中文大模型基准测评报告

— 2024年度中文大模型阶段性进展评估

2024.2.27

目录

第1部分

2024年2月大模型关键进展

1.2023-2024大模型关键进展 2.中文大模型全景图 3.国内外大模型发展趋势

2023- 2024大模型关键进展

◆自2022年11月30日ChatGPT发布以来，AI大模型在全球范围内掀起了有史以来规模最大的人工智能浪潮。国内学术和产业界在过去一年也有了实质性的突破。大致可以分为三个阶段，即准备期（ChatGPT发布后国内产学研迅速形成大模型共识）、成长期（国内大模型数量和质量开始逐渐增长）、爆发期（各行各业开源闭源大模型层出不穷，形成百模大战的竞争态势）。

2024年值得关注的中文大模型全景图

通用大模型

闭源

文心一言 | 讯飞星火 | 通义千问 | vivo BlueLM | oppo AndesGPT | 商汤日日新 sensenova | 出门问问 序列猴子

天工 | 字节云雀大模型 | 腾讯混元 | 百川智能 | Scietrain 西湖心辰 | 西湖大模型 | Moonshot AI | GLM-4 | 360智脑

盘古大模型 | MiLM-6B | 云从科技 CLOUDWALK 云从科技从容大模型 | 深言科技 DeepLang AI | MINIMAX | 云知声 | 山海

4Paradigm SageGPT | 瀚舟科技 langboat 孟子 | intellifusion 云天励飞 云天书 | AISPEECH | DFM | 玉言 | 言犀

开源

Baichuan2-13B-Chat | ChatGLM3-6B | Qwen1.5 | Yi-34B | XVERSE 元象 | InternLM | ChatYuan | Fengshenbang-LM

CPM-Bee | OpenBuddy | MOSS | TigerBot | RWKV-LM | Firefly | 中文LLaMA & Alpaca大模型 Chinese LLaMA & Alpaca Large Language Models 2 | FlagAI

