TabRank: Chain-of-Thought Distillation for Table Re-Rankers

Anonymous EMNLP submission

Abstract

The ability to retrieve relevant tables for answering questions is a key task for structured information retrieval. Multi-stage retrieval systems rely heavily on rerankers to refine candidate lists produced by efficient first-stage retrievers. As a result, neural rerankers and LLM-based reranking methods have become increasingly important due to their superior capacity for semantic understanding and reasoning compared to conventional sparse or dense retrieval models. Recently, Large Reasoning Models (LRMs) equipped with explicit chain-of-thought (CoT) reasoning have shown strong improvements in ranking quality in unstructured passage retrieval.

In this work, we present **TabRank**, a framework for training reasoning rerankers for Tabular Retrieval. We first present a comprehensive dataset of 6728 reasoning traces for tabular reranking on the Natural Questions Tables dataset. We then explore two variants of training a compact reasoning model on these reasoning traces: explicit CoT distillation and conditioning the student reranker on the teacher's reasoning trace within the prompt. We stress-test **TabRank** on several out-of-distribution generalization settings on diverse domains and multi-table scenarios. Our approach significantly improves performance across a variety of table retrieval datasets, increasing Acc@10 by 30.5% on HybridQA, 15.2% on SQA, 52.9% on TabFact, and 13.1% on TATQA subsets of the Multi-Table QA Benchmark compared to the base model. Notably, TabRank generalizes effectively to multi-table reasoning. Our code, data and models are available at https://anonymous.4open.science/r/TabRanker-46DC/

1 Introduction

Multi-stage retrieval has become the dominant architecture for open-domain question answering over structured data: a lightweight first-stage retriever produces a coarse candidate set using embedding-based methods, which a reranker refines before passing results to a reader or generator any downstream task. Because the reranker operates at the critical bottleneck between retrieval and reasoning, its effectiveness directly governs end-to-end system performance. Table retrieval, however, presents challenges fundamentally different from those of unstructured passages. Their semantics are encoded not only in surface lexical content, but also in structural properties such as column schemas, row alignments, and inter-cell relationships. Consequently, reranking models need to capture the relational and compositional signals important for effective table retrieval. that has received relatively limited attention in existing reranking literature.

The broader reranking literature has instead focused almost exclusively on passage retrieval, driving a wave of neural reranking research spanning pointwise, pairwise, and listwise paradigms, from cross-encoder architectures such as MonoT5 (Nogueira et al., 2020) and RankT5 (Zhuang et al., 2023), to instruction-tuned generative rerankers including RankGPT (Sun et al., 2023) and RankZephyr (Pradeep et al., 2023b), as well as contrastive approaches that leverage hard negatives to learn fine-grained relevance distinctions.

A major shift emerged with the introduction of Large Reasoning Models (LRMs). Systems such as OpenAI's *o1* (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025) demonstrated that models trained with reinforcement learning to generate extended chain-of-thought (CoT) reasoning before producing answers achieve substantially stronger performance on complex multi-step tasks. The information retrieval community rapidly adopted this paradigm. Rank1 (Weller et al., 2025) showed that fine-tuning on LRM-generated ranking rationales yields strong reranking performance with notable out-of-distribution generalization gains, while sub-

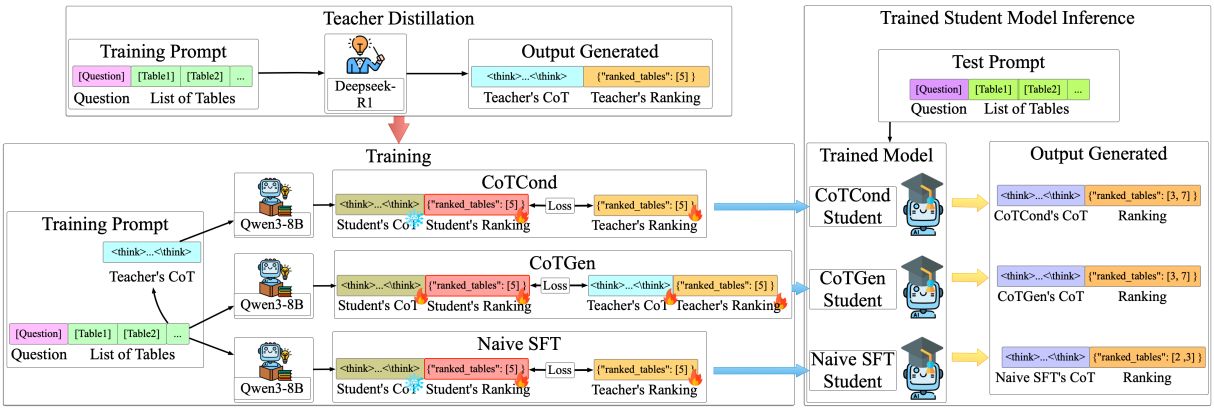

Figure 1: Teacher distillation and CoTCond, CoTGen, and Naive SFT training and inference pipeline for table ranking. DeepSeek-R1 produces teacher chain-of-thought and ranking labels from the training prompt, which guide the training of CoTCond, CoTGen, and Naive SFT students. The ❄ icon marks components excluded from loss computation, while the 🔥 icon marks components used for loss computation.

