arXiv:2601.00926v1 [cs.IR] 1 Jan 2026

MACA：将可信大语言模型蒸馏为高效检索器的框架

Satya Swaroop Gudipudi[1],Sahil Girhepuje[1],Ponnurangam Kumaraguru[2],和 Kristine Ma[1]

[1] 摩根大通[2] 海得拉巴国际信息技术学院satyaswaroop.gudipudi@chase.com, sahil.girhepuje@chase.com, pk.guru@iiit.ac.in, kristine.ma@chase.com

摘要。 现代企业检索系统必须处理简短且信息不充分的查询，例如“境外交易手续费退款”和“近期支票状态”。在这些情况下，语义细微差别和元数据至关重要，但针对每个查询进行大语言模型（LLM）重排序和人工标注成本高昂。我们提出元数据感知跨模型对齐（MACA），将经过校准的元数据感知大语言模型重排序器蒸馏为紧凑型学生检索器，从而避免在线大语言模型调用。元数据感知提示词通过检查排列下的一致性和对释义的鲁棒性来验证教师模型的可信度，然后提供列表式评分、困难负例和校准后的相关性间隔。学生模型通过 MACA 的 MetaFusion 目标函数进行训练，该目标函数结合了元数据条件排序损失与跨模型间隔损失，使其学会将正确答案排在语义相似但主题、子主题或实体不匹配的候选项之前。在专有消费者银行常见问题语料库和 BankFAQs 上，MACA 教师模型的准确率@1 比 MAFA 基线分别高出五个百分点和三个百分点。MACA 学生模型的表现显著优于预训练编码器；例如，在专有语料库上，MiniLM 的准确率@1 从 0.23 提升至 0.48，同时推理过程无需调用大语言模型，并支持检索增强生成。

关键词： 元数据感知 · 可信度 · 高效检索。

1 引言

现代企业检索系统必须在准确率和延迟之间取得平衡，同时处理诸如“境外交易手续费退款”或“近期支票状态”这类信息不完整的用户查询。这些查询较短，仅部分指明其元数据；它们暗示了用户的需求，却往往省略主题、子主题、意图或实体等关键信息。因此，元数据分类体系中带有不同标签的多个候选答案仍然看似合理。例如，传统语义检索器可能会混淆信用卡的境外交易手续费与借记卡的境外 ATM 手续费，也可能把“查询交易状态”与“近期支票状态”这类查询混为一谈，因为它们主要依赖表层文本相似度，而对候选项一侧的结构化元数据利用有限。

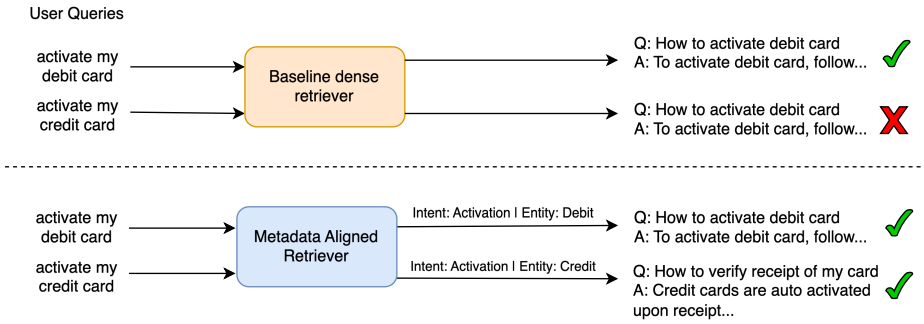


图 1：MACA 可减少简短银行业查询中的实体漂移。基线模型对两个查询都返回相同的借记卡常见问题；MACA 则在排名第 1 位选出正确的借记卡/信用卡常见问题（推理时无需使用大语言模型）。

如图 1 所示，基线稠密检索器可能会将“激活我的借记卡”和“激活我的信用卡”都匹配到同一个借记卡常见问题。元数据感知检索器则会利用意图和卡类型，为每个查询选择不同的答案，而且无需在推理时调用大语言模型 (LLM)。

因此，生产环境中的流水线依赖多阶段检索：首先由 BM25 [1] 等词法模型生成候选结果，再由交叉编码器、ColBERT 或大语言模型重排序器等神经重排序器调整顺序，以捕捉更细微的语义差异 [2–4]。在企业环境中，为了区分用户意图、产品或费用类型的细微差异（例如，区分境外交易手续费和境外 ATM 手续费），通常需要进行此类优化。然而，针对每个查询执行大语言模型推理成本高、对延迟敏感，而且在运维上容易出问题 [2]。大规模人工标注也既昂贵又不一致，尤其是在必须可靠推断意图、主题、子主题和实体等潜在元数据时。

以往系统通常采用结合大语言模型重排序器的混合检索 [3, 4] ，或将大型教师模型通用地蒸馏到小型学生模型中 [5, 6], ，但往往没有充分利用显式领域元数据。在分类体系庞大的场景中，若将所有元数据不加区分地注入大语言模型提示词，可能会超出上下文限制，并稀释相关性推理。因而，有效的检索需要一种监督信号，既能突出最显著的元数据信号，又不会给模型造成过重负担。

