

CRAFT: Training-Free Cascaded Retrieval for Tabular QA

*Adarsh Singh¹ *Kushal Raj Bhandari² Jianxi Gao²
†Soham Dan³ †Vivek Gupta¹

¹Arizona State University ²Rensselaer Polytechnic Institute ³Microsoft
asing725@asu.edu bhandk@rpi.edu

Abstract

Open-Domain Table Question Answering (TQA) involves retrieving relevant tables from a large corpus to answer natural language queries. Traditional dense retrieval models, such as DTR and DPR, not only incur high computational costs for large-scale retrieval tasks but also require retraining or fine-tuning on new datasets, limiting their adaptability to evolving domains and knowledge. In this work, we propose **CRAFT**, a zero-shot, cascaded retrieval approach that first uses a sparse retrieval model to filter a subset of candidate tables before applying more computationally expensive dense models as re-rankers. To improve retrieval quality, we enrich table representations with descriptive titles and summaries generated by *Gemini Flash 1.5*, enabling richer semantic matching between queries and tabular structures. Our method outperforms state-of-the-art (SOTA) sparse, dense, and hybrid retrievers on the NQ-Tables dataset. It also demonstrates strong zero-shot performance on the more challenging OTT-QA benchmark, achieving competitive results at higher recall thresholds, where the task requires multi-hop reasoning across both textual passages and relational tables. This work establishes a scalable and adaptable paradigm for table retrieval, bridging the gap between fine-tuned architectures and lightweight, plug-and-play retrieval systems.

1 Introduction

Tables serve as a vital format for representing structured knowledge, yet enabling open-domain question answering (QA) systems to retrieve and interpret tabular information effectively remains a non-trivial challenge. Unlike traditional Tabular QA (TQA) settings that assume access to gold tables, open-domain QA must first identify relevant

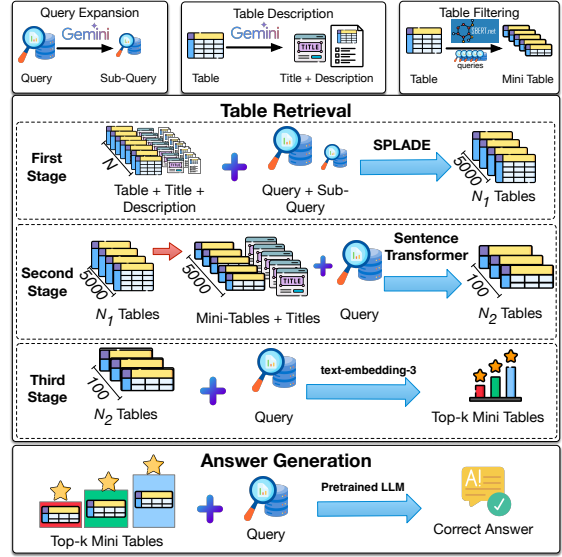


Figure 1: Overview of the CRAFT Framework.

tables from a large corpus before reasoning over their structure to derive accurate answers (Liang et al., 2024). Simple approaches that linearize tables into text and apply standard text-retrieval methods can lose the syntactic and structural information encoded in rows and columns (Badaro et al., 2023; Min et al., 2024). As a result, effective table retrieval has been identified as a key bottleneck for scalable QA over tabular data (Li et al., 2025).

Recent work has started to tackle open-domain TQA by developing specialized retrieval models. Early approaches adopted sparse or dense text retrievers on flattened tables (e.g., BM25 or bi-encoder models) (Fan et al., 2024; Robertson and Zaragoza, 2009). The Dense Table Retriever (DTR) introduced by Herzig et al. (2021) augments a bi-encoder with row and column features to better capture table structure. Several methods employ multi-stage cascades to prune candidates before applying costly scoring, with each stage acting as a separate ranking component. For example, (Wang et al., 2011) introduced a two-stage document ranking cascade composed of a candi-

*These authors contributed equally to this work.

†These authors jointly supervised this work.

date generation step and a learned re-ranker, which are trained jointly as a single model. In a similar vein, the "Retrieve, Read, Rerank" framework from (Hu et al., 2019) embeds retrieval, answer extraction, and re-ranking within a single shared Transformer network that is fine-tuned end-to-end. The Re²G system (Glass et al., 2022) implements a retrieve-rerank-generate pipeline where the re-ranking and generation stages are merged into one BART-based sequence-to-sequence model fine-tuned with knowledge distillation.

More recent methods explicitly incorporate fine-grained structure. Jin et al. (2023) proposed a syntax and structure-aware retriever that represents each question phrase and each table header and values separately, achieving state-of-the-art results on the NQ-Tables benchmark. Likewise, Li et al. (2025) introduces THYME, a field-aware hybrid retriever that tailors matching strategies to different table fields (e.g., cell values vs. captions). It significantly outperforms prior baselines on open-domain table retrieval, highlighting the gains from table-specific modeling. Similarly, Pan et al. (2022) introduced T-RAG, an end-to-end table QA framework that combines table retrievers with sequence-to-sequence generators for retrieval-augmented generation.

These advances, however, often rely on complex architectures or extensive fine-tuning on the TQA dataset. This raises an intriguing question: *Can off-the-shelf pretrained models, when integrated into a carefully designed retrieval pipeline, achieve competitive performance to expensive domain-specialized, fine-tuned methods?* To this end, we propose **CRAFT**, a modular and scalable multi-stage retrieval framework for TQA that harnesses pretrained models without relying on dataset-specific fine-tuning.

Our contributions are as follows :

- We design a modular, cascaded architecture that strategically combines complementary retrieval signals (lexical and semantic) to maximize recall while maintaining efficiency.
- Through comprehensive ablation studies, we demonstrate the robustness of CRAFT across varying retrieval thresholds N , highlighting its adaptability and stability.
- CRAFT achieves state-of-the-art retrieval recall and end-to-end performance on NQ-Tables, and strong zero-shot performance on

OTT-QA, with particularly competitive results at higher recall thresholds, demonstrating the effectiveness of training-free retrieval in real-world, open-domain settings.

2 Retrieval Benchmarks

We used two datasets (a.) NQ-Tables (Herzig et al., 2021), and (b.) OTT-QA (Chen et al., 2020) to evaluate our **CRAFT** approach. Both datasets are paired with extensive table corpora. NQ-Tables focuses on single-hop factual queries, whereas OTT-QA requires more intricate multi-hop reasoning across texts and tables. We use the subset relevant to the table for retrieval evaluation. Each query in both datasets is associated with approximately one gold table, ensuring consistency in supervision and facilitating direct comparison of retrieval performance. Table 1 provides an overview of the NQ-Tables and OTT-QA benchmarks.

	NQ-Tables		OTT-QA	
	Train	Test	Train	Test
Total Query Count	9,594	966	41,469	2,214
Unique Query Count	9,594	919	41,469	2,214
Avg. # Words	8.94	8.90	21.79	22.82
Table Count	169,898		419,183	
Avg. # Rows	10.70		12.90	
Avg. # Columns	6.10		4.80	
# Golden Tables / Query	1.00	1.05	1.00	1.00

Table 1: Statistics of the NQ-TABLES and OTT-QA benchmarks.

Together, these statistics underline the scalability and diversity of the evaluation setting, where NQ-Tables offer a more controlled environment and OTT-QA presents a challenging, open-domain scenario that better tests model generalization.

