中国工业互联网研究院
China Academy of Industrial Internet

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

中国经济信息社
CHINA ECONOMIC INFORMATION SERVICE

人工智能大模型工业应用准确性测评

2024年3月版

一、前言

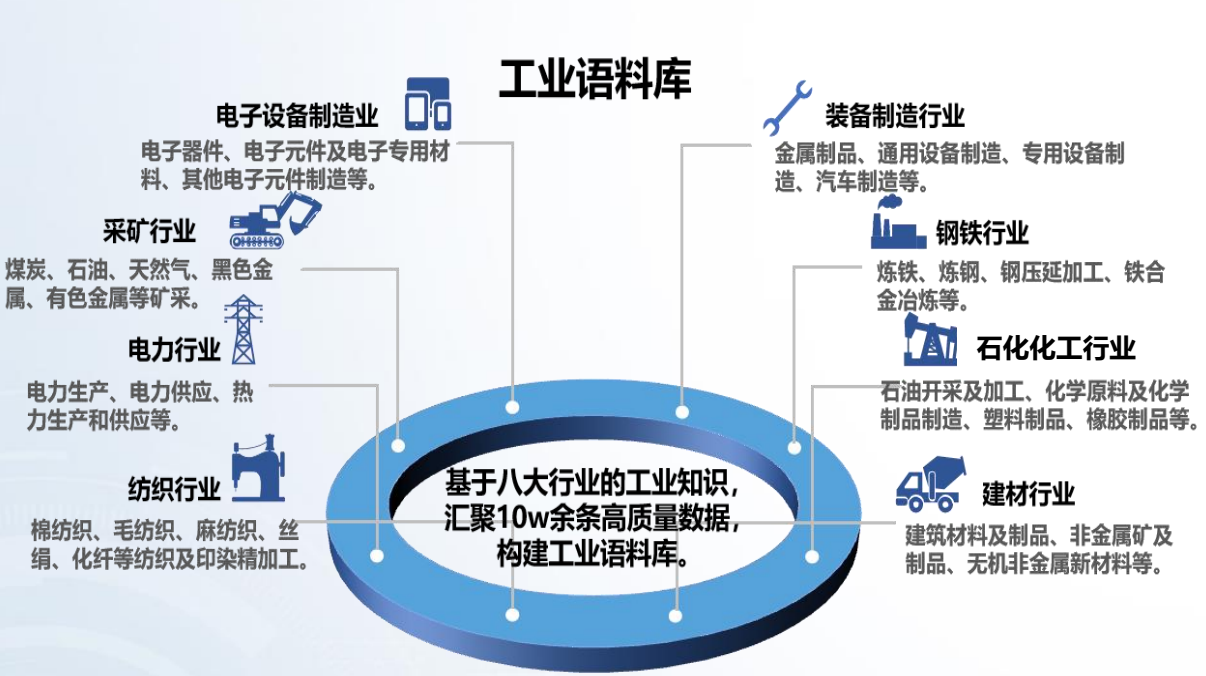
为贯彻落实党中央国务院关于促进人工智能发展的决策部署，中国工业互联网研究院依托通用人工智能与工业融合创新中心（简称“中心”），联合香港科技大学、中国经济信息社，深入研究人工智能大模型在工业领域的应用性能、技术架构、标准体系，并在此基础上，形成本报告。

结合工业企业大模型应用情况调研，本报告在原有**工业知识问答准确性测评**的基础上，新增**数据分析、工程建模、文档生成、代码理解**等四大场景，构建测试数据集，对国内外具有代表性的大模型进行测试，发布新一轮的准确性测评报告，供业界进行参考。

本报告测评结果虽经中心专家委论证，但因大模型迭代速度快，技术复杂，囿于工作团队专业知识和能力，报告难免存在分析结论不足等问题，且测评结果仅适用于测试期间，欢迎大家批评指正。

二、测评内容

2023年初至今，大模型技术发展突飞猛进，已逐步渗透至工业领域诸多环节，涵盖了知识问答、工程建模、数据分析、文档生成、代码理解等场景，正快速成长为工业转型升级和创新发展的重要动力。



- 依托国家工业互联网大数据中心，聚焦重点工业行业，汇集高质量语料，形成工业语料库，支撑大模型在工业领域应用测评；
- 结合工业企业调研，在原有知识问答基础上，新增四类工业应用测评场景，开展大模型在各应用场景的准确性测评。

三、测评方法

➢ 测评流程

01 筛选题目

根据场景、难度、行业，选取有标准答案的题目，经人工校验后形成测试题。

02 进行问答

调用待测试大模型API，收集大模型答案。

03 生成判分标准[1]

利用GPT4将原有标准答案整理为评分标准，并通过人工校验提升判分标准科学性。

04 进行判分[2]

利用GPT4，根据评分标准，按步骤赋分。

➢ 评分标准

1. **题目类型**：每个场景抽取若干题目进行测试，题型以问答题为主。
2. **题目数量**：
   - 知识问答：144 道
   - 数据分析：20 道
   - 工程建模：100 道
   - 文本生成：40 道
   - 代码理解：150 道

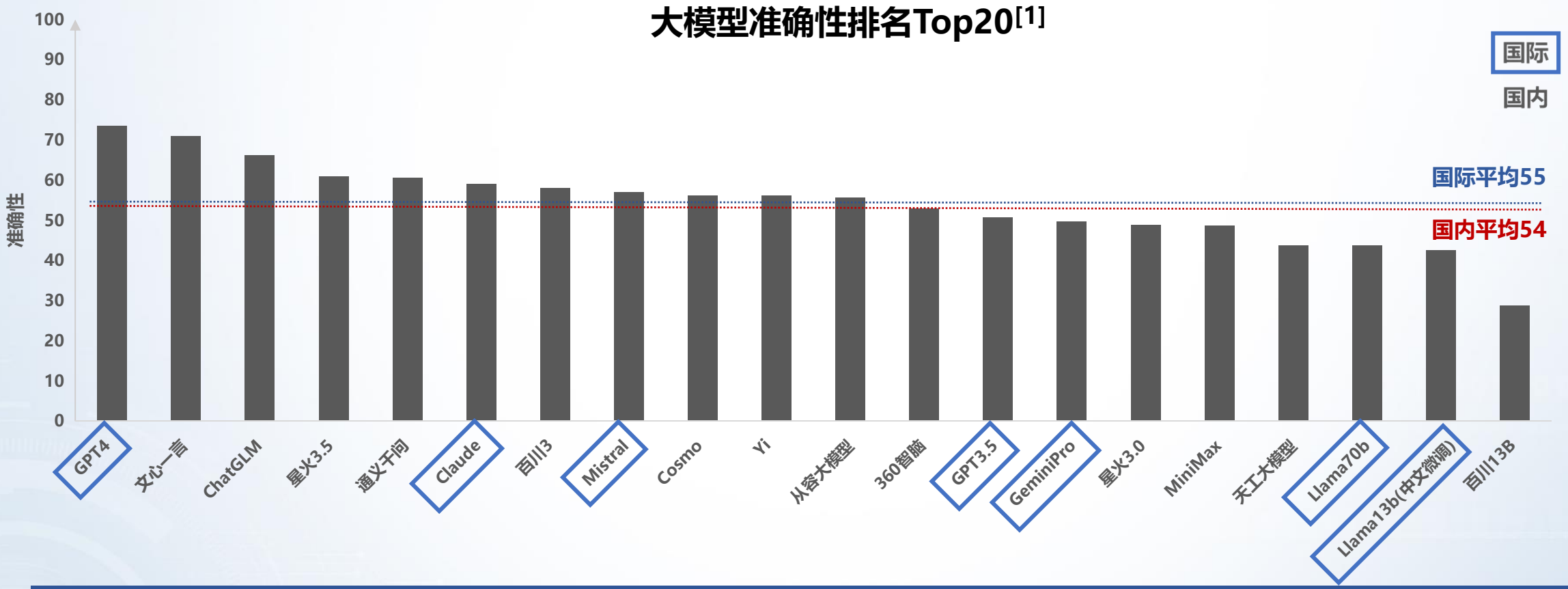
   注：各场景题目数量虽不一致，但考察要点总量保持在同一个数量级。
3. **题目得分**：需要结合具体题目的评分细则，按照步骤进行赋分，赋分后分数进行归一化处理。
4. **场景得分**：
   - 场景得分为题目总分百分化处理后的分数。
   - 若有细分场景，则场景总分为细分场景的平均成绩。
5. **综合评分**：由各场景算数平均分计算得出。

