به نام خدا

بررسی مدل‌های زبانی بزرگ LLMs معماری، کاربردها و چالش‌ها

تحلیل و مقایسه ۶ مقاله پژوهشی

دانشجو: نوید نژادنوراله

استاد راهنما: دکتر احمد شریف

مقاله اول: Summary of ChatGPT-Related Research

موضوع اصلی: این مقاله پژوهشی درباره قابلیت‌ها و محدودیت‌های مدل‌های زبانی بزرگ (LLM)، به‌ویژه ChatGPT (نسخه‌های GPT-۳.۵ و GPT-۴) است.
تمرکز اصلی بر روی برنامه‌های کاربردی در حوزه‌های آموزش، پزشکی، علوم انسانی و علوم پایه است.
ویژگی‌ها:
ارائه تحلیل از ۱۹۴ مقاله مرتبط.
ارزیابی عملکرد ChatGPT در حوزه‌هایی مثل ریاضیات و فیزیک.
بررسی جنبه‌های اخلاقی و کاربردی مدل‌ها.
مزایا:
تحلیل گسترده از حوزه‌های مختلف.
استفاده از داده‌های آماری جامع.
معایب:
تمرکز محدود بر معماری و جزئیات فنی مدل‌ها.
نپرداختن عمیق به مشکلات خاص مثل محدودیت‌های محاسباتی.

مدل‌های بررسی‌شده:

ChatGPT (GPT-۳.۵ و GPT-۴):

این مدل‌ها از معماری Transformer استفاده می‌کنند و با استفاده از یادگیری پیش‌نمایشی بر اساس داده‌های گسترده وب آموزش دیده‌اند.

ویژگی‌ها:

توانایی درک زبان طبیعی و تولید متن.

قابلیت‌هایی مثل یادگیری از بازخورد انسانی (Reinforcement Learning from Human Feedback - RLHF).

کاربردها:

در زمینه‌های مختلفی مثل آموزش، علوم پزشکی، تعامل انسان و ماشین، و تولید محتوا استفاده شده‌اند.

InstructGPT:

نسخه‌ای پیشرفته‌تر از GPT که به‌طور خاص برای دریافت دستورالعمل‌های انسانی طراحی شده است.

تمرکز بر یادگیری از بازخورد کاربران برای بهبود پاسخ‌ها.

تشریح مدل‌ها:

Transformer Architecture:

اساس معماری مدل‌های GPT است که از مکانیزم Self-Attention برای پردازش متون استفاده می‌کند.

RLHF:

یکی از ویژگی‌های کلیدی مدل‌ها، که باعث تنظیم مدل بر اساس اولویت‌ها و ارزش‌های انسانی می‌شود.

مقاله دوم: A Review on Large Language Models

موضوع اصلی: این مقاله به بررسی کلی مدل‌های زبانی بزرگ، معماری‌ها، تاریخچه، چالش‌ها و کاربردهای آنها می‌پردازد.
ویژگی‌ها:
توضیحات جامع در مورد معماری مدل‌ها (مثل Transformers) و فرآیندهای یادگیری.
دسته‌بندی کاربردها در حوزه‌هایی چون پزشکی، آموزش، کشاورزی، و تجارت.
مزایا:
تحلیل دقیق معماری مدل‌ها و منابع داده.
ارائه دیدگاه‌های دقیق در مورد مشکلات فنی و آینده‌پژوهی.
معایب:
کمتر به مثال‌های عملی و ارزیابی عملکرد مدل‌ها پرداخته شده است.
ارزیابی‌های آماری محدودتر نسبت به مقاله اول.

