بنیاد ملی علم ایران

عنوان فراخوان

نحوه توسعه مدلهای زبانی بزرگ مناسب برای زبان فارسی و قابل استفاده در یک چت‌بات سازمانی

توضیحات فراخوان

هدف اصلی این پژوهش نیازمندی شرکت در نحوه استفاده از LLMهای فارسی در تکامل محصولات چت‌بات جاری است. که با استفاده از آن بتوان نیازهای متنوع مشتریان چت‌بات را پاسخ داد. مجری باید بتواند با استفاده از LLMهای موجود (داخلی و خارجی)، راه‌حل‌های مجزایی به شرکت ارائه نماید که بتوان برای هر مشتری بصورت جداگانه چت‌بات مخصوص آن را ساخت. این تحقیق باید بتواند کارهایی از قبیل پرسش و پاسخ، تشخیص مقصود، جداسازی اسلات‌ها، تشخیص مفهوم متن، خلاصه‌سازی، جستجوی معنایی و ... را با استفاده از مدل‌های زبانی بزرگ (LLM) انجام دهد.

تبیین و تشریح مسئله پژوهشی

درخصوص LLMها، مدلهای Open Source نسبتاً زیادی وجود دارد و برخی از آن‌ها فارسی را نیز پشتیبانی می‌کنند. نیاز فعلی ما توسعه LLM فارسی و عمومی نیست بلکه هدف آن است که از یک LLM فارسی (چه این سورس خارجی و یا نسخه‌های فارسی موجود داخلی) و چه سرویس‌هایی مانند ChatGPT استفاده کرد و چت‌بات‌های منطبق با نیاز مشتریان را توسعه داد. لذا در این پروژه به دنبال Fine Tuning کل مدل LLM نیستیم و هدف، استفاده از LLM ها برای پاسخ‌دهی به نیازهای متنوع مشتریان است. حال ممکن است در این راه نیاز به Fine Tune هم باشد که این عملیات یک Fine Tunning بسیار کوچک و در قالب دامنه محتوایی مشتری انجام می‌شود که بتوان یک چت‌بات مخصوص آن مشتری تهیه نمود. لذا LLM پایه‌ای انتخابی می‌تواند هرکدام از LLMهای Open Source داخلی یا خارجی باشد، و اصلاً یکی از وظایف مجری مقایسه LLMهای موجود و انتخاب بهترین آن برای پاسخ‌دهی به نیاز مشتری می‌باشد.

مجری لازم است که نحوه بکارگیری مدل‌های زبانی بزرگ را برای هر مشتری بصورت جداگانه انجام دهد. حال این روش می تواند از Zero Shot تا Fine Tune به تناسب نیاز هر مشتری متفاوت باشد. لازم به ذکر است مجری باید

معماری و نحوه انجام این عملیات را بصورت کامل برای کارفرما انجام دهد، تا کارفرما توانایی انجام این نوع کارها را برای سایر مشتریان نیز بدست آورد. بطور خلاصه چند نمونه از نیاز مشتریان به مجری اعلام و مجری برای هر کدام بصورت جداگانه باید یک LLM و همچنین راهکار استفاده بهینه از آن را ارائه دهد.

پژوهشگر بایستی بتواند در این پژوهش راهکار عملیاتی با ارائه نتایج پژوهشی مناسب برای هر یک از بخشهای زیر که برای مشتریان متفاوت است، را ارائه دهد:

۱- Open domain یا close domain باشد

۲- Inter personal باشد

۳- Task oriented یا chitchat باشد

۴- LLM based باشد که در همه حالتهای ۵گانه زیر قابلیت ارائه راهکار مناسب باشد:

a. ارائه راهکار NLU مبتنی بر LLM

b. ارائه راهکار NLG مبتنی بر LLM

c. استفاده از zero/few shot برای چت‌بات

d. ارائه LLM فارسی fine tune شده برای یک چت‌بات

۵- ارائه روش مناسب مبتنی بر RAG برای چت‌بات فارسی

۶- بهترین مدلهای LLM فارسی موجود

۷- مدلهای LLM کد باز با معماری قابل استفاده مجدد فارسی

۸- نحوه استفاده از LLM در محصولات چت‌بات فارسی

۹- نحوه بهره‌برداری از LLM در افزایش دقت محصولات چت‌بات فارسی

۱۰- سایر سرویسها از قبیل

a. تشخیص مقصود

b. جداسازی اسلاتها

c. تشخیص موجودیت

d. خلاصه‌سازی

e. انتخاب بهترین پاسخ بین چند پاسخ
f. جستجوی معنایی
g. ساخت پاسخ
h. تشخیص مفهوم
i. Topic Modelling
j. Question Answering
k. Data Argumentation

