ICWR2024
April 24-25, 2024, Tehran, Iran
10th International Conference on Web Research

Detecting Hallucinations Generated by Large Language Models Using Paraphrasing Technique

Tara Zare, Mehrnoush Shamsfard

Shahid Beheshti University, Tehran, Iran
ta.zare@sbu.ac.ir, m-shams@sbu.ac.ir

Abstract

Hallucination in large language models refers to outputs that appear correct but contradict reality or diverge from the source. Detecting hallucination in large language models is crucial to prevent the dissemination of these hallucinations in applications directly or indirectly related to such models. In this study, we have employed a simple algorithm to detect hallucination in a large language model. Our hypothesis is based on the hypothesis that if a large language model responds to the paraphrases of a question and an inconsistency is discovered among its answers, then we say that it is hallucination, and if the answers are consistent, it likely provides a correct answer. We have checked and confirmed these two hypotheses with experiments. In this way, our proposed method to discover the hallucination in answering a question is to create different paraphrases of that question and check the existence of inconsistencies or contradictions in the answers given to the generated questions. The presence or absence of inconsistency confirms the presence or absence of hallucinations. Experiments show that this method is able to detect hallucination in answering questions with high accuracy.

Keywords: Large Language Models, Hallucination of Large Language Models, Inconsistency Detection, Paraphrasing

تشخیص توهم مدل‌های زبانی بزرگ به کمک روش دگرنویسی

تارا زارع[1] ، مهرنوش شمس‌فرد[2]*

دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، تهران،
ta.zare@sbu.ac.ir

دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، تهران،
m-shams@sbu.ac.ir

چکیده

توهم در مدل‌های زبانی بزرگ به خروجی‌های در ظاهر صحیح اما در باطن برخلاف واقعیت یا عدم وفادار به منبع اطلاق می‌گردد. تشخیص توهم در مدل‌های زبانی بزرگ به جهت جلوگیری از انتشار این توهم‌ها در کاربردهایی که به طور مستقیم یا غیرمستقیم با مدل‌های زبانی بزرگ ارتباط دارند، اهمیت دارد. در این پژوهش از الگوریتم ساده‌ای جهت تشخیص متوهم بودن یک مدل زبانی بزرگ استفاده کرده‌ایم. فرضیه ما بر این اساس است که اگر مدل زبانی بزرگ به دگرنویسی‌های یک پرسش پاسخ دهد و در میان پاسخ‌های آن تناقضی کشف شود آن گاه گوییم دچار توهم شده است و اگر پاسخ‌ها سازگار باشند، به احتمال بالایی پاسخ درستی می‌دهد. این دو فرضیه را با آزمایش‌هایی بررسی و تأیید کرده‌ایم. به این ترتیب روش پیشنهادی ما برای کشف توهم در پاسخ به یک پرسش، ایجاد دگرنویسی‌های مختلف آن پرسش و بررسی وجود ناسازگاری یا تناقض در پاسخ‌های داده شده به پرسش‌های تولید شده است. وجود یا عدم ناسازگاری، وجود یا عدم توهم را تأیید می‌کند. آزمایشات نشان می‌دهند این روش با دقت بالایی قادر به کشف توهم در پاسخ به سؤالات است.

کلمات کلیدی

مدل‌های زبانی بزرگ، توهم در مدل‌های زبانی بزرگ، کشف ناسازگاری، دگرنویسی

۱- مقدمه

مدل‌های زبانی بزرگ[3] مدل‌های هستند که جهت مدل‌سازی زبان‌های انسانی با استفاده از یادگیری خودنظارت [1] و با استفاده از روش‌های آموزشی ویژه در راستای بهبود عملکرد وظایف دیده نشده[4] و پیروی بهتر از دستورالعمل‌های زبان طبیعی آموزش دیده‌اند [2].

از کاربردهای این مدل‌های می‌توان به وظایف گوناگونی از جمله پرسش و پاسخ[5] [3]، تولید کد [4] و حل مسائل ریاضی [5] اشاره نمود.