我们通过**MACA**（元数据感知跨模型对齐）弥补这些不足。这是一个两阶段框架，可将经过校准的元数据感知大语言模型重排序器蒸馏为紧凑的单次检索器。在第一阶段，元数据感知提示词首先通过检查候选项排列下的一致性以及对释义的鲁棒性，验证教师模型的可信度。随后，它会生成列表式评分、校准后的相关性间隔，以及困难负例；这些负例对应的常见问题措辞相似，但结构化元数据不同。在第二阶段，紧凑的 MACA 学生模型通过 MetaFusion 目标函数从这些信号中学习。该目标函数结合了：(i) 元数据条件排序损失；(ii) 跨模型间隔损失（RCMA），用于迁移教师模型在近似竞争项上的间隔。在

推理时仅运行学生模型，即可实现低延迟和稳定的质量，且无需在线调用大语言模型。

在专有消费者银行业务常见问题数据集和公开的 BankFAQs 数据集上， [7], MACA 教师在准确率@1 上分别比 MAFA [8] 智能体基线高出约 5 个百分点（专有数据集）和 3 个百分点（BankFAQs）。随后，经过 MACA 蒸馏的学生模型大幅超越预训练编码器；例如，在专有语料库上，MiniLM 的准确率@1 从 0.23 提升至 0.48，同时仍接近教师模型。这使推理过程无需调用大语言模型，也与受监管领域中检索增强生成的延迟和治理约束十分契合。

贡献。 我们的贡献有三方面：(i) 提出一种可信度协议和元数据感知提示词，沿性能、一致性和释义鲁棒性这几个维度校准大语言模型重排序器；(ii) 提出 MACA 元融合目标，将元数据条件化排序与跨模型间隔对齐相结合，并使用 MACA 判定器选出的近似负例；(iii) 在两个银行常见问题语料库上开展实证研究，表明经过校准的大语言模型教师能够监督紧凑型检索器，使其在无需为每个查询调用大语言模型的情况下，取得接近教师模型的大部分性能提升。

2 相关工作

多阶段检索与融合。 现代信息检索系统通常结合高召回率的第一阶段检索器与更强大的重排序器 [9]。词法 BM25 [1] 仍是稳健的稀疏基线，而 ColBERT [4] 等晚期交互模型则能以可控的延迟提升效果。BEIR [10]等基准测试记录了稀疏检索、稠密检索和重排序等方法在效果与延迟之间的权衡，而倒数排名融合（RRF） [11]通常通过聚合异构候选结果进一步提升效果。多阶段基于 BERT 的排序器 [9]通过叠加稠密检索和交叉编码器组件，进一步拓展了这一模式。MACA 在第一阶段采用这种混合设计，但最终以单个蒸馏检索器取代重排序阶段。

作为重排序器和评判器的大语言模型。通过提示词引导的大语言模型已成为强大的零样本和少样本重排序器：成对排序提示（PRP） [12], RankGPT [13],和 RankZephyr [2] 表明，在 BEIR 风格的基准测试中，成对提示和列表式提示的表现可媲美甚至超越有监督的交叉编码器。与此同时，大语言模型即评审研究（例如 MT-Bench 及相关工作 [14]„ 以及《我们能信任大语言模型作为相关性评判者吗？》等研究 [15] ）指出，大语言模型判断可能会受到提示词措辞、候选项顺序和领域迁移的影响。当大语言模型被用来替代人工评估者时，这种情况尤为明显。近期关于基于基准测试的大语言模型评估在改述下的鲁棒性的研究发现，即使模型排名相对稳定，绝对分数仍可能大幅下降 [16]。MACA 与这一研究方向一致：我们不预设大语言模型重排序器天生可靠，而是明确

使用 Accuracy@k、改述鲁棒性@k 和顺序一致性@k 审查候选提示词，并且只将最稳定的元数据感知提示词冻结为教师模型。

面向高效检索器的引导式蒸馏。 蒸馏将能力强但速度慢的教师模型压缩为轻量级学生模型 [5]。对于神经排名，MarginMSE 通过对齐学生模型与教师模型的间隔，持续提升高效排序器的表现 [17]；RocketQA [18] 则将困难负样本挖掘和去噪结合起来，用于稠密检索。除了排名之外，大语言模型生成的监督信号也推动了指令微调和数据增强的发展：Self-Instruct [19],、Alpaca [20],和 AugGPT [21]。我们的工作在此基础上进一步探索：(i) 在标注前校准元数据感知的大语言模型提示词，以提升可靠性；(ii) 通过跨模型间隔目标，使紧凑型学生模型与提示词的偏好保持一致。

元数据感知引导与多智能体标注。 元数据可以提升准确率，但往往会增加延迟。近期一种用于极端分类的方法 MOGIC 展示了早期融合元数据“预言机”如何在不增加服务时开销的情况下，引导低延迟学生模型 [22]。与我们的设置更为接近的是，MAFA [8] 采用多个专门智能体及一个判别器进行常见问题标注，并报告了在专有银行数据集和公开银行数据集上的显著提升。MACA 的不同之处在于，它在推理时无需逐查询调用大语言模型：我们首先评估并选择用于标注的元数据感知提示词，然后将其判断蒸馏到一个学生检索器中，并通过跨模型间隔学习使其与之对齐。