مدل‌های بررسی‌شده:

1. **BERT (Bidirectional Encoder Representations from Transformers):**
    1. مدلی دوجهته که برای درک متن در دو جهت (قبل و بعد از یک کلمه) طراحی شده است.
2. **کاربردها:**
    1. پاسخ به سوالات و استنتاج زبان.
2. **GPTسری‌ها (GPT-2, GPT-3, GPT-3.5, GPT-4):**
    1. تمرکز بر تولید متن طبیعی و پاسخ‌دهی در مکالمات.
3. **T5 (Text-to-Text Transfer Transformer):**
    1. یک مدل تبدیل متن به متن که برای انجام کارهای مختلف NLPطراحی شده است.
4. **RoBERTaو DistilBERT:**
    1. نسخه‌های بهینه‌سازی‌شده و ساده‌شده از BERTبرای کارایی بیشتر.

2. **تشریح مدل‌ها:**

1. **تاریخچه:**

1. توسعه مدل‌ها از رویکردهای آماری (مانند مدل‌های n-gram) آغاز شده و با ظهور شبکه‌های عصبی پیشرفته‌تر، به مدل‌های دوجهته و خودتوجهی گسترش یافته است.

2. **مقیاس‌پذیری:**

1. مدل‌های بزرگ‌تر با پارامترهای بیشتر عملکرد بهتری در وظایف زبان طبیعی نشان داده‌اند.

مقاله سوم: A Survey on Large Language Models

1. موضوع اصلی: این مقاله به مرور پیشرفت‌ها، چالش‌ها و کاربردهای عملی مدل‌های زبانی بزرگ می‌پردازد.
2. ویژگی‌ها:
    1. تأکید بر مسائل عملی، مثل استفاده از مدل‌ها در حوزه‌های پزشکی و مالی.
    2. بررسی محدودیت‌ها و مشکلات اخلاقی، مثل سوگیری و نیازهای محاسباتی.
3. مزایا:
    1. نگاه جامع به کاربردهای عملی و چالش‌های پیاده‌سازی.
    2. ارائه تکنیک‌هایی برای بهبود قابلیت‌ها و کاهش مشکلات.
4. معایب:
    1. جزئیات محدودتر در مورد تاریخچه و معماری مدل‌ها.
    2. تمرکز کمتر بر ارزیابی‌های آماری.

مدل‌های بررسی‌شده:

GPT (Generative Pre-Trained Transformers):

شامل نسخه‌های GPT-۳، GPT-۳.۵ و GPT-۴.

تمرکز بر مکالمات و تولید متن.

LLaMA (Large Language Model Meta AI):

یک مدل پیشرفته برای پردازش زبان.

طراحی شده برای کاهش منابع محاسباتی.

BERT و ELMo (Embeddings from Language Models):

مدل‌هایی با تاکید بر بردارهای معنایی و درک متنی.

Flan-T۵ و Minerva:

مدل‌هایی که برای یادگیری دستورالعمل‌های خاص و حل مسائل پیچیده ریاضی طراحی شده‌اند.

تشریح مدل‌ها:

Masked Language Modeling (MLM):

تکنیکی که در برخی مدل‌ها مانند BERT استفاده می‌شود، که در آن کلمات به‌طور تصادفی مخفی می‌شوند و مدل باید آن‌ها را پیش‌بینی کند.

Multimodal Integration:

برخی مدل‌ها، مثل GPT-۴، قابلیت پردازش داده‌های چندگانه (تصویر و متن) را دارند.

مقاله چهارم: ChatGPT and Large Language Models in Academia: Opportunities and Challenges

تمرکز: استفاده از مدل‌های زبان بزرگ (LLMs) مانند ChatGPT در آکادمیا، همراه با فرصت‌ها و چالش‌های مرتبط.
فرصت‌ها:
بهبود کارایی در نوشتار علمی و آموزش.
ارائه‌ی ابزارهایی برای کمک به تحقیقات و برنامه‌نویسی.
چالش‌ها:
دقت پایین و احتمال تولید اطلاعات نادرست ("توهم").
مسائل اخلاقی مانند نحوه استفاده از متن‌های تولیدشده و احتمال تقلب.
نیاز به مستندسازی نحوه استفاده از این ابزارها.
کاربردها: کمک به نگارش مقالات، پیشنهاد ایده‌ها، و حتی طراحی پیشنهادات پژوهشی.