3 Our CRAFT Approach

To create a retrieval setup without dataset-specific fine-tuning, we adopt a modular framework that leverages off-the-shelf models with multi-stage retrieval, with more complex models for each successive stage. As shown in Figure 1, CRAFT involves pre-processing followed by stage-wise table retrieval using increasingly expressive models, culminating in answer generation via a pre-trained LLM. Our approach demonstrates that pre-trained models can effectively retrieve relevant tabular data and generate accurate answers. The pipeline consists of leveraging pre-processing followed by cascaded table retrieval and end-to-end

answer generation to maximize semantic alignment between input queries and tabular content using off-the-shelf embedding models.

3.1 Preprocessing

The preprocessing pipeline comprised three key operations designed to enhance both query and table representations for effective retrieval. **Firstly**, Gemini-1.5-Flash (Team et al., 2025) was employed to generate sub-questions from each original query, decomposing complex queries into more granular and interpretable components. **Second**, Gemini-1.5-Flash was also used to generate informative table titles and detailed descriptions, enriching the contextual understanding and enabling more accurate downstream retrieval. **Finally**, all table rows were ranked according to their semantic relevance to the queries using model (*all-mpnet-base-v2*) from Sentence Transformer (Reimers and Gurevych, 2019) for the NQ-Tables dataset, and Jina Embeddings v3 (Sturua et al., 2024) for the OTT-QA dataset.

3.2 Table Retrieval

The table retrieval step follows a unified three-stage pipeline applied to NQ-Tables and OTT-QA, with dataset-specific inputs and model choices.

In the **first stage**, SPLADE (Formal et al., 2021), a pre-trained sparse lexical-expansion retriever, was applied to a comprehensive dataset consisting of 169K and 419K Tables in NQ-Tables and OTT-QA dataset, respectively, leveraging generated table titles, column headers, individual cell values, and table descriptions to filter and narrow down the candidate tables to the top 5,000 entries. SPLADE leverages sparse lexical expansion to efficiently score candidate passages using minimal memory, making it well-suited for handling large-scale retrieval tasks.

In the **Second Stage**, **mini-tables** are constructed by appending column headers to each of the top 5 rows and then scoring these mini-tables to produce a compact shortlist. For NQ-Table, the Sentence Transformer *all-mpnet-base-v2*, a 768 dimensional pre-trained bi-encoder model optimized for semantic similarity with solid zero/few-shot retrieval performance across heterogeneous text, is used for fine-grained semantic matching due to strong sentence-level retrieval quality and robust generalization. For OTT-QA, Jina Embeddings v3, a multilingual long-context embedding model with task adapters that improve robustness

on mixed schema/numeric text, is used to better handle longer row-level contexts and numeric content typical of diverse table formats in OTT-QA. This stage selects the top K mini-tables for final reranking.

Lastly, **Stage three**, the top- K mini-tables are **re-ranked** using state-of-the-art embedding models optimized for semantic precision and computational efficiency. For the NQ-Table dataset, we employ *text-embedding-3-small* (OpenAI, 2024) and *text-embedding-3-large*, which provide compact, cost-effective 1536- and 3072-dimensional representations, respectively, achieving strong semantic accuracy while maintaining high throughput—making them well-suited for large-scale reranking. For the OTT-QA dataset, re-ranking is performed using *gemini-embedding-001* (Lee et al., 2025), designed for long-context understanding and production-grade scalability. The Gemini embedding space enables robust semantic scoring without additional query or document augmentation, preserving the integrity of dense representations and preventing noise amplification.

Overall, this cascaded retrieval framework progressively transitions from high-recall sparse retrieval to high-precision semantic reranking, achieving efficient large-scale performance while minimizing information loss between stages.

3.3 End-to-End Answer Generation

For the end-to-end table question-answering, we use the final retrieval result from CRAFT Stage 3, using three different pre-trained models: *Llama3-8B* (Grattafiori et al., 2024), *Qwen2.5-7B* (Qwen et al., 2025), and *Mistral-7B* (Jiang et al., 2023). Each LLM generates answers from prompts that include a varying number of tables ($n = \{1, 3, 5, 8, 10\}$). Each prompt contains the table title, column headers, and mini-tables (top 5 rows) for n tables, offering a comprehensive yet concise context to support accurate answer generation. Additionally, we experimented with the inclusion of a few-shot prompt to guide the text generation process. We evaluate model performance and accuracy by computing the F1 score, which quantifies the overlap between the generated answers and the reference answers, enabling a comparative analysis across different models and table settings.

4 Results and Analysis

4.1 Retrieval Performance.

Table 2 and Table 3 present the retrieval performance of three categories of methods Sparse, Dense, and Hybrid retrievers on the NQ-Tables and OTT-QA datasets, respectively. We conduct a comprehensive comparison across traditional sparse models such as BM25 and SPLADE, dense retrievers including TaPas (Herzig et al., 2020), DPR (Karpukhin et al., 2020), BIBERT (Yates et al., 2021), DTR, T-RAG, and SSDR, as well as hybrid approaches like Re2G, THYME, DHR (Lin et al., 2023a), and BIBERT+SPLADE.

Model	R@1	R@10	R@50
Sparse Retrieval Models			
BM25	18.49	36.94	52.61
SPLADE	39.84	83.33	94.65
Dense Retrieval Models			
BIBERT	43.78	82.25	93.71
DPR	45.32	85.84	95.44
TAPAS	43.79	83.49	95.10
DTR [§]	32.62	75.86	89.77
DTR (M) + HN [†]	47.33	80.96	91.51
DTR (S) + HN [†]	43.53	77.80	90.21
SSDR(im)	45.47	84.00	95.05
T-RAG*	46.07	85.40	95.03
Hybrid Retrieval Models			
Re2G [†]	30.47	66.49	83.35
BIBERT+BM25	35.87	79.63	94.56
DHR	43.67	84.65	95.62
BIBERT+SPLADE	45.62	86.72	95.62
THYME	<u>48.55</u>	86.38	96.08
CRAFT Stage-wise			
Stage 1	34.38	72.90	91.62
+ Stage 2	36.65	82.91	96.08
+ Stage 3 ^{*s}	41.13	87.16	<u>96.84</u>
+ Stage 3 ^{*l}	49.84	<u>86.83</u>	97.17
CRAFT	49.84	<u>86.83</u>	97.17

Table 2: Table retrieval performance on NQ-Tables. All models except *CRAFT* are trained on NQ-Tables. *: results on the full test set (966 queries). ^{*l} and ^{*s}: use *text-embedding-3-large* and *text-embedding-3-small* in Stage 3, respectively. [§] denotes models trained with the THYME setup. [†] denotes the use of the original released model checkpoints. **R** = Recall. Best results in **bold**; second-best are underlined.

CRAFT demonstrates strong retrieval performance across both datasets, achieving state-of-the-art results on NQ-Tables and competitive zero-shot generalization on OTT-QA.

Retrieval on NQ Tables. On NQ-Tables, CRAFT progressively improves through its three-stage cascade: In Stage 1, R@10 is 72.90, and in Stage 2, it is boosted to 82.91, and Stage 3’s neural reranker further refines top-ranked results,

raising R@1 to 41.13 and R@10 to 87.16. Using the higher-dimensional *text-embedding-3-large* model increases R@1 to 49.84.

Retrieval on OTT-QA. CRAFT maintains competitive recall at higher retrieval depths despite operating entirely without task-specific training. While fine-tuned hybrid retrievers such as THYME and BIBERT+SPLADE achieve superior Recall@1 through dataset-specific optimization, CRAFT approaches their performance at R@10 (89.88 vs. 91.10) and R@50 (96.07 vs. 96.34), with gaps of only 1.22 and 0.27 points respectively. The stage-wise improvements mirror those observed on NQ-Tables: Stage 2 delivers a substantial gain of +21.0 points at R@10 over Stage 1, while Stage 3 contributes an additional +8.0 points. This consistent progression across both datasets demonstrates that CRAFT’s modular, training-free cascade effectively amplifies recall and maintains strong cross-dataset generalization, even when scaling from 169K tables (NQ-Tables) to 419K tables (OTT-QA) with more complex multi-hop reasoning requirements.