3 方法论（MACA）

我们围绕两个问题展开： (研究问题1) 能否对大语言模型进行校准，使其能够可靠地执行具有元数据感知能力的相关性评分；以及 (研究问题2) 这样的教师模型能否监督一个紧凑型检索器，从而使推理无需在线大语言模型调用？每个项目 $f \in \mathcal{F}$ 由一个问题、一个答案以及由算法1诱导出的分类体系标签（主题、子主题、意图、实体）组成。用户查询通常较短且含义模糊，因此获取细粒度的人工相关性标签成本高昂。图2概述了MACA流程以及第一阶段与第二阶段之间的交互。

3.1 自动分类体系创建与标注

我们使用大语言模型构建涵盖主题、子主题、意图和实体的元数据分类体系，并为查询和常见问题段落添加标签。基于消费者银行业务的先验知识，大语言模型首先为每个未标注查询分配一个粗粒度主题。第二轮在该主题的条件下，预测子主题、意图和显著实体。由此得到的（主题、子主题、意图、实体）元组构成自动分类体系，我们使用相同的提示方案为常见问题及其答案添加标签，并复用该体系。在实验中，我们完全依赖这一自动分类体系。在生产系统中，也可以选择由领域专家（SMEs）对其进行一次优化。算法1总结了这一流程。

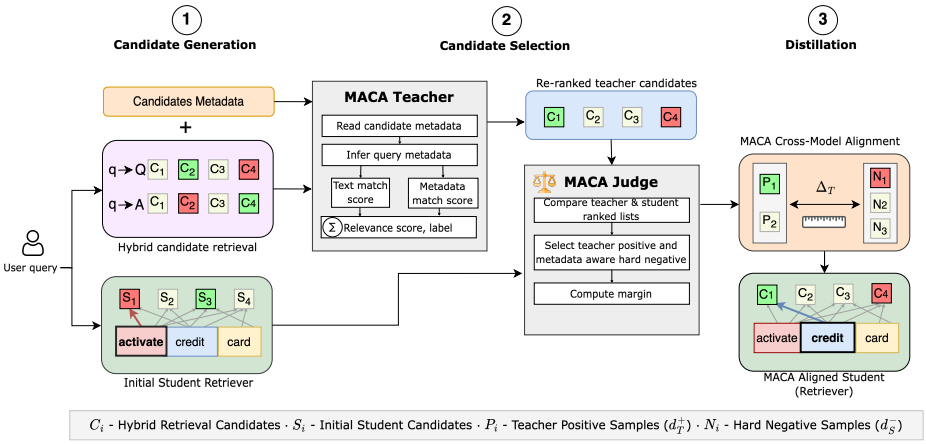


图 2：**MACA流程**。第一阶段：来自查询到问题视图$q\to Q$ 和查询到答案视图 $q\to A$ 的候选项通过倒数排名融合进行合并，再由经过校准的 MACA 教师模型使用自动分类体系创建步骤中的元数据进行重排序（算法1）。第二阶段：MACA 判定器选出一个教师正样本d_T^+ 和一个学生困难负样本 d_S^-，并生成一个边际值 Δ_T。随后使用这些三元组，通过 MNRL 和 RCMA 训练紧凑型学生模型。在推理时，仅使用学生模型。

3.2 第一阶段：经过校准的大语言模型重排序器（研究问题1）

第一阶段构建一个元数据感知大语言模型重排序器，作为教师模型。要让 MACA 发挥作用，这个教师模型必须值得信赖：它对留出标签的判断应准确（性能），在候选列表发生无害变化时应保持稳定（一致性），并且不受同一查询的不同释义影响（鲁棒性）。如果教师模型对“card activation”和“activate card”给出不同的首选答案，就只会把不稳定性带入学生模型。因此，我们首先校准并选定一个在这三个维度上表现良好的提示词，然后才固定该教师模型用于蒸馏。

候选项生成。对于每个查询 q，我们基于常见问题语料库中的两个稠密视图构建一个高召回率候选集。查询到问题视图 $q\to Q$仅搜索常见问题，查询到答案视图 $q\to A$ 仅搜索常见问题答案。 $q\to Q$在查询与常见问题标题相似时效果最佳，而 $q\to A$则能在关键信息（例如卡类型或境外交易手续费）明确写在答案而非问题中时提高覆盖率。我们分别对每个视图进行排名，再通过倒数排名融合（RRF） [11]将其合并；合并后的前K 个候选项随后被传递给大语言模型重排序器。

元数据感知重排序。 元数据感知提示词要求大语言模型 (i) 推断查询级元数据（意图、主题、子主题、实体），并 (ii) 综合文本与元数据对齐情况为每个候选项评分。大语言模型采用 TREC 风格的四级评分（精确匹配、部分匹配、较不相关、不相关），我们将其映射

算法1 基于大语言模型的自动分类体系创建与标注

要求： 查询集 Q、常见问题集 $\mathcal{F}$、领域先验 D（消费者银行业务）、大语言模型 L
确保： 分类体系 $\mathcal{T}$ 涵盖（主题、子主题、意图、实体）以及以下对象的元数据标签 Q 和 $\mathcal{F}$

