

“中国电子杯”第三届高校 ICT 产教融合创新大赛

赛题五

一、赛题名称

面向行业数据的智能实体挖掘与关系建模

二、命题单位

中国软件与技术服务股份有限公司

三、所需知识背景与技术要求

- 1.掌握中文文本预处理技术，优先采用基于 Transformer 架构的中文实体识别与关系抽取技术；
- 2.熟悉开源大模型使用与提示工程，熟悉 Qwen/DeepSeek 系列开源国产大模型，掌握轻量化微调、RAG 等领域适配方法；
- 3.掌握知识图谱构建、图数据建模与图检索增强技术，推荐采用 Neo4j 完成业务场景的关系建模；
- 4.了解政府采购基础业务逻辑，清楚了解采购单位、中标供应商、投标参与方等市场主体角色定位与业务往来。

四、赛题背景

中国软件与技术服务股份有限公司（简称“公司”）作为横跨中国电子信息产业集团有限公司（CEC）先进计算板块和网信产业板块的重要企业，聚焦基础软件及党政核心应用解决方案两大板块，在服务国家网络强国和数字中国建设中发挥网信事业“国家队”、“主力军”的战略作用。业务覆盖 60%以上的中央部委，31 个省、市、自治区的 200 多个地级市，服务近千万最终用户，具备全栈数智产品能力。

随着行业数字化转型的深入，海量非结构化行业数据中蕴含着巨

大的知识价值。如何通过智能实体挖掘、关系建模与知识图谱技术，将分散的业务数据转化为结构化知识资产，进而支撑智能分析与决策，是当前数据智能领域的核心课题，也是公司推进数智化转型过程中的关键技术需求。

本赛题以政府采购招投标领域真实数据为实践场景，聚焦行业数据智能分析能力的综合考察，旨在培养学生数据工程与知识建模能力，要求参赛团队构建领域数据分析引擎，完成实体深度挖掘、关系抽取与业务场景关联建模，探索大模型时代下行业数据智能应用的技术路径与落地方案。

五、赛题描述

本题目需构建行业数据分析引擎，分为三个具体任务，分别对应引擎的两大主要功能及查询平台，竞赛题目涉及的相关术语解释见“九、参考资料”部分：

（一）任务一：实体识别与自动化提取

1.任务说明：基于赛事提供的招采标讯公告及附件文件，完成标的物关键信息的精准识别与自动化提取，需提取字段包含产品服务名称、品目、品牌（产品供应商）、规格型号、单价、数量、总价等共 7 个。

2.数据集形式：招采标讯公告为 html 文件集合；附件以 zip 压缩包提供，一篇公告对应一个 html 文件和一个附件压缩包（如有），内含 doc/docx/xlsx/pdf 等非结构化文本及图片附件。

3.任务要求：优先采用中文实体提取、基于 Transformer 架构的 NLP 技术、Qwen/DeepSeek 系列开源国产大模型的提取方案，可使用

轻量化微调、提示工程优化效果，不建议仅用规则匹配完成核心提取，确保多格式文件下的信息提取高效、准确。建议参赛团队自行对数据集进行抽样，采用算法标注人工核验的方式先形成标准验证集，用于支撑算法的精度自检。

（二）任务二：用户画像建模与关系分析

1.任务说明：基于招采标讯公告中的竞标与得分数据，提取中标、投标和参与投标的主体信息，构建采购单位-中标供应商-投标参与方关系模型，完成指定场景的用户画像、竞争合作关系分析与可视化检索：