Model	R@1	R@10	R@50
Sparse Retrieval Models			
BM25	23.98	51.94	69.11
SPLADE	62.74	89.52	95.21
Dense Retrieval Models			
BIBERT	56.82	86.50	94.26
DPR	53.43	85.95	93.22
TAPAS	57.86	86.77	94.04
DTR	42.10	75.75	88.80
SSDR(im)	56.96	86.22	93.55
Hybrid Retrieval Models			
BIBERT+BM25	59.49	86.81	94.67
DHR	63.64	88.48	95.30
BIBERT+SPLADE	<u>64.72</u>	<u>91.01</u>	96.34
THYME	66.67	91.10	<u>96.16</u>
CRAFT Stage-wise			
Stage 1	40.74	60.89	88.17
+ Stage 2	47.33	81.88	91.82
+ Stage 3	55.56	89.88	96.07
CRAFT	55.56	89.88	96.07

Table 3: Table Retrieval Performance on OTT-QA. **R** denotes recall score. Best performances are shown in **bold**, and second-best are underlined. All models, except *CRAFT*, are trained on OTT-QA. SPLADE in the table is the supervised baseline, whereas CRAFT Stage 1 uses the base SPLADE model.

4.2 Robustness of CRAFT Retrieval.

To assess robustness under query paraphrasing, we generated perturbed versions of each question using *Gemini 2.5 Flash*, while keeping the ground-truth answers unchanged. For example:

Original : Philadelphia is known as the city of what?
Perturbed : What descriptor follows “the city of”
when referencing Philadelphia?
Answer : Brotherly Love

The prompt used for generating the perturbed queries is detailed in Figure 5.

Analysis. Table 4 reports recall scores on perturbed queries for DTR, trained on NQ-Tables with and without hard negatives, and our three-stage **CRAFT** pipeline. We chose it as a baseline because it is a traditional, well-known, and widely adopted model. The checkpoint used in our experiments is publicly available and has been fine-tuned on the NQ-Tables dataset, making it an ideal baseline for comparing our Cascade Approach against a standard dense embedding method. Additionally, it serves as a common baseline used in many papers that report scores on the NQ-Tables Dataset, like LI-RAGE, DTR, THYME, CLTR, SSDR, and T-RAG, facilitating easier comparison.

Model	R@1	R@10	R@50	Δ (Avg.)
Original Query				
DTR (M)	38.74	75.73	88.36	-
DTR (M) + HN	47.33	80.96	91.51	-
DTR (S)	39.28	73.88	87.60	-
DTR (S) + HN	43.53	77.80	90.21	-
CRAFT	41.13	87.16	96.84	-
Perturbed Query				
DTR (M)	27.09	66.27	84.33	- 8.38
DTR (M) + HN	37.87	76.06	88.47	- 5.80
DTR (S)	25.25	61.81	78.24	- 11.82
DTR (S) + HN	35.04	70.73	85.85	- 6.64
CRAFT Stage-wise				
Stage 1	33.40	68.22	87.05	- 3.41
+ Stage 2	34.05	73.18	88.92	- 6.5
+ Stage 3 **	41.02	86.83	96.08	- 0.04

Table 4: Performance of DTR vs CRAFT with perturbed queries on NQ-Tables. **M** and **S** denote Medium and Small model sizes respectively. Δ is the average change in Recall relative to the original query scores, positive values indicate improvement, and negative values indicate degradation. **HN** represents a model trained with Hard Negatives. ** indicate the use of *text-embedding-3-small* model in Stage 3.

With the original queries, **CRAFT** already outperforms the DTR baselines. When using the perturbed queries, both DTR Medium and Small suffer substantial average drops of -8.38 and -11.82 points, respectively. The DTR version trained with hard negatives reduces but does not eliminate this degradation (avg = -5.80 for DTR (M) and -6.64 for DTR (S)). By contrast, **CRAFT**’s full three-stage pipeline nearly fully recovers its original performance with a slight decrease in paraphrasing.

Stage-wise Analysis. Analysis over stages further shows incremental gains: Stage 1 drops performance under perturbation by 3.4 %, Stage 2 partly recovers it, and Stage 3 restores the original recall, as shown in Table 4. These results highlight **CRAFT**’s superior robustness to query rephrasing compared to DTR.

4.3 End-to-End Results.

We evaluate the end-to-end performance of **CRAFT** across multiple pretrained language models and retrieval configurations. For end-to-end QA, our approach aligns with recent work, such as THYME and EE-BIBERT, which report results on larger open-source models like Llama3-8B, Qwen2.5-7B, and Mistral-7B and closed-source model GPT-4o.

Results on NQ-Tables. As shown in Table 5, **CRAFT** consistently outperforms several state-of-the-art baselines across different models, for varying numbers of retrieved tables ($n = 1, 3, 5$). These results highlight the robustness of our modular, training-free approach. Beyond higher recall, mini-table filtering and compact context windows reduce irrelevant content and allow more evidence (more retrieved tables) per query, so even semantically correct non-gold tables can still support higher end-to-end F1.

Effect of Larger LLMs. Table 6 further emphasizes the effectiveness of **CRAFT**, particularly when employing larger, instruction-tuned LLMs. Compared to prior dataset-specific baselines, **CRAFT** consistently delivers superior F1 scores. We find that using just a single retrieved table ($n = 1$), *GPT 4o* achieves an F1 score of 56.76, already surpassing the best-performing baseline, LI-RAGE (54.17), even though it benefits from deeper retrieval ($n = 5$). With increased retrieval depth ($n = 5$), *GPT-4o* further improves to 67.72, highlighting significant gains from additional context. Similarly, *Mistral-Small-Instruct* and *Llama-3.1-70B-Instruct* achieve notable scores of 57.14 and 56.94 at $n = 5$, clearly outperforming LI-RAGE (current SOTA).

Effect of Few-Shots. To evaluate the scalability and robustness of **CRAFT**, under zero-shot and few-shot prompting as the number of retrieved tables increases ($n = 1, 3, 5, 8, 10$). As shown in Table 14 (Appendix Section C), all models exhibit a clear upward trend in F1 score with deeper retrieval, demonstrating that structured multi-stage

Retriever	n=1			n=3			n=5		
	Mistral	Llama3	Qwen	Mistral	Llama3	Qwen2.5	Mistral	Llama3	Qwen2.5
BIBERT	32.93	32.66	34.80	34.6	33.16	37.92	35.30	34.28	37.01
SPLADE	29.61	32.07	31.90	35.42	37.17	35.79	34.95	33.88	37.35
SSDR_in	32.76	34.25	33.42	36.95	38.14	37.14	36.60	37.89	33.92
BIBERT-SPLADE	32.67	32.66	33.24	33.59	33.66	36.76	35.71	33.92	37.02
THYME	35.48	36.14	37.28	37.59	39.16	40.28	37.20	39.29	41.20
CRAFT	39.13	38.31	39.73	45.28	40.76	43.52	44.53	40.55	46.49

Table 5: Performance comparison (F1) across different retrievers, LLMs, and values of n with NQ-Tables. The evaluated models include *Mistral-7B-Instruct*, *Llama3-8B-Instruct*, and *Qwen2.5-7B-Instruct*, corresponding to **Mistral**, **Llama3**, and **Qwen**, respectively. **Bolded** values indicate top results.

