

人工智能大模型在工业领域知识问答： **稳定性评测**

中国工业互联网研究院
香港科技大学

2023年9月

前言

人工智能大模型作为实现通用人工智能的重要途径，正在对制造业的产业发展和应用模式产生变革性影响。近期，中国工业互联网研究院联合香港科技大学、新华指数，共同研究AI大模型在工业领域的应用性能、技术架构与标准体系，并发布评测报告。

工业领域对人工智能大模型的应用稳定性要求极高。人工智能大模型在工业领域应用的稳定性研究对于应用安全性保障、稳健性提升、可信赖性增强、决策优化以及合规性等方面都具有重要的指导意义。本报告聚焦评测大模型在工业领域的**知识问答稳定性**，通过选取工业领域典型的八大行业，构建基础知识测试集与变体测试集，对国内外具有代表性的AI大模型在文法、数据、精简表达、扩充表达四大类八个子维度展开深入评估。选取性能前五名的模型进行公布，并发布前十名的数据，供业界进行参考。

本报告评测结果虽经专家组认真论证，但因大模型迭代速度快，评测结果仅适用于测试期间。报告难免存在分析结论片面与不足，欢迎大家批评指正。

01

第一部分 评测规则

02

第二部分 评测结果

03

第三部分 总结

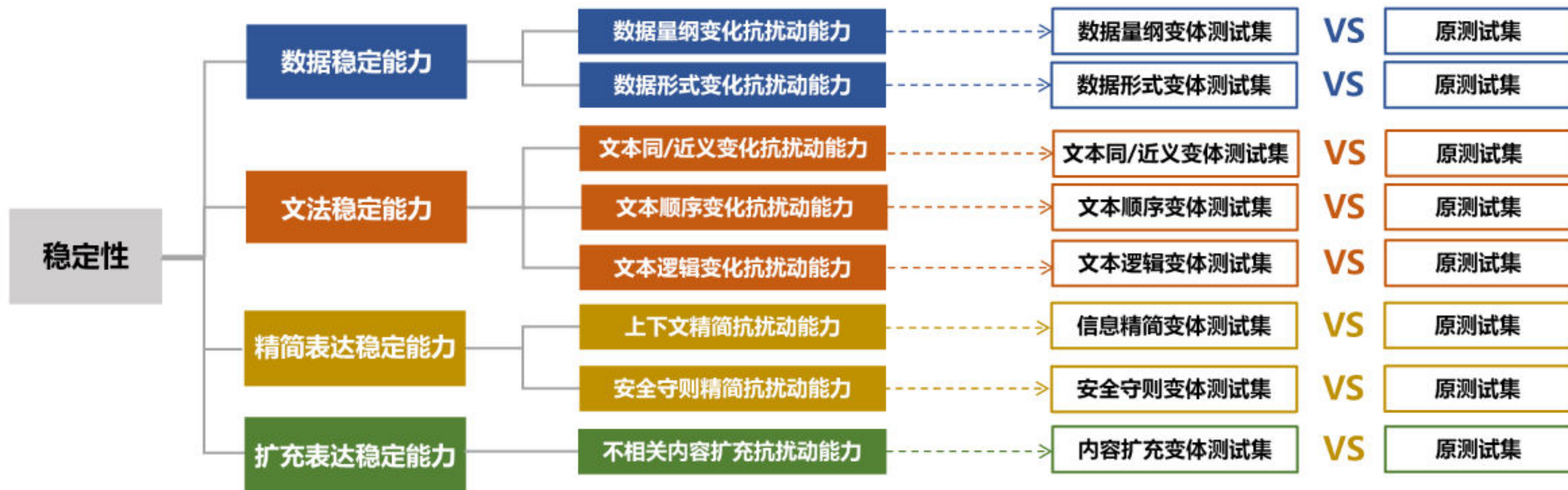
工业领域知识问答稳定性评测方法

稳定性定义：

- 指人工智能大模型在工业知识问答领域，具备较高回答准确性能的前提下，针对一定参数扰动，维持某些性能特性的能力，即在非理想的工作环境或应用场景下也能够良好运行的能力。

评测体系：

- 稳定性评测围绕AI大模型在数据、文法、精简表达、扩充表达四大类八个子维度展开深入评估。



注：经调研及评测专家组认定，该评测体系符合大模型工业日常应用稳定性基本逻辑。

工业领域知识问答稳定性评测方法

评分标准:

- 原问题与对应变体问题进行对比测试, 题型包含单选题和判断题;
- 所有进行变体对比的测试题应为模型测试回答准确的测试题;
- 若变体与原问题回答**一致**时, 得分**(1分)**; 若变体与原问题回答**不一致**时, 不得分**(0分)**。

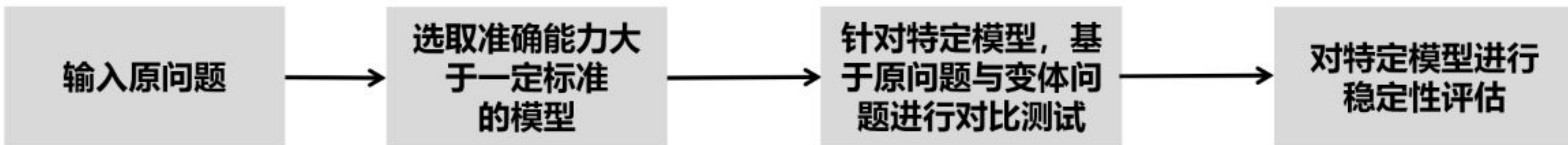
模型稳定性能力指数定义:

稳定性能力指数 = 抗扰动能力指数算术平均值

$$\text{各测试维度抗扰动能力指数} = \frac{\text{各维度模型实际得分}}{\text{各维度模型理论最高得分}} \times 100$$

$$\text{各行业稳定性综合能力指数} = \frac{\text{各行业模型实际得分}}{\text{各行业模型理论最高得分}} \times 100$$

稳定性测试流程:



工业领域知识问答稳定性评测数据集

测试集

电子设备制造业



电子器件、电子元件及电子专用材料、其他电子元件制造等。

采矿行业



煤炭、石油、天然气、黑色金属、有色金属等。

电力行业



电力生产、电力供应、热力生产和供应等。

纺织行业



棉纺织、毛纺织、麻纺织、丝绢、化纤等纺织及印染精加工。

装备制造行业



金属制品、通用设备制造、专用设备制造、汽车制造等。



钢铁行业

炼铁、炼钢、钢压延加工、铁合金冶炼等。



石化化工行业

石油开采及加工、化学原料及化学制品制造、塑料制品、橡胶制品等。



建材行业

建筑材料及制品、非金属矿及制品、无机非金属新材料等。

根据OT经验，结合外部数据源，
构建8大行业
测试数据集。

专家组成员



唐立新，中国工程院院士，IEEE Fellow，东北大学副校长，第十四届全国人大代表，工业智能与系统优化国家级前沿科学中心主任和首席科学家。



刘云浩，ACM Fellow，IEEE Fellow，长江学者，ACM主席奖、CCF王选奖获得者，清华大学教授，创新学院院长。



张涛，IET Fellow，清华大学教授，信息科学技术学院副院长，自动化系系主任。



刘劼，IEEE Fellow，ACM杰出科学家，哈尔滨工业大学（深圳）讲席教授，国际人工智能研究院院长。



张大庆，欧洲科学院院士，IEEE Fellow，北京大学讲席教授，CCF普适计算专委会主任。



刘江川，加拿大工程院院士，IEEE Fellow，加拿大西蒙菲莎大学计算机学院正教授和大学杰出教授。

专家组成员



张燕咏, IEEE Fellow, 中国科学技术大学教授, 计算机科学与技术学院副院长, ACM中国副主席。科技部下一代人工智能重大项目负责人。



聂再清, 清华大学国强教授, AIR首席研究员, 微软学术搜索发起人, 曾任微软自然语言理解平台LUIS技术负责人。



杨铮, IEEE Fellow, 清华大学副教授、博士生导师, 国家万人计划。



黄河燕, 北京理工大学计算机学院教授, 国家科学技术进步奖一等奖获得者。



刘云新, 清华大学国强教授, 智能产业研究院首席研究员, 前微软亚洲研究院主任研究员。国家重点研发计划项目负责人。



贾佳亚, IEEE Fellow, 香港中文大学终身教授, 全球计算机视觉、人工智能领域权威专家, 国家科技部“新一代人工智能”重大项目核心专家。



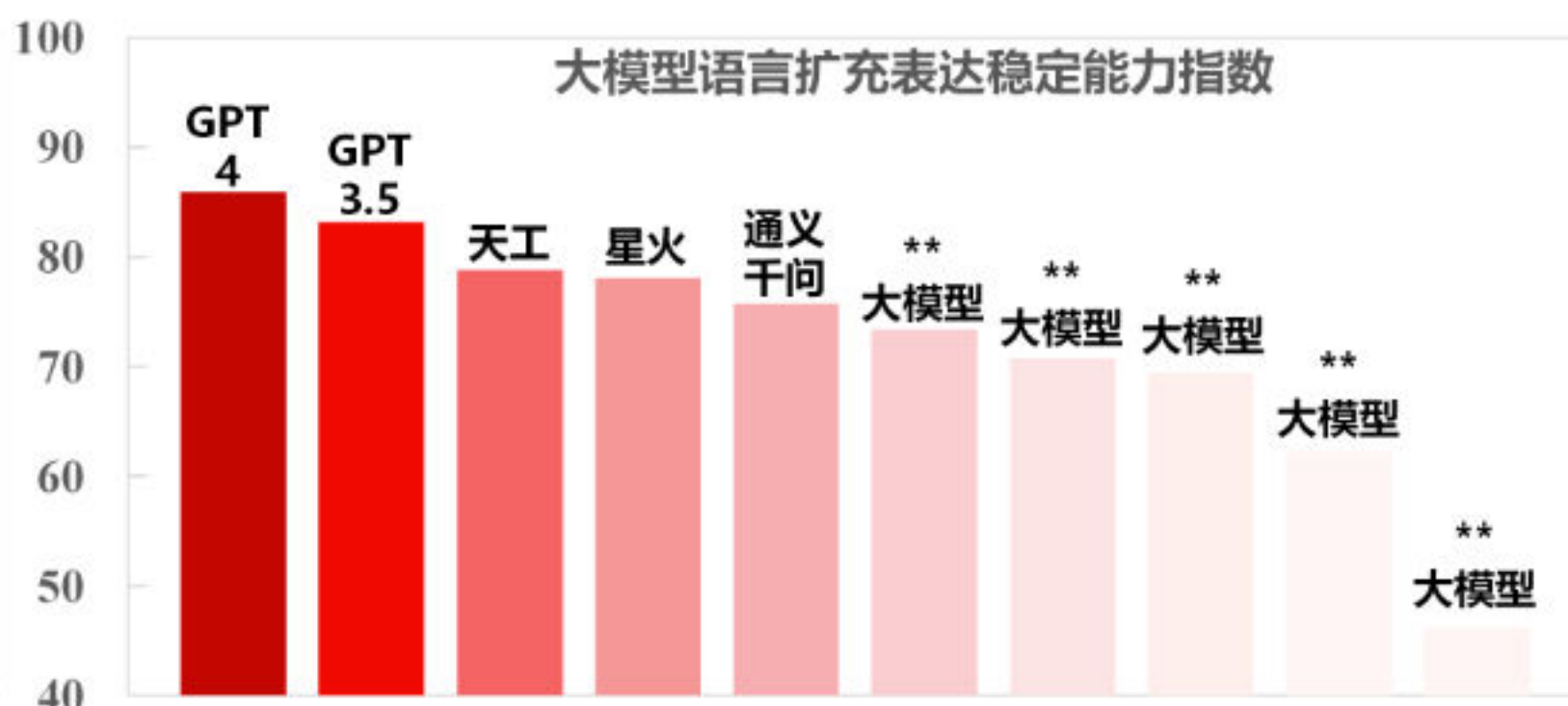
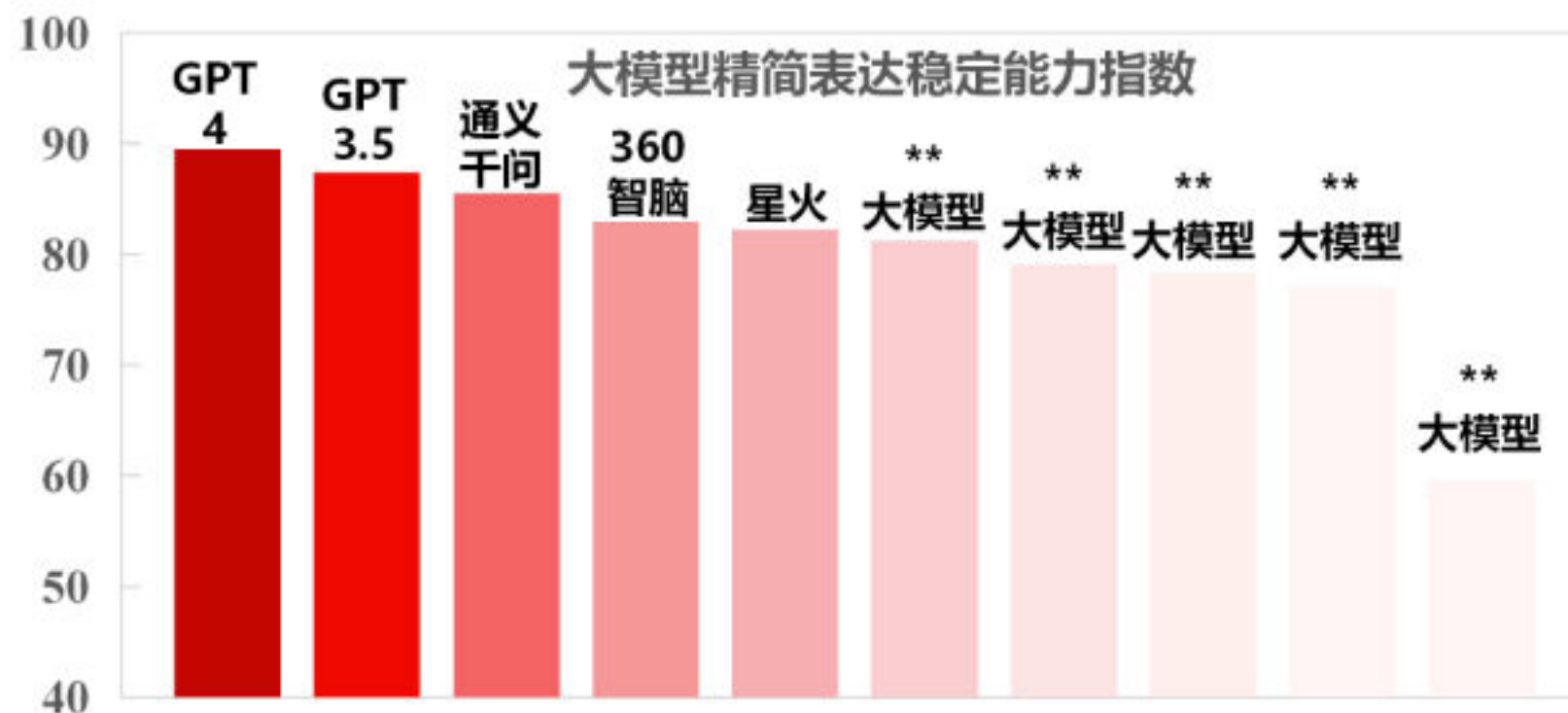
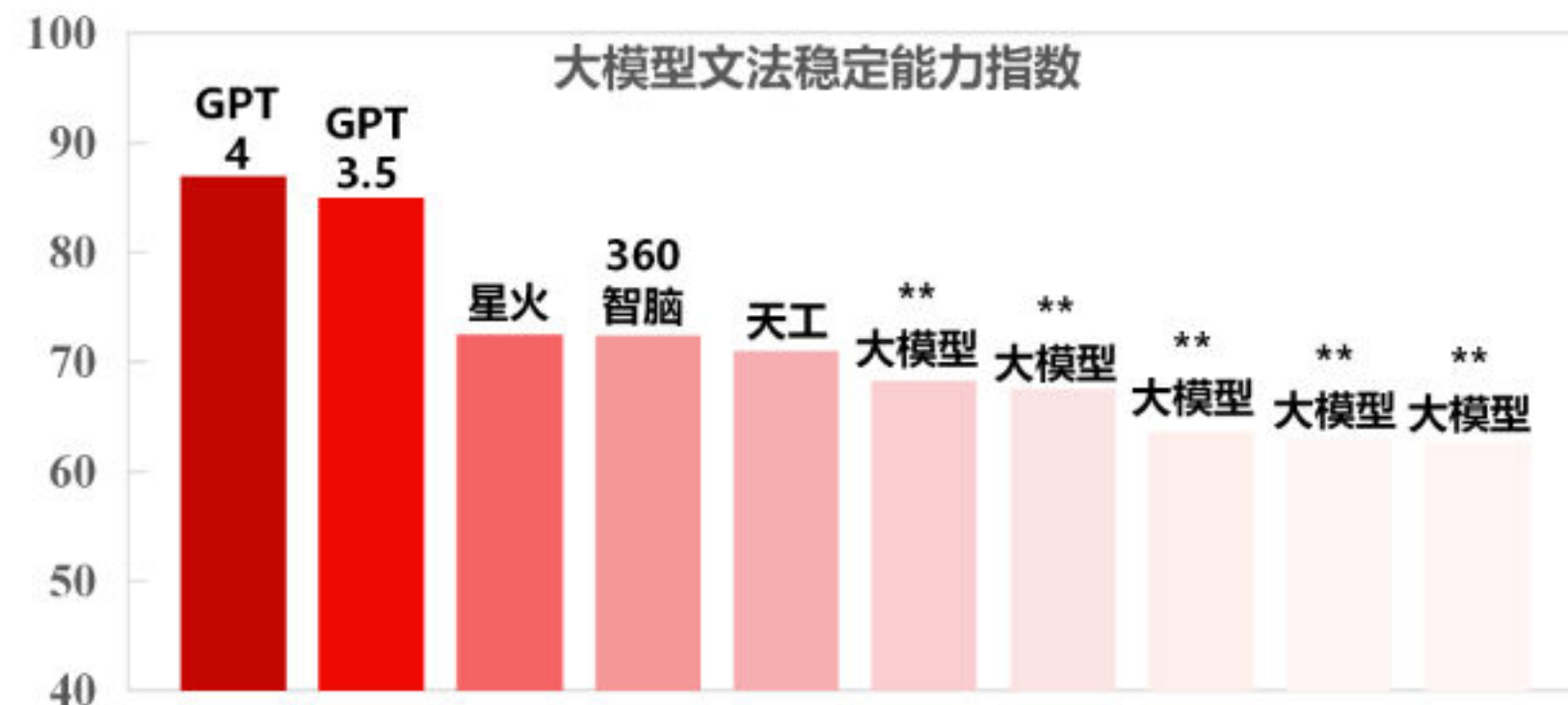
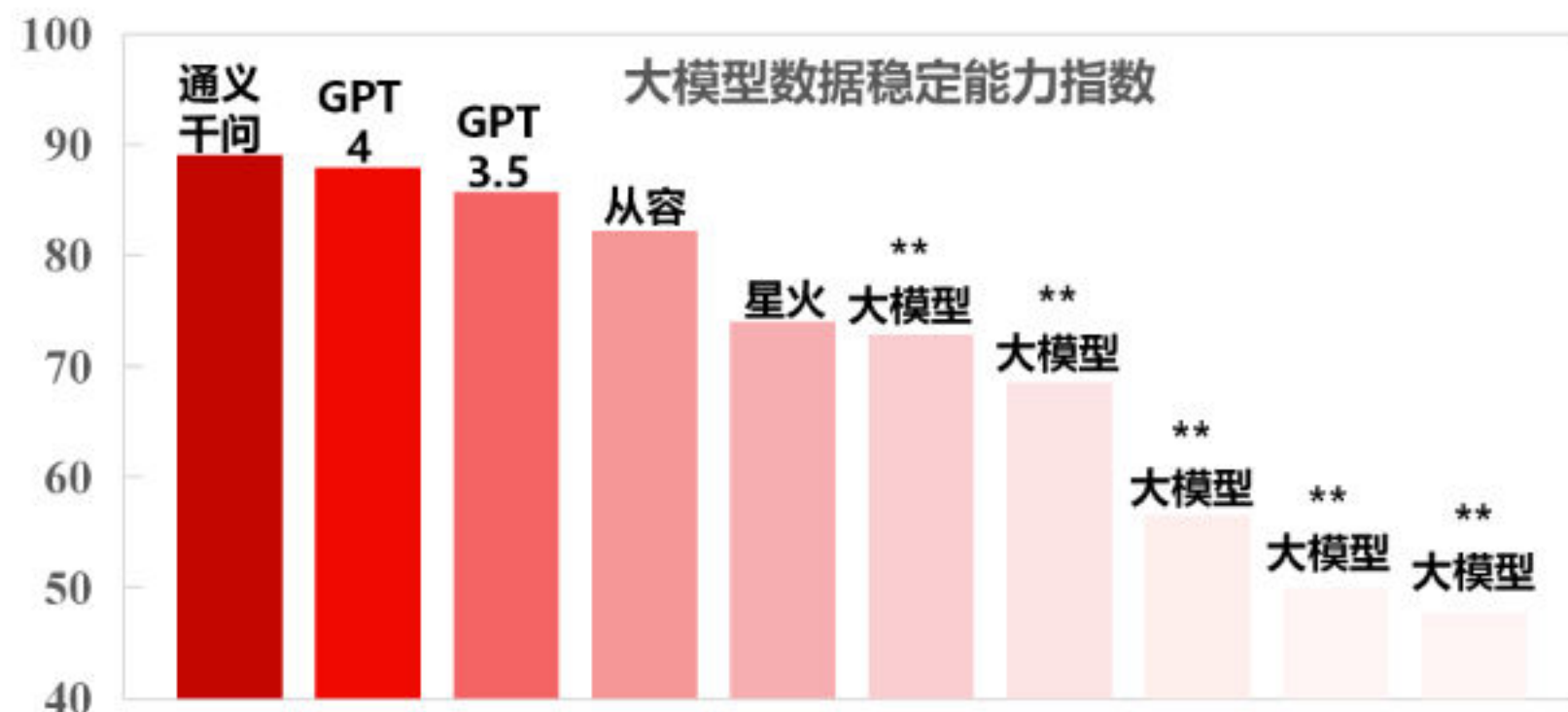
张晓明, 北京雁栖湖应用数学研究院工业与应用首席研究员, 美国麻省理工学院博士。

01 第一部分 评测规则

02 第二部分 评测结果

03 第三部分 总结

工业领域知识问答稳定性评测结果



工业领域知识问答稳定性评测结果

四大能力维度稳定性



八大细分能力维度稳定性



行业维度稳定性



● 国内大模型能力区间 ● GPT4 ● GPT3.5

总体来看:

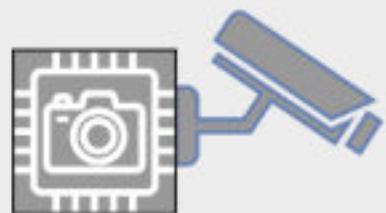
- 从能力维度, 国内大模型与国外头部大模型在数据稳定、精简表达、扩充表达能力差距较小, 在文法稳定方面存在一定差距;
- 从行业应用, 国内大模型与国外在钢铁、采矿、装备制造等方面应用差距较小, 其他行业具有较大提升空间。

一、数据稳定能力

- 数据稳定能力指大模型在工业应用中涉及数据形式或数据量纲变化的输出稳定性。

工业场景的潜在应用

1. 数据形式扰动

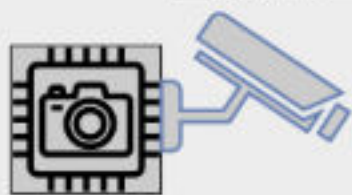


产品1生产日期
检查：正确

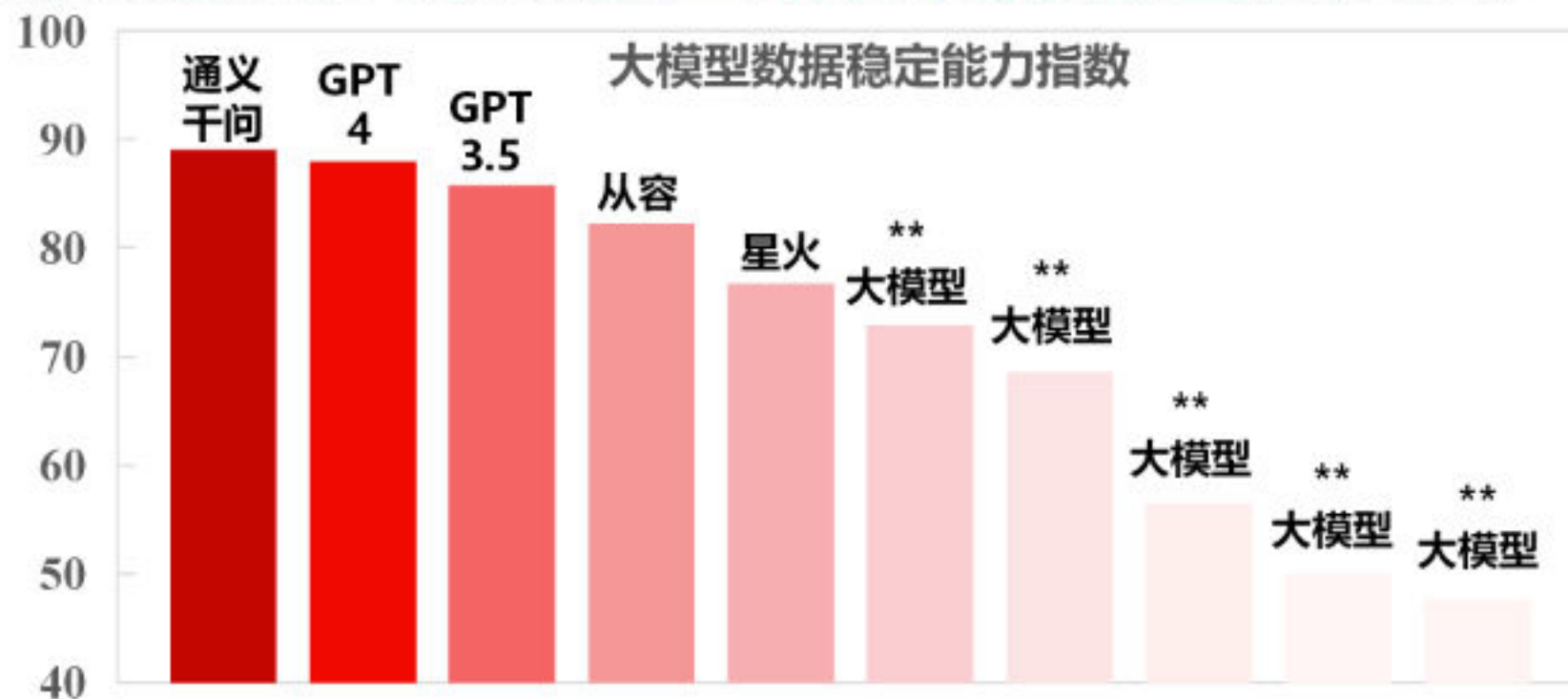
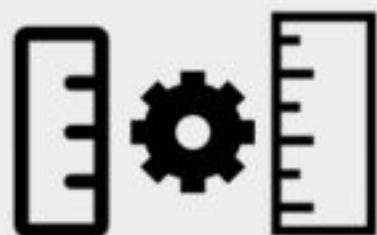
产品2生产日期
检查：正确