sequent approaches such as ReasonRank (Liu et al., 2025) and Rank-R1 (Zhuang et al., 2025) further integrated reasoning via reinforcement learning objectives and rationale-guided supervision. However, this line of work has exposed a fundamental tension regarding how reasoning should be incorporated into reranking models. The dominant paradigm, which we refer to as **CoTGen distillation**, trains a student model to autoregressively reproduce the teacher's complete reasoning trace as a supervised target alongside the final prediction. While effective in-domain, recent studies suggest that this approach may encourage overfitting to dataset-specific reasoning trajectories. (Lu et al., 2025) demonstrated that such models exhibit degraded cross-domain transferability, while (Jedidi et al., 2025) show that reasoning-based rerankers tend to produce overly polarized relevance estimates, negatively affecting partial-relevance calibration. These risks are especially acute for table retrieval, where reasoning over structural properties differs substantially from passage-level reasoning, and where domain shift, across table layouts, schemas, and question types, is the norm rather than the exception.

In this work, we address these limitations by introducing **TabRank**, a reranking model for structured tabular retrieval that substantially outperforms both standard supervised fine-tuning and CoTGen distillation in generalizing to out-of-distribution table settings, including multi-table retrieval and financial-domain datasets such as TAT-QA. TabRank employs a conditional reasoning distillation framework for listwise table reranking: rather than supervising the model to generate teacher reasoning traces token-by-token, TabRank prepends DeepSeek-R1-generated reasoning tokens directly into the input prompt and conditions the reranker on this reasoning context while computing loss only over the final ranking output. This formulation encourages the model to leverage intermediate reasoning signals without explicitly imitating the teacher's reasoning trajectory, thereby decoupling structured reasoning benefits from the overfitting risks associated with reasoning trace generation.

We train TabRank using single-table supervision and evaluate its generalization across four diverse table reasoning benchmarks, including out-of-distribution settings and multi-table retrieval tasks. Our contributions are summarized as follows:

- We show that TabRank's conditional CoT distillation, treating teacher reasoning as contextual input rather than an autoregressive generation target, improves Acc@10 by 30.5% on HybridQA, 15.2% on SQA, 13.1% on TaTQA and 52.9% on TabFact subsets of Multi-Table QA Benchmark relative to the base model in out-of-distribution settings.
- We demonstrate that TabRank although trained exclusively with single-table supervision generalizes naturally to multi-table retrieval without architectural modification.
- We release a comprehensive data including distilled DeepSeek-R1 reasoning traces and supervision signals used for fine-tuning, to support reproducibility and future research on reasoning-aware table retrieval.

2 Related Work

Neural reranking has progressed through cross-encoders that jointly encode query–passage pairs for relevance scoring. MonoT5 (Nogueira et al., 2020) framed reranking as sequence-to-sequence generation, while RankT5 (Zhuang et al., 2023) strengthened this with direct ranking loss optimization. The advent of large language models introduced listwise reranking: RankGPT (Sun et al., 2023) demonstrated that prompting LLMs to output passage permutations is surprisingly competitive in a zero-shot setting, and open-source models such as RankVicuna (Pradeep et al., 2023a) and RankZephyr (Pradeep et al., 2023b) distilled this capability at a fraction of the cost. FIRST (Reddy et al., 2024) further reduced inference overhead by scoring from the first generated token.

The recent integration of chain-of-thought (CoT) reasoning into reranking marks a qualitative shift in this trajectory. ReasonIR (Shao et al., 2025) incorporated reasoning at the retrieval stage, ReasonRank (Liu et al., 2025) combined rationale supervision with reinforcement learning ranking rewards, and Rank-R1 (Zhuang et al., 2025) optimized reranking end-to-end through answer-aware reinforcement learning, reducing reliance on supervised reasoning traces altogether.

Despite this momentum, a critical limitation of the dominant CoT distillation paradigm has emerged. (Lu et al., 2025) show that training models to reproduce teacher reasoning traces improves in-domain performance but degrades cross-domain generalization, while (Jedidi et al., 2025) show that inference-time reasoning produces overly polarized relevance scores that hurt partial-relevance estimation. **TabRank** addresses both failure modes: we are the first to bring reasoning-augmented reranking to structured tabular data, and we propose conditional reasoning distillation as a principled alternative that retains the benefits of structured reasoning while avoiding the generalization costs of trace imitation.

3 Problem Setup

Given a natural language query q and a set of k candidate tables returned by a first-stage retriever,

$$\mathbf{T} = \{t_1, t_2, \ldots, t_k\},$$

the goal of listwise table reranking is to learn a reranker $f_\theta(q, \mathbf{T})$ that produces a ranked ordering

$$\hat{\sigma} = (\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_k),$$

such that relevant tables are assigned higher positions in the ranking. Each candidate table t_i is associated with a binary relevance label $y_i \in \{0, 1\}$, where $y_i = 1$ indicates that table t_i contains the information required to answer query q.

For the single-table retrieval setting, each query is associated with exactly one relevant table. Let g denote the index of the gold table. Then,

$$y_g = 1, \qquad \sum_{i=1}^{k} y_i = 1.$$

In contrast, for multi-table retrieval, multiple candidate tables may jointly contribute to answering the query.

