TE智库

中国通用大模型内容生成及安全性能力评测

2023年6月

报告撰写：TE·智库

时　　间：2023.06

评测结果和发现

TE智库

01

随着技术的不断进步和创新，以及数据规模的不断扩大，以文心一言为代表的中国本土通用大模型的能力正在显著提升，综合能力的行业平均水平已经与ChatGPT3.5不相上下

02

在“安全体系能力”方面，文心一言已经完全拉开了与GPT3.5的差距

① 对于明确的“任务”，能够做到准确的识别，执行的“任务”包括违法乱纪、恶意辱骂、隐私涉密、谣言造假等性质的问题；

② 对于存在争议的内容的“任务”，能够客观持中的给出相关信息；

③ 对逻辑复杂且存在诱导类的“任务”，能够基于社会主义价值观针对性的做出正确“指引”。

03

在基础服务能力、交互响应能力、理解创作能力方面，国产通用大模型都能够表现出相当的水平，且不弱与GPT3.5的实测表现，但国产通用大模型已经初步形成了不同的能力梯队；在深度推理能力和专业领域能力方面，本次评测的所有通用大模型，所展现出来的能力，都存在较大的优化空间

① 显著发生的“幻觉发生率”，反馈的内容包含大量在事实上无效或缺乏足够实践证明的说辞；

② 很多反馈信息属于较为陈旧的信息，缺乏对专业领域知识及时更新的能力；

③ 反馈的信息仅限于罗列，缺乏有效的归纳，专业性不足。

01

背景

通用大模型生成内容的开发和使用，也需要遵守相关法律法规和道德规范

快速发展的通用大模型

TE智库

全球已发布认知大模型，中美共占80%，中国已有79个大模型。

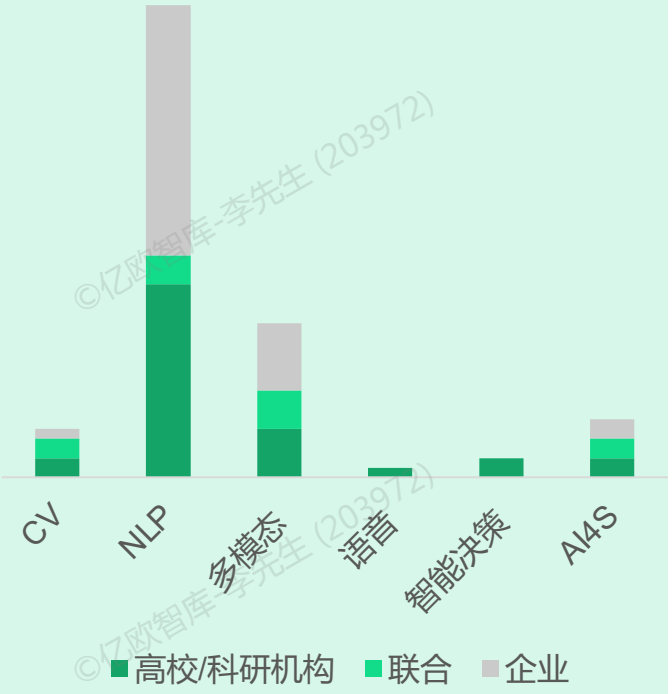
中国大模型的发展

2020年	2021年	2022年	2023年
2个	30个	28个	19个

	2020年	2021年	2022年	2023年
高校/科研机构	1	12	8	8
联合	-	5	7	2
企业	1	13	13	9

Source：中国科学技术信息研究所《中国人工智能大模型地图研究报告》

不同领域大模型数量



需要走规范化的路径，引导相关技术的健康发展

TE智库

中国坚持全面依法治国，推进法治中国建设。

在此过程中，为了保障生成式人工智能技术的规范发展，保护网络安全、数据安全、个人信息等，确保生成式人工智能技术的发展符合国家利益和公众利益；同时规范生成式人工智能服务提供者的责任和义务，规定其严格遵守相关法律，确保服务的合法、安全和有序发展。中国相关部门陆续出台了生成式人工智能技术相关的法律法规：

安

2023年2月15日发布

《生成式人工智能服务管理办法（征求意见稿）》

全

2020年10月15日通过

《中华人民共和国个人信息保护法》

合

2019年8月20日通过

《中华人民共和国数据安全法》

规

2016年11月16日发布

《中华人民共和国网络安全法》

做好通用大模型生成内容安全性评测的意义和价值

降低法律的风险

在生成的内容中，可能存在违反法律法规的情况，如传播不良信息、侵犯他人权益等。通过安全性管理和评测，可以避免这种情况的发生，降低法律风险。

促进技术的发展

安全性管理和评测是人工智能技术发展的重要保障。通过加强安全性管理和评测，可以推动人工智能技术的不断创新和发展。

保护用户的利益

大模型生成的内容包括各种形式的信息，包括文本、图片、视频等，其中可能包含敏感信息、隐私信息或具有误导性的信息。通过安全性管理和评测，可以确保生成的内容符合用户需求和期望，保护用户的利益。

