关于政务大模型（政务咨询场景）

第二次阶段性测评情况的报告

朱琴并彬华同志：

为持续提升政务大模型的问答能力，我们委托中国信通院于11月上旬对百度、阿里、智谱3家大模型在公积金、交管、市场监管、不动产登记等4个场景的训练效果开展第二次阶段性测试。现将有关情况报告如下。

一、本次测评组织的基本情况

本次阶段性测评的主要目的是持续跟踪检验大模型训练效果。测评指标体系与第一次阶段性测评保持一致，为更全面掌握各模型的实际成效，大幅增加了测试集数量，使测试用例覆盖高中低频提问并按照难度进行适度分布，同时，将现场测评更改为通过API接口批量机跑单轮问答后交业务专家评阅打分（除打样部门业务专家外，市级综窗、丰台区综窗、通州区综窗也参与了评阅），同时，多轮部分由业务专家登录强化学习界面进行问答。测试集和业务专家评阅情况见附件。

业务专家评阅完成后，我们分别组织大模型企业、打样部门对错题进行了复盘分析。

二、本次测评结果

**（一）大模型业务能力测评**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 公积金场景业务能力评分 | | | | | | |
| 轮次 | 单轮 | | | 多轮 | | |
| 指标 | 准确性 | 完整性 | 友好性 | 准确性 | 完整性 | 友好性 |
| 智谱 | 2.96 | 2.69 | 2.7 | 2.78 | 2.81 | 2.64 |
| 百度 | 3.51 | 3.07 | 3.1 | 3.6 | 3.72 | 4.14 |
| 阿里 | 2.66 | 2.3 | 2.3 | 3.18 | 3.03 | 2.88 |
| 交管场景业务能力评分 | | | | | | |
| 轮次 | 单轮 | | | 多轮 | | |
| 指标 | 准确性 | 完整性 | 友好性 | 准确性 | 完整性 | 友好性 |
| 智谱 | 1.93 | 1.97 | 2.1 | - | - | - |
| 阿里 | 2.68 | 2.64 | 2.77 | - | - | - |
| 市场监管场景业务能力评分 | | | | | | |
| 轮次 | 单轮 | | | 多轮 | | |
| 指标 | 准确性 | 完整性 | 友好性 | 准确性 | 完整性 | 友好性 |
| 百度 | 4.66 | 4.7 | 5 | 4.36 | 4.4 | 5 |
| 不动产登记场景业务能力评分 | | | | | | |
| 轮次 | 单轮 | | | 多轮 | | |
| 指标 | 准确性 | 完整性 | 友好性 | 准确性 | 完整性 | 友好性 |
| 智谱 | 2.14 | 2.28 | 2.64 | 3.33 | 3.5 | 4.42 |
| 阿里 | 3.17 | 3.26 | 3.49 | 3.67 | 4.08 | 3.83 |

**（二）大模型通用能力测评**

在业务分类测试中，智谱、阿里（均为88%）略高于百度（86%）；在意图分类测试上，阿里、百度（均为96.5%）平均准确率高于智谱（94%）；在指代消歧测试上，百度（85%）表现优于阿里（80%）和智谱（75%）；在阅读理解测试中，三家大模型表现相近，均未出现明显知识点错误，从生成答案与标准答案相似度来看，通过每题得分横向对比，百度8题得最优分、智谱和阿里分别6题得最优分。

**（三）第一次、第二次测评结果对比**

**一是本轮测评得分下滑，但不代表模型能力倒退。**本次测评中，各单位直观评价各模型的表现均有较大提升。但是，测评结果计分上看，三个模型的整体综合得分出现不同程度下滑，个别模型下滑较明显。主要原因：一是本次测评测试集用例数量较第一次大幅增加，业务覆盖较全面，所考察的知识点和模型能力点多；二是较难提问数量明显增加，测评难度提升。其中，市公积金中心在测评过程中补充组织了一次专项测试，提供公积金接线员月考测试场景共计127道题目，智谱、阿里的判断正确率分别为65%、69%（百度未参与），准确率提升较为明显。

**二是本轮多轮问答能力整体提升较高。**对于条件提供不全的提问，能够逐步澄清，帮助用户确定真实诉求。百度和阿里在公积金场景多轮效果上表现较为明显，无论准确度、完整性得分均高于第一次。

**三是从大模型表现来看，百度表现较为突出，**在两个结对场景单轮和公积金场景多轮均降幅较小，无论在公积金场景还是市场监管场景得分均较高。**阿里在本轮进步明显，**尤其在交管场景的完整性和准确性上得分均有提升。**智谱本轮表现失常，**在各结对场景的单轮和多轮表现均不稳定，得分偏低，下降最为明显（智谱表示本轮表现失常的原因是代码上传错误）。