The reranker is trained to maximize the ranking quality of relevant tables within the candidate list. We evaluate reranking performance using Recall@K, nDCG@K, and Accuracy@K. Details regarding these metrics are provided in Appendix A.

4 Data Generation

4.1 Training Data Sampling

We use the NQ-Tables dataset (Herzig et al., 2021) and construct training samples from its training split, which contains 9,574 queries. Since our training setup focuses exclusively on single-table retrieval, each training query contains exactly one relevant table.

For each query, we construct a ranked list of candidate tables by combining the outputs of three retrieval pipelines using Reciprocal Rank Fusion (RRF): the lexical matching algorithm BM25, the sparse embedding model SPLADE-V3 (Lassance et al., 2024), and the dense embedding model all-mpnet-base-v2 (Song et al., 2020). Each training sample contains between 10 and 20 tables, where the exact number is uniformly sampled. We retain only samples where the gold table appears within the retrieved candidate set. For every query–candidate pair, we construct an instruction-style prompt consisting of the query and the ranked candidate table list. The generated prompt is subsequently passed to a teacher reasoning model for synthetic reasoning generation.

4.2 Data Generation

We generate synthetic reasoning traces using DeepSeek-R1, leveraging its explicit reasoning capabilities to produce intermediate ranking rationales before generating the final ranked output. The

teacher model is prompted with the query and candidate tables and instructed to reason within a dedicated <think> block prior to producing the final ranking. Our prompt is shown in Figure 4.

Unlike prior reranking formulations that rely on pairwise or pointwise reasoning, our prompt structure naturally extends to both single-table and multi-table retrieval settings without requiring architectural modifications. This unified formulation enables the model to generalize to multi-table retrieval despite being trained only on single-table supervision.

We generate training data using temperature 0.7, maximum generation length of 8192 tokens. We intentionally allow long generation lengths to encourage deeper reasoning before producing the final ranking output.

After generation, we apply several filtering steps by removing examples containing fewer than 9 candidate tables, samples exceeding 25,000 total tokens, incomplete generations, and malformed rankings or duplicated predictions. The final filtered dataset contains 6,728 queries, with an average of 20.06 candidate tables and 2,304 reasoning tokens per query.

5 Reasoning Distillation Strategies

We finetune the base model using the generated reasoning data under three distinct training paradigms.

Naive SFT. The first baseline, which we refer to as Naive Distillation (Naive SFT), trains the base model using only the final ranked output generated by the teacher model. The reasoning traces are removed entirely, and the model is optimized solely to predict the final ranking sequence.

Formally, given input $(q, \mathbf{T})$ and target ranking $\hat{\sigma}$, the training objective minimizes the autoregressive loss over ranking tokens:

$$\mathcal{L}_{\text{rank}} = -\sum_{t=1}^{n} \log p_\theta(\hat{\sigma}_t \mid q, \mathbf{T}, \hat{\sigma}_{<t}). \quad (1)$$

This setup resembles standard supervised finetuning for reranking without explicit reasoning supervision.

Standard Chain-of-Thought Distillation. Our second setup follows standard chain-of-thought distillation approaches commonly used in reasoning models. We denote this method as **CoTGen**.

Here, the model is trained to explicitly generate both the reasoning trajectory and the final ranking output. The generated <think> tokens are treated as supervised targets, and loss is computed jointly over reasoning and ranking tokens:

$$\mathcal{L}_{\text{CoTGen}} = \mathcal{L}_{\text{reason}} + \mathcal{L}_{\text{rank}}. \quad (2)$$

Conditional Reasoning Distillation. We propose **CoTCond**, a conditional reasoning distillation framework for reranking. Instead of training the model to generate reasoning traces, we prepend teacher-generated thinking tokens directly into the input prompt and condition the model on the reasoning context during training.

Formally, the model input becomes:

$$x = [q; \mathbf{T}; r], \quad (3)$$

where r denotes the teacher-generated reasoning trace. The model is trained only to predict the final ranking conditioned on the provided reasoning:

$$\mathcal{L}_{\text{CoTCond}} = -\sum_{t=1}^{n} \log p_\theta(\hat{\sigma}_t \mid q, \mathbf{T}, r, \hat{\sigma}_{<t}). \quad (4)$$

Importantly, no loss is computed over reasoning tokens themselves. The reasoning sequence acts purely as contextual conditioning information rather than an autoregressive generation target.

This training formulation differs fundamentally from standard CoT distillation. By removing the requirement to imitate teacher reasoning token-by-token, the model is encouraged to learn ranking-relevant abstractions rather than memorizing dataset-specific reasoning trajectories. We observe that this significantly improves cross-domain generalization while also reducing inference-time reasoning overhead.

Additionally, CoTCond produces substantially shorter reasoning traces at inference time, resulting in lower latency and reduced generation cost compared to CoTGen-style models.

6 Training Details

All models are initialized from the same pretrained checkpoint and finetuned using LoRA adapters with identical optimization settings for fair comparison.

