CRAFT

Training-Free Cascaded Retrieval for Tabular QA

Adarsh Singh[1*α] Kushal Raj Bhandari[2*] Jianxi Gao[2] Soham Dan[3δ] Vivek Gupta[1δ]

[1]Arizona State University [2]Rensselaer Polytechnic Institute [3]Microsoft Research

*Equal Contribution δJoint Supervision

α *on the industry job markets for ML Engineer*





ACL 2026 SAN DIEGO JULY 2–7

Retrieval & Its Significance in RAG

Retrieval bounds everything downstream, the generator(LLM) can only reason over what the retriever finds.

Table Retrieval

Table Retrieval focuses on retrieving relevant tables that best match a natural language query.

It bridges the gap between unstructured user queries and structured tabular data.

***Fig.** A query from **NQ-Tables Dataset**, the answer is in a single cell single cell.*

Query: Philadelphia is known as the city of what?

169K tables in corpus

↓

Retrieve the Gold Table

↓

City	State	Nickname	
Boston	MA	Beantown	…
Philadelphia	PA	Brotherly Love	
Chicago	IL	Windy City	…
…	…	…	…

Ans: Brotherly Love

Prior Fine-tuned Approaches

Contrastive Learning

Dense bi-encoder trained with row/column **positional tokens**, trained on with hard negative mining.

→ **High recall@1**

\+ **Field-aware Matching**

Field-aware hybrid matching: sparse for cells, **dense for titles**. Tailors retrieval strategy per table field.

→ **State-of-the-art retrieval**

Joint Retriever-Reader

Joint retriever-reader training with late interaction and explicit relevance signals for end-to-end QA.

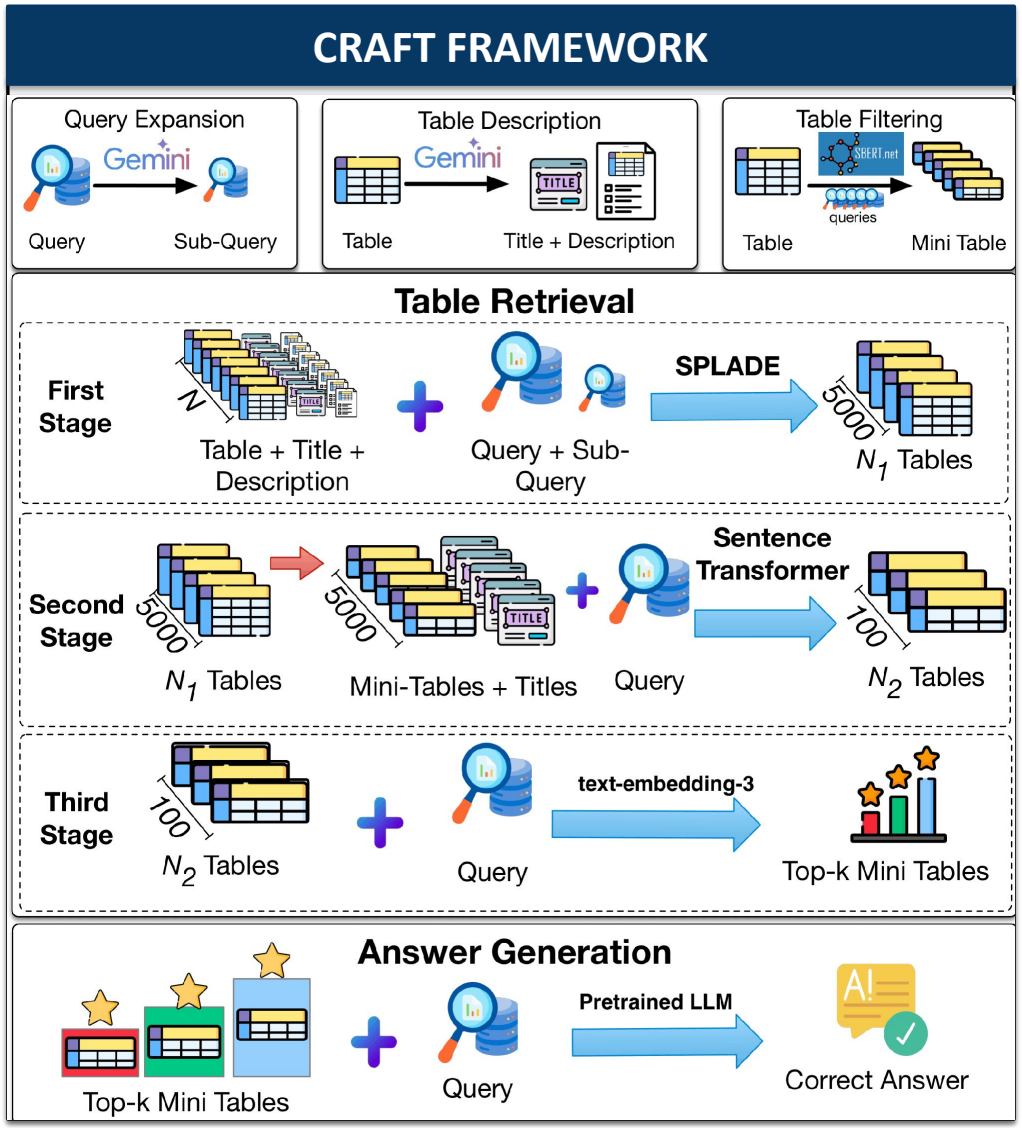
→ **state-of-the-art end-to-end qa (F1)**