Model	n=1	n=5
RAG (Lewis et al., 2020)	-	39.67
DPR-RAGE (Lin et al., 2023b)	-	49.68
LI-RAGE (Lin et al., 2023b)	-	54.17
CRAFT		
- Qwen2.5-72B-Instruct	48.75	49.17
- Llama-3.3-70B-Instruct	46.21	55.16
- Llama-3.1-70B-Instruct	49.75	56.94
- Mistral-Small-Instruct	<u>50.26</u>	<u>57.14</u>
- GPT-4o	56.76	67.72

Table 6: F1 scores of **CRAFT** on NQ-Tables with larger open source and closed source LLMs compared to dataset-specific baselines for $n = 1$ and $n = 5$ retrieved tables. Bold indicates the best performance.

retrieval alone substantially narrows the gap with dataset-specific fine-tuning. However, incorporating a small number of in-context examples does not consistently yield significant improvements across models. Moreover, performance gains taper beyond $n = 5$, suggesting an optimal retrieval depth that balances information richness with the risk of noise accumulation. This effect becomes more pronounced as both n and the number of few-shot examples increase, leading to greater performance variability.

Result with OTT-QA. We further provide end-to-end performance on the OTT-QA dataset under an oracle setting, where retrieved tables are paired with the gold passage. Under this oracle setup, where retrieved tables are paired with their corresponding gold passages to isolate the effect of retrieval quality, our end-to-end OTT-QA evaluation shows that Stage 3 (MT) consistently achieves the best balance between recall and precision. As this setup is not directly comparable to prior baselines, we present these results in the Appendix Section C for completeness.

These results underscore that **CRAFT**’s structured, training-free retrieval methodology not only outperforms established baselines but also scales effectively with increased retrieval depth and few-shot prompting. This highlights the strength of **CRAFT**’s modular design in enabling adaptable, efficient, and high-performing Table-QA without

requiring any dataset-specific fine-tuning.

4.4 Computational Efficiency

CRAFT is a training-free retrieval pipeline that optimizes per-query computation cost without requiring dataset-specific fine-tuning. Unlike end-to-end trained baselines that shift most computation to offline training and then rely on a single, fully embedded corpus at inference time, **CRAFT**’s cascaded architecture strategically concentrates expensive computation only where it provides the greatest benefit. This modular design offers two key advantages: (1) it avoids costly re-training when adapting to new domains, and (2) it enables seamless integration of emerging embedding models by simply swapping components without architectural changes.

Retrieval Efficiency. **CRAFT** shifts the bulk of computation to an offline pre-computation phase, drastically reducing online embedding calls. We build a SPLADE inverted index over all tables using sparse, lightweight lexical representations and pre-encode each table row with the Stage 2 dense ranker model exactly once. At inference, **CRAFT** executes only three query embeddings plus 5,100 mini-table embeddings per query:

- Stage 1 (Sparse Retrieval): One query encoding + inverted-index lookup over 169,898 (NQ-Tables) / 419,183 (OTT-QA) tables.
- Stage 2 (Dense Pruning): One *all-mpnet-base-v2* query embedding + sublinear-time ANN search over all precomputed row embeddings to select the top 5,000 mini-tables, avoiding the $\mathcal{O}(N)$ cost of brute-force search.
- Stage 3 (Re-ranking): One *text-embedding-3-small* query embedding + embeddings for the top 100 mini-tables.

By contrast, a fully dense baseline requires 169,898 table embedding calls. Stage 1 pruning

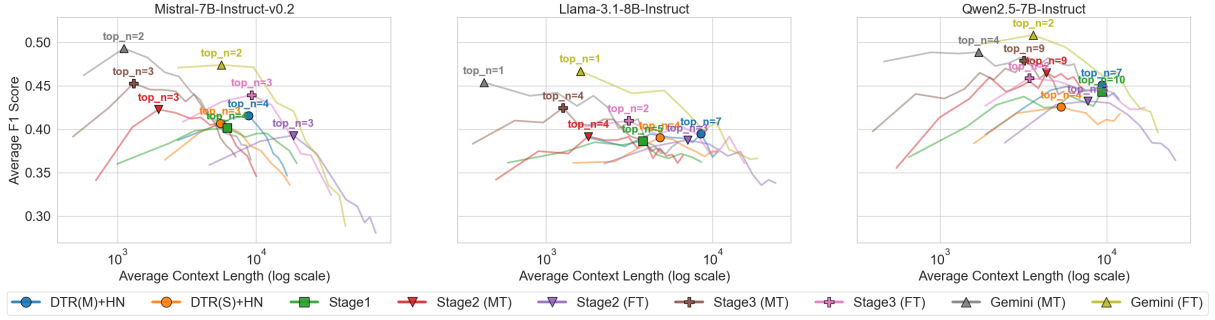


Figure 2: End-to-end performance comparison across context lengths with the NQ dataset. The figure illustrates the relationship between average context length and F1 score across three models, evaluated over ten different numbers of retrieved tables (n) on the NQ dataset. FT denotes the use of the Full Table for end-to-end evaluation, while MT represents the Mini-Table variant.

reduces the candidate set by over 97% while retaining 99.56% recall Table 8 cutting embedding calls by over $33\times$ without sacrificing accuracy.

Computational Trade-offs and Adaptability.

Stage 2 employs a lightweight dense encoder to produce embeddings for the generated mini-tables on-the-fly during inference. While this increases per-query latency compared to retrieving from a pre-embedded corpus, it substantially reduces irrelevant content, thereby saving valuable context tokens for downstream processing. This design manages the trade-off between per-query cost and end-to-end performance by filtering noise early in the pipeline, which significantly benefits Stage 3 retrieval and LLM-based QA, both highly sensitive to token budgets. CRAFT’s modular architecture is particularly well-suited to evolving embedding ecosystems. Recent closed-source systems have introduced increasingly large and capable embedding models, such as OpenAI’s *text-embedding-3* series and Gemini Embedding. CRAFT’s Stage 3 reranking component can be replaced with any such model (API-based or open-source) without incurring substantial additional cost, because the principal bottlenecks in retrieval pipelines are the token budget and the number of per-query embedding calls, not the specific embedding architecture. By reducing both of these quantities, CRAFT remains adaptable to future embedding models while maintaining efficient and robust retrieval.

End-to-End Efficiency. While cascading retrieval improves performance with reduced offline computational overhead, computational efficiency also remains critical for the end-to-end QA task.

Effect of Mini-Tables. Table 7 compares the average token counts for each model when using

Model	n	Sub-Table	Full Table
Llama-3.1-8B-Instruct	1	362.70	1782.80
	3	958.79	4841.75
	5	1561.34	8231.99
	8	2544.36	12831.89
	10	3213.31	15798.45
Mistral-7B-Instruct-v0.2	1	476.58	3314.03
	3	1305.60	10183.81
	5	2158.50	18252.36
	8	3566.04	29321.20
	10	4537.62	36191.87
Qwen2.5-7B-Instruct	1	393.68	1951.08
	3	1060.45	5319.05
	5	1731.25	9009.43
	8	2824.83	14046.00
	10	3568.37	17292.88

Table 7: Comparison of average token counts between sub-table and full-table contexts across models and top- k values.

only the top- k sub-table context versus concatenating the entire table. Across all three models and $k \in 1, 3, 5$, sub-table inputs (CRAFT mini-tables) reduce token length by over 70%. For example, Llama3 drops from 1782.80 tokens at $k = 1$ to 362.70; Mistral from 3,314.03 to 476.58; and Qwen2.5 from 1,951.08 to 393.68. This substantial reduction directly lowers inference costs and improves runtime efficiency. By selectively presenting only the most relevant table rows fragments, **CRAFT** retains essential semantic content while avoiding the diminishing returns and latency overhead associated with full-table encoding.