- **为更贴合应用场景实际，进一步评价模型的多维能力，本期测评题型以问答题为主；**
- **为保障判分的一致性与准确度，问答题的评分方式由人工判分改为大模型判分，并按步骤赋分。**

[1] 对于GPT4，先获取其回答，再用其生成标准答案、进行判分，避免信息泄露；
[2] GPT4的API承诺不记录数据用于训练，参考业界成熟方案，使用GPT4的API生成标准答案和判分结果，减少测评误差。

四、测评结果-综合排名

➢ 测评成绩

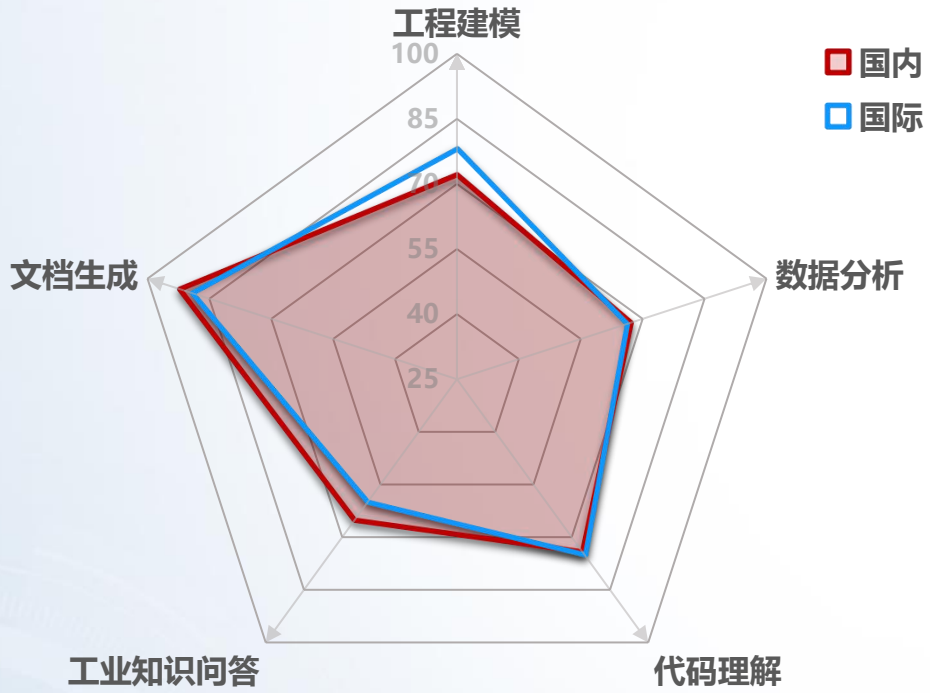


- 综合能力上，GPT4处于领先地位，国内大模型文心一言、ChatGLM紧随其后；
- 对于国内大模型，多个模型综合能力超过GPT3.5，包括文心一言、ChatGLM、星火3.5、通义千问等；
- 对于国外大模型，GPT4领先优势明显，其余模型差距较大。

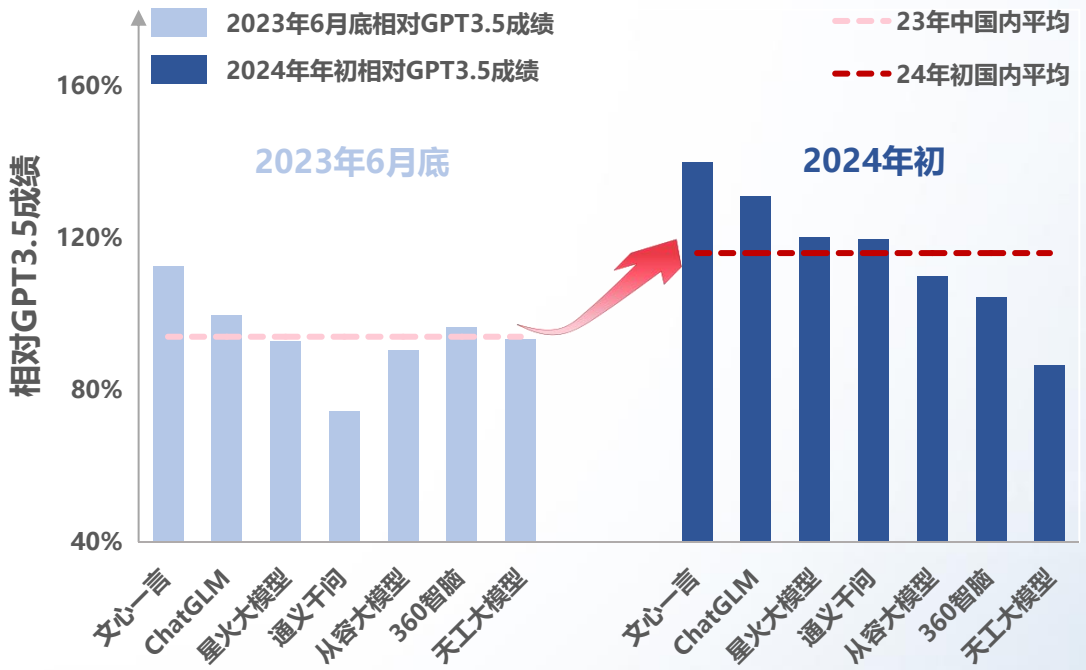
[1] 模型版本号参见附录1。

四、测评结果-能力对比与变化趋势

➢ 各维度大模型最佳能力对比图[1]



➢ 国内大模型发展趋势[2]



