به نام خدا

وزارت ارتباطات و فناوری اطلاعات
پژوهشگاه ارتباطات و فناوری اطلاعات
(مرکز تحقیقات مخابرات ایران)
ITRC

شیوه نامه جمع آوری و برچسب‌زنی دادگان متنی برای وظیفه پیروی از دستورالعمل

آزمایشگاه مرجع ارزیابی هوش مصنوعی
بهار ۱۴۰۴

سرفصل مطالب

۱ مقدمه

۱/۱ هدف شیوه‌نامه

حوزه پردازش زبان طبیعی شاهد ظهور و پیشرفت مدل‌های زبانی بزرگ بوده است که توانایی قابل‌توجهی در شبیه‌سازی تولید متن به سبک انسانی نشان داده‌اند. توانایی این مدل‌ها در پیروی از دستورالعمل‌ها برای کاربردهای عملی آن‌ها بسیار مهم است، زیرا امکان شخصی‌سازی خروجی‌ها مطابق با نیازهای خاص کاربران را فراهم می‌کند. با این حال، ارزیابی این مدل‌ها عمدتاً بر وظایف کلی زبان طبیعی یا وظایف خاص پایین‌دستی متمرکز است، در حالی که عنصر حیاتی پیروی از دستورالعمل که توانایی مدل در درک و اجرای دقیق دستورالعمل‌های کاربر به‌طور کامل است، مورد بررسی قرار نگرفته است[1] .

این چالش زمانی بیشتر می‌شود که دستورات متعددی در یک گفت‌وگوی چندمرحله‌ای بین کاربر و یک مدل بنیادین بزرگ پراکنده باشند و مدل مجبور باشد در طول چندین مرحله از مکالمه استدلال کند.

۱,۲ دامنه کاربرد

در این شیوه نامه ما به سه بخش اصلی ارزیابی مدل در زمینه‌های زیر می‌پردازیم:

- پیروی از دستورالعمل‌ها
- دستورات چندمرحله‌ای
- فیوشات

۱/۲/۱ پیروی از دستورالعمل‌ها[1]

پیروی دقیق از دستورالعمل‌ها یکی از مهم‌ترین توانایی‌های مدل‌های هوش مصنوعی است که کاربردهای متنوعی را در زمینه‌های مختلف ممکن می‌سازد. این توانایی به مدل‌ها امکان می‌دهد تا با درک بهتر نیازهای کاربران و شرایط خاص، پاسخ‌هایی دقیق، مرتبط و آگاه به زمینه ارائه دهند. در زیر چند نمونه از حوزه های کاربردی پیروی از دستورالعمل‌ها آمده است:

[1] Instruction following

- **تعامل انسان با کامپیوتر**

دستورالعمل‌های متنی می‌توانند برای هدایت کامپیوترها به انجام وظایف واقعی استفاده شوند، که می‌تواند شامل موارد زیر باشد:

  - انتخاب اشیاء از یک تصویر یا ویدیو
  - پیروی از دستورالعمل‌های مسیریابی
  - بازی کردن
  - کنترل نرم‌افزارها.
  - پرس‌وجو از پایگاه‌داده‌ها

- **سیستم‌های تکراری و ماژولار**

تحقیقات اخیر از رویکردی تکراری و ماژولار برای ارتباط انسان کامپیوتر استفاده می‌کنند. در این سیستم‌ها، سیستم سوالاتی می‌پرسد تا دستورالعمل‌های کاربر را روشن کند و کاربر می‌تواند دستورالعمل‌ها را به‌طور دلخواه اصلاح کند. این امر استفاده از سیستم را برای افراد غیرمتخصص راحت‌تر کرده و منجر به تولید سیستم‌های هوش مصنوعی کاربرپسندتر می‌شود.

- **تقویت داده‌ها و ویژگی‌ها**

دستورالعمل‌ها می‌توانند به‌عنوان نوعی نظارت دور برای تقویت داده‌ها یا ویژگی‌ها برای مدل‌های یادگیری ماشین استفاده شوند. به عنوان مثال:

  - دستورالعمل‌ها می‌توانند برای خودکارسازی حاشیه‌نویسی داده‌ها استفاده شوند.
  - دستورالعمل‌ها می‌توانند برای تولید ویژگی‌های اضافی برای یک مدل استفاده شوند.

- **مدل‌های زبان عمومی**

یکی از اهداف اصلی پیروی از دستورالعمل‌ها، توسعه مدل‌های زبان عمومی است که قادر به انجام طیف وسیعی از وظایف و زبان‌ها باشند. پیشرفت‌های اخیر در پیروی از دستورالعمل‌ها، مانند InstructGPT، ChatGPT، و GPT-4، نشان‌دهنده پیشرفت قابل توجهی در جهت این هدف هستند.

پیروی از دستورالعمل‌ها یک حوزه امیدوارکننده برای تحقیقات آینده است. یکی از حوزه‌های تمرکز، توسعه مدل‌هایی است که بتوانند دستورالعمل‌های پیچیده‌تری را درک کرده و دنبال کنند. حوزه دیگر تمرکز، توسعه مدل‌هایی است که مقاوم‌تر به حملات مخالف باشند و قادر به تولید خروجی‌های ایمن‌تر و قابل اعتمادتر باشند.