مدل‌ها:

1. **ChatGPT:**
    1. بر اساس معماری **GPT (Generative Pre-trained Transformer)** توسعه داده شده است.
    2. نسخه‌های مختلفی از آن وجود دارد:
        1. **GPT-3.5:** نسخه رایگان **ChatGPT** که قابلیت‌های اولیه پاسخ‌دهی و پردازش متن را دارد.
        2. **GPT-4:** نسخه پولی با قابلیت‌های پیشرفته‌تر، مانند پذیرش ورودی‌های تصویری.
    3. ویژگی‌ها:
        1. تعامل مکالمه‌ای و توانایی پیگیری مکالمات گذشته.
        2. تولید متن با کیفیت انسانی و پاسخ‌های منطقی.
        3. طراحی‌شده برای کاربردهای عمومی، از جمله نوشتن کد، پاسخ به سوالات، و ایجاد محتوای متنی.
2. مدل‌های جایگزین:
    1. **BLOOM:**
        1. یک مدل چندزبانه با ۱۷۶ میلیارد پارامتر.
        2. متن‌باز و قابل استفاده برای توسعه‌دهندگان.
    2. **Google Bard:**
        1. مدل دیگری که توسط گوگل توسعه یافته است و شباهت‌هایی با **ChatGPT** دارد.

- **جزئیات:**
- **قابلیت‌ها:**
  - تولید محتوای آکادمیک (مانند نوشتن مقاله یا مرور ادبیات).
  - کمک به برنامه‌نویسی و حل مسائل کدنویسی.
- **چالش‌ها:**
  - احتمال ارائه اطلاعات نادرست یا "توهم".
  - محدودیت‌هایی در دقت و صحت داده‌های علمی.

مقاله پنجم: Large Language Model-Based Chatbots in Higher Education

تمرکز: پتانسیل و کاربرد چت‌بات‌های مبتنی بر مدل‌های زبان بزرگ در آموزش عالی.
فرصت‌ها:
ایجاد تجربه‌های یادگیری شخصی‌سازی‌شده.
بهبود دسترسی و شمولیت در آموزش.
کمک به مربیان در طراحی مواد آموزشی و ارزیابی‌های فرمی و بازخورددهی.
چالش‌ها:
مشکلات اخلاقی مانند تقلب آکادمیک و مسائل مربوط به حریم خصوصی داده‌ها.
محدودیت‌های دقت اطلاعات تولید شده توسط AI.
مثال‌ها: استفاده از چت‌بات‌ها برای کمک به تولید مواد آموزشی، ارزیابی دانش‌آموزان و بازخورددهی.

مدل‌ها:

1. GPT (Generative Pre-trained Transformer):
    1. این مدل‌ها، پایه‌ای برای چت‌بات‌های مبتنی بر LLMهستند.
    2. توسعه مدل‌ها از) GPT-1در سال ۲۰۱۸) شروع شده و به GPT-4رسیده است.
    3. تعداد پارامترها:
        1. GPT-1: حدود ۱۱۷ میلیون پارامتر.
        2. GPT-2: 1.5میلیارد پارامتر.
        3. GPT-3: 175میلیارد پارامتر.
        4. GPT-4: تعداد دقیق پارامترها افشا نشده است، اما بر اساس یادگیری تقویتی با بازخورد انسانی ((RLHFتوسعه یافته است.
2. مدل‌های دیگر:
    1. BERT (Bidirectional Encoder Representations from Transformers):
        1. از معماری دوطرفه برای تجزیه‌وتحلیل متون استفاده می‌کند.
    2. RoBERTaو ALBERT:
        1. بهینه‌سازی‌شده برای سرعت و دقت بالاتر.
    3. T5 (Text-to-Text Transfer Transformer):
        1. برای تبدیل تمامی مسائل پردازش زبان به یک مسئله‌ی ورودی-خروجی متنی طراحی شده است.