- 在工业知识问答、文档生成等领域，国内大模型已取得领先，数据分析、代码理解等领域能力接近；
- 在工程建模领域，国内大模型与国际存在一定差距。

- 对比往期测评，2023年下半年国内大模型能力提升明显（以GPT3.5为基准）。

[1] 选取国内外各能力维度性能最佳的大模型进行对比；
[2] 国内大模型发展趋势统计规则见附录2。

五、场景测评一：工业知识问答

大模型可结合自身知识，回答不同工业领域问题，将用于员工培训、故障诊断、客服咨询、市场调研等交互场景，协助企业员工熟悉生产流程，帮助用户了解产品特性。

知识快速获取

工艺辅助优化

数字人售后服务

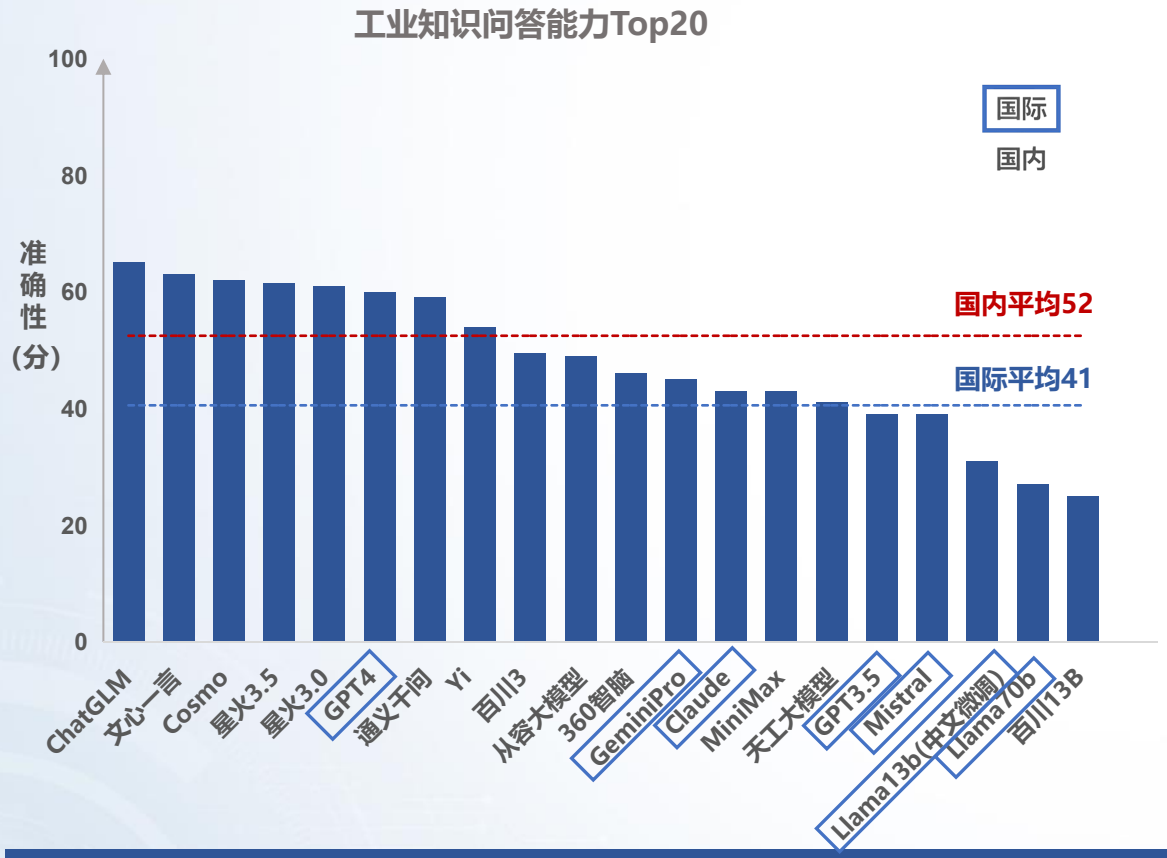
员工自助培训

应用场景研判

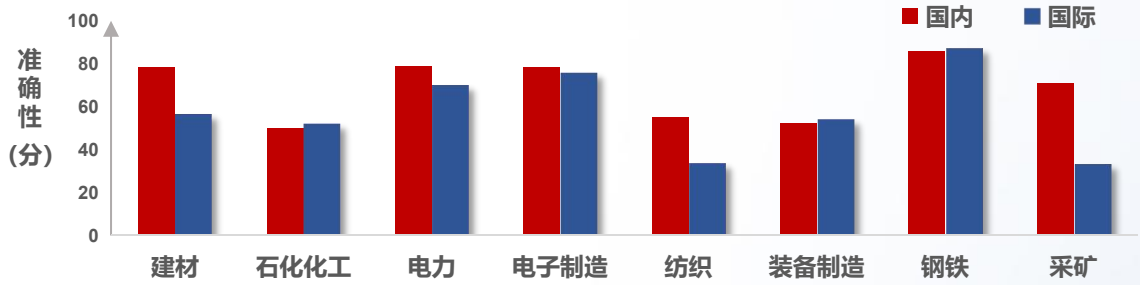
- **研发设计环节**：研发工程师可基于大模型快速、便捷获取高质量知识，提升研发效率；
- **生产制造环节**：产线工人可实时向大模型查询生产工艺经验，辅助其进行制造工艺优化；
- **售后服务环节**：企业可基于大模型，通过数字人实时向客户提供售后咨询服务；
- **技能培训环节**：新员工可通过大模型了解企业信息、学习生产技能。

五、场景测评一：工业知识问答

➢ 测评结果



➢ 行业能力对比[1]



➢ 题目样例

问题：

你知道哪些常用逻辑电平？TTL与CMOS电平可以直接互连吗？

评分标准：

(1) **常用逻辑电平包括：12V，5V，3.3V。**（1分，给出标准中同样或近似的回答则得1分，否则不得分。）

(2) TTL和CMOS电平是否可以直接互连：**不可以直接互连。**（1分，给出标准中同样或近似的回答则得1分，否则不得分。）

(3) TTL和CMOS电平互连的条件：**CMOS输出可以直接接到TTL，而TTL接到CMOS需要在输出端口加一上拉电阻接到5V或者12V。**（1分）

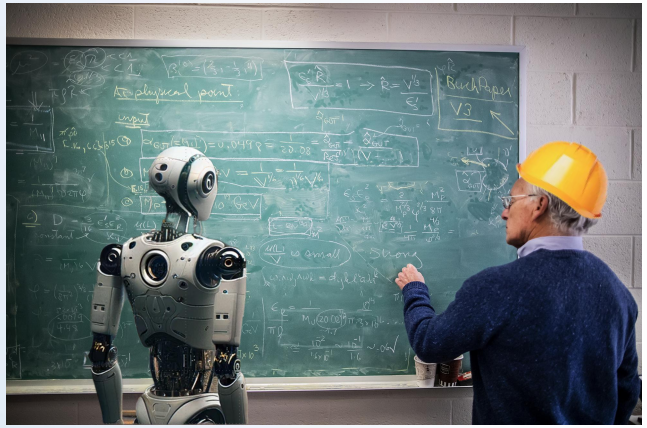
本题共3小项，每个小项1分，满分3分。对于每个小项，如果描述有差距，或者详细程度不足，酌情给0.3或者0.5分或者0.8分。

- **在知识问答领域国内大模型已具备一定优势，ChatGLM、文心一言等多个大模型实现对GPT4超越；**
- **国内大模型在建材、采矿等行业具有显著优势，在装备制造、钢铁等行业与国际水平接近；**
- **对比不同行业，国内外大模型在钢铁、电力等行业有较好的知识储备，对于纺织、装备制造等行业仍需加强训练。**

