why coT falls short

论文笔记:《Rethinking Reasoning in Document Ranking: Why Chain-of-Thought Falls Short》

作者: Xuan Lu, Haohang Huang, Rui Meng, Yaohui Jin, Wenjun Zeng, Xiaoyu Shen
机构: Shanghai Jiao Tong University, Eastern Institute of Technology, Ningbo

1. 研究背景

1.1 问题定义

- 文档重排序 (Document Reranking) 是 IR 系统的关键组件，用于精炼检索结果，提升排序质量。

- 受大型推理模型 (LRMs, 如 DeepSeek-R1, o1) 的启发，近期研究开始将显式的"思维链"(CoT) 推理引入 LLM Reranker。

  > via supervised fine-tuning Weller et al. (2025); Ji et al. (2025); Yang et al. (2025) or using reinforcement learning Zhang et al. (2025); Zhuang et al. (2025); Liu et al. (2025).

- **核心问题：** 尽管大家_假设_ CoT 对排序有益，但这种有效性尚未被系统性地探索。CoT 是否真的能提升 Reranker 性能?

  > 之前((Rank1,rank-k,ReasonRank))没有与非推理基线模型的公平比较;
  > 新证据(Don't "Overthink" Passage Reranking: Is Reasoning Truly Necessary?-认为llm限制逐点的文章相关性 可能引入噪声) 但这些分析范围有限，仅关注使用监督目标训练的逐点reranker,并不系统

1.2 研究目标

- 本文旨在对 Reranker 中的"推理"进行**首次系统性研究**。

- **目标：** 公平地对比"推理" (CoT) 模型与"直接" (Direct-output) 模型在不同范式 (Pointwise, Listwise) 和不同训练策略 (SFT, RL) 下的性能。

- 解答关键问题：显式推理对 Reranker 来说是必要的吗？还是仅仅增加了不必要的推理成本?

约定
所有重排序器均在 MS MARCO 数据集上进行训练，并使用 DeepSeek-R1 生成的 CoT 链增强推理

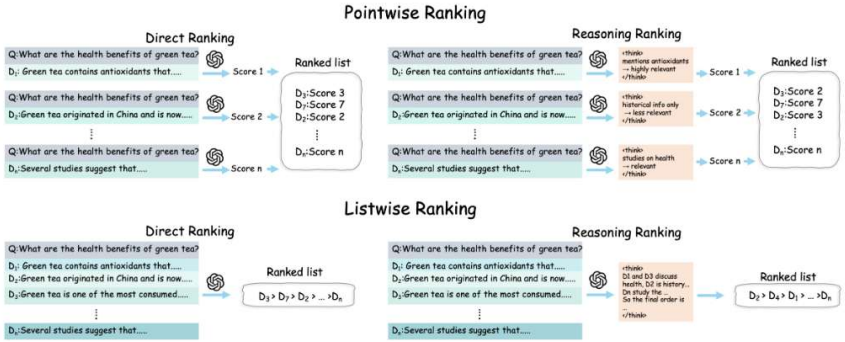
能力。
在两个互补的基准数据集上评估了模型：BRIGHT（侧重于推理密集型查询）和 BEIR（一套标准的检索数据集）

背景知识

2. 研究设计 (Study Design)

2.1 核心对比范式 (Paradigms)

- **Pointwise Reranker (逐点式):** 独立评估每个 `(Query, Document)` 对 。模型输出一个相关性分数 (e.g., "True" / "False" 的 logits) 。
- **Listwise Reranker (列表式):** 联合考虑整个候选文档集 。模型直接自回归地生成一个排序后的文档 ID 列表 (e.g., `[3] > [5] > [1]`) 。



2.2 核心对比模型 (Model Variants)

- **Direct-Point (非推理, 逐点):** 直接输出 "True" / "False" 决策 。
- **Reason-Point (推理, 逐点):** 先生成 CoT 理由，*然后_再输出 "True" / "False" 决策* 。
- **Direct-List (非推理, 列表):** 直接输出排序列表 (e.g., `<answer>[3] > [1]</answer>`) 。
- **Reason-List (推理, 列表):** 先在 `<think>` 标签中生成 CoT，*然后_再在 `<answer>` 标签中输出排序列表* 。

2.3 模型训练 (Training Details)