2. 数据量纲扰动



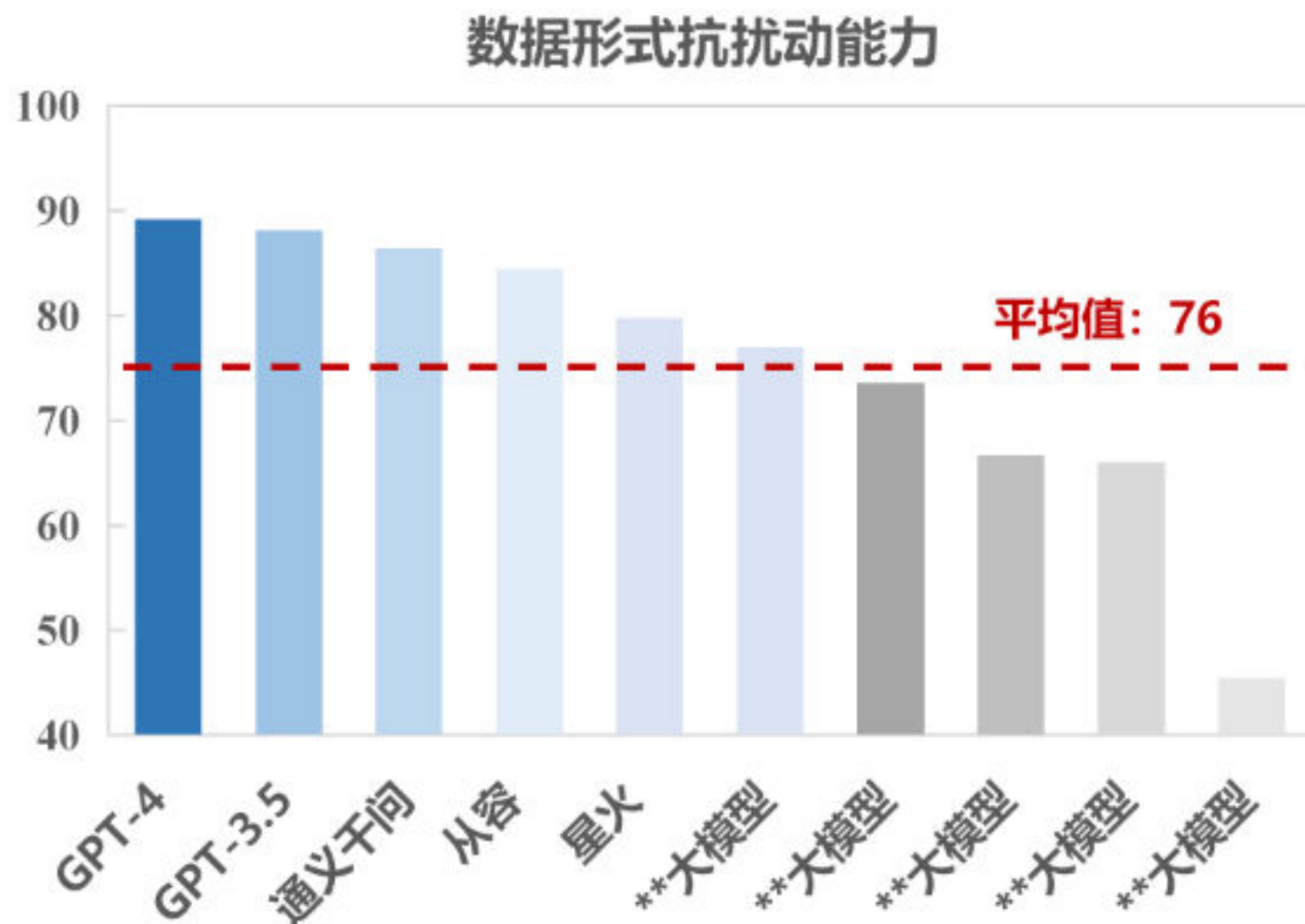
测量结果1为2cm,
测量结果2为20mm,
两次测量结果一致



- 整体看，国内外头部大模型能力接近90分，说明大模型在数据上稳定性较高；
- 对比看，国内头部大模型基于中文语境优势，数据能力已实现超越；国内尾部大模型能力低于50分，表现具有较大提升空间；
- 供给角度看，国内尾部大模型需要引入训练语料、知识库、工具函数集等，对数据进行专门强化；
- 应用角度看，鼓励工业企业将大模型应用于数据处理，但需要进行结果校验，尤其是对数据单位要进行多重描述，防止理解偏差。

数据稳定能力—数据形式抗扰动

- 大模型对工业领域不同数值的数据形式的回答稳定性。



测试样例:

原提问: 分度头的主轴轴心线能相对于工作台平面向上 ()和向下10°。

A、10° B、45° C、90° D、120°

扰动提问: 分度头的主轴轴心线能相对于工作台平面向上 ()和向下10°。

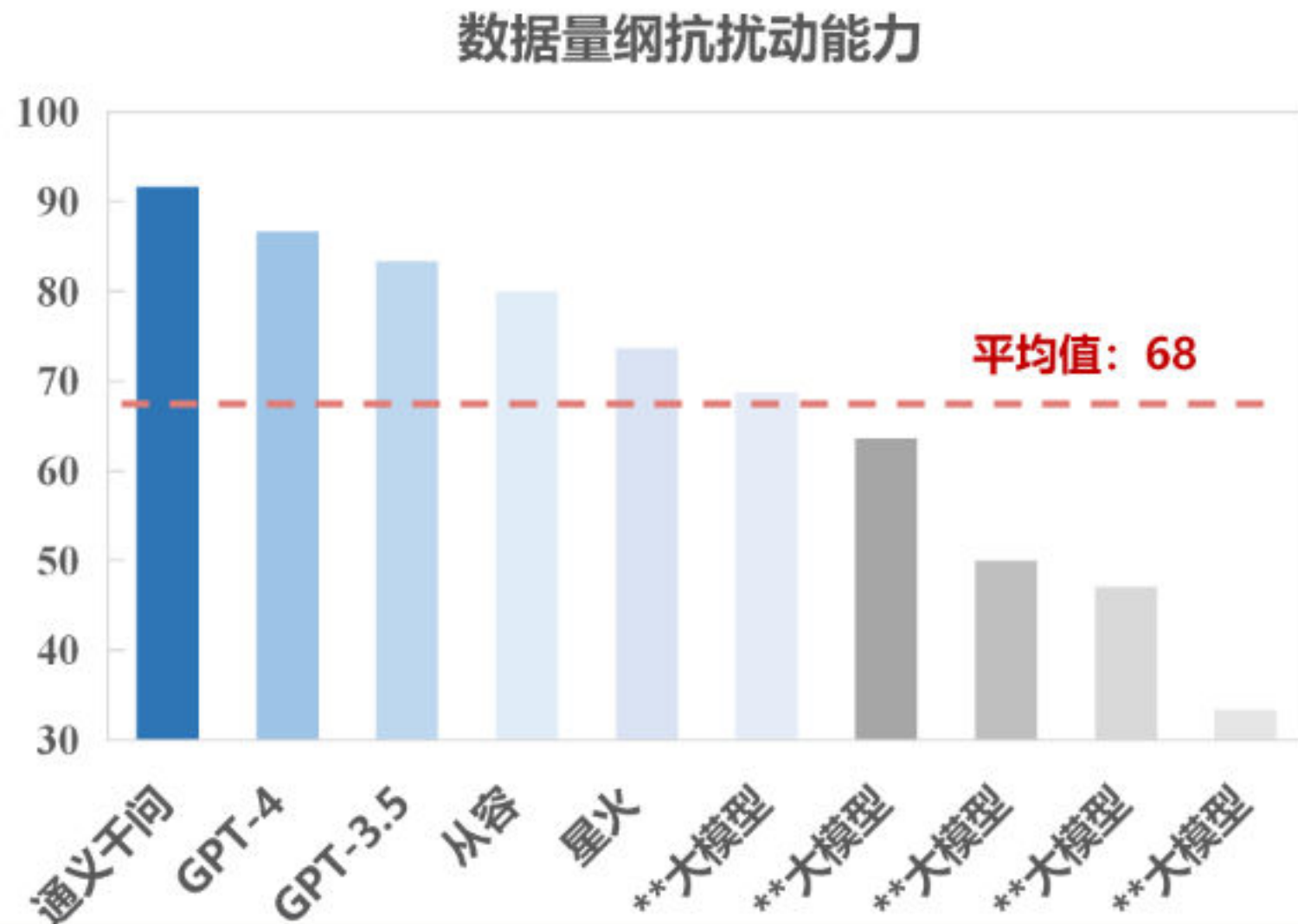
A、10.0° B、45.0° C、90° D、120°

注: 数据形式变化测试变体涉及“有效数字”、“百分比”和“科学记数法”等形式变化。

- 对于数据形式抗扰动能力, GPT-4表现最佳, 当前国内外头部梯队大模型差距较小, 国内大模型间存在一定差距;
- 从模型供应商角度看, 尾部模型需对数据形式稳定性进行专门优化, 可通过数据预处理或引入工具函数集等方式强化数据稳定能力;
- 从工业企业角度看, 在应用大模型时, 对于数据密集型任务, 需使用数据形式稳定性较高的大模型。

数据稳定能力—数据量纲抗扰动

- 大模型对工业领域不同数值数据量纲的回答稳定性。



测试样例:

原提问: 在220kV带电区域中的非带电设备上检修时, 工作人员正常活动范围与带电设备的安全距离应大于() 。
A、 0.35m B、 0.6m C、 1.5m **D、 3.0m**

扰动提问: 在220kV带电区域中的非带电设备上检修时, 工作人员正常活动范围与带电设备的安全距离应大于() 。
A、 0.35m B、 0.6m C、 1.5m **D、 300cm**

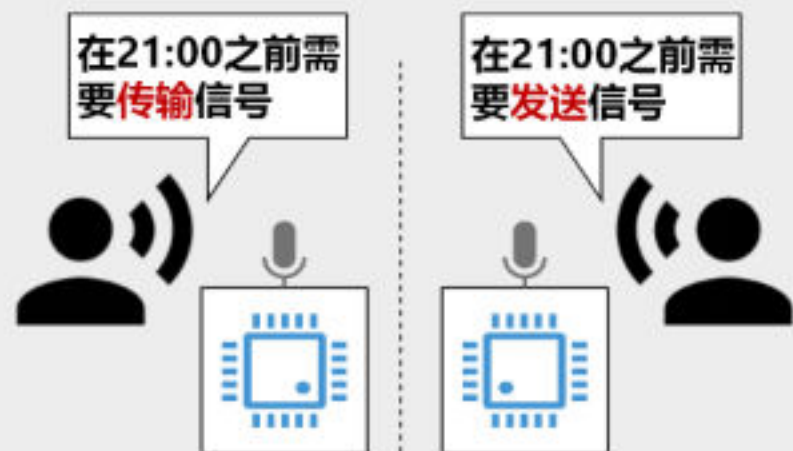
- 对于数据量纲的变化, 国内头部大模型的稳定性已实现赶超, 大模型间能力表现差距明显;
- 从供应商角度看, 尾部模型供应商需针对量纲进行专项优化, 对问题中数据进行预处理;
- 对于工业企业, 使用大模型时应关注数据量纲抗扰动能力, 尽量保持量纲的一致性, 或同步文字或字母描述。