For CoTCond, teacher-generated reasoning traces are included in the prompt but excluded from the loss computation, so the model learns only from

the final ranking tokens while still conditioning on the reasoning context. CoTGen instead optimizes over both reasoning and ranking tokens, requiring the model to generate reasoning sequences during training and inference. Naive SFT removes reasoning entirely and trains only on the final ranking output. Overall, CoTCond provides reasoning guidance without requiring explicit reasoning generation, leading to better generalization and lower inference cost than CoTGen.

All reranker models are finetuned using the LLaMA-Factory framework with LoRA adaptation. Training is performed on 2 NVIDIA H200 GPUs for approximately 18 hours using an effective batch size of 64 and learning rate of 5×10^{-5}.

7 Result and Analysis

Table 1 reports out-of-distribution reranking performance across HybridQA, SQA, TAT-QA, and TabFact subsets of Multi-Table QA Benchmark (Zou et al., 2025). Details regarding these datasets are provided in Appendix C.

Across all four datasets, CoTCond is the best-performing method. The base model here is Qwen3-8B (Team, 2025) fine-tuned exclusively on NQ-Tables, a Wikipedia-derived single-table benchmark. Three of the four evaluation datasets (HybridQA, SQA, TabFact) share this Wikipedia provenance, while TAT-QA is drawn from financial annual reports and represents a fundamentally different table structure, reasoning style, and domain. Despite this, CoTCond generalizes consistently across all four settings, which we discuss in turn.

Magnitude of the CoTCond gains. CoTCond improves substantially over the base reranker on all benchmarks. The largest relative gain appears on **TabFact**, where Accuracy@10 increases from 0.3816 to 0.5835, corresponding to a 52.9% improvement. **HybridQA** shows the next largest Accuracy@10 gain at 30.5%, rising from 0.6200 to 0.8091. **SQA** and **TAT-QA** show smaller but still meaningful gains, with Accuracy@10 improving by 15.2% and 13.1%, respectively.

These gains are strongest on the datasets where the base reranker struggles most. TabFact begins with the lowest base Accuracy@10, and CoTCond produces the largest relative improvement. **TAT-QA** also presents a challenging transfer setting, yet CoTCond remains the only method to improve every metric decisively. This behavior suggests that conditional reasoning helps most when lexical or surface-level matching provides a weak signal, and the reranker must rely on structural or compositional evidence.

Conditioning outperforms generation. The comparison between CoTGen and CoTCond isolates the role of the distillation objective. Both methods use the same teacher-generated reasoning data, but they present it to the student in different ways. CoTGen trains the model to generate both the reasoning trace and the final ranking. CoTCond instead places the reasoning trace in the input and computes loss only over the final ranking. This single design change produces gains on every benchmark.

On **HybridQA**, CoTCond improves over CoTGen from 0.7659 to 0.7833 on Recall@5 and from 0.7624 to 0.8091 on Accuracy@10. On **SQA**, Recall@10 rises from 0.8446 to 0.8840. On **TAT-QA**, Recall@5 rises from 0.5124 to 0.5520 and nDCG@10 rises from 0.5203 to 0.5459. On **TabFact**, CoTCond improves nDCG@10 from 0.6661 to 0.6890 and Accuracy@10 from 0.5614 to 0.5835. The repeated margin over CoTGen shows that reasoning supervision helps most when it acts as a conditioning context rather than a token sequence to imitate.

This result challenges the assumption that richer token-level supervision necessarily improves reasoning distillation. In this setting, forcing the student to reproduce long teacher traces appears to dilute the ranking objective. The model must allocate capacity to matching teacher phrasing, intermediate steps, and potentially dataset-specific reasoning habits. CoTCond avoids that pressure. It gives the model access to the teacher's reasoning signal while allowing optimization to concentrate on the ranking decision. The result is a cleaner learning signal for reranking.

Naive distillation exposes the value of reasoning context. Naive_SFT provides a useful lower bound on distillation without reasoning. It trains only on the teacher's final ranking outputs, discarding the reasoning traces entirely. This strategy improves over the base model on several metrics, but its gains are weaker and less stable than those of the reasoning-based methods. **TAT-QA** shows the clearest failure mode. Naive_SFT slightly improves Recall@5 and Recall@10, but nDCG@5 drops by 1.2% and nDCG@10 drops by 1.8% rela-