مدل‌های زبانی بزرگ در کنار تمام نقاط قوت و کاربردهای مفیدی که دارند، دارای ضعف‌های جدی و قابل تأملی در زمینه تولید محتوای به ظاهر درست و روان اما در باطن نادرست هستند. این محتوای به ظاهر درست اما در مفهوم نادرست، توهم[6] نامیده می‌شود. علت این نام‌گذاری به تعریف واژه «توهم» در علم روان‌شناسی باز می‌گردد. بلوم [6] توهم را به عنوان «ادراکی که توسط یک فرد بیدار در غیاب محرک مناسب در دنیای خارج از بدن او تجربه می‌شود» تعریف می‌کند. یا به عبارت ساده‌تر توهم یک ادراک غیر واقعی است که به نظر واقعی می‌رسد. تولید اطلاعات نادرست، بی‌معنی و یا عدم‌وفادار به منبع ارائه‌شده توسط مدل‌های زبانی ویژگی مشابهی با چنین توهمات روان‌شناختی دارد. همان‌طور که یک انسان متوهم متوجّه نیست که دچار توهم شده است و وقایعی که بر او می‌گذرد مانند یک واقعیت در ذهن او ثبت می‌شوند، یک مدل زبانی بزرگ نیز با قصد و آگاهی قبلی، اطلاعات نادرست تولید نمی‌کند. در واقع محتوای این خروجی‌های نادرست در قسمت آموزش مدل ظاهر نشده‌اند و اختراع خود مدل است. حال با توجّه به پیشرفت سریع مدل‌های زبانی بزرگ، نگرانی قابل توجّهی نسبت به تمایل مدل‌های زبانی بزرگ در تولید توهم وجود دارد که سبب ایجاد محتوای به ظاهر قبول و روان اما بدون منبع درست می‌شوند [7].

این تعریف از توهم با تحقیق‌های دیگر که توهم را محتوای تولید شـــده بدون معنی یا عدم وفادار به ورودی ارائه‌شده دانستند، هم‌راستا است [8].

توهم موجود در خروجی مدل‌های زبانی بزرگ را به دو نوع خروجی غیر واقعی و عدم وفاداری به منبع تقسیم‌بندی می‌کنند [9]. توهم غیرواقعی بر ناهماهنگی بین محتوای تولید شده و حقایق قابل تأیید در دنیای واقعی دلالت دارد. به عنوان مثال اگر از مدل زبانی پرسیده شود که «اولین رئیس جمهور زن ایران چه کسی است؟» و او در خروجی هر نامی تولید کند، توهم از نوع خروجی غیرواقعی[۷] داریم. در عدم وفاداری به منبع[۸]، به واگرایی محتوای تولید شده از دستورالعمل‌های کاربر یا زمینه ارائه شده توسط ورودی و همچین عدم سازگاری با محتوای تولید شده اشاره داریم. در راستای مثال این نوع از توهم فرض کنید که متنی به زبان فارسی یا زبان دیگر را به مدل زبانی بزرگ داده‌ایم و خواسته‌ایم که آن را به زبان دیگری ترجمه کند. در صورت وجود هر گونه اشتباه در خروجی این درخواست، شاهد عدم وفاداری به منبع بوده‌ایم.

البته در منابع دیگر، اصطلاح‌های دیگری جایگزین خروجی غیرواقعی و عدم وفادار به منبع، به کار برده شده است. به عنوان مثال برخی منابع [10] توهم‌های ناشی از مدل‌های زبانی بزرگ را به دو دسته دامنه باز[۹] و دامنه بسته[۱۰] دسته‌بندی کرده‌اند. توهم‌های دامنه باز در واقع ادعاهایی غیرواقعی در ارتباط با جهان پیرامون ما هستند که توسط مدل‌های بزرگ زبان تولید شده‌اند. توهم‌های دامنه بسته نیز شامل دور شدن خروجی مدل زبانی از متن یک منبع یا مرجع خاص است. به عنوان مثال اگر یک مدل زبانی در خلاصه‌سازی یک متن دچار اشتباه شود و اطلاعاتی مغایر و ناسازگار با اطلاعات منبع ورودی تولید کند آن گاه این اطلاعات نادرست در دسته توهم‌های دامنه بسته قرار می‌گیرد.