Effect of Token Reduction. Figure 2 compares the end-to-end performance of three instruction-tuned models across varying context lengths on the NQ dataset. The curves reveal a clear trade-off between context breadth and noise accumulation: smaller retrieved sets (low n) often underperform because they may exclude the gold table, while larger retrieved sets (high n) increase

the risk of including irrelevant content that dilutes attention and lowers F1. This pattern is particularly evident for Qwen2.5-7B-Instruct, where performance peaks at top_n = 9 using the Mini-Table (MT) configuration, showing that our multi-stage retrieval strategy effectively identifies an optimal context length that balances relevance and compactness. Stage 3 (MT) consistently achieves superior results at shorter context lengths, highlighting the benefit of reducing redundant information while retaining sufficient evidence for reasoning.

5 Ablation Study

As Table 8 shows, our three-stage cascade smoothly shifts from broad coverage to pinpoint accuracy. In Stage 1, the SPLADE filter finds nearly all relevant tables at deeper ranks (91.62 % at R@50) but only 34.38 % at the very top. However, SPLADE achieves a R@5000 of 99.59%, indicating performance saturation and leaving minimal scope for further enhancement. Stage 2’s Sentence-Transformer reranker then delivers most of the mid-level gains, raising R@10 by over 10 points and boosting R@50 above 96%, while still keeping high deep recall. The top 5,000 tables returned by Stage 1 are passed to Stage 2, which performs dense semantic reranking over these candidates. In Stage 2, we achieve the highest recall of 97.82%, which is a significant improvement over the Stage 1 retrieval. These 100 tables are carried forward to Stage 3, our final retrieval step. Finally, Stage 3’s neural reranker sharpens the top results, lifting R@1 by 4.5 points to 41.13 % and increasing R@10 to 87.16 %. Each stage builds upon the last, improving accuracy step by step without requiring any additional training.

Stage 1: Sparse Retrieval with Context Enrichment We test the effect of table descriptions and query decomposition using BM25 and SPLADE. The results for Stage 1 are shown in Table 9 and Table 10. Both retrievers exhibit improvement when employing Question Decomposition. Specifically, SPLADE achieves gains of 2.18% and 4.03% at R@10 and R@50, respectively. With the table description, SPLADE achieved a consistent increase in recall. Notably, SPLADE achieves an R@5000 of 99.59%, indicating performance saturation and leaving minimal scope for further enhancement. In contrast, BM25 demonstrates substantial improvement with the inclusion of table descriptions, clearly benefiting from context-

enriched lexical matching.

Table	R@1	R@10	R@50	R@1k	R@5k
BM25					
w/o Desc	17.95	34.38	47.44	70.73	82.48
+ Desc	22.74	45.16	62.57	86.61	93.14
SPLADE					
w/o Desc	31.06	69.05	86.85	98.45	99.59
+ Desc	32.64	70.62	87.59	98.91	99.67

Table 9: **Stage 1:** Effect of Table Descriptions on Retrieval Performance over NQ-Tables. **R** denotes Recall Score. N=169,898 Tables in Corpus.

Question	R@1	R@10	R@50	R@1k	R@5k
BM25					
Original	22.74	45.16	62.57	86.61	93.14
+ SubQues	24.15	53.54	71.38	91.08	96.84
SPLADE					
Original	32.64	70.62	87.59	98.91	99.67
+ SubQues	34.38	72.90	91.62	98.91	99.56

Table 10: **Stage 1:** Effect of Query Decomposition-Retrieval Performance on NQ-Tables. **R** denotes Recall Score. N=169,898 Tables in Corpus.

Stage 2: Dense Semantic Reranking and Irrelevant Context Reduction We apply the Stage 2 dense model to rerank the top 5000 tables. Since dense embedding models have limited context length, we rank the rows of each table and retain the 5 most relevant rows per table. Table 11 shows recall gains from filtering irrelevant rows, with R@10 rising from 79.21 to 82.91 and R@50 from 94.66 to 96.08. Dense scoring better captures the semantic relevance missed in Stage 1.

Stage 3: Neural Reranking Finally, we rerank the top 300 and 100 mini-tables. Table 12 shows the reranking results. Even without fine-tuning, neural reranking significantly sharpens final table rankings.

These findings confirm the effectiveness of the cascaded design, which begins with broad lexical and semantic coverage, progressively refines relevance through dense pruning, and culminates in lightweight neural reranking for precise top-rank. Collectively, they highlight how each stage contributes a distinct function within the pipeline, yielding an accurate and computationally efficient retrieval framework.

Error Analysis To further understand the performance gaps, we provide a qualitative error analysis in Appendix D. We find that most failure cases arise from near-duplicate tables with highly similar schemas or ambiguous entity mentions, which

Model	NQ-Tables						OTT-QA					
	R@1	R@10	R@50	R@100	R@500	R@5K	R@1	R@10	R@50	R@100	R@500	R@5K
Stage 1	34.38	72.90	91.62	94.89	98.91	99.56	40.74	60.89	88.17	91.64	96.93	99.28
Stage 2	36.65	82.91	96.08	97.82	98.91	-	47.33	87.35	91.82	94.21	97.38	-
Stage 3	41.13	87.16	96.84	-	-	-	55.56	89.88	96.07	-	-	-
CRAFT	41.13	87.16	96.84	-	-	-	55.56	89.88	96.07			

Table 8: Table Retrieval Performance on NQ-Tables and OTT-QA datasets with various stages of CRAFT. R denotes Recall Score. Best performances are in **Bold**. Boxed values indicate the number of tables passed to the next stage for re-ranking.

Table Text	R@1	R@10	R@50	R@100	R@1k
Original*	37.64	79.76	94.45	97.82	99.45
-Description	34.82	79.21	94.66	97.39	99.45
-Irrelevant Rows	36.65	82.91	96.08	97.82	98.91

Table 11: **Stage 2:** Effect of Irrelevant Context Reduction- Retrieval using NQ-Tables. Original* is the title, header, all cell values, and description. N2=5000 Tables in Corpus. Best performances are in **Bold**.

N3	R@1	R@10	R@50
300	39.03	86.02	96.89
100	41.13	87.16	96.84

Table 12: **Stage 3:** Effect of Neural Reranking: Retrieval Performance using *text-embedding-3-small* model on NQ-Tables. N3 is the number of tables in the corpus. Best performances are in **Bold**.

are difficult to disambiguate without task-specific supervision. Additionally, even when the gold table is retrieved, using full tables can introduce noise that hinders answer extraction, whereas our mini-table representation isolates relevant rows and enables the LLM to generate correct answers more reliably.

6 Conclusion

We present CRAFT, a modular, training-free cascaded retrieval framework for Table Question Answering that achieves competitive performance compared to state-of-the-art systems that require dataset-specific fine-tuning. CRAFT achieves new state-of-the-art retrieval Recall@1 on NQ-Tables, maintains high robustness against query paraphrasing, and delivers strong end-to-end QA performance when paired with modern LLMs. By decoupling retrieval from domain-specific fine-tuning, CRAFT offers a scalable and adaptable paradigm for dynamic tabular domains where rapid deployment and data scarcity are critical issues. Future work will extend CRAFT’s capabilities to enhance its reasoning through a fully integrated, multi-hop retrieval system to tackle complex benchmarks, such as OTT-QA, without relying on oracle passages. These directions will fur-

ther advance CRAFT toward a unified, efficient, and generalizable solution for open-domain table understanding.