Dataset	Method	Recall @5	Recall @10	nDCG @5	nDCG @10	Acc @10	Metadata Avg Tokens	Metadata Fails
HybridQA	-	0.7649	0.8488	0.7041	0.7393	0.7827	-	-
	Base Model	0.6921	0.7617	0.6771	0.7072	0.6200	2369	66
	Naive SFT	0.7371 Δ+6.5%	0.8253 Δ+8.3%	0.7052 Δ+4.2%	0.7434 Δ+5.1%	0.7157 Δ+15.4%	1241	175
	CoTGen	0.7659 Δ+10.7%	0.8526 Δ+11.9%	0.7322 Δ+8.1%	0.7698 Δ+8.9%	0.7624 Δ+23.0%	**2983**	124
	CoTCond	**0.7833** Δ+13.2%	**0.8849** Δ+16.2%	**0.7432** Δ+9.8%	**0.7864** Δ+11.2%	**0.8091** Δ+30.5%	2634	13
SQA	-	0.6745	0.7432	0.6018	0.6326	0.6892	-	-
	Base Model	0.6914	0.8041	0.6483	0.6951	0.7095	1807	0
	Naive SFT	0.7038 Δ+1.8%	0.8423 Δ+4.8%	0.6575 Δ+1.4%	0.7141 Δ+2.7%	0.7770 Δ+9.5%	1282	**9**
	CoTGen	0.7038 Δ+1.8%	0.8446 Δ+5.0%	0.6633 Δ+2.3%	0.7205 Δ+3.7%	0.7703 Δ+8.6%	**3318**	0
	CoTCond	**0.7387** Δ+6.8%	**0.8840** Δ+9.9%	**0.6782** Δ+4.6%	**0.7381** Δ+6.2%	**0.8176** Δ+15.2%	2145	0
TAT-QA	-	0.4107	0.4788	0.3774	0.4037	0.3260	-	-
	Base Model	0.5014	0.5953	0.4657	0.5051	0.3785	2794	19
	Naive SFT	0.5152 Δ+2.8%	0.6017 Δ+1.1%	0.4601 Δ−1.2%	0.4959 Δ−1.8%	0.3867 Δ+2.2%	1297	29
	CoTGen	0.5124 Δ+2.2%	0.6234 Δ+4.7%	0.4751 Δ+2.0%	0.5203 Δ+3.0%	0.4144 Δ+9.5%	**3704**	23
	CoTCond	**0.5520** Δ+10.1%	**0.6639** Δ+11.5%	**0.5000** Δ+7.4%	**0.5459** Δ+8.1%	**0.4282** Δ+13.1%	2473	**0**
TabFact	-	0.4633	0.5342	0.4536	0.4846	0.3444	-	-
	Base Model	0.5543	0.6302	0.5659	0.6003	0.3816	2193	489
	Naive SFT	0.5848 Δ+5.5%	0.6780 Δ+7.6%	0.5810 Δ+2.7%	0.6220 Δ+3.6%	0.4716 Δ+23.6%	1365	825
	CoTGen	0.6277 Δ+13.3%	0.7342 Δ+16.5%	0.6190 Δ+9.4%	0.6661 Δ+10.9%	0.5614 Δ+47.1%	**3716**	583
	CoTCond	**0.6490** Δ+17.1%	**0.7678** Δ+21.8%	**0.6368** Δ+12.5%	**0.6890** Δ+14.8%	**0.5835** Δ+52.9%	2454	**45**

Table 1: Out-of-distribution retrieval performance on the MultiTableQA dataset. The base model is Qwen3-8B trained exclusively on the NQ-Tables training set. Δ% is relative to Base. Best values per dataset are **bolded**, second-best are underlined. All rerankers rerank the top 25 tables from the first stage retriever.

tive to the base model. This combination suggests that the model retrieves more relevant candidates somewhere in the list while placing them less effectively near the top.

The contrast with CoTCond indicates that teacher reasoning contains information that final rankings alone do not transmit. Ranking labels describe the desired order, but they do not explain which table fields, rows, schema elements, or relational cues justify that order. Conditional reasoning supplies this missing scaffolding during training. Because CoTCond masks reasoning tokens from the loss, the model can use this scaffolding without having to learn to reproduce it verbatim.

Token efficiency. CoTCond also improves the accuracy and reliability profile without incurring the full generation cost of CoTGen. Across all four datasets, CoTGen produces the longest outputs, ranging from 2983 tokens on HybridQA to 3716

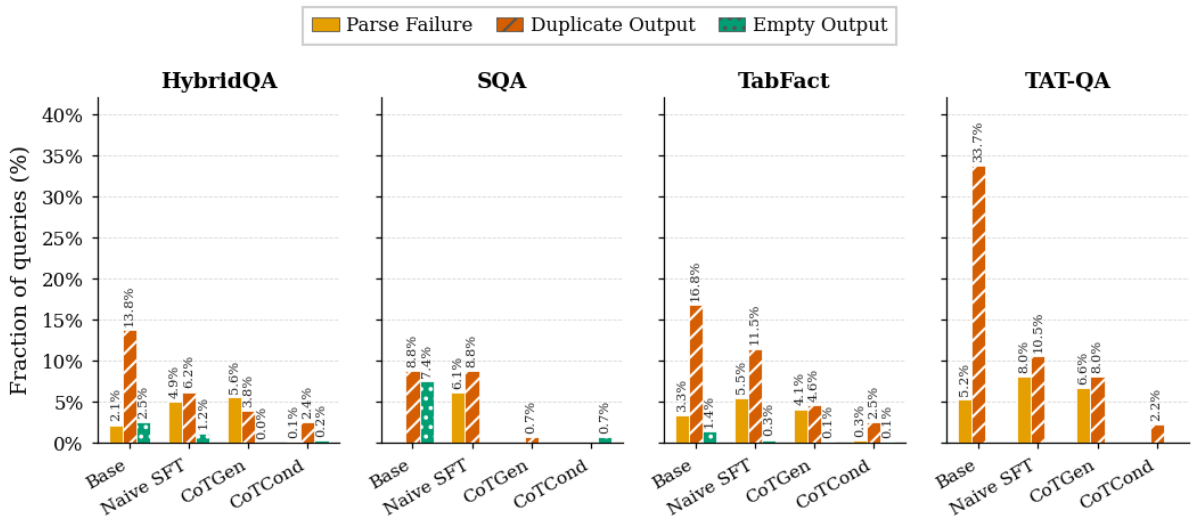

Figure 2: Distribution of per-query outcomes across datasets and reranking methods. Each bar decomposes model behavior into correct ranking, parse failures, empty outputs, and duplicate outputs.