- **基座模型 (Backbone):** Qwen3-4B 和 Qwen3-8B 。

- **训练数据 (Data):** 基于 MS MARCO。Pointwise 使用 RANK1 语料库； Listwise 使用 REASONRANK 语料库 。CoT 理由(查询-段落对)由 DeepSeek-R1 生成 。
- **训练策略 (Regimes):**
  1. **SFT (监督微调,使用LLaMA-Factory3):** 使用交叉熵损失进行训练 。
     For Reason-Point, we perform supervised fine-tuning on quadruples ⟨query, passage, rationale, answer⟩. For Direct-Point, we ablate the rationale and fine-tuning on⟨query, passage, answer⟩, training the model to emit a single token in {TRUE, FALSE}
  2. **SFT + GRPO (强化学习):** 在 SFT 后，使用 GRPO (一种 RL 算法) 结合一个复合排序奖励 (NDCG@10, Recall@10, RBO) 进一步优化 。

3. 实验设计与结果

3.1 实验设置

- **评测基准 (Benchmarks):**
  1. **BRIGHT:** 一个需要强推理能力的 IR 数据集 。
  2. **BEIR:** 一个标准的、异构的 IR 评测集 。
- **对比模型 (Baselines):** Rank1-7B/14B , TFRank-4B/8B , Rank-R1-7B/14B , ReasonRank-7B 等。
- **核心指标 (Metric):** 主要指标是 NDCG@10

这是针对listwise范式的，第一阶段执行监督微调（SFT），以训练模型输出排序序列，第二阶段使用广义重加权策略优化（GRPO）（Guo等人，2025）对SFT模型进行优化

$$\$R_m = \mathrm{NDCG@10}(y^{\mathrm{list}}, y') + \phi \cdot \mathrm{Recall@10}(y^{\mathrm{list}}, y') + \gamma \cdot \mathrm{RBO}(y^{\mathrm{list}}, y')$$

最后指标: Rm
y^list 模型预测排名序列，y^' 标准排名(gold)，
NDCG@10衡量排名前十文档相关性和排序位置质量，Recall@10衡量前十中包含多少相关文档，
RBO强调顶部排名重叠度，

$$\$\mathrm{RBO}(y^{\mathrm{list}}, y') = (1-p)\sum_{d=1}^{|y^{\mathrm{list}}|} p^{d-1} \frac{|y^{\mathrm{list}}_{1:d} \cap y'_{1:d}|}{d}$$

权重参数ϕ和γ分别对覆盖率和重叠度进行加权

遵循 REASONRANK，使用简单的格式验证器对Rm 进行门控以稳定学习：

$$\$R = \begin{cases} R_m, & \text{both output and answer formats are valid,} \\ 0, & \text{only the output format is valid,} \\ -1, & \text{otherwise,} \end{cases}$$

GRPO目标函数优化更新策略

$$\$\mathcal{J}_{\text{GRPO}}(\theta) = -\sum_{i,t} \min(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_{i,t}) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$$

第一项（裁剪损失）： 目标是最大化加权优势 $\hat{A}_{i,t}$。通过 min 函数和 clip 函数，它**限制了模型参数更新的幅度**

第二项（正则化）： KL 散度项作为**正则化**项，惩罚当前策略 π_θ 离原始 SFT（Supervised Fine-Tuning）参考策略 π_{ref} 太远

3.2 实验结果 (见 Table 1 & 2)

- **核心结论：推理在重排序中是不必要的 (Reasoning Is Unnecessary)** 。

- 在所有模型大小 (4B/8B)、训练策略 (SFT/GRPO) 和评测基准 (BRIGHT/BEIR) 上，**Direct (无 CoT) 模型** 的性能**一致且显著地优于 Reason (有 CoT) 模型** 。

- **举例 (BRIGHT, 8B):** Direct-Point (26.8) 远超 Reason-Point (20.7) 。

- **举例 (BEIR, 4B):** Direct-Point (45.4) 远超 Reason-Point (40.1) 。

- 本文提出的 Direct 模型性能也优于所有现有的（基于 CoT 的）SOTA 基线 。

Table 1: Performance comparison on BRIGHT across different reranker variants. We report results for Direct-Point, Reason-Point, Direct-List, and Reason-List under both SFT and GRPO training, together with representative pointwise and listwise baselines.