Major Drawback

All require **training data, hard negative mining,** and **domain-specific fine-tuning.**

Overview

- Modular
- Cascaded Retrieval
- **No fine-tuning**

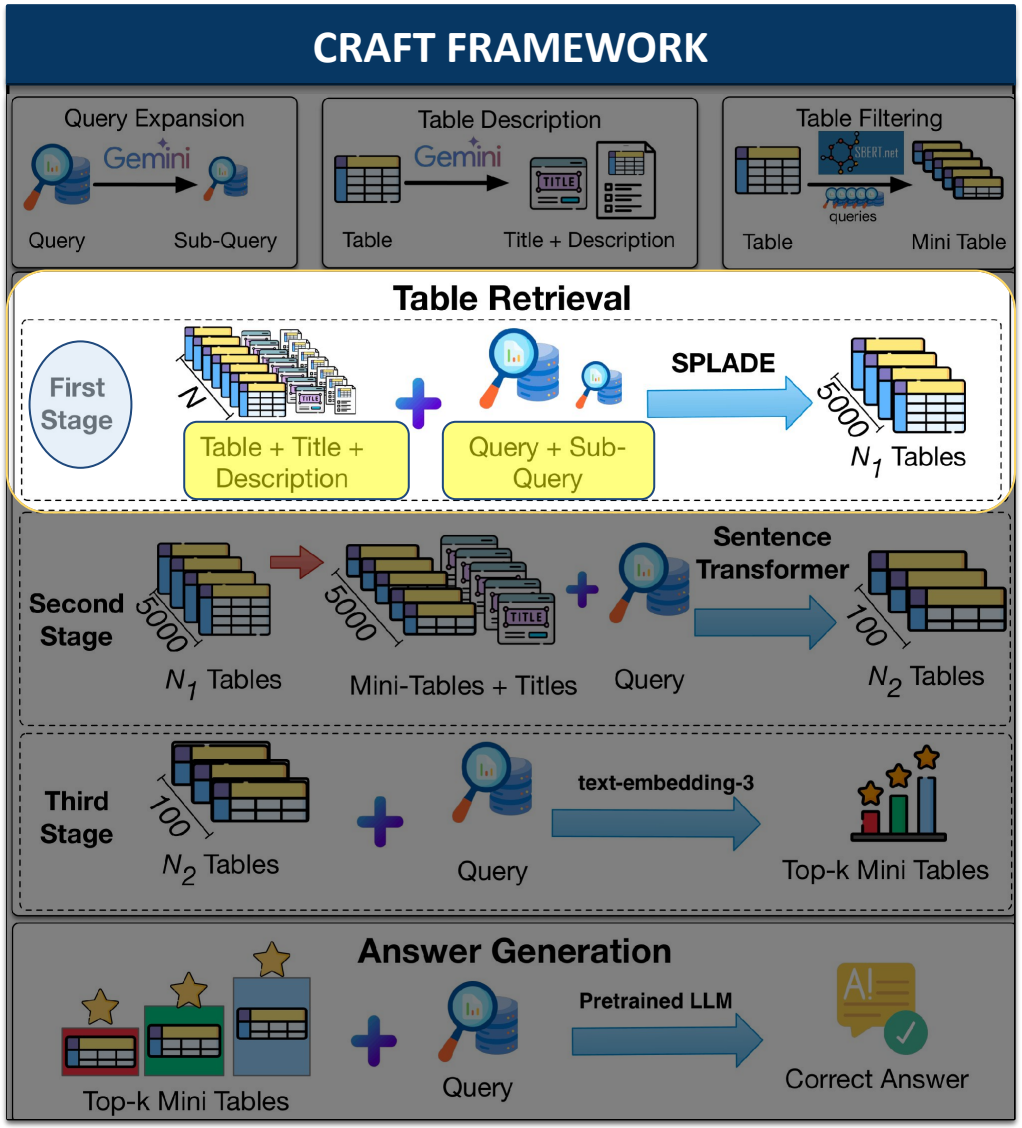


CRAFT - Stage 1

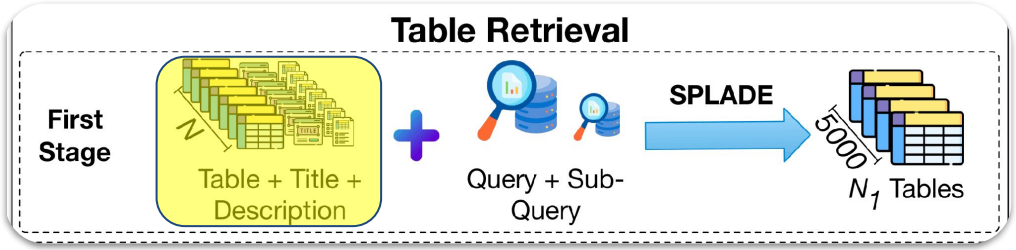
LLM Generated Abstractions

① Table Descriptions - Offline preprocessing

② Query Expansion - Online processing



Stage 1: Table Descriptions



Model: Gemini Flash 1.5

- Generates informative **title** + **paragraph summary** per table.
- Enriches sparse index with semantic context beyond just column headers.

