北邮 AI 星火科研助手说明手册

一、快捷入口

快捷入口 1: 使用图书馆主页中轮播图的"科研助手"链接,点击图片直接进入"科研助手",用学工号登录,密码同信息门户。

快捷入口 2: 在图书馆首页导航栏中,依次点选"信息资源"——"AI 科研助手",点击对应链接进入,用学工号登录,密码同信息门户。



二、使用限制

功能		限制/天,失败的不算		
		限制说明	普通	北邮
			用户	用户
成果调研	论文检索	不限	不限	不限
	对话	/	100/天	100/天
	综述生成	30 篇/次	3次/天	20 次/天
论文研读	论文研读	30 轮/篇	10 篇/天	50 篇/天
	多文档对比	5 篇/次	5次/天	50 次/天
学术写作	文本翻译	2000 字符/次	不限	不限
	文本润色	2000 字符/次	不限	不限
	文档翻译	300s/篇, 6000 字/篇	5 篇/天	20 篇/天

三、使用说明

(一) 成果调研

星火科研助手中的"成果调研"功能旨在帮助用户快速 了解某一研究方向的论文成果、学者和研究机构,具体包括 如下功能:

综述生成:选择要生成综述的文章,一键生成综述。点击综述正文中的参考文献链接可查看论文的摘要等信息。

学者论文调研:调研指定学者的论文成果。

研究方向论文调研:调研指定研究方向的论文成果。例如,搜索"谣言检测"显示结果后,右边对话框中输入:帮 我汇总一下研究"谣言检测"的相关论文。

学者论文总结:对指定学者的论文进行概括总结。

研究方向论文总结:对指定研究方向的论文进行概括总结。例如,搜索"故事生成"显示结果后,右边对话框中输入:故事生成领域最新的研究是什么,有什么优缺点?总结一下。

研究方向学者/机构推荐:针对指定研究方向推荐学者或机构。例如:搜索"文本生成"显示结果后,右边对话框中输入:近年研究文本生成领域的机构有哪些?

以下以"量子计算"为例——



图 1 成果调研使用界面

如图 1 所示,当在搜索框输入"量子计算"后,星火科研助手将在左侧呈现出相关文献,并在右侧的对话框里提供对左侧文献的概要总结。通过选中左侧文献并点击"综述生成"按键,即可快速获得该领域文献的前沿概况。目前,星火科研助手最多支持对 30 篇选中的论文生成综述。

用户可以通过在对话框中进行提问来获取研究方向或 学者的论文成果情况。例如,可以询问"国内外量子计算的 机构有哪些?请列出前 10 名"通过点击末尾的"论文循证" 按键,系统可在左侧显示对话框回复内容涉及到的学术论文。

(二) 论文研读

星火科研助手中的"论文研读"功能支持用户上传论文进行交互式问答,主要功能如下:

论文知识问答:根据论文中内容进行回答。

论文外问答: 开放式问答, 需要论文外知识。

片段翻译、总结:对选中片段进行翻译或总结。

以 OpenAI 经典论文 InstructGPT 为例——

在完成论文上传后,系统会自动用中文给出论文的摘要、方法概述和结论信息(如图2所示)。在论文研读对话框中,用户可以询问"这篇论文的创新点是什么?""这篇论文有哪些贡献"等论文内知识(如图3所示)。



图 2 论文研读界面

这篇论文的主要创新点是什么

这篇论文的主要创新点是使用人类偏好数据(即InstructGPT)对大型语言模型进行微调,以使模型能够更好地遵循用户意图。通过使用来自标签者的反馈作为奖励信号来微调模型,作者成功地将模型的行为与特定人群(主要是标签者和研究人员)的陈述偏好相一致,而不是更广泛的"人类价值观"。此外,作者还通过在公共NLP数据集上进行评估,展示了这些模型在遵循指令、生成适当输出以及在特定任务上的改进等方面相对于GPT-3的优势。然而,尽管取得了一定的进展,但这些模型在安全性和可靠性方面仍有很大的改进空间。

常性方面仍有很大的改进空间。

② 重新生成

② 重新生成

② 1 Introduction

请输入内容(enter 发送,enter + shift 换行)



发送

图 3 论文内知识问答

通过对论文中片段进行选中,系统可以帮助用户获得片 段摘要或将片段翻译成中文或英文,并可以在对话框中引用 文本进行针对片段的询问(如图 4 所示)。

Abstract

Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not aligned with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through the OpenAI API, we collect a dataset of labeler demonstrations of the desired model behavior, which we use to fine-tune GPT-3 using supervised learning. We then collect a dataset of rankings of model outputs, which we use to further fine-tune this supervised model using reinforcement learning from human feedback. We call the resulting models InstructGPT. In human evaluations on our prompt distribution, outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3, despite having 100x fewer parameters.

Moreover, InstructGPT models show improvements in truthfulness and reduction in toxic output generation while having minimal performance regressions of NLP datasets. Even though InstructGPT still makes simple mistakes, our show that fine-tuning with human fee

language models with human intent.