Model	Training	StackExchange							Coding		Theorem-based			Avg.
		Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.	
Pointwise														
BM25	/	18.9	27.2	14.9	12.5	13.6	18.4	15.0	7.9	24.4	6.2	4.9	10.4	14.5
Rank1-7B	SFT	31.4	36.7	18.3	25.4	13.8	17.6	24.8	16.7	9.5	6.1	9.5	11.6	18.5
Rank1-14B	SFT	29.6	34.8	17.2	24.3	18.6	16.2	24.5	17.5	14.4	5.5	9.2	10.7	18.5
TFRank-4B	SFT+GRPO	33.2	45.9	17.6	29.5	21.0	20.9	18.3	25.0	9.1	9.5	9.8	7.3	20.6
TFRank-8B	SFT+GRPO	33.7	46.2	23.7	26.0	24.1	20.1	23.6	28.8	12.5	10.8	11.4	9.7	22.6
Reason-Point-4B	SFT	23.6	29.0	15.0	23.7	16.7	12.2	18.3	18.4	12.4	8.9	11.0	9.4	16.5
Direct-Point-4B	SFT	34.9	45.1	23.3	31.8	26.6	23.6	30.7	18.5	35.4	7.2	13.6	15.2	<u>25.5</u>
Reason-Point-8B	SFT	24.9	34.6	17.5	26.2	25.9	22.4	19.7	11.9	36.6	9.3	6.5	12.6	20.7
Direct-Point-8B	SFT	33.9	46.4	24.6	31.6	25.8	25.9	32.0	25.3	35.5	12.0	13.5	15.2	**26.8**
Listwise														
Rank-R1-7B	GRPO	26.0	28.5	17.2	24.2	19.1	10.4	24.2	19.8	4.3	4.3	8.3	10.9	16.4
Rank-R1-14B	GRPO	31.2	38.5	21.2	26.4	22.6	18.9	27.5	20.2	9.2	9.7	9.2	11.9	20.5
REARANK-7B	GRPO	23.4	27.4	18.5	24.2	17.4	16.3	25.1	27.0	8.0	7.4	7.9	9.5	17.7
ReasonRank-7B	SFT+GRPO	36.3	44.2	24.8	31.7	30.7	24.9	32.8	28.7	17.5	12.0	18.5	14.0	<u>26.4</u>
Reason-List-4B	SFT	30.7	37.3	18.7	27.7	27.9	19.8	28.5	28.1	13.7	9.1	13.9	13.3	22.4
Direct-List-4B	SFT	32.7	38.6	20.0	28.4	28.6	20.5	31.2	30.9	15.1	10.4	17.8	15.6	24.1
Reason-List-8B	SFT	31.9	39.6	22.4	29.0	29.9	23.4	34.5	26.8	18.9	9.7	15.6	12.1	24.5
Direct-List-8B	SFT	32.6	38.4	21.3	28.9	31.9	22.6	31.8	28.9	16.9	11.1	18.5	15.4	24.9
Reason-List-4B	SFT+GRPO	33.6	40.8	21.6	28.0	33.3	26.0	29.3	31.0	13.3	11.4	16.5	15.4	25.0
Direct-List-4B	SFT+GRPO	33.8	41.5	23.4	29.3	34.0	23.9	34.2	33.4	13.7	11.9	17.1	14.6	25.9
Reason-List-8B	SFT+GRPO	32.1	40.3	26.7	32.1	30.0	25.5	33.8	28.8	19.4	9.8	18.0	14.0	25.9
Direct-List-8B	SFT+GRPO	35.2	42.7	23.1	30.6	34.0	27.6	33.9	29.2	22.9	12.1	17.9	15.8	**27.1**

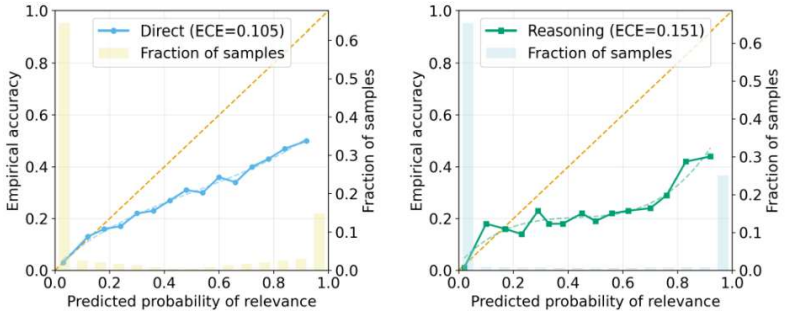
3.3 消融实验 (Failure Mode Analysis)

为什么 CoT 会失败?