三、测评分析

本次测评难度较大，但也更接近真实场景。通过对本次测评结果以及两次测评的对比分析，结合与各单位、模型企业复盘、分析的情况，在大模型训练和工作推进上有如下特点：

**（一）大模型技术问题还需突破。**第一次摸底测试发现的问题中，除多轮得到明显优化外，其余问题依然未能在测评期内得到有效解决。此外，又发现了同一诉求涉及跨业务时跳出当前语境、知识树在某种程度上对简单问题反而产生干扰等新问题。总的来看，目前主要有三大类原因，分别是基础模型能力不足或衔接不好、外挂知识库能力缺陷、业务知识理解不透彻，需要分别加以解决。比如，政务术语口语化匹配依然是导致无法识别用户问题的一大障碍，对政务知识进行专题化处理（比如，按大模型指令的方式提供政务术语表）可增进模型对业务知识的理解。

**（二）大模型知识更新能力还需优化。**正式测试前1天市公积金中心下发了新政及配套解读，我们特别编写10道测试用例，智谱、百度、阿里分别答对8题、6题和4题，说明各模型已经具有了一定的知识更新能力，但是，若考题答案涉及新旧政策组合的情况，3家大模型企业均无法做出正确或完整的答复。各模型企业需与业务部门共同探索通过创新技术路径来实现与业务的高度融合，寻找治理知识的最佳路径。

**（三）测试集及强化学习量需扩大。**前期由于各模型总体效果不尽如人意，不具备大规模开展强化学习的基础。从当前实际测评效果和各模型进展来看，业务部门强化学习的数量对模型“改错”迭代影响较大。大模型企业反映，目前收集到的强化学习和测评集数据量较小（仅几百个），不具备对基础模型进行大规模训练的条件，至少需要几万数据来帮助大模型理解政务专业名词、增强推理能力，提升训练效果。同时，由于部分场景（比如规自委）专业性太强，大模型企业无法自行法优化，仍然需要业务部门深度参与。

**（四）模型稳定性问题较为突出。**在组织大模型企业与打样部门复盘的过程中，我们发现，除百度外，阿里和智谱模型的问答效果持续发生变化，企业反馈原因为：模型和外挂知识库一直处于不断技术调试、上下文指令改写、知识补充替换清理以及算法优化过程中。这从另一个方面说明，两家大模型企业技术实现方式仍然没有完全成熟。

四、下一步工作建议

**（一）启动场景内部试用。**目前，公积金中心、市市场监管局均表示可启动内部试用。计划按场景制定内部试用方案，尽快推动“边用边测边改”，确保收集足够量的真实提问。我们将做好统筹工作，对打样部门的内部试用方案提出规范和程序，联合业务单位、大模型企业持续定期测评、复盘、分析、补充和治理知识，探寻最佳技术提升路径和业务使用模式，确保大模型表现稳定。

**（二）优化测评方式，使测评工作更具可对比性。**总结两次测评工作情况，对测试集、业务专家评测方式、大模型企业配合方式、计分方式等进行细致复盘、优化，建立全市统一的测试集，使测试工作常态化，跟踪监测模型成长变化趋势，在帮助模型能力提升同时，绘制模型成长曲线，为下一步模型选择提供数据基础。

**（三）开展安全测评。**由市委网信办根据政务问答的特点制定内容安全、舆情安全、网络安全方面的评测指标和测评方案，开展安全方面的测评工作。

**（四）组织专家指导技术优化问题。**与市科委、市经信局联合，围绕大模型表现以及技术落地方式，寻找技术、高校、企业、业务等各方面的专家，共同探讨有关垂直模型能力强化、表现不稳定以及波动等方面的解决办法。

**（五）启动研究大模型落地模式。**自启动至今，“赛马”已近8个月，各大模型均取得一定成效，拟联合市经信局、市科委、市委网信办共同研究大模型选择、落地部署以及项目化方面的工作。

特此报告。

附件：1.关于第二次摸底测评测试集及业务专家评阅的情况说明

2.测评报告（2023年11月1日）

政府网站建设管理处

首都之窗运行管理中心

2023年11月14日

（承办：赵珣阳；复核：韩昉，51958；审核：张莉、刘莎）