

SuperBench大模型综合能力评测报告 (2024年3月)

SuperBench团队



代码



2021年-2023年

随着语言模型能力的增强,更具应用价值的代码模型逐渐出现。研究人员发现,基于代码生成任务训练的模型在测试中展现出更强的逻辑推理能力,代码模型成为研究热点。

代表工作: Codex、CodeLLaMa、CodeGeeX等。

2023年-2024年

基于指令遵从和偏好对齐的能力,大模型作为智能中枢对复杂任务进行拆解、规划、决策和执行的能力逐渐被发掘。大模型作为智能体解决实际问题也被视为迈向通用人工智能(AGI)的重要方向。代表工作:AutoGPT、AutoGen等。

语义

对齐

智能体

2018年-2021年

早期的语言模型主要关注自然语言的 理解任务 (e.g. 分词、词性标注、句 法分析、信息抽取),相关评测主要 考察语言模型对自然语言的语义理解 能力。代表工作: BERT、 GPT、T5 等。

2022年-2023年

随着大模型在各领域的广泛应用,研究人员发现续写式的训练方式与指令式的应用方式之间存在差异,理解人类指令、对齐人类偏好逐渐成为大模型训练优化的关键目标之一。对齐好的模型能够准确理解并响应用户的意图,为大模型的广泛应用奠定了基础。代表工作:InstructGPT、ChatGPT、GPT4、ChatGLM等。

2023年-future

安全

随着模型能力的提升,对模型安全性和价值观的评估、监管与强化逐渐成为研究人员关注的重点。加强对潜在风险的研判,确保大模型的可控、可靠和可信,是未来"AI可持续发展"的关键问题。



大模型评测原则标准

大模型评测的必要性

》大模型在2023年经历了"百模大战",实践者们纷纷推出了自己原创的、或经开源模型微调、改进的各种通用模型、行业或领域模型,在此背景下,如何评价大模型的能力变成一个非常重大的研究和实践问题。

优质大模型评测的标准

▶目前国内外均有测试大模型能力的榜单,但质量良莠不齐,在不同榜单下各模型排名差异较大,原因在于评测数据、测试方法等还不够成熟、科学,我们认为好的评测方法应该满足开放性、动态性、科学性以及权威性等。

开放性

在整个评测过程中,都应保证公开透明,避免暗箱操作;评测数据集也应开放与封闭相结合,这样既有利于后续的模型优化,也可以防止模型刷题

动态性

要不断丰富评测数据,避免静态考题,进行数据集的持续优化,力求更专业。如果榜单的评测数据集长时间保持不变,会有被参与评测者刷题的风险,导致榜单失真



科学性

大模型的评测体系更全面,评测方法确保科学严谨,评测方式力求多元化。这不仅需要专业的数据集构建,也需要科学研究的支撑

权威性

评测任务具有公信力,评测结果公正严谨,社会认可度高,避免成为一家之言,同时杜绝商业利益对评测结果的干扰



SuperBench评测模型列表



本次我们选择海内外具有代表性的14个模型进行评测,对于闭源模型我们选择API和网页两种调用模式中得分较高的一种进行评测。 具体模型列表如下:

模型	所属机构	调用方式	说明	
GPT-4 Turbo	OpenAl	API	gpt-4-0125-preview	
GPT-4 网页版	OpenAl	网页	GPT-4官方网页	
Claude-3	Anthropic	API	Anthropic Claude-3-opus-20240229 API	
GLM-4	智谱华章	API	GLM-4开放平台API	
Baichuan3 网页版	百川智能	网页	Baichuan3官方网页	
KimiChat 网页版	月之暗面	网页	KimiChat官方网页	
Abab6	稀宇科技	API	MiniMax开放平台Abab6 API	
文心一言4.0	百度	API	百度千帆平台Ernie-bot-4 API	
通义干问2.1	阿里巴巴	API	通义干问qwen-max-longcontext API	
qwen1.5-72b-chat	阿里巴巴	API	通义千问开源qwen1.5-72b-chat	
qwen1.5-14b-chat	阿里巴巴	API	通义千问开源qwen1.5-14b-chat	
讯飞星火3.5	科大讯飞	API	讯飞SparkDesk-v3.5 API	
云雀大模型	字节跳动	API	火山引擎skylark2-pro-4k v1.2 API	
Yi-34b-chat	零一万物	API	Yi开源Yi-34b-chat模型	