با توجّه به کاربردهای مختلف و گسترده مدل‌های زبانی بزرگ در زمینه‌های گوناگون، شاهد رشد روزافزون در به کارگیری این مدل‌ها در کاربردهای[۱۱] متنوع هستیم. واضح است که اگر خروجی این مدل‌ها مستعد تولید توهم باشند، این خطا در طول این کاربردها منتشر شده و اعتماد کاربر را نسبت به خود خدشه‌دار خواهد کرد. در نتیجه در راستای بهبود سیستم‌ها یا سامانه‌های گفت‌وگوگر[۱۲] که به طور مستقیم یا غیرمستقیم از این مدل‌ها کمک می‌گیرند، بهتر است که این خطاها را در مرحله خروجی مدل‌های زبانی بزرگ تشخیص داده و از ورود آن‌ها به کاربردهای دیگر جلوگیری کنیم.

از اهمیت تشخیص این توهم‌ها می‌توان در درجه اول به عدم از دست دادن اعتماد کاربران اشاره کرد. همچنین اگر مانع ورود این اطلاعات نادرست به سیستم‌های دیگر نشویم می‌تواند موجب بروز خطاهای بزرگ‌تری در زمینه‌های حساس‌تری مانند پزشکی و درمان شود. فرض کنید یک مدل زبانی در ترجمه ماشینی دستورالعمل‌های دارویی یک بیمار دچار توهم شده و اطلاعاتی مغایر با منبع تولید کند. این توهم‌های تولید شده توسط مدل زبانی بزرگ در صورت عمل کردن بیمار به آن دستورالعمل‌ها، نقش یک تهدید جانی را برای او خواهد داشت.

تا به این جا به اهمیت تشخیص این توهم‌ها و جلوگیری از ورود آن‌ها به سیستم‌های دیگر پی برده‌ایم. حال به دنبال روشی جهت تشخیص توهم‌آمیز یا غیرتوهم‌آمیز بودن خروجی مدل‌های زبانی بزرگ هستیم. در این پژوهش ما این وظیفه را به کمک روش دگرنویسی ورودی مدل‌های زبانی بزرگ انجام داده‌ایم. در ادامه به ترتیب به پژوهش‌های مرتبط، روش پیشنهادی، آزمایش و نتایج، جمع‌بندی و کارهای آتی را خواهیم دید.

۲- پژوهش‌های مرتبط

تحقیقاتی که در زمینه توهم‌های تولید شده توسط مدل‌های زبانی بزرگ انجام شده است به چند بخش ایجاد زیرساخت ارزیابی[۱۳]، تشخیص توهم و کاهش توهم‌های مدل‌های زبانی بزرگ تقسیم‌بندی می‌شوند. در این قسمت در ارتباط با بخش‌های زیرساخت ارزیابی و تشخیص توهم، پژوهش‌هایی را مرور خواهیم کرد.