- **جزئیات:**
- **ویژگی‌ها:**
  - استفاده در تولید مواد آموزشی، طراحی آزمون‌ها و ارزیابی‌ها.
  - شخصی‌سازی یادگیری از طریق مکالمات طبیعی.
- **چالش‌ها:**
  - افزایش تقلب آکادمیک.
  - مسائل مربوط به حریم خصوصی و دقت خروجی‌ها.

مقاله ششم: Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data

تمرکز: استفاده از LLMها برای ساخت چت‌بات‌هایی که اطلاعات خودگزارشی کاربران را جمع‌آوری می‌کنند.
فرصت‌ها:
جمع‌آوری داده‌ها از طریق مکالمات طبیعی.
ایجاد چت‌بات‌های منعطف و با قابلیت‌های پاسخگویی انسانی.
چالش‌ها:
طراحی مناسب پرامپت‌ها برای هدایت مکالمات به سمت جمع‌آوری اطلاعات مشخص.
احتمال اشتباه در مکالمات یا انحراف از موضوع اصلی.
کاربردها: جمع‌آوری داده‌های مرتبط با موضوعات سلامتی (مانند خواب، تغذیه، و ورزش) از طریق مکالمات چت‌باتی.

- مدل‌ها:

1. **GPT-3:**
    1. این مدل به عنوان زیرساخت اصلی چت‌بات‌ها استفاده شده است.
    2. قابلیت یادگیری از ورودی‌های متنی و ارائه پاسخ‌های طبیعی.
2. **HyperCLOVA:**
    1. یک مدل توسعه‌یافته توسط NAVERکه به صورت خاص برای زبان‌های آسیایی بهینه‌سازی شده است.
    2. برای ایجاد مکالمات طبیعی و محتوای بومی طراحی شده است.

- جزئیات:

- طراحی پرامپت‌ها:
    - مدل‌ها بر اساس طراحی پرامپت‌ها (Prompt Design)هدایت می‌شوند.
    - پرامپت‌ها شامل توضیحات شخصیت (مثلاً "یک متخصص خواب") و اسلات‌های اطلاعاتی (مثلاً زمان خواب) هستند.
- کاربردها:
    - جمع‌آوری داده‌های شخصی (مانند کیفیت خواب، رژیم غذایی، فعالیت بدنی).
    - تعاملات طبیعی با کاربران و کاهش بار ورودی دستی.

مقایسه کلی:

تمرکز موضوعی:
مقاله اول بیشتر بر روی عملکرد و کاربردهای ChatGPT متمرکز است.
مقاله دوم دیدگاهی جامع‌تر نسبت به معماری و تاریخچه مدل‌ها ارائه می‌دهد.
مقاله سوم بر جنبه‌های کاربردی و عملیاتی مدل‌ها تأکید دارد.
تحلیل آماری و داده‌ها:
مقاله اول از تحلیل داده‌ها و آمار گسترده‌ای استفاده می‌کند.
مقاله دوم بیشتر بر جنبه‌های مفهومی و تکنیکی تمرکز دارد.
مقاله سوم بین این دو قرار دارد، اما تمرکز بیشتری بر مشکلات عملی دارد.
کاربردها و چالش‌ها:
مقاله اول و سوم به مسائل کاربردی و چالش‌های اخلاقی پرداخته‌اند.
مقاله دوم تحلیل دقیق‌تری از معماری مدل‌ها و داده‌ها دارد.

مقایسه کلی:

- **ChatGPT و GPT-3:**
- هسته اصلی مقالات چهارم و ششم.
- بهبودهای نسخه‌های مختلف از جمله GPT-4 مورد تاکید است.
- **BERT و مدل‌های مشابه:**
- مقاله پنجم به بررسی مدل‌های دوجهته مانند BERT و مشتقات آن می‌پردازد.
- بر کارایی در درک زبان و پردازش متون پیچیده تمرکز دارد.
- **مدل‌های منطقه‌ای و تخصصی:**
- در مقاله ششم، مدل‌های خاصی مانند HyperCLOVA برای زبان‌ها یا حوزه‌های خاص بررسی شده‌اند.