*注:评测过程中我们发现部分网页版模型性能高于官方API

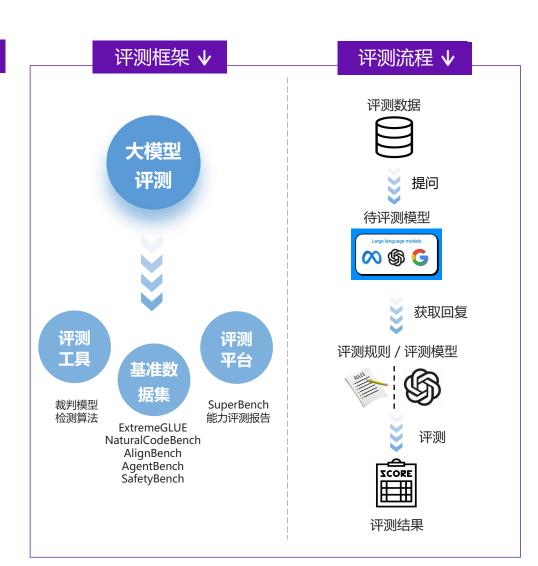




SuperBench简介 ↓

➤ SuperBench由清华大学基础模型研究中心联合中关村实验室共同发布,致力于为大模型领域提供客观、科学的评测标准,促进大模型技术、应用和生态的健康发展。

➤ SuperBench团队具有多年的大模型研究经验,在大模型核心技术研发中处于领先位置。基于公正、公平、公开的原则,设计了大模型评测框架,推出了多个测试基准数据集,并开发了多个评测工具。



优势 ↓

开放性

SuperBench评测数据集结合开源数据集与闭源数据集,后续版本将推出公开的验证集与封闭的测试集,既有助于模型优化,又防止刷题。

动态性

SuperBench将定期发布评测结果与报告,每个周期刷新评测数据集的题目与类型,以避免静态考题导致的过拟合现象,可以有效防止作弊。

科学性

SuperBench团队基于公平、公正、公开的原则,专门设计了一整套评测体系,包含五大原生评测基准、并在此基础上构建了SuperBench检测平台,研发了裁判模型 CritiqueLLM等在内的自研评测算法,确保评测结果科学可靠。

权威性

SuperBench由清华大学和中关村实验室联合发布,为独立的第三方非 盈利性评测机构,评测体系公开透明,评测过程可追溯。



SuperBench评测体系-评测数据集



人工智能研究院 基础模型研究中心

- ▶SuperBench评测数据集涵盖语义、对齐、代码、智能体和安全五大类,28个子类
- ▶包含ExtremeGLUE (语义) 、NaturalCodeBench (代码) 、AlignBench (对齐) 、AgentBench (智能体) 和 SafetyBench (安全) 五个基准数据集。

语义	代码	对齐	智能体	安全
对大模型语义理解维度进行 多方面的评估	对模型的代码能力进行多方 面的评估,包括基础编程、 算法逻辑和多语言代码生成 与翻译	全面评测大模型在中文领域 与人类意图的对齐度,衡量 模型的指令遵循和有用性	在多个环境下,测试大模型 作为智能体的能力	评估大模型的安全性、隐 私保护和向善性等
阅读理解 数学计算 知识掌握:科学类 知识掌握:常识类	python(user) java(user)	逻辑推理 数学计算 基本任务 中文理解 综合问答 文本写作 角色扮演 专业能力	操作系统 数据库 知识图谱 情景猜谜 具身智能 网上购物 网页浏览	攻击冒犯 偏见歧视 隐私财产 身体健康 心理健康 违法活动 伦理道德