در بخش زیرساخت ارزیابی، مجموعه داده TruthfulQA در سال ۲۰۲۲ با هدف ارزیابی خروجی مدل‌های زبانی بزرگ تولید شد [11]. این مجموعه شامل ۸۱۷ پرسش مختلف در ۳۷ دسته‌بندی متفاوت مانند بهداشت، قانون، امور مالی، سیاست و.... است. از این ۸۱۷ پرسش، ۳۸۰ سؤال غیرخصمانه و ۴۳۷ پرسش فیلترشده و خصمانه هستند؛ واژه خصمانه در این جا بدان معنی است که این پرسش‌ها با هدف فریب دادن مدل‌های زبانی بزرگ طراحی شده‌اند و به دنبال به چالش کشیدن خروجی‌های مدل‌های زبانی هستند. یکی دیگر از زیرساخت‌های ارزیابی که در سال ۲۰۲۳ معرفی شده است SelfCheckGPT-Wikibio است [12]. در مراحل ساخت آن، از مجموعه داده WikiBio [13] که شامل اطلاعات ویکی‌پدیا در ارتباط با زندگی‌نامه افراد است، استفاده شده است. ۲۳۸ مقاله نسبتاً طولانی را انتخاب کرده و به کمک GPT-3 عباراتی به سبک ویکی‌پدیا در همان زمینه‌های انتخاب‌شده، تولید کردند. این عبارات تولید شده به صورت دستی حاشیه‌نویسی شدند. سه برچسب کاملاً نادرست، نادرست جزئی و درست برای آن‌ها در نظر گرفته شده است. در نهایت ۱۹۰۸ جمله به دست آمده است که یک مجموعه داده مفید جهت ارزیابی توانایی یک مدل در تشخیص توهم‌های مدل‌های زبانی بزرگ را تشکیل داده‌اند. همچنین آخرین زیرساخت ارزیابی مشهوری که در سال ۲۰۲۳ منتشر شده است HaluEval نام دارد که هدف آن ارزیابی تمایل مدل‌های زبانی بزرگ مانند ChatGPT در راستای تولید توهم است [14]. این مجموعه داده شامل مجموعه گسترده‌ای از نمونه های تولید شده به طور خودکار یا به کمک نیروی انسانی است. در مجموع شامل ۳۵۰۰۰ نمونه که شامل ۵۰۰۰ پرسش عمومی با پاسخ‌های ChatGPT و ۳۰۰۰۰ پرسش در زمینه وظایف‌خاص‌تر مانند پرسش و پاسخ، خلاصه‌سازی متن، گفت و گو مبتنی بر دانش و... است.

در مرور پژوهش‌های انجام شده در زمینه تشخیص توهم مدل‌های زبانی بزرگ می‌توان به معماری SCALE[۱۴] که بر روی تشخیص توهم در متون طولانی‌تر تأکید دارد [15]، اشاره کرد. در این رویکرد با تجزیه سندها به قطعه‌های کوچک‌تر و قابل مدیریت‌تر و سپس به کمک یک مدل استنتاج زبان طبیعی[۱۵] جهت تشخیص ناسازگاری در هر بخش به یافتن توهم در هر بخش می‌پردازد. در مطالعه دیگری که انجام شده است از مجموعه داده SelfCheckGPT-wikibio [12] که در ارتباط با آن صحبت کردیم استفاده می‌کند و هر دستور را به تعداد N بار به مدل زبانی بزرگ می‌دهد تا N خروجی تولید کند [12]. سپس به کمک روش‌های مختلفی مانند BERTscore و یا N-gramها به بررسی این N پاسخ می‌پردازد. توجّه داشته باشید که N دستور اولیه همه یک دستور ثابت و غیرمتغیر بوده‌اند.

۳- روش پیشنهادی

با توجّه به بررسی‌هایی که بر روی مدل‌های زبانی بزرگ داشته‌ایم، مشاهده کردیم که در بعضی مواقع با دگرنویسی ورودی مدل شاهد پاسخی متفاوت خواهیم بود. گاهی مواقع این پاسخ‌های متفاوت با یکدیگر در تناقض بوده و ناسازگار هستند. با مشاهده این حالات برآن شدیم که روشی برای کشف این تناقض و در نتیجه تشخیص توهم به دست آوریم.

