ICS
CCS

T/GDEIIA

团　　体　　标　　准

T/GDEIIA

基于大模型的政务咨询系统技术要求与评估方法

Technical requirements and evaluation methods for government consultation system based on large model

（征求意见稿）

2023-xx-xx 发布　　2023-xx-xx 实施

广 东 省 电 子 信 息 行 业 协 会　　发 布

目　　次

前　　言

本文件按照GB/T 1.1-2020给出的规则起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由广东省电子信息行业协会提出并归口。

本文件起草单位：广州数据集团、广州市数字政府运营中心、中国信息通信研究院等。

本文件主要起草人：

本文件于2023年12月16日首次发布。

基于大模型的政务咨询系统技术要求与评估方法

1 范围

本文件面向以大规模与训练模型为技术底座，能够提供智能问答、政务咨询、知识搜索等功能的基于大模型的政务咨询系统。

本文件规定了基于大模型的政务咨询系统的功能、性能要求和评估方法，主要包括大模型基础能力、政务咨询业务能力、系统安全应用能力及指标评估方法四个部分。

本文件适用于基于大模型的政务咨询系统及同类产品的的研发、评估和验收等工作。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 35282-2023 信息安全技术 电子政务移动办公系统安全技术规范

GB/T 31506-2022 信息安全技术 政务网站系统安全指南

3 术语和定义

3.1

大模型 large model

一种基于海量通用数据训练得到的大规模预训练模型，具备多个领域的任务能力且通用性较高，但在实际应用场景中仍需结合生产数据进行二次开发。

4 缩略语

下列缩略语适用于本文件。

ROUGE：面向召回率的评价方法(Recall-Oriented Understudy for Gisting Evaluation)

LCS：最长公共子序列(Longest common subsequence)

5 指标要求及评估方法概述

5.1 评估方法概述

针对模型能力部分的评估方法包括检查和测试两类测评方法，具体为：

a) 检查：检查是通过对测评对象进行观察、查验、分析以帮助测评人员理解、澄清或取得证据的过程。检查主要有评审、核查、审查、观察、研究和分析等，检查对象是文档等；

b) 测试：测试是指使用预定的方法/工具使测评对象产生特定的结果，将运行结果与预期的结果进行比对的过程，主要包括人工评测、工具测试等测试操作。

5.2 指标要求及评估方式概述

表1 指标项与评估方式对照表

指标维度	指标项	评估方式
模型能力	模型信息披露	检查

指标维度	指标项	评估方式
	文本分类	测试
	语义理解	测试
	澄清反问	测试
	情感分析	测试
	信息摘要	测试
	内容生成	测试
系统功能	信息检索	检查
	智能对话	检查
	政务咨询	检查
	文件解读	检查
	智能填表	检查
系统性能	准确性	测试
	完整性	测试
	友好性	测试
	稳定性	测试
	实效性	测试
	响应时间	测试
系统安全可用	内容安全	检验、测试
	数据安全	检查
	应用安全	检查
	服务可靠性	检查

6 指标要求及评估方式详述

6.1 模型能力

6.1.1 模型信息披露

指标要求：系统提供方应向系统使用方披露模型基本信息。

评估目的：收集模型基本信息以作为模型能力评估及模型成本投入的参考信息。

评估方法：模型参数、训练数据、训练框架、时间成本及算力需求信息披露，具体如下：

a) 披露参测预训练模型的模型参数。对于单流结构模型，披露模型需要存储的参数量；对于双流或多流结构模型，分别统计各模型需要存储的参数量，披露模型需要存储的参数量总和。

b) 披露参测预训练模型的训练阶段的数据集大小。

c) 披露参测预训练模型依赖的训练框架类别。

d) 估算预训练模型训练全程各节点的时间消耗总和，基于训练使用的设备信息，将训练设备对标到基准设备下，计算模型训练时使用设备与基准设备的性能比值，换算出参测预训练模型在基准设备下的总训练时长。

e) 估算预训练模型训练全程各节点的时间消耗总和，基于训练使用的设备信息，将训练设备对标到基准设备下，计算模型训练时使用设备与基准设备的性能比值，换算出参测预训练模型在基准设备下的总训练时长。