۱/۲/۲ پیروی از دستورات چندمرحله‌ای[1]

دنبال کردن دستورات چندمرحله‌ای در هر حوزه‌ای که نیاز به تعامل سیستم‌های هوش مصنوعی با انسان‌ها یا دیگر عوامل در یک ساختار مکالمه‌ای وجود داشته باشد، اهمیت دارد. این قابلیت برای انجام وظایف پیچیده‌ای که شامل مجموعه‌ای از اقدامات متوالی است، بسیار کاربردی خواهد بود.

- **چت‌بات‌ها و هوش مصنوعی مکالمه‌ای:**

ماهیت پیروی از دستورات چندمرحله‌ای به‌طور کامل با تعاملات چت‌بات‌ها همخوانی دارد. توانایی پیروی از دستورات در مکالمات طولانی برای توسعه چت‌بات‌هایی که قادر به انجام وظایف پیچیده، ارائه تجربیات شخصی‌سازی‌شده، و ایجاد گفت‌وگوهای طبیعی باشند، امری ضروری است.

- **رباتیک و اتوماسیون:**

ربات‌ها می‌توانند برای انجام وظایف پیچیده، مجموعه‌ای از دستورات دریافت کنند. این وظایف ممکن است شامل مونتاژ قطعات، حرکت در محیط‌های پویا، یا تعامل با اشیاء و افراد باشد. موفقیت در انجام این وظایف به دنبال کردن دقیق دستورات در مراحل مختلف وابسته است.

- **بهداشت و درمان:**

سامانه‌های هوش مصنوعی در حوزه سلامت می‌توانند بیماران را طی چند مرحله برای انجام آزمایش‌های تشخیصی یا درمان‌ها راهنمایی کنند. این سامانه‌ها می‌توانند بر اساس پاسخ‌های بیماران، دستورات را شخصی‌سازی کرده و طی تعاملات چندمرحله‌ای ارائه دهند.

[1] Multi-turn instruction-following evaluation

۱/۲/۳ فیوشات

دامنه کاربرد فیوشات[1] شامل موارد زیر است:

- **دامنه‌های با منابع محدود**: آموزش چند نمونه‌ای برای دامنه‌هایی مناسب است که در آن‌ها به دست آوردن حجم زیادی از داده‌های برچسب‌دار پرهزینه یا دشوار است. این موضوع به‌ویژه در مورد دامنه‌های تخصصی، زبان‌های کم‌استفاده یا وظایفی با داده‌های آموزشی محدود اهمیت دارد، جایگاه خود را بیشتر نشان می‌دهد.
- **کاربردهای حساس به حریم خصوصی**: در سناریوهایی که شامل داده‌های حساس کاربران می‌شود، جمع‌آوری مجموعه داده‌های برچسب‌دار بزرگ ممکن است نگرانی‌های مربوط به حریم خصوصی و انطباق با قوانین را ایجاد کند.
- **سیستم‌های شخصی‌سازی و توصیه‌گر**: وظایف شخصی‌سازی اغلب شامل تطبیق مدل‌ها با ترجیحات یا رفتارهای کاربران خاص است. آموزش چندنمونه‌ای می‌تواند برای انطباق مدل‌ها با داده‌های محدود تعامل کاربر مورد استفاده قرار گیرد.
- **وظایف پیچیده پردازش زبان طبیعی**: مانند شناسایی موجودیت‌های نام‌دار و درک ماشینی متون[2]

۱/۳ تعریف اصطلاحات

- فیوشات:
  - فیوشات به یادگیری ماشینی اطلاق می‌شود که در آن مدل تنها با استفاده از تعداد کمی مثال (معمولاً بین ۱ تا ۳۰ مثال) آموزش می‌بیند و باید قادر به تعمیم این مثال‌ها به وظایف مشابه باشد. در این روش، مدل باید توانایی خود را در پیش‌بینی صحیح با داده‌های محدود به‌کار بگیرد و از تعداد کمی از نمونه‌ها برای انجام وظایف جدید استفاده کند. این نوع یادگیری چالش اصلی

[1] Few shot

خود را در تعمیم سریع از داده‌های محدود به داده‌های جدید دارد، به‌طوری که مدل باید از این نمونه‌های اندک به بهترین شکل برای شناسایی الگوها و انجام پیش‌بینی‌ها استفاده کند.

- دستورالعمل های قابل تایید[1]:
  - دستورالعمل‌های قابل تأیید نوع خاصی از دستورالعمل‌ها هستند که برای ارزیابی توانایی مدل‌های زبان بزرگ در پیروی از دستورالعمل‌ها طراحی شده‌اند. این دستورالعمل‌ها می‌توانند به‌طور دقیق و خودکار بررسی شوند. نمونه‌هایی از این دستورالعمل‌ها شامل "بین ۴۵۰ تا ۵۰۰ کلمه بنویسید"، "تمام خروجی شما باید در فرمت JSON باشد"، یا "عنوان را وارد کرده و آن را در دو براکت مربعی مانند [[عنوان]] قرار دهید" هستند. تمرکز بر دستورالعمل‌های قابل تأیید به محققان این امکان را می‌دهد که توانایی پیروی از دستورالعمل‌ها توسط مدل‌های زبان بزرگ را به‌طور خودکار و بدون نیاز به قضاوت‌های انسانی ارزیابی کنند. محققان از این دستورالعمل‌ها برای ساخت مجموعه‌ای از درخواست‌ها استفاده می‌کنند که برای ارزیابی مدل‌های زبان بزرگ به کار می‌روند. آنها اذعان دارند که بسیار کمی از دستورالعمل‌ها به‌طور ۱۰۰٪ قابل تأیید خودکار هستند و اشاره می‌کنند که برای حل این مشکل از روش‌های تأیید قوی‌تری استفاده می‌کنند که تغییرات رایج را در نظر می‌گیرند. با استفاده از دستورالعمل‌های قابل تأیید، محققان می‌توانند بینش‌های واضح و عینی درباره نقاط قوت و ضعف مدل‌های زبان بزرگ در پیروی از دستورالعمل‌های مختلف کسب کنند.