二、文法稳定能力

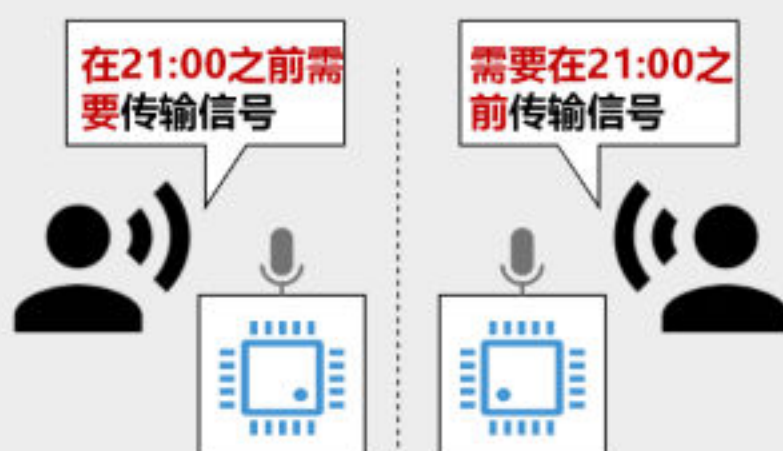
- 指大模型在工业应用中涉及文本内容或逻辑变化的输出稳定性。

工业场景的潜在应用

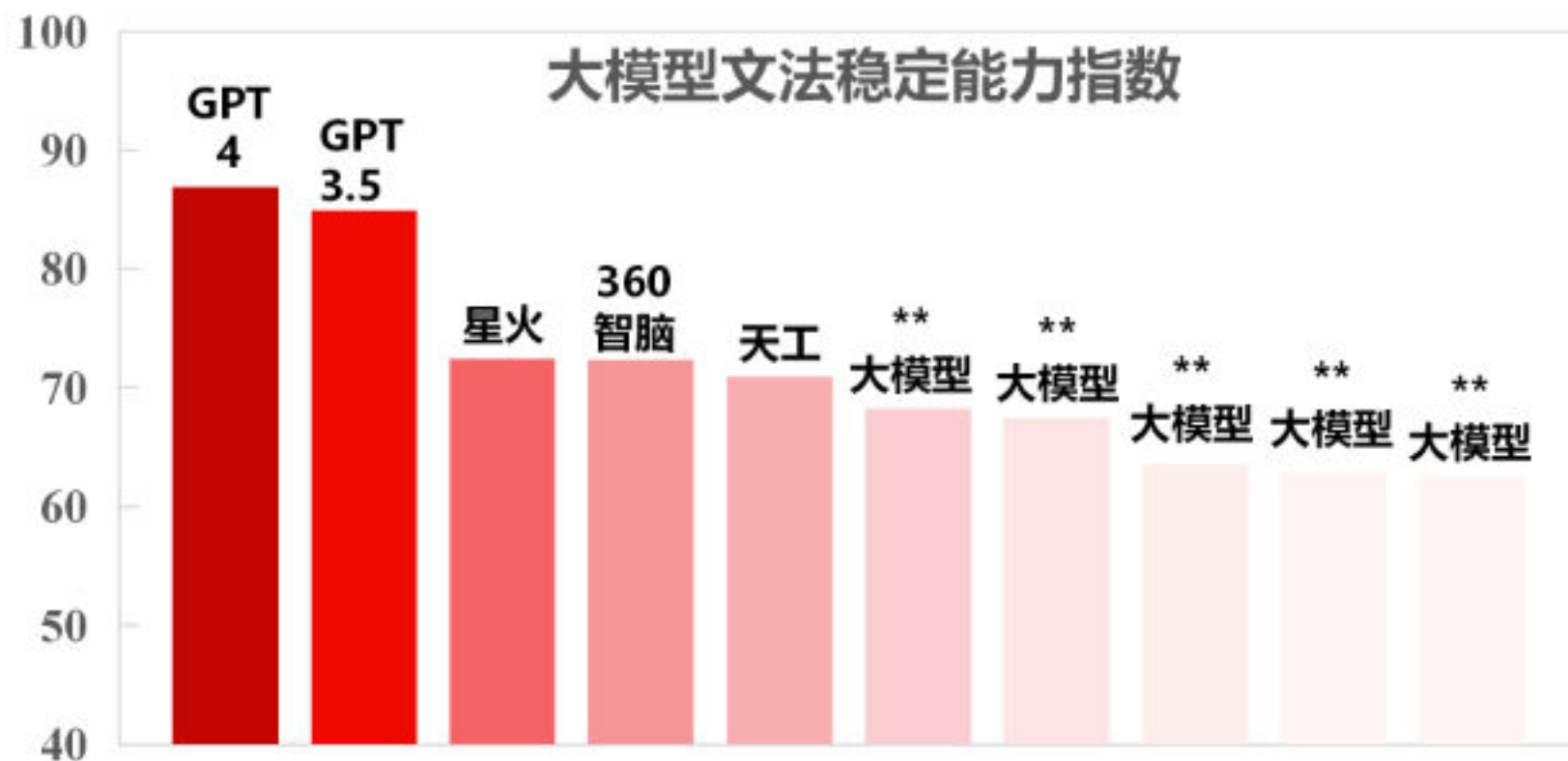
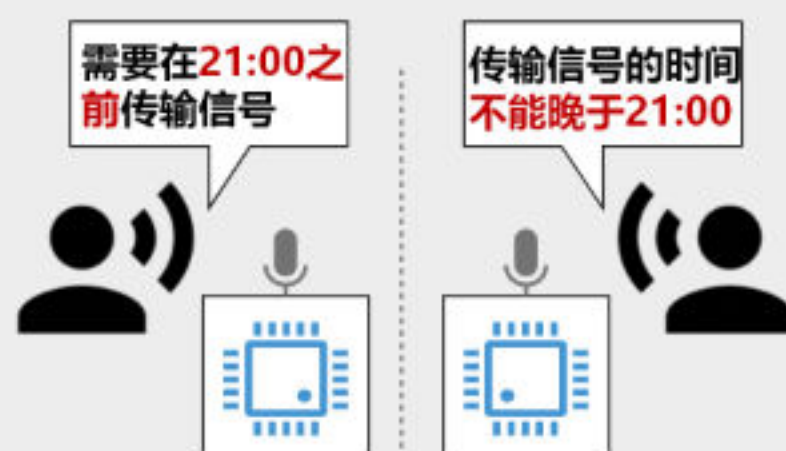
1. 表达同/近义变化



2. 表达顺序变化



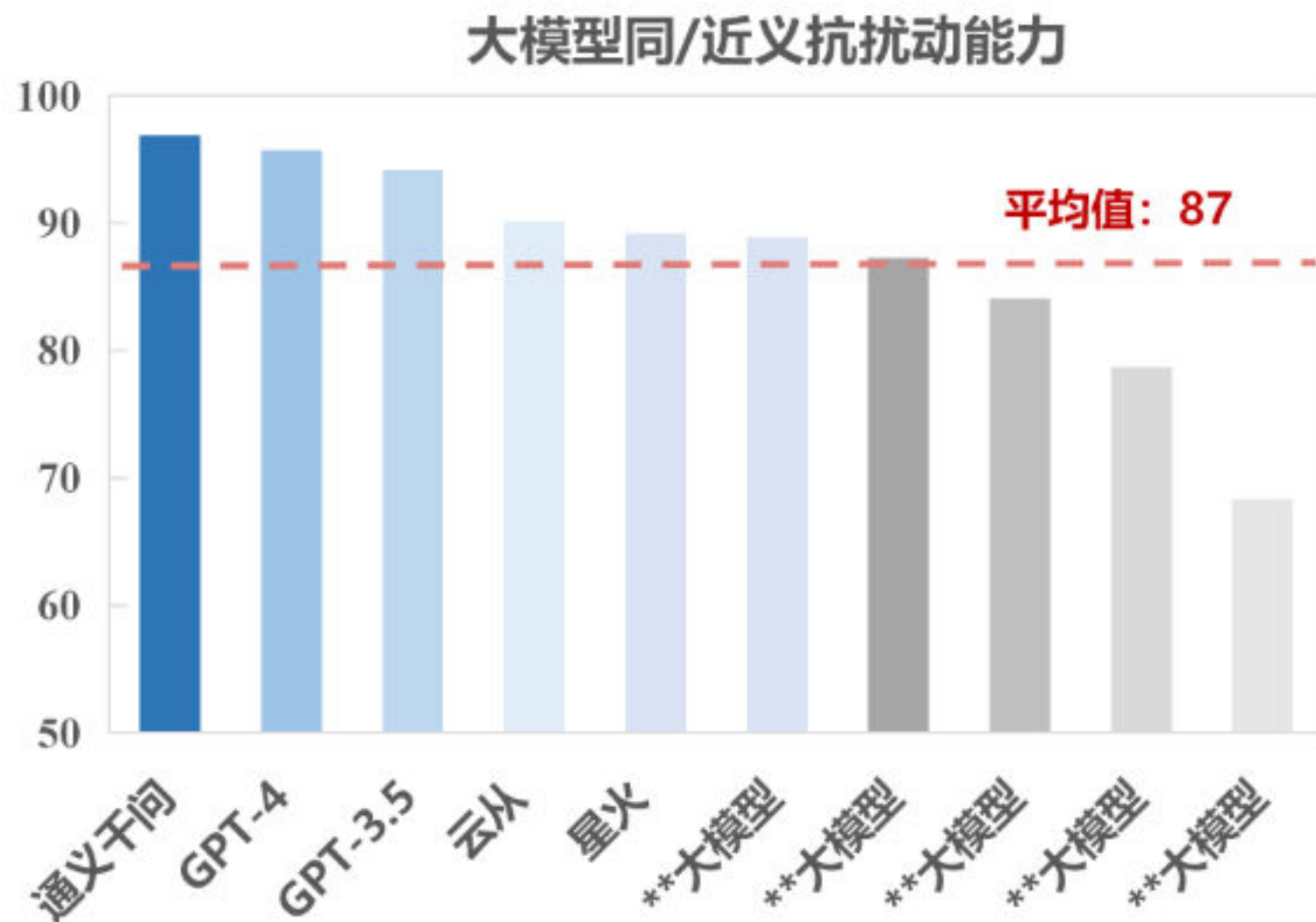
3. 表达逻辑变化



- 整体看, GPT-3.5/4 在85分以上, 在文法上稳定性较高;
- 对比看, 国内外大模型的文法稳定能力差距较大, GPT3.5/4 优势明显, 国内大模型具有较大追赶空间;
- 供给角度看, 国内大模型在同/近义词的稳定性较强, 对逻辑变化和对顺序变化的稳定性需要进一步提升, 说明国内大模型的对问题的理解能力还有待加强, 需针对性进行优化;
- 应用角度看, 工业企业对于文字简单处理的应用建议尝试使用大模型, 如文献整理和收集, 但对逻辑密集型的应用需谨慎使用。

文法稳定能力—同/近义抗扰动

- 指大模型对工业领域相关词汇表达同义或近义的回答稳定性。



测试样例:

原提问: 在计算机通信中, **传输**的是信号, 把直接由计算机产生的数字信号进行**传输**的方式为()**传输**。

A.基带 B.宽带 C.调制 D.解调

扰动提问: 在计算机通信中, **发送**的是信号, 把直接由计算机产生的数字信号进行**发送**的方式为()**发送**。

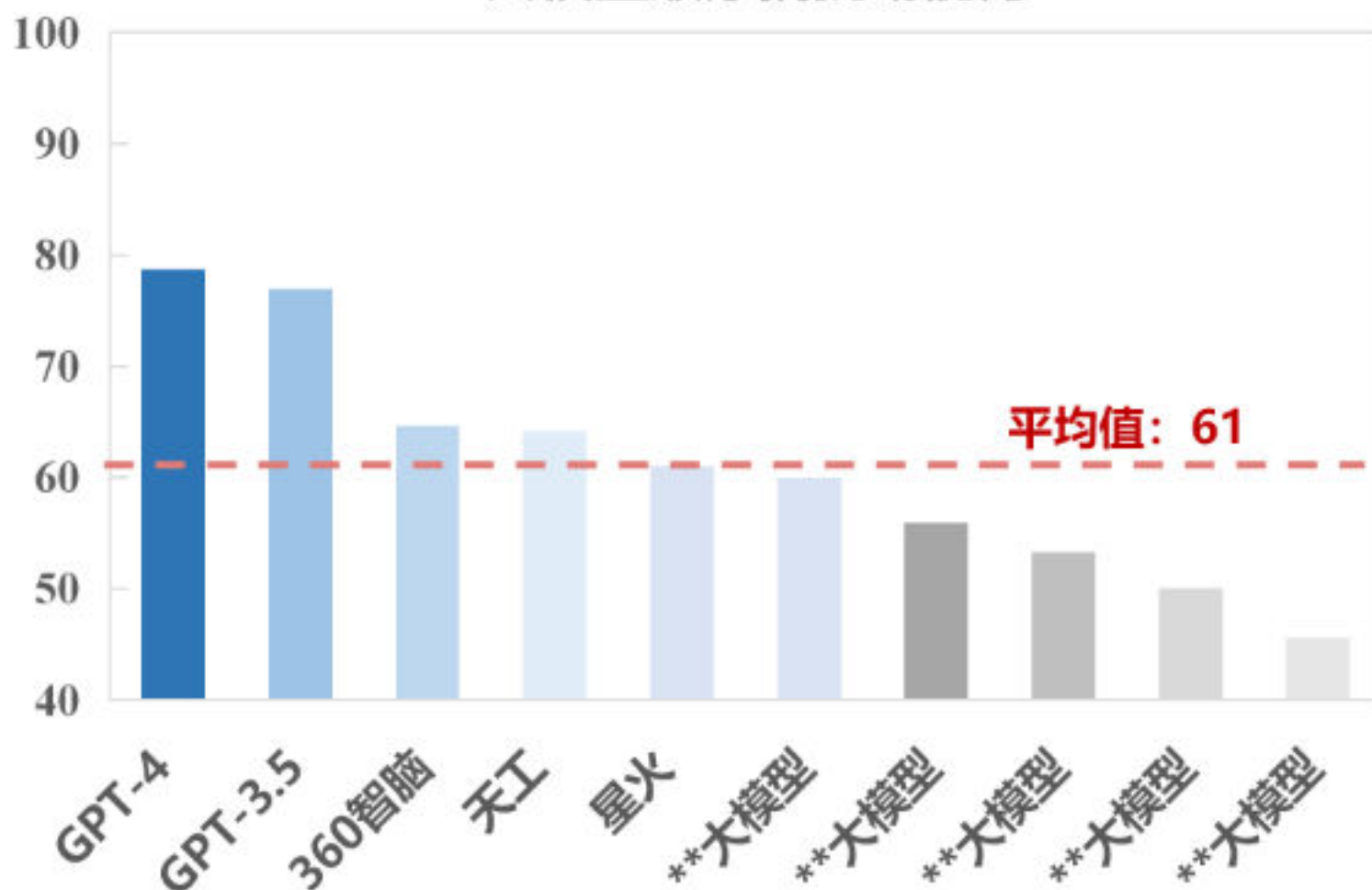
A.基带 B.宽带 C.调制 D.解调

- 对于同/近义变化, 当前国内外头部梯队大模型抗扰动能力较强, 已实现对GPT4/3.5的赶超;
- 对于尾部大模型供应商, 需要通过增加预训练语料的丰富度等方式提升语言理解能力;
- 对于工业企业, 为获得高质量回答, 要加强培训, 尽量使用常用词汇或进行多维度描述以减少歧义。