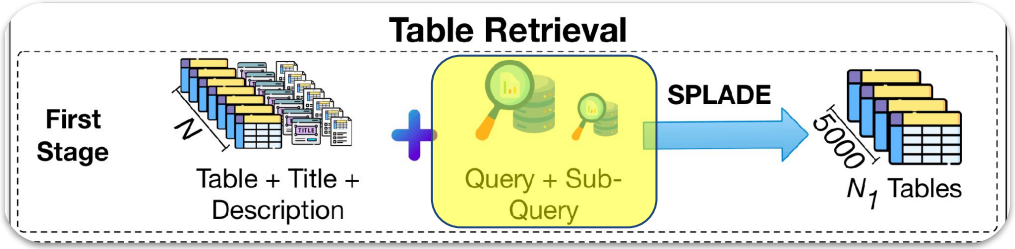
Effect of Table Descriptions

Configuration	*R@1*	*R@10*	*R@50*
SPLADE (base)	*31.06*	*69.05*	*86.85*
+ Table Descriptions	***32.64***	***70.62***	***87.59***

+2.73% Recall@10

Table. Retrieval Performance of SPLADE on NQ-Tables

② Query Expansion



Expanding each query gives the sparse index a richer lexical signal to match.

NL Query: Who won three stanley cups in a row?

Expanded Query: Which **NHL teams** have achieved three **consecutive** stanley cup victories?