- **Pointwise 失败原因：破坏校准度 (Breaks Calibration)** 。
  推理虽然可以提高相关性预测的准确性，但会破坏分数校准，并导致误报率上升，最终降低排序性能

1. **过度自信:** CoT 模型的预测概率与真实准确率严重脱节（见 Figure 2），ECE (预期校准误差 ECE 是所有区间内"平均置信度"与"实际准确率"之间差异的加权平均值) 更高 (0.151 vs 0.105，越低越好)，表明模型过度自信 。

2. **正类偏见:** (见 Table 3) CoT 模型更倾向于预测"True (相关)" 。这导致 TPR (真正率) 上升，但 **TNR (真负率) 暴跌** 。在负样本为主的 Reranker 任务中，这会导致大量**"假正例" (False Positives)** 被排到前面，严重损害 NDCG 。
   评估池:100pos 200neg



(a) Direct-Point-8B (b) Reason-Point-8B

Figure 2: Calibration curves of pointwise rerankers: predicted probabilities vs. empirical accuracies.

Table 3: Class-conditional performance on pointwise rerankers. We report TPR (%) and TNR (%).

Model	Biology		MS MARCO		Avg.
	TPR	TNR	TPR	TNR	
DeepSeek-R1	**52.4**	96.1	**40.8**	85.1	**68.6**
Reason-Point-4B	43.7	91.3	38.7	79.4	63.3
Direct-Point-4B	34.0	93.2	30.7	85.7	60.9
Reason-Point-8B	50.5	98.1	35.9	85.5	67.5
Direct-Point-8B	31.1	**100.0**	25.5	**94.2**	62.7

Table 4: Listwise (GRPO) performance on MS MARCO (NDCG@10).

Model	MS MARCO	
	DL19	DL20
Direct-List-4B	**73.77**	68.97
Reason-List-4B	70.76	68.71
Direct-List-8B	73.00	**71.38**
Reason-List-8B	72.60	69.81

- **Listwise 失败原因：损害泛化性 (Hurts Generalization)** 。
  推理虽然可以增强领域内训练的拟合度，但也会增加预测方差，并损害领域外泛化能力，

即使通过 GRPO 缩短推理过程也是如此。

1. **过拟合:** (见 Figure 3) 在**训练集**上，CoT 模型的 NDCG@10 反而**更高** (e.g., 87.55 vs 86.93) 。但离散度更大

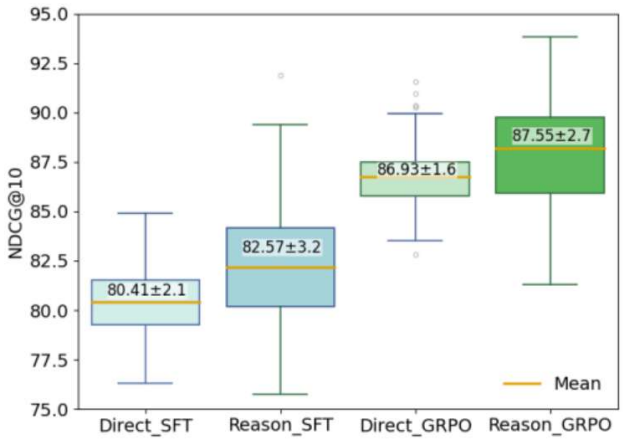


Figure 3: Training-split listwise performance of four 8B variants. Reasoning improves mean NDCG@10 but increases variance.

