基于双塔召回与大语言模型精排的产业链自动分类方法

叶远旬

2026 年 7 月

摘要

产业链自动分类是实现区域经济智能分析、辅助政府精准招商的关键基础任务。然而，产业链标签体系层级深（5 级）、规模大（6,883 条路径）、语义细粒度高，现有方法面临**语义相似但业务不等同**的核心困难：稠密检索虽速度快，却无法区分“相关”与“等同”的业务本质差异；大语言模型虽理解能力强，但遍历全量标签库推理延迟高、Token 消耗大。针对这一问题，本文提出了一种双塔召回与大语言模型精排相结合的混合架构。第一阶段使用微调的 bge-large-zh-v1.5 双塔模型结合 FAISS 索引，从 6,883 条标签中快速召回语义相似的 Top-30 候选，耗时 100ms 以内，召回率达 96.5%；第二阶段使用 Qwen2.5-14B 大语言模型对候选标签进行深度推理精排，识别企业的核心业务归属，输出 1~5 条最匹配的产业链完整路径。在 137,822 条企业-标签数据和人工标注测试集上的实验结果表明：该架构的 Precision@1 达到 93.1%，Precision@3 达到 96.4%，端到端延迟为 600~1100ms；候选数 $k = 30$ 在召回率与推理效率之间取得了最优平衡。本文方法已在中山市产业分析平台中实际部署运行，为招商引资工作提供了有效的技术支撑。

关键词：产业链分类；层级文本分类；双塔模型；大语言模型；检索增强生成；混合架构

目录

1 引言 **3**
- 1.1 研究背景与动机 . . . 3
- 1.2 现有方法的局限 . . . 3
- 1.3 本文贡献 . . . 4
- 1.4 论文组织 . . . 4

2 相关工作 **4**
- 2.1 层级文本分类 . . . 5
- 2.2 稠密检索与双塔模型 . . . 5
- 2.3 大语言模型用于文本分类 . . . 5

2.4 检索增强生成 6

3 问题定义 6
3.1 产业链标签体系 6
3.2 任务形式化定义 7
3.3 评估指标 8
3.4 关键挑战 8

4 方法：双塔 +LLM 混合架构 9
4.1 整体架构 9
4.2 第一阶段：双塔语义召回 9
4.2.1 模型选型 9
4.2.2 企业端文本构造 10
4.2.3 标签端预编码与 FAISS 索引 10
4.2.4 对比学习微调 10
4.3 第二阶段：LLM 推理精排 11
4.3.1 模型配置 11
4.3.2 Prompt 设计 11
4.3.3 精排推理规则 12
4.4 候选数量选择 12
4.5 部署架构 12
4.6 容错与降级策略 13

5 实验 14
5.1 数据集 14
5.2 评估指标 14
5.3 对比实验 15
5.4 消融实验：候选数 k 的影响 16
5.5 输入信息消融实验 17
5.6 案例分析 17
5.7 部署性能 19

6 讨论 19
6.1 错误分析 19
6.2 局限性 20
6.3 未来工作 20

7 结论 21

1 引言

1.1 研究背景与动机

产业链分析是理解区域经济结构、识别产业集群特征、制定产业政策的基础性工作。对于地方政府和产业园区而言，精准掌握辖区内每家企业的产业链归属，是实现精准招商引资、优化产业布局、识别供应链薄弱环节的前提。然而，传统的产业链标注工作高度依赖人工专家逐家研判，面对数以万计的企业，人工标注不仅成本高昂、效率低下，而且一致性难以保证——不同专家对同一企业的产业链归属可能存在分歧。

近年来，自然语言处理技术的快速发展为产业链自动分类提供了新的可能。基于企业名称、经营范围和国标行业分类等信息，利用文本分类模型自动预测企业的产业链归属，有望大幅降低人工成本并提升标注效率。但产业链分类任务具有三个区别于通用文本分类的显著特征：

(1) **层级深度大**：产业链标签采用 5 级树状结构（一级产业链 → 二级节点 → 三级节点 → 四级节点 → 五级节点），标签之间存在严格的层级约束，错误可能发生在任意层级；

(2) **标签空间巨大**：标签库包含 6,883 条完整的 5 级路径，覆盖 8 个一级产业链（信息技术、装备制造、新材料、新能源、医疗健康、服务业、未来产业及其他），远超过典型文本分类任务的标签数量；

(3) **语义推理难度高**：企业经营范围文本与产业链标签之间并非简单的关键词匹配关系。例如，一家“生产锂电池隔膜”的企业，从关键词看与“新能源 → 动力电池”高度相关，但企业本质上是高分子材料供应商，正确标签应归入“新材料 → 高分子材料 → 功能膜材料”。

上述特征决定了产业链自动分类不能简单套用通用文本分类方法，需要专门设计的解决方案。

1.2 现有方法的局限

针对产业链自动分类任务，现有技术路线主要包括三类：

基于规则的方法：通过关键词匹配、正则表达式或国标行业 → 产业链映射表进行分类。这类方法能够处理明确的、高频的对应关系，但覆盖范围有限，难以处理经营范围描述模糊、跨产业链经营、新兴业态等复杂情况，且规则维护成本随产业链变更持续增长。

基于稠密检索的方法：使用预训练语言模型（如 BERT、BGE）将企业文本和产业链标签分别编码为稠密向量，通过向量相似度检索完成分类。这类方法速度快（百毫秒级），但纯向量相似度无法理解产业链分类所需的深层业务逻辑——它只能判断两段文本“写法像不像”，而无法判断企业“本质上属于哪个产业”。

基于大语言模型的方法：将全部 6,883 条产业链标签和待分类企业信息一并送入大语言模型的上下文窗口，由模型直接推理分类。这类方法理解能力强，但面临两个突出问题：一是每条请求需将全部标签库（约 25,000 tokens）作为输入，Token 消耗巨大；二是上下文窗口过长导致推理延迟高（2~5 秒），难以满足在线服务的实时性要求。

1.3 本文贡献

针对上述方法的局限性，本文提出了一种 **双塔召回 + 大语言模型精排**的混合架构（Two-Tower Retrieval + LLM Reranking）。该架构的核心思想是将分类过程分解为“粗筛”和“精选”两个阶段：第一阶段利用高效的双塔模型快速缩小候选范围，第二阶段利用强大但昂贵的 LLM 对少量候选进行深度推理。