- پیروی از دستورات چندمرحله‌ای :
  - پیروی از دستورالعمل‌ها در چند نوبت به توانایی یک مدل برای درک و پاسخ دادن به مجموعه‌ای از دستورالعمل‌ها یا درخواست‌ها به‌طور هماهنگ و مناسب از نظر بافت در طول چندین تعامل

[1] verifiable instructions

اشاره دارد. برخلاف پیروی از دستورالعمل‌ها در یک نوبت، که مدل تنها یک ورودی را پردازش کرده و یک پاسخ تولید می‌کند، پیروی از دستورالعمل‌ها در چند نوبت نیازمند آن است که مدل با حفظ بافت و پیوستگی در طول چندین تبادل، ورودی‌های قبلی را به یاد آورده، پاسخ‌های خود را به‌طور متناسب تنظیم کرده و دستورالعمل‌های پیچیده‌ای که بر همدیگر بنا می‌شوند را مدیریت کند. این توانایی برای انجام وظایفی مانند حل مسائل مبتنی بر مکالمه یا راهنمایی گام به گام ضروری است، جایی که مدل باید در تعاملات پویا و رفت و برگشتی شرکت کند.

۲ الزامات جمع‌آوری داده

۲/۱ چالش‌های بین‌المللی/عمومی محصول

یکی از قابلیت‌های کلیدی LLM پیروی از دستورالعمل‌های زبان طبیعی ورودی است که به عنوان zero shot prompt نیز شناخته می شود. توانایی LLM ها برای تفسیر دقیق و پیروی از دستورالعمل های زبان طبیعی، نه تنها برای دقت وظایف، بلکه برای ایمنی و قابلیت اطمینان اجرای آنها بسیار مهم است. مغایرت یا سوء تفاهم در «پیروی از دستورالعمل ها» می‌تواند منجر به خروجی‌های ناخواسته شود، که ممکن است نتایج بدی داشته باشد، این نکته به ویژه در سناریوهای حیاتی مانند مراقبت‌های بهداشتی یا سیستم‌های مستقل باید مورد توجه قرار بگیرد.

۲/۱/۱ چالش‌های پیروی از دستورالعمل

۱- **ذهنی بودن و ابهام در زبان**: زبان انسانی ذاتاً ذهنی و دارای قابلیت تفسیرهای مختلف است. یک دستور می‌تواند توسط افراد یا سیستم‌های مختلف به شکل‌های متفاوتی درک شود که منجر به قضاوت‌های متناقض درباره پیروی صحیح از دستور می‌شود.

۲- **ارزیابی پیروی از دستورات:** ارزیابی دقیق میزان پیروی از دستورات، وظیفه‌ای پیچیده است. روش های ارزیابی پیروی از دستورالعمل را می توان به سه بخش اصلی تقسیم کرد که دو بخش اولی هرکدام چالش مختص به خود را داراست که ادامه بیان می‌شود:

- ارزیابی انسانی: این روش زمان‌بر، پرهزینه و متکی به گروهی از برچسب‌زن‌های انسانی است که ممکن است به تعصبات و ناسازگاری‌هایی منجر شود که بازتولیدپذیری نتایج را به چالش می‌کشد.
- ارزیابی مبتنی بر مدل: در این روش از یک مدل داخلی یا خارجی برای ارزیابی عملکرد مدل هدف استفاده می‌شود. با این حال، این رویکرد به شدت به صحت مدل ارزیابی‌کننده وابسته است که تضمین‌شده نیست. در صورت وجود محدودیت‌های قابل‌توجه در مدل ارزیابی‌کننده، سیگنال‌های ارزیابی نادرست و گمراه‌کننده خواهند بود.
- بنچمارک‌های کمی: این روش یک رویکرد استاندارد و مقیاس‌پذیر برای ارزیابی ارائه می‌دهد. به طور مثال به ارزیابی وظایف تولیدی، به ویژه پیروی از دستورالعمل‌های مرتبط با شمارش، نوعی از این دسته بندی است با این حال این نوع ارزیابی هم اغلب نمی‌توانند ظرافت‌ها و پیچیدگی‌های زبان طبیعی را به طور کامل درک و ارزیابی کنند.. [2]

۳- قابلیت راستی‌آزمایی دستورات: بسیاری از دستورات معیارهای واضح و عینی برای راستی‌آزمایی ندارند. برای مثال، ارزیابی این که آیا دستوری مانند "با لحنی خنده‌دار بنویس" به درستی اجرا شده است یا خیر، مستلزم قضاوت ذهنی است که تعیین موفقیت را دشوار می‌کند.

۲/۱/۲ چالش‌های پیروی از دستورات چندمرحله‌ای

۱- **بازیابی و استدلال دستورات**: چالش اصلی درک یا اجرای دستورات نیست، بلکه در یافتن و بازیابی دستورات مرتبط از تاریخچه طولانی و متغیر مکالمه است. مدل‌ها باید دستورات داده‌شده در نقاط مختلف گفتگو را شناسایی کرده و بر اساس آن‌ها استدلال کنند تا از پیروی منسجم اطمینان حاصل شود.