文法稳定能力—顺序抗扰动

- 大模型对工业领域语句表达顺序变换的回答稳定性。

大模型顺序抗扰动能力



测试样例：

原提问：物体受热时发生体积膨胀的现象称为()
A:热稳定性 B:热膨胀性 C:透气性 D:化学稳定性

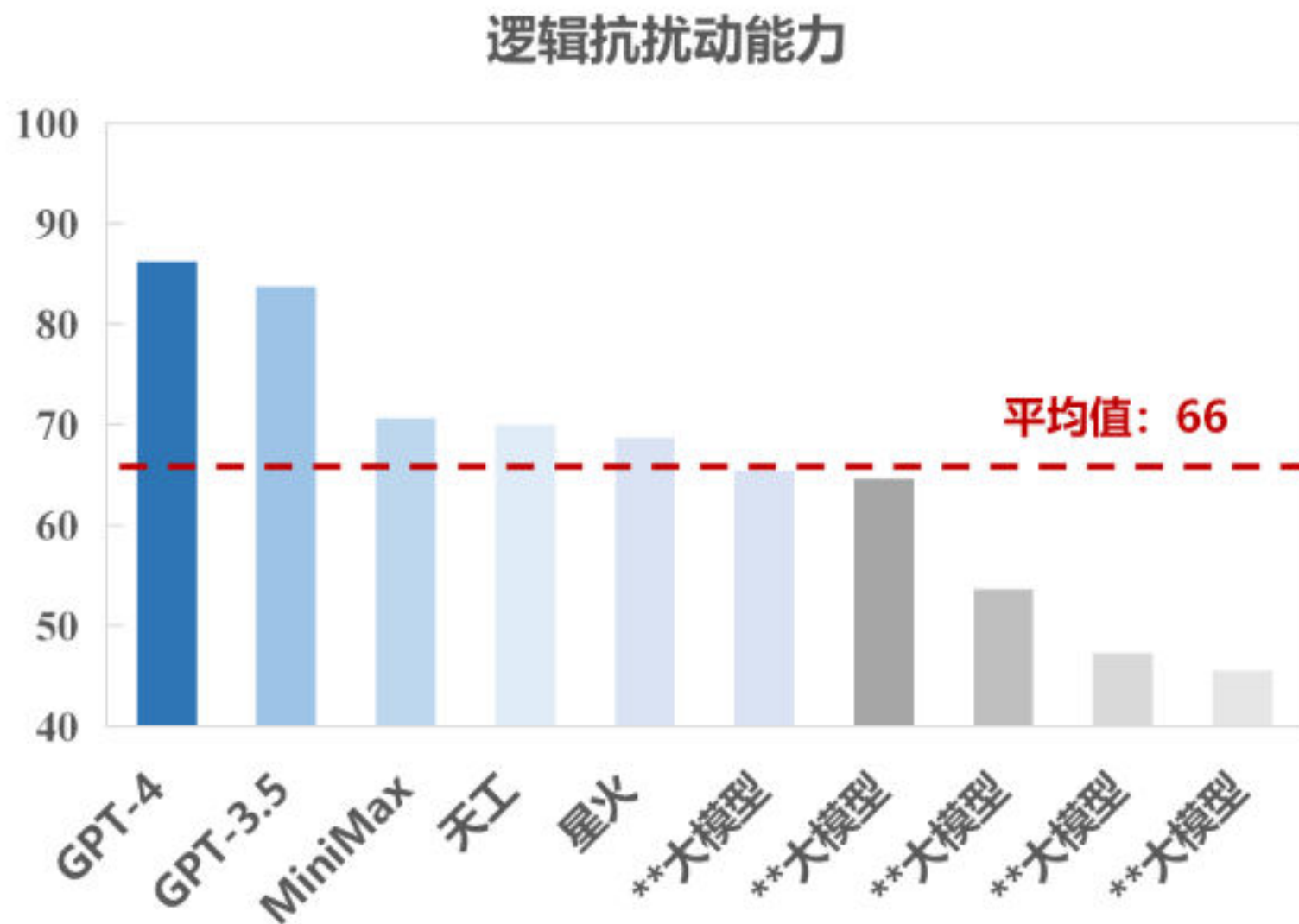
扰动提问：物体受热时发生体积膨胀的现象称为()
A:热稳定性 B:透气性 C:化学稳定性 D:热膨胀性

注：顺序变化测试变体涉及到“选项内容顺序”、“选项符号顺序”和“选项顺序”变化。

- 对于顺序变换，整体大模型稳定能力提升空间较大，且国外头部大模型表现能力相对较好；
- 对于尾部大模型供应商，需要通过语料变换等方式提升模型真正理解问题的能力，而不是记忆题目；
- 对于工业企业，要注意对测试语料的保护，使用充分变换的语料进行实际应用效果测试。

文法稳定能力—逻辑抗扰动

- 大模型对工业领域不同逻辑表达形式的回答稳定性。



测试样例:

原判断题: 受力物体内一点只要不受力, 就不会发生塑性变形。

扰动判断题: 如果物体内的某点没发生塑性变形, 那么该点一定没有受力。

注: 逻辑变化测试变体涉及“否定”、“双重否定”和“逆否”变化。

- 对于逻辑变换, GPT-4/3.5表现突出, 国内大模型的整体水平需进一步提升;
- 对于大模型供应商, 需持续改善模型逻辑能力, 如丰富逻辑语料, 完善提示词, 引入知识库, 工具函数集等;
- 对于工业企业, 在逻辑复杂场景中应谨慎使用大模型, 安全性要求高的场景建议进行人工检验。

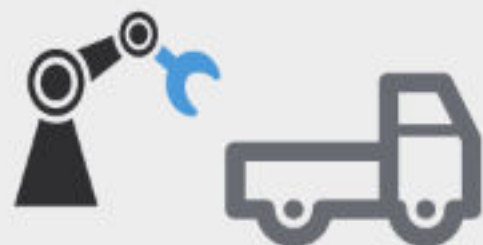
三、精简表达稳定能力

- 指大模型在工业应用中表达变化精简的输出稳定性。

工业场景潜在应用

1. 上下文精简扰动

指令：这是一批来自A工厂的货物，工厂那边昨天下雨，货物可能受潮，请将受潮货物单独挑出来



指令：请将货物中的受潮货物单独挑出来

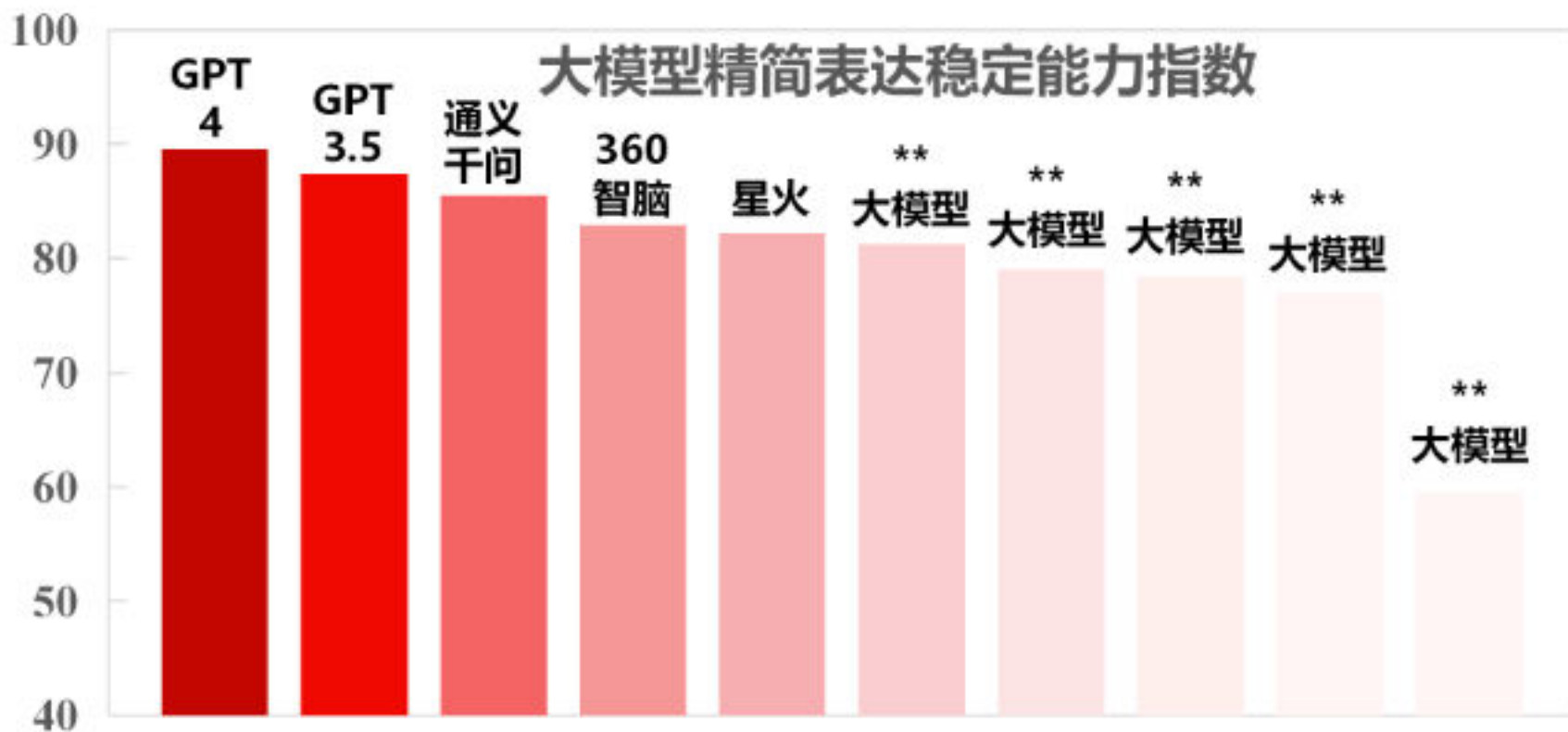


2. 安全守则精简扰动

指令：请按照《冶金企业安全生产监督管理规定》进行除尘器卸灰操作



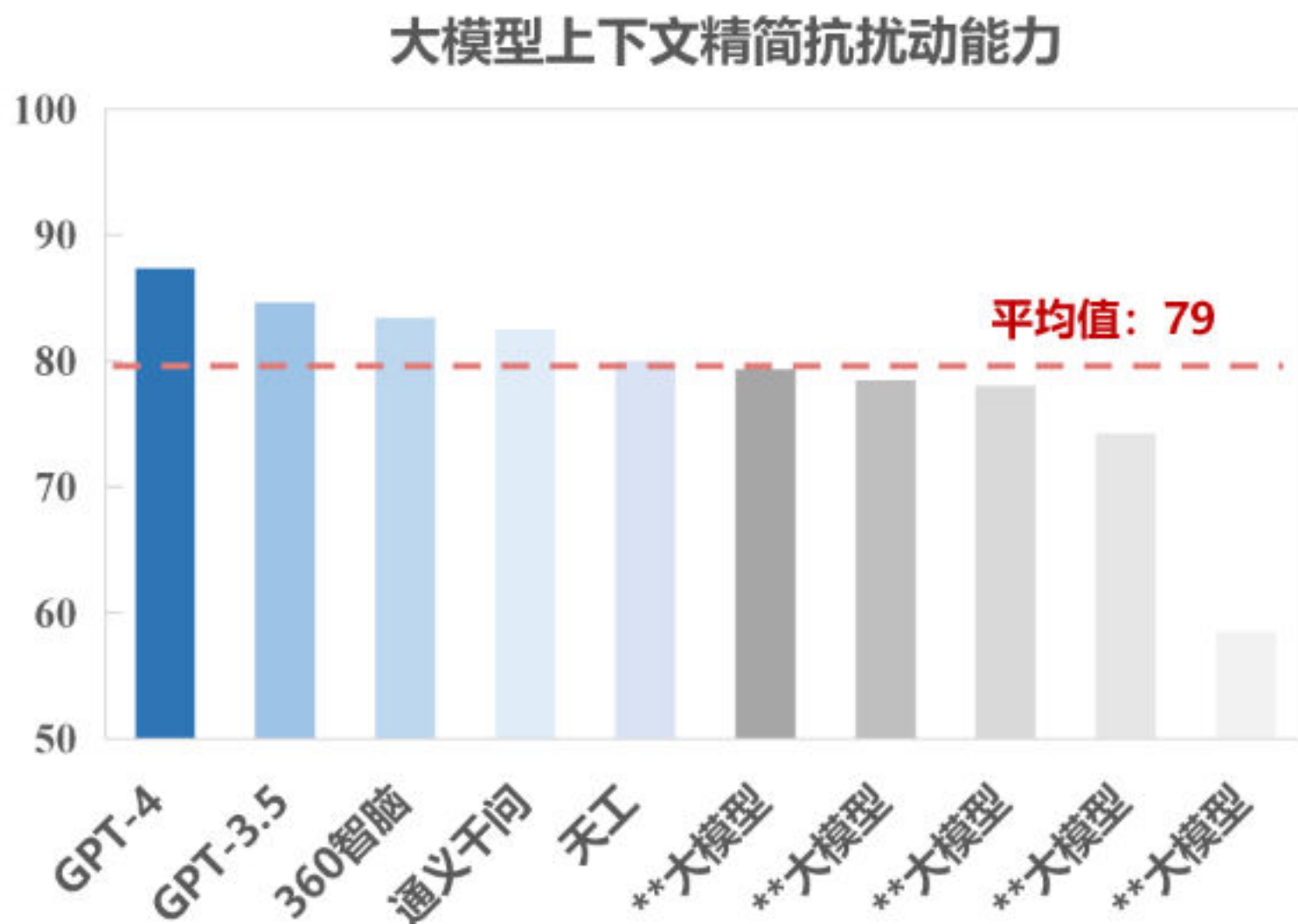
指令：请进行除尘器卸灰操作



- 整体看，国内外头部大模型能力在90分左右，能力较强。
- 对比看，GPT-4/3.5比较领先。国内头部能力在85分左右，国内尾部大模型低于60分，提升空间较大；
- 供给角度看，国内尾部大模型需要加强上下文处理能力和知识提取能力；
- 应用角度看，工业企业在安全操作要求较高的场景应慎重评估大模型的应用能力。

精简表达稳定能力—上下文精简抗扰动

- 大模型应对输入信息上下文精简变化的稳定性。



测试样例:

原提问: 在机械装备行业中, 根据基础物理研究, 物体受热时发生体积膨胀的现象称为()

A:热稳定性 B:热膨胀性 C:透气性 D:化学稳定性

扰动提问: 物体受热时发生体积膨胀的现象称为()