چالش‌های کلیدی نیاز فناورانه

یکی از موانع جدی این تحقیق، دقت مدل‌های LLM در زبان فارسی است. با توجه به اینکه در دنیا LLMهای زیادی بصورت چندزبانه وجود دارد، ولی دقت این مدل‌ها در زبان فارسی بسیار کمتر از سایر زبان‌های رایج از قبیل انگلیسی و زبان‌های لاتین است. لذا افزایش دقت این مدل‌ها در زبان فارسی از اهمیت زیادی برخوردار است. لازم به ذکر است که افزایش دقت بصورت عمومی در زبان فارسی مدنظر نیست. بلکه افزایش دقت در زبان فارسی برای کاربردهای مختلف مشتریان کارفرما مدنظر است. و این افزایش دقت باید در پاسخ‌دهی به سؤالات کاربران در محتوای مخصوص هر مشتری ایجاد شود.

چالش دوم و بسیار مهم، Fine Tune کردن یا آموزش مجدد این مدل‌های زبانی بزرگ است که باید طوری انجام شود که نیاز به حداقل سخت‌افزار GPU داشته باشد. و بصورت Agile باشد که بتوان برای هر نوع داده که از مشتری دریافت می‌گردد ما با سرعت مناسبی آموزش مجدد را انجام دهیم.

چالش سوم مقدار مصرف GPU در این مدل‌ها است. در تحقیق باید دقیقاً مشخص شود که برای افزایش سرعت این مدلها چه راهکارهایی وجود دارد.

چالش چهارم نحوه بکارگیری LLMها برای هر کدام از مشتریان است. مشتریان ممکن است نیازمندی‌های متفاوتی داشته باشند که نتوان با یک نوع LLM (حتی Fine Tune شده) نیز نتوان به تمامی آن‌ها پاسخ داد. لذا باید برای هر نیاز مشتری، یک راهکار متناسب با آن نیاز ارائه شود تا به بهترین دقت رسید.

گام‌های پژوهشی

مراحل پژوهش بصورت ذیل پیشنهاد می گردد:

- **فاز ۱:** تحقیقات اولیه در مورد مدل‌های زبانی بزرگ، مدل‌های متن‌باز، زیرساخت‌های مشابه جهانی

خروجی فاز: گزارش‌های کامل از تحقیقات اولیه، مقایسه مدل‌های مختلف از نظر دقت، سرعت و قابلیت

- **فاز ۲:** طراحی معماری زیرساخت و یکپارچه‌سازی و اتصال آن

خروجی فاز: ارائه معماری کامل، نحوه ارتباط اجزای مختلف، نحوه ارتباط با ماژول‌های بیرونی

- **فاز۳:** پیاده‌سازی نسخه‌ی اول زیرساخت

خروجی فاز: تحویل نرم‌افزار اجرایی بصورت عملیاتی

- **فاز ٤:** پیاده‌سازی و تحویل نسخه‌ی نهایی زیرساخت

خروجی فاز: تحویل نرم‌افزار اجرایی بصورت عملیاتی و اشکال‌زدایی شده

خروجی پژوهش

خروجی‌های این پروژه شامل موارد زیر خواهد بود:

- کدهای نرم‌افزاری پروژه: شامل تمامی برنامه‌های نوشته شده به زبان‌های برنامه‌نویسی مختلف، تنظیمات برنامه‌نویسی، تمامی مخازن پروژه‌های برنامه‌نویسی و پکیج‌های آن‌ها
- طراحی معماری: شامل شرح و طراحی معماری کلی سیستم نهایی و مؤلفه‌های مختلف آن و نحوه و پروتکل ارتباطی آن‌ها
- مستندات اتصال و یکپارچه‌سازی: شامل مستندات لازم برای یکپارچه‌سازی سیستم با سایر سیستم‌ها مانند API Reference
- مستندات استقرار: مستندات راهنمای استقرار سیستم اعم از منابع سخت‌افزاری و شبکه‌ای مورد نیاز، برنامه‌های سیستمی لازم برای راه‌اندازی پروژه و سیستم عامل لازم برای استقرار
- پژوهش‌های علمی در راستای مدل‌های زبانی: تمامی پژوهش‌هایی که در راستای شناخت بهتر ظرفیت‌های مدل‌های زبانی و مدل‌های زبانی مولد برای اجرای این پروژه صورت گرفته، اعم از مطالعات تطبیقی یا توسعه‌ی دانش جدید در این حوزه

تسهیم مالکیت فکری

- **مالکیت معنوی**: مجری در مالکیت معنوی ناشی از اجرای پژوهش سهیم خواهد بود و انتشار مقاله مشترک توسط مجری و متقاضی در ژورنال‌های داخلی و خارجی، ارائه مقاله در کنفرانس‌ها و سمینارها با موافقت و اشاره به‌نام همه دست‌اندرکاران مجاز خواهد بود.
- **مالکیت منافع مادی**: با توجه به مدل کسب‌وکار و اجرا و اثبات دستاوردهای حاصل از طرح توسط شرکت متقاضی، منافع مالی ناشی از توسعه این فناوری برای شرکت متقاضی خواهد شد اما مطابق تراضی بین شرکت متقاضی و مجری، قابل اشتراک بین آنها خواهد بود.

