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

人工智能大模型
工业应用准确性测评

2024年3月版

一、前言

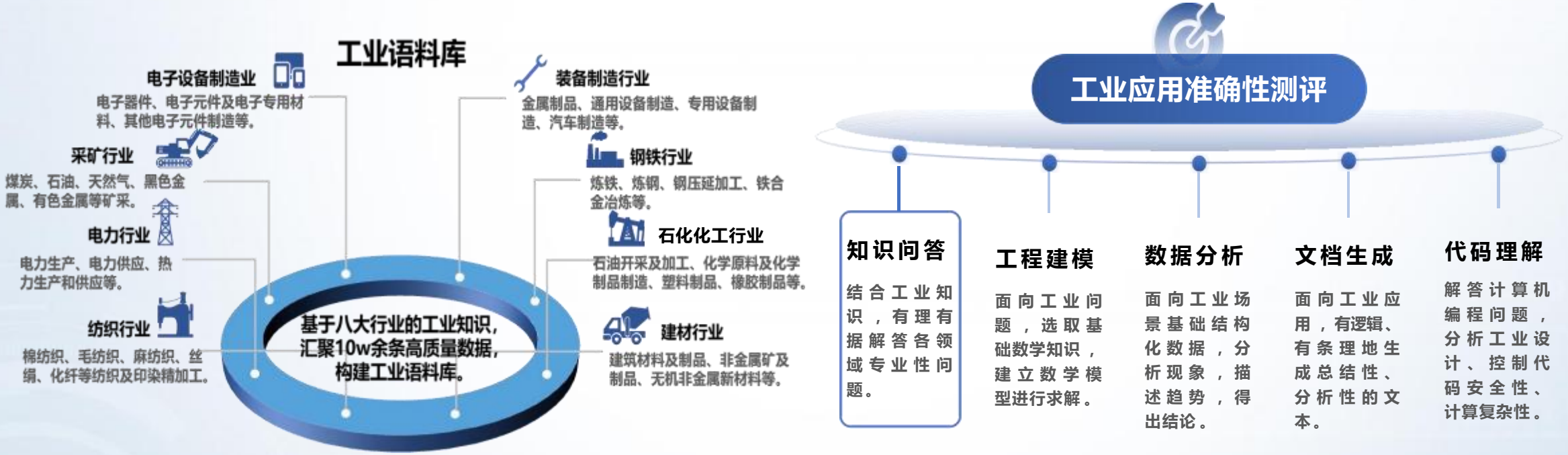
为贯彻落实党中央国务院关于促进人工智能发展的决策部署，中国工业互联网研究院依托通用人工智能与工业融合创新中心（简称“中心”），联合香港科技大学、中国经济信息社，深入研究人工智能大模型在工业领域的应用性能、技术架构、标准体系，并在此基础上，形成本报告。

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

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

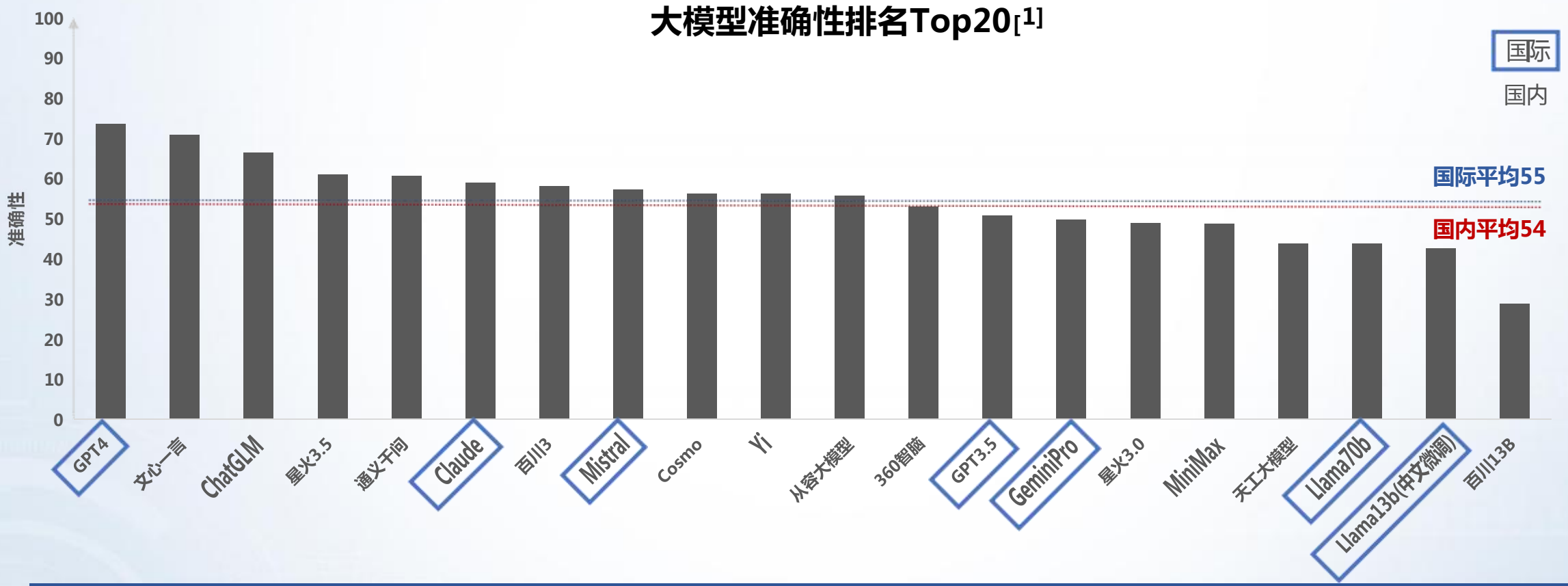
   注：各场景题目数量虽不一致，但考察要点总量保持在同一个数量级。
3. **题目得分：**需要结合具体题目的评分细则，按照步骤进行赋分，赋分后分数进行归一化处理。
4. **场景得分：**
   - 场景得分为题目总分百分化处理后的分数。
   - 若有细分场景，则场景总分为细分场景的平均成绩。
5. **综合评分：**由各场景算数平均分计算得出。

