

“中国电子杯”第三届高校 ICT 产教融合创新大赛

赛题八

一、赛题名称

多模态数据驱动的可解释精准问数/问答智能体

二、命题单位

中电云计算技术有限公司

三、所需知识背景

本赛题聚焦于大数据与人工智能的交叉前沿，旨在构建一个能够理解用户意图、精准解析用户问数/问答需求，并在复杂数据环境下提供可解释结果的智能系统。参赛者需具备或快速学习以下核心领域的知识：

1. 数据平台与存储

- ❖ 结构化数据：理解关系型数据库、数据仓库、数据平台的基本概念，熟悉 SQL 语言及多表关联查询。
- ❖ 非结构化数据：理解对象存储、向量数据库的基本原理，掌握处理 PDF、文档、图片等非结构化数据的流程。
- ❖ 多模态数据湖仓：了解现代数据平台如何统一管理结构化与非结构化数据，实现“查数”与“搜知识”的融合。

2. 知识问答与信息检索

- ❖ NL2SQL（自然语言转 SQL）：将用户的自然语言问题转化为精确的结构化

查询语句，涉及语义解析、槽位填充、Schema Linking（模式链接）等技术。

- ❖ RAG（检索增强生成）：基于大语言模型（LLM），通过从外部知识库检索相关信息来增强生成内容的准确性和事实性，应对非结构化知识的问答。
- ❖ 多跳问答：针对需要结合多个文档或数据片段进行推理才能回答的复杂问题，例如，“对比 A 项目去年的营收和今年市场分析报告中的预期增长率”。

3. AI 智能体（AI Agent）

- ❖ 感知与规划：智能体如何理解复杂的用户意图，将高层级任务分解为可执行的子任务（如先查表，再搜文档，最后总结）。
- ❖ 工具使用与记忆：智能体如何调用 NL2SQL、文档解析、OCR 等工具，并拥有短期和长期记忆，以在多轮交互中保持上下文连贯性。
- ❖ 可解释性：智能体不仅要给出答案，还要展示其思考过程、工具调用链、数据来源，让答案“有迹可循”，解决大模型“黑箱”难题。

四、赛题背景

在信息产业高速发展的今天，企业的核心竞争力越来越依赖于对全域数据的深度洞察。当前面临两大挑战：

1. 结构化问数对数据平台的意义与挑战

- ❖ 意义：“问数”能力是将数据平台民主化的关键一步，它让不具备 SQL 技能的业务人员也能直接触达 PB 级数据，实现“人人用数”，极大提升决策效率。
- ❖ 挑战：核心在于“精准性”与“复杂性”的矛盾。业务术语（别名、黑话）与数据库规范表名、字段名之间存在鸿沟；复杂分析需求常常涉及多表关联、

聚合、窗口函数，自动生成的 SQL 质量极不稳定，一个错误的 JOIN 可能导致全量数据泄露或计算错误，在金融、政务等领域这是不可接受的。

2. 非结构化知识问答对知识库的意义与挑战

- ❖ 意义：“问答”能力解锁了占比超过 80%的非结构化数据（PDF 报告、技术文档、规章制度）的价值，实现基于知识的精准对话，是构建企业级智能客服、知识中台的核心。
- ❖ 挑战：核心在于“幻觉”与“解析”的瓶颈。大模型容易生成看似合理但事实错误的内容。同时，真实文档格式千奇百怪（扫描件、复杂表格、多级目录错乱），准确解析并理解其结构逻辑（如识别并应用文档中的计算公式）是保证知识问答质量的基石。

本赛题旨在融合上述两大方向，打造一个能够同时应对精准问数与知识问答双重挑战的、具备高度可解释性的智能体系统。

五、赛题描述

参赛队伍需设计并实现一个多模态数据环境下可解释精准问数/问答智能体。系统需接收用户的自然语言查询，该查询可能涉及结构化数据（数据库表）和/或非结构化数据（知识库文档）。系统需要自主规划任务路径，精准执行查询与检索，并通过推理整合所有信息，最终生成一个准确、完整且具备可解释性的答案。

核心考察能力：

- ❖ 复杂意图理解与泛化能力：对模糊、有歧义、包含业务别名或专业术语的输入进行精准纠错与改写。

- ❖ 异构数据源精准定位能力： 准确识别问题所涉及的数据来源，是在结构化数据库的哪张表、哪个字段，还是在哪个知识库文档的哪一章节。
- ❖ 融合推理与分析能力： 能够执行跨源查询，例如“查询 2024 年销售冠军的业绩，并总结他的成功方法论”，这一查询需要先从数据库找出冠军是谁，再去文档库检索他的经验分享。
- ❖ 鲁棒性与可解释性： 系统在应对低质量数据（模糊扫描件、错别字、颠倒图片）时保持稳定，并清晰展示从问题输入到最终答案的全链路推理过程。