7 Limitations

While this study demonstrates the effectiveness of a training-free, cascaded approach, its scope has certain boundaries. The evaluation focuses primarily on single-hop question answering over the NQ-Tables and OTT-QA datasets. The framework’s performance on more complex, multi-hop reasoning tasks requiring synthesizing information across multiple tables and text passages was not fully explored. A potential extension is incorporating a lightweight contrastive objective after Stage 3 to improve R@1 by better distinguishing structurally similar yet semantically distinct tables. This direction is left for future work within the training-free framework.

Additionally, the end-to-end question answering evaluation for OTT-QA was conducted in an oracle setting, relying on gold passages to isolate retrieval performance. This approach, while useful for analysis, does not represent a full open-retrieval pipeline. The current system is also tailored specifically for text-based table retrieval not multimodal tables that include non-textual data like images. Finally, although CRAFT’s offline pre-computation is efficient, the multi-stage nature of the pipeline results in increasing per-query processing time.

8 Acknowledgements

We thank the anonymous reviewers and the Area Chair for their constructive feedback. We are grateful to the members of the Complex Data Analysis and Reasoning Lab for their helpful discussions and contributions to the project’s ideation. We also thank Aman Singh and Prasham Titiya for their insightful input during the early stages of this work.

References

- Gilbert Badaro, Mohammed Saeed, and Paolo Papotti. 2023. [Transformers for Tabular Data Representation: A Survey of Models and Applications](#). *Transactions of the Association for Computational Linguistics*, 11:227–249.
- Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Wang, and William W Cohen. 2020. Open question answering over tables and text. *arXiv preprint arXiv:2010.10439*.
- Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. [A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501, Barcelona Spain. ACM.
- Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. [SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking](#). In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292, New York, NY, USA. Association for Computing Machinery.
- Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. [Re2G: Retrieve, rerank, generate](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- 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.
- Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training. *arXiv preprint arXiv:2004.02349*.
- Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. 2019. Retrieve, read, rerank: Towards end-to-end multi-document reading comprehension. *arXiv preprint arXiv:1906.04618*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *CoRR*, abs/2310.06825.
- Nengzheng Jin, Dongfang Li, Junying Chen, Joanna Siebert, and Qingcai Chen. 2023. [Enhancing Open-Domain Table Question Answering via Syntax- and Structure-aware Dense Retrieval](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Nusa Dua, Bali. Association for Computational Linguistics.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftexhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, et al. 2025. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20*, Red Hook, NY, USA. Curran Associates Inc.
- Da Li, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2025. [Tailoring Table Retrieval from a Field-aware Hybrid Matching Perspective](#). *Preprint*, arXiv:2503.02251.
- Xinyi Liang, Rui Hu, Yu Liu, and Konglin Zhu. 2024. [Open-Domain Question Answering over Tables with Large Language Models](#). In De-Shuang Huang, Yijie Pan, and Jiayang Guo, editors, *Advanced Intelligent Computing Technology and Applications*, volume 14873, pages 347–358. Springer Nature Singapore, Singapore.
- Sheng-Chieh Lin, Minghan Li, and Jimmy Lin. 2023a. [Aggretriever: A simple approach to aggregate textual representations for robust dense passage retrieval](#). *Transactions of the Association for Computational Linguistics*, 11:436–452.
- Weizhe Lin, Rexhina Blloshmi, Bill Byrne, Adria de Gispert, and Gonzalo Iglesias. 2023b. [LI-RAGE: Late Interaction Retrieval Augmented Generation](#)

with Explicit Signals for Open-Domain Table Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1557–1566, Toronto, Canada. Association for Computational Linguistics.

Dehai Min, Nan Hu, Rihui Jin, Nuo Lin, Jiaoyan Chen, Yongrui Chen, Yu Li, Guilin Qi, Yun Li, Nijun Li, and Qianren Wang. 2024. [Exploring the Impact of Table-to-Text Methods on Augmenting LLM-based Question Answering with Domain Hybrid Data](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 464–482, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI. 2024. [text-embedding-3-small](#). Model card, accessed 19 May 2025.

Feifei Pan, Mustafa Canim, Michael Glass, Alfio Gliozzo, and James Hendler. 2022. [End-to-End Table Question Answering via Retrieval-Augmented Generation](#). *Preprint*, arXiv:2203.16714.

Qwen, An Yang, Baosong Yang, Beichen Zhang, and Hui. 2025. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.

Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Stephen Robertson and Hugo Zaragoza. 2009. [The Probabilistic Relevance Framework: BM25 and Beyond](#). *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. [jina-embeddings-v3: Multilingual embeddings with task lora](#). *Preprint*, arXiv:2409.10173.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, and Katie Millican. 2025. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.

Lidan Wang, Jimmy Lin, and Donald Metzler. 2011. A cascade ranking model for efficient ranked retrieval. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 105–114.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. [Pretrained transformers for text ranking:](#)

[BERT and beyond](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 1–4, Online. Association for Computational Linguistics.

A Hyperparameter Configuration

Configuration Parameter	Value
Batch Size	144
Learning Rate	1×10^{-5}
Training Epochs	50
Optimizer	Adam
Initialization	BERT-base
THYME Dropout (p_{sem}, p_{lex})	0.15
THYME FLOPS Reg. (λ_q, λ_t)	1×10^{-4}

Table 13: Training and inference configuration used for all baselines.

Table 13 outlines the training and inference configuration adopted for baseline retrievers. These configurations were chosen to ensure stable convergence across both datasets while maintaining generalization in retrieval performance. The combination of moderate dropout, lightweight regularization, and a conservative learning rate enabled consistent optimization across heterogeneous data distributions, striking a balance between efficiency and representational expressiveness. For Re2G, we directly used the publicly released checkpoints fine-tuned on NQ Dataset.

B Prompt Used For Experiments

The prompt used for end-to-end question answering in Figure 3, the table description prompt in Figure 4 and the question decomposition prompt is shown in Figure 5.

```
You are given a set of tables along with their
titles. Based on the information provided in the
most relevant table(s) answer the following
question.
Do **not** include any explanations or extra
text.
Ensure the final answer format is **only**:
'FINAL ANSWER: Answer Value'
[FEWSHOT EXAMPLES]
Now, answer the following problem, given the
relevant tables:
[TABLES]
QUESTION: [QUESTION]
FINAL ANSWER:
```

Figure 3: Prompt used for End-to-End question answering.

You are an advanced AI specialized in analyzing and summarizing structured data within tables.
Your task is to generate a meaningful title and description for the provided table.
The title should be concise and clearly reflect the table's primary content, while the description should summarize the context and provide insights into the key data.
The description must be tailored for query ranking, meaning it should help identify the table's relevance to possible questions.

Guidelines:

1. Table Title:
 - a. Provide a concise, clear title that reflects the content and context of the table.
 - b. Be mindful of all shorthand terms in the table and expand them in the title and description.
 - c. If there are any explicit content, then make sure it is not included.
2. Table Description:
 - a. Summarize the main purpose of the table, the key columns, and the types of data it contains.
 - b. Must include abbreviations or terms that might be unclear, expand them in the description.
 - c. Identify and explain any patterns or notable trends. For example, if certain data points frequently appear together or if there are gaps, mention them.
 - d. Point out the significance of key relationships in the data (e.g., correlations between columns or rows).
 - e. If there are columns with categorical data (e.g., 'Region', 'Category'), mention their distribution and any specific categories of interest.
 - f. If there are any explicit content, then make sure it is not included.