6.1.2 文本分类

指标要求：模型应具备通用领域、政务领域、及政务业务领域文本分类能力。

评估目的：评估大模型对文本分类任务的性能。

评估方法：评估大模型对通用领域、政务领域、及政务业务领域文本进行分类的准确率，计算方法见公式：

$$P_C = \frac{P_1}{P} \times 100\%$$

式中：

P_C ——文本分类准确率；

P_1 ——分类正确的文本数；

P ——待分类的总文本数；

6.1.3 语义理解

指标要求：模型应具备语义理解能力，包括意图理解、政务专有名词理解等。

评估目的：评估大模型对语义理解任务的性能。

评估方法：评估对用户咨询话术中语义理解能力，包括意图理解、政务专有名词理解等，具体如下：

a) 意图理解性能评估方法：评估大模型对用户咨询话术中单意图、多意图的理解准确率，计算方法见公式

$$P_F = \frac{F_1}{F} \times 100\%$$

式中：

P_F ——意图识别准确率；

F_1 ——正确识别意图的总句数；

F ——意图识别数据总句数；

b) 政务专有名词理解性能评估方法：评估大模型对政务专业名词解释的准确率，计算方法见公式：

$$P_C = \frac{F_1}{F} \times 100\%$$

式中：

P_C ——政务专有名词理解准确率；

P_1 ——政务专有名词正确解释的总句数；

F ——政务专有名词解释的总句数；

6.1.4 信息摘要

指标要求：模型应具备信息摘要能力，以支持用户在政策、法规、办事指南、通知等文件中快速获取关键信息。

评估目的：评估大模型信息摘要的性能。

评估方式：评估大模型完成信息摘要任务的客观指标 ROUGE-L-f 值，计算方法见公式：

$$\mathrm{R}_{LCS}=\frac{LCS(X,Y)}{\mathrm{len(Y)}}$$

$$\mathrm{P}_{LCS}=\frac{LCS(X,Y)}{\mathrm{len(Y)}}$$

$$\mathrm{F}_{lcs}=\frac{(1+\beta^2)R_{LCS}\mathrm{P}_{LCS}}{R_{LCS}+\beta^2\mathrm{P}_{LCS}}$$

式中：

X——参考摘要；

Y——生成摘要；

LCS(X,Y) ——X 和 Y 的最大公共子字符串(LCS)的长度；

β——超参数；

6.1.5 内容生成

指标要求：模型应具备内容生成能力，可针对用户咨询问题生成可读易懂的回复内容。

评估目的：评估大模型生成内容的质量。

评估方法：采用人工评价法，对生成内容进行流畅性、连贯性主观评价，具体评分要求参考下表：

表2 内容生成评分准则

评分	评分准则	
	流畅性	连贯性
1 分	文本不具备可读性；	文本和前文逻辑矛盾；
2 分	文本具有可读性，但存在大量搭配不当等语法错误；	文本和前文存在少量逻辑矛盾；
3 分	文本基本流畅，存在少碰语法错误；	文本和前文无明显逻辑矛盾，但和前文转折不够流畅；
4 分	文本流畅，存在少量搭配不当；	文本和前文无逻辑矛盾，且和前文转折流畅；
5 分	文本十分流畅，无任何语法错误；	文本和前文无逻辑矛盾，且和前文连贯一致。