2. **泛化失败:** 这种在训练集上的优势**无法泛化**到（OOD的）BRIGHT/BEIR 甚至（In-Domain 的）MS MARCO 测试集 。
   Taable4上面

3. **RL 的影响:** GRPO (RL) 确实能缓解"过度思考"(Overthinking)，将 CoT 长度从 397 压缩到 172 ，但 CoT 模型的泛化性依然不如 Direct 模型 (偷懒)。Taable4上面

4. 贡献与启示

4.0 相关工作(LLm 和 LRM的进展，有需要可以浏览)

4.1 主要贡献

1. **首次系统性研究:** 第一次对 CoT 在 Reranker 中的作用进行了大规模、受控的公平对比 。
2. **清晰的负面证据:** 提供了明确证据，表明 CoT（显式推理）**有害** Reranker 性能，且成本高昂 。
3. **深入的失败分析:** 揭示了 CoT 失败的两种不同机制：Pointwise (破坏校准度) 和 Listwise (损害泛化性) 。

4.2 启示 (Implications)

1. **回归 Direct:** Reranker 任务应优先考虑**高效、鲁棒**的 Direct 打分/排序模型，而不是复杂的推理模型 。
2. **CoT 不万能:** 显式推理在很多 NLP 任务中有用，但**不能想当然地认为它对所有任务都有益** 。

5. 未来工作方向

- **Pointwise:** 研究"校准度感知"(Calibration-Aware) 的打分目标，以修复 CoT 模型的过度自信问题 。Evaluation
- **Listwise:** 探索"简洁、有针对性"(Concise, Targeted) 的推理策略，而不是依赖又长又导致过拟合的 CoT，以平衡可解释性与泛化性 。
  设计推理策略

6. 结论

本文系统地评估了 CoT 在 Pointwise 和 Listwise 重排序中的作用 。研究发现，与"直接输出"的模型相比，**CoT 模型性能更差、成本更高** 。

失败的根本原因在于：(i) 在 Pointwise 中，CoT 破坏了模型的**校准度**并引入了**正类偏见** ；(ii) 在 Listwise 中，CoT 导致**过拟合**，损害了**泛化能力** 。

结论是：Direct (非推理) 模型更稳定、更有效 。

关键词: 文档重排序 (Document Reranking), 思维链 (CoT), Pointwise, Listwise, 校准度 (Calibration), 泛化性 (Generalization), 强化学习 (GRPO)

coT会造成过拟合的原因是因为什么

过度依赖训练数据中的推理模式和风格

- **模仿而非真正的推理：** CoT通过让模型生成一系列中间步骤来解决问题，但模型学习到的可能不是深层的逻辑原理，而是**模仿**训练样本中特定的推理**风格或模式**。
- **缺乏泛化性：** 如果训练数据中的CoT示例具有特定的格式、措辞或解题路径，模型可能会过度适应这些特定的"思维链"，导致在遇到具有不同风格、稍微改变格式或需要不同解题策略的新问题时，表现下降（即泛化能力差）。

CoT导致过拟合的核心原因在于，模型可能学会了**表面**的、特定于训练数据的 **"解题套路"** ，而不是

获得可泛化到未知情境的**深层逻辑推理能力**。这就像一个学生只记住了某些题目的具体解法步骤，换个问法就不会做了一样。

discussion

任务不匹配——Reranker 究竟是"判别"任务还是"生成"任务?

- CoT 擅长的是**"生成式" (Generative) 任务**，其"思考过程"本身就是答案的一部分，或者对推导出答案至关重要 。

- 而 Reranker 本质上是一个**"判别式" (Discriminative) 任务**。它的目标不是"生成"一个理由，而是"判断"一个（Query, Doc）对的相关性（Pointwise），或是"对比" N 个文档的优劣（Listwise）

评估的悖论——我们是否用错了"尺子"?

我们看的**综述（2308...）在 8.6 节**明确提出了一个未来方向，即传统的 nDCG 指标已经不够用了 。在 RAG 时代，我们需要的是**"面向生成的排序评估" (Generation-oriented ranking evaluation)** 。

- NDCG@10 只关心文档的**"主题相关性"** 。

- 而"面向生成的评估"关心的是文档是否**"适合生成答案"**（**比如，事实清晰、论证严谨**）。

"思考"的矛盾——"过度思考" vs "测试时扩展"

真正的未来——"简洁推理" (连接论文的 Future Work)

- Pointwise 的未来：它提议"校准度感知的打分" (calibration-aware scoring) 。

- Listwise 的未来：它提议"设计**简洁且有针对性的推理策略**" (design of concise, targeted reasoning strategies)