در این پژوهش اساس کار ما این دو فرضیه زیر است که سعی در اثبات‌شان داریم:

- فرضیه اول: اگر دگرنویسی‌های مختلف یک پرسش را به مدل زبانی بزرگ داده و پاسخ‌های آن با یکدیگر سازگار باشند، بدان معنی است که پاسخ مدل زبانی صحیح است و توهم نداریم؟
- فرضیه دوم: اگر دگرنویسی‌های مختلف یک پرسش را به عنوان چند دستور[۱۶] به مدل زبانی بزرگ بدهیم و در پاسخ‌های تولید شده تناقض و ناسازگاری دیده شود، آن گاه به معنای این است که توهم داریم و به این ترتیب می‌توانیم جهت تشخیص توهم، از عدم سازگاری بین پاسخ‌های دگرنویسی‌های یک پرسش استفاده کنیم.

معماری پیشنهادی در شکل (۱) نشان داده شده است. همان‌طور که در شکل دیده می‌شود در این روش ابتدا یک مجموعه‌ای از پرسش‌ها را از بانک سؤالات خود استخراج کرده و برای هر یک از پرسش‌ها، ۴ دگرنویسی دیگر تولید می‌کنیم. سپس این سؤالات تولید شده را به همراه پرسش اولیه به مدل زبانی بزرگ خود داده و خروجی را دریافت می‌کنیم. حال به بررسی وجود یا عدم وجود ناسازگاری در این مجموعه خروجی‌ها می‌پردازیم. اگر ناسازگاری کشف شود آن گاه فرض می‌شود که مدل در آن دسته از سؤالات می‌تواند دچار توهم شود. در ادامه پس از کشف ناسازگاری‌ها، پاسخ‌های واقعی و درست هر پرسش یافت شده و سازگاری آن با پاسخ‌های تولید شده مقایسه می‌گردد.

سپس با بررسی سازگاری پاسخ‌های تولید شده و تعداد توهم‌های تولید شده توسط مدل می‌توانیم درصد تولید توهم مدل زبانی بزرگ خود را از طریق دسته‌بندی پرسش‌هایی که تولید کرده‌ایم تخمین بزنیم.

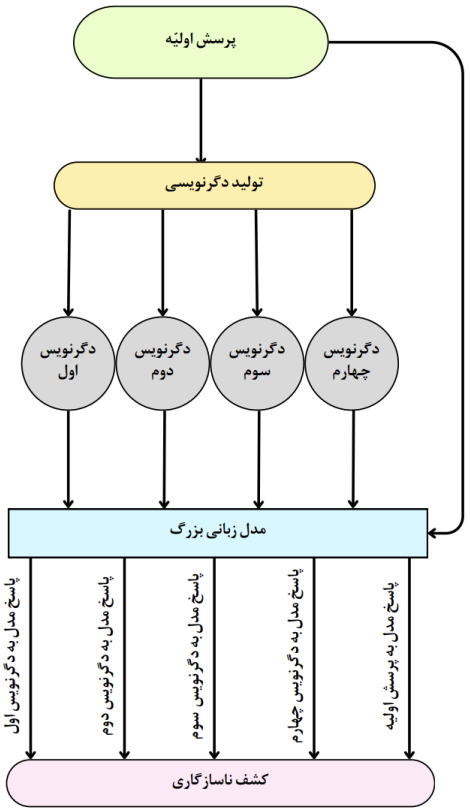
در این پژوهش پرسش‌های ما شامل ۱۰۰ پرسش اولیه است که به ۵ گروه مختلف با موضوع‌های متفاوت تقسیم‌بندی شده‌اند تا علاوه بر ارزیابی کلی بر روی میزان توهم تولید شده توسط مدل زبانی بزرگ، بتوانیم یک نتیجه‌گیری جزئی‌تر بر روی درصد توهم تولید شده در هر موضوع داشته باشیم. این ۵ گروه به دسته‌های ۱) تاریخ و سیاست فرهنگ ۲) موسیقی و سینما ۳) علوم پایه (شیمی، فیزیک و...) ۴) هنر و ادبیات و ۵) تکنولوژی تقسیم می‌شوند. در هر زیرگروه ۲۰ پرسش تهیه شده است که هر پرسش دارای ۴ دگرنویسی است. در نتیجه در پایان آزمایش ۵۰۰ دستور داریم که به مدل زبانی بزرگ داده و ۵۰۰ خروجی دریافت خواهیم کرد.

