# 微软合作项目-V1.3

# 项目概述

本项目旨在通过与微软亚洲研究院的合作，建立一个AI产品专家模型，能对云学堂的产品知识进行学习并深刻理解，能够准确回答用户对于产品功能的咨询问题，并能对产品中涉及到的业务进行世界知识的检索，并给出能提升产品专业度的有效建议。

# 项目描述

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **一、产品知识问答** | | | | | | |
| **一、项目基本情况** | | | | | | |
| **项目名称：** | **产品知识问答** | | **项目编号：** | |  | |
| **制作人：** |  | | **审核人：** | |  | |
| **项目经理：** |  | | **制作日期：** | |  | |
| **二、项目描述** | | | | | | |
| **1. 项目背景与目的（所有的项目均起始于某个商业问题，该部分简要描述这些问题）** | | | | | | |
| 在日常服务客户的过程中，每天都会收到数百条来自客户的咨询问题，这些问题涵盖了功能操作、场景实现、系统提示信息解读，甚至包括忘记登录账号和密码等各类情况。为了解决这些问题，提升客户服务体验，并减轻产品经理和研发人员的工作压力，需要提供AI产品知识问答助手的功能，以确保能够及时且准确地回应客户的需求。 | | | | | | |
| **2. 项目目标（包含质量目标，工期目标、费用目标和交付产品特征与特征的主要描述）** | | | | | | |
| 利用AI产品知识问答助手自动回复客户的咨询问题，以此释放产品经理和研发人员的人力资源，使他们能够集中精力处理更为复杂和重要的任务。 | | | | | | |
| **三、项目里程碑计划（包含里程碑的时间和成果）** | | | | | | |
| **选择一：**   * 阶段1：在一个特定场景实现L1级别的问答，预期第1-2个月. * 阶段2：在另外1个场景实现L1级别的问答，预期第3-4个月。 * 阶段3：在一个特定场景实现L2级别的问答，预期第3-5个月。 * 阶段4：在另外1个场景实现L2级别的问答，预期第6-7个月。   **选择二：**   * 阶段1：在一个特定场景实现L1级别的问答，预期第1-2个月. * 阶段2：在另外1个场景实现L1级别的问答，预期第3-4个月。 * 阶段3：在一个特定场景实现L2级别的问答，预期第3-5个月。 * 阶段4：总结，部署，解决应用中的问题，预期第5-6个月   **依赖数据**   1. **产品定位及业务描述** 2. **产品模块定位及业务描述** 3. **模块包含的页面** 4. **操作链路日志（页面Id/页面信息/操作点位/功能名称/操作结果） 注：这个数据量很大，能够覆盖所有的需求场景。如果有特别的长尾需求，可以让后台先操作一遍。（需要有对应操作的目的或者完成的功能描述）** 5. **功能点位参考资料（PRD/测试用例/操作手册）** 6. **现有FAQ（问题和答案，覆盖L1和L2）。 注1：对于目前没有答案的L2问题，系统自动推理生成答案 注2：已有的L2问题将被作为示例和测试用例。 注3：不包含截图（流程图等）生成。** | | | | | | |
| **四、评价标准（说明项目成果在何种情况下将被接受）** | | | | | | |
| 对于涉及产品功能的问题，回答中必须包含相关的截图或详细的操作路径。例如，如果问题是关于 “如何创建一个项目”，则回答应指明创建项目的具体位置（即所属的导航菜单），并附上详细的操作步骤及相应的截图。  我们指定随机100条客户服务中碰到的问题，回答的准确性高于95%视作合格。  微软注：回答准确性与具体问题的分布及评价标准关系密切，在看到实际的问题及初步试验完成前，无法预估回答准确性。 | | | | | | |
| **五、项目主要利益干系人（包括高管、客户、职能部门主管、供应商、项目赞助人、项目经理、项目组成员等干系人）** | | | | | | |
| **姓名** | | **类别** | | **部门** | | **职务** |
|  | |  | |  | |  |
|  | |  | |  | |  |
|  | |  | |  | |  |
|  | |  | |  | |  |
|  | |  | |  | |  |

# 备注

微软亚洲研究院计划投入的人员及其时间分配如下：

1. 首席研究员或者首席研究经理：0.3
2. 高级研究员，高级应用科学家：0.5
3. 研究工程师：0.3

注：

1. 以上时间分配量可能有一个或者多个同级别人员共同组成。
2. 微软可能根据项目执行需求或者人员分配情况调整具体分配在该项目的人员，但总分配量不低于上述计划。

**关于所提供方案在公开数据集准确性的承诺：**

微软亚洲研究院所提供的方案在下述公开数据集能够达到下表中的准确性指标，可以在客户的Azure订阅环境中复现。

A table with numbers and letters

Description automatically generated

* 所有以上结果基于GPT-4-1106-preview
* 评价指标：
  + EM – 精确匹配
  + Acc – 用GPT-4来确认答案给标准答案的语义一致性（推荐）
* 指标计算器：  
  <https://github.com/RUC-NLPIR/FlashRAG/blob/main/flashrag/evaluator/metrics.py>

**HotpotQA**

* 领域：多跳问答（Multi-hop Question Answering）
* 问题性质：HotpotQA包含约113,000个基于维基百科的问题-答案对。问题需要通过多个支持文档进行查找和推理才能回答。它测试系统在多文档推理、事实比较和解释预测方面的能力。

**2WikiMultiHopQA**

* 领域：多跳问答（Multi-hop Question Answering）
* 问题性质：2WikiMultiHopQA使用结构化和非结构化数据，包含推理路径的证据信息。问题需要通过多个推理步骤才能回答，旨在全面评估推理步骤。

**MuSiQue**

* 领域：多跳问答（Multi-hop Question Answering）
* 问题性质：MuSiQue通过组合单跳问题来创建多跳问题，包含约25,000个2-4跳的问题。它旨在测试系统在多跳推理中的真实能力，并包含不可回答的对比问题以增加数据集的严格性。