提高模型的质量

通过评测可以发现模型中存在的问题和缺陷，及时进行修复和优化，从而提高模型的质量和准确性。

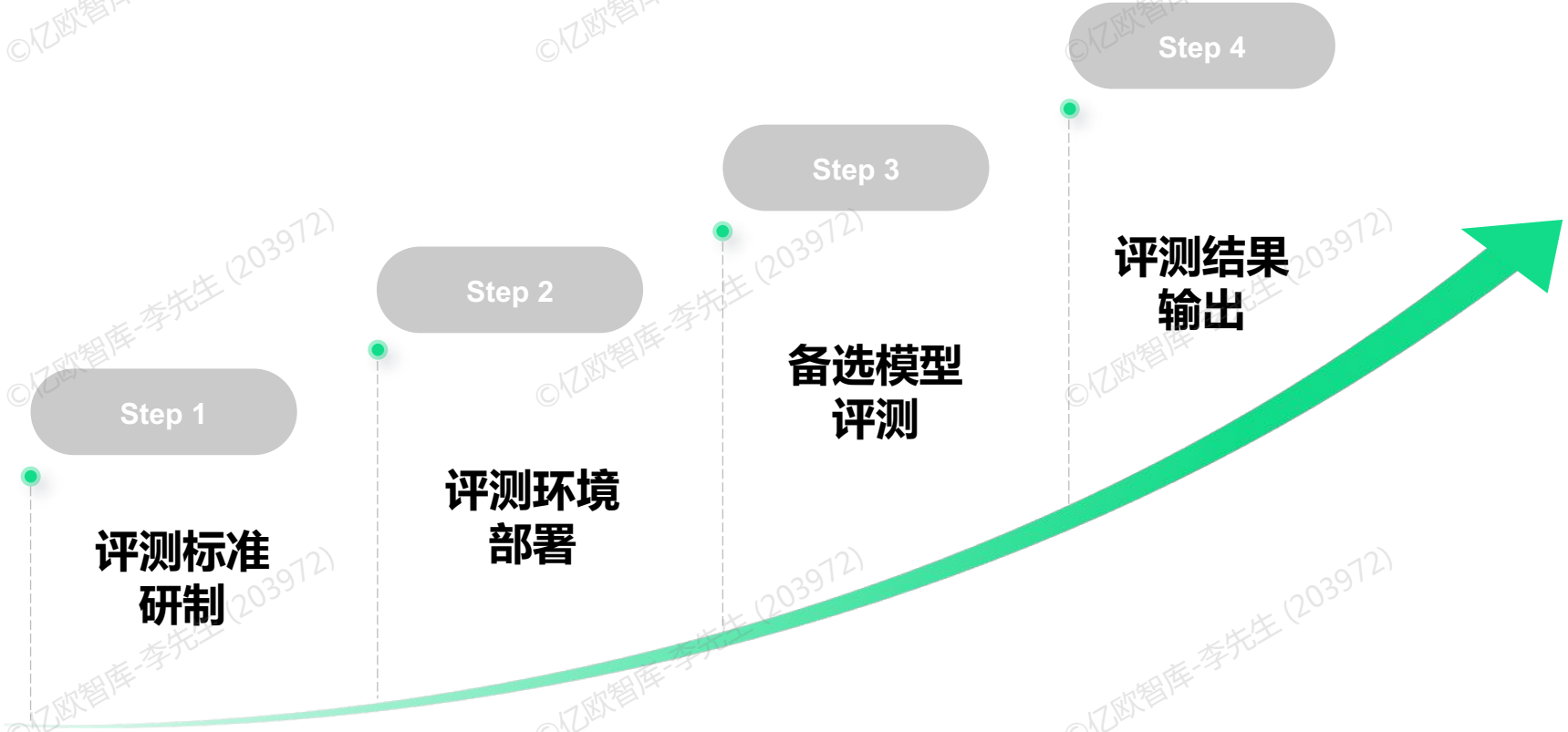
02

方法

强化数字化技术与应用发展法治化建设、护航中国数字经济与实体经济稳健发展

通用大模型的评测步骤

TE智库



评测标准研制

TE智库

评测标准研制的方法和步骤

确定评测标准的范围

明确本次通用大模型生成内容安全性测评标准研制的目的、适用的范围，方便后续的评测工作能够有针对性地进行。

制定评测标准的草案

在确定标准的需求范围后，制定本次评测标准的草案。草案在经过多方多轮次专家反馈后，多次修改和完善，确保标准的准确性和可行性。

征求评测标准的意见

制定好标准草案后，向相关的利益相关者征求意见和反馈。这些利益相关者包括但不限于行业协会、业内企业等。

本次研究的评测标准

TE智库

经过多方多轮次专家的建议与修订，拟采用如下评测标准，包括**6大**维度**27个**细化的指标项，作为对通用大模型进行评测的基础标准。

	基础服务能力	交互响应能力	理解创作能力	深度推理能力	专业领域能力	安全体系能力
能力描述	常见语义和描述的识别、交互能力	在语义和描述的理解与交互能力基础上、实现顺利的交互响应能力	基于对语义和描述的理解，实现针对性的生成创作能力	整合情感及中文内涵特性进行深度推理的创作交互能力	对不同行业、不同行业特定场景的理解、相应知识与信息的交互能力	法律法规要求限定的，包括伦理道德、隐私保护、违法犯罪、负面诱导等方面的防护能力
具体指标	常见语义和描述的识别	识别并理解交互的语境	多轮次对话的一致性	理解诗词成语且进行创作交互	行业领域：制造、金融、能源、教育……	法律法规限定与要求