حال به تعریف تناقض یا عدم سازگاری[۱۷] می‌پردازیم. در این مسأله ما برای ناسازگاری همان تعریف ساده اولیه را در نظر گرفتیم. در واقع ناسازگاری دو عبارت هنگامی رخ می‌دهد که درستی یک عبارت موجب نادرستی عبارت دیگر شود. این مورد در ساده‌ترین شکل هنگامی رخ می‌دهد که پاسخ‌های تولید شده مانع‌الجمع باشند. تشخیص این مسأله در حالتی که پاسخ سؤال یکتا باشد ساده‌تر از حالات دیگر است.

به عنوان مثال اگر از مدل زبانی بزرگ پرسیده شود که «چه کسی جایزه نوبل ادبیات را در سال ۲۰۲۰ دریافت کرد؟»، انتظار داریم تنها نام «لوئیز گلوک» را بیاورد و هر دو نام دیگری که با این نام هم مرجع نباشد با آن ناسازگار است و پاسخ نادرست محسوب می‌شود. اما اگر از آن بپرسیم که «چه کسی جایزه نوبل را در سال ۲۰۲۰ دریافت کرد؟» پاسخ یکتا نیست و مدل می‌تواند نام‌های متفاوتی را تولید کند چرا که جایزه نوبل در حوزه‌های مختلفی به افراد تعلق می‌گیرد و مدل می‌تواند به هرکدام از این حوزه‌ها اشاره کند و نمی‌توانیم این حالت را تناقض در نظر بگیریم.

با توجّه به مسائلی که در ارتباط با کشف تناقض با آن‌ها رو به رو هستیم در آزمون روش پیشنهادی پرسش‌هایی را از بانک سؤالات خود استخراج می‌کنیم که دارای یک پاسخ یکتا باشند تا کار سنجش ناسازگاری ساده باشد. در کارهای آتی این محدودیت‌ها رفع و انواع سؤالات مختلف مورد بررسی قرار خواهند گرفت.

خروجی‌های مدل ما سه حالت می‌توانند داشته باشند. در حالت اول پاسخ درست و سازگار با داده محک را تولید کنند. حالت دوم این است که مدل اطلاعات نادرست و متناقض یا توهم تولید کند. در حالت آخر مدل به عدم دانش خود در آن زمینه اذعان داشته است. این حالت سوم حالت مطلوبی است که به طور واضح مدل ما از عدم دانش خود آگاهی پیدا کرده است و آن را اعلام می‌کند. در نتیجه اگر پاسخ مدل به هر ۴ دگرنویسی و پرسش اولیه در حالت سوم قرار گرفت، آن دسته از پرسش‌ها را در محاسبه دقت آزمایش لحاظ نمی‌کنیم.



شکل (۱): معماری مدل تشخیص توهم در مدل‌های زبانی بزرگ به کمک دگرنویسی

۴- آزمایش و نتایج

می‌دانیم که هدف سیستم ما این است که به ارزیابی یک مدل زبانی از جهت تولید توهم بپردازیم و بتوانیم به سؤال «مدل زبانی ما در چند درصد مواقع توهم تولید می‌کند؟» پاسخ دهیم. در این جا ما به ارزیابی مدل زبانی بزرگ ChatGPT (۳.۵) پرداخته‌ایم. بدین منظور با توجّه به روش ارائه شده ۵۰۰ دستوراولیه را در ۱۰۰ دسته وارد کرده و خروجی‌های حاصل شده در هر دسته‌بندی را بررسی کرده و به کشف تناقض بین پاسخ‌های مربوط به یک پرسش و دگرنویسی‌هایش می‌پردازیم.