جدول مقایسه‌ای مدل‌های زبانی بزرگ بر اساس سه مقاله

ویژگی‌ها	مقاله اول	مقاله دوم	مقاله سوم
مدل‌های بررسی‌شده	ChatGPT (GPT-3.5, GPT-4), InstructGPT	BERT, GPT (2, 3, 3.5, 4), T5, RoBERTa, DistilBERT	GPT (3, 3.5, 4), LLaMA, BERT, ELMo, Flan-T5, Minerva
مزایا	- گستردگی دامنه کاربرد. - استفاده از RLHF برای تعامل بهتر با کاربران.	- تحلیل عمیق معماری مدل‌ها. - پوشش گسترده تاریخچه و منابع داده.	- تمرکز بر مسائل عملی و کاربردی. - شناسایی چالش‌های خاص مثل منابع محاسباتی.
معایب	- اطلاعات محدود درباره معماری مدل‌ها. - بررسی ناقص محدودیت‌های محاسباتی.	- کمتر به مثال‌های عملی پرداخته شده است. - تحلیل آماری کمتر.	- اطلاعات محدود درباره تاریخچه مدل‌ها. - جزئیات کمتر درباره برخی جنبه‌های فنی.
کاربردها	- آموزش و یادگیری. - پزشکی (پاسخ به سوالات بیماران). - تعامل انسان و ماشین.	- پزشکی و بیومدیکال. - آموزش و کشاورزی. - تجارت و رسانه‌های اجتماعی.	- پزشکی. - امور مالی. - مهندسی. - تولید محتوا در چندین دامنه.
محدودیت‌ها و چالش‌ها	- سوگیری در پاسخ‌ها. - عملکرد پایین در مسائل پیچیده ریاضی و منطقی.	- نیاز به منابع محاسباتی بالا. - حساسیت به داده‌های کم و غیرمعمول.	- مصرف زیاد منابع محاسباتی. - مشکلات اخلاقی (مثل سوگیری و حفظ حریم خصوصی).
تکنیک‌های کلیدی	RLHF (یادگیری تقویتی از بازخورد انسانی).	Masked Language Modeling (MLM). Self-Attention Mechanism.	Multimodal Integration (ادغام داده‌های چندحالته). Masked Language Modeling (MLM).

جدول مقایسه‌ای مدل‌های زبانی بزرگ بر اساس سه مقاله دوم

ویژگی‌ها	مقاله چهارم	مقاله پنجم	مقاله ششم
مدل زبانی اصلی	ChatGPT (نسخه‌های 3.5 و 4)	GPT-1، GPT-2، GPT-3، GPT-4	GPT-3، HyperCLOVA
مدل‌های دیگر	- BLOOM - Google Bard	- BERT - RoBERTa - ALBERT - T5 - LaMDA - OPT	- هیچ مدل دیگری معرفی نشده
مزایا	- افزایش کارایی در نوشتار علمی و برنامه‌نویسی - کمک به مرور ادبیات و پیشنهاد ایده‌های پژوهشی	- شخصی‌سازی یادگیری برای دانشجویان - تسهیل در تولید مواد آموزشی و ارزیابی‌ها	- جمع‌آوری داده‌های شخصی از طریق مکالمات طبیعی - ایجاد چت‌بات‌های انعطاف‌پذیر و انسانی
معایب	- دقت پایین در برخی موارد و تولید اطلاعات نادرست ("توهم") - مشکلات اخلاقی و احتمال تقلب	- احتمال تقلب آکادمیک و کاهش صداقت علمی - دقت پایین و مسائل مربوط به حریم خصوصی	- طراحی پرامپت‌های پیچیده برای هدایت چت‌بات‌ها - احتمال انحراف مکالمات از موضوع اصلی
کاربردها	- نوشتن مقاله و مرور ادبیات - طراحی پیشنهادهای پژوهشی - کمک به برنامه‌نویسی	- طراحی برنامه‌های درسی - آماده‌سازی مواد آموزشی - ارائه بازخورد به دانشجویان	- جمع‌آوری اطلاعات مرتبط با خواب، تغذیه، فعالیت بدنی از کاربران - تعاملات طبیعی با کاربران