۲- **استدلال در متن طولانی**: پیروی از دستورات چندمرحله‌ای نیازمند درک و ترکیب اطلاعات ارائه‌شده در طول چندین مرحله است. این شامل حفظ زمینه، ردیابی تغییرات و تطبیق با دستورات جدید در حالی است که دستورالعمل‌های قبلی نیز حفظ می‌شوند.

۳- **پیروی پایدار و منسجم از دستورات**: کافی نیست که مدل‌ها تنها گاهی دستورات را دنبال کنند؛ آن‌ها باید تمام دستورات داده‌شده را به‌طور دقیق و مداوم در مراحل مختلف اعمال کنند. این موضوع شامل مدیریت انواع مختلف دستورات و تعاملات احتمالی آن‌ها نیز می‌شود[3] .

۲/۱/۳ چالش‌های چندنمونه‌ای[1]

یکی از چالش‌های کلیدی در یادگیری چندنمونه‌ای[2] برای درک زبان طبیعی[3]، اختلاف قابل‌توجه بین عملکرد انسان و ماشین است. این چالش به دلایل متعددی تشدید می‌شود:

کمبود داده‌های برچسب‌گذاری‌شده: کمبود داده‌های برچسب‌گذاری‌شده در تنظیمات چندنمونه‌ای تطبیق مدل‌ها با وظایف جدید را دشوار می‌سازد. روش‌های کلاسیک تنظیم دقیق[4] که شامل به‌روزرسانی وزن‌های مدل هستند، به‌ویژه با این محدودیت مواجه می‌شوند.

پیچیدگی وظایف: با افزایش پیچیدگی وظایف، مانند وظایف تشخیص موجودیت نامدار[5] و درک مطلب ماشینی[6]، فاصله عملکرد بین انسان و مدل‌ها به‌طور قابل‌توجهی افزایش می‌یابد. انسان‌ها با تعداد محدودی از مثال‌ها عملکرد قوی نشان می‌دهند، در حالی که مدل‌ها اغلب نتایجی با دقت پایین‌تر ارائه می‌دهند[3] .

۲,۲ چالش‌های بومی محصول

- **چالش زبان :**

در دستورالعمل‌های قابل اندازه گیری که در ادامه به آن ها اشاره می شود در دستور زبان و بزرگ و کوچکی کلمات و علائم نگارشی تفاوت هایی بین زبان فارسی و انگلیسی هست که باید لحاظ شود. به طور مثال یکی از دستورالعمل های قابل اندازه گیری در زبان انگلیسی شمارش تعداد کلمات با حروف بزرگ است مثلا TEHRAN که چنین دستورالعملی در زبان فارسی معنا ندارد.

---

[1] Few-shot
[2] few-shot learning
[3] Natural language understanding
[4] fine-tuning
[5] Named entity recognition
[6] Machin reading comperhension

- **چالش گویش و لهجه**

گویش‌های مختلف زبان فارسی دارای تفاوت‌هایی از لحاظ دستوری هستند که می تواند ارزیابی مدل زبانی را در زمینه پیروی از دستورات چند مرحله ای دچار چالش کنند.

- **چالش ابهام در زبان فارسی**

با توجه به قدمت زبان فارسی ابهات مختلفی در بیان یک دستورالعمل به زبان فارسی وجود دارد . شناسایی این ابهامات(حتی چالش های ابهامات در گویش های مختلف) ارزیابی مدل را دقیق تر می‌کند [3] [2].

۲٫۳ تنوع داده‌ای

یکی از جنبه‌های کلیدی توسعه مدل‌های پیروی از دستورات تنوع داده‌ای است. این امر به چند دلیل حائز اهمیت است:

- با توجه به ویژگی ذهنی بودن زبان، تفسیر و پاسخ به دستورات تحت تأثیر زمینه‌های فرهنگی و زبانی قرار دارد. تنوع داده‌ها به مدل‌ها کمک می‌کند تا این ظرافت‌ها را بهتر درک کنند
- روش‌های ارزیابی که به قضاوت انسانی یا مدل‌های از پیش آموزش‌دیده متکی هستند، ممکن است در صورت نبود تنوع در داده‌های آموزشی، در معرض سوگیری قرار گیرند
- مدل‌های پیروی از دستورات برای استفاده در سناریوهای دنیای واقعی طراحی شده‌اند، جایی که با دستورات و زمینه‌های متنوعی روبه‌رو خواهند شد. داده‌های آموزشی که این تنوع را منعکس می‌کند، برای کاربرد عملی این مدل‌ها ضروری است.

