中银证券-大模型平台POC方案

# 方案概述

本测试方案对各厂商基于智能问答，文档召回，内容溯源等大模型搜索增强的能力应用在文件检索、文件审核、报告撰写及数据问答场景进行测试评估。本规范规定了各厂商在POC中的测试内容及其方法；测试环境由各厂商在上证云服务器上自行搭建（私域模型），测试结果作为系统选型的参考依据。

上证云通过申请机制，限制白名单内ip访问上证云。申请用户通过专有网络VPC访问对应的资源集和地域，实现不同用户间的逻辑隔离、网络隔离。并且每个用户只能影响自己有权限的资源集，实现用户权限隔离。

# 测试内容

本次POC的测试内容包括客观系统功能点以及业务场景建设能力。包含文件检索、文件审核、报告撰写及数据问答四个能力测试。

# 测试数据

POC业务样例数据列表：

1. 外部管理制度文档、外部监管信息文档、公司对外发布的合规文档。
2. 对外发布的宣传营销物料
3. 市场公开的公募基金报告
4. 市场公开的行情数据、舆情数据

# 测试流程

参与测试的厂商需要准备好相关系统，基于我司提供的数据完成问答及业务操作页面的调优工作，并根据我司的安排，在测试完成后进行测试和展示。本次测试为内部测试，测试结果和最终评分有我司评委给出。总体流程

1. 交流
2. 厂商搭建环境
3. 厂商根据现有环境展示产品功能，并进行业务功能测试
4. 厂商方进行POC测试成果汇报
5. 相关人员对各家实施商的POC情况进行评审并打分。

# 系统功能点测试项

## 后台管理功能

|  |  |
| --- | --- |
| 功能类别 | 功能点 |
| 用户及用户组 | 支持在后台管理支持新建用户和用户组。 |
| 支持配置用户的角色是否为管理员用户。 |
| 业务空间管理 | 支持在系统后台对所有业务空间进行管理，可修改、添加业务空间管理员。 |
| 系统应用 | 支持后台创建系统应用，系统应用可通过应用 ID 和密钥使用 API 调用系统功能。 |

## 知识库管理功能

|  |  |
| --- | --- |
| 功能类别 | 功能点 |
| 创建知识库 | 可创建知识库，自定义知识库名称、知识库介绍，设置知识库标签，设置搜索参数以及Embedding参数等； |
| 文件上传及处理 | 支持TXT/MD/Excel/HTML/Doc/DocX/PDF/JPEG/PNG/BMP/TIF 等不同格式文件上传，提取文件中的文本内容。支持将文档中的图片、图表抽取为链接形式。 |
| 支持多种文件切分方式，如智能切片、固定长度切片、按照文本语义动态长度切片等，对切片进行向量化处理，建立索引并落库存储 |
| 支持基于表格识别能力增强、基于OCR的文件识别能力增强 |
| 支持批量文件上传、批量下载。 |
| 支持文件分类树对上传的文件进行分类管理。 |
| 支持对上传失败的文件进行重试、批量重试。 |
| 支持对上传的文档进行解析内容的编辑修正。 |
| 支持对上传的文档(pdf, html)进行预览和问答。 |
| 问答对(QA对)管理 | 支持人工编写/批量上传问答对（QA对），问答对加入知识库系统； |
| 通过语义理解、引导词等自动从历史问答/文本中生成问答对（QA对），问答对加入知识库系统； |
| 支持配置问答对的相似问题，新建、编辑时自动生成相似问题。 |
| 支持问答对答案的智能优化。 |
| 支持对自动生成的问答对、手动录入、批量导入的问答对进行审核，只有通过审核入库后方能生效。 |
| 支持配置和分配待审核任务。 |
| 权限管控 | 支持设置知识库的可见、可使用范围，支持配置本空间可用、部分空间可用或全部可用。 |
| 如果助手绑定的知识库对某用户不可用（无权限）的情况下，给出明确提示，如只使用部分知识作答，部分知识无权限，需要申请。 |
| 运营及审计 | 支持对知识库中的数据进行统计展示，包括知识数量、问答对数量，历史更新数量等。 |
| 支持记录并展示知识库中的操作记录。 |