$\mathcal{T} \leftarrow \emptyset$
▷步骤 1：为查询分配主题
对所有 $q \in Q$ **执行**
  $t_q \leftarrow L(D,\ q)$
  存储 t_q 用于 q；更新 $\mathcal{T}$.主题
结束循环
▷步骤 2：查询的细粒度元数据
对所有 $q \in Q$ **执行**
  $(s_q, i_q, E_q) \leftarrow L(D, q, t_q)$
  为 q存储 $(t_q,\ s_q,\ i_q,\ E_q)$
  使用 $(t_q,\ s_q,\ i_q,\ E_q)$ 更新 $\mathcal{T}$
结束循环
▷ 步骤 3：使用诱导分类体系为常见问题添加标签
对于所有 $(Q_f,\ A_f) \in \mathcal{F}$**执行**
  $(t_f,\ s_f,\ i_f,\ E_f) \leftarrow L(\mathcal{T},\ Q_f,\ A_f)$
  存储 $(t_f,\ s_f,\ i_f,\ E_f)$ 以用于 f常见问题
  更新 $\mathcal{T}$ 中 $(t_f,\ s_f,\ i_f,\ E_f)$ 的支持计数
结束循环
▷ 第 4 步（可选）：生产系统的主题专家精炼
可选地通过合并近重复标签并剪除低支持度的条目
返回 $\mathcal{T}^*$ 以及 Q 和 $\mathcal{F}$的所有元数据标签

为数值标签 3，2，1，0，遵循 TREC DL [23]。此外，它还会根据文本覆盖度与元数据重叠度（意图、主题、子主题、实体杰卡德）的加权组合，为每个候选项 d 生成连续相关性分数 $S_T(d)$；这些分数随后用于确定正样本与近似竞争项之间的间隔。

我们探索了多种提示词策略，包括思维链、Reflex- ion 和 ReAct [24–26], ，并将其可信度作为一个涵盖 Accuracy@k、一致性@k 和鲁棒性@k 的三维概况进行衡量（第 4.1 节）。随后，针对研究问题1，我们选择在这三个维度之间实现最佳权衡的提示词，并将其固定为 MACA 教师。

3.3 第二阶段：候选选择与蒸馏（研究问题2）

候选选择（*MACA* 判定器）。给定一个查询 q，冻结的教师重排序器返回一个包含 k_T 个条目的列表及其分数 S_T（·），学生检索器则返回一个包含 k_S 个条目的列表及其分数 S_S(·)。MACA 判定器是一个大语言模型提示词，它可以同时查看这两个带有元数据和教师标签的列表，并选择一个正样本与一个困难负样本组成的样本对。

(1)教师正样本。判别器将 d_T^+ 设为得分最高且标签为精确匹配的教师候选项；如果不存在，则选择得分最高的部分匹配项。

(2)元数据感知困难负例。随后，判别器会构建一个由差一点匹配的学生候选项组成的集合

$$\mathcal{N} = \{\, d \in \text{student top } k_S : d \neq d_T^+,\ d \text{ not marked relevant by the judge},\ \text{overlap}(d_T^+, d) \geq 1 \,\}, \tag{1}$$

其中，重叠测试用于判断主题或意图是否相同，并允许子主题或实体存在差异（例如境外交易手续费与境外 ATM 手续费）。如果 $\mathcal{N} = \varnothing$，则回退为判别器标记为不相关的学生条目中分数最高的一项。困难负样本是

$$d_S^- = \arg\max_{d \in \mathcal{N}} S_S(d) \tag{2}$$

（3）校准边距。对于（d_T^+， d_S^-），判别器使用教师分数计算

$$\Delta_T = S_T(d_T^+) - \max\big(S_T(d_S^-),\, \gamma\big), \tag{3}$$

并采用标签感知下限 γ，同时截断至 $[-m_{\max}, m_{\max}]$，以抑制异常值。

蒸馏。学生模型通过以下方式训练：

$$\mathcal{L}_{\text{total}} = \alpha\, \mathcal{L}_{\text{MNRL}} + \beta\, \mathcal{L}_{\text{RCMA}} \tag{4}$$

此处，$\mathcal{L}_{\text{MNRL}}$ 是标准的批内多负例排序损失 [27]„ 而 RCMA 是一种 MarginMSE 风格的蒸馏项 [17]:

$$\Delta_S = S_S(q, d_T^+) - S_S(q, d_S^-), \qquad \mathcal{L}_{\text{RCMA}} = \big(\Delta_S - \Delta_T\big)^2 \tag{5}$$

这种对齐将教师模型对正样本与困难负样本的区分迁移过来，促使学生模型复现元数据感知的分差。

总而言之，MACA 包括一个基于大语言模型的自动元数据诱导步骤（算法1）、一个经校准的元数据感知大语言模型重排序器，用于为候选项打分（第一阶段），以及一个 MACA 判定器，用于选择正样本、困难负样本和边际值，并通过 MetaFusion 训练紧凑型学生模型（第二阶段）。在推理时，仅使用学生检索器。