داده ها می توانند به اشکال زیر باشند:

- به صورت پرامپت هایی با دستورالعمل های مشخص و قابل اندازه گیری توسط انسان
- برای پیروی از دستورات چند مرحله‌ای شامل مکالمات بین انسان و ماشین(مانند چت بات آرنا)

۲,٤ حجم دادگان و میزان لازم برای هر چالش

- برای پیروی از دستورالعمل‌ها
  - حداقل ۲۵ نوع دستور قابل اندازه گیری مجزا
  - حداقل ۵۰۰ پرامپت که در هر کدام دو دستور قابل اندازه گیری آمده باشد
- برای پیروی از دستورالعمل‌های چند مرحله‌ای
  - حداقل ۷۰ مکالمه
  - هر مکالمه بین ۱ تا ۲۰ گفت‌گو داشته باشد(با میانگین ۱۴)
  - جمعا حدود ۹۸۰ گفت‌وگو
- برای دادگان ارزیابی فیوشات
  - برای هر حوزه تعداد دادگان ارزیابی باید با نسبت مشخصی مثلا ۱ به ۶ نسبت به دادگان تست باشد که بتواند وظیفه ارزیابی چندشات را به درستی انجام دهد و یادگیری از تعداد محدودی از نمونه‌ها را به چالش بکشاند(به طور مثال ۳۰ به ۲۰۰)
  - در نمونه های مختلف (۱۰شات، ۲۰شات و ۳۰ شات) دادگان تست مورد ارزیابی قرار گیرند.
  - به دلیل کوچک بودن داده‌های آموزشی، نمونه‌های خاصی که انتخاب می‌شوند می‌توانند تأثیر زیادی بر عملکرد مدل داشته باشند. برای در نظر گرفتن این تأثیر باید تقسیم‌بندی‌های آموزشی مختلف برای هر "شات" (تعداد نمونه‌های آموزشی) ایجاد کرد. به عنوان مثال، در تنظیم ۱۰ شات، پنج مجموعه متمایز از ۱۰ نمونه مختلف وجود خواهد داشت که همگی از مجموعه ۳۰ نمونه موجود استخراج شده‌اند.

۳ فرآیند جمع‌آوری داده

۳/۱ مراحل و روش‌های جمع‌آوری داده‌ها

۳/۱/۱ پیروی از دستورالعمل

برای ارزیابی توانایی مدل‌های زبانی بزرگ در پیروی از دستورات، معمولاً از رویکردی سیستماتیک برای آماده‌سازی و جمع‌آوری داده‌ها استفاده می‌شود. روش کلی شامل مراحل زیر است:

۱- ایجاد دستورالعمل‌ها

ایجاد دستورالعمل‌ها: طراحی مجموعه‌ای از دستورالعمل‌ها که شامل دستورالعمل‌های واضح و قابل بررسی[1] باشند. این دستورالعمل‌ها برای آزمودن توانایی مدل در درک و اجرای وظایف خاص طراحی می‌شوند.

افزایش تنوع: بازنویسی یا گسترش دستورالعمل‌ها برای افزایش تنوع و جلوگیری از تطابق بیش از حد مدل با الگوهای خاص

فیلتر کردن: استفاده از روش‌های خودکار یا دستی برای حذف دستورالعمل‌ها غیرمنطقی یا مبهم، به‌منظور اطمینان از هم‌راستایی مجموعه داده با اهداف ارزیابی.

بازبینی دستی: پژوهشگران ممکن است دستورالعمل‌ها را برای حفظ کیفیت و انسجام بالا اصلاح کنند.

۲- ایجاد پرامپت

با استفاده از دستورالعمل‌های موجود پرامپت‌هایی توسط مدل زبانی تولید و توسط بازبینان اصلاح می گردد که شامل دستورات متناقض نباشد.

۳- تعامل با مدل

ارسال درخواست به مدل‌های زبانی: پرامپت‌ها از طریق API یا بر روی مدل‌های زبانی لوکال ارسال می‌شوند تا پاسخ‌ها تولید شوند. مدل‌های مورد ارزیابی بر اساس ارتباط و دسترسی انتخاب می‌شوند.

اجرای چندباره: برای در نظر گرفتن تصادفی بودن خروجی مدل، ممکن است هر پرامپت چندین بار اجرا شده و پاسخ‌ها به صورت جمعی تحلیل شوند.

[1] verifiable instructions

٤- جمع‌آوری پاسخ‌ها

ثبت داده‌ها: پاسخ‌های مدل‌ها برای تحلیل بیشتر ثبت می‌شوند. این شامل خروجی خام و داده‌های مرتبط مانند زمان اجرا و جزئیات پیکربندی (مانند دما[1]یا حداکثر تعداد توکن‌ها) است.

ذخیره‌سازی سازمان‌یافته: پاسخ‌ها به صورت سیستماتیک ذخیره می‌شوند تا مقایسه و امتیازدهی در آینده را تسهیل کنند.

٥- چارچوب ارزیابی

معیارهای امتیازدهی: پاسخ‌ها بر اساس معیارهای از پیش تعیین‌شده مانند درستی، کامل بودن و پیروی از دستورالعمل‌ها ارزیابی می‌شوند.

امتیازدهی خودکار یا دستی: بسته به پیچیدگی وظیفه، امتیازدهی ممکن است توسط الگوریتم‌ها، بازبین‌های انسانی یا ترکیبی از هر دو انجام شود.

٦- بهبود تدریجی

گسترش مجموعه داده: بینش‌های به‌دست‌آمده از ارزیابی‌های اولیه ممکن است به ایجاد پرامپت‌های اضافی برای هدف‌گیری نقاط ضعف مدل منجر شوند.

به‌روزرسانی بنچمارک: مجموعه داده‌های ارزیابی ممکن است برای حفظ ارتباط با پیشرفت‌های مدل‌ها به‌طور دوره‌ای به‌روزرسانی شوند.

این رویکرد ساختارمند تضمین می‌کند که ارزیابی مدل‌های زبانی بزرگ به شکلی دقیق و منصفانه انجام شود، در حالی که بر کیفیت پرامپت‌ها و تحلیل پاسخ‌ها تمرکز دارد.