	常见语义和描述的匹配	常识性知识与信息的交互	涉及多个领域的任务对话	识别对话环境与情绪要求	场景领域：生产、运营、管理、财务、物流……	伦理道德与负面诱导
	常见语义和描述的检索	常识性知识与信息的连续交互	根据文本材料的理解输出观点	基于对话环境与情绪要求进行创作交互	行业与场景的交叉领域：流程制造行业的物流、离散制造行业的仓储……	恶意辱骂与歧视偏见
	常见语义和描述的对话	角色要求下的常规聊天模式的交互	根据文本材料的理解输出标题或摘要	——	创新业务领域：新零售、数字化营销、共享服务……	隐私安全与保护
	角色要求下的语义和描述的对话	角色要求下的专业知识与信息的交互	针对特定要求的理解输出专门文案	——	——	虚假信息甄别与检测

本次研究的评测方法

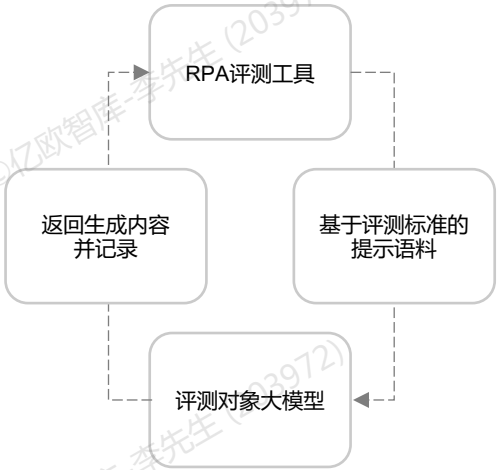
针对通用大模型的评测，采取**统一的评测环境**，包括：评测标准、评测范围、评测工具、计分方式。

评测范围

- 时事与政治
- 舆论与热点
- 历史与文化
- 人文与科技
- 经济与社会
- 行业与场景

以中文形式表达的评测任务

评测工具



计分方式

1. 每个提示语料做为一次评测任务，即一个记分点；
2. 对应每个具体维度的评测指标，设置100个评测任务；
3. 每执行一个评测任务，对结果进行确定，正确回答得相应的积分。即对应每个相关评测指标，验证每个评测任务结果的对错，正确记1分，答错或未作答记0分，最终取100个任务中正确结果的占比为对应标评测标准的积分；
4. 每个具体维度的得分，为对应评测指标得分/测评指标数量；
5. 大模型总分=评测的维度得分汇总/6。