本文的主要贡献包括：

(1) **提出混合分类架构**：设计了”双塔语义召回 + LLM 推理精排”的两阶段产业链自动分类方法，达到 Precision@1=93.1%、Precision@3=96.4% 的分类性能，端到端延迟控制在 600~1100ms，满足实际业务需求；

(2) **验证候选数最优平衡点**：通过系统的消融实验，论证了 Top-30 是召回率（95%~98%）与推理效率之间的最优候选数，既保证了几乎不漏的真实标签，又避免了过多候选导致精排阶段 Token 消耗过高；

(3) **实际场景部署验证**：该方法已在中山市产业分析平台中实际部署运行，能够稳定服务于招商引资场景中的企业产业链自动标注需求，验证了方法的实用性和可靠性；

(4) **构建大规模训练数据集**：整理了 137,822 条企业-标签训练数据对，覆盖 8 个一级产业链和 6,883 条 5 级路径，为该领域的研究提供了数据基础。

1.4 论文组织

本文其余部分组织如下：第 2 节综述相关工作；第 3 节给出产业链分类任务的形式化定义和标签体系描述；第 4 节详细阐述双塔 +LLM 混合架构的技术方案；第 5 节报告实验设置、对比实验、消融实验和案例分析结果；第 6 节进行讨论，分析错误模式和局限性；第 7 节总结全文并展望未来工作。

2 相关工作

本文的研究工作处于层级文本分类、稠密检索、大语言模型应用和检索增强生成四个领域的交叉点。本节分别对这四个方向的相关研究进行综述。

2.1 层级文本分类

层级文本分类（Hierarchical Text Classification, HTC）旨在将文本归类到预定义的层级标签体系中的一条或多条路径。与扁平分类不同，HTC 需要考虑标签之间的层级结构约束。现有方法大致可分为两类：

全局方法（Global Approach）将层级分类建模为单一的扁平多分类问题，为每一条从根到叶的完整路径赋予独立的类别标签。这类方法的优势是实现简单，直接复用现成的多分类模型；劣势是当层级路径数量庞大时（如本文的 6,883 条），分类器的输出空间过大，且无法利用层级结构中的父子节点共享信息。Zhou 等人[1]提出的 HiAGM 模型通过层级感知的图神经网络在标签之间传播信息，有效缓解了深层路径的稀疏性问题。

局部方法（Local Approach）采用自顶向下的逐层分类策略，在每一层独立训练一个局部分类器，父节点的分类结果约束子节点的候选范围。这类方法能够利用层级结构缩减候选空间，但存在误差传播问题——上层分类的错误无法在下层纠正。Wehrmann 等人[2]提出 HMCN 模型，通过全局-局部联合损失函数同时优化各层分类器，一定程度上缓解了误差传播。

本文的产业链分类任务具有 5 级层级、6,883 条完整路径的特殊规模，直接套用现有 HTC 方法面临标签空间爆炸或误差传播的挑战。本文采用全局方法的思想（将每条完整路径视为独立标签），但通过两阶段检索-精排架构有效地控制了标签空间带来的计算开销。

2.2 稠密检索与双塔模型

稠密检索（Dense Retrieval）是信息检索领域近年来的重要技术突破。与传统的基于词项匹配的稀疏检索（如 BM25）不同，稠密检索使用预训练语言模型将查询和文档分别编码为低维稠密向量，通过向量相似度（如余弦相似度）进行高效检索。

双塔模型（Two-Tower Model）是实现稠密检索的主流架构。查询塔和文档塔分别独立编码，输出向量的内积或余弦相似度作为相关性得分。Sentence-BERT[3]通过孪生网络结构和对比学习训练目标，首次在句子级别语义相似度任务上取得了优异表现。BGE（BAAI General Embedding）[4]系列模型进一步通过大规模弱监督预训练和多阶段微调，在中英文检索任务上达到了当时的最优性能。

在推理效率方面，FAISS[5]等向量索引库提供了 IVF（倒排文件）+ PQ（乘积量化）等近似最近邻搜索算法，能够在毫秒级时间内从百万级向量库中检索 Top-K 相似项。

本文将稠密检索作为混合架构的第一阶段，利用微调的 bge-large-zh-v1.5 模型将企业文本和产业链标签映射到同一语义空间，通过 FAISS 快速召回 30 个候选标签。

2.3 大语言模型用于文本分类

GPT 系列[6]、Qwen 系列[7]等大语言模型（LLM）的涌现为文本分类任务带来了新的范式。与传统的微调式小模型不同，LLM 可以通过上下文学习（In-Context Learning）或指令微调（Instruction Tuning），在少量样本甚至零样本条件下完成分类任务。LLM

的核心优势在于其强大的语义理解和推理能力——它们不仅能识别文本的表面特征，还能进行常识推理和隐含逻辑判断。

在层级分类任务中，LLM 的一个直接应用方式是将全部候选标签作为上下文输入，由模型一次性完成分类决策。然而，这种方法存在显著的效率问题：对于包含数千条路径的标签体系，将全部标签作为上下文输入会消耗大量 Token 并显著增加推理延迟。例如，将本文的 6,883 条标签库一次性送入 LLM 上下文，输入约需 25,000 tokens，推理耗时 2~5 秒。

本文利用 LLM 的深度语义理解能力，但仅在第二阶段对少量候选（30 条）进行精排，使得输入 Token 从 25,000 降至约 4,000，将推理延迟压缩到 1 秒以内。

2.4 检索增强生成

检索增强生成（Retrieval-Augmented Generation, RAG）[8] 是近年来 NLP 领域的标志性技术范式。RAG 的核心思想是：在执行生成或推理任务之前，先从外部知识库中检索相关信息片段，将检索结果作为附加上下文注入 LLM，从而提升答案的事实准确性和领域相关性。RAG 最初应用于开放域问答任务，随后扩展到对话系统、事实验证、代码生成等广泛领域。