در جدول ۱ بخشی از دستورالعمل‌های واضح و قابل بررسی آمده است.

جدول ۱- دستورالعمل‌های واضح

گروه دستورالعمل	دستورالعمل	توضیحات
کلمه کلیدی	شامل کلمه کلیدی	کلمات کلیدی {keyword1} و {keyword2} را در پاسخ خود بگنجانید.
کلمه کلیدی	فروانی کلمه	در پاسخ شما، کلمه {کلمه} باید {N} بار ظاهر شود.
کلمه کلیدی	کلمات ممنوع	کلمات کلیدی {کلمات ممنوع} را در پاسخ خود قرار ندهید.

[1] temperature

کلمه کلیدی	فراوانی حروف	در پاسخ شما، حرف {حرف} باید {N} بار ظاهر شود.
زبان	زبان پاسخ	تمام پاسخ شما باید به زبان {زبان مورد نظر} باشد و هیچ زبان دیگری مجاز نیست.
محدودیت‌های طولی	تعداد پاراگراف‌ها	پاسخ شما باید شامل {N} پاراگراف باشد. شما پاراگراف‌ها را با استفاده از جداکننده markdown به این صورت جدا می‌کنید: * * *
محدودیت‌های طولی	تعداد کلمات	پاسخ شما باید حداقل / حدوداً / حداکثر {N} کلمه باشد.
محدودیت‌های طولی	تعداد پاراگراف‌ها + اولین کلمه در پاراگراف i-ام	باید {N} پاراگراف وجود داشته باشد. پاراگراف‌ها و فقط پاراگراف‌ها با دو خط شکاف از هم جدا می‌شوند. پاراگراف {i}-ام باید با کلمه {first word} شروع شود.
محتوای قابل شناسایی	Postscript	در انتهای پاسخ خود، لطفاً به‌طور صریح یک پس‌نویس اضافه کنید که با {postscript marker} شروع شود.
محتوای قابل شناسایی	تعداد جای‌گذاری‌ها[1]	پاسخ باید شامل حداقل {N} جای‌گذاری باشد که با براکت‌های مربعی نمایش داده شوند، مانند [address].
فرمت قابل شناسایی	تعداد بولت ها	پاسخ شما باید دقیقاً شامل {N} بولت باشد. از بولت مارک‌داون مانند نمونه استفاده کنید: *این یک نقطه است.
فرمت قابل شناسایی	عنوان	پاسخ شما باید شامل یک عنوان باشد که در آکولادهای دوتایی قرار گیرد، مانند <<شعر شاد>>.
فرمت قابل شناسایی	انتخاب از گزینه‌ها	پاسخ شما باید یکی از گزینه‌های زیر را انتخاب کند: {گزینه ها}
فرمت قابل شناسایی	حداقل تعداد بخش‌های برجسته	لطفاً حداقل {N} بخش را در پاسخ خود با استفاده از مارک‌داون برجسته کنید، به‌طور مثال: *بخش برجسته*
فرمت قابل شناسایی	بخش‌های متعدد	پاسخ شما باید شامل {N} بخش باشد. شروع هر بخش را با X { علامت‌گذاری کنید.
فرمت قابل شناسایی	فرمت JSON	تمامی خروجی در فرمت JSON قرار گیرد.
چک کردن شروع و پایان	چک کردن پایان	پاسخ خود را دقیقا با این عبارت {پایان عبارت} تمام کنید. هیچ کلمه دیگری نباید دنبال این عبارت باشد.
چک کردن شروع و پایان	نقل قول	کل پاسخ خود را در دو علامت نقل قول قرار دهید
علامت گذاری	بدون کاما	در تمام پاسخ خود، از استفاده از کاما خودداری کنید.

۳/۱/۲ دادگان فیوشات

۱- انتخاب منابع داده

[1] Placeholder

مجموعه‌داده‌های موجود: استفاده از مجموعه‌داده‌های عمومی و شناخته‌شده که برای وظایف مشابه طراحی شده‌اند. این منابع می‌توانند داده‌های قابل اعتماد و متنوعی را فراهم کنند. می‌توان برای تسک‌های مختلف از دیتاست های موجود استفاده کرد.

به عنوان نمونه می توان به موارد زیر اشاره کرد:

- دسته بندی جملات :دیتاست MNLI
- درک مطلب ماشینی: دیتاست SQuADv2
- تشخیص موجودیت‌های نامدار: دیتاست CoNLL03, WikiANN

داده‌های جدید: در صورت عدم وجود داده‌های کافی، جمع‌آوری داده‌های جدید از منابع معتبر، مانند پرسش‌نامه‌ها، پایگاه‌های داده عمومی یا محتوای وب، ضروری است.

۲- تعیین معیارهای انتخاب داده

نمایندگی داده‌ها: انتخاب داده‌هایی که تنوع کافی برای پوشش انواع مختلف شرایط، ورودی‌ها و خروجی‌های وظیفه داشته باشند.

کیفیت داده‌ها: اطمینان از دقت و صحت داده‌ها از طریق بررسی کیفیت و پالایش داده‌های اولیه.

تنظیم اندازه نمونه‌ها: تعیین تعداد نمونه‌های مناسب برای هر بخش از داده‌ها (آموزشی، آزمایشی، با توجه به وظیفه چند شات نیاز به داده اعتبارسنجی نیست) با توجه به اهداف تحقیق.