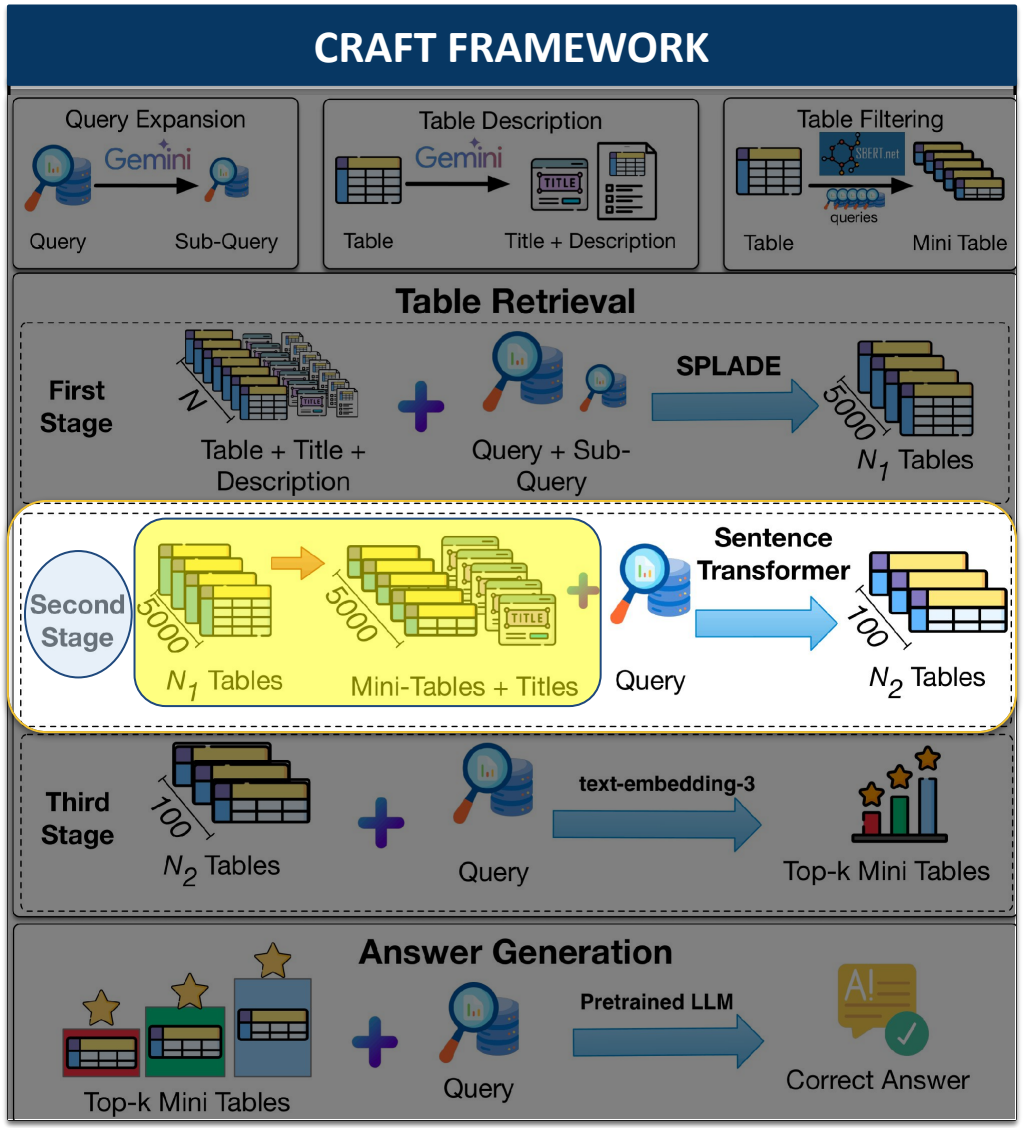
Effect of Query Expansion in Stage 1 (NQ-Tables)

Configuration	R@1	R@10	R@50
SPLADE (base)	31.06	69.05	86.85
+ Table Descriptions	32.64	70.62	87.59
+ Table Descriptions + Expanded Query	**34.38**	**72.90**	**91.62**

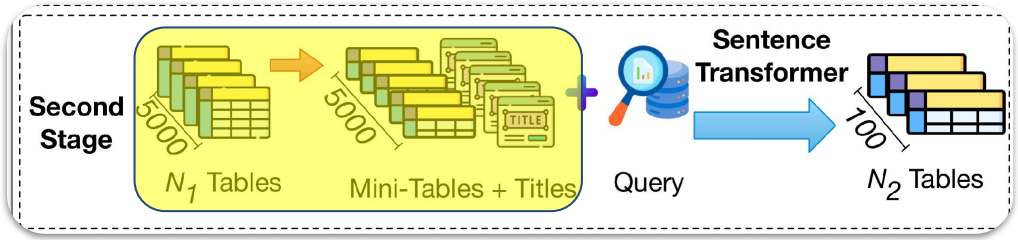
+ 5.61% Recall ↑

CRAFT - Stage 2

① Mini-Table



③ Mini-Table Creation



Retains relevant rows/table by semantic similarity to the query

- Retains top-5 most relevant rows per table
- Trims noise while improving Recall
- Lets more tables fit in the context window.