[1] 图中数据为各行业国内外性能最佳大模型成绩。

五、场景测评二: 工程建模

大模型具备基础建模能力，将帮助工程师和企业管理人员在实际工程设计、生产运维等领域进行数学建模，寻求最佳的解决方案。



工程数学建模

预测模型优化生产计划

优化员工班次布局提高人效

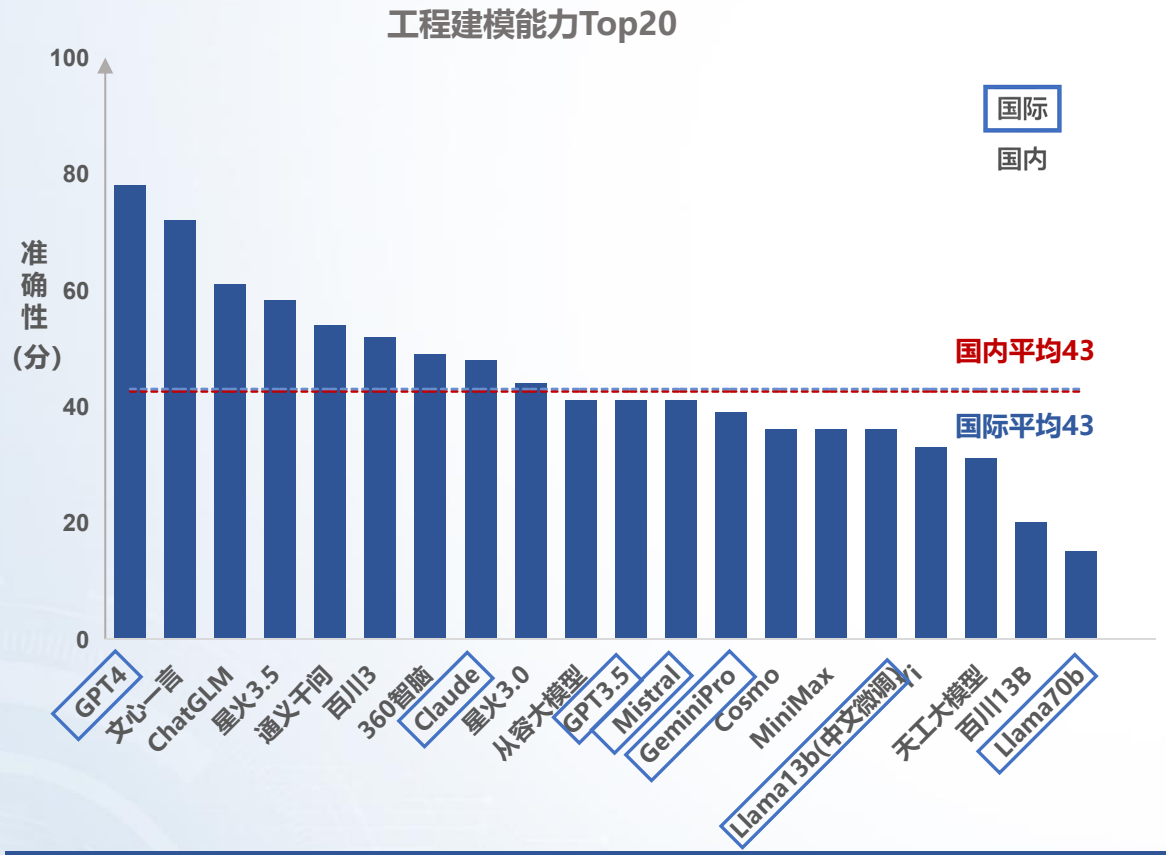
营销收益建模节约销售成本

应用场景研判

- **研发设计环节**：基于历史实践，建立成本模型，指导新项目的规划和预算编制，提高项目成功率；
- **生产制造环节**：建立时序预测、异常检测模型，基于预测优化生产计划，提高施工效率和安全性；
- **运维管理环节**：建立运筹模型，对工厂生产人员进行排版优化，提升人员效能；
- **营销宣传环节**：建立营销收益模型，提升营销效率，节约营销成本。

五、场景测评二: 工程建模

➢ 测评结果



➢ 题目样例

问题:

某公司在2018年年初预订x万产量的目标，2018年6月已完成计划的60%，此后按照上半年月均产量生产，则2018年超出计划产量300万. 那么该公司2018年年初预订的产量为多少万元?

评分标准:

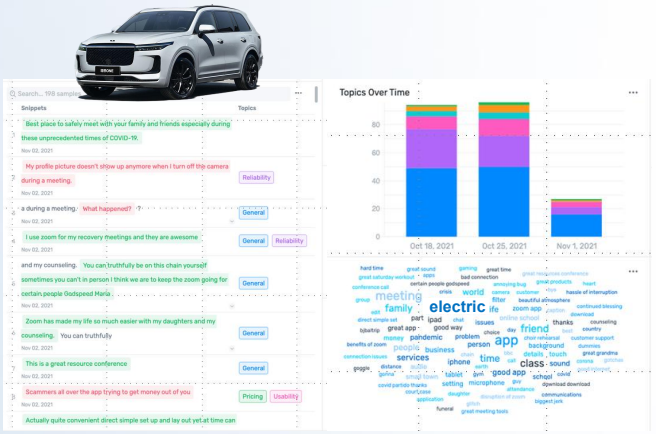
1. 如果能正确列出**完成计划的60%即为 $0.6x$ 万的关系，得1分**;
2. 如果能正确列出**下半年产量也为 $0.6x$ 万的关系，得1分**;
3. 如果能正确**列出并解方程 $0.6x + 0.6x - x = 300$，得1分**;
4. 如果能正确**解出 $x=1500$，得1分**;

本题共四个得分点，满分为4分，得分情况为（得分/满分）。

- **在工程建模领域，GPT4、文心一言处于领先地位，对比其它模型具有显著优势;**
- **国内外平均成绩均为43分，大模型建模能力整体处于较低水平，可收集数学建模专业语料进行强化训练，也可以使用代码解释器等增强工具提升大模型建模能力。**