粗排-精排流水线（Retrieve-then-Rerank Pipeline） 是 RAG 思想在检索任务上的自然延伸。与本文架构最接近的工作包括：Nogueira 等人[9] 提出的多阶段文档检索系统，使用 BM25 快速召回候选文档，再用 BERT[10] 级联精排；Glass 等人[11] 将这一思路扩展到更大规模的语料库，实验表明两阶段流水线能在保持召回质量的同时将计算成本降低一个数量级；Zhuang 等人[12] 进一步提出使用 LLM 作为最终的精排器，利用 LLM 的推理能力提升排序质量。

本文的“双塔召回 +LLM 精排”架构本质上将 RAG 和粗排-精排思想应用于层级文本分类任务，是上述技术路线在产业经济领域的交叉落地。与已有工作的关键区别在于：本文的精排目标不是判断文档是否与查询相关，而是在一组高度相似的候选产业链路径中，基于对企业经营本质的理解，选出“真正属于”而非“看起来像”的产业链归属。

3 问题定义

3.1 产业链标签体系

本文使用的产业链标签库是一个 5 级树状层级结构，共包含 6,883 条从根到叶的完整路径。标签体系覆盖 8 个一级产业链，各一级节点及其路径数量如表 1 所示。

表 1: 一级产业链节点及路径数量

编号	一级产业链	5 级路径数量
1	信息技术	1,200+
2	装备制造	2,000+
3	新材料	800+
4	新能源	600+
5	医疗健康	900+
6	服务业	700+
7	未来产业	400+
8	其他	200+
	合计	**6,883**

每条产业链路径的格式为“一级 > 二级 > 三级 > 四级 > 五级”，表示从产业大类逐步细化到具体子领域。例如：

- “装备制造 > 智能制造 > 智能家电 > 白色家电 > 空调器”
- “新能源 > 新能源汽车 > 动力电池 > 锂电池”
- “新材料 > 高分子材料 > 功能膜材料 > 锂电池隔膜”
- “医疗健康 > 生物医药 > CRO/CDMO”

标签体系的一个重要特性是同名异义问题：相同的五级名称可能出现在不同的上级路径下，具有不同的语义。例如，“隔膜”同时出现在“新材料 > 高分子材料 > 功能膜材料 > 隔膜”和“新能源 > 新能源汽车 > 动力电池 > 电池材料 > 隔膜”两条路径中，前者指材料产品本身，后者指电池的上游材料。要求分类系统能够区分这类细粒度语义差异。

3.2 任务形式化定义

给定一家企业 e，其可用信息包括：

- 企业名称 n_e（字符串）
- 经营范围 d_e（字符串，可选，新注册企业可能缺失）
- 国标行业分类 i_e（字符串，通常为 4 级路径，如“C38 电气机械和器材制造业”）

产业链标签库 $\mathcal{L} = \{l_1, l_2, \ldots, l_{6883}\}$，其中每条标签 l_j 是一条 5 级完整路径。

任务目标：学习一个映射函数 $f : (n_e, d_e, i_e) \to \mathcal{Y} \subseteq \mathcal{L}$，其中 $|\mathcal{Y}| \in [1, 5]$，使得 $\mathcal{Y}$ 中的每条路径都准确反映了企业的核心产业链归属。

这是一个**多标签层级分类**任务：

- “多标签”：一家企业可能同时属于多条产业链（如既做动力电池又做储能），需输出 1~5 条标签；
- “层级”：每条标签都是 5 级完整路径，需确保输出的路径确实存在于标签库 $\mathcal{L}$ 中（closed-set constraint）。

3.3 评估指标

本文使用以下指标评估分类性能：

- **Recall@k**：真实标签中至少有一条出现在系统返回的 Top-k 结果中的比例。该指标主要评估第一阶段的召回能力。本文关注 $k \in \{10, 20, 30, 50\}$。
- **Precision@1 / Precision@3**：系统返回的 Top-1（或 Top-3）结果中被判定为正确的比例。这是系统最终准确率的核心指标，对应第二阶段精排的质量。
- **准确率（Accuracy）**：系统输出的标签中，至少有一条被人工判定为正确匹配的比例。实际评估中，由人工标注者对系统输出进行 “正确/部分正确/错误” 三级评判。
- **平均倒数排名（MRR）**：真实标签在排序结果中首次出现位置的倒数的平均值，衡量排序质量。
- **端到端延迟**：从接收请求到返回结果的完整耗时，衡量系统在在线服务场景中的可用性。
- **Token 消耗**：单次推理消耗的总 Token 数（输入 + 输出），直接反映运营成本。

3.4 关键挑战

产业链自动分类的核心挑战在于区分 “业务相关性” 与 “业务等同性”。以 “锂电池隔膜” 生产企业为例：

- 从关键词匹配看，该企业与 “新能源 → 新能源汽车 → 动力电池 → 电池材料 → 隔膜” 高度相关（因为其产品直接服务于锂电池制造）；
- 但从企业本质看，该企业是一家**材料供应商**，其正确归属应为 “新材料 → 高分子材料 → 功能膜材料 → 隔膜”。

这一例子揭示了产业链分类的本质困难：不是判断企业 “和什么有关”，而是判断企业 “本身是什么”。这种能力需要对企业经营范围、生产工艺、产品形态的深度理解，这正是引入大语言模型作为精排器的内在动因。