提取摘要 翻译文本 引用文本

添加笔记

1 Introduction

Large language models (LMs) can be "prompted" to perform a range of natural language processing (NLP) tasks, given some examples of the task as input. However, these models often express unintended behaviors such as making up facts, generating biased or toxic text, or simply not following user instructions (Bender et al., 2021; Bommasani et al., 2021; Kenton et al., 2021; Weidinger et al., 2021; Tamkin et al., 2021; Gehman et al., 2020). This is because the language modeling objective

英语

中文

图 4 划线功能展示

除了论文内知识问答,用户也可以在对话框中询问论文 外的知识内容,例如"介绍 GPT 模型的发展历程"(如图 5 所示)。

^{*}Primary authors. This was a joint project of the OpenAI Alignment team. RL and JL are the team leads. Corresponding author: lowe@openai.com

[†]Work done while at OpenAI. Current affiliations: AA: Anthropic; PC: Alignment Research Center.

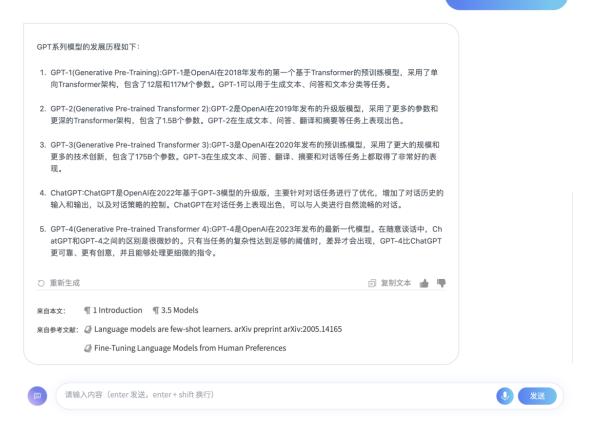


图 5 论文外知识问答

(二) 学术写作

星火科研助手中的"学术写作"功能具有学术翻译和英文润色能力,可以帮助用户更好地完成科研写作任务。

以李德毅院士提出的"人工智能基础问题"为例——

在点击"学术翻译"后,用户可以通过直接复制文本或上传中文文档,系统会将其翻译为英文,或者用户上传英文文档,系统可以将其翻译成中文(如图 6 所示)。学术翻译功

能支持 docx 格式文档的上传,输出结果同样会保存为 docx 文档供用户下载使用。

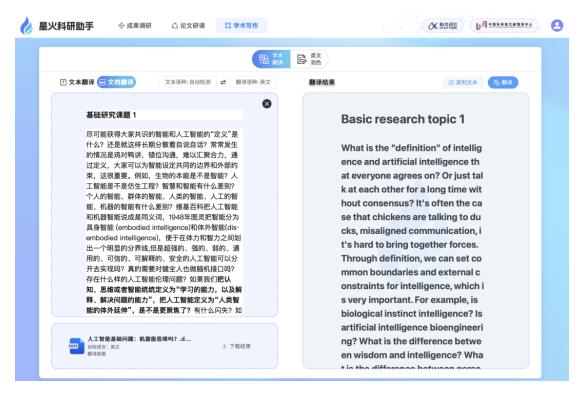


图 6 文档翻译功能

在英文润色功能,用户可以输入撰写好的英文论文,星火科研助手可以对输入进行润色修改,更加符合英文学术写作的表述。如图 7 所示,即一位研究生撰写的学术论文片段,其中存在多种表述问题。例如,第二句 "However, it depends on the amount of speech data because of the speaker-dependent model training method." 的 "it"表意不明存在歧义,在提到实验结果时反复使用 "can achieve"等问题。而在右侧的系统输出中,以上问题都得到了解决。

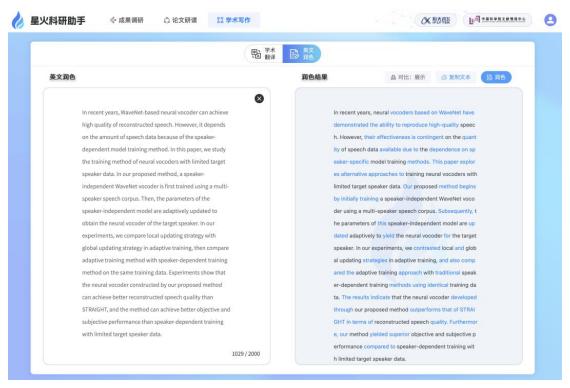


图 7 英文润色

四、使用咨询

如对星火科研助手有完善建议,请点击右上角头像-意见 反馈,将有工程师收集整理。试用过程中如有困难,也可积 极向图书馆反馈。

咨询电话: 66605327

电子邮箱: hanwm@bupt.edu.cn

联系人:韩老师

图书馆&信息化技术中心 2024年3月26日