4 实验

4.1 设置

数据集 我们使用两个常见问题语料库。（i）专有消费者银行业务：15k 个未标注查询用于蒸馏，250 个专家标注查询用于评估；其元数据分类体系（∼120 子主题和 ∼150 实体）通过算法1自动诱导，并在提示中作为枚举值使用，同时用于元数据感知评估。（ii）BankFAQs：2,400 个问答对；我们使用 GPT-4o 改写问题，以获得 2,400 个合成训练查询和一个包含 250 个查询的留出集。

教师模型和学生模型 第一阶段（教师模型）使用基于 OpenAI text-embedding-3-large 的稠密 $q \to Q$和 $q \to A$ 检索，通过倒数排名融合 [11], 进行融合，然后使用元数据感知的 GPT-4o 重排序器（MACA 教师，仅在线下用于标注）。第二阶段将其蒸馏到三种生产检索中广泛使用的现成编码器: all-MiniLM-L6-v2、all-mpnet-base-v2 和 msmarco-distilbert-base-tas-b [28–30]。在表 2 中，我们分别将它们称为 MiniLM、MPNet 和 TAS-B。我们报告使用 MACA–MNRL 和 MACA–MetaFusion（MNRL+RCMA）微调的模型。

指标与可信度**Accuracy@k** 是指至少有一个相关候选项出现在前-k 个结果中的查询所占比例（$k \in \{1, 3, 5, 10, 15\}$）。

对于每个模型，我们报告其在 Accuracy@k、一致性@k（C@k）和鲁棒性@k（R@k）上的三维可信度概况。设 Q为评估查询集合。对于排列，令 topk$_p(q)$ 表示在固定候选列表的 p 排列下，查询 q 的前-k 个集合，每个查询有 P_q 种排列。对于 q的释义 $\{q^{(i)}\}_{i=1}^{M_q}$ ，令 topk($q^{(i)}$) 表示释义 i的前-k 个集合。我们定义

$$
\begin{aligned}
C@k &= \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{P_q} \max_{S} \#\{\, p : \mathrm{topk}_p(q) = S \,\}, \\
R@k &= \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{M_q} \max_{S} \#\{\, i : \mathrm{topk}(q^{(i)}) = S \,\}.
\end{aligned}
\tag{6}
$$

因此，C@k 是与众数前-k 个集合一致的排列所占比例的平均值，而 R@k 则是释义中相应比例的平均值。

训练除非另有说明，学生模型使用 Adam 训练 4 个 epoch（学习率2×10^{-5}，批量大小 64，线性预热）；MetaFusion 权重（α，β）在验证集上选定。代码发布须经组织审批。

4.2 实验结果

研究问题*1*：值得信赖的大语言模型教师。表 1 比较了将不同大语言模型提示策略用作重排序器时的表现。在专有银行数据集上，MACA 在所有截断位置均取得最高准确率（Acc@1/3/5 $= 0.61/0.79/0.84$），Acc@1 比思维链和 Reflexion 高 3–4 个百分点，比 MAFA 高 5 个百分点。在 BankFAQs 上，由于查询是常见问题的释义，任务相对容易，各种提示策略都取得了较高准确率，但 MACA 的 top-1 性能仍然最强（Acc@$1 = 0.72$，而思维链/ReAct 为 0.71，Reflexion/MAFA 为 0.66–0.68），同时 Acc@3/5 也达到最佳水平。图 3 展示了稳定性方面的类似趋势：在 $k \in \{1, \ldots, 5\}$ 上，MACA 的一致性@k 最高，且对释义的鲁棒性最强；与思维链、ReAct 或 Reflexion 相比，其专有数据集与 BankFAQs 曲线之间的差距更小。总体而言，MACA 在性能、一致性和鲁棒性之间实现了最佳权衡，支持将其用作冻结教师进行蒸馏。

表 1：在消费者银行业务（专有数据）和 BankFAQs 上，对不同大语言模型提示策略进行提示词级消融实验：准确率@{1,3,5}。

	消费者银行业务（专有）			BankFAQs		
提示词	Acc@1	Acc@3	Acc@5	Acc@1	Acc@3	Acc@5
思维链	0.57	0.73	0.79	0.71	0.88	0.92
ReAct	0.53	0.74	0.77	0.71	0.89	0.92
Reflexion	0.58	0.73	0.81	0.66	0.87	0.90
MACA（我们的方法）	0.61	0.79	0.84	0.72	0.89	0.92
MAFA	0.56	0.70	0.77	0.68	0.81	0.86