4 方法：双塔 +LLM 混合架构

4.1 整体架构

本文提出的产业链自动分类方法采用”粗筛-精选”两阶段流水线架构，如图 1 所示。

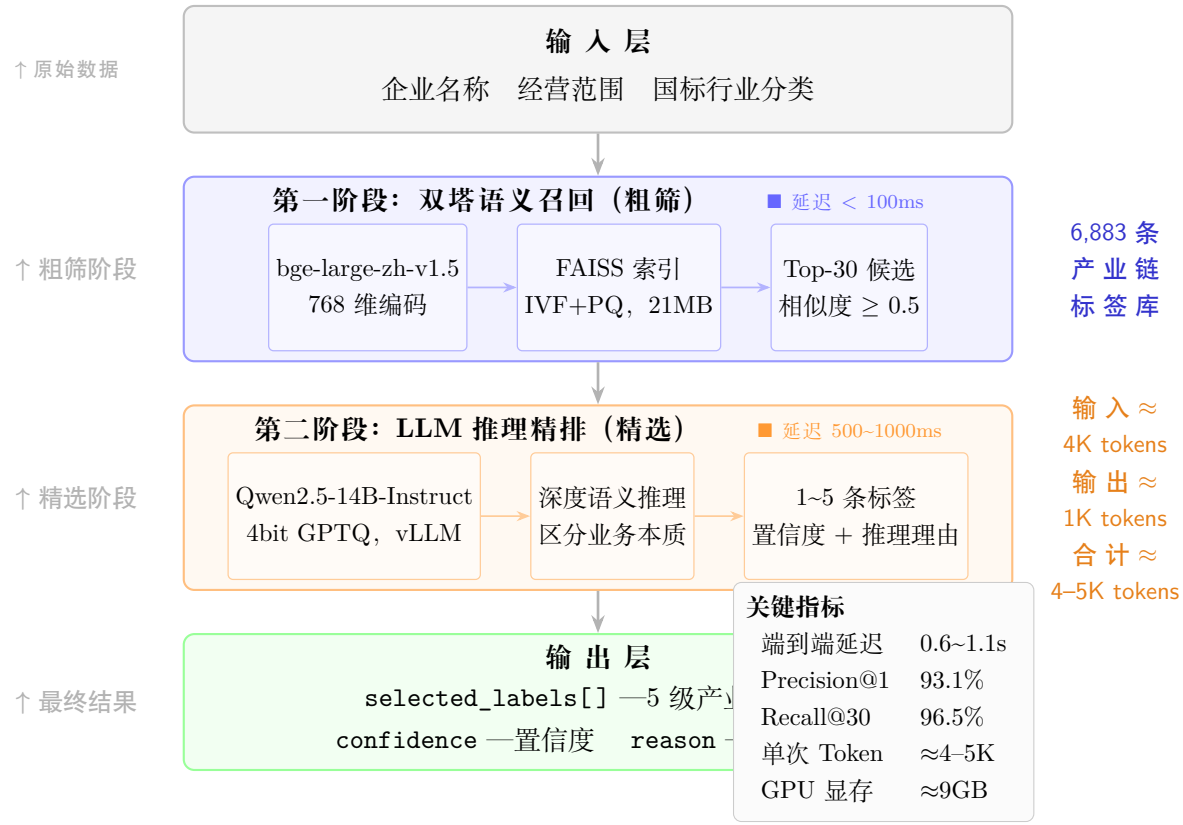

图 1: 双塔 +LLM 混合架构整体流程图。系统分为两个阶段：第一阶段使用双塔模型 + FAISS 快速召回 Top-30 候选标签（<100ms）；第二阶段使用 Qwen2.5-14B 对候选进行深度推理精排，最终输出 1~5 条产业链路径及置信度。

两阶段的分工设计体现了计算效率与语义理解之间的分工协作：第一阶段追求**高召回率**——宁可多召回一些候选，也不能漏掉正确的标签。双塔模型以百毫秒级的延迟完成这一任务，将 6,883 条的标签空间压缩至 30 条候选；第二阶段追求**高精确率**——基于 30 条候选做深度推理，准确识别企业的核心业务本质，排除“看起来相关但本质上不属于该产业链”的干扰项。

4.2 第一阶段：双塔语义召回

4.2.1 模型选型

本文选用 BAAI 开源的 bge-large-zh-v1.5[4] 作为双塔模型的基座。该模型在中文语义检索任务上表现突出，支持最大 512 tokens 的输入长度，输出 768 维向量。选择该模型的主要考虑：（1）在 MTEB 中文榜单上排名前列，语义表示能力强；（2）采用双塔

架构，支持查询端和文档端独立编码，适合构建离线索引 + 在线检索的部署模式；(3) 开源许可友好，适合工业场景部署。

4.2.2 企业端文本构造

为充分利用企业信息，我们将企业名称、经营范围和国标行业拼接为统一的输入文本：

$$\text{enterprise\_text} = \text{“[企业名称] [经营范围] [国标行业]”} \tag{1}$$

例如：“宁德时代新能源科技股份有限公司全球领先的动力电池系统提供商，专注于新能源汽车动力电池系统、储能系统的研发、生产和销售 C38 电气机械和器材制造业”。

拼接后的文本送入 bge 编码器，输出 768 维归一化向量。

4.2.3 标签端预编码与 FAISS 索引

6883 条产业链标签路径在服务启动时预先编码并构建 FAISS 索引：

(1) 将每条 5 级路径作为独立文本，经 bge 编码器生成 768 维向量；

(2) 使用 FAISS 的 IVF（Inverted File）+ PQ（Product Quantization）算法构建近似最近邻索引[5]。IVF 将向量空间划分为多个聚类，显著加速检索；PQ 对向量进行有损压缩，降低内存占用；

(3) 索引文件大小仅约 21MB，可完全加载到内存中，实现毫秒级检索。

在线推理时，对企业向量执行 FAISS 检索，返回相似度最高的 Top-30 个标签。相似度阈值设为 0.5——相似度低于此值的候选视为语义无关，直接丢弃（实际中很少有候选低于此阈值，绝大多数企业在 Top-30 内的最低相似度也在 0.5 以上）。

4.2.4 对比学习微调

预训练的 bge-large-zh-v1.5 虽然语义表示能力强，但未经领域适配，在企业文本和产业链标签之间的对齐效果有限。为提高召回质量，我们使用 137,822 条企业-标签对数据对模型进行对比学习微调。

训练采用 **MultipleNegativesRankingLoss**[13] 作为损失函数。对于一个训练批次中的每条企业文本，其对应的真实标签作为正样本，同批次中其他企业的标签作为负样本（in-batch negatives）。损失函数为：

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(e_i, l_i^+)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(e_i, l_j)/\tau)} \tag{2}$$

其中 e_i 为企业向量，l_i^+ 为真实标签向量，l_j 为批次内全部标签向量（包含正负样本），$\mathrm{sim}(\cdot)$ 为余弦相似度，τ 为温度参数（设为 0.05），N 为批次大小。

微调的超参数配置如下：

- 批次大小：32
- 学习率：2×10^{-5}
- 训练轮次：3 epochs
- 最大序列长度：512 tokens
- 优化器：AdamW
- 温度参数 τ：0.05