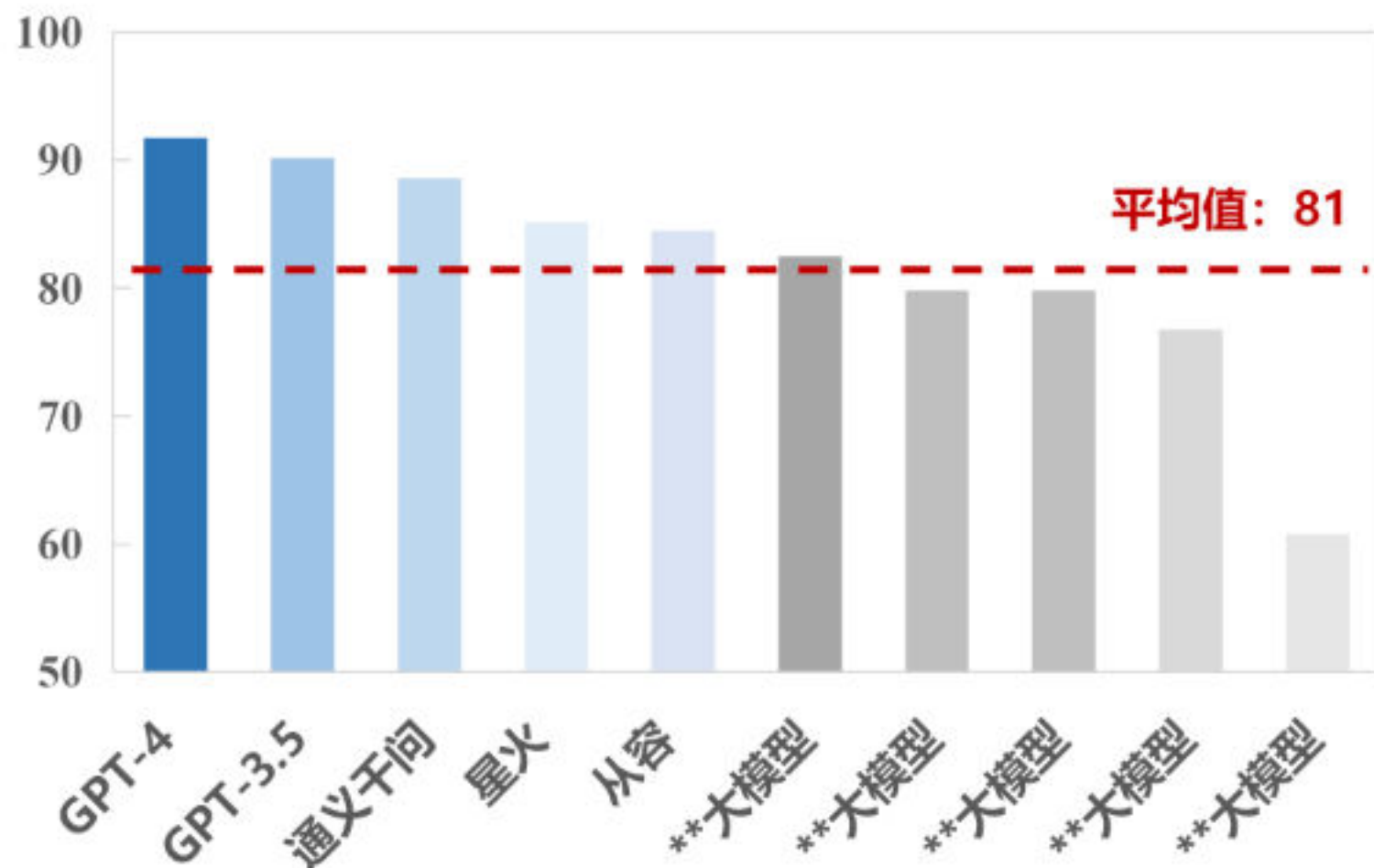
A:热稳定性 B:热膨胀性 C:透气性 D:化学稳定性

- 对于上下文变化的扰动, 国内头部梯队大模型的稳定性和GPT-3.5接近, 但尾部模型提升空间较大;
- 对于大模型供应商, 需进一步研究大模型提示词的机理, 提升模型对上下文的理解能力;
- 对于工业企业, 企业应用时应为员工进行培训, 提升提示词质量, 从而获得高质量的回答。

精简表达稳定能力—安全守则精简抗扰动

大模型应对安全生产指令的稳定性。

安全指令抗扰动能力



测试样例：

原提问：结合中华人民共和国工业和信息化部钢铁行业生产经营规范，对于钢铁行业，判断正确或错误：一般富氧1.0%，可提高理论燃烧温度35~45℃，增加喷煤率4.0%。

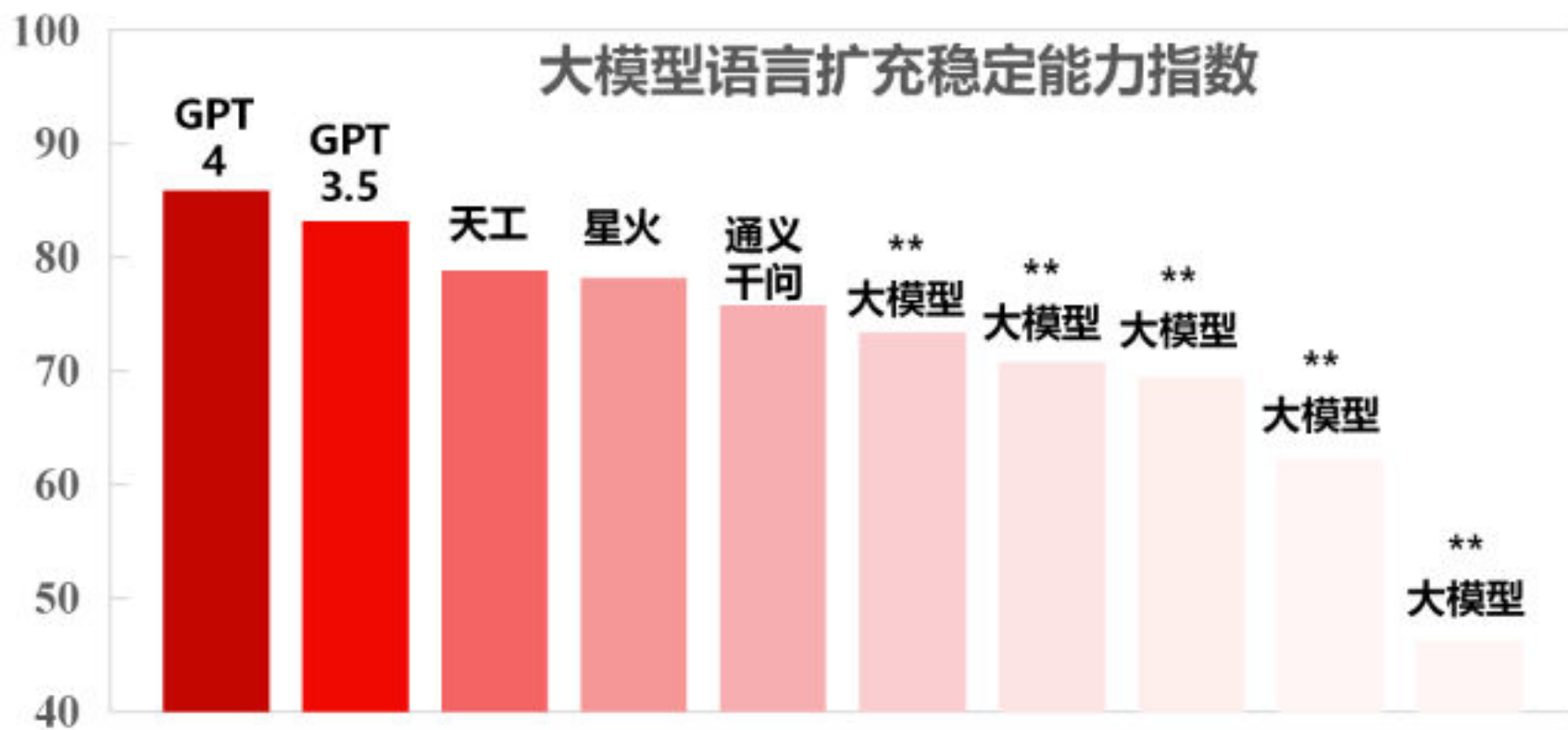
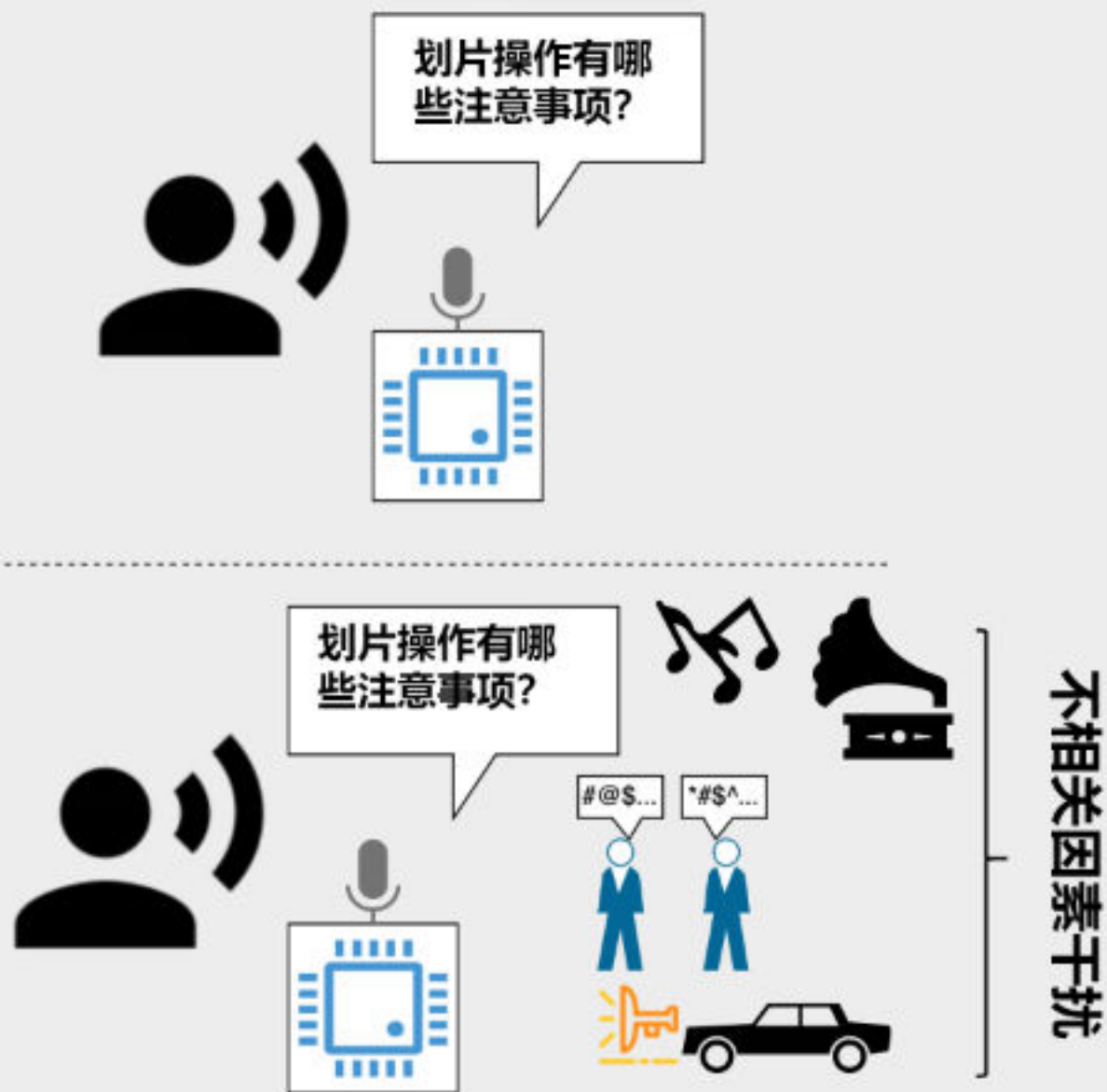
扰动提问：对于钢铁行业，判断正确或错误：一般富氧1.0%，可提高理论燃烧温度35~45℃，增加喷煤率4.0%。

- 对于安全生产指令变动的的影响，国内外大模型稳定性差距较小，尾部模型稳定性提升空间较大；
- 对于大模型供应商，需进行知识提取型任务的强化和优化，例如引入安全生产知识库等，以进一步提升模型稳定性；
- 对于工业企业，在安全性要求高的领域使用大模型需特别谨慎，尽量丰富表达。例如撰写施工计划等，需引入人工校验。

四、扩充表达稳定能力

- 指大模型在工业应用扩充表达的输出稳定性。

工业场景的潜在应用

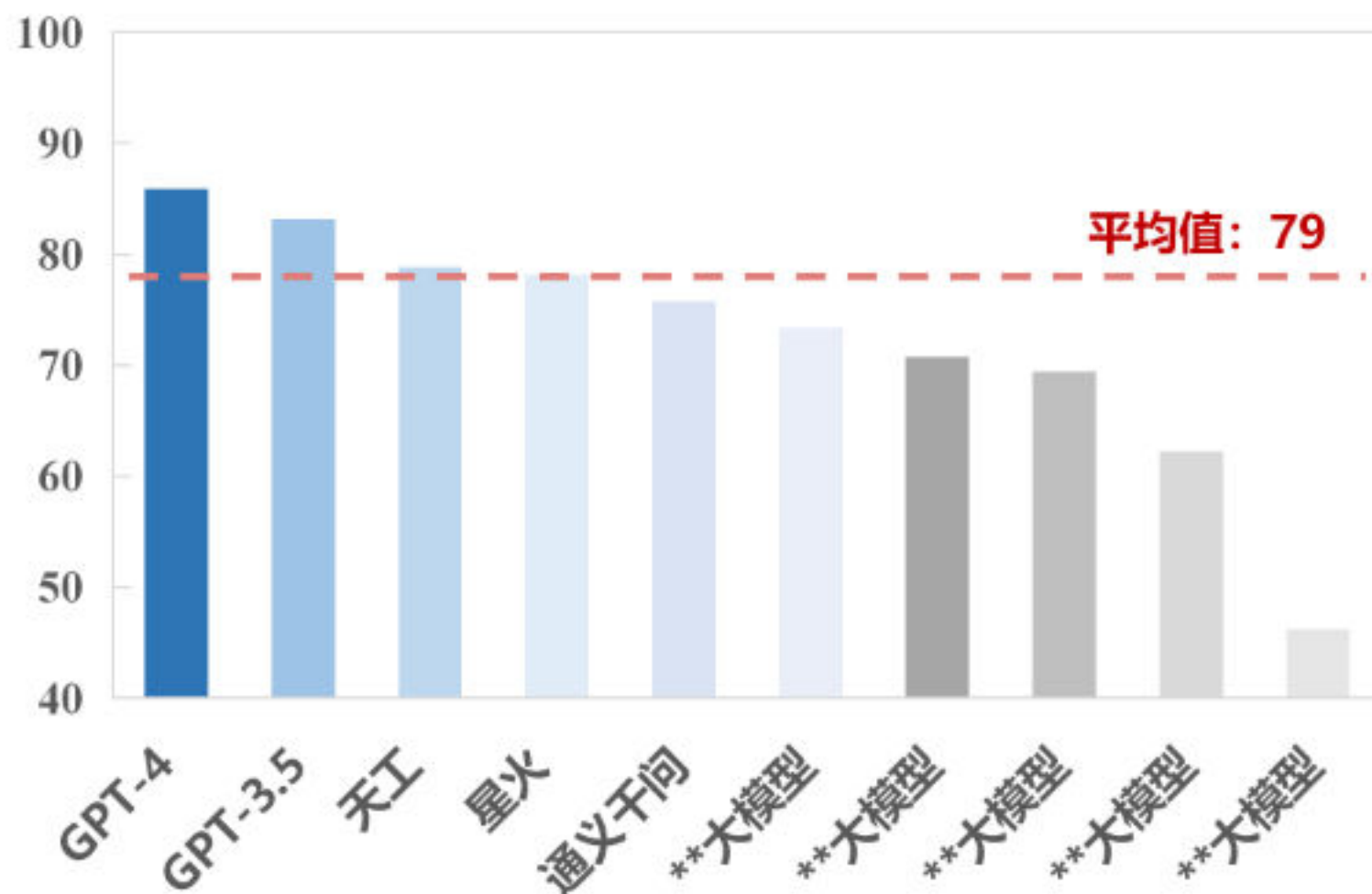


- 整体看，国内外头部大模型能力在80分以上，能够在较复杂上下文中保持了对关键词的敏感性；
- 对比看，GPT-4/3.5 比较领先，国内头部能力在80分左右，国内尾部大模型低于50分，有进一步提升空间，大模型间差距明显；
- 供给角度看，国内尾部大模型需扩大上下文窗口和强化对齐能力；
- 应用角度看，企业应用中可以通过精确指令和背景提升回答稳定性。