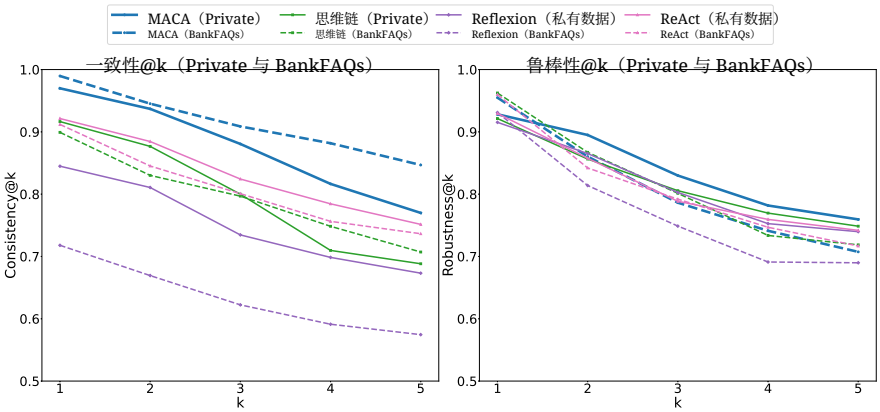


图 3：四种提示在两个数据集上的可信度（一致性@1、鲁棒性@1）；MACA 的综合表现最为均衡。

研究问题*2*：蒸馏后的检索质量。表 2 报告了从 MACA 教师蒸馏出的三个学生模型的检索准确率。在专有语料库上，与 BM25 及各骨干网络未经微调的版本相比，MACA微调显著提升了所有骨干网络的表现：采用 MetaFusion 后，MiniLM/MPNet/TAS-B 的 Acc@1 从 0.23/0.32/0.26 提升至 0.48/0.47/0.46，大幅缩小了与教师模型（0.61）之间的差距。在更高的截断位置也有类似提升，例如 MiniLM 的 Acc@5 从 0.61 提升至 0.77，MPNet 和 TAS-B 也呈现类似趋势。在 BankFAQs 上，由于预训练编码器的 Acc@$1 \geq 0.58$ 和 Acc@$5 \geq 0.89$ 已经较高，MACA 带来的提升幅度较小，但稳定可见（例如，MiniLM 的 Acc@1 从 0.62 提升至 0.70，MPNet 从 0.64 提升至 0.70）。在两个数据集上，MACA 学生模型都保留了大语言模型教师的大部分效能，同时仅需运行一次编码器，且无需在线大语言模型调用。

消融实验：*MetaFusion* 损失。 从表 2 可以看出，相较于 BM25 和预训练学生模型，大部分性能提升来自 MNRL，它使专有数据集上的 Acc@1 大致翻倍（例如，MiniLM 达到 $0.23 \rightarrow 0.46$）。随后，MetaFusion（+MF）在所有骨干网络的最高排名位置带来了幅度较小但稳定的提升：消费者银行业务上的 Acc@1 提高了 2–4 个百分点，BankFAQs 上则提高了约 1 个百分点，

表 2：基线模型和 MACA 蒸馏学生模型在消费者银行业务（专有数据集）和 BankFAQs 上的 Accuracy@{1,5,10,15}。MACA 教师（大语言模型重排序器）仅在线下用于标注训练数据；推理时使用学生检索器。MiniLM、MPNet 和 TAS-B 分别表示 all-MiniLM-L6-v2、all-mpnet-base-v2 和 msmarco-distilbert-base-tas-b；“+MNRL”和“+MF”表示分别采用 MNRL 和 MACA–MetaFusion 进行 MACA 微调。

	消费者银行业务（专有数据集）				BankFAQs			
模型	Acc@1	Acc@5	Acc@10	Acc@15	Acc@1	Acc@5	Acc@10	Acc@15
BM25	0.19	0.44	0.51	0.55	0.38	0.65	0.71	0.75
MiniLM	0.23	0.61	0.71	0.79	0.62	0.90	0.93	0.95
MiniLM+MNRL	0.46	0.77	0.80	0.83	0.69	0.93	0.96	0.97
MiniLM+MF	0.48	0.77	0.82	0.83	0.70	0.94	0.96	0.97
MPNet	0.32	0.57	0.70	0.75	0.64	0.92	0.95	0.97
MPNet+MNRL	0.44	0.76	0.78	0.81	0.70	0.94	0.96	0.97
MPNet+MF	0.47	0.76	0.78	0.79	0.70	0.94	0.96	0.97
TAS-B	0.26	0.46	0.58	0.65	0.58	0.89	0.95	0.97
TAS-B+MNRL	0.42	0.76	0.79	0.81	0.64	0.92	0.97	0.97
TAS-B+MF	0.46	0.75	0.79	0.81	0.65	0.91	0.97	0.97

同时 Acc@5 几乎没有变化。这与 RCMA 的设计相符：通过对齐学生模型与教师模型在正样本和元数据感知困难负例之间的间隔，MetaFusion 将学习重点放在决定严格顶层检索的边界样本对上。MACA 判别器对近似错误样本的筛选确保这些样本对的差异主要在于子主题或实体，因此学生模型能够学会保留对排名靠前答案至关重要的元数据敏感型区分。

5 讨论与未来工作

MACA 旨在检验大语言模型能否经过校准，成为可信的元数据感知相关性教师（研究问题1），以及这样的教师能否监督紧凑型检索器，从而免去在线大语言模型调用（研究问题2）。我们在银行常见问题解答场景中的结果表明，这两个目标都可以实现：MACA 提示词在准确率、一致性和鲁棒性方面均优于思维链/ReAct/Reflexion 和 MAFA 智能体基线；MACA 蒸馏学生模型的表现大幅超过 BM25 和现成的稠密编码器，同时保留了教师的大部分效能。