（1）指定任意采购单位，查询并输出与其存在长期/大量合作关系的中标供应商、产品供应商信息，如合作次数、交易总金额等；

（2）指定任意采购单位，查询并输出其历史招标项目中高频（TOP5）参与投标的主体，以及高频协同参与投标的主体组合；

（3）指定任意中标供应商，查询并输出其历史项目中高频（TOP5）共同竞标的主体及全部参与投标主体信息，如频次、总金额等；

（4）指定任意两个或多个中标供应商，查询并输出与上述主体均存在合作关系的采购单位、合作频次及合作金额等；

（5）指定任意两个或多个中标供应商，查询并输出其历史共同参与竞标的项目信息与竞标结果，包含项目数量、项目总金额等核心指标。

注：为了实现上述场景建模，参赛团队可能需要在任务一的数据提取环节，增加采购单位、中标供应商、参与投标人、项目名称、项目金额等额外关键字段的提取以支撑模型构建。

2.数据集形式：招采标讯公告为 html 文件集合。

3.任务要求：优先采用 Qwen/DeepSeek 系列开源国产大模型，结合信息提取、知识图谱建模、图检索增强技术，可采用轻量化微调、提示工程优化效果，保障关系分析准确、检索高效。

（三）任务三：可视化分析检索平台构建

1.任务说明：整合任务一实体提取、任务二关系建模成果，搭建一体化可视化检索平台，实现数据集自动化处理、多维度数据检索与业务场景可视化查询，完成全流程分析能力落地。

2.平台支持的数据输入形式：复用任务一、任务二的官方数据集格式，支持批量上传同规范标准的数据集进行自动化处理。

3.任务要求：对接任务一提取结果，实现数据集上传后自动解析、实体抽取与结构化入库，提供标的物全字段检索、筛选与查看能力；对接任务二关系模型，实现五大业务场景的可视化查询、统计与结果展示；界面简洁可用、响应高效，推荐前端采用 Vue 架构，后端采用 Python 技术栈，鼓励创新交互设计。

六、作品提交要求

1.汇报 PPT：完整演示 PPT，需设置核心内容预览页，重点呈现功能实现、技术创新、性能测试与成果亮点；

2.汇报视频：1080P 横屏 MP4 格式视频，时长≤10 分钟，用于讲解 PPT 并展示作品核心价值；

3.系统演示视频：1080P 横屏 MP4 格式视频，时长≤5 分钟，完整演示任务三平台中，有关任务一、二要求的全功能操作流程；

4.系统设计文档：同步提交 Word 与 PDF 版本，含核心亮点预览

页，包含业务需求分析、技术选型依据、数据模型架构、功能详细设计、功能及测试效果五大模块：

（1）业务需求分析：详细定义需要实现的功能理解以及需要达到的业务指标和性能指标；

（2）技术选型依据：阐述相关方法的技术选型依据，实现原理和技术创新型等；

（3）数据模型架构：介绍数据模型的构成、模块划分以及数据处理建模流程；

（4）功能详细设计：业务流程图，输入输出及关键功能的实现原理；

（5）功能及测试效果：各项任务的业务功能完成情况，成果展示，以及准确性测试结果。

5.过程性文档：同步提交 Word 与 PDF 版本，含技术路线对比分析报告和数据质量分析报告：

（1）技术路线对比分析报告：对比至少 2 种不同的实体提取方案，说明为什么选最终方案，以及调优过程中的关键尝试与效果数据；

（2）数据质量分析报告：如脏数据类型、缺失值分布、附件解析难点等，可附迭代的提取效果对比表。

6.源码及配套资料：提交一份打包的压缩包文件，内容包括以下部分：

（1）提交工程文件，标注软件版本和运行环境，并在关键部分进行注释；

（2）官方数据集的提取结果，包括任务一提取的数据表，任务

二构建的数据模型；

（3）必要的说明文档，包括源码的功能说明、执行步骤等。

7.有利于展示作品的其他资料。

以上为对应作品提交内容要求细则，各阶段提交作品要求详见“七、评价指标”章节。

七、评价指标

本赛题为开放技术场景构建类赛题，鼓励参赛团队在基于国产大模型的技术框架下，采用多样化的实现路径完成实体提取与关系建模。为充分发掘创新方案、提升入选提取建模技术路线的丰富度。