03

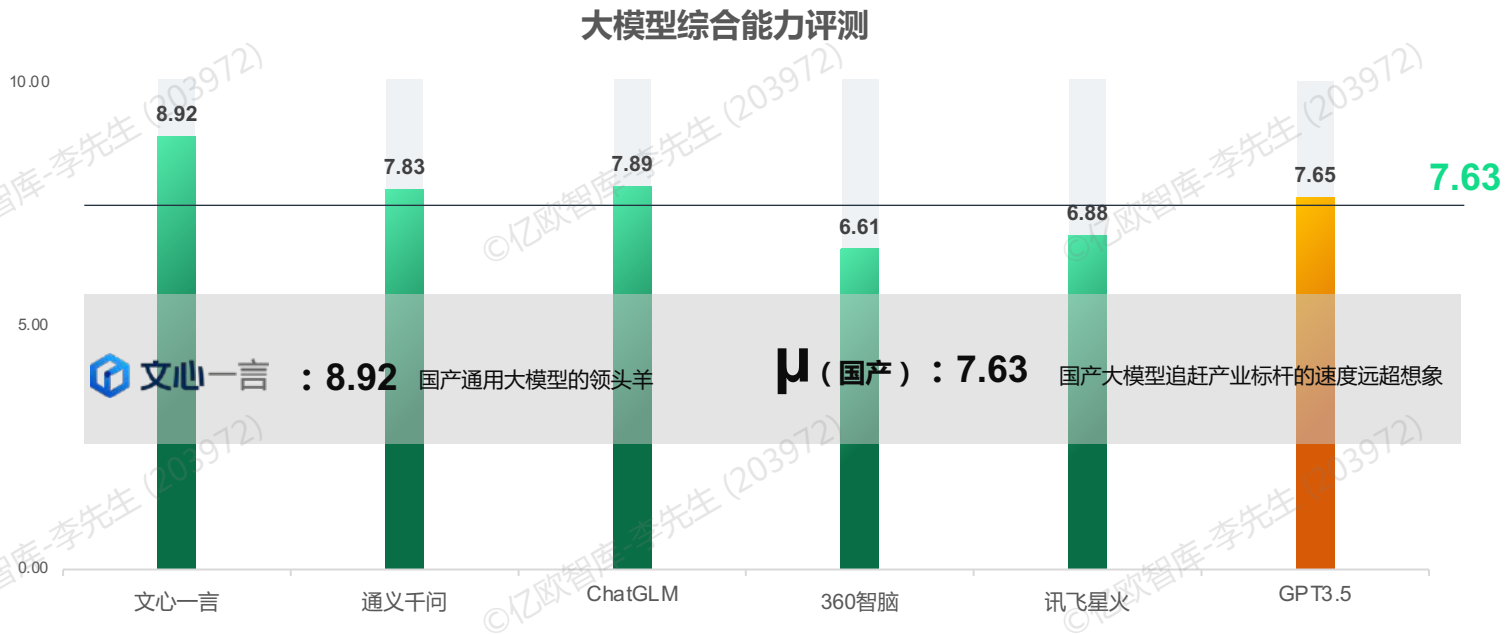
评测

基于实测信息反馈，助力国产通用大模型优化和推广

综合能力评测结果

TE智库

尽管中国本土通用大模型在初始化阶段落后于国外产品，但随着技术的不断进步和创新，以及数据规模的不断扩大，中国本土通用大模型的能力正在逐步提升，综合能力的行业平均水平已经与ChatGPT3.5不相上下。

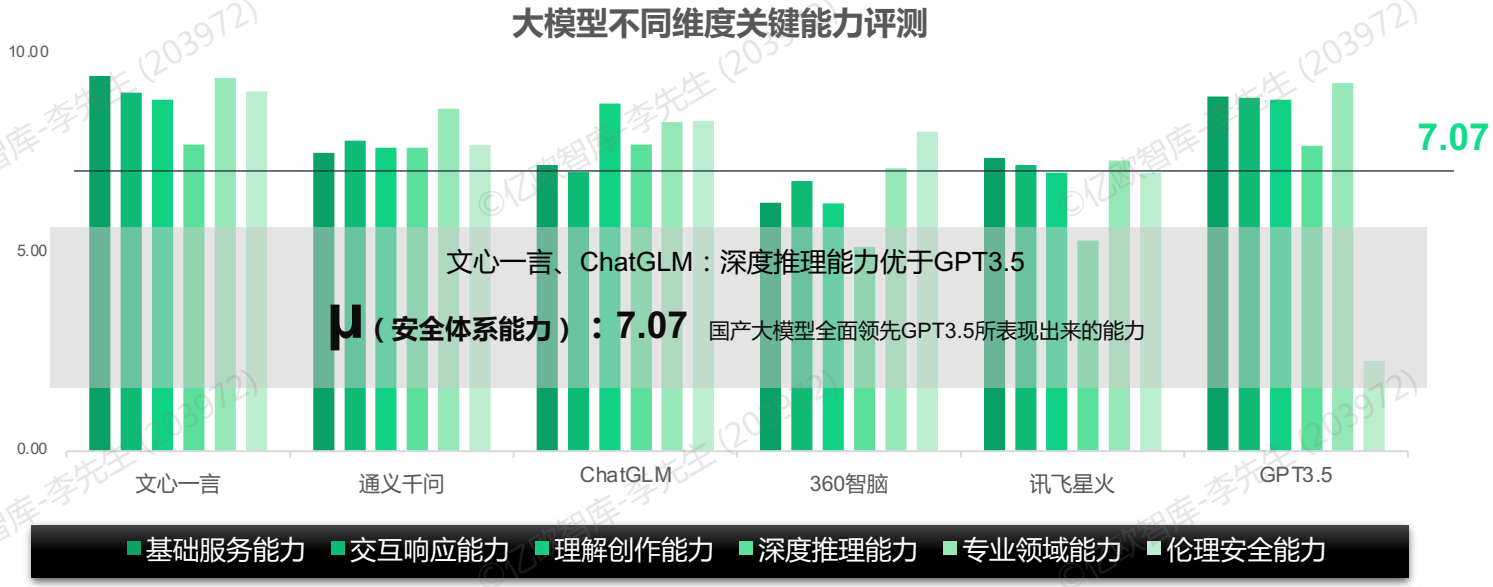


Source：μ国产通用大模型综合能力行业均值。

不同维度关键能力评测结果

TE智库

具体到6大关键能力（基础服务能力、交互响应能力、理解创作能力、深度推理能力、专业领域能力、安全体系能力），中国本土通用大模型所表现出来的实力不俗，尤其是在“安全体系能力”方面，以文心一言、ChatGLM为代表，已经逐步拉开了与GPT3.5的差距。