六、评价指标

为吸引广泛参与并有效选拔人才，大赛设置“基础任务”、“中级任务”，评分权重分别为 40%、60%（注：高级任务为额外加分，总分可超 100 分）。

1. 基础任务（40 分）：基础问数/问答能力搭建与演示

目标： 跑通基本流程，实现一个能处理简单、明确问题的端到端系统。

要求：

- ❖ 搭建一个可交互的 Demo。
- ❖ 实现单表简单查询的 NL2SQL 功能，并能正确执行返回结果。
- ❖ 实现单文档的简单检索问答（RAG）功能。
- ❖ 系统能够对答案进行基本的来源展示，例如高亮显示引用的文档片段或展示生成的 SQL 语句。

2. 中级任务（60 分）：核心难题攻克

目标：在一系列真实世界的难题中取得突破，系统需具备更强的鲁棒性与准确性。参赛队伍可从下列难题中选择若干进行重点攻克（初赛 ≥ 1 个、决赛 ≥ 3 个），并在技术文档中详述解决方案。

结构化问数赛道难题：

- 1) 要素精准识别与问题改写：在不限定具体实体类型的前提下，构建高可用性的要素泛化模型（如“北京市”的别名“北京”、“帝都”的识别），并基于此对不标准的用户输入进行智能纠错与标准化改写。
- 2) 可扩展的要素提取：设计一套不依赖将大量示例写入提示词（Few-Shot Prompting）的要素提取方案，以保证在面对新增业务领域时，系统能低成本、快速地扩展识别能力。
- 3) 智能语义校验与澄清：当用户问题存在维度缺失（如问“增长率”却未指定时间范围），系统需主动发起澄清，引导用户补全信息，而非直接生成错误查询。
- 4) 精准表定位（Schema Linking）：解决如何根据提取的要素，在数十张表构成的业务库中精准定位到正确的表和字段，排除语义相近但实际无关的干扰项。
- 5) 高质量多表关联 NL2SQL：保证涉及多表 JOIN、聚合、嵌套查询的 SQL 生成质量和稳定性，重点解决表关联路径的正确选择问题。

非结构化知识问答赛道难题：

- 6) 文档公式识别与计算：

(6a) 如何从非结构化文档（如 PDF）中，精确识别出哪些文本内容是计算公式

的输入参数。

(6b) 如何正确识别并解析公式本身的结构与逻辑，确保计算结果无误。

7) 不标准目录结构的鲁棒识别：解决 PDF 文档中目录格式混乱、层级丢失的问题，能够正确识别各级标题，并按正确的层级结构组织文档切片，这对于多级标题下的长文档问答至关重要。

8) 文档复杂度评估：设计一套文档复杂度评估模型，包括但不限于评分规则设计与实现路径，用于预估解析难度，指导后续处理策略。

9) 文档质量综合评估：实现对文档常见质量问题的自动检测与告警，包括但不限于：页面颠倒/倾斜、插图清晰度不足、繁体字/错别字识别并校正等。

3. 高级任务 (20 分, 加分项)：卓越突破与创新辨识 (仅决赛适用)

本部分不设固定任务清单，旨在辨识并奖励那些在中级任务基础上，展现出超乎寻常的技术深度、独创性或应用价值的作品。它可能是一个点的极致突破，也可能是一条简捷的新路径。无论最终整体得分如何，卓越的单项贡献都应被看见。

评审将重点关注，参赛作品是否至少满足以下一类特质：

1) 技术突破领先性。哪怕问题尚未被完美解决，但你提出的方法在解决路径、核心指标或思想框架上，达到了业界公认的领先水平。具体表现 (不限于此)：

❖ 指标突破：在公认的基准数据集或大赛提供的测试集上，于某项核心难题 (如复杂 NL2SQL 执行准确率、扭曲表格识别 F1-score) 上取得了显著高于现有公开最优结果的成绩。