## 知识库检索功能

|  |  |
| --- | --- |
| 功能类别 | 功能点 |
| 知识库检索策略 | 支持配置混合检索、向量检索、全文检索的检索模式，支持配置混合检索权重比例。 |
| 多知识库检索模式，支持配置串行检索、并行检索。 |
| 智能体知识库检索设置，支持配置检索全部、自动匹配检索。 |
| 支持 QA 优先检索模式，支持配置 QA 匹配阈值，支持 QA 匹配提示词设置，若匹配 QA 则直接输出答案。 |
| 支持使用 rerank 模型对召回结果进行精排。 |
| 支持对召回的上下文进行自动扩充填补。 |
| 知识库检索测试 | 支持多知识库共同进行检索测试，通过选定多知识库，选定检索设置参数，给出检索问题后，会基于当前选定的知识库及参数给出知识库检索结果。 |
| 保存历史检索测试记录。 |

## 数据库管理功能

|  |  |
| --- | --- |
| 功能类别 | 功能点 |
| 数据库配置 | 支持数据库录入，输入数据库地址、用户名、密码，保存为数据库信息。 |
| 支持数据库连接测试，通过连接测试后方能使用。 |
| 支持接入 MySQL、SQL Server、PostgreSQL、Oracle、GaussDB、GoldenDB、Hive 数据库。 |
| 支持自动导入数据库表结构。当数据库表结构有变更时，支持同步数据库表结构。 |
| 支持编辑数据库库、表的描述，支持编辑数据库表字段的描述。这些描述信息提供给大模型智能体识别使用。 |
| 支持选择哪些表提供给大模型进行访问，如果超出表范围，则进行提示，并拒绝查询。 |
| 支持写操作保护，识别大模型生成的 SQL 中包含修改数据库信息、表结构等危险操作，进行提示并拒绝执行。 |
| 查询示例，支持导入问题和 SQL 查询示例，智能体调用数据库时自动查询相关示例，通过 In Context Learning 提升生成 SQL 准确率。 |

## 工作流管理功能

|  |  |
| --- | --- |
| 功能类别 | 功能点 |
| 工作流配置 | 支持在业务空间中以图形方式配置工作流。 |
| 支持工作流调试和试运行，支持前端展示每个节点的运行状态。 |
| 工作流节点支持 | 支持大模型节点、插件节点、知识库节点、代码节点、选择器节点、助手节点及工作流节点。 |

## 插件管理功能

|  |  |
| --- | --- |
| 功能类别 | 功能点 |
| 系统内置插件 | 支持代码解释器、指标绘图、网络搜索等系统内置插件。 |
| 插件创建与引用 | 支持配置及集成外部插件，完成配置后可以API形式集成到大模型的应用场景中； |
| 插件与工具的管理 | 支持插件中配置多个工具，支持工具的配置、调试、发布等，已发布插件可在助手配置时集成 |
| 权限管控 | 支持配置插件的使用范围，包括只允许本空间可用、部分空间可用、公开可用。 |

## 助手配置管理功能

|  |  |
| --- | --- |
| 功能类别 | 功能点 |
| 多类型助手创建 | 支持基于模版新建助手，支持文档问答助手、数据库问答助手、智能体助手、摘要助手等，可配置助手的名称、介绍、标签等 |
| 版本管理 | 支持助手的版本管理，记录助手版本历史信息。支持选择某个历史版本进行发布。 |
| 灵活的助手配置项 | 支持对助手进行多种配置，如添加大模型并自定义prompt提示词、关联知识库及设置RAG配置、关联插件工具、关联数据库、自定义助手的开场白、开场问题等 |
| 支持配置推荐问，推荐问根据用户的上下文进行推荐，优先推荐 QA 库中存在的问题，或者历史召回片段中助手能够回答的问题。 |
| 测试运行及调试 | 可视化测试运行环境，测试提示工程大模型问答效果； |
| 对话日志 | 支持查看对话日志并导出，对话日志包含消息总数、用户反馈内容及原因等。 |
| 支持根据来源、用户反馈、标注反馈对对话日志进行筛选。 |
| 支持对话日志导出，支持仅导入用户点踩的对话日志。 |
| 权限管控 | 对助手的使用进行权限管控，支持配置仅本空间可用、部分空间可用及公开可用。 |
| 敏感词管理 | 支持系统、空间、助手三级敏感词管理控制。助手可以关联使用系统、空间中的敏感词信息，或者单独配置敏感词使用。 |
| 当用户问题或者答案命中敏感词时，给出相关提示。问题或回答命中的提示可以单独进行配置。 |
| 可在助手管理后台查看助手命中敏感词的历史记录。 |
| 数据标注 | 支持对助手对话历史进行标注。 |
| 支持对标注结果进行审核。审核完成后进入指定知识库的问答对中，保证后续问答效果的持续提升。 |
| 支持按对话历史时间段进行标注任务分配。 |
| 标注及审核时，支持检索知识库等对标注审核进行辅助。 |
| 分析看板 | 支持对助手的使用数据和评价指标进行统计分析。 |