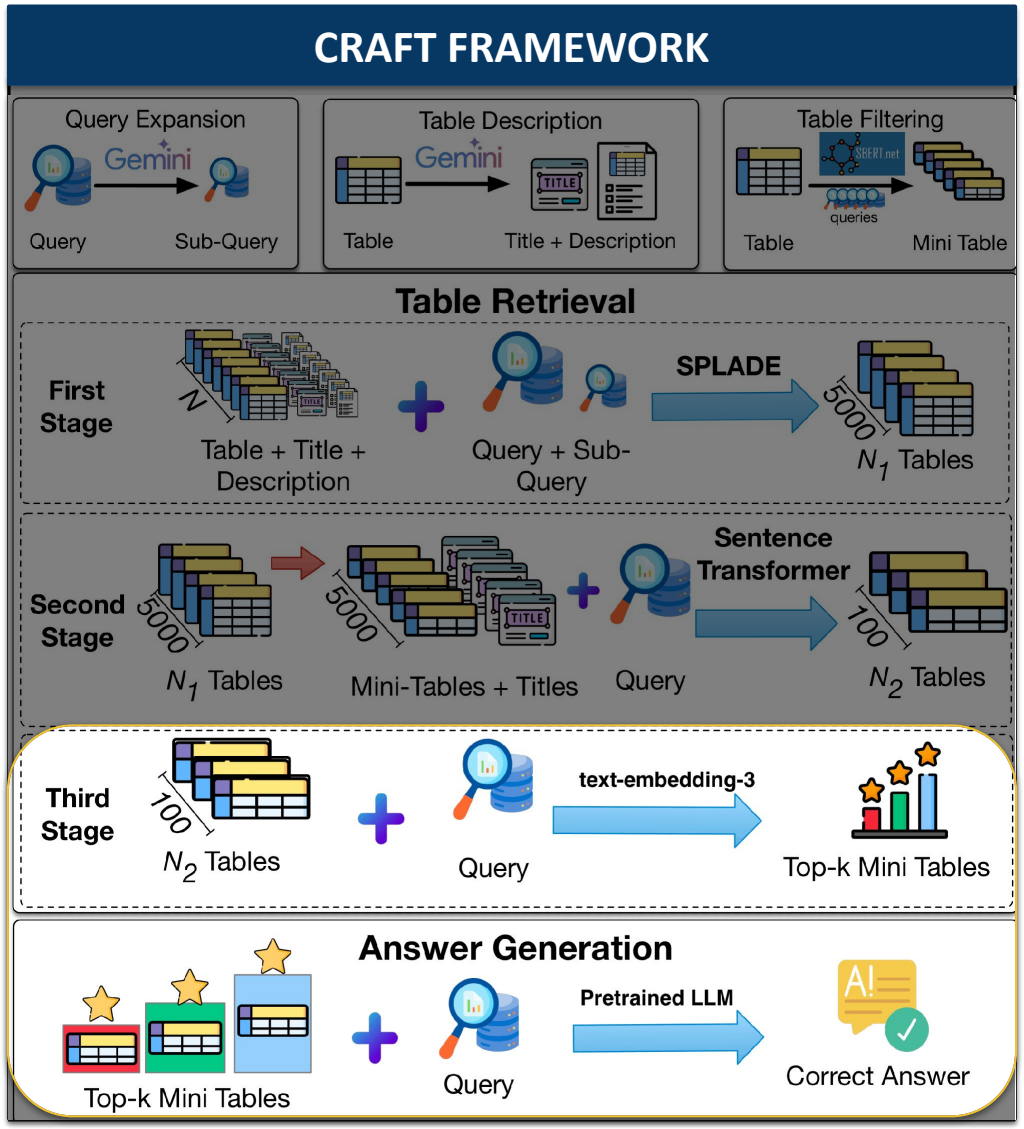
Effect of Rows Reduction in Stage 2 (NQ-Tables)

Configuration	R@1	R@10	R@50
Full table	37.64	79.76	94.45
– Irrelevant Rows (Mini-Table)	36.65	**82.91**	**96.08**

→ **↑ Recall Increases**

CRAFT: Stage 3

- Plug and Play
- Mini-tables reduce embedding cost.