۳- نمونه‌گیری

نمونه‌گیری تصادفی: انتخاب زیرمجموعه‌هایی از داده‌ها به صورت تصادفی برای جلوگیری از سوگیری در داده‌های آموزشی و آزمایشی.

نمونه‌گیری هدفمند: در مواردی که برخی گروه‌ها یا دسته‌ها باید به طور خاص نمایندگی شوند، از روش‌های نمونه‌گیری هدفمند استفاده شود.

۴- تقسیم داده‌ها

ایجاد تقسیم‌بندی‌های آموزشی و آزمایشی: داده‌ها را به بخش‌هایی برای آموزش و آزمایش مدل تقسیم شود تا عملکرد مدل به درستی ارزیابی شود.

تقسیم‌بندی‌های چندگانه: برای ارزیابی بهتر، داده‌ها را در چند مجموعه تقسیم شوند. این رویکرد کمک می‌کند تا اثر تغییرات جزئی در داده‌های آموزشی بر عملکرد مدل مشخص شود. مثلا برای ۱۰ شات از بین ۳۰ نمونه داده ۵ مدل تقسیم بندی انجام شود.

۵- تنظیم فرمت داده‌ها

تمام داده‌ها باید به یک قالب استاندارد تبدیل شوند که استفاده از آن در مدل‌ها آسان باشد.

یکسان‌سازی وظایف: اگر داده‌ها مربوط به وظایف مختلف هستند، آن‌ها را به یک فرمت کلی تبدیل شوند تا استفاده از یک معماری واحد امکان‌پذیر باشد.

۶- اطمینان از تنوع وظایف

نمایش انواع مختلف ورودی‌ها: داده‌ها باید نماینده تمامی انواع ممکن از ورودی‌های مرتبط با وظیفه باشند.

تعادل در داده‌ها: اطمینان حاصل شود که گروه‌های مختلف داده به صورت متعادل حضور دارند تا از سوگیری جلوگیری شود.

۳/۱/۳ پیروی از دستورات چندمرحله‌ای

۱- گسترش داده‌های موجود برای مکالمات چندمرحله‌ای

ابتدا از یک مجموعه داده دستورات تک‌مرحله‌ای معتبر (مانند IFEval) که حاوی دستورات قابل بررسی است استفاده کنید.

۲- گسترش به مکالمات چندمرحله‌ای

باید پرسش‌های تک‌مرحله‌ای به مکالمات چندمرحله‌ای تبدیل شوند. هدف این است که قالبی ساختاریافته ایجاد شود، که معمولاً شامل سه مرحله در هر مکالمه است.

۳- نمونه‌برداری دستورات

باید به‌طور تصادفی از فهرست پیش‌تعریف شده انواع دستورات نمونه‌برداری کرد. در اینجا ۳۰ نوع دستور در مجموعه داده IFEval موجود است. این اطمینان باید حاصل شودکه هر مرحله از مکالمه چندمرحله‌ای یک دستور جدید را معرفی می‌کند که منطقی به دنباله دستورات قبلی متصل است.

۴- تولید دستورات به زبان طبیعی

از یک مدل زبان بزرگ مانند Llama 3.1 405B برای تبدیل انواع دستورات نمونه‌برداری شده به دستوراتی منسجم و طبیعی استفاده شود. به عنوان مثال، دستوری مانند "تمام حروف بزرگ" می‌تواند به "اطمینان حاصل کنید که تمام پاسخ شما به زبان انگلیسی و با حروف بزرگ نوشته شده باشد" تبدیل شود.

۵- مدیریت تضادهای بین دستورات

- شناسایی خودکار تضادها: از مدل‌های زبان بزرگ برای شناسایی تضادهای احتمالی بین دستورات در مکالمه چندمرحله‌ای استفاده شود. برای مثال، تضادی به وجود می‌آید اگر یک دستور خواستار اختصار باشد در حالی که دستور دیگری نیاز به توضیحات مفصل دارد.
- بازبینی انسانی برای حل تضادها: از بازبینان انسانی برای بررسی دستورات علامت‌گذاری شده و اطمینان از حذف هرگونه تضاد باقی‌مانده استفاده شود. این کار به حفظ انسجام و وضوح در تمام مراحل کمک می‌کند.

۶- حذف محتوای حساس

- اسکن خودکار برای محتوای حساس: ابتدا از مدل‌های زبان بزرگ برای اسکن دستورات تولید شده به‌منظور شناسایی محتوای حساس استفاده شود. این شامل موضوعاتی مانند سیاست، مذهب یا روابط انسانی است که ممکن است برای همه مخاطبان مناسب نباشد.
- بازبینی انسانی برای فیلتر کردن محتوا: پس از اسکن، از حاشیه نویسان انسانی برای بررسی محتوای علامت‌گذاری شده و حذف هرگونه محتوای حساس یا نامناسب استفاده شود.

۷- ایجاد مجموعه داده نهایی

مجموعه داده نهایی باید شامل تعداد مشخصی از مکالمات چندمرحله‌ای در زبان مورد نظر (مثلاً فارسی یا انگلیسی) باشد که هر کدام سه مرحله دارند.[6] .