校内赛、决赛小组赛评分框架一致，建议在校内赛阶段，各参赛高校参考决赛评分要求，以提取率、准确性等效果指标作为核心衡量因素，以便更多具备功能实现能力的团队脱颖而出。

进入决赛阶段后，建议各团队在保障效果达标的基础上，进一步关注平衡准确度和处理效率，以及查询响应效率等性能优化，通过工程化手段提升作品的实用化水平。

（一）校内赛阶段（参考建议）

1.评分要求

大项	内容	分值	评分要求
技术指标-任务一：实体识别与自动化提取（40分）			
功能指标	覆盖率	10	标的物提取覆盖公告原文和附件两类，每实现一类得5分，全实现得10分。
	提取率	10	7个关键字段全提取得10分；5≤字段数<7，得5分；3≤字段数<5，得3分；字段数<3不得分。

性能指标	提取准确性	15	按准确率/精确率/召回率综合排名赋分：前 15%得 15 分；16%-35%得 12 分；36%-60%得 10 分；61%-85%得 5 分；其余得 2 分。
	处理速率	5	按单篇公告（含附件）处理平均速度排名赋分：前 15%得 5 分；16%-35%得 4 分；36%-60%得 3 分；61%-85%得 2 分；其余得 1 分。时间越短排名越靠前。
技术指标-任务二：用户画像建模与关系分析（25 分）			
功能指标	业务场景达成率	25	5 个核心业务场景，每实现 1 个得 5 分，全实现得 25 分。
技术指标-任务三：可视化分析检索平台构建（25 分）			
功能指标	平台功能完整性	15	支持任务一和任务二要求功能的自动处理和检索，每完整实现一项得 7.5 分，全实现得 15 分。
	技术栈合规性	5	采用 Vue+Python 技术栈得 5 分，未按推荐栈但功能完整得 3 分，未按推荐且功能不完整不得分。
性能指标	查询效率	5	单次查询操作响应时间 < 1 秒，得 5 分；响应时间 ≥ 1 秒，< 3 秒，得 3 分；响应时间 ≥ 3 秒，< 5 秒，得 1 分；响应时间 ≥ 5 秒，不得分。
综合指标（10 分）			
过程性材料	过程性文档质量	2	过程性文档规范完整、内容详实、逻辑清晰得 2 分；存在内容缺失/结构混乱酌情扣分。
文档质量	设计文档质量	3	设计文档规范完整、内容详实、逻辑清晰得 3 分；存在内容缺失/结构混乱酌情扣分。
演示效果	作品汇报演示质量	5	汇报逻辑清晰、演示流畅完整、交互逻辑人性化、突出核心成果得 5 分；表达/演示存在问题酌情扣分。

2.提交作品内容

（1）技术文档--设计报告：提交 Word 版本及 PDF 版本各 1 份，

建议封面或目录后的第 1-2 页项目为重要内容预览页，包含作品重要核心亮点。具体内容建议围绕包含业务需求分析、技术选型依据、数据模型架构、功能详细设计、功能及测试效果等进行叙述；

（2）汇报 PPT 及视频：提交完整演示汇报 PPT，需设置核心内容预览页，重点呈现功能实现、技术创新、性能测试与成果亮点。同步提交汇报 PPT 讲解及系统演示视频；

（3）源码及配套资料：提交一份打包的压缩包文件，内容包括：提交工程文件，标注软件版本和运行环境，并在关键部分进行注释；官方数据集的提取结果，包括任务一提取的数据表，任务二构建的数据模型；必要的说明文档，包括源码的功能说明、执行步骤等。

（4）过程性文档：同步提交 Word 与 PDF 版本，含技术路线对比分析报告和数据质量分析报告等。

（二）决赛小组赛阶段

1.评分要求

大项	内容	分值	评分要求
技术指标-任务一：实体识别与自动化提取（40 分）			
功能指标	覆盖率	10	标的物提取覆盖公告原文和附件两类，每实现一类得 5 分，全实现得 10 分。
	提取率	10	7 个关键字段全提取得 10 分；5≤字段数<7，得 5 分；3≤字段数<5，得 3 分；字段数<3 不得分。
性能指标	提取准确性	10	按准确率/精确率/召回率综合排名赋分：前 15%得 10 分；16%-35%得 8 分；36%-60%得 6 分；61%-85%得 4 分;其余得 2 分。
	处理速率	10	按单篇公告（含附件）处理平均速度排名赋分：前 15%得 10