❖ 方法独创：提出了一种新颖且有效的模型架构、训练范式或算法 (例如，一种无需标注数据的跨域 Schema Linking 方法，或一种新的公式识别图神经

网络），即使计算量巨大、当前不易产业化，但具备重要的学术启发价值。

- ❖ 极限挑战：勇于在极端设定的“高混淆”或“零样本”场景下进行测试，并证明了方法在该场景下的有效性与局限性，为领域提供了宝贵的研究边界。

2. 实用创新与转化价值。你的方法未必是理论最优，但因其巧妙的设计，能在现有技术条件下直接产生应用效益，展现出极高的工程智慧。具体表现（不限于此）：

- ❖ 方案创新：针对某个难题，提出了一种出奇简单、易于实现和维护的解决方案，显著降低了系统复杂度或计算成本，并能在实际产品环境中稳定运行。例如，用一个精巧的规则和传统小模型的组合，就大幅提升了别名识别泛化能力。
- ❖ 工程化创新：在系统设计层面有突出贡献，如设计了一套极低延迟的推理链路、一个鲁棒的反馈自愈机制，或一个优雅的多智能体协同框架，实现了从“可用”到“好用”的跨越。
- ❖ 直击痛点的产品化：交付的作品功能完整，用户体验极佳，并能清晰阐述其针对特定行业（如金融、政务、医疗）痛点的商业化路径和价值闭环，已具备 Demo 级的产品成熟度。

提交与评审方式：

参赛队伍若认为自身作品符合以上特质，需在提交材料中专门撰写一份《突破与创新性自述》报告（篇幅不超过 5 页），在其中明确指出：

- ❖ 定位：声明申报的是“技术突破领先性”还是“实用创新与转化价值”。
- ❖ 核心贡献：用精炼的语言阐述突破点是什么。

- ❖ 举证：提供翔实的证据，如与**公开最优结果**的详细对比实验、场景适配说明、成本效益分析、系统架构图解等。

评委将根据这份自述，结合系统实测与答辩表现，进行综合评判。此部分得分将与基础、中级任务得分累加，形成最终排名。

七、参赛平台

参赛平台不做强制要求，但为了方便参赛和评比，这里提供一些建议。

- ❖ 开发环境建议采用 Python + Vue，注意版本管理以便环境部署及应用重现。
- ❖ 数据集建议采用开源公开数据集。

推荐使用 OHR-Bench，它是一个用于评估 OCR（光学字符识别）对 RAG（检索增强生成）系统级联影响的基准测试，下载地址：<https://hf-mirror.com/datasets/pendant/OHR-Bench>。数据库推荐使用 Chinook，它是一个数字媒体商店的样例数据库，下载地址：<https://github.com/lerocha/chinook-database>。复杂案例可以选用数据集 AdventureWorks (PG)，它有现成 Docker 镜像 chrisaton/adventureworks:postgres。

- ❖ 大模型建议订购国产 Token，小模型采用开源产品（若需）。

八、校内初赛作品要求及评分标准

1.校内初赛作品提交要求

(1) 技术文档——设计报告：提交 Word 版本及 PDF 版本各 1 份，总篇幅建议不超过 20 页（含封面、目录、正文及附录）。封面或目录后的第 1-2 页必须为“**项目核心亮点预览页**”，集中展示作品相对于赛题基础任务和中级任务要求所实现的核心能力，包括但不限于：系统整体架构图、多模态数据融合处理链路、核心创新点摘要、已攻克的中高级任务清单及实现效果关键数据（如 NL2SQL 执行准确

率、RAG 检索命中率、澄清交互成功率等）。

(2) 技术文档——技术实现说明书：格式要求同上，总篇幅建议不超过 30 页。

内容应系统阐述以下模块：

- ❖ 系统总体架构设计
- ❖ NL2SQL 模块技术实现（包含 Schema Linking 方案、要素识别模型、SQL 生成与校验机制）
- ❖ RAG 问答模块技术实现（包含文档解析与切片策略、向量检索方案、生成融合机制）
- ❖ 多源异构数据融合推理策略（结构化数据库与非结构化知识库的协同查询与结果整合）
- ❖ 可解释性实现方案（包含推理链路追踪、数据来源标注、工具调用链可视化）
- ❖ 中级任务攻克方案详述
- ❖ 系统测试与性能评估（包含测试数据集说明、评估指标定义、测试结果及分析）

(3) 技术数据：作品原始文件（ZIP 数据包），应包含：

- ❖ 完整项目源代码（需包含清晰的目录结构说明和 README 文件）
- ❖ 数据库 schema 定义文件
- ❖ 知识库文档样本（不少于 10 份，涵盖不同类型文档格式）
- ❖ 模型权重文件（若适用）
- ❖ 运行环境配置说明（包括依赖包列表及版本号）
- ❖ 可执行 Demo 的快速启动脚本