五、场景测评三: 数据分析

大模型可将结构化数据提炼为核心结论，对复杂业务数据进行自动分析，更全面、及时地帮助企业管理者运营和决策，提升工作效率和运营质量。



分析用户评价

分析生产时序数据

分析库存数据

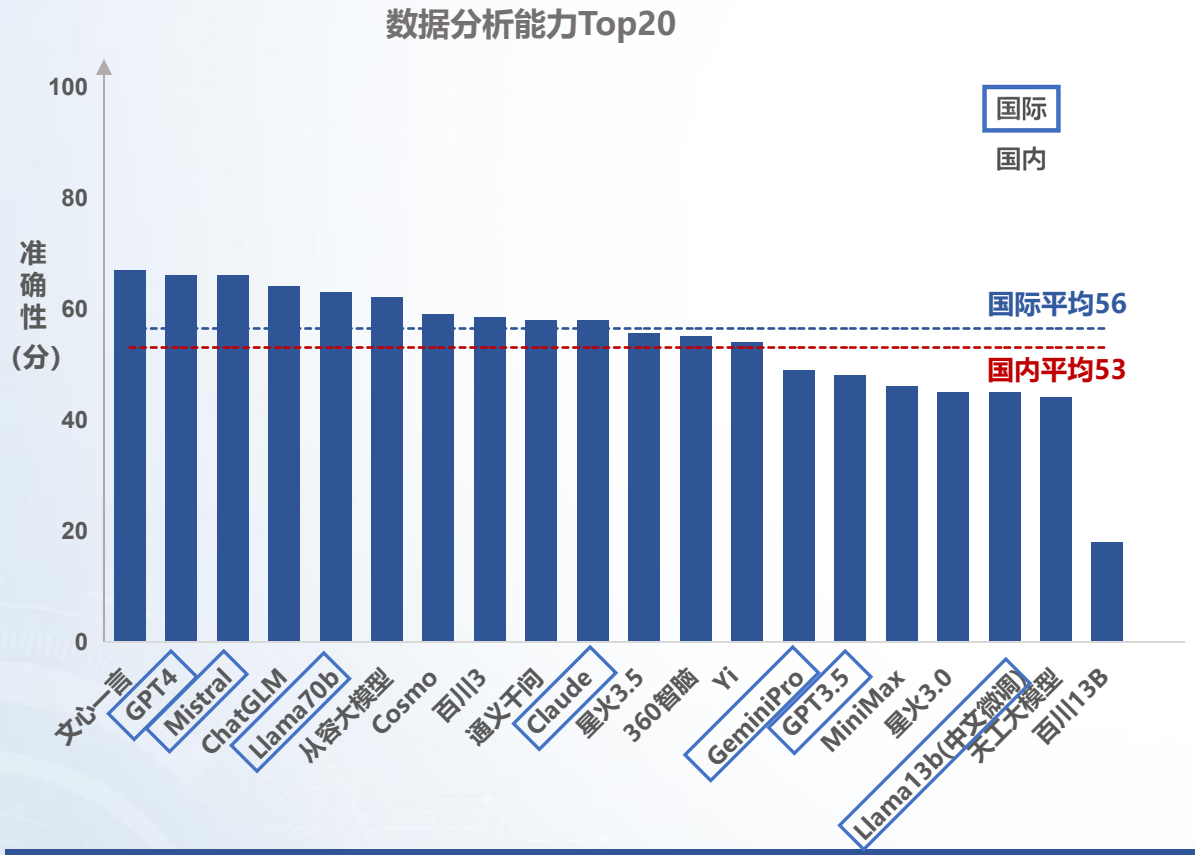
分析安全数据

应用场景研判

- **研发设计环节**：在海量产品评价数据中提取共性问题，改进产品设计，提升产品品质；
- **生产制造环节**：自动分析工业生产时序数据，发现数据异常或潜在风险，及时预警或报错；
- **运维管理环节**：辅助分析库存数据，进行呆滞库存和缺料提醒，提升管理效率；
- **人员培训环节**：分析事故数据，杜绝生产事故，消除安全隐患。

五、场景测评三: 数据分析

➢ 测评结果



➢ 题目样例

问题：您需要撰写一份简短的报告，介绍下面的图表/表格/图形的主要特征。您应该执行以下任务：概括数据，描述过程的各个阶段等等，请使用中文进行撰写。下表为2001-2010年几种型号电话年产量。

Year	PhoneA	Phone B	...
2001	200	700	...
...	...	...	...
2010	700	475	...

评分标准:

(1). 文章对比了2001年至2010年几种电话的年均产量变化。（1分，如果有相关的全局性描述，则得1分，否则不得分。）

(2). 在这10年期间，B电话稳步下降，而A电话支出迅速增长。（1分**必须有B稳步下降的描述，且有A电话迅速增长的描述性语言**，只给出数据不进行对比描述不得分。）

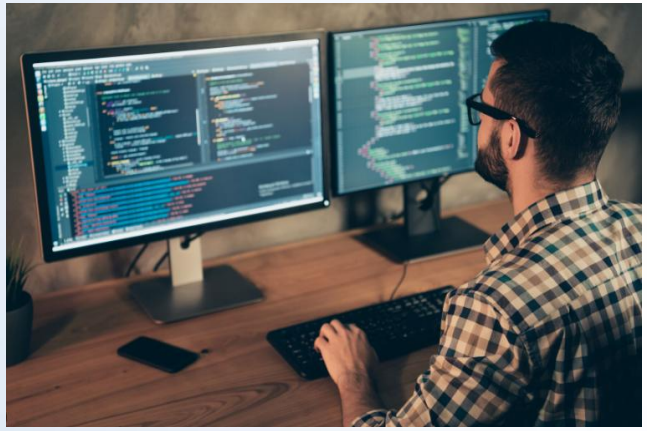
(3). 2007年是A产量超过B电话产量的转折点。（1分,**必须指出2007年A电话超过B的关键节点**，只给数据出数据不描述不得分。）(4)... (5)...

本题共5小项，对于每个小项，如果学生的回答中有和该项一致的语句，则得1分，如果描述有差距，或者详细程度不足，酌情给0.3或者0.5分或者0.8分。

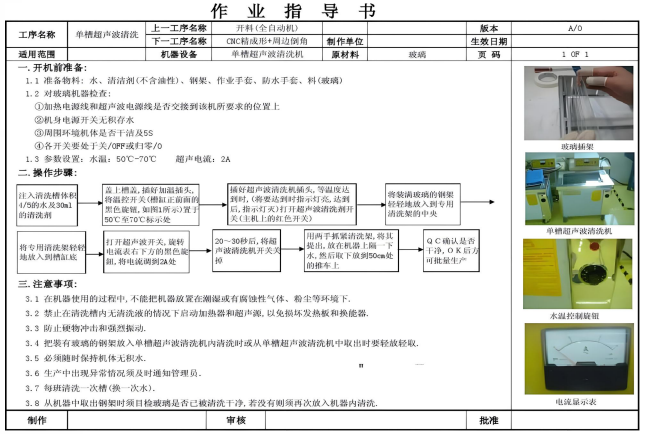
- **在数据分析领域，文心一言能力最佳，与GPT4、Mistral等构成第一梯队；**
- **国内外大模型分数均较低，大模型直接用于数据分析可能造成部分信息遗漏或描述偏差，实际应用中可使用优秀数据分析案例进行微调，或将案例加入到提示词中，利用大模型小样本学习能力提升效果。**

五、场景测评四: 文档生成

大模型将帮助用户快速、高效处理和生成各类文档，如宣传文案、操作手册、技术文档、施工方案等，提高工作效率和质量。



生成技术文档



作 业 指 导 书

工序名称	单槽超声波清洗	上一工序名称	开料（全自动机）			版本	A/0
		下一工序名称	CNC精成形+周边倒角	制作单位		生效日期	
适用范围		机器设备	单槽超声波清洗机	原材料	玻璃	页 码	1 OF 1

一、开机前准备：

1.1 准备物料：水、清洁剂（不含油性）、钢架、作业手套、防水手套、料（玻璃）

1.2 对玻璃机器检查：

①加热电源线和超声波电源线是否交接到该机所要求的位置上

②机身电源开关无积存水

③周围环境机体是否干洁及5S

④各开关要处于关/OFF或归零/O

1.3 参数设置：水温：50℃-70℃　　超声电流：2A

二、操作步骤：

注入清洗槽体积4/5的水及30ml的清洗剂 → 关上槽盖，插好加温插头，将温控开关（槽缸正前面的黑色旋钮，如图1所示）置于50℃至70℃标示处 → 插好超声波清洗机插头，等温度达到时，（将要达到时指示灯亮，达到后，指示灯灭）打开超声波清洗剂开关（主机上的红色开关） → 将装满玻璃的钢架轻轻地放入到专用清洗架的中央 → 将专用清洗架轻轻地放入到槽缸底 → 打开超声波开关，旋转电流表右下方的黑色旋钮，将电流调到2A处 → 20～30秒后，将超声波清洗机开关关掉 → 用两手抓紧清洗架，将其提出，放在机器上面一下水，然后取下放到50cm处的推车上 → Q C 确认是否干净，O K 后方可批量生产