DATASETS: NQ-Tables & OTT-QA

Metric	NQ-Tables Single-Hop	OTT-QA Multi-Hop
Queries (Train / Test)	966	2,214
Tables in corpus	169,898	419,183
Avg. Rows / Table	10.70	12.90
Avg. Columns / Table	6.10	4.80
Gold Tables / Query	1.05	1.00
Avg. Query Words	8.94	22.82

Retrieval Results

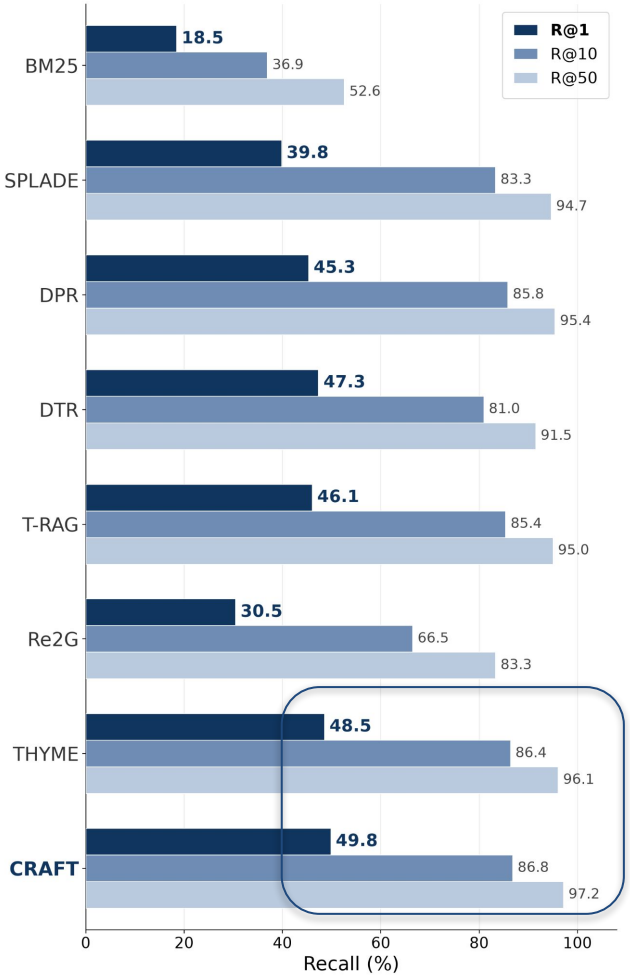


Fig. CRAFT achieves the **highest** recall performance on **NQ-Tables**, while remaining competitive at deeper depths (@10,@50) on **OTT-QA**.