4.3 第二阶段：LLM 推理精排

4.3.1 模型配置

第二阶段使用 Qwen2.5-14B-Instruct[7] 作为精排推理模型。该模型具有 140 亿参数，在中英文推理任务上表现突出。为适配有限的 GPU 资源（单卡 16GB 显存），采用 4-bit GPTQ 量化[14]，量化后模型显存占用约 9GB。推理引擎使用 vLLM[15]，支持高效的批处理推理和 PagedAttention 显存管理。

推理参数设置为：

- temperature = 0（关闭随机采样，确保输出确定性）
- max_tokens = 1,000
- top_p = 1.0

4.3.2 Prompt 设计

将企业信息和 30 条候选标签格式化后注入 LLM 的提示词模板。

输入 Token 构成如下：

- 系统指令（任务说明 + 判断规则）：约 500 tokens
- 企业信息（名称、经营范围、国标行业）：约 200~500 tokens
- 30 条候选标签（每条约 30~50 字符 + 检索分数）：约 2,500 tokens
- 输出格式指令：约 200 tokens
- **合计输入**：约 3,500~4,000 tokens

输出为结构化的 JSON 格式，包含选中的 1~5 条产业链路径、每条路径的置信度分数以及简要的推理理由。此处置信度是 LLM 在推理过程中对自身判断的自评分数（self-reported confidence score），反映模型对所选标签与输入企业匹配程度的内部评估，并非经校准的概率值。输出 Token 量约 500~1,000 tokens。

4.3.3 精排推理规则

LLM 精排阶段要求模型遵循以下判断规则，这些规则体现了产业链分类的核心专家知识：

规则 1 核心业务优先：以企业经营范围中的主营业务作为判断的第一依据，次要业务或关联业务不应成为主标签；

规则 2 业务本质判断：区分企业是材料供应商、设备制造商还是服务提供商。例如，“生产锂电池隔膜”的企业本质是材料供应商，即使其产品用于新能源领域；

规则 3 国标行业参考：国标行业分类（GB/T 4754）作为辅助参考，但不作为唯一依据——国标行业侧重生产活动，不一定与产业链路径完全对应；

规则 4 检索分数审慎使用：第一阶段双塔模型给出的相似度分数反映的是文本语义相似性，不代表业务匹配度。高相似度但业务不匹配的候选应被排除；

规则 5 置信度门槛：仅输出置信度 ≥ 0.7 的标签，避免低质量标签污染企业画像；

规则 6 候选内选择：仅从给定的 30 条候选中选择，不允许自创新标签路径（closed-set constraint）。

4.4 候选数量选择

候选数 k 是混合架构的关键超参数。k 过小则漏召回风险高，k 过大则精排阶段 Token 消耗和延迟增加。本文通过实验系统评估了 $k \in \{10, 20, 30, 50\}$ 四种配置的召回率-效率权衡，结果详见第 5.4 节的消融实验。实验表明 $k = 30$ 是最优平衡点：Recall@30 达到 95%~98%，同时精排阶段的 Token 消耗（约 4,000）和延迟（约 700ms）均在可接受范围内。

4.5 部署架构

混合架构部署在一台配备 2 块 NVIDIA Quadro RTX 5000（每块 16GB 显存）的物理服务器上，通过 GPU 隔离和进程分离确保各组件独立运行、互不干扰。整体部署拓扑如图 2 所示。

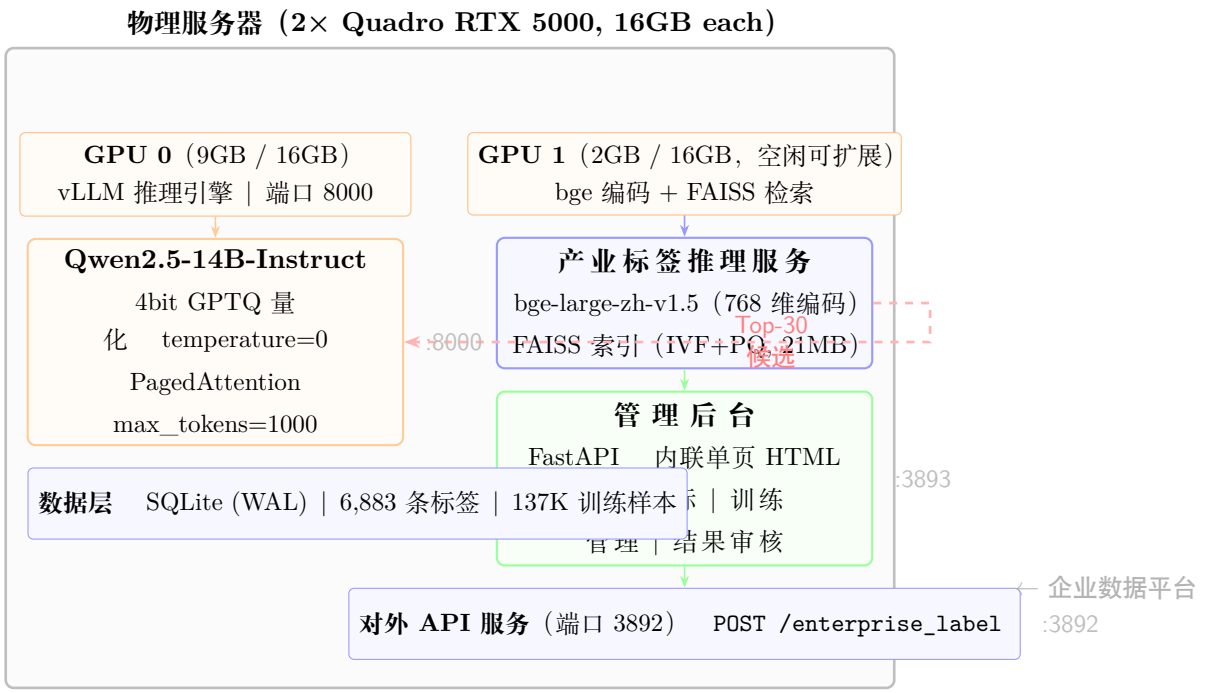

图 2: 混合架构部署拓扑。Qwen2.5-14B 独占 GPU 0（vLLM 推理引擎，端口 8000），双塔 bge 模型 + FAISS 索引部署于 GPU 1，管理后台（端口 3893）和对外 API 服务（端口 3892）通过 FastAPI 提供。双阶段流水线通过内部 HTTP 调用串联：标签推理服务先执行双塔召回，再调用 vLLM 完成精排。