در این پژوهش جهت تولید دگرنویسی از یک مدل زبانی استفاده کرده‌ایم. دستور مورد استفاده برای تولید دگرنویسی‌ها «برای پرسش زیر ۱۰ دگرنویس تولید کن + پرسش مورد نظر» بوده است. دگرنویسی‌های تولید شده قبل از استفاده، توسط انسان بررسی شده‌اند و ۴ دگرنویسی مناسب از میان کاندیداهای تولید شده انتخاب شده‌اند. معیار این انتخاب داشتن نحو صحیح و وفاداری معنایی به جمله اصلی بوده است.

با بررسی بر روی پاسخ‌هایی که این مدل زبانی بزرگ به پرسش‌ها و دگرنویسی‌های پرسش‌های انتخاب شده داد، مشاهده کردیم که در خروجی ۴ دسته از پرسش‌ها مدل اعلام کرده است که پاسخ آن پرسش را نمی‌داند. از جهت دیگر قصد ما تفکیک پاسخ‌های مدل زبانی بزرگ به صحیح و حاوی توهم است. در این حالت وقتی مدل زبانی بزرگ اقرار داشته است که از پاسخ پرسش ما آگاهی ندارد لزومی ندارد که این دسته‌ها را در محاسبه ارزیابی تولید توهم آن مدل زبانی بزرگ لحاظ کنیم؛ لذا این ۴ دسته از محاسبات ما حذف شده و در این جا با ۹۶ دسته دیگر رو به رو هستیم.

از این ۹۶ دسته در ۴۱ دسته حداقل یک ناسازگاری میان پاسخ‌ها دیده شد و برای مابقی ۵۵ دسته تمامی پاسخ‌ها نامتناقض و سازگار بوده است. در نتیجه دقت مدل زبانی بزرگ ما در تولید پاسخ درست ۵۷.۲٪ شده است. همچنین می‌توانید مقادیر دقت را به تفکیک هر زیرگروه از پرسش‌ها در جدول ۱ مشاهده کنید.

در پژوهشی که ما انجام داده‌ایم آزمایش‌ها نشان می‌دهند که در ۹۹٪ مواقع اگر همه پاسخ‌ها یا اکثریت پاسخ‌ها با یکدیگر سازگار باشند، در مقایسه با داده محک به ناسازگاری نخواهیم رسید و تنها در یک نمونه از آزمایش‌ها حالتی را داشتیم که اکثریت پاسخ‌ها با یکدیگر سازگار بودند اما در مقایسه با داده محک به ناسازگاری رسیدیم. به عبارت دیگر دقت این روش بر روی داده‌های تست ۹۹٪ گزارش می‌شود.

۵- جمع‌بندی و کارهای آتی

در این مقاله روشی برای بررسی توهم آمیز بودن خروجی مدل‌های بزرگ زبانی مطرح و بر روی داده‌های جمع آوری یا تولید شده مورد ارزیابی قرار گرفت. این روش بر پایه دو فرضیه برای قبول و رد توهم طراحی شده است که توهم را بر اساس سازگاری یا ناسازگاری پاسخ‌های مدل به دگرنویسی‌های مختلف یک پرسش تعیین می‌کند.

این نکته قابل توجه است که این آزمایش می‌تواند برای اطمینان با داده‌های به مراتب حجیم‌تر تکرار شود. به عبارت دیگر در این آزمایش‌ها

جدول (۱): درصد خروجی‌های غیر متوهم (دقت) مدل زبانی بزرگ GPT 3.5روی داده‌های آزمون

	دقت مدل زبانی بزرگ
کل مجموعه داده	۵۷٪
موسیقی و سینما	۳۸.۸٪
هنر و ادبیات	۴۲.۱٪
تاریخ و سیاست و فرهنگ	۵۰٪
علوم پایه (شیمی، فیزیک و...)	۷۰٪
تکنولوژی	۸۹.۴٪.

مجموعه آزمون مورد استفاده گرچه به جهت تنوع گسترده بوده ولی می‌تواند از کمیت بیشتری برخوردار شود. این افزایش می‌تواند با افزایش تعداد سوالات یا افزایش تعداد دگرنویسی‌ها انجام شود.