这些发现表明，检索增强生成在受监管领域具有实际优势。经过校准的教师与 MetaFusion 相结合，可生成快速且具备元数据感知能力的单次处理学生模型，避免逐查询进行大语言模型调用。此外，我们的可信度指标（Accuracy@k、一致性@k、鲁棒性@k）为在大规模标注中使用大语言模型重排序器之前对其进行审计提供了具体流程，这对必须说明模型选择依据的从业者很有价值。

作为未来工作，我们计划将 MACA 从银行业扩展到医疗和法律搜索等领域，并纳入更丰富的元数据，例如

用户或渠道上下文。我们还希望拓展对可信度的定义，将校准和感知不确定性的标注纳入其中，并在完整的 RAG 流水线中评估 MACA，以衡量其对端到端答案质量的影响。

免责声明

本文所表达的观点和意见仅代表作者本人，并不一定反映其现任或前任雇主的官方政策或立场。

A MACA 提示

MACA 重排序器提示词（摘要）

任务。为查询对候选常见问题进行排名
仅使用提供的分类体系元数据。
输入。 查询；带有意图的候选项，
主题、子_主题、实体。
步骤。
1. 推断查询元数据：意图、主题、
   子_主题（slug $\leq$ 3 个词或“未知”）、实体
   （$\leq$ 5 个词元）。
2. 对每个候选项：计算意图/-
   主题/子主题匹配；实体重叠
   $|\cap|$ 和杰卡德 $|\cap|/|\cup|$。
3. 分数：覆盖度 $\in [0,1]$ （尽量少读；
   不确定时为 0）。元数据 $= 0.40\,[\text{意图}] + 0.25\,[\text{主题}] + 0.20\,[\text{子主题}] + 0.15 \times$ 杰卡德。相关性 $= 0.65 \times$ 覆盖率 $+ 0.35 \times$ 元数据。
4. 标签：如果意图以及（主题或
   子主题）匹配且重叠≥1，则为精确匹配；如果存在任意匹配但并非精确匹配，则为部分匹配；如果相关性较低，则为
   主题匹配但主要实体不同；
   否则则不相关。
5. 歧义：如果意图不明确，优先选择可采
   取行动的内容（如何_操作/故障排除/政策
   icy_限制/资格），当分数
   关闭 ·

输出（严格 JSON，仅显示排名第一的结果）。

```
{"faq_id":"F123","label":"精确匹配",
 "相关性_评分":0.87,
 "top_matching_entity":"激活"}
```

MACA-Judge 提示词（摘要）

任务。选择教师正样本 $+$ 学生困难样本
负样本；输出校准后的间隔。
输入。 查询；教师模型排名前-k_T （分数 S_T）；
学生排名前-k_S （分数 S_S）；元数据。
策略。
1. **正样本** d_T^+： 教师模型中最高排名的项目标记为精确匹配；否则，
   标记为部分匹配的最佳项目。
2. **近似未命中集合** $\mathcal{N}$：学生条目未被
   教师模型标记为相关，且共享主题
   或与 d_T^+ 有相同意图；去重。
3. **困难负样本** d_S^-： $\operatorname{argmax}_{d\in\mathcal{N}} S_S(d)$；
   平局判定依据元数据接近度（主题 $>$
   意图 $>$ 实体），然后依据 $S_T(d)$。
4. **间隔** $\Delta_T = S_T(d_T^+) - \max(S_T(d_S^-), \gamma)$，
   截断至 $[-m_{\max}, m_{\max}]$。

输出（严格 JSON）。

```
{"query":"...","pos_id":"F123",
 "neg_id":"F987","delta_T":0.42}
```

图 4：MACA 的紧凑型提示词卡片。左：元数据感知重排序器。右：用于三元组选择和间隔校准的确定性判定器。

参考文献

1. S. Robertson 和 H. Zaragoza。《概率相关性框架：BM25 及其扩展》。发表于《信息检索基础与趋势》，3(4):333–389，2009 年。DOI：https://doi.org/10.1561/1500000019。