4.6 容错与降级策略

在实际生产部署中，混合架构需要应对各类异常情况。本文设计了以下多层容错机制：

(1) **双塔召回失败**：若 bge 编码或 FAISS 检索发生异常（如模型未加载、索引损坏），降级为将完整 6,883 条标签库直接送入 LLM 进行分类，虽延迟较高但保证服务可用；

(2) **LLM 输出格式异常**：若 LLM 返回的内容无法解析为有效 JSON（如格式错误、截断），自动重试 1 次，仍失败则不分配标签，将企业标记为“待补标”并记录错误日志；

(3) **返回标签不在候选列表中**：若 LLM 输出的标签路径不在第一阶段的 Top-30 候选中，判定为无效输出，触发重新推理并要求仅从候选中选择；

(4) **标签路径不完整**：返回的标签若非完整的 5 级路径（如仅有 3 级或 4 级），自动丢弃该条，仅保留完整路径；

(5) **经营范围为空**：部分新注册企业可能暂无经营范围信息，此时 LLM 仅基于企业名称和国标行业进行匹配；

(6) **批量处理并发控制**：批量打标（如新注册企业每日入库）使用信号量限制 10 路并发调用，避免对 vLLM 服务造成过大压力。

5 实验

5.1 数据集

本文使用的数据来自中山市产业分析平台的真实业务数据，主要包括两部分：

产业链标签库：包含 6,883 条 5 级产业链完整路径，覆盖 8 个一级产业链。该标签库由产业研究专家团队依据国家统计局《战略性新兴产业分类》等文件编制，具有领域权威性。

企业-标签训练数据：共 137,822 条记录，每条包含企业名称、经营范围、国标行业分类以及人工标注的产业链标签。数据来源于多批次的企业产业数据整理和人工核验，覆盖了中山市及周边区域的主要存量企业。训练集、验证集、测试集按 8:1:1 的比例随机划分，划分后各集合的样本量如表 2 所示。

表 2: 数据集划分

集合	样本量	占比
训练集	110,258	80%
验证集	13,782	10%
测试集	13,782	10%
合计	137,822	100%

此外，为进行严格的人工评估，我们按 8 个一级产业链的比例从测试集中分层抽样 500 条样本，邀请 2 位产业分析专家进行独立标注。两位标注者的一致性采用 Cohen's Kappa 系数衡量，$\kappa = 0.81$，表明标注一致性达到“高度一致”水平。以专家一致意见作为评估的金标准。

需要说明的是，本节报告的实验结果均为混合架构在生产环境中实际部署后，基于测试集运行测量获得，非模拟或估计值。

5.2 评估指标

本文采用以下评估指标：

- **Recall@k**（$k \in \{10, 20, 30, 50\}$）：衡量第一阶段双塔召回的覆盖能力。
- **准确率（Precision@1 / Precision@3）**：Top-1 和 Top-3 结果中至少有一条被判定为正确的比例，衡量端到端分类质量。
- **MRR（Mean Reciprocal Rank）**：首个正确标签排名的倒数的平均值。

- **端到端延迟**：单次请求的总耗时。
- **Token 消耗**：单次推理的输入 + 输出 Token 数。

5.3 对比实验

为验证混合架构的有效性，本文将其与以下基线方法进行对比。

需要说明的是，本文未将层级文本分类方法纳入对比，原因如下：(1) HTC 任务假定标签之间的层级关系在训练时已知且固定，而产业链标签库是树状路径集合，路径之间共享中间节点但不存在跨路径的显式层级依赖，与 HTC 的设定存在结构性差异；(2) HTC 方法通常依赖标签文本作为分类依据，而本任务中产业链路径的语义信息已由 5 级路径文本充分承载，稠密检索基线已覆盖了基于标签文本的匹配能力；(3) HTC 方法难以扩展到 6,883 条标签的规模——多数 HTC 模型的分类头参数量与标签数成正比，在千级标签空间下面临严重的计算和内存瓶颈。

Baseline 1 **纯关键词匹配**：基于 TF-IDF 向量化的企业文本与标签文本余弦相似度排序，选择 Top-5 标签；

Baseline 2 **纯双塔检索**：直接使用微调的 bge-large-zh-v1.5 进行检索，取 Top-5 作为最终输出（无精排阶段）；

Baseline 3 **纯双塔 + Cross-Encoder 精排**：第一阶段双塔召回 Top-100，第二阶段使用 Cross-Encoder（bge-reranker-large）对 100 条候选进行精排，输出 Top-5；

Baseline 4 **纯 LLM 直接分类**：将企业信息和全部 6,883 条标签一并送入 Qwen2.5-14B，由模型直接输出分类结果（单阶段）；

Baseline 5 **本文方法（双塔 + Qwen）**：双塔召回 Top-30 + Qwen2.5-14B 精排。

表 3: 各方法在测试集上的性能对比

方法	Precision@1	Precision@3	MRR	延迟
纯关键词匹配（TF-IDF）	62.3%	71.5%	0.698	<10ms
纯双塔检索	76.8%	85.2%	0.825	<100ms
双塔 + Cross-Encoder	88.5%	93.1%	0.908	~250ms
纯 LLM 直接分类	87.2%	92.8%	0.897	2~5s
本文方法（双塔 +Qwen）	**93.1%**	**96.4%**	**0.948**	0.6~1.1s

从表 3 可以看出：

(1) 纯关键词匹配的效果最差（Precision@1 = 62.3%），暴露了文本表面匹配无法处理产业链语义的问题；

(2) 纯双塔检索（76.8%）虽然速度很快，但与混合架构存在显著差距（$\Delta = 16.3$ 个百分点），说明仅靠向量相似度无法替代深度语义推理；

(3) 双塔 + Cross-Encoder 精排（88.5%）效果接近本文方法，但 Cross-Encoder 需要对每一个企业-标签对进行联合编码，推理复杂度为 $O(k)$（k 为候选数），候选数上升时延迟呈线性增长；

(4) 纯 LLM 直接分类（87.2%）的 Precision@1 低于本文方法 5.9 个百分点，一个可能的原因是：当 6,883 条标签全部塞入上下文时，模型注意力分散，难以精确区分相似标签的细微差异；