			分; 16%-35%得 8 分; 36%-60%得 6 分; 61%-85%得 4 分; 其余得 2 分。时间越短排名越靠前。
技术指标-任务二：用户画像建模与关系分析（20 分）			
功能指标	业务场景达成率	20	5 个核心业务场景，每实现 1 个得 4 分，全实现得 20 分。
技术指标-任务三：可视化分析检索平台构建（20 分）			
功能指标	平台功能完整性	5	支持任务一和任务二要求功能的自动处理和检索，每完整实现一项得 2.5 分，全实现得 5 分。
	技术栈合规性	5	采用 Vue+Python 技术栈得 5 分，未按推荐栈但功能完整得 3 分，未按推荐且功能不完整不得分。
性能指标	查询效率	10	单次查询操作响应时间 < 1 秒，得 10 分；响应时间 ≥ 1 秒，< 3 秒，得 7 分；响应时间 ≥ 3 秒，< 5 秒，得 3 分；响应时间 ≥ 5 秒，不得分。
综合指标（20 分）			
过程性材料	过程性文档质量	5	过程性文档规范完整、内容详实、逻辑清晰得 5 分；存在内容缺失/结构混乱酌情扣分。
文档质量	设计文档质量	5	设计文档规范完整、内容详实、逻辑清晰得 5 分；存在内容缺失/结构混乱酌情扣分。
演示效果	作品汇报演示质量	10	汇报逻辑清晰、演示流畅完整、交互逻辑人性化、突出核心成果得 10 分；表达/演示存在问题酌情扣分。

2.提交作品内容

（1）答辩 PPT：提交完整答辩 PPT，需设置核心内容预览页，重点呈现功能实现、技术创新、性能测试与成果亮点；

注意：为确保比赛公平公正，PPT 中不能出现特殊识别信息。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）系统演示：现场演示或采用录屏形式。视频要求提交 1080P

横屏 MP4 格式视频，时长 ≤ 5 分钟，完整演示任务三平台中，有关任务一、二要求的全功能操作流程；

(3) 系统设计文档：同步提交 Word 与 PDF 版本，含核心亮点预览页，包含业务需求分析、技术选型依据、数据模型架构、功能详细设计、功能及测试效果五大模块，word 模板由主办单位提供，请关注组委会和赛题交流群通知；

(4) 源码及配套资料：提交一份打包的压缩包文件，内容包括：提交工程文件，标注软件版本和运行环境，并在关键部分进行注释；官方数据集的提取结果，包括任务一提取的数据表，任务二构建的数据模型；必要的说明文档，包括源码的功能说明、执行步骤等。

(5) 过程性文档：同步提交 Word 与 PDF 版本，含技术路线对比分析报告和数据质量分析报告等。

(三) 评分细则

1. 提取准确性计算方式

提取准确性 = 准确率 $\times 0.4$ + 精确率 $\times 0.3$ + 召回率 $\times 0.3$

2. 处理速率的测量定义

处理速率 = 系统完成全部公告解析与实体抽取的总耗时（含附件处理时间）/ 基准数据集公告总篇数

总耗时以命题方统一平台环境计时为准，从数据集导入完成时刻起算，至全部提取结果写入数据库止；

3. 查询效率的测量定义

查询效率以任务三平台中，任务二五大业务场景各执行一次查询的平均响应时间计；

响应时间=前端发起请求至结果完全渲染完毕的总时长。

4.排名赋分边界处理规则

当多个团队综合得分完全一致且恰好处于排名比例边界(如第15%分位)时,均按较高分段赋分(即“就高不就低”原则),以鼓励竞争积极性。

5.业务场景达成率的测量定义

任务二中业务场景达成率测定方式为业务场景查询结果与官方提供的基准数据进行比对,结果不一致的场景视为未实现。

6.开源模型使用要求

赛题要求使用 Qwen/DeepSeek 系列开源国产大模型,采用基座模型+外挂调优方案,应使用赛事组提供的标准模型资源,并声明模型版本和参数规模;采用模型训调方案,应使用符合赛事要求的基座模型,并出具相关佐证材料。其他模型方案视为不满足赛事要求。