三、注意事项：

3.1 在机器使用的过程中，不能把机器放置在潮湿或有腐蚀性气体、粉尘等环境下。

3.2 禁止在清洗槽内无清洗液的情况下启动加热器和超声源，以免损坏发热板和换能器。

3.3 防止硬物冲击和强烈振动。

3.4 把装有玻璃的钢架放入单槽超声波清洗机内清洗时或从单槽超声波清洗机中取出时要轻放轻取。

3.5 必须随时保持机体无积水。

3.6 生产中出现异常情况须及时通知管理员。

3.7 每班清洗一次槽（换一次水）。

3.8 从机器中取出钢架时须目检玻璃是否已被清洗干净，若没有则须再次放入机器内清洗。

制作		审核		批准	

生成作业指导书

生成设备运行状态报告

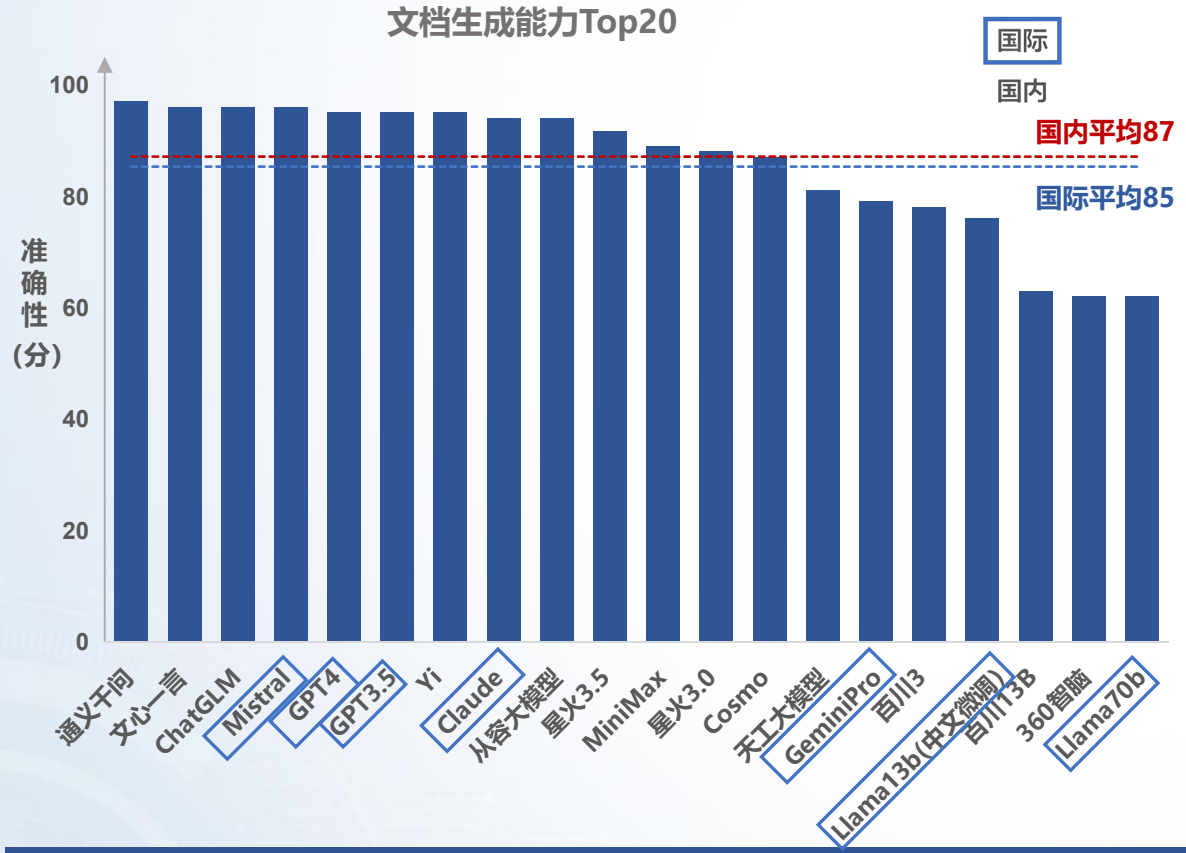
生成库存管理运营文档

应用场景研判

- **研发设计环节**：大模型可基于本地知识库，辅助工作人员生成技术方案和设计方案，帮助研发人员提升效率，为设计人员提供灵感；
- **生产制造环节**：大模型可基于生产订单和生产计划，自动生成作业指导书，提高生产效率；
- **运维管理环节**：大模型可根据设备运行情况，自动编写运行报告；可根据供应链库存情况生成库存管理报告文档，提升运营效率。

五、场景测评四: 文档生成（要点总结）

➢ 测评结果



➢ 题目样例

问题：分析以下文字，总结B公司企业创新的启示。

B公司专门成立了热效率技术攻关团队，通过大量的仿真和台架试验，经过上千种方案的探索分析，不断尝试与改进，最终把发动机各个方面的功能发挥到极致，实现了热效率突破 50%。思路决定出路，以往一些科技企业遭遇挫败是因为单纯以技术为主导按已有的技术去做产品，再去找销路，结果市场并不认可。……

评分标准：

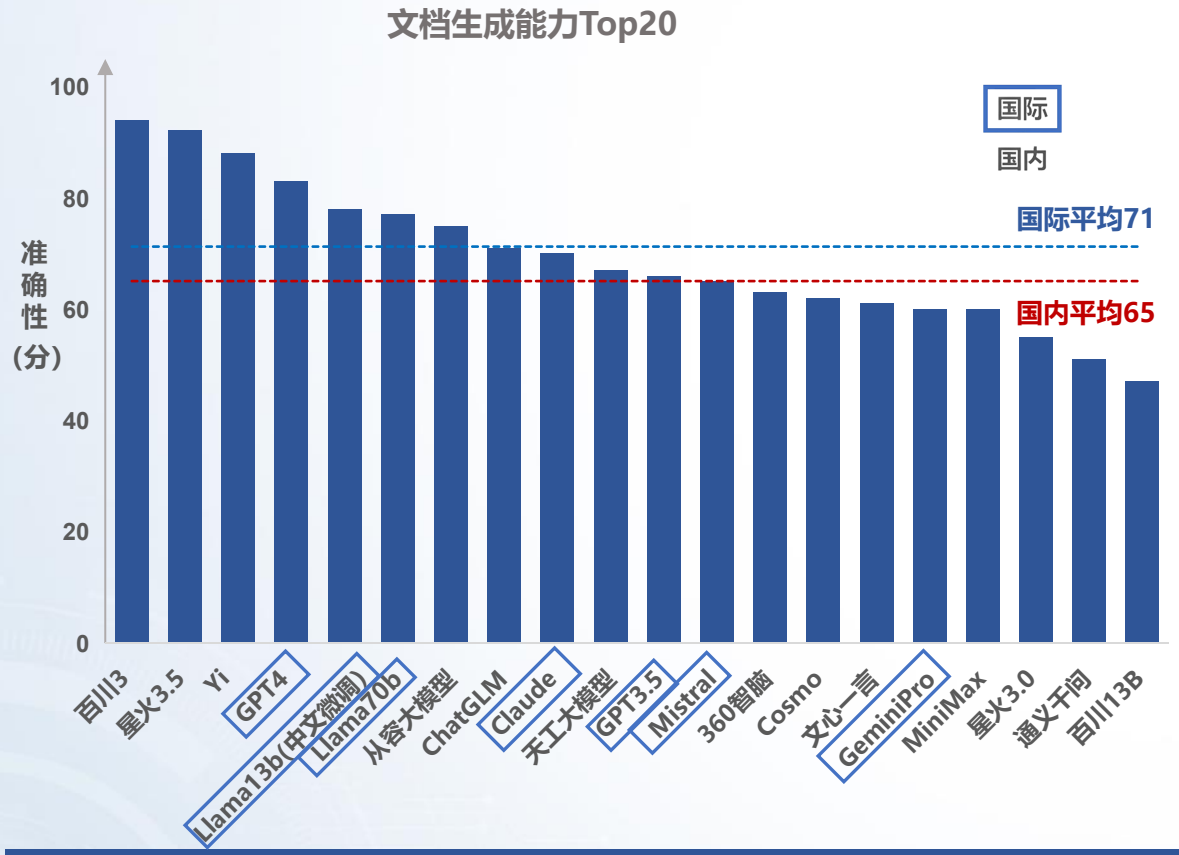
1. 敢于超前研发。树立首创精神, 敢为人先, 形成差异化竞争力，抵御风险，赢得优势。（1分，**必须有关于超前研发，敢于创新的的近似描述**，否则不得分）
2. 加强技术攻关。成立专门团队, 进行大量试验, 不解探索分析,不断尝试改进。（1分，**必须有关于技术公关，探索尝试的的近似描述**，否则不得分）
3 . 市场需求导向。转变技术指导市场的思路, 从客户需求出发,确定产品创新方向。（1分，**必须有关于市场导向，重视调研，技术指导市场的相关描述**，否则不得分）
4. … 5. …