tokens on TabFact. CoTCond uses fewer tokens than CoTGen on every benchmark while achieving better retrieval performance. The savings are substantial on TAT-QA, where CoTCond uses 2473 tokens compared with 3704 for CoTGen, and on TabFact, where it uses 2454 tokens compared with 3716.

Naive_SFT remains the cheapest method in token count, but its lower ranking quality and higher failure rates make that efficiency less useful. CoTCond occupies a stronger point on the accuracy-cost comparison. It spends more tokens than Naive_SFT, but it buys substantially higher ranking quality and far better output validity. It also spends far fewer tokens than CoTGen while outperforming it across all metrics.

8 Error Analysis

Aggregated retrieval metrics conceal the structural failures that drive the gains in Table 1. Figure 2 therefore decomposes predictions into four mutually exclusive categories: correct rankings, parse failures, empty outputs, and duplicate predictions.

Two patterns emerge. First, the base reranker suffers from substantial output reliability failures, especially on **TabFact** and **TAT-QA**. Duplicate predictions dominate these errors, reaching 16.8% on TabFact and 33.7% on TAT-QA. This explains the gap between the base model's recall and ranking accuracy: relevant tables are often retrieved, but the generated rankings are structurally invalid.

Second, reasoning supervision substantially improves output reliability. **CoTGen** sharply reduces duplicate and empty outputs across all datasets, suggesting that reasoning traces regularize the decoding process and encourage more coherent ranked lists. However, parse failures remain persistent on more difficult datasets such as TAT-QA and TabFact, likely because long autoregressive reasoning chains increase opportunities for formatting drift.

CoTCond achieves the cleanest reliability profile overall. It nearly eliminates structural failures across all benchmarks, reducing duplicate predictions to 2.2% on TAT-QA and below 3% on every other dataset. By conditioning on teacher reasoning instead of generating it token-by-token, CoTCond preserves the benefits of reasoning supervision while avoiding the long generation chains responsible for many CoTGen failures.

Overall, the results show that most improvements from reasoning distillation come from eliminating structural failures rather than refining fine-grained ranking order. Conditioning-based distillation provides the strongest balance between ranking quality and output reliability.

Overall, the error analysis shows that the metric improvements attributed to CoTCond in Table 1 stem from eliminating structural failures rather than from marginal reordering gains. Once malformed outputs, empty rankings, and duplicate predictions are suppressed, the reranker reliably places at least one gold table in the top-10 on the vast majority of queries. This finding reframes the contribution of reasoning distillation for listwise reranking. Rea-