7.诚信参赛要求

参赛团队须承诺作品为独立完成,不得委托外部机构代做或直接使用商业化招投标系统相关源码,一经发现取消比赛资格。

八、参赛平台

任务基准数据集在赛题发布后的10个工作日内由命题单位统一发放,内含待提取的招标公告及相关附件约1000条;评分基准数据规模约100条,内容结构同任务基准数据集,但不对外发放,仅用于评分。

统一平台运行环境由命题方/赛事主办方在赛事评比前10个工作日提供,包含云环境和包含一定额度通用大模型资源的API Key(使

用预训练、微调等模型的团队需要自行筹备相关资源), 参赛团队可在此部署系统并开展场景 POC 验证。

相关指标评审基于平台内运行产生的结果, 作为成果准确度、处理速率评测的基准依据, 参赛团队需要提供评审方登录系统的操作说明及账号, 用于评审方操作系统验证功能及使用基准数据测试性能指标, 验证过程将全程录屏。

九、参考资料

(一) 公告及附件样式参考

中国政府采购网: <https://www.ccgp.gov.cn/>, 点击“搜标题”进入详情页面, 每一条公告对应的 html 文件为基准数据样式。

(二) 名词及术语解释

1. 市场主体类

招标单位/采购单位: 依法开展招标采购活动的国家机关、事业单位、团体组织或企业, 是招标项目的发起与需求方;

中标供应商: 响应招标、通过评审并最终被确定为中标人的法人、其他组织, 是项目的成交服务/供货方, 若项目涉及联合体投标、分标段招标等, 则中标供应商可能有多个;

投标参与方: 响应招标公告、提交投标文件并参与竞争的法人或其他组织, 本题目中特指参与投标但未中标的主体;

产品供应商: 为招标项目提供硬件/软件产品的生产厂商或授权经销商, 属于标的物的品牌提供方。

2. 业务对象类

招采标讯: 中标公示、结果公示等公开信息的统称, 是本次赛题

的核心数据来源；

标的物：招标采购的对象，含货物（软硬件产品）、工程、服务三类，是本次实体提取的核心目标统称；

品目：按政府采购品目分类目录划分的采购项目类别，用于标准化归类标的物；

规格型号：标的物的技术参数、型号、配置等详细标识，用于精准界定产品/服务标准。

3.审查与流程类

资格性审查：对投标主体的资质、资格证明等进行审核，判断其是否具备投标资格；

符合性审查：对投标文件是否满足招标文件实质性要求进行审查，判断投标是否有效。

4.行为与关系类

竞标主体：同一项目中参与投标竞争的所有投标人；

协同/协作投标：多家投标人在同一或多个项目中共同参与投标的行为；

合作关系：采购单位与中标供应商之间形成的项目成交与履约关系。

5.数据指标类

中标金额：中标供应商最终成交的项目合同金额；

合作频次：特定主体之间在历史项目中发生合作或共同参与竞标的次数。

（三）提取示例

以中国政府采购网-广州市增城区中心医院二期信息化建设项目
(CZ2026-0696)(采购包1)结果公告为例:

1.抬头和“公告概要”子页面可提取到的数据字段

广州市增城区中心医院二期信息化建设项目 (CZ2026-0696) (采购包1) 结果公告

2026年07月31日 15:20 来源: 中国政府采购网广东分网 【打印】 [【显示公告概要】](#)