## 助手评测功能

|  |  |
| --- | --- |
| 功能类别 | 功能点 |
| 数据集管理 | 支持上传评测使用的数据集。 |
| 支持知识问答数据集、关键词测评数据集、召回率数据集、插件参数差错率数据集等。 |
| 助手评测管理 | 支持使用给定的数据集对助手的指定的版本进行评测。 |
| 支持知识问答评测、关键词评测、召回率评测、插件调用参数评测等。 |

# 场景能力测试

## 文件检索

基于上传到知识库的各类文档数据，调用大模型能力准确高效返回用户查询的片段内容。

## 文件审核

1. 公募基金报告核对：对比审核多篇公募基金报告，对同一报告中同一概念的数据进行核对，若数据不一致进行标记。不同报告中同一章节的内容进行核对，若与给定的标准内容不一致进行标记。
2. 宣传物料审核：提取PDF、图片等宣传文件中文字；对照审核标准进行合规评审并反馈修改意见

## 报告撰写

2.按照指定日期，根据市场行情、行业动态、政策舆情等公开数据，编写投资市场综合性评述。

## 数据问答

根据用户需求描述，能正确从wind库提取相应数据字段（包含不同表的不同字段组合）。

# 业务效果评价方法

基于参评厂商的各项评测返回结果采用人工评价指标——MOS（Measure of Software Quality），基于一组标准人工评估大模型的性能。

|  |  |  |
| --- | --- | --- |
| 场景 | 评估指标 | 评估口径 |
| 文件检索 | 准确率 | 根据标准答案比较文本重叠度 |
| 时效性 | 完整答案时间长度 |
| 召回率 | top 10命中片段占比 |
| 文件审核 | 准确率 | 基于业务标准规则，查看审核出的部分占比 |
| 报告撰写 | 业务满意度 | 业务主观打分 |
| 数据问答 | 准确率 | 返回指标数量是否全、数值是否正确 |

根据人工评测的指标维度，由参与者以分数的形式来进行评分。指标：对每条数据的总体评分进行加和平均，得出一个总体评价得分。

|  |  |
| --- | --- |
| 分数 | **问答评分标准** |
| 5分 | 回答正确且质量高，结果真实，非常符合用户期望，无冗余。 |
| 4分 | 回答正确但质量稍差，结果真实，较符合用户期望。 |
| 3分 | 大部分回答正确，结果真实，正确部分符合用户期望。 |
| 2分 | 大部分回答不正确或结果不真实，只有很少一部分符合用户期望。 |
| 1分 | 有结果，但回答基本错误或回答相关度很低。 |
| 0分 | 结果为空、完全错误或回答无关，或回答内容违反安全和伦理要求。不在知识库的内容回答不知道。 |



# 测试周期

各家公司的测试周期时长均相等，单一公司准备周期为x个工作日(已经完成)。x月x日厂商开始部署环境准备POC测试，x月x日-x月x日业务开始POC结果验收。

# 公共说明

## 测试地点

本次POC项目基于上证云环境部署

## 联系方式

POC项目测试期间如有任何问题，可联系以下客户联系人。

姓名：

手机号码：

电子邮箱：

## 评分

科技人员：

1根据功能点有无，评客观分；

业务人员：

1根据业务效果，评客观分

2根据业务使用易用性，评主观分