Here are the table headers,
[TABLE HEADERS]

with some table rows:
[SUBSET OF TABLES]

The output should be of format and nothing else.
"Table Title":"" | "Table Description":""

Figure 4: Prompt used for Table Decomposition.

You are an advanced AI specialized in analyzing and reasoning any type of question.
Your task is to generate a meaningful sub-questions and description for the provided question.

The question relates to addressing an information on a table.
The sub-question should be a different concise and clearly reflect the question's primary objective to extract more information from the table, while the description should expand upon the context and provide insights into the key data.
The description must be tailored for query ranking, meaning it should help identify the table's relevance to possible questions.

Guidelines:

1. Sub-Questions:
 - a. Provide a different but a concise, clear questions that reflects the theme and key information required from the table.
 - b. Be mindful of all key words in the question and rationalize them as a logical step to extract information.
2. Question Description:
 - a. Summarize the main purpose of the question and the types of data the table might contains.
 - b. Must include abbreviations or terms that might be unclear, expand them in the description.
 - c. Discuss the significance of key relationships in the question (e.g., relation between the key words).

Here is the Question: **[QUESTION]**

The output must just be of the following format:
"Sub-Questions":"" | "Question Description":""

Figure 5: Prompt used for Query Expansion.

Retriever	n=1			n=3			n=5			n=8			n=10		
	Mistral	Llama3	Qwen	Mistral	Llama3	Qwen2.5	Mistral	Llama3	Qwen2.5	Mistral	Llama3	Qwen2.5	Mistral	Llama3	Qwen2.5
Zero fewshot	39.13	38.31	39.73	45.28	40.76	43.52	44.53	40.55	46.49	44.01	41.11	46.53	41.50	39.15	47.24
1-Fewshot	40.43	37.54	37.94	48.28	37.35	44.69	46.11	38.27	45.07	44.85	40.40	45.46	42.39	39.94	45.05
2-Fewshot	41.56	42.73	38.98	49.29	37.77	43.16	46.69	36.83	46.02	45.35	40.54	45.03	43.53	39.35	44.86
3-Fewshot	40.70	39.41	40.68	48.60	36.90	44.27	47.43	36.94	45.59	44.83	39.97	44.89	43.08	38.87	44.77
4-Fewshot	41.58	40.96	40.99	48.77	37.41	43.86	47.59	36.72	46.19	45.56	40.55	44.64	43.72	40.06	43.37
5-Fewshot	41.17	41.72	40.36	48.50	39.39	43.65	47.70	39.27	45.81	44.94	41.27	44.55	42.31	41.07	44.21

Table 14: F1 performance scores of different LLMs (Mistral-7B, Llama3-8B, and Qwen2.5-7B) across various retrieval configurations using **CRAFT**. Results are shown for increasing numbers of retrieved tables ($n = 1, 3, 5, 8, 10$) and different few-shot prompting strategies. Each row represents a different few-shot variant, with `zero.fewshot` indicating zero-shot performance.

Retriever	n=1			n=3			n=5			n=8			n=10		
	Mistral	Llama3	Qwen2.5	Mistral	Llama3	Qwen	Mistral	Llama3	Qwen	Mistral	Llama3	Qwen	Mistral	Llama3	Qwen
Small Models															
Gold	43.60	41.10	52.87	-	-	-	-	-	-	-	-	-	-	-	-
Stage 2(MT)	38.83	35.58	46.49	41.71	25.78	48.37	40.54	29.11	48.97	38.09	30.38	47.76	36.96	29.53	46.67
Stage 2(FT)	37.72	34.62	46.69	39.46	25.73	47.31	38.40	29.09	49.06	36.20	29.53	47.32	35.12	28.35	46.02
Stage 3 Gemini(MT)	39.38	36.60	49.38	41.02	26.99	50.15	39.14	29.81	48.88	36.16	31.59	46.77	35.44	30.75	47.93
Stage 3 Gemini(FT)	38.28	36.08	48.54	39.38	26.29	49.91	38.10	29.24	49.16	34.95	30.00	47.16	34.32	30.03	46.93
Large Models															
Gold	61.38	59.30	65.31	-	-	-	-	-	-	-	-	-	-	-	-
Stage 2(MT)	54.47	52.06	56.36	56.90	50.52	61.59	57.18	50.44	62.22	54.84	51.43	62.03	53.11	51.45	61.34
Stage 2(FT)	54.80	51.50	56.79	57.52	50.52	60.74	58.12	50.94	62.59	55.66	51.55	61.45	53.25	51.00	61.69
Stage 3 Gemini(MT)	55.43	53.24	59.24	59.41	51.95	63.55	56.77	50.99	63.34	54.12	51.42	63.16	52.86	50.61	62.80
Stage 3 Gemini(FT)	55.06	53.28	58.35	59.21	51.17	62.95	57.27	51.13	62.17	54.65	51.08	61.92	53.72	50.69	60.06

Table 15: F1 performance scores of different LLMs (Top(Small Models): *Mistral-7B-Instruct*, *Llama3-8B-Instruct*, and *Qwen2.5-7B-Instruct*, and Bottom(Large Models): *Mistral-Small-Instruct*, *Llama-3.3-70B-Instruct*, and *Qwen2.5-72B-Instruct*) across various retrieval configurations using CRAFT for OTT-QA dataset. Results are shown for increasing numbers of retrieved tables ($n = 1, 3, 5, 8, 10$) and different table retrievals. Each row represents a different table retrieved using different table embeddings, with **Gold** indicating end-to-end score with the gold table and passage.

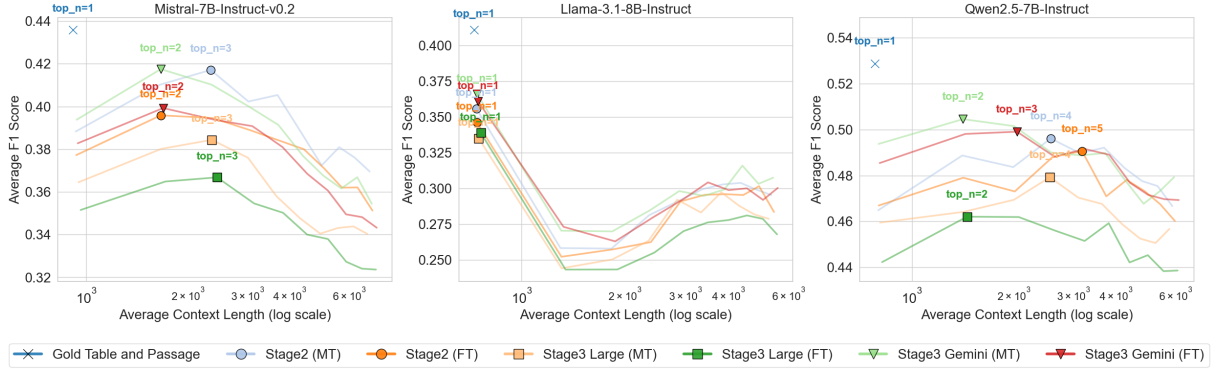


Figure 6: End-to-end performance comparison across context lengths with OTTQA dataset. The figure illustrates the relationship between average context length and F1 score for three models (small size models), evaluated over 10 different n -retrieved tables with the OTT-QA datasets.

C Further End-to-End Analysis

End-to-End OTT-QA Performance Table 15 presents the end-to-end F1 performance of different instruction-tuned LLMs on the OTT-QA dataset using our multi-stage retrieval framework. It compares multiple retrieval strategies across varying depths, highlighting how retrieval design impacts downstream QA accuracy.