- **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

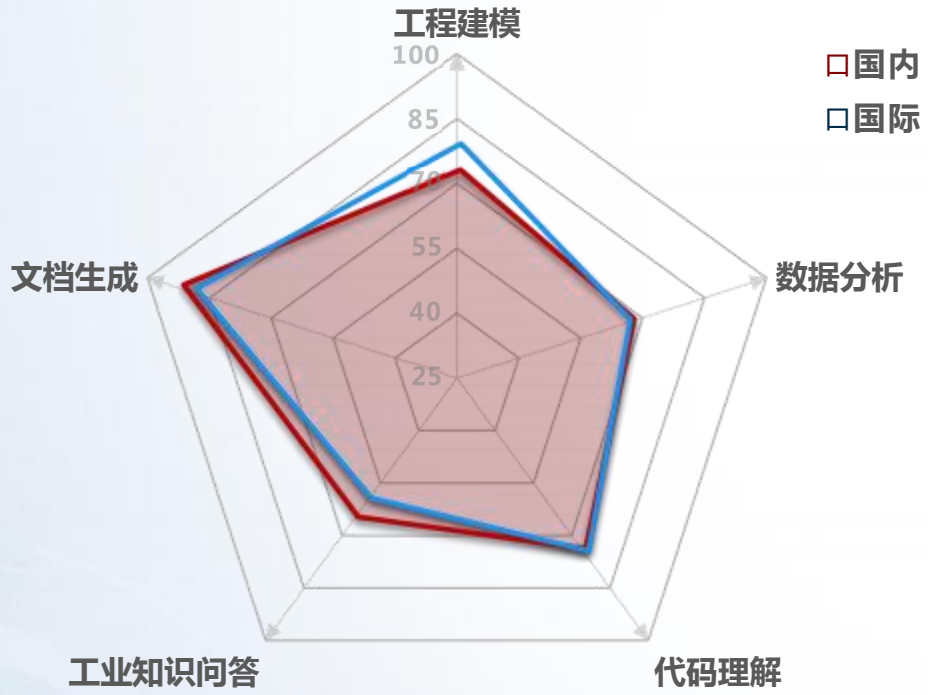


- 综合能力上，GPT4处于领先地位，国内大模型文心一言、ChatGLM紧随其后；
- 对于国内大模型，多个模型综合能力超过GPT3.5，包括文心一言、ChatGLM、星火3.5、通义千问等；
- 对于国外大模型，GPT4领先优势明显，其余模型差距较大。

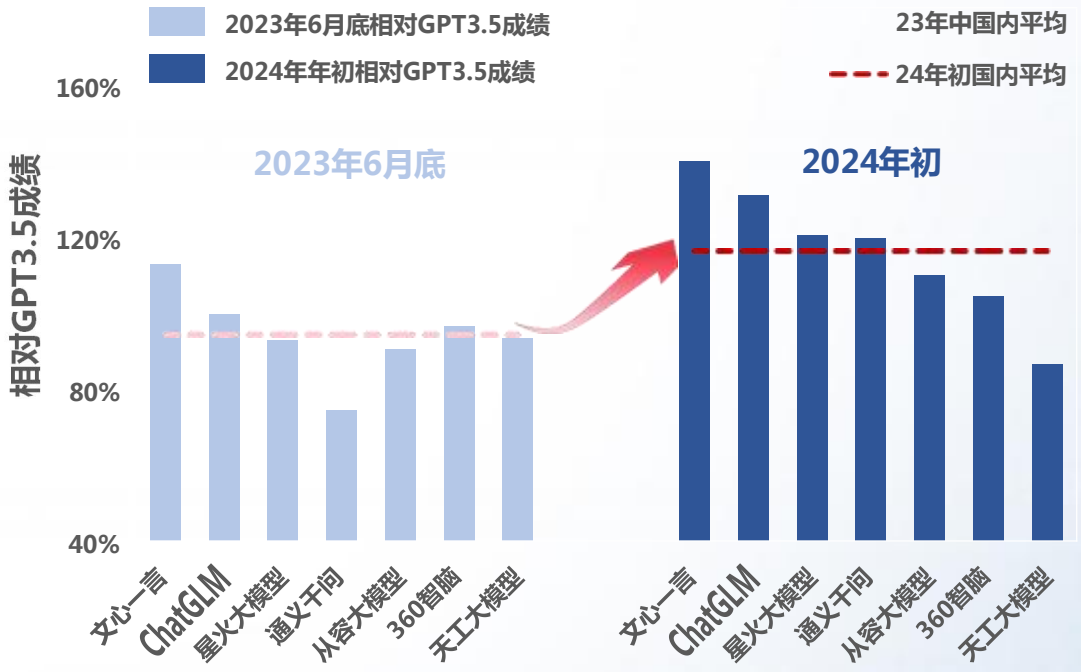
[1] 模型版本号参见附录1。

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。