扩充表达稳定能力——不相关扩充抗扰动

- 大模型应对输入不相关文本信息量增加的回答稳定性。

信息提取抗扰动能力



测试样例:

原提问: 噪声最小的织机是 ()

A、剑杆织机 B、有梭织机 C、喷气织机 D、喷水织机

扰动提问1: 噪声最小的织机是 ()

A、剑杆织机 B、有梭织机 C、喷气织机 D、喷水织机

E、以上选项都相关

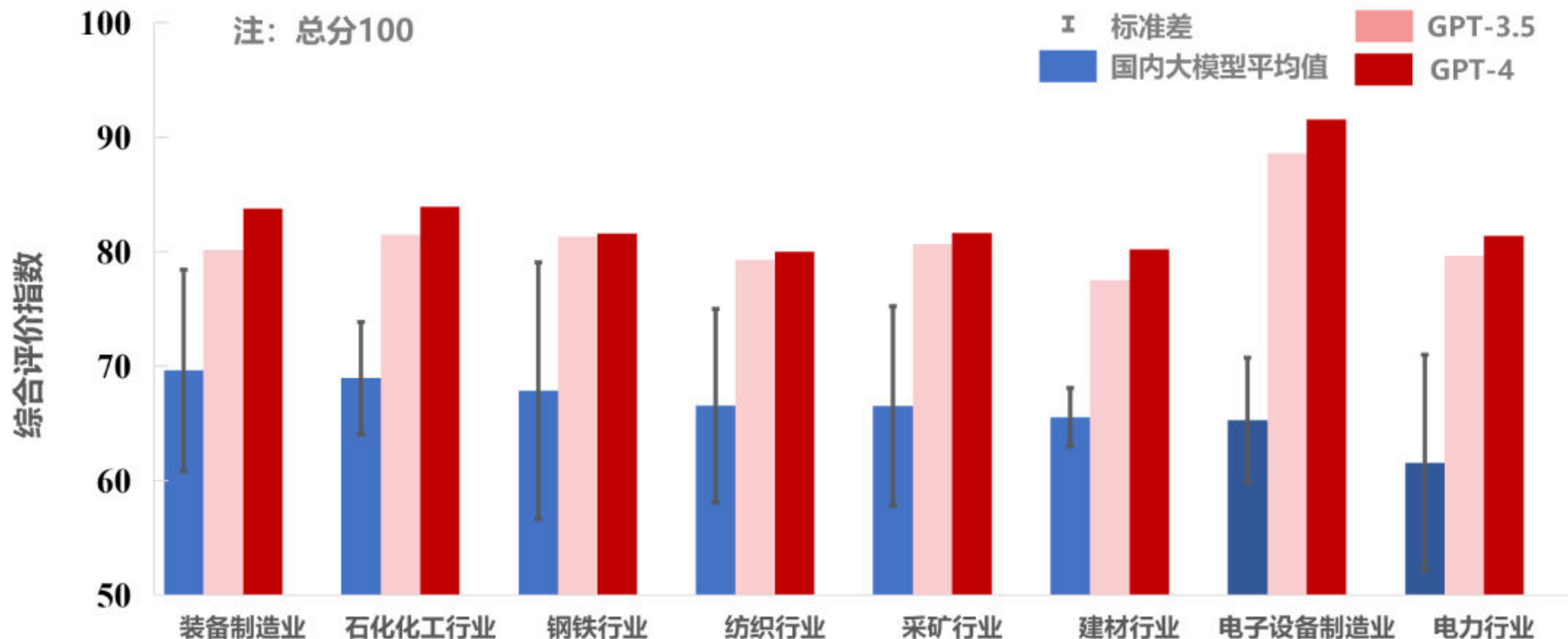
扰动提问2: **今天天气不太好**, 噪声最小的织机是 ()

A、剑杆织机 B、有梭织机 C、喷气织机 D、喷水织机

- 对于不相关扰动能力, 国内外头部梯队差距较小, 尾部能力提升空间较大;
- 对于尾部大模型, 需强化对上下文的理解能力和指令对齐能力;
- 对于工业企业, 在应用中要对使用人员进行提示工程培训, 使用精简的背景和指令提升稳定性。

工业领域知识问答稳定性评测结果

大模型在不同行业的知识问答稳定性评测

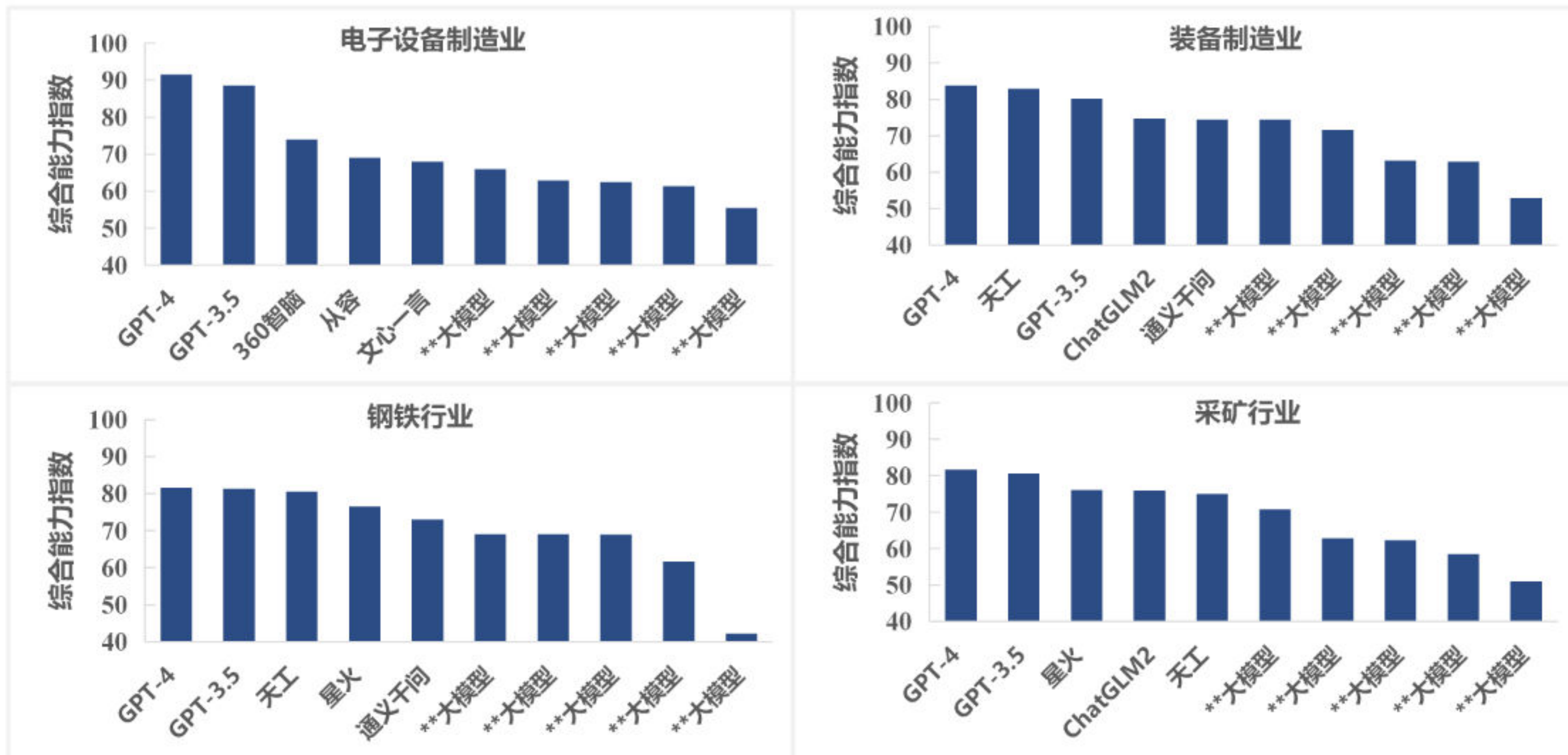


从行业维度看：

- 国内大模型整体水平较国外头部模型具有较大提升空间；
- 国外头部大模型在电子设备行业稳定性表现突出，国内大模型在八个行业知识问答稳定性差距较小，但波动性较大。