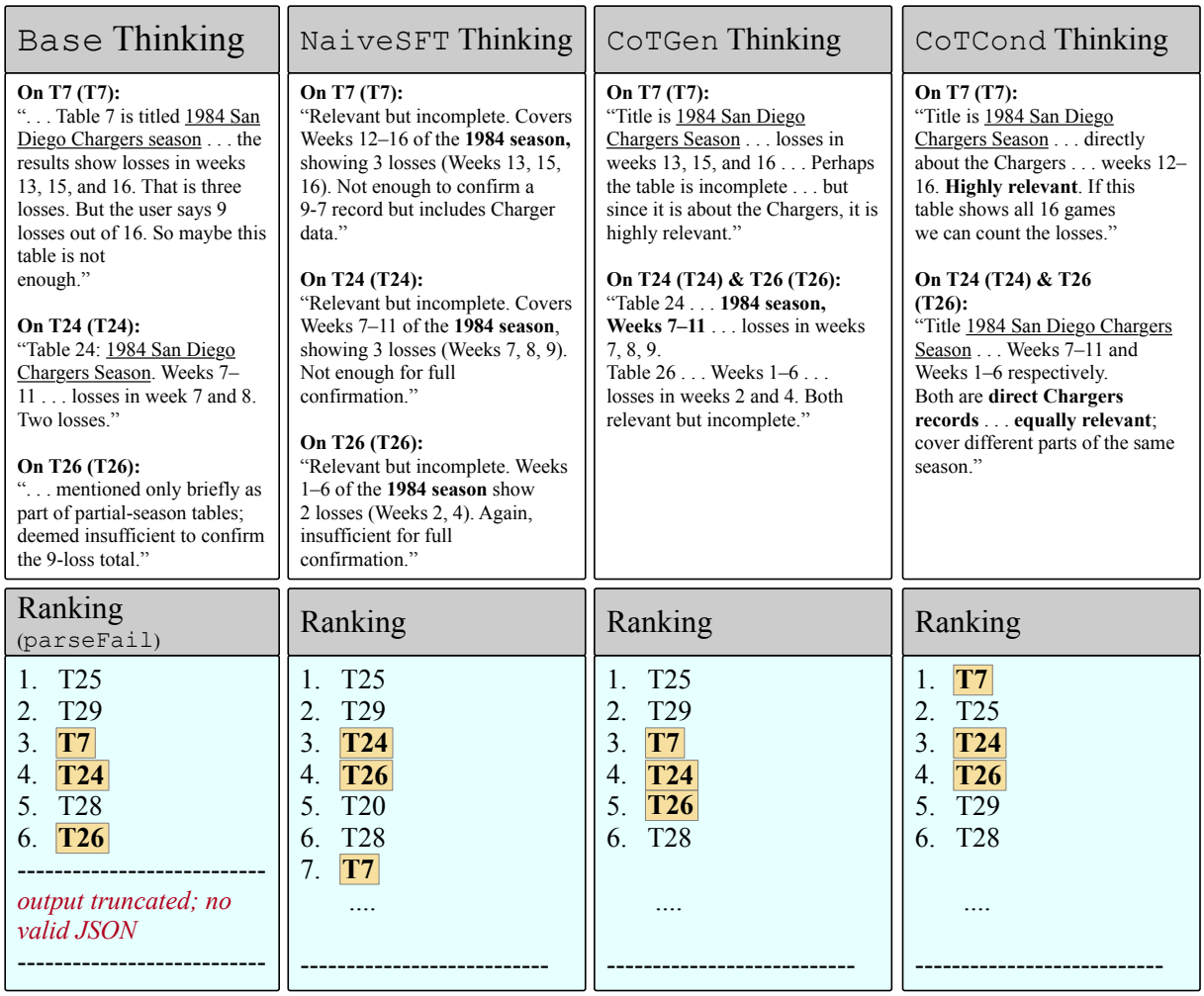

Figure 3: Chain-of-thought reasoning excerpts (top rows) and produced table rankings (bottom rows) across four prompting strategies for a query from TabFact (downstream task is fact verification). The input claim is: *"The San Diego Chargers lost 9 games out of 16."* Each column shows the model's reasoning about the three gold tables **T7 (id 4381)** , **T24 (id 4380)** , **T26 (id 4379)** . All three gold tables are partitions of the same *1984 San Diego Chargers Season* table, containing weeks 1–6, 7–11, and 12–16 respectively.

soning supervision matters primarily because it teaches the model to produce coherent ranked lists over long candidate sets, and CoTCond captures that benefit without paying the generation cost of CoTGen. A qualitative ranking example for this is provided in Figure 3.

9 Conclusion

In this paper, we present **TabRank**, a reasoning-aware framework for table reranking that includes a dataset of 6,728 synthetic reasoning traces for NQ-Tables Reranking and a set of compact distilled reranking models for tabular retrieval. We demonstrate that incorporating reasoning supervision significantly improves retrieval quality across a diverse collection of out-of-domain and multi-table question answering benchmarks. In particular, our proposed conditional reasoning distillation strategy allows models to leverage teacher-generated reasoning without explicitly reproducing long reasoning traces, resulting in better generalization, stronger ranking accuracy, and lower inference cost.

In addition to improving ranking performance, TabRank also reduces structural generation failures such as malformed outputs and duplicate predictions, leading to more reliable reranking behavior. Overall, our results highlight the effectiveness of conditional reasoning for general-purpose table reranking and suggest promising future directions for reasoning-enhanced retrieval systems in tabular question answering and structured information retrieval.

10 Limitations

Our work focuses exclusively on text-based table retrieval and assumes that tables are available in a structured serialized format. Many real-world tables, however, are embedded in scanned documents, PDFs, spreadsheets, or images where layout and visual structure play a significant role. Extending reasoning-based reranking to multimodal table representations remains an important direction for future work.

Additionally, our training setup relies on synthetic reasoning traces generated by a single teacher model. While these traces provide effective supervision, they may also inherit biases or reasoning artifacts specific to the teacher. Exploring more diverse teacher models or reinforcement-learning-based objectives could improve robustness further.

Finally, our models are trained primarily on English-language benchmarks and evaluated on question answering and fact checking workloads. The extent to which conditional reasoning distillation generalizes to multilingual settings, semi-structured enterprise data, or broader retrieval tasks beyond QA and Fact Checking remains open for future investigation.

11 Ethics Statement

This work focuses on improving table reranking models for question answering and information retrieval research. The datasets used in this study are publicly available academic benchmarks, including Natural Questions Tables, HybridQA, SQA, TabFact, and TAT-QA. We do not collect or release any personally identifiable information, and our experiments are conducted solely for research purposes.

Our approach relies on synthetic reasoning traces generated by large language models. While these traces can improve reranking performance, they may also inherit biases, factual inconsistencies, or reasoning artifacts present in the teacher models. Such biases could affect downstream retrieval behavior and generalization across domains. We therefore encourage careful evaluation of reasoning-based retrieval systems before deployment in high-stakes or real-world applications. Our work explores conditional reasoning distillation techniques that improve efficiency while reducing inference overhead.

Additionally, we used AI-assisted writing tools to improve the clarity and readability of the manuscript by obtaining feedback and suggestions during the writing process.