We consider the following retrieval settings:

- **Gold:** Uses the ground-truth table (oracle evaluation setup) paired with gold passages to isolate the effect of table retrieval and provide an upper-bound reference.
- **FT (Full Table):** Uses the top- n retrieved table IDs (from Stage 2 or Stage 3) in their full form as context for answering the question.

- **MT (Mini Table):** Uses compact representations of the retrieved tables (based on table IDs), retaining only the most relevant content to reduce noise and improve reasoning efficiency.

Since OTT-QA is a complex hybrid multi-hop benchmark that requires reasoning across both tables and passages, we employ an oracle evaluation setup that pairs the retrieved tables with the corresponding gold passages. This approach isolates the effect of table retrieval quality and enables direct comparison with an upper-bound performance using the gold table (**Gold**). Results are reported for five retrieval depths ($n = 1, 3, 5, 8, 10$), allowing us to assess how the number of retrieved tables influences downstream QA accuracy.

Across both small and large models, highlighted by Figure 6 and Figure 7 respectively,

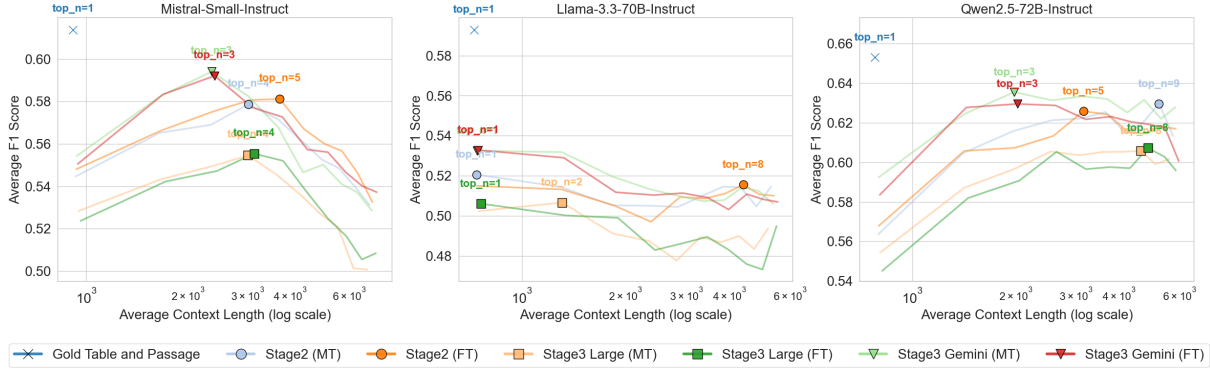


Figure 7: End-to-end performance comparison across context lengths with OTT-QA dataset. The figure illustrates the relationship between average context length and F1 score for three models (large size models), evaluated over 10 different n -retrieved tables with the OTT-QA datasets.

performance trends reveal a clear trade-off between recall and precision. For smaller n , the retrieved tables may not include the gold table, leading to suboptimal results. Conversely, higher n increases the likelihood of retrieving the relevant table but also introduces redundant or noisy tables that dilute reasoning accuracy. Among all settings, Stage 3 (MT) consistently achieves strong performance, particularly for Qwen2.5-72B-Instruct, which reaches its best score at $n = 3-5$, closely matching the gold-table upper bound. This demonstrates the benefit of the Mini-Table (MT) strategy, which compactly represents relevant content while minimizing contextual noise. For large-scale models such as Qwen2.5-72B-Instruct, Stage 3 Gemini (MT) yields the most balanced results, highlighting how our approach generalizes effectively across model sizes by achieving a “sweet spot” between retrieval breadth and contextual precision.

D Error Analysis

Our error analysis reveals several consistent patterns in CRAFT’s behavior under imperfect retrieval conditions. We find that retrieved tables often exhibit high topical similarity to the query, yet do not necessarily contain the specific evidence required for multi-hop reasoning, indicating that surface-level relevance is an unreliable proxy for answerability. Moreover, the rank of the gold table does not directly correlate with end-to-end QA performance, therefore even lower-ranked tables can support correct predictions when paired with effective mini-table representations that isolate relevant evidence. We also observe that standard retrieval metrics such as R@1 and R@5 tend to underestimate overall system capability,

as downstream reasoning components can compensate for imperfect retrieval by extracting useful signals from non-top-ranked candidates. Finally, using full tables as input introduces substantial noise, frequently leading to missed answers or spurious predictions, whereas mini-table representations reduce irrelevant context and enable more accurate reasoning. Taken together, these findings suggest that improving QA performance requires joint optimization of retrieval and reasoning, with intermediate representations such as mini-tables playing a central role.

Case Study I: Full Table Fails Despite Correct Retrieval. Table 16 presents an example where the gold table (*1997_AFL_draft_1*) is ranked third among the top-5 retrieved tables, yet the full-table approach fails to extract the correct answer, responding with *Pass* (i.e., unable to find relevant information). In contrast, the mini-table design successfully identifies *Jamie Elliott* as the answer. This illustrates how CRAFT can exhibit low Recall@1 while still producing correct final answers through focused context construction.

Case Study II: Full Table Hallucination with Closely-ranked Retrieval. Table 17 presents a more nuanced failure mode. The gold table (*List_of_songs_recorded_by_Elisa_0*) is ranked second—a relatively strong retrieval result—yet answer quality diverges sharply between the two approaches. The full-table approach returns *Lotus (2003)*, a plausible but incorrect answer drawn from a different row in the retrieved context. The mini-table approach correctly returns *Asile’s World (2000)* by isolating the relevant row corresponding to Elisa’s most popular original song.

Taken together, these case studies highlight that

Attribute	Value
QID	<i>e4197b9b9bd82bdd</i>
Question	Who was the St. Kilda pick 11 in the 1997 AFL draft?
Gold Answer	Jamie Elliott
Gold Table	<i>1997_AFL_draft_1</i>
Gold Table Rank	3
Top 5 Retrieved TableIDs	
1	<i>1998_AFL_Draft_3</i>
2	<i>List_of_St_Kilda_Football_Club_records_and_statistics_8</i>
3	<i>1997_AFL_draft_1</i> ✓
4	<i>1998_AFL_Draft_2</i>
5	<i>List_of_St_Kilda_Football_Club_records_and_statistics_9</i>
LLM QA Step	
Full Table	Pass (no answer found) ×
Mini-Table	Jamie Elliott ✓

Table 16: Case Study I: The gold table is ranked third. The full-table approach fails to extract the correct answer, while the mini-table design succeeds by isolating the relevant row.

the benefit of mini-tables is not limited to weak retrieval scenarios. Even when relevant tables are retrieved at high ranks, full-table context can introduce sufficient noise through topically similar but irrelevant rows to cause model failure or hallucination. Mini-table construction acts as a targeted denoising step, enabling the downstream QA model to focus on the precise evidence needed for multi-hop inference.

Attribute	Value
QID	<i>bde8169851a494cf</i>
Question	What album did Elisa’s most popular original song come out on?
Gold Answer	Asile’s World
Gold Table	<i>List_of_songs_recorded_by_Elisa_0</i>
Gold Table Rank	2
Top-5 Retrieved Table Ids	
1	<i>List_of_songs_recorded_by_Elisa_1</i>
2	<i>List_of_songs_recorded_by_Elisa_0</i> ✓
3	<i>Elisa_(Italian_singer)_1</i>
4	<i>Elisa_(Italian_singer)_0</i>
5	<i>Raquel_Olmedo_1</i>
LLM QA Step	
Full Table	Lotus (2003) ×
Mini-Table	Asile’s World (2000) ✓

Table 17: Case Study II: Even with the gold table ranked second, the full-table approach hallucinates an incorrect answer, while the mini-table approach correctly isolates the relevant information.