# آموزش مدل‌های زبانی برای پیروی از دستورالعمل‌ها با بازخورد انسانی

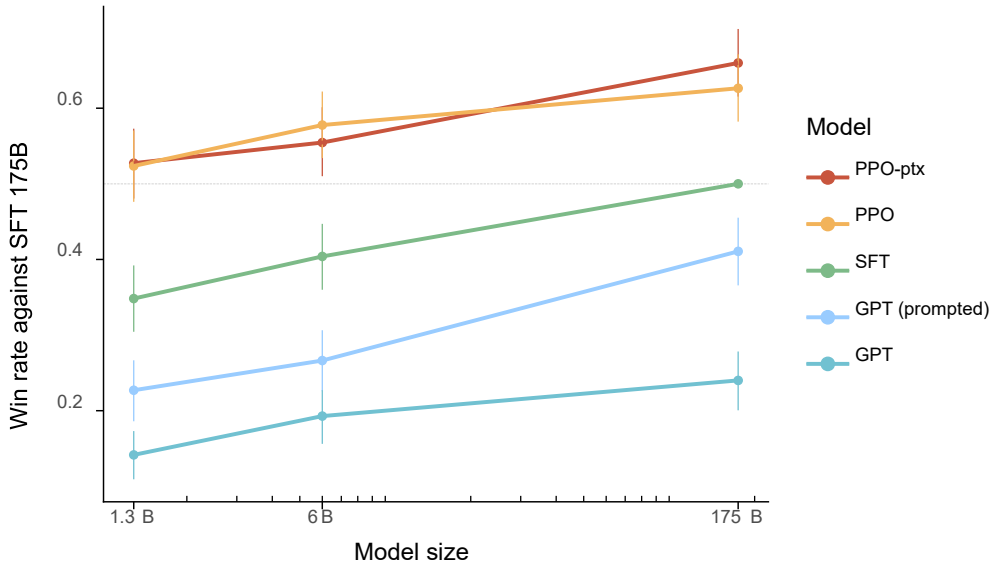
ساختن مدل‌های زبانی بزرگ به تنهایی باعث بهبود عملکرد آنها در دنبال کردن نیاز کاربر نمی‌شود. به عنوان مثال، مدل‌های زبانی بزرگ ممکن است خروجی‌هایی ایجاد کنند که ناصحیح، سمی یا به طور ساده، برای کاربر مفید نباشند. به عبارت دیگر، این مدل‌ها با نیازهای کاربران هم‌تراز نیستند. در این مقاله، روشی برای هم‌تراز کردن مدل‌های زبانی با نیاز کاربر با استفاده از بازخورد انسانی ارائه شده است. با شروع از مجموعه‌ای از پیشنهادهای نوشته شده توسط برچسب‌گذار و پیشنهادهای ارسال شده از طریق API مدل زبانی، مجموعه‌ای از نمونه‌های دمویی رفتار مورد نظر مدل توسط برچسب‌گذار جمع آوری شده است. سپس با استفاده از یادگیری نظارت شده، مدل GPT-3 را fine-tune کرده‌ایم. به دنبال آن، مجموعه‌ای از رتبه‌بندی خروجی مدل توسط بازخورد انسانی جمع آوری شده و با استفاده از یادگیری تقویتی، مدل Supervised را نیز بهبود داده‌ایم. نتیجه گرفته شده، مدل InstructGPT است. در ارزیابی انسانی بر روی توزیع prompt ما، خروجی‌های حاصل از مدل InstructGPT با ۱.۳ Bپارامتر ترجیح داده شدند نسبت به خروجی‌های مدل GPT-3با ۱۷۵ Bپارامتر، با وجود داشتن ۱۰۰ برابر کمترین پارامتر. علاوه بر این، مدل‌های InstructGPT بهبودهایی در صداقت و کاهش ساخت خروجی سمی را نشان می‌دهند، در حالی که در مجموعه‌داده‌های NLP عمومی عملکرد کمتری نشان نمی‌دهند. با این حال، با وجود این‌که مدل InstructGPT هنوز اشتباهات ساده را می‌کند، نتایج نشان می‌دهند که fine-tuning با بازخورد انسانی یک جهت آینده‌گرا برای هم‌تراز کردن مدل‌های زبانی با نیاز کاربر است.