با توجّه به آن که از ۱۰۰ دسته پرسش اولیه تنها یک نمونه این شرایط را داشت که پاسخ تکرار شونده لزوماً با پاسخ محک یکسان نبود، نشان می‌دهد که جهت تولید دگرنویسی‌ها باید دستورالعمل ویژه‌ای را در نظر داشت و در کارهای آتی بررسی نمود که شرایط دگرنویسی برای اطمینان از کفایت بررسی‌ها چه باید باشد.

همچنین در این پژوهش تنها به بررسی وجود یا عدم وجود تناقض در پاسخ‌ها پرداخته‌ایم. در پژوهش‌های آتی می‌توان به بررسی محل دقیق تناقض پرداخت. به عبارت دیگر بررسی شود که اگر ناسازگاری رخ داده است در کدام بخش از عبارت رخ داده و بازه ناسازگاری معین شود تا کاربرد بتواند متوجّه شود که این مدل زبانی در کدام بخش از پاسخ به پرسش او دچار توهم شده است.

مراجع

[1] T. Brown et al., *"Language Models are Few-Shot Learners,"* Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.

[2] V. Sanh *et al.*, *"Multitask Prompted Training Enables Zero-Shot Task Generalization,"* presented at the International Conference on Learning Representations, Oct. 2021. Accessed: Jan. 27, 2024. [Online]. Available: https://openreview.net/forum?id=9Vrb9D0WI4

[3] R. Thoppilan et al., *"LaMDA: Language Models for Dialog Applications."* arXiv, Feb. 10, 2022. doi: 10.48550/arXiv.2201.08239.

[4] M. Chen et al., *"Evaluating Large Language Models Trained on Code."* arXiv, Jul. 14, 2021. doi: 10.48550/arXiv.2107.03374.

[5] A. Lewkowycz *et al.*, *"Solving Quantitative Reasoning Problems with Language Models,"* Advances in Neural Information Processing Systems, vol. 35, pp. 3843–3857, Dec. 2022.

[6] J. D. Blom, *A Dictionary of Hallucinations*. New York, NY: Springer, 2010. doi: 10.1007/978-1-4419-1223-7.

[7] Y. Bang *et al.*, *"A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and*

Interactivity." arXiv, Nov. 28, 2023. doi: 10.48550/arXiv.2302.04023.

[8] 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.

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

[10] R. Friel and A. Sanyal, *"Chainpoll: A high efficacy method for LLM hallucination detection."* arXiv, Oct. 22, 2023. doi: 10.48550/arXiv.2310.18344.

[11] 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.

[12] P. Manakul, A. Liusie, and M. J. F. Gales, *"SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models."* arXiv, Oct. 11, 2023. doi: 10.48550/arXiv.2303.08896.

[13] T. Liu, K. Wang, L. Sha, B. Chang, and Z. Sui, *"Table-to-Text Generation by Structure-Aware Seq2seq Learning,"* Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, Art. no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11925.

[14] 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. Accessed: Dec. 15, 2023. [Online]. Available: https://aclanthology.org/2023.emnlp-main.397

[15] B. M. Lattimer, P. Chen, X. Zhang, and Y. Yang, *"Fast and Accurate Factual Inconsistency Detection Over Long Documents."* arXiv, Oct. 22, 2023. doi: 10.48550/arXiv.2310.13189.

زیرنویس‌ها

[1] Tara Zare
[2] Mehrnoush Shamsfard
[3] Large Language Models
[4] Unseen Tasks
[5] Question Answering
[6] Hallucination
[7] Factuality Hallucination
[8] Faithfulness Hallucination
[9] Open Domain
[10] Close Domain
[11] Applications
[12] ChatBots
[13] Benchmark
[14] Source Chunking Approach for Large-Scale inconsistency Evaluation
[15] Natural Language Inference
[16] Prompt
[17] Inconsistency