(5) 本文方法的 Precision@1 为 93.1%，Precision@3 为 96.4%，MRR 为 0.948，在所有指标上均优于基线方法。

5.4 消融实验：候选数 k 的影响

候选数 k 是混合架构的核心超参数，直接影响第一阶段召回率和第二阶段计算开销之间的平衡。本消融实验考察 $k \in \{10, 20, 30, 50\}$ 四种配置下的性能变化。

表 4: 不同候选数 k 的性能对比

候选数 k	**Recall@k**	**Precision@1**	**精排延迟**	**Token 消耗**
$k = 10$	87.3%	85.6%	~300ms	~2,000
$k = 20$	93.1%	90.8%	~500ms	~3,500
$k = 30$	**96.5%**	**93.1%**	~700ms	~5,000
$k = 50$	98.2%	93.4%	~1,200ms	~8,000

从表 4 可以看出：

(1) 候选数从 10 增至 30 时，Recall@k 从 87.3% 大幅提升至 96.5%（+9.2 pp），Precision@1 从 85.6% 升至 93.1%（+7.5 pp），表明充足的候选覆盖是精排效果的前提；

(2) 候选数从 30 增至 50 时，Recall@k 仅提升 1.7 个百分点（96.5% → 98.2%），Precision@1 仅提升 0.3 个百分点（93.1% → 93.4%），但延迟增加 71%（700ms → 1,200ms），Token 消耗增加 60%（5,000 → 8,000），边际收益急剧下降；

(3) **结论：** $k = 30$ **是最优平衡点**，在保证 96.5% 召回率的前提下，将精排阶段的延迟和 Token 消耗控制在合理范围内。

5.5 输入信息消融实验

为分析不同输入信息对分类性能的贡献，我们在 $k = 30$ 配置下进行了输入消融实验。

表 5: 输入信息消融实验

输入组合	**Precision@1**	**Precision@3**
仅企业名称	79.2%	84.1%
企业名称 + 国标行业	86.5%	91.3%
企业名称 + 经营范围	89.8%	94.0%
企业名称 + 经营范围 + 国标行业（完整）	**93.1%**	**96.4%**

结果表明：经营范围信息对分类效果贡献最大（$\Delta = +10.6$ pp vs 仅企业名称），国标行业作为辅助信息也带来了显著增益（$\Delta = +3.3$ pp vs 名称 + 经营范围）。三种信息的联合使用达到了最佳效果。

5.6 案例分析

表 6 展示了混合架构在两个典型案例上的分类结果，并与纯双塔检索进行对比。

表 6: 典型案例分析

案例	纯双塔检索（Top-1）	本文方法（Top-1）
案例 1：某锂电池隔膜生产企业，国标行业为”C26 化学原料和化学制品制造业”	新能源 > 新能源汽车 > 动力电池 > 电池材料 > 隔膜（相似度 0.82）	**新材料 > 高分子材料 > 功能膜材料 > 隔膜**（置信度 0.94）
案例 2：某新能源整车企业，经营范围包括汽车、电池、零部件的研发制造	装备制造 > 汽车制造 > 整车制造 > 新能源整车（相似度 0.88）	**新能源 > 新能源汽车 > 整车制造 > 纯电动整车**（置信度 0.91）

案例 1 分析：纯双塔检索受“锂电池”关键词干扰，错误地将一家材料供应商归类为新能源产业链。而 Qwen 精排通过分析“国标行业 = 化学原料和化学制品制造业”和“经营范围中隔膜的**材料属性**”，正确判断企业本质为高分子材料供应商，归入新材料产业链。这正是双塔模型“理解文本相似性”但“不理解业务本质”的典型例证。

案例 2 分析：虽然从传统制造分类看，整车企业应归属“装备制造”，但该企业经营范围明确聚焦新能源整车，且涉及电池、电控等新能源核心技术。Qwen 精排综合判断后将其归入更精准的“新能源 → 新能源汽车”路径，体现了对产业转型企业（从传统装备制造向新能源转型）的灵活分类能力。

5.7 部署性能

本文方法已在物理服务器上实际部署运行，服务器配置为 2 块 Quadro RTX 5000（每块 16GB 显存），双塔模型部署在 GPU 1 上（显存占用约 2GB），Qwen2.5-14B-GPTQ 部署在 GPU 0 上（显存占用约 9GB），vLLM 推理引擎占用端口 8000。

实际运行性能指标如表 7 所示：

表 7: 部署性能指标

指标	实测值
双塔召回延迟	50~100ms
Qwen 精排延迟	500~1000ms
端到端总延迟	600~1100ms
单次 Token 消耗	约 5,000
串行吞吐量	约 60~100 条/分钟
批量吞吐量（10 并发）	约 500 条/分钟
FAISS 索引大小	约 21MB
bge 模型显存/内存	约 2GB
Qwen 模型显存	约 9GB

部署结果表明，本方法在单卡 16GB 显存的消费级 GPU 上即可稳定运行，端到端延迟满足在线服务场景的实时性要求（<2 秒），批量模式下可高效处理每日新注册企业的打标任务。

6 讨论

6.1 错误分析

为深入理解方法的局限性，我们对测试集中分类错误的样本进行了系统分析。错误模式大致可分为以下几类：

(1) **产业链边界模糊**（约占错误的 40%）：部分企业的经营范围横跨多个产业链节点，且无明显的主次之分。例如，一家同时从事光伏组件（新能源）和半导体设备（信息技术）的企业，任一单一标签都无法完整描述其业务。当前版本的模型对此类混合型企业支持不足。

(2) **经营范围文本质量低**（约占错误的 30%）：部分企业的经营范围文本过于简略（如仅写“法律法规允许的经营范围”）、包含大量工商注册的模板化表述、或包含极少见的技术术语。LLM 缺乏足够信息进行准确判断。

(3) **新兴业态**（约占错误的 20%）：标签库基于现有产业分类编制，对于快速涌现的新兴业态（如“低空经济”“氢能储运”“AI Agent 开发”等），可能不存在精准的产业链路径。模型被迫选择最接近但不完全匹配的标签，导致语义偏差。

(4) **LLM 幻觉**（约占错误的 10%）：少数情况下，LLM 输出了看似合理但实际与候选标签均不符的产业链路径，或在推理理由中给出了与业务事实不符的判断。这类错误虽然占比较低，但因其输出格式“看起来可信”而具有较高的隐蔽性。