2. 校内初赛评分标准（总分 100 分）

大项	内容	分值	评分要求
功能指标	1. 基础 NL2SQL 单表查询能力：系统能正确解析用户的简单查询意图并生成可执行的 SQL 语句	10	采用至少 10 个测试用例进行客观评估，每个正确执行得 1 分
	2. 基础 RAG 单文档问答能力：系统能从给定文档中正确检索并生成答案	10	采用至少 10 个测试用例进行客观评估，每个正确回答得 1 分
	3. 答案可解释性实现：系统能展示 SQL 语句、检索片段或推理链路	5	具备可解释性展示得 3 分，展示质量高得 5 分
	4. 交互 Demo 可用性：系统可通过界面或 API 进行交互	5	基础交互可用得 3 分，体验良好得 5 分
性能指标	1. NL2SQL 执行准确率：在给定测试集上的查询结果与预期结果匹配度	10	准确率 $\geq 70\%$ 得 6 分， $\geq 80\%$ 得 8 分， $\geq 90\%$ 得 10 分
	2. RAG 答案忠实度：生成答案与检索内容的相关性和事实准确性	10	准确率 $\geq 70\%$ 得 6 分， $\geq 80\%$ 得 8 分， $\geq 90\%$ 得 10 分
	3. 系统效率(用户输入到答案输出的端到端时延)设计证明	5	参赛队伍需在技术文档中提供系统效率分析，包括： ① 系统各模块的时间复杂度理论分析（如 NL2SQL 生成、RAG 检索与生成）； ② 在至少两种不同数据规模下（如小规模 1 千行、中规模 1 万行）的实测或预估响应性能对比； ③ 阐述瓶颈及优化策略。 内容完整、分析合理、优化策略清晰者得 5 分，有分析但深度不足得 3 分，仅简单提及或无分析得 0 分。
优化指标	1. 中级任务攻克数量与质量	15	每攻克 1 个中级任务基础得分 5 分，解决质量突出额外加分，总分不超过 15 分
	2. 系统鲁棒性：对模糊输入、错别字等非理想输入的容错能力	10	基础容错得 5 分，鲁棒性良好得 7 分，优异得 10 分
	3. 多轮对话上下文保持能力	5	具备多轮能力得 3 分，连续 5 轮以上准确得 5 分
文档质量	技术文档完整性、规范性、可读性	10	结构清晰、内容完整得 6 分，图表规范、逻辑严谨得 8 分，具有工业文档水准得 10 分

大项	内容	分值	评分要求
演示效果	线上提交作品的功能演示（可附演示视频或截图）	5	能够清晰展示核心功能得 5 分

评分说明： 校内初赛总分 100 分。参赛作品必须至少完成 1 个中级任务方可视为有效参赛作品。评委将根据作品材料质量与功能实现程度综合评定。初赛排名前列的队伍晋级决赛小组赛。

九、决赛小组赛作品要求及评分标准

1. 决赛小组赛作品提交要求

(1) **答辩 ppt：**凝练作品内容、方案、亮点、创新点，并阐述作品相较于基础任务要求的超越与突破。总页数建议不超过 20 页。PPT 应包含但不限于以下内容：

- ❖ 赛题理解与需求分析
- ❖ 系统总体架构与核心技术方案
- ❖ 中级任务攻克详情（重点呈现）
- ❖ 高级任务创新点阐述（若有）
- ❖ 系统测试结果与性能评估
- ❖ 总结与展望

注意： 为确保比赛公平公正，PPT 中不能出现队伍单位、人名等识别信息。

答辩 PPT 模板由主办单位提供，请关注组委会和赛题交流群通知。

(2) **技术文档（决赛版）：**提交 Word 版本及 PDF 版本各 1 份，是在校内初赛技

术文档基础上的深化版，需补充在初赛评审反馈基础上的改进内容，以及决赛现场演示的系统优化与完善。篇幅建议不超过 50 页。word 模板由主办单位提供，请关注组委会和赛题交流群通知。

(3) **技术数据（决赛版）**：作品最终原始文件（ZIP 数据包），内容要求同校内初赛，但需包含完善后的全部代码和数据。

(4) **突破与创新性自述**（仅申报高级任务加分的队伍需要提交）：单独提交一份 PDF 文档，篇幅不超过 5 页。内容须包括：

- ❖ 定位声明：申报“技术突破领先性”或“实用创新与转化价值”
- ❖ 核心贡献：用精炼的语言阐述突破点是什么
- ❖ 举证材料：与公开最优结果的详细对比实验数据、场景适配说明、成本效益分析、系统架构图等
- ❖ 评审价值说明：阐述该突破为何值得加分认可