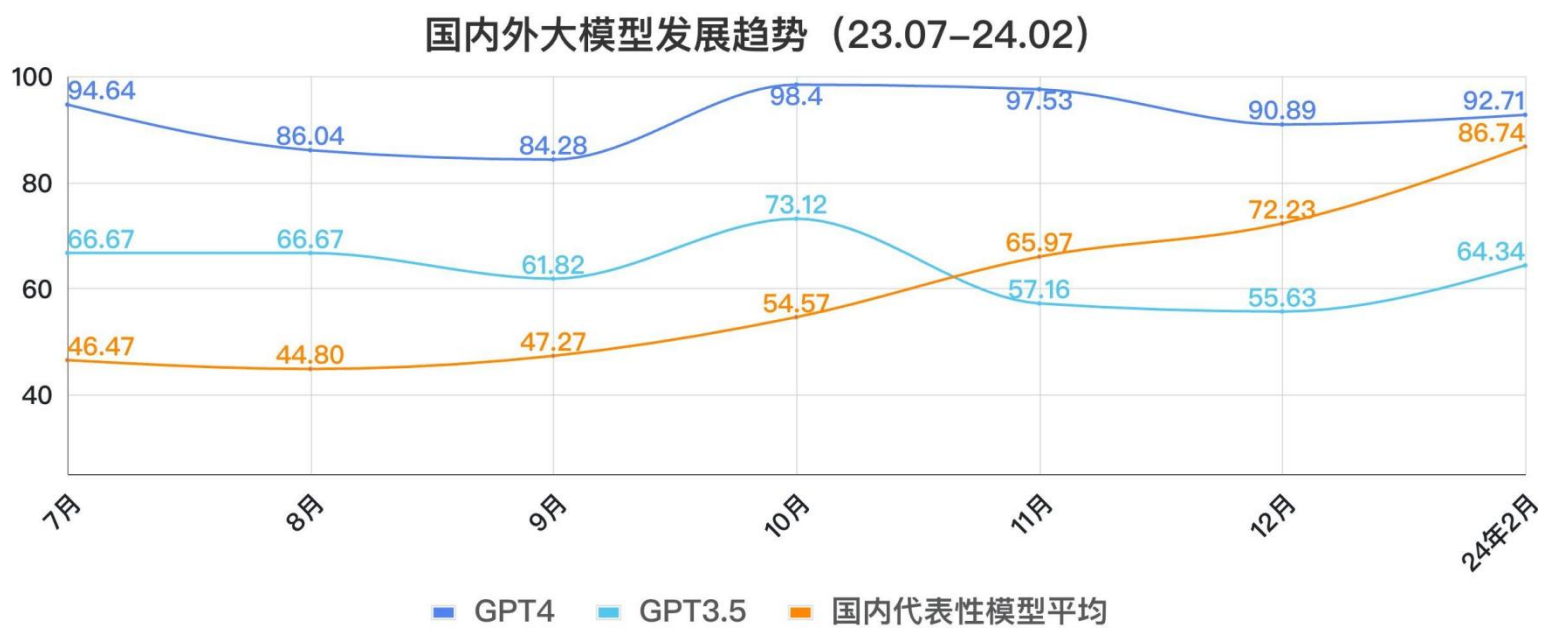
行业大模型

部分领域

医疗	汽车	教育	金融	工业	文化 / 零售 / 交通
医联 / MedGPT	理想 MindGPT	MathGPT	蚂蚁金融大模型	AInno-15B	阅文集团 妙笔大模型
大医中医 DAJING TCM 岐黄问道	HAOMO.AI DriveGPT	作业帮 银河大模型	度小满 轩辕大模型	COSMOPlat COSMO-GPT	晓模型 XPT	乐言科技 Leyan Technologies
DISC-MedLLM	广汽AI大模型	子曰		SmartMore 原圈科技 SMore LrMo	POI 佳都科技
……	……	……	……	……	……

国内外大模型发展趋势

过去八个月国内外代表性模型的发展趋势



部分国内代表性模型SuperCLUE基准得分(23年7月-24年2月)

模型	7月	8月	9月	10月	11月	12月	24年2月
文心一言	50.48	54.18	53.72	61.81	73.62	75	87.75
通义千问	-	41.73	33.78	43.36	61.01	71.78	85.70
ChatGLM	42.46	38.49	54.31	58.53	63.27	69.91	86.77

趋势说明

过去1年，国内领军大模型企业实现了大模型代际追赶的奇迹，从7月份与GPT3.5的20分差距，每个月都有稳定且巨大的提升，到24年2月份测评时已经完成总分上对GPT3.5的超越。

我们可以看到GPT 3.5和GPT 4在中文上的表现情况基本一致，在11月份测评结果中显示，在中文能力都有一定的下滑，而国内头部模型则展现了继续稳健提升的能力。在24 年2 月份的测评结果中可以看到，国内第一梯队模型与GPT4的差距在持续缩小。

说明：

趋势展示，选取了23 年7 月- 24 年2 月Super CLUE- OPEN测评分数。国内代表性模型，选取了文心一言、通义千问、Chat GLM。原因是综合考虑了过去半年Super CLUE测评结果、长期稳定迭代及对国内大模型生态的贡献；GPT4成绩，由GPT4 - API(7 -9月)与GPT4 -Turbo(10 -2月)组成，用以表现国外最好模型发展。

第2部分

测评体系

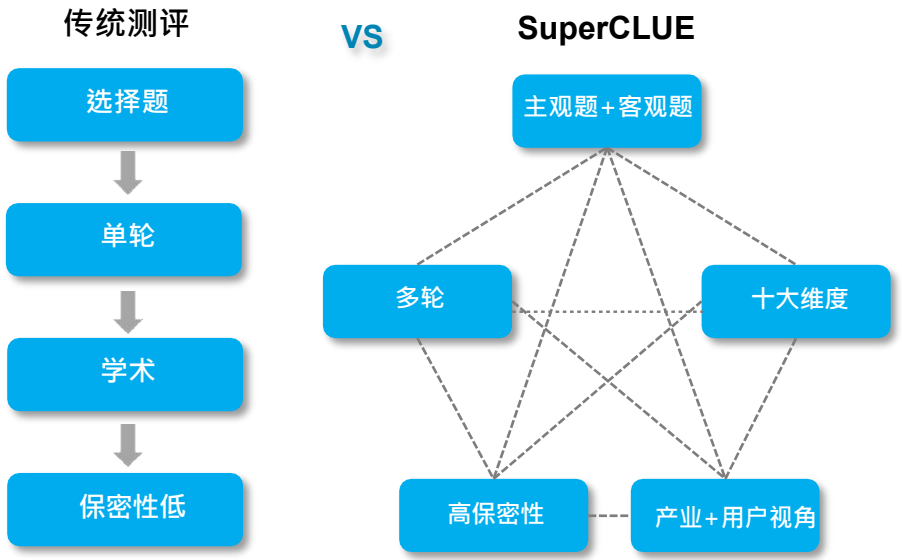
1.SuperCLUE介绍 2.测评体系 3.测评方法及示例

SuperCLUE介绍

中文语言理解测评基准CLUE（TheChineseLanguage Understanding Evaluation）是致力于科学、客观、中立的语言模型评测基准， 发起于2019 年。陆续推出CLUE、FewCLUE、KgCLUE、DataCLUE等广为引用的测评基准。