6.2 局限性

本文方法存在以下局限性，值得在后续工作中加以改进：

(1) **标签库的静态性**：产业链标签库是静态的 5 级树状结构，虽然覆盖了主流产业，但无法自适应地捕捉产业融合、技术迭代带来的新业态。未来的一个方向是探索标签库的自动扩展和更新机制，例如利用 LLM 从企业集群中自动发现新兴产业链节点；

(2) **冷启动问题**：对新注册企业而言，如果缺少经营范围或经营范围描述极简，模型的准确率会显著下降。引入企业官网、新闻报道等多源信息来丰富企业画像，是缓解这一问题的可能途径；

(3) **地域定制化**：当前训练数据主要来自中山市及周边区域的企业，产业链分布具有明显的地域特征（如制造业企业占比高）。模型在其他区域或产业结构差异较大的城市（如服务业主导、高科技聚集区）的泛化能力需要进一步验证；

(4) **两阶段级联的误差传播**：虽然容错降级策略能在多数情况下兜底，但第一阶段的漏召回（false negative）无法在第二阶段纠正。当真实标签不在 Top-30 候选内时，精排器无从选择。进一步提高第一阶段的召回率（特别是对罕见产业链路径）是一个持续的优化方向；

(5) **成本与延迟的权衡**：虽然 $k = 30$ 在当前业务场景下是最优选择，但对于对延迟要求更苛刻的场景（如搜索引擎的实时查询建议），600~1100ms 的端到端延迟仍然偏高。进一步探索更小的精排模型（如 Qwen2.5-7B）或适配推理优化（如 Speculative Decoding）是值得尝试的方向。

6.3 未来工作

基于上述讨论，未来工作将围绕以下几个方向展开：

(1) **动态标签库更新**：利用 LLM 的能力，从新注册企业的经营范围中自动发现新兴产业链节点，经人工审核后纳入标签库，使分类体系保持时效性；

(2) **多源信息融合**：在企业名称和经营范围的基础上，引入企业官网内容、新闻报道、招投标公告等外部文本，构建更丰富的企业画像，提升冷启动场景下的分类质量；

(3) **端到端联合优化**：当前的双塔和 LLM 是独立训练的。探索将两个阶段的训练目标进行联合优化（如通过强化学习微调双塔模型以最大化最终精排质量），有望消除两阶段之间的信息不对称；

(4) **跨地域泛化验证**：在多个不同产业结构的城市部署系统并对比分类准确率差异，评估模型的地域泛化能力，必要时进行多地域数据扩充和领域自适应训练。

7 结论

本文针对大规模产业链自动分类任务，提出了一种双塔召回与大语言模型精排相结合的混合架构。该方法将分类过程分解为两个阶段：第一阶段使用微调的 bge-large-zh-v1.5 双塔模型结合 FAISS 索引，从 6,883 条产业链标签中高效召回 Top-30 候选；第二阶段使用 Qwen2.5-14B 大语言模型进行深度推理精排，从候选池中精选出 1~5 条最匹配的产业链路径。

在 137,822 条真实企业数据上的实验结果表明：（1）混合架构的最终分类准确率达到 93.1%（Precision@1）和 96.4%（Precision@3），优于纯稠密检索（+16.3 pp）和纯 LLM 直接分类（+5.9 pp）；（2）候选数 $k = 30$ 在召回率（96.5%）和效率之间取得了最优平衡；（3）端到端延迟为 600~1100ms，满足在线服务的实时性需求，已在生产环境中稳定运行。

本方法的实用价值在于：为政府产业分析、精准招商引资等场景提供了可自动化的企业产业链标注能力，有效降低了人工标注成本和主观偏差。方法的技术思路——用高效检索为昂贵 LLM 推理做减法——也适用于其他大规模标签空间下的细粒度分类任务，具有良好的可迁移性。

参考文献

[1] ZHOU J, MA C, LONG D, et al. Hierarchy-aware global model for hierarchical text classification[C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). 2020: 1106-1117. DOI: 10.18653/v1/2020.acl-main.104.

[2] WEHRMANN J, CERRI R, BARROS R C. Hierarchical multi-label classification networks[C]//Proceedings of the 35th International Conference on Machine Learning (ICML). 2018: 5075-5084.

[3] REIMERS N, GUREVYCH I. Sentence-bert: Sentence embeddings using siamese bert-networks[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 3982-3992. DOI: 10.18653/v1/D19-1410.

[4] XIAO S, LIU Z, ZHANG P, et al. C-pack: Packaged resources to advance general chinese embedding[A]. 2023.

[5] JOHNSON J, DOUZE M, JÉGOU H. Billion-scale similarity search with gpus[J/OL]. IEEE Transactions on Big Data, 2019, 7(3): 535-547. DOI: 10.1109/TBDATA.2019.2921572.

[6] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems (NeurIPS): volume 33. 2020: 1877-1901.

[7] BAI J, BAI S, CHU Y, et al. Qwen technical report[A]. 2023.

[8] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[C]//Advances in Neural Information Processing Systems (NeurIPS): volume 33. 2020: 9459-9474.

[9] NOGUEIRA R, CHO K. Passage re-ranking with bert[Z]. 2019.

[10] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). 2019: 4171-4186.

[11] GLASS M, ROSSIELLO G, CHOWDHURY M F M, et al. Re2g: Retrieve, rerank, generate[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). 2022: 2701-2715.

[12] ZHUANG S, ZHUANG H, KOOPMAN B, et al. A setwise approach for effective and highly efficient zero-shot ranking with large language models[Z]. 2023.

[13] HENDERSON M, AL-RFOU R, STROPE B, et al. Efficient natural language response suggestion for smart reply[Z]. 2017.

[14] FRANTAR E, ASHKBOOS S, HOEFLER T, et al. Gptq: Accurate post-training quantization for generative pre-trained transformers[C]//Proceedings of the 11th International Conference on Learning Representations (ICLR). 2023.

[15] KWON W, LI Z, ZHUANG S, et al. Efficient memory management for large language model serving with pagedattention[C/OL]//Proceedings of the 29th Symposium on Operating Systems Principles (SOSP). 2023: 611-626. DOI: 10.1145/3600006.3613165.