2. 决赛小组赛阶段评分标准（总分 100 分）

大项	内容	分值	评分要求
功能指标	1. 多源异构数据融合问数/问答能力：系统能同时处理涉及结构化数据库和非结构化知识库的联合查询需求	10	能正确回答至少 3 种跨源问题得 6 分，5 种以上得 8 分，应对复杂多跳问题得 10 分
	2. 智能意图理解与澄清：系统能识别模糊/歧义/信息缺失的输入并主动交互澄清	10	具备意图识别得 4 分，主动澄清得 7 分，澄清精准且体验好得 10 分
	3. 答案可解释性：系统完整展示推理链路、工具调用、数据来源及溯源证据	5	展示完整链路得 3 分，可视化清晰、用户友好得 5 分
性能	1. NL2SQL 复杂查询准确率（含多表关联、聚合、嵌套查询）	10	准确率≥70%得 5 分，≥80%得 7 分，≥90%得 10 分

大项指标	内容	分值	评分要求
	2. RAG 多文档/多模态问答准确率	10	准确率$\geq 70\%$得 5 分，$\geq 80\%$得 7 分，$\geq 90\%$得 10 分 参赛队伍需在技术文档或答辩 PPT 中提供以下内容： ① 复杂度分析：系统各核心链路（意图解析→数据定位→查询/检索→生成答案）的时间复杂度与空间复杂度； ② 多规模性能对比：至少在 3 种不同数据规模下（例如结构化数据：1 千行、1 万行、10 万行；非结构化文档：10 页、100 页、500 页）的实测或仿真性能数据，展示响应时间随数据量增长的变化趋势； ③ 吞吐量评估：在典型配置下（需说明软硬件环境）系统的最大并发支持能力估计； ④ 优化策略：针对瓶颈所采取的具体优化措施（如缓存、索引、查询改写、并行处理等）。 以上内容完整、逻辑严谨、有实测数据支持、优化策略有效者得 5 分；有分析但缺乏实测或数据不充分得 3 分；仅简单提及得 1 分。
	3. 系统效率设计与伸缩性证明	5	
	4. 系统鲁棒性：低质量数据（模糊扫描件、错别字、非标准格式文档）下的稳定性	5	基础鲁棒得 2 分，较强鲁棒得 4 分，极端条件下仍稳定得 5 分
	4. 系统鲁棒性：低质量数据（模糊扫描件、错别字、非标准格式文档）下的稳定性	5	
优化指标	1. 中级任务攻克广度与深度	15	攻克 3 个中级任务得 9 分，4 个得 12 分，5 个及以上得 15 分（以技术文档详述为准）
	2. 中级任务攻克质量：相对于基础方案的改进幅度与有效性	10	改进显著得 6 分，有独创性方案得 8 分，达到业界领先水平得 10 分
文档质量	技术文档完整性、规范性、图表清晰度	5	文档完整、结构清晰得 5 分
演示效果	1. 现场答辩表现：逻辑清晰、表达流畅、时间控制	5	答辩良好得 3 分，表现优秀得 5 分
	2. 系统现场演示：功能完整展示、运行流畅、应对评委提问	5	演示成功得 3 分，互动优秀得 5 分
综合评	产品设计、技术实现、团队协作、论述清晰等	5	综合评估

大项	内容	分值	评分要求
价			

3. 高级任务加分标准（20 分，额外加分）

本部分不设固定任务清单，旨在辨识并奖励在中级任务基础上展现出卓越技术突破或实用创新价值的作品。评审将基于队伍提交的《突破与创新性自述》结合系统实测与答辩表现，按照以下两个维度综合评判：

加分方向	加分范围	评判要点
技术突破领先性	0~10 分	在 NL2SQL 复杂查询准确率、多跳推理、Schema Linking、文档解析等核心难题上达到业界领先水平；提出了具有学术启发价值的创新方法；在极限挑战场景下证明了方法的有效性
实用创新与转化价值	0~10 分	提出了出奇简单、易于部署的解决方案；在工程化设计、系统架构、用户体验方面有突出贡献；作品具备清晰的商业化路径和行业应用价值

合计总分上限：**100 分（基础得分）+20 分（加分）=120 分**

评分说明： 决赛小组赛总分以 100 分计（不含加分），高级任务加分直接累加至最终成绩。获奖等次及单项奖依据总分排名确定。未完成中级任务的参赛作品不具有晋级冠军争夺赛资格。

文档版本信息：

- ❖ 发布日期：2026 年 5 月
- ❖ 文档说明：本文件为杯赛阶段作品要求及评分标准正式文档，如有调整以组委会最新通知为准。请参赛队伍关注组委会和赛题交流群获取最新动态。