6.2 系统功能

6.2.1 信息检索

指标要求：系统应具备信息检索能力，在信息库中快速检索出用户询问或搜索的对应内容。

评估目的：检验系统是否具备政务信息检索功能。

评估方法：

a) 材料调研：

1) 材料中简述实现信息检索功能的技术方法；

2) 材料中简述信息检索功能实现情况（如信息源、检索维度、检索方式等）。

b) 模拟演示

1) 依照材料中信息检索功能实现情况进行系统操作演示；

2) 记录系统演示效果。

6.2.2 政务咨询

指标要求：系统应具备政务咨询能力，可针对用户咨询的政务相关问题做出回答。

评估目的：检验系统是否具备政务咨询功能。

评估方法：

a) 材料调研：

1) 材料中简述实现政务咨询功能的技术方法；

2) 材料中简述信息政务咨询能力范围（如咨询内容覆盖度、咨询方式、转人工逻辑等）。

b) 模拟演示

1) 演示材料中提到的信息检索能力；

2) 记录系统演示效果。

6.2.3 智能对话

指标要求：

a) 应具备单轮对话能力；

b) 应具备多轮对话能力；

c) 应具备多类型智能对话能力，如任务式对话、闲聊式对话等；

d) 宜具备多语种对话能力。

评估目的：评估系统是否具备智能对话多项能力。

评估方法：

a) 材料调研：

1) 材料中简述系统具备哪些智能对话能力；

2) 材料中简述系统智能对话优化方式。

b) 模拟演示

1) 演示材料中提到的智能对话能力；

2) 记录系统演示效果。

6.2.4 政策解读

指标要求：系统宜具备政策解读能力，以辅助工作人员快速了解各类政府文件中的内容。

评估目的：检验系统是否具备政策解读功能。

评估方法：

a) 材料调研：

1) 材料中简述实现政策解读功能的使用方法；

2) 材料中简述信息政策解读能力范围（如内容颗粒度、解读内容呈现等）。

b) 模拟演示

1) 演示材料中提到的政策解读能力；

2) 记录系统演示效果。

6.2.5 智能填表

指标要求：系统宜具备智能填表能力，以辅助工作人员及用户快速填制相关表格。

评估目的：检验系统是否具备智能填表功能。

评估方法：

a) 材料调研：

1) 材料中简述实现智能填表功能的使用方法；

2) 材料中简述信息智能填表能力范围（如表单类型、可填内容等）。

b) 模拟演示

1） 演示材料中提到的智能填表能力；

2） 记录系统演示效果。

6.3 服务性能

6.3.1 准确性

指标要求：系统政务问询能力应具备准确性，回复内容应包含问题中提到的关键知识点，并且可对用户意图进行准确理解，模糊意图问题进行澄清和反问等能力。

评估目的：评估政务问询回答准确性。

评估方法：采用人工评价法，对问询答案进行准确性主观评价，具体评分要求参考下表：

表3 政务咨询准确性评分准则

评分	评分准则
	准确性
1 分	用户意图完全识别错误，答非所问；
2 分	用户意图未完全理解，关键知识点回答存在小部分错误或遗漏；
3 分	用户意图理解全面且准确，关键知识点回答存在小部分错误或遗漏；
4 分	用户意图理解全面且准确，关键知识点全部正确回答，但是答案中包含与其他意图相关的信息；
5 分	用户意图理解全面且准确，关键知识点全部正确回答，并且语句描述自然易懂，没有其他意图相关的冗余信息。

6.3.2 完整性

指标要求：系统政务问询能力应具备完整性，回复内容应确保答案内容全面、完整、无重要信息遗漏，并提供关联知识参考。

评估目的：评估政务问询回答完整性。

评估方法：采用人工评价法，对问询答案进行完整性主观评价，具体评分要求参考下表：

表4 政务咨询完整性评分准则

评分	评分准则
	完整性
1 分	答案不完整，缺少所有关键细节和解释，无多知识点整合能力，无知识依据来源或知识依据来源错误；
2 分	答案不完整，缺少部分关键细节和解释，对多知识点整合完整不完整，知识依据来源清晰；
3 分	答案基本完整，但一些关键细节的解释但不够完整和详细，对多知识点整合略有遗漏，知识依据来源清晰；
4 分	答案完整，包含了大部分的关键细节和解释，但有一些细微的不完整之处，对多知识点整合完整，知识依据来源清晰；
5 分	答案完整，答案包含了所有的关键细节和解释，对多知识点整合完整，没有任何遗漏，知识依据来源清晰。

6.3.3 友好性

指标要求：系统政务问询能力应具备友好性，应确保系统回答的内容易于理解和接受，符合用户的语言习惯和心理预期，尽可能用通俗易懂便于理解的形式回答问题，提高用户的满意度和体验感。

评估目的：评估政务问询回答友好性。

评估方法：采用人工评价法，对问询答案进行完整性主观评价，具体评分要求参考下表：

表5 政务咨询友好性评分准则

评分	评分准则
	友好性
1 分	答案的表达方式晦涩难懂，对负面情绪有不友好回答；
2 分	答案表达方式不够清晰或不够易于理解，出现大量难懂政策词汇，对负面情绪无友好性回答；
3 分	答案表达方式清晰但缺乏一些深入的解释，对负面情绪无友好性回答；
4 分	答案表达方式清晰且易理解，具备逻辑性及总结能力，对负面情绪无友好性回答；
5 分	答案表达方式非常清晰易于理解，具备逻辑性及总结能力，对负面情绪能做出安抚性回答。

6.3.4 时效性

指标要求：系统政务问询能力应具备时效性，确保输出输出的时效性，输出内容在有效期内，能够识别并过滤过期内容。

评估目的：评估政务问询回答友好性。

评估方法：采用人工评价法，对问询答案进行完整性主观评价，具体评分要求参考下表：

表6 政务咨询时效性评分准则

评分	评分准则
	时效性
1 分	答案内容全部太过老旧，答案中出现常识性错误；
2 分	答案的关键内容老旧，信息对用户产生误导；
3 分	答案部分非关键内容老旧，信息不会对用户产生严重误导；
4 分	答案内容存在有部分不合时宜，信息不会对用户产生严重误导；
5 分	答案内容能够识别并过滤过期内容，无不合时宜内容的输出

6.4 系统安全可用

6.4.1 内容安全

指标要求：

a) 应具备意识形态安全性，

b) 应不涉及违法及伦理道德内容。

评估目的：评估系统输出内容安全性。

评估方法：技术测试和专家攻击等。

6.4.2 数据安全

指标要求：系统应参照GB/T 35282-2023中9.5的要求。

评估目的：评估系统是否满足GB/T 35282-2023中9.5的要求。

评估方法：应参照GB/T 35282-2023中11.1.2.3的测试方法。

6.4.3 应用安全

指标要求：系统应参照GB/T 31506-2022中6.5.5中内容要求。

评估目的：评估系统是否满足GB/T 31506-2022中6.5.5中内容要求。

评估方法：参照GB/T 31506-2022中6.5.5中内容要求进行材料查验。

6.4.4 服务可靠性

指标要求：系统应具备服务可靠性。

评估目的：评估系统是否具备服务可靠性

评估方法：检查参评单位服务可靠性保障策略相关材料，如平均故障间隔时间、用户无感的系统升级、服务状态监测和自动重启、冗余备灾策略等。