本题共5要点，对于每个要点，如果学生回答中有和该项一致的语句，则得1分，如果描述有差距，或者详细程度不足，酌情给0.3或者0.5分或者0.8分。

- **在文档生成（要点总结）领域，国内外性能最佳大模型成绩接近满分，基本可成熟应用于该场景；**
- **国内外平均成绩相对较高，文档生成（要点总结）场景属于当前大模型较擅长领域。**

五、场景测评四: 文档生成（观点分析）

➢ 测评结果



➢ 题目样例

问题：阅读以下观点，回答你是否同意，如果你不同意，请说明哪种情况会削弱下面的观点：

过去的一年, QM的工伤事故比邻近的工厂多 30%, 邻近工厂每班工作时间比我们公司短 1 小时。专家称许多工伤事故的主要原因是疲劳和睡眠不足。因此, 为减少QM的工伤事故数量, 从而提高生产效率, 我们需要把 3 个班次的工作时间缩短 1 小时, 这样我们的员工可以获得充足的睡眠。

评分标准:

总结提炼后，评分标准如下：

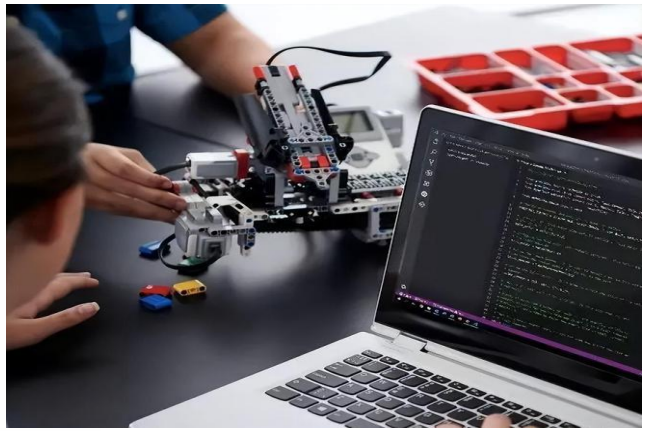
- (1) 两家公司是否具有可比性, 没有给出具体的分析...
- (2) 去年的情况今年是否依然持续...
- (3) 倒班时间缩短一小时，不能保证员工获得充足的睡眠；
- (4) 即便缩短倒班时间能够保证员工获得充足的睡眠，员工的工伤数量也并不一定会下降...
- 对于以上四点，**每个分论点在作文中有所体现得1分，共计4分。**

- **在文档生成（观点分析）领域，百川3、星火3.5、Yi优势明显，已实现对GPT4的领先；**
- **国际大模型平均超出国内较多，国内模型需整理高质量语料进行强化训练，提升观点分析成效。**