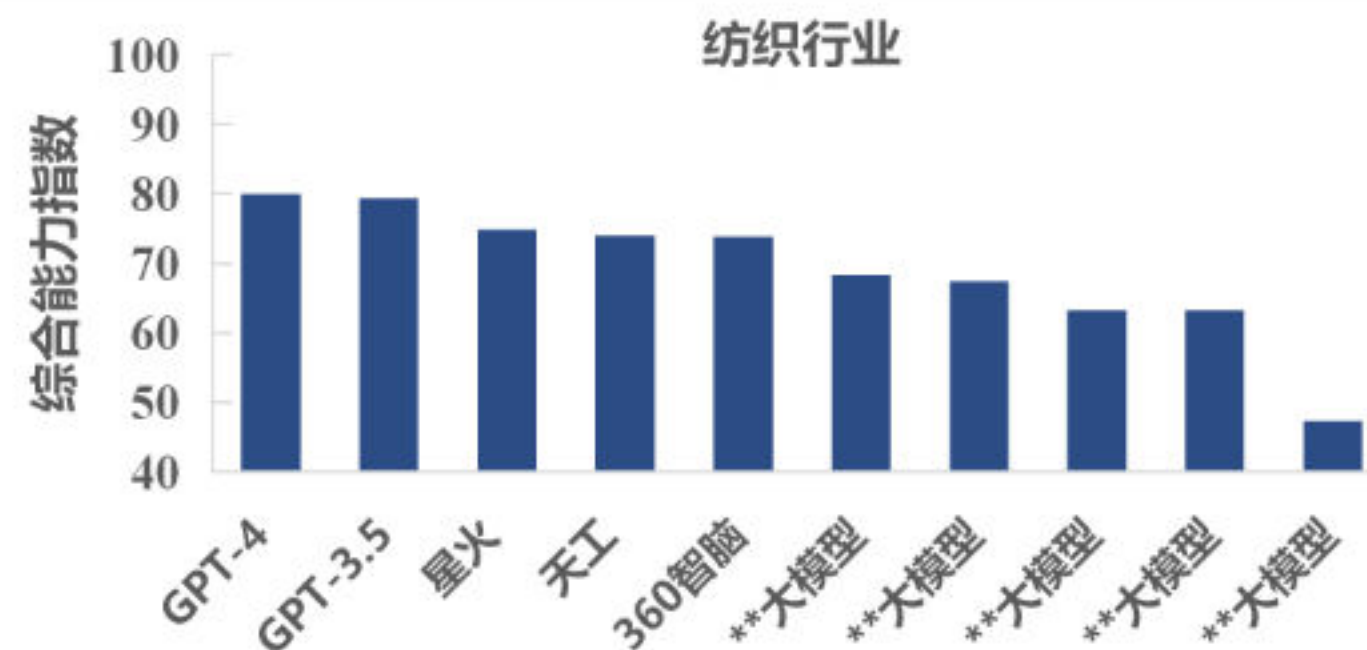
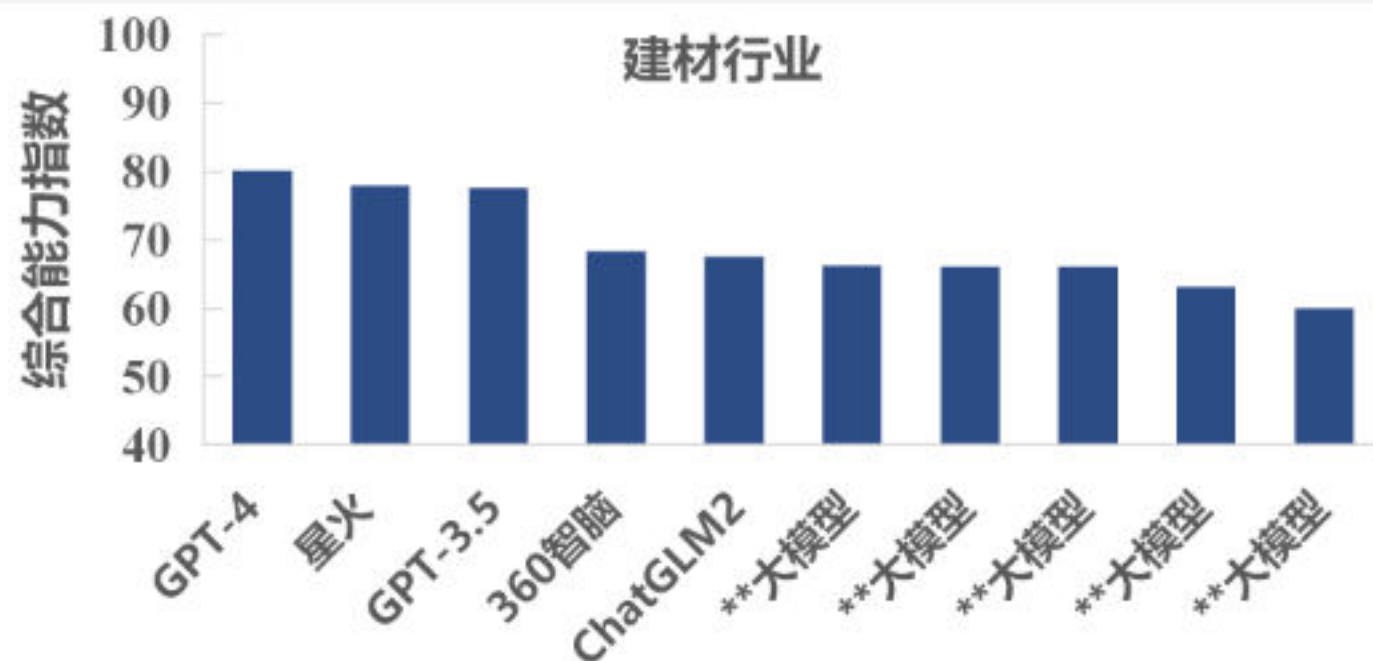
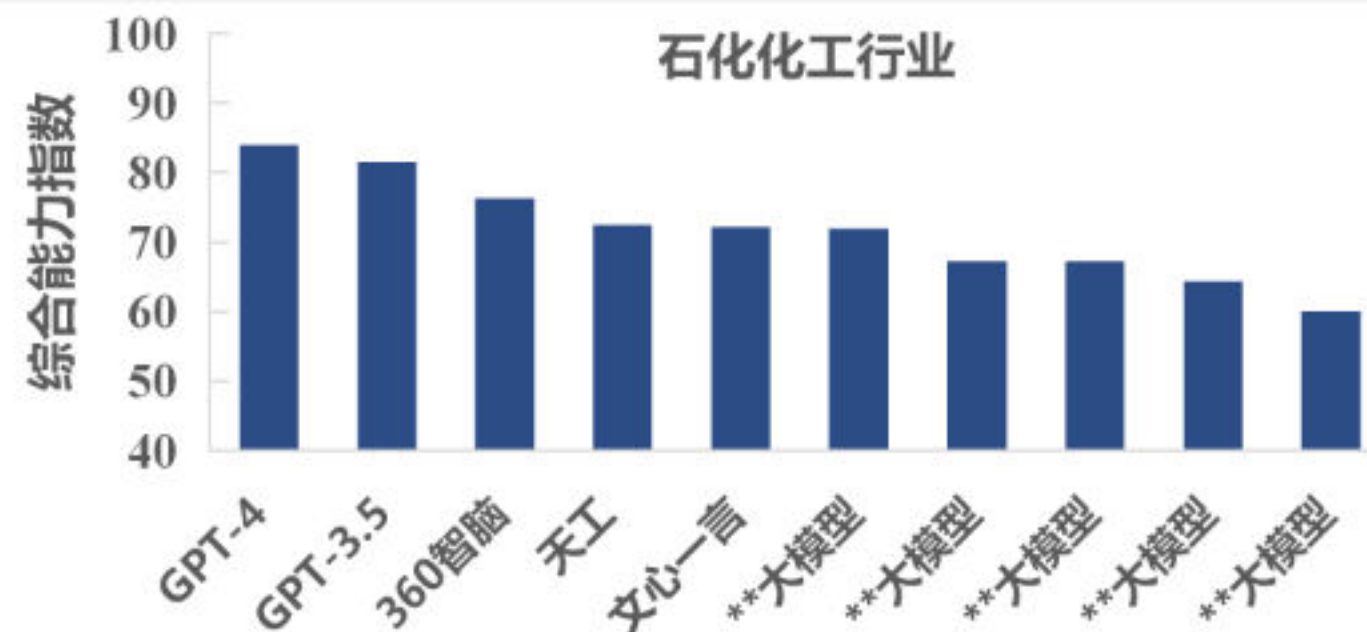
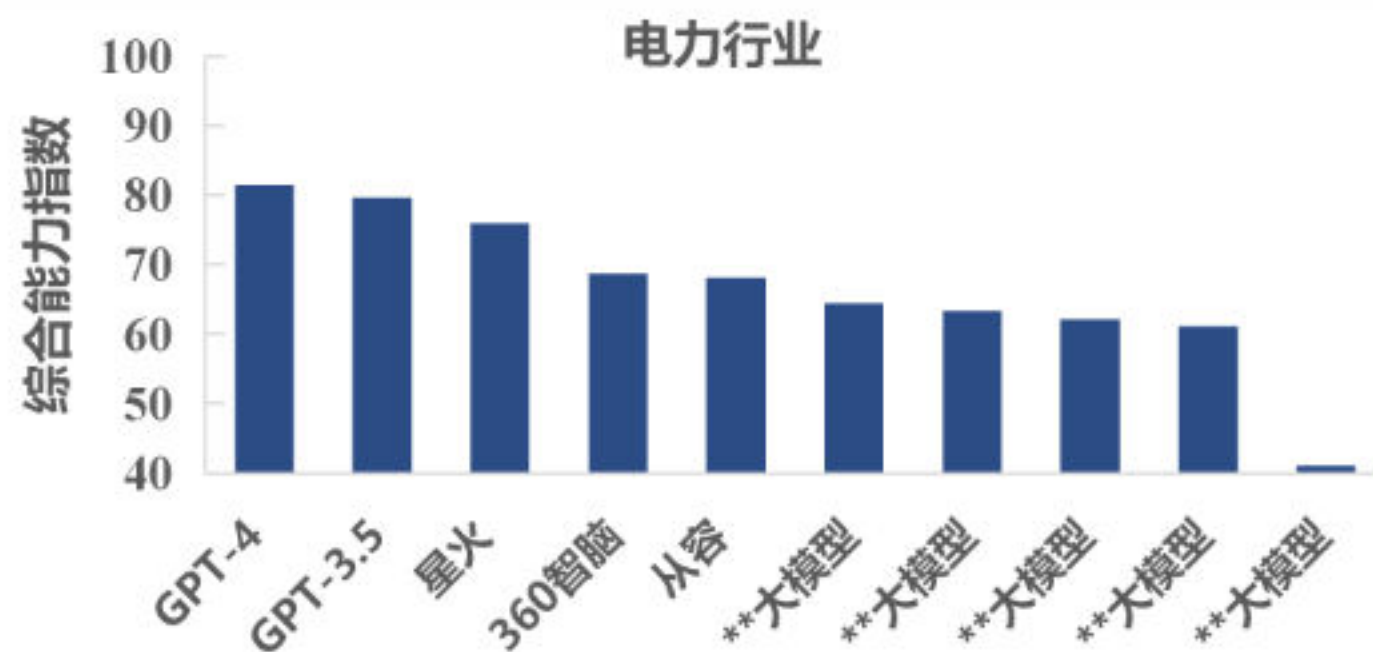
工业领域知识问答稳定性评测结果

各行业下大模型知识问答稳定性测评TOP5



工业领域知识问答稳定性评测结果

各行业下大模型知识问答稳定性测评TOP5



01

第一部分 评测规则

02

第二部分 评测结果

03

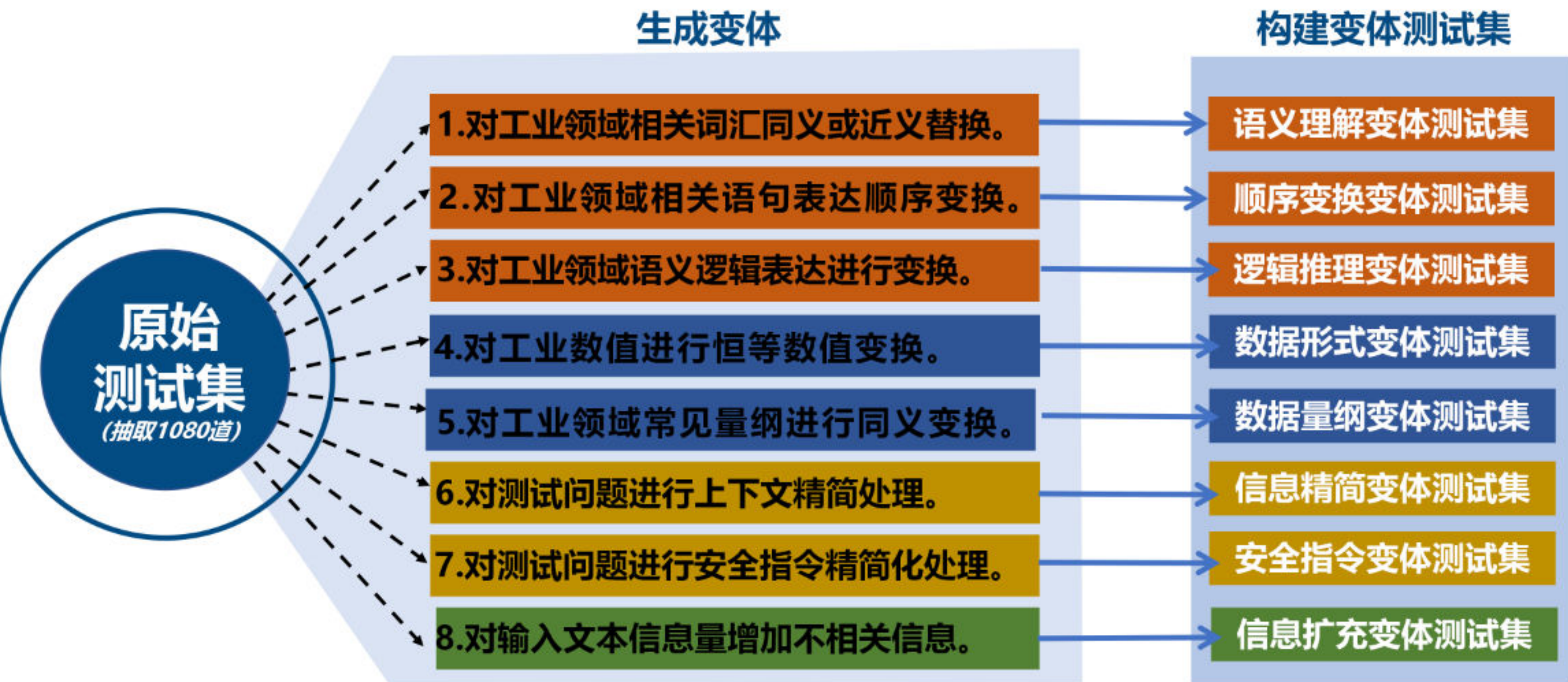
第三部分 总结

测评结论与建议：

- 整体上，国外头部大模型稳定性表现优异，国内大模型存在一定提升空间；
- 能力维度上，国内外头部大模型在数据稳定、精简、扩充表达能力差距较小，在文法稳定方面国内大模型提升空间较大；
 - 对于**数据稳定能力**，**建议**国内尾部大模型引入训练语料、知识库、工具函数集等，对数据专门进行强化；**提倡**工业企业将大模型应用于数据处理，但需要进行结果校验，尤其是对数据单位要进行多重描述，防止理解偏差。
 - 对于**文法稳定能力**，**建议**国内大模型对逻辑变化和对顺序变化的稳定性需进一步提升，并针对性进行优化；**提倡**工业企业对于文字简单处理的应用尝试使用大模型，但对逻辑密集型的应用谨慎使用。
 - 对于**精简表达稳定能力**，**建议**国内尾部大模型加强上下文处理能力和知识提取能力；**警示**工业企业在安全操作要求较高的场景应慎重评估大模型的应用能力。
 - 对于**扩充表达稳定能力**，**建议**国内尾部大模型需扩大上下文窗口和强化对齐能力；**提倡**工业企业应用通过精确指令和背景提升回答稳定性。
- 行业维度上，国内大模型整体水平较国外头部模型具有较大提升空间；国外头部大模型在电子设备行业稳定性表现突出，国内大模型在八个行业知识问答稳定性差距较小。

附录1：稳定性评测数据集

- **变体测试集：**从原始测试集中抽取客观题生成8类变体题目，构建变体测试集。



附录2：大模型评测版本

大模型名称	公司名称	版本信息	Web链接
GPT-4	OpenAI	gpt-4-0613	https://chat.openai.com
GPT-3.5	OpenAI	gpt-3.5-turbo-0301	https://chat.openai.com
通义千问	阿里巴巴	v1.0.5	https://qianwen.aliyun.com
文心一言	百度	v2.2.2	https://yiyan.baidu.com
星火	科大讯飞	v2.0	https://xinghuo.xfyun.cn
ChatGLM2	智谱华章	6b	https://open.bigmodel.cn
Minimax	Minimax	V5.5	https://api.minimax.chat
天工*	昆仑万维	v3.5.20230705.a	https://tiangong.kunlun.com
从容*	云从科技	v20230518	https://maas.cloudwalk.com
360智脑	360	V4.0	https://chat.360.cn

注：1.本研究实测模型包括但不限于上述通用大模型，此处只列举部分效果优异的模型版本信息。排名不分先后；
 2.标*处采用Web端测试方式，其余为API调用测试；
 3.行业大模型相关研究在后期进行。

下一步，中国工业互联网研究院将联合港科大、新华指数共同发布“中国AI大模型工业应用指数”，加速推动人工智能技术与实体经济的融合创新。在大模型问答性能、应用安全、标准制定等方面展开研究。欢迎交流合作。

转载请注明来源：中国工业互联网研究院。

联系人：叶老师 13661350566
吴老师 15210188099

中国工业互联网研究院

地址：北京市朝阳区酒仙桥北路
甲10号403号楼
电话：010-87901276
网址：www.china-aii.com