نحوه پذیرش

پذیرش طرح‌ها رقابتی است و از بین پروپوزال‌های دریافتی، موردی که شرایط زیر را داشته باشد، در اولویت خواهد بود:

۱. ترکیب متخصصین تیم پیشنهادی مرتبط باشد.

۲. افراد پیشنهادشده، دارای سابقه پژوهشی و فنی در آن موضوع باشند.

۳. زمان‌بندی، هزینه و شرح خدمات، متناسب و مرتبط با پژوهش موردتقاضا باشد. (در این بخش، مجری می‌تواند برآورد اولیه خود را اعلام کند اما بدیهی است جزئیات اجرایی در ابتدای امر مشخص نیست و مجری و کارفرما با علم به این موضوع وارد این توافق خواهند شد)

٤. پروپوزال، طبق فرمت پیشنهادی بنیاد، تهیه و از طریق سامانه کایپر ارسال شده باشد.

٥. فونت حروف و اعداد فارسی B Nazanin و اندازه قلم ۱۳ و فونت حروف و اعداد انگلیسی، Times New Roman و اندازه قلم ۱۱ باشد.

هزینه‌های قابل قبول

- ✓ حق‌التحقیق نیروی انسانی؛

- ✓ هزینه‌های نرم‌افزاری؛
- ✓ تست‌ها و آنالیزها؛
- ✓ خدمات؛

حوزه های اولویت دار

مهندسی کامپیوتر/هوش مصنوعی

واجدین شرایط

پژوهشگر اصلی تیم لازم است عضو هیئت‌علمی فعال یکی از دانشگاه‌ها و مؤسسات آموزش عالی کشور باشد. پس از دریافت پروپوزال از طریق سامانه، ارزیابی انجام گرفته و در صورت کسب امتیاز بالا، تیم برگزیده جهت مذاکره با بنیاد و شرکت متقاضی دعوت خواهد شد.

فایل پیوست

فرم درخواست پیشنهاده (پروپوزال)

متن فراخوان

تاریخ فراخوان

کلیه افراد واجد شرایط تا ۳۱ فروردین فرصت دارند که پروپوزال خود را از طریق سامانه کاپیر برای بنیاد ملی علم ایران ارسال نمایند.

توجه: تاریخ این فراخوان تمدید نخواهد شد و فقط پروپوزال‌های ارسالی در بازه زمانی اعلام شده در فراخوان، به مرحله داوری خواهند رفت.

مبلغ حمایت

پژوهش پیشنهاد شده تا سقف ۸۰ درصد، حداکثر ۲/۵ میلیارد تومان، توسط بنیاد ملی علم ایران حمایت خواهد شد.

بدیهی است که مابقی هزینه‌ها باید توسط شرکت متقاضی ارائه دهنده پژوهش تأمین شود.

شیوه ثبت نام و ارسال درخواست

متقاضیان گرامی جهت ثبت‌نام می‌توانند به سامانه کایپر به نشانی rtms.insf.org مراجعه و از طریق بخش متقاضیان/ پژوهشگران اقدام نمایند. درصورتی‌که در این سامانه پروفایل مشخصات فردی ندارید ابتدا ثبت‌نام نموده و سپس به‌وسیله نام کاربری (Email) و رمز عبور اعطا شده وارد سامانه شوید. پس از ورود در بخش ارسال طرح جدید می‌توانید از کارتابل پژوهش عمیق شرکت‌های دانش‌بنیان اقدام به ارسال طرح نمایید.

مسئول پاسخگو

پژوهشگران محترم پس از مطالعه توضیحات فراخوان و آیین نامه‌های مربوطه در پورتال بنیاد علم، در صورت داشتن هر گونه ابهام یا سؤال در خصوص فرایند ارسال طرح، شرایط و محتوای علمی فراخوان می‌توانند از پروفایل خود در سامانه کایپر با کارگروه دانش‌بنیان از طریق تیکت، یا با ایمیل jandili.a@insf.org سؤالات خود را مطرح نمایند و یا با شماره تلفن ۰۲۱۸۲۱۶۱۱۵۰(آقای جندیلی) تماس حاصل فرمایند.