2. R. Pradeep、S. Sharifymoghaddam 等。《RankZephyr：高效且稳健的零样本列表式重排序》。arXiv:2312.02724，2023 年。可访问：https://arxiv.org/abs/2312.02724。3. Rodrigo Nogueira、Wei Yang、Kyunghyun Cho 和 Jimmy Lin。《使用 BERT 进行多阶段文档排序》。arXiv:1910.14424，2019 年。预印本：https: //arxiv.org/abs/1910.14424。4. O. Khattab 和 M. Zaharia。《ColBERT：通过 BERT 上的上下文化晚期交互，实现高效且有效的段落检索》。发表于 SIGIR 论文集，2020 年。DOI：https://doi.org/10.1145/3397271.3401075。预印本：https://arxiv. org/abs/2004.12832。5. Geoffrey Hinton、Oriol Vinyals 和 Jeff Dean。《蒸馏神经网络中的知识》。arXiv:1503.02531。网址：https://arxiv.org/abs/1503.02531。6. Victor Sanh 等。《DistilBERT：BERT 的蒸馏版本，更小、更快、更便宜、更轻量》。arXiv:1910.01108，2020 年。网址：https://arxiv.org/abs/1910.01108。7. Kshirasagar, Somanath。《BankFAQs》。https://www.kaggle.com/datasets/ somanathkshirasagar/bankfaqs。8. M. Hegazy、A. Rodrigues 和 A. Naeem。《MAFA：用于标注的多智能体框架》。arXiv:2505.13668，2025 年。PDF：https://arxiv.org/pdf/2505.13668。9. R. Nogueira 和 K. Cho。《使用 BERT 进行段落重排序》。NAACL（行业演示），2019 年。预印本：https://arxiv.org/abs/1901.04085。10. N. Thakur 等。《BEIR：用于信息检索模型零样本评估的异构基准》。NeurIPS 数据集与基准，2021 年。预印本：https://arxiv.org/abs/2104.08663。11. G. V. Cormack、C. L. A. Clarke 和 S. Büttcher。《倒数排名融合优于孔多塞方法和单独的排名学习方法》。发表于 SIGIR，2009 年。https: //doi.org/10.1145/1571941.1572114。预印本：https://cormack.uwaterloo. ca/cormacksigir09-rrf.pdf。12. Z. Qin 等。《通过成对排名提示，大语言模型可有效充当文本排序器》。arXiv:2306.17563，2023 年。网址：https://arxiv.org/abs/ 2306.17563。13. W. Sun 等。《ChatGPT 擅长搜索吗？探究大语言模型作为重排序智能体（RankGPT）的表现》。发表于 EMNLP Findings，2023 年。预印本：https: //arxiv.org/abs/2304.09542。14. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Chen, C., Li, X., Gonza- lez, J. E., & Stoica, I. (2023)。《以 MT-Bench 和 Chatbot Arena 评判大语言模型即评审》。*arXiv* 预印本 *arXiv:2306.05685*。https://openreview.net/forum? id=uccHPGDlao 15. Faggioli, G., Gialampouki, A., Koutsoumpos, A., Mizzaro, S., & Paletta, L. (2023)。《我们能信任大语言模型作为相关性评判者吗？》。载于：第 *13* 届意大利数字图书馆研究会议（*IRCDL*）论文集，119-130。https://sol.sbc.org.br/index.php/sbbd/article/download/30724 30527 16 / . Lunardi, R., Della Mea, V., Mizzaro, S., & Roitero, K. (2024)。《大语言模型基于基准的评估的稳健性与可靠性》。*arXiv* 预印本 *arXiv:2509*04013 17*04013*2509。。https://arxiv.org/pdf/ . . S. Hofstätter 等。《通过跨架构知识蒸馏改进高效神经排序模型》。arXiv:2010.02666，2020 年。网址：https:// arxiv.org/abs/2010.02666。18. Y. Qu 等。《RocketQA：面向开放域问答的稠密段落检索优化训练方法》。发表于 NAACL 论文集，2021 年。DOI：https://doi.org/10.18653/v1/2021.naacl-main.466。预印本：https://arxiv.org/abs/2010.08191。

19. Y. Wang 等. Self-Instruct：用自生成指令对齐语言模型. arXiv:2212.10560，2022. 网址：https://arxiv.org/abs/2212.10560。20. R. Taori 等. Alpaca：一种强大且可复现的指令遵循模型. arXiv:2303.16199（注：Alpaca 报告/技术说明；数据集已发布），2023。21. H. Dai 等. AugGPT：利用 ChatGPT 进行文本数据增强. arXiv:2302.13007，2023. https://arxiv.org/abs/2302.13007。22. S. C. Prabhu 等. MOGIC：用于改进极端分类的元数据注入式预言机引导. 载于国际机器学习大会论文集，2025。预印本：https://manikvarma.org/pubs/prabhu25.pdf。23. Nick Craswell 等. *TREC 2021* 深度学习赛道概述. TREC 2021 概述报告，2021。可查阅：
h
t
t
p
s://
w
w
w.
mi
crosoft.com/en-us/research/wp-content/uploads/2022/05/trec2021-deeplearning-overview.pdf。24. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi EH, Le QV, Zhou D（2022）思维链提示可激发大语言模型的推理能力。载于：神经信息处理系统大会论文集（神经信息处理系统大会）。25. Shinn, N., Labash, F., Gopinath, A.：Reflexion：通过语言反馈学习的语言智能体。arXiv 预印本 arXiv:2303.11366（2023）。26. Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, Cao Y（2023）ReAct：在语言模型中协同推理与行动。载于：国际学习表征会议。27. Henderson M 等（2017）用于智能回复的高效自然语言回复建议。载于：第 26 届万维网伴会论文集（WWW Companion）。28. Wang W, Wei F, Dong L, Bao H, Yang N, Zhou M（2020）MiniLM：面向任务无关压缩预训练变换器的深度自注意力蒸馏。载于：神经信息处理系统大会论文集（神经信息处理系统大会）。29. Song K, Tan X, Qin T, Lu J, Liu T-Y（2020）MPNet：用于语言理解的掩码与置换预训练。载于：神经信息处理系统大会论文集（神经信息处理系统大会）。30. Hofstätter S, Lin S-C, Yang J-H, Zeng X, Lin J, Hanbury A（2021）通过平衡主题感知采样，高效训练高性能稠密检索器。载于：SIGIR 论文集。