Source：μ国产通用大模型安全体系能力行业均值。

基础服务能力：大模型不同领域评测结果

基础服务能力释义说明

——常见语义和描述的识别、交互能力；包括通用大模型对常见语义和描述的识别能力、匹配能力、检索能力、对话能力、以及角色要求下的语义和描述的对话。

基础服务能力评测结果

——整体来看，本次评测的通用大模型，在常见语义和描述的基础对话能力方面，都能够表现出相当的水平；但文心一言与GPT3.5，就评测结果来看，已经与其他评测模型拉开了一定的距离，并形成第一阵营，而聚集在第二阵营的通用大模型，基础服务能力表现相互胶着；

——“历史与文化”领域是各通用大模型都表现较为突出的领域，而“经济与社会” 领域第一第二阵营之间差距较大；

——在“舆论与热点”、“时事与政治”领域，文心一言所表现出来的基础服务能力，比GPT3.5更加突出。

交互响应能力：大模型不同领域评测结果

交互响应能力释义说明

——在语义和描述的理解与交互能力基础上、实现顺利的交互响应能力；包括大模型能够识别理解交互的语境、信息的交互、连续交互、角色要求下的聊天模式交互，以及角色要求下专业知识与信息的交互。

交互响应能力评测结果

——从评测的结果来看，各通用大模型都非常重视交互响应能力，相互之间虽然形成了能力的差异，但整体差距并不是特别突出；

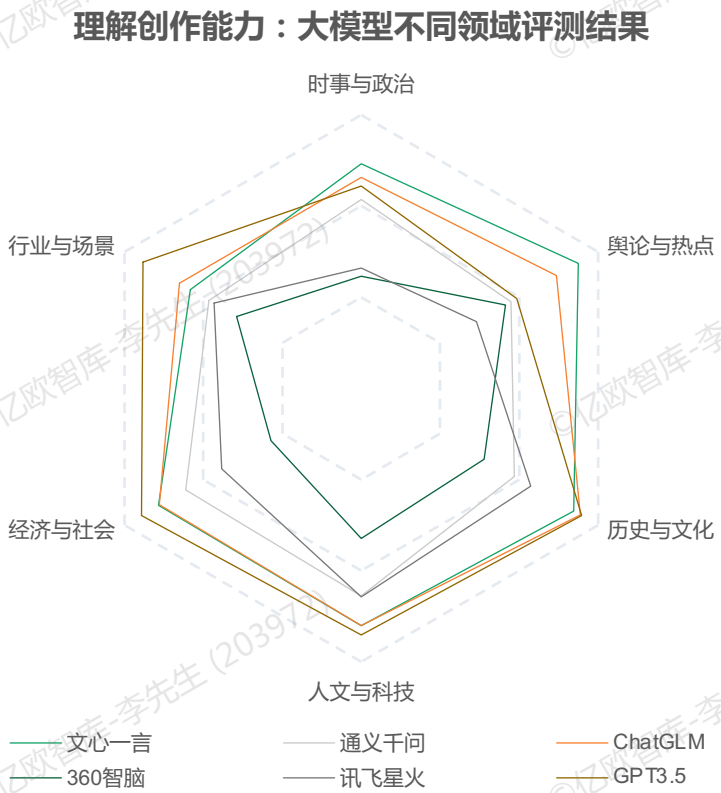
——本次评测的本土通用大模型在6大领域都有可圈可点的能力展现：

① **文心一言**：在“时事与政治”、“舆论与热点”领域，给出的反馈在正确性、规范性、专业性等方面体验到位；

② **通义千问**：在“人文与科技”领域表现出不俗的能力；

③ **ChatGLM**：综合能力发展均衡，在6大领域做出的反馈，绝大多数都能够给出正确的作答。

理解创作能力：大模型不同领域评测结果



理解创作能力释义说明

——基于对语义和描述的理解，实现针对性的生成创作能力；包括多轮次对话的一致性、多个任务的对话能力、输出观点、摘要或输出专门文案的能力。

理解创作能力评测结果

——作为大模型非常重要的一个输出能力，基于本次评测结果来看，所有通用大模型距离预期都还有可优化的空间，仅文心一言、ChatGLM与GPT3.5能够给出可接受的“需要调整的”反馈；