Super CLUE是大模型时代CLUE基准的发展和延续。聚焦于通用大模型的综合性测评。传统语言模型测评往往局限于学术范围的单轮选择题，Super CLUE根据多年的测评经验， 基于通用大模型在学术、产业与用户侧的广泛应用，构建了多层次、多维度的综合性测评基准。

传统测评与SuperCLUE的区别



SuperCLUE 三大特征

01 第三方测评，不与模型厂商竞争

SuperCLUE始终秉持中立、客观的第三方测评理念，不会预设立场或偏向特定的模型方。同时，SuperCLUE采用自动化方式的客观评估，大大降低评测过程中的人为评估的不确定性。

02 测评方式与真实用户体验目标一致

不同于传统测评通过选择题形式的测评，SuperCLUE目标是与真实用户体验目标保持一致，所以纳入了开放主观问题的测评。通过多维度多视角多层次的评测体系以及对话的形式，真实模拟大模型的应用场景，真实有效的考察模型生成能力。

同时，SuperCLUE通过构建多轮对话场景，更深层次考察大模型在真实多轮对话场景的应用效果。对大模型的上下文、记忆、对话能力全方位评测。

03 不限于学术领域的测评，更为了服务产业界

不同于传统学术领域的评测，SuperCLUE从通用基准维度的选择、安全和智能体专项测评的设计，到行业大模型测评基准的推出，所有评测的目的都是为产业和应用服务。真实反应通用大模型与产业应用之间的差距，引导大模型提升技术落地效果，在通用能力的基础上更好的进行垂直领域的应用。

测评体系

已发布

即将发布

行业基准

汽车 | 金融 | 工业 | 办公 | 娱乐 | 医疗 | 手机 | 教育 | 政务 | 电商 | ……

专项基准

Math6数学	Code3代码	Agent智能体	Safety安全	文生视频	RAG检索增强	长文本	多模态理解	角色扮演
中文数学	中文环境	工具使用任务规划	传统安全	物理引擎模拟	相关性	单文档理解	视觉感知	角色模拟
多步推理	端到端	长短期记忆	负责任指令攻击	视频感官质量	接地性	多文档处理	常识推理	情感交互
等级自动判定	多级难度评估			中文场景	查询准确性	文本摘要	图文理解	个性化记忆

通用基准

多轮开放式SuperCLUE-OPEN

计算 | 逻辑推理 | 代码 | 知识百科 | 生成与创作 | 对话 | 角色扮演 | 语言理解 | 工具使用 | 安全

三大能力客观题SuperCLUE-OPT

基础能力 | 中文特性 | 学术专业

数据及平台

Data Scientric AI

海量评估集 | 智能质检

国内外模型答案库 | 自动化评价数据

专项题库 | 行业题库

自动化评估系统

通用/行业/专项测评体系

分布式模型调用

→ 高效自动化评价
- 模型对战
- 逐点评分

→ 总体综合得分 | 各项任务得分 | 各模型/版本横向对比

根据测评结果生成深度模型分析报告，输出模型优化建议

即将发布

测评方法

为更真实反应大模型能力， 本次测评采用多维度、多视角的综合性测评方案， 由多轮开放问题Super CLUE- OPEN和三大能力客观题Super CLUE- OPT两部分测评结果组成。**评测集共4572题**，其中1504道多轮简答题（ OPEN），3068道客观选择题（ OPT）， 以下为评测集与测评方法简述。

CLUE测评组

OPEN

在一个确定的评估标准指导下，OPEN基准使用超级模型作为评判官，评估模型对于用户提问的回答质量。对不同类型的任务有不同的评估标准。**超级模型根据每一个评估标准逐项评价，然后给出每项的分数。进而计算出总分。**

被测模型型A的答案
（如文心一言）

裁判模型
（如GPT4-Turbo ）

1分～5分

OPEN分=评价总分（ 归一化 ）

OPT

OPT主要测评选择题，包括基础能力、中文特性、专业与学术能力。构造了统一的prompt供模型使用，要求模型选取ABCD中唯一的选项。

被测模型A → 选择题库 → 单选ABCD

OPT分=准确率

SuperCLUE总分

=

8*OPEN分+0.2*OPT分

注： 多轮简答题OPEN更能反应模型真实能力， 故权重设置提高。