References

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Eisenschlos. 2021. Open domain question answering over tables via dense retrieval. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 512–519, Online. Association for Computational Linguistics.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. arXiv preprint arXiv:2412.16720.

Nour Jedidi, Yung-Sung Chuang, James Glass, and Jimmy Lin. 2025. Don't" overthink" passage reranking: Is reasoning truly necessary? arXiv preprint arXiv:2505.16886.

Carlos Lassance, Hervé Déjean, Thibault Formal, and Stéphane Clinchant. 2024. Splade-v3: New baselines for splade.

Wenhan Liu, Xinyu Ma, Weiwei Sun, Yutao Zhu, Yuchen Li, Dawei Yin, and Zhicheng Dou. 2025. Reasonrank: Empowering passage ranking with strong reasoning ability. arXiv preprint arXiv:2508.07050.

Xuan Lu, Haohang Huang, Rui Meng, Yaohui Jin, Wenjun Zeng, and Xiaoyu Shen. 2025. Rethinking reasoning in document ranking: Why chain-of-thought falls short. arXiv preprint arXiv:2510.08985.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 708–718, Online. Association for Computational Linguistics.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023a. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. arXiv preprint arXiv:2309.15088.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023b. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! arXiv preprint arXiv:2312.02724.

Revanth Gangi Reddy, JaeHyeok Doo, Yifei Xu, Md Arafat Sultan, Deevya Swain, Avirup Sil, and Heng Ji. 2024. First: Faster improved listwise reranking with single token decoding. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 8642–8652.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen-tau Yih, Pang Wei Koh, et al. 2025. Reasonir: Training retrievers for reasoning tasks. arXiv preprint arXiv:2504.20595.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: masked and permuted pre-training for language understanding. In Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. In Proceedings of the 2023 conference on empirical methods in natural language processing, pages 14918–14937.

Qwen Team. 2025. Qwen3 technical report.

Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. 2025. Rank1: Test-time compute for reranking in information retrieval. arXiv preprint arXiv:2502.18418.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. Rankt5: Fine-tuning t5 for text ranking with ranking losses. In Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval, pages 2308–2313.

Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2025. Rank-r1: Enhancing reasoning in llm-based document rerankers via reinforcement learning. arXiv preprint arXiv:2503.06034.

Jiaru Zou, Dongqi Fu, Sirui Chen, Xinrui He, Zihao Li, Yada Zhu, Jiawei Han, and Jingrui He. 2025. Gtr: graph-table-rag for cross-table question answering. arXiv e-prints, pages arXiv–2504.

A Evaluation Metrics

We report three standard information retrieval metrics, each computed at cutoff K.

Recall@K. Recall@K measures the fraction of relevant items recovered within the top-K retrieved results. Given a query with ground-truth relevant set $\mathcal{R}$:

$$\text{Recall@}K = \frac{|\mathcal{R} \cap \mathcal{S}_K|}{|\mathcal{R}|} \tag{5}$$

where $\mathcal{S}_K$ denotes the set of top-K retrieved candidates. A score of 1 indicates complete coverage of all relevant items within the top K results.

nDCG@K. Normalized Discounted Cumulative Gain evaluates ranking quality by assigning higher credit to relevant items appearing at earlier positions via a logarithmic position discount:

$$\text{DCG@}K = \sum_{i=1}^{K} \frac{\text{rel}_i}{\log_2(i+1)} \tag{6}$$

where $\text{rel}_i \in \{0, 1\}$ is the binary relevance label of the item at rank i. DCG@K is then normalized by the Ideal DCG (IDCG@K)—the score achieved by a perfect oracle ranking:

$$\text{nDCG@}K = \frac{\text{DCG@}K}{\text{IDCG@}K} \in [0, 1] \tag{7}$$

A score of 1 indicates that all relevant items are ranked optimally.

Accuracy@K. Accuracy@K measures the proportion of queries for which *all* relevant items are retrieved within the top-K results:

$$\text{Acc@}K = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbf{1}\left[\mathcal{R}_q \subseteq \mathcal{S}_K^{(q)}\right] \tag{8}$$

where $\mathcal{Q}$ is the set of evaluation queries, $\mathcal{R}_q$ denotes the set of relevant items for query q, $\mathcal{S}_K^{(q)}$ represents the top-K retrieved results, and $\mathbf{1}[\cdot]$ is the indicator function. The metric assigns a value of 1 only when every relevant item for a query appears within the top-K retrieved results, and 0 otherwise. This provides a strict evaluation of retrieval completeness at rank K.

B Prompt Template

```
You are a table relevance expert. Given a
question and a set of candidate tables, rank
them from most to least useful for answering
the question. Reason carefully about each
table's content, schema, and how directly it
addresses the question.
Output your final ranking as JSON:
{"ranked_tables": [1, 2, 3, ...]} where the
numbers are the Table IDs shown in the prompt.

Question: {question}

Candidate tables to rank:

### Table 1
| ... | ... | ... |


### Table N
| ... | ... | ... |
```

Figure 4: Prompt used for data generation.

C Dataset Statistics

Dataset	# Queries	# Tables
TabFact	15,106	34,351
HybridQA	6,106	17,229
SQA	148	320
TaTQA	362	4,754

Table 2: Statistics of the datasets used in our experiments.