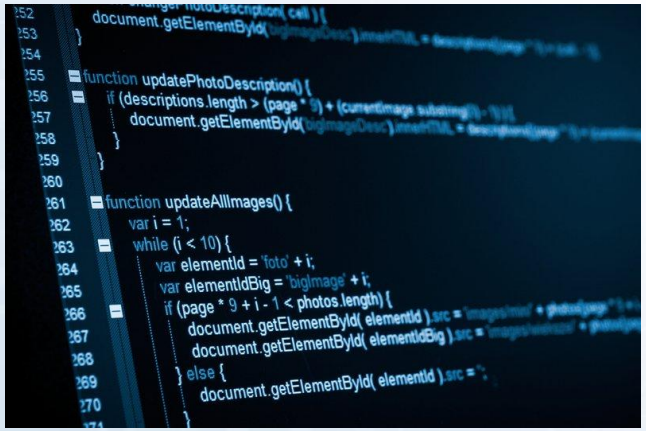
五、场景测评五: 代码理解

大模型将面向工业需求编写代码，回答计算机编程相关问题，辅助代码功能性和安全性检测，提升工程师编码效率，保障程序安全、平稳运行。

代码生成与自动编程



代码错误检测与修正



代码注释生成

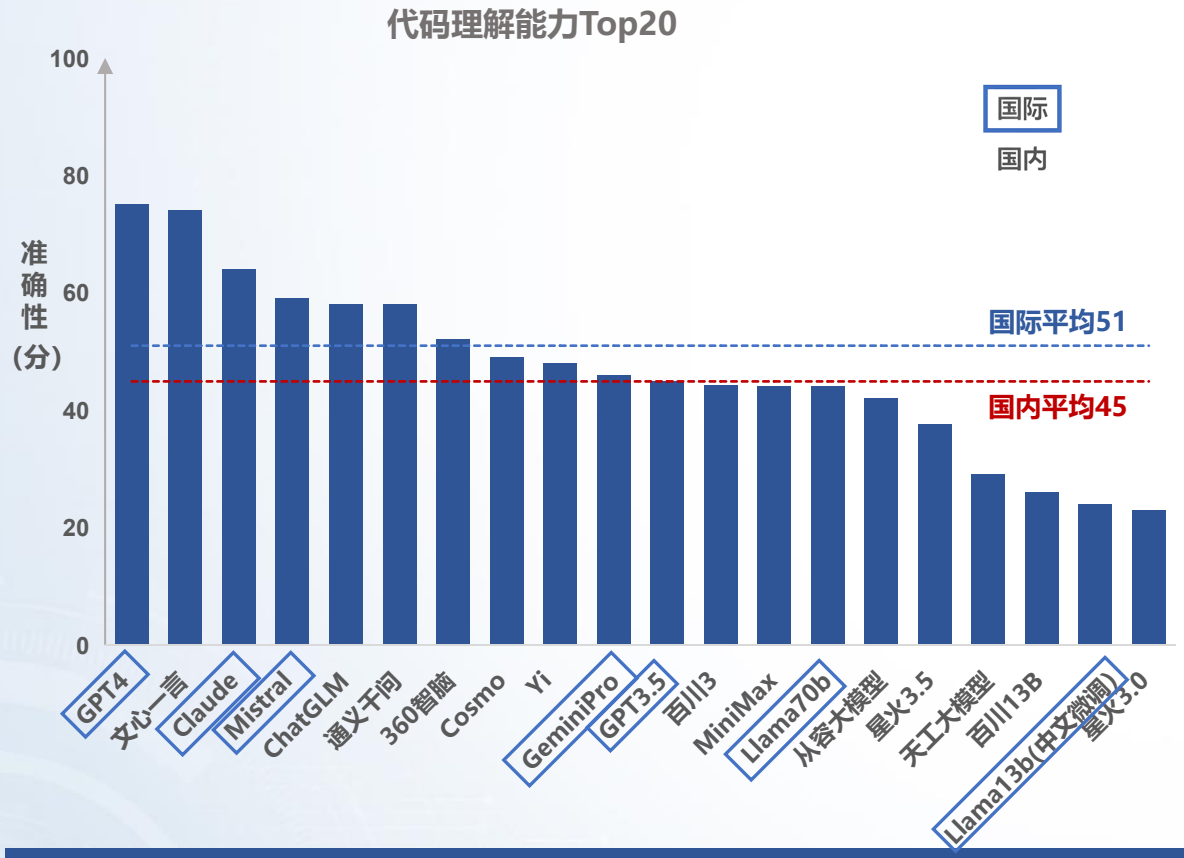
工控代码审查

应用场景研判

- **研发设计环节**：大模型可根据自然语言描述，自动生成工业代码，辅助编写自动化脚本、复杂的业务逻辑代码，提升编码效率；
- **生产制造环节**：大模型可对生产设备的控制代码进行安全审查，排查代码中的错误和漏洞，保障生产安全；
- **运维管理环节**：大模型可自动生成代码注释，帮助工控运维人员高效理解和维护代码，提升运维效率。

五、场景测评五: 代码理解

➢ 测评结果



➢ 题目样例

问题： 回答下列选择题，并给出解析。下列代码中存在什么安全问题？

```
public  void doPost(HttpServletRequest request, HttpServletResponse response)
        throws ServletException, IOException {
    javax.servlet.http.Cookie[] theCookies = request.getCookies();
...
   java.util.Properties benchmarkprops = new java.util.Properties();
    String algorithm = "MD5";
    java.security.MessageDigest.getInstance(algorithm);
    byte[] input = {(byte) '?'};
    Object inputParam = param;
    if (inputParam instanceof String) input = ((String) inputParam).getBytes();
}
```

评分标准：

代码中使用已知的**弱哈希算法MD5**，代码如下：

```
String algorithm = "MD5";
java.security.MessageDigest md =
java.security.MessageDigest.getInstance(algorithm);
```

弱哈希算法有MD5、SHA-1 和 SHA-2 等哈希函数。

（回答中如果能指出安全问题是弱哈希算法得1分，否则不得分）

- **在代码理解领域，GPT4和文心一言准确度较高，相对其他模型优势明显；**
- **国内外平均成绩相对偏低，编程相关知识掌握薄弱，应全面提升模型训练集中代码语料的数量和质量，代码解释器模块可能对理解代码的能力有较大帮助，建议更多大模型引入。**

六、总体评价与后续规划

➢ 各场景第一梯队与点评

	第一梯队			点评
知识问答	智谱·AI ChatGLM	Baidu 百度 文心一言	卡奥斯 COSMOPlat 卡奥斯	• 国内大模型已具备一定优势，ChatGLM、文心一言等多个大模型已超越GPT4；
工程建模	OpenAI GPT4	Baidu 百度 文心一言	智谱·AI ChatGLM	• GPT4处于领先地位，大模型整体建模能力处于较低水平，有较大提升空间；
数据分析	Baidu 百度 文心一言	OpenAI GPT4	MISTRAL MISTRAL	• 文心一言能力最佳，与GPT4、Mistral构成第一梯队，但整体水平偏弱；
文档生成 要点总结	Alibaba.com 通义千问	Baidu 百度 文心一言	MISTRAL MISTRAL	• 国内大模型保持领先，性能最佳大模型已经能够较完善地完成文本总结任务；
文档生成 观点分析	百川智能 BAICHUAN AI 百川3	科大讯飞 iFLYTEK 星火3.5	Open-source 开源模型 Yi	• 大模型在观点分析上处于及格水平，还存在明显提升空间；
代码理解	OpenAI GPT4	Baidu 百度 文心一言	Claude	• GPT4和文心一言在代码理解领域较为领先，具有一定优势，其他大模型仍有较大提升空间。

六、总体评价与后续规划

➢ 总体评价

场景成熟度

- 大模型在文档生成领域应用成熟度较高，在工业知识问答、数据分析、工程建模、代码理解场景应用成熟度相对较低；
- 国内外大模型在文档生成、数据分析、代码理解场景准确度差异较大。

行业知识掌握

- 大模型在钢铁、电力等行业有较好的知识储备，对于纺织、装备制造等行业仍需加强训练；
- 国内大模型在建材、采矿等行业优势显著，在装备制造、化工等行业与国际接近。

发展趋势

- GPT4仍处于领先地位；
- 近半年国内大模型能力显著提升，与GPT4差距不断缩小，部分场景应用能力已赶超。

➢ 后续计划

针对工业应用场景，汇聚整理工业知识语料库，支持大模型预训练或微调；

开展大模型多模态能力测评，包括图像识别、视频理解等，挖掘更多大模型工业潜在应用场景；

面向大模型当前应用成熟度较低的场景，提供稳定性、准确性等能力优化指导；

面向工业应用开展行业大模型测评工作，在重点领域遴选推广一批优秀的行业大模型。

中国工业互联网研究院
China Academy of Industrial Internet

做让人尊重的奋斗者

心怀大局　创新笃行　专业精微　团结清廉

联系人：叶老师 13661350566 邱老师 18823660419

附录1：报告涉及的大模型及其版本号

编号	大模型	公司	版本号
1	GPT4	OpenAI	GPT4-Preview-1104
2	GPT3.5	OpenAI	GPT-3.5-turbo
3	文心一言	百度	Ernie-bot-4.0
4	星火大模型	科大讯飞	spark-V3.5；V3
5	Yi	零一万物	Yi-34B
6	GeminiPro	Google	Gemini-Pro
7	通义千问	阿里巴巴	Qwen-Max
8	360智脑	360	360GPT_S2_V9
9	ChatGLM	智谱华章	GLM-4
10	Claude	Anthropic	Claude-2.1
11	llama	Meta	llama-70B; (开源) llama-13B-中文微调(开源)
12	Mistral	Mistral	Mistral-Medium
13	从容大模型	云从科技	20240104版
14	天工大模型	昆仑万维	20240112版
15	MiniMax	MiniMax	ChatCompletion-abab5.5-chat
16	Cosmo	卡奥斯	20240124版
17	Baichuan	百川智能	Baichuan-3;Baichuan-13b (开源)

注：本研究实测模型包括但不限于上述大模型，此处只列举部分模型版本号。

附录2：评分规则

➢ 提升问答题比例的原因

选择型题目的局限性

- **随机给出答案**：部分模型随机给出答案，即使选择正确也无法证明模型能够给出准确的解答过程。
- **过程评价缺失**：有的大模型选择虽然错误，但能够提供建设性的思路，有一定的参考意义。

问答型题目优势和问题

- **优势**：问答题更贴近实际，对回答步骤判分更加科学、合理。
- **问题**：对比选择判断类题目，传统人工判分的方法效率低。

问答型题目判分问题解法

- **评分标准保障一致性**：依据标准回答，生成判分标准，提升判分准确度，保障一致性。
- **大模型提升判分效率**：用逻辑性好的大模型进行判分，在确保判分准确性前提下提升判分效率。

➢ 国内大模型发展趋势统计规则

1. 以GPT3.5为基准，依据在相同测评的相对成绩，计算发展趋势

$$\text{相对GPT3.5成绩} = \frac{\text{某模型测评分数}}{\text{GPT3.5测评分数}} \times 100\ \%$$

➢ 问答题评分步骤

生成标准答案 → 输入问题获取回答 → 按照标准答案和步骤赋分 → 分维度进行分数统计分析