一、项目编号: CZ2026-0696

二、项目名称: 广州市增城区中心医院二期信息化建设项目

三、采购结果

合同包1(软件集成实施服务):

供应商名称	供应商地址	中标(成交)金额
联通数智医疗科技有限公司	广州市海珠区国际生物岛合景科盛广场	26,680,000.00元

四、主要标的信息

合同包1(软件集成实施服务):

货物类 (联通数智医疗科技有限公司)

广州市增城区中心医院二期信息化建设项目（CZ2026-0696）（采购包1）结果公告

2026年07月31日 15:20 来源：中国政府采购网广东分网 【打印】 [【显示公告正文】](#)

公告概要：

公告信息:			
采购项目名称	广州市增城区中心医院二期信息化建设项目		
品目			
采购单位	广州市增城区中心医院		
行政区域	广州市	公告时间	2026年07月31日 15:20
评审专家名单	梁汝雯,文海荣,袁野,黄鸿猷,张树霞,张庆亚,陈剑		
总中标金额	¥ 2668.000000 万元（人民币）		
联系人及联系方式:			
项目联系人	王立尹、林子程		
项目联系电话	020-26215526、020-28866416		
采购单位	广州市增城区中心医院		
采购单位地址	广州市增城区宁西街创新大道28号		
采购单位联系方式	020-62707061		
代理机构名称	广州市政府采购中心		
代理机构地址	广东省广州市天河区天润路445号		
代理机构联系方式	020-26215526、020-28866163		
附件:			
附件1	合同包1: 本国产品声明函或相关证明材料 (联通数智医疗科技有限公司) .pdf		
附件2	合同包1: 报价明细附件 (联通数智医疗科技有限公司) .pdf		

可提取到的字段为项目名称、采购单位、项目金额（总中标/成交金额）、中标供应商。

2.内容可提取到的评分项数据字段

四、主要标的信息

合同包1(软件集成实施服务):

货物类 (联通数智医疗科技有限公司)

品 目 号	品目名称	采购标的	品牌	规格型号	数量 (单位)	单价(元)	总价(元)
1-2	信息化设备	自助综合机 (多功能自助终端及系统配套设备)	长城医疗	MBS2000K-3	1.0000(批)	1,336,000.0000	1,336,000.0000
1-3	信息化设备	自助报到机 (多功能自助终端及系统配套设备)	长城医疗	MBS516K-8	1.0000(批)	271,000.0000	271,000.0000
1-4	信息化设备	自助清单机 (多功能自助终端及系统配套设备)	长城医疗	MBS909	1.0000(批)	116,000.0000	116,000.0000

可提取到产品服务名称（采购标的）、品目、品牌、规格型号、单价、数量、总价。附件的情况类似，不再额外说明。

3.内容可提取到的建模数据字段（部分）

八、其他补充事宜

合同包1(软件集成实施服务):

供应商	资格 性审 查	符合 性审 查	技术得 分	商务得 分	综合诚 信得分	价格得 分	综合得 分	得分 排名	推荐 排名
联通数智医疗科技有限公司	通过	通过	48.71	30.00	4.59	10.22	93.52	1	1
中国移动通信集团广东有限公司	通过	通过	32.93	28.00	4.78	10.20	75.91	2	2
中国电信股份有限公司广东分公司	通过	通过	29.50	28.50	4.79	10.22	73.01	3	
广州云奕技术有限公司	通过	通过	30.07	5.00	4.54	15.00	54.61	4	
广州天锐医健信息科技有限公司	通过	通过	29.57	8.00	4.68	10.20	52.45	5	

1.公共资源交易服务费的支付及发票开具

根据该数据，图中所有供应商组成的集合统称为竞标主体，可提取到本项目的中标供应商为“联通数智医疗科技有限公司”，“中国电信股份有限公司广东分公司、广州云奕技术有限公司”等为本项目的参与投标方（参与但未中标），所有参与投标方均为中标供应商的潜在协作投标单位。