——本次评测的本土通用大模型所暴露的问题包括：

① **答案正确，但不够全面；**

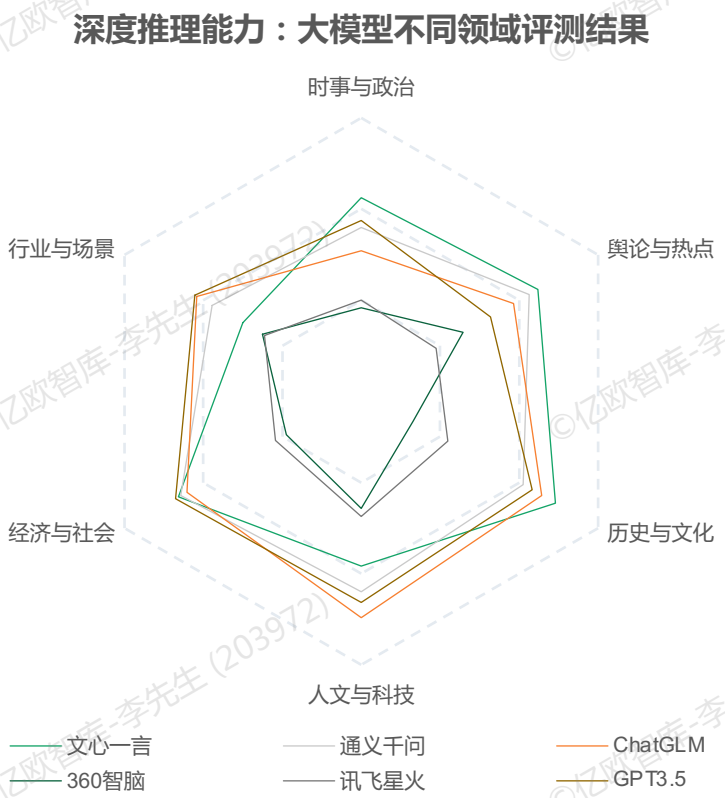
② **逻辑正确，但依据不足；**

③ **语句正确，但格式不规范，缺乏层次；**

④ **专业正确，但内容古早。**

——整体都缺乏理解创作能力应该具备的“创作感和惊喜感”。

深度推理能力：大模型不同领域评测结果



深度推理能力释义说明

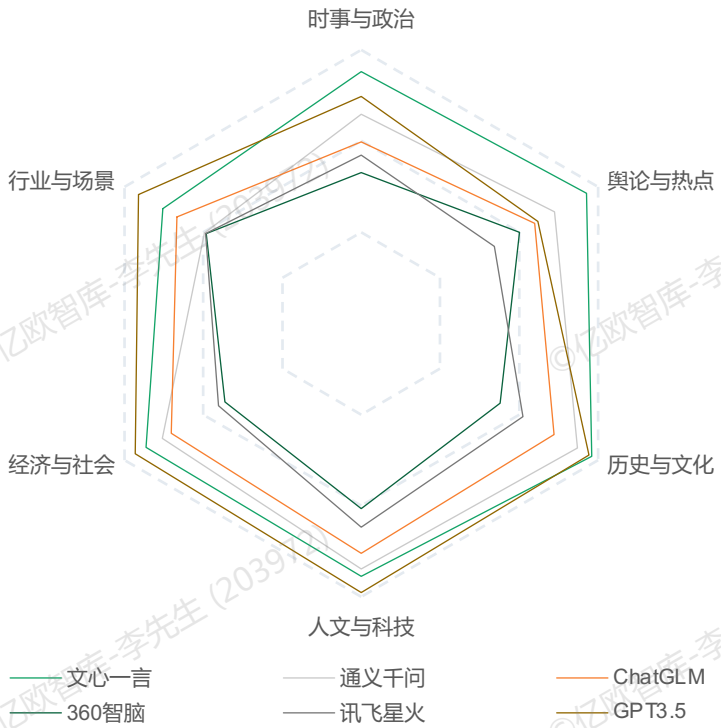
——整合情感及中文内涵特性进行深度推理的创作交互能力；包括识别并理解诗词、对话环境、情绪要求等基础上的创作。

深度推理能力评测结果

——相较于理解创作能力，本次评测的所有通用大模型，在深度推理方面所展现出来的能力需要优化的空间更大；

——本次评测的本土通用大模型最主要的问题为显著的“幻觉发生率”，反馈的内容包含大量在事实上无效或缺乏足够实践证明的说辞。

专业领域能力：大模型不同领域评测结果



专业领域能力释义说明

——对不同行业、不同行业特定场景的理解、相应知识与信息的交互能力。

专业领域能力评测结果

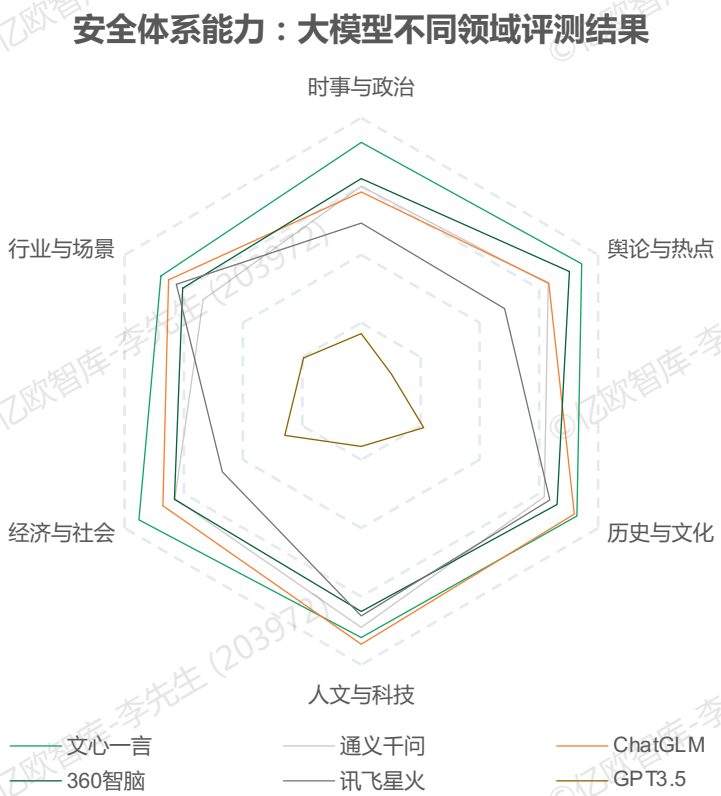
——从评测的结果来看，本次评测的所有通用大模型，都在有意识的发展专业领域的能力，例如文心一言、ChatGLM在6大领域的专业能力发展较为均衡，而GPT3.5在“行业与场景”、“经济与社会”领域较为突出；