سخن دانشجو: (در این مقاله به دنبال آن هستیم که مدل زبانی را ارائه کنیم که فارق از حجم و ابعاد آن بعد از استفاده و در سیستم پاسخی درست و صادقانه ارائه کند به همین منظور روشی مطرح می‌گردد با عنوان "آموزش مدل زبانی با بازخورد انسانی" به این صورت که فرد خبره مجموعه از نمونه های رفتاری مورد نظر را برچسب گذاری کرده و در نهایت آن را با استفاده از یادگیری نظارت شده در مدل GPT-3 را با داده های برچسب گذاری شده، fine-tune می‌کند. بعد از آن مجموعه از رتبه بندی مدل برای پاسخ ها با بازخورد انسانی انجام شده که با استفاده از یادگیری عمیق (ژرف) مدل یادگیری با ناظر را تقویت میکند که حاصل ان مدل InstructGPT است که با ۱.۳ میلیارد پارامتر نتیجه بهتری به نسبت GPT-3 داشته است که با ۱۷۵ میلیارد پارامتر آموزش داده شده است.

این مقاله به دنبال این است که نشان دهد یک NLP زمانی که با داده های برچسب گذاری توسط فرد خبره fine-tune شود نتیجه بهتری دارد و عملکرد آن بهتر است)

مدل‌های زبانی بزرگ (LMs) با دادن نمونه‌هایی از یک وظیفه به عنوان ورودی، می‌توانند برای انجام یک دامنه گسترده از وظایف پردازش زبان طبیعی(NLP) ، مورد استفاده قرار بگیرند. با این حال، این مدل‌ها اغلب رفتارهای ناخواسته مانند اختراع واقعیت‌های جدید، تولید متن متحیز و سمی، یا عدم رعایت دستورالعمل‌های کاربر را از خود نشان می‌دهند.

سخن دانشجو: (این به دلیل این است که هدف مدلسازی زبان، پیش‌بینی کلمه بعدی در یک دنباله با داشتن کلمات قبلی است و دقیقاً با نتیجه مورد نظر کاربر برای یک وظیفه خاص، هم‌خوانی ندارد. علاوه بر این، داده‌های آموزشی استفاده شده برای آموزش این مدل‌ها، اغلب حاوی تعصبات و عدم دقت است که ممکن است در فرایند تولید مدل، تشدید شود. برای حل این موضوعات، پژوهشگران به تکنیک‌های مختلفی از جمله مهندسی پرسمان، تنظیم دقیق روی داده‌های خاص دامنه و ادغام محدودیت‌ها و رهنمودها برای تولید زبان، پرداخته‌اند. این رویکردها هدف دارند برای وظایف خاص، کنترل و دقت LMs را بهبود بخشند و رفتارهای ناخواسته را کاهش دهند. به عنوان یک ربات چت AI، من نیز با استفاده از تکنیک‌های مشابه، طراحی شده‌ام تا اطمینان حاصل شود که پاسخ‌های من مربوط، دقیق و با احترام به کاربران هستند. با این حال، مانند هر فناوری دیگری، مهم است محدودیت‌های AI را در نظر گرفته و تلاش برای بهبود قابلیت‌های آن را ادامه دهیم، در عین حال که احتمال مسائل اخلاقی مرتبط را مورد توجه قرار دهیم.)




شکل ۱: ارزیابی‌های انسانی مدل‌های مختلف بر روی توزیع پرسمان API، با بررسی اینکه در چه میزان خروجی‌های هر مدل نسبت به مدل 175B SFT ارجحیت دارند. مدل InstructGPT که نسحه با پیش پردازش شده است (PPO-ptx) و نسخه آن بدون پیش پردازش مخلوط (PPO)، به شدت عملکرد بالاتری نسبت به مدل‌های پایه GPT مخصوصا GPT prompted , GPT-3 دارند؛ خروجی‌های مدل 1.3B PPO نیز به خروجی‌های مدل 175B GPT-3 ترجیح داده می‌شود. تابع زیان در سراسر مقاله بازه اطمینان ۹۵٪ را نشان می‌دهند.