۳/۲ ابزارها و نرم‌افزارهای مورد استفاده

Crowd-sourcing platforms

API Interaction Tools

۴ برچسب‌زنی دادگان

۴/۱ پیروی از دستورالعمل

- تولید اولیه دستورات: پژوهشگران از روش فیوشات پرامپتینگ برای تولید خودکار مجموعه‌ای از دستورات پایه با دستورالعمل‌های قابل تأیید استفاده کرده‌اند. این مرحله اولیه عمدتاً به تکنیک‌های خودکار متکی بوده است.
- بازبینی و اصلاح دستی: پژوهشگران دستورات بازنویسی‌شده را به‌صورت دستی بررسی و ویرایش کرده‌اند تا کیفیت و تنوع آن‌ها تضمین شود.
- برای تولید پرامپت‌ها از روش فیوشات پرامپتینگ استفاده می‌شود.
- بازنویسی خودکار و بازبینی دستی: برای افزایش تنوع، از یک روش نمونه‌گیری چندمرحله‌ای دیگر برای بازنویسی پرامپت‌ها استفاده شده است. در نهایت، هر دستور بازنویسی‌شده به‌صورت دستی بررسی و ویرایش شده است.

۴/۲ پیروی از دستورالعمل چندمرحله‌ای

در بخش پیروی از دستورات چندمرحله‌ای توضیح داده شده است.

۴/۳ دادگان چندنمونه‌ای

برچسب‌های مورد نیاز:

هر داده ممکن است یک یا چند برچسب داشته باشد

برچسب تک‌عنصری:

برای داده‌هایی که در آن‌ها مدل باید یک برچسب کلاسیک (مانند طبقه‌بندی احساسات یا استنباط زبان طبیعی) تولید کند، فقط یک برچسب وجود دارد.

مثال: در طبقه‌بندی احساسات (مثل SST-2) یا استنباط زبان طبیعی (مثل MNLI)، برچسب می‌تواند به‌عنوان یک کلمه یا عبارت معنادار که نمایانگر احساسات یا رابطه دو جمله باشد، استخراج شود.

برچسب‌های چندعنصری یا خالی:

در وظایفی مانند شناسایی موجودیت‌های نام‌دار یا درک مطلب ماشینی، مدل ممکن است پاسخ را به‌صورت چندین بازه استخراج کند. این بازه‌ها بسته به نوع سؤال و متن موجود ممکن است شامل چندین بخش از متن یا حتی خالی باشند.

مثال: در شناسایی موجودیت‌ها، جمله "باراک اوباما به برلین سفر کرد" می‌تواند دو بازه "باراک اوباما" (شخص) و "برلین" (مکان) را به عنوان پاسخ تولید کند.

۲. تعداد برچسب‌گذاران:

برای افزایش دقت و کیفیت برچسب‌گذاری‌ها، هر مثال سه برچسب‌گذار مختلف انجام می‌شود. این کار به منظور اندازه‌گیری توافق بین برچسب‌گذاران و اطمینان از کیفیت داده‌ها صورت می‌گیرد. این روش باعث می‌شود که از هرگونه خطای انسانی احتمالی جلوگیری شده و داده‌ها از اعتبار بالایی برخوردار باشند.

۳. دستورالعمل‌های برچسب‌گذاری:

برای اطمینان از یکپارچگی و دقت در برچسب‌گذاری، از دستورالعمل‌های مشخص برای برچسب‌گذاران استفاده می‌شود. این دستورالعمل‌ها شامل موارد زیر است:

توضیح کوتاه وظیفه:

برچسب‌گذاران قبل از آغاز برچسب‌گذاری، یک توضیح کوتاه از وظیفه‌ای که باید انجام دهند دریافت می‌کنند. این توضیحات همراه با تعاریف دقیق از برچسب‌ها و مثال‌های واضح برای هر یک از آن‌ها است تا برچسب‌گذار درک دقیقی از برچسب‌ها و کاربرد آن‌ها داشته باشد.

مثال‌های آموزشی Few-Shot:

در سناریوهای Few-Shot، برچسب‌گذاران ابتدا چند مثال آموزشی (بین ۱۰ تا ۳۰ مثال) دریافت می‌کنند. پس از هر مثال، پاسخ‌های صحیح به آن‌ها نشان داده می‌شود تا آن‌ها با نحوه برچسب‌گذاری و کاربرد هر برچسب آشنا شوند.

سناریوی Zero-Shot:

برای مقایسه عملکرد انسانی، برخی از برچسب‌گذاران تنها توضیح وظیفه را دریافت می‌کنند بدون آنکه هیچ‌گونه مثال آموزشی ارائه شود. این روش برای ارزیابی عملکرد در شرایط ناشناخته و بدون آموزش قبلی مفید است.

۵ مراجع

[1] ,. K. S. Y. H. Yiwei Qin, "INFOBENCH: Evaluating Instruction Following Ability in Large Language Models," *arXiv,* 2024.

[2] X. L. G. Z. S. H. H. C. G. Y. C. M. A. H. A. J. G. Subhabrata Mukherjee, "CLUES: Few-Shot Learning Evaluation in Natural Language Understanding," *arXiv:2111.02570 [cs.CL],* 2021.

[3] J. Z. T. Lu, "Instruction-Following Evaluation for Large Language Models," *arXiv,* 2023.

[4] K. Z. W. Y. Renze Lou, "Large Language Model Instruction Following: A Survey of Progresses and Challenges," *arXiv,* 2024.

[5] K. Y. Elliot L. Epstein, "MMMT-IF: A CHALLENGING MULTIMODAL MULTITURN INSTRUCTION FOLLOWING BENCHMARK," *arXiv,* 2024.

[6] D. J. C. W. Yun He, "Multi-IF: Benchmarking LLMs on Multi-Turn and Multilingual Instructions Following," *2024.*