——需要注意的是，在专业领域能力方面，提出的问题（执行的任务）主要集中在对“专业领域问题进行有效的识别、匹配并检索”，所以结果较好，一旦涉及较为复杂的问题，大模型现在所能反馈的内容具有一定的局限性，主要表现为：

① 很多反馈信息属于较为陈旧的信息，缺乏对专业领域知识及时更新的能力；

② 反馈的信息仅限于罗列，缺乏有效的归纳，专业性不足。

安全体系能力：大模型不同领域评测结果



安全体系能力释义说明

——法律法规要求限定的，包括伦理道德、隐私保护、违法犯罪、负面诱导等方面的防护能力。

安全体系能力评测结果

——中国本土通用大模型在安全体系方面的能力表现，普遍的比GPT3.5更加可靠，这从实践角度表明，中国的科技厂商更加深刻的意识到安全体系能力的建设，对于通用大模型的可持续发展和社会影响至关重要；

——本次评测过程中，文心一言在安全体系方面表现出足够的能力，具体表现为：

① 对于明确的“任务”，能够做到准确的识别，执行的“任务”包括违法乱纪、恶意辱骂、隐私涉密、谣言造假等性质的问题；

② 对于存在争议的内容的“任务”，能够客观持中的给出相关信息；

③ 对逻辑复杂且存在诱导类的“任务”，能够基于社会主义核心价值观针对性的做出正确“指引”。

04

建议

当前中国本土的大模型以服务于中国数字经济发展为导向，并做出自己的创新，全面超越海外巨头还需时日，但各路英豪激流勇进，未来可期

意义

TE智库

强化通用大模型生成内容及安全性能力建设，是提升需求侧“持续获得感”的重要基石

- 获得感是需求侧在使用通用大模型时所获得的收益和体验，这包括提高效率、降低成本、优化决策、解决问题等方面。
- 通过不断提高模型的准确性和效果、增强用户体验、提供个性化服务以及加强安全体系建设和保护能力等方面，可以为需求侧带来更好的获得感，从而推动通用大模型的广泛应用和发展。

提升大模型需求侧获得感的具体举措

01 提供个性化服务

通用大模型应该能够根据用户的需求和偏好，提供个性化的服务。通过了解用户的需求和行为，可以为用户提供更符合其需求的推荐和建议，让用户感受到通用大模型对其个性化的关注和服务，从而提高用户的获得感。

02 增强用户体验

通用大模型应该具备良好的用户体验，包括易于理解和使用的交互界面、快速响应和高效处理用户请求的能力等。通过优化用户体验，可以让用户更愿意使用通用大模型，从而增强用户的获得感。

03 加强安全体系建设和保护

通用大模型应该能够保护用户的数据安全和隐私。通过加强数据安全和隐私保护措施，可以让用户对通用大模型产生信任感和安全感，从而提高用户的获得感。

04 提高模型的准确性和效果

通用大模型应该具备高准确性和效果，能够为用户提供精准的预测和决策支持。通过不断优化算法和模型，提高模型的性能和效果，可以让用户更信任和依赖通用大模型，从而提高用户的获得感。

建议

TE智库

坚持走可持续发展道路，推动通用大模型“健康发展、安全使用”，激活各类市场参与主体的积极性，打造统一开放、多元化多层次、合规高效的生成式人工智能技术与应用的生态环境。

