational Linguistics ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 14379–14391. doi: 10.18653/v1/2024.findings-acl.854.
[39] K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg, "Inference-Time Intervention: Eliciting Truthful Answers from a Language Model," presented at the Thirty-seventh Conference on Neural Information Processing Systems, Nov. 2023. Accessed: Oct. 24, 2024. [Online]. Available: https://openreview.net/forum?id=aLLuYpn83y
[40] S. Zhang, T. Yu, and Y. Feng, "TruthX: Alleviating Hallucinations by Editing Large Language Models in Truthful Space," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 8908–8949. doi: 10.18653/v1/2024.acl-long.483.
[41] M. Yuksekgonul *et al.*, "Attention Satisfies: A Constraint-Satisfaction Lens on Factual Errors of Language Models," presented at the The Twelfth International Conference on Learning Representations, Oct. 2023. Accessed: Oct. 21, 2024. [Online]. Available: https://openreview.net/forum?id=gfFVATffPd
[42] D. Zhu *et al.*, "PoLLMgraph: Unraveling Hallucinations in Large Language Models via State Transition Dynamics," in *Findings of the Association for Computational Linguistics: NAACL 2024*, K. Duh, H. Gomez, and S. Bethard, Eds., Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 4737–4751. doi: 10.18653/v1/2024.findings-naacl.294.
[43] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3214–3252. doi: 10.18653/v1/2022.acl-long.229.
[44] J. Li, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen, "HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6449–6464. doi: 10.18653/v1/2023.emnlp-main.397.