توضیح مختصر درباره هر مقاله:

مقاله اول:
بر کاربردهای عملی ChatGPT متمرکز است.
قابلیت‌های ویژه مانند تعاملات انسانی و استفاده در آموزش و پزشکی بررسی شده‌اند.
چالش‌های مرتبط با دقت در ریاضیات و موضوعات پیچیده مشخص شده است.
مقاله دوم:
بررسی تاریخی و تکنیکی مدل‌های زبانی از BERT تا GPT-۴.
توضیح در مورد معماری‌های پیشرفته مانند Transformer و تکنیک‌های جدید مانند MLM.
تمرکز بیشتر بر معماری و تکنیک‌های پیشرفته.
مقاله سوم:
به چالش‌ها و کاربردهای عملی مدل‌ها می‌پردازد.
استفاده از مدل‌های مدرن مانند LLaMA برای بهینه‌سازی منابع محاسباتی.
تأکید بر چالش‌های اخلاقی و منابع مورد نیاز برای توسعه مدل‌ها.

جمع‌بندی و مقایسه مدل‌ها

1. تمرکز معماری:
    1. مقاله اول بیشتر بر روی قابلیت‌های GPT-3.5 و GPT-4 تمرکز دارد.
    2. مقاله دوم تحلیل گسترده‌تری از تاریخچه و معماری مدل‌ها (مانند Transformer و BERT) ارائه می‌دهد.
    3. مقاله سوم بر مقایسه مدل‌ها از نظر کاربرد عملی و معماری تاکید دارد.
2. مقایسه مدل‌ها:
    1. مدل‌های سری GPT به‌دلیل اندازه و توانایی تولید متن طبیعی در مقیاس وسیع برجسته هستند.
    2. مدل‌های BERT و T5 برای وظایفی که نیاز به درک عمیق‌تر زبان دارند مناسب‌تر هستند.
    3. مدل‌های LLaMA و Flan-T5 بر کاهش منابع و بهینه‌سازی تمرکز دارند.

- هر مقاله جنبه متفاوتی از مدل‌های زبانی را تحلیل کرده است و ترکیب این دیدگاه‌ها می‌تواند درک جامع‌تری از وضعیت کنونی LLMs ارائه دهد.

نتیجه‌گیری:

- برای تحقیقات شما، اگر هدف بررسی کاربردها و محدودیت‌های عملی مدل‌هاست، مقاله سوم مناسب‌تر است. اگر به دنبال درک عمیق‌تری از معماری و تاریخچه مدل‌ها هستید، مقاله دوم بهترین انتخاب است. در نهایت، اگر به بررسی گسترده عملکرد مدل‌ها در حوزه‌های مختلف علاقه دارید، مقاله اول مناسب‌تر خواهد بود.

مقالات در گیت هاب به ترتیب زیر ذخیره شده

- **مقاله اول**: fdp.niam-۲۹۵۰۱۶۲۸۲۳۰۰۰۱۷۶S-۲.۰s-۱
- **مقاله دوم**: A_Review_on_Large_Language_Models_Architectures_Applications_Taxonomies_Open_Issues_and_Challenges.pdf
- **مقاله سوم**: A_survey_on_large_language_models_Applications,_challenges,_limitations.pdf
- **مقاله چهارم**: فایل: pdf.۹-۰۰۳۳۹-۰۲۳-۱۳۰۴۰s
- **مقاله پنجم**: فایل: - Based Chatbots in Higher Education.pdf-IedoM egaugnaL egraL - icgiY -۲۰۲۴ Advanced Intelligent Systems
- **مقاله ششم**: فایل: Leveraging_large_language_models_to_power_chatbots_for_collecting.pdf

ممنون از توجه زیبای شما