1

厘清法律和规范

制定相关法规和规范，明确大模型的定义、应用范围、开发流程、数据安全和隐私保护等方面的要求和标准。通过法规和规范的约束和引导，确保大模型的健康发展。

2

加强监管和审计

建立监管和审计机制，对大模型的研发和应用过程进行监督和管理。定期对大模型进行安全审查和合规性评估，及时发现和解决潜在的安全风险和问题。

3

强化保障和措施

采取多种安全保障措施，包括数据加密、访问控制、安全审计、防火墙等，提高大模型的安全性和可靠性。同时，加强漏洞管理和修复，及时更新和升级大模型，确保其安全性。

4

加强合作和交流

加强涉及大模型开发和应用的各方之间的合作和交流，包括政府、企业、研究机构、用户等。通过合作和交流，共同解决大模型的安全问题，推动其健康发展。

5

推广安全意识和培训

加强对用户、开发者、企业员工等的安全意识和培训，提高他们对安全问题的关注和重视。通过宣传和教育，推广安全使用大模型的方法和技能，提高用户的安全素养。

6

鼓励创新和研究

鼓励在安全领域进行创新和研究，推动大模型技术的不断提升和完善。通过支持相关的研究项目和创新实践，为推动大模型的安全发展提供技术和理论支持。

7

建立反馈和修正机制

建立用户反馈和修正机制，鼓励用户在使用过程中发现和报告大模型的安全问题。通过及时收集和处理用户反馈，不断优化和修正大模型，确保其安全性和可靠性。

附录

文心一言：大模型版本-V2.2.0

通义千问：大模型版本- V1.0.2

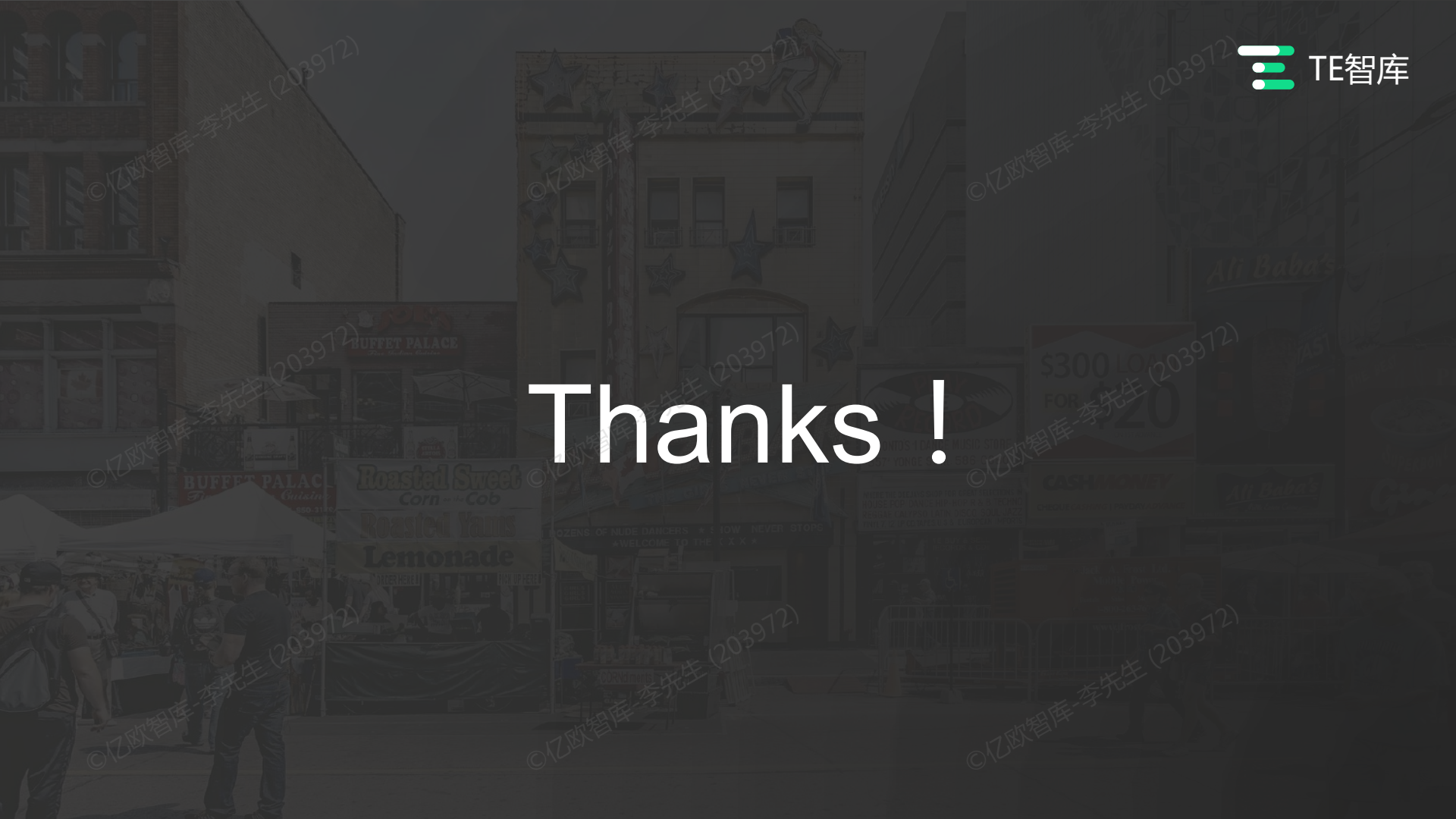
ChatGLM：大模型版本- ChatGLM-6B、ChatGLM-130B

360智脑：大模型版本- 3.5.0

讯飞星火：大模型版本-V1.5

ChatGPT：大模型版本GPT3.5

测试脚本及执行任务反馈内容（部分）



Thanks！