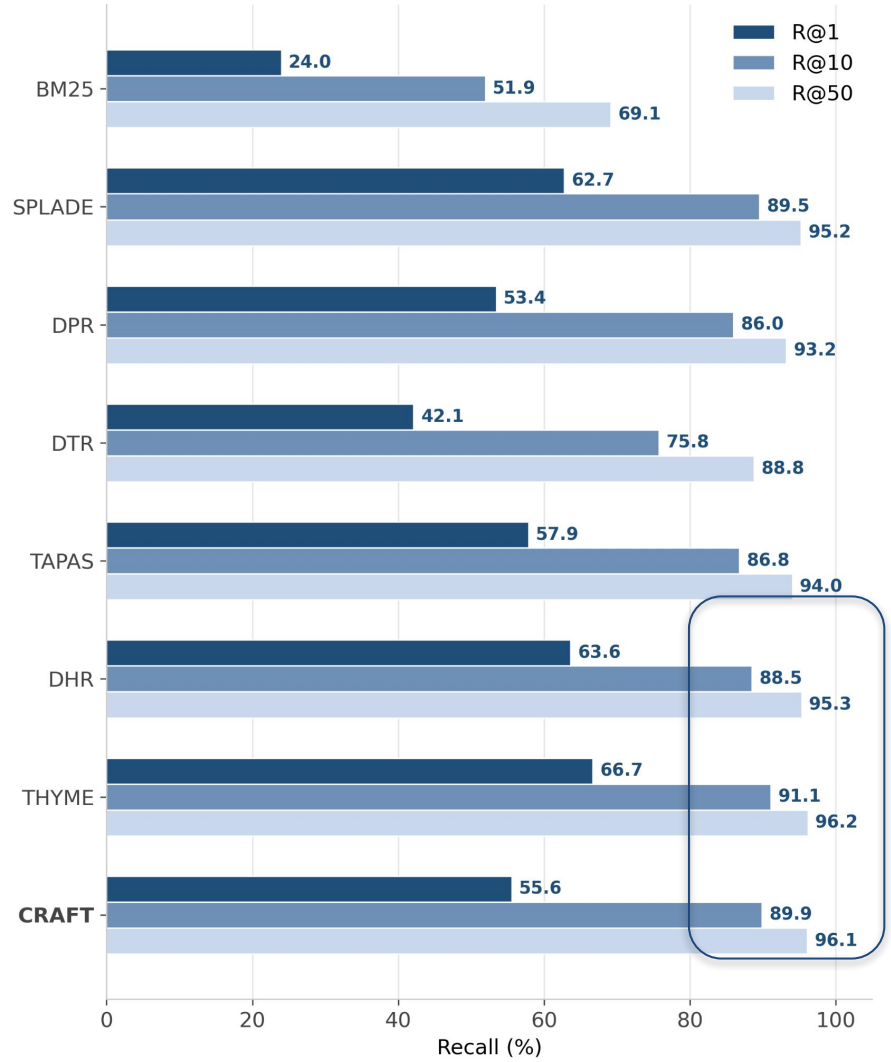
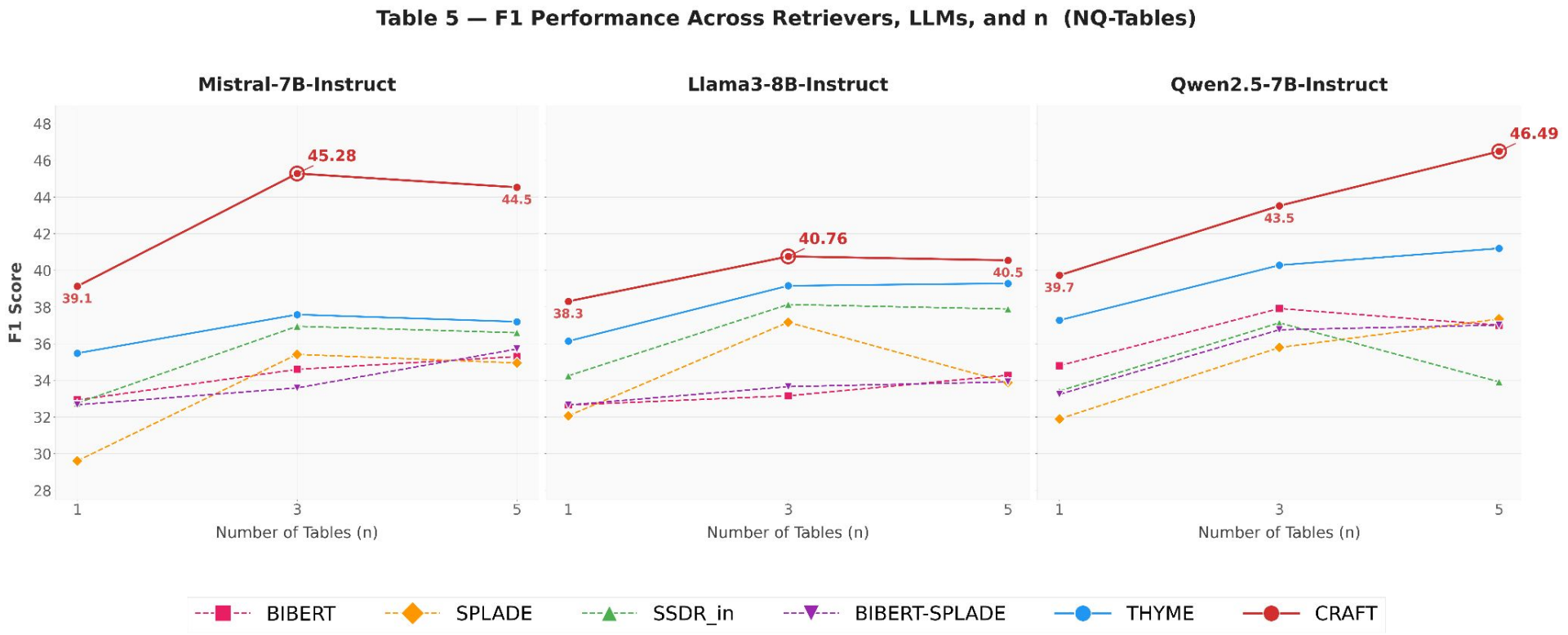


Table. Retrieval performance on **NQ-Table**

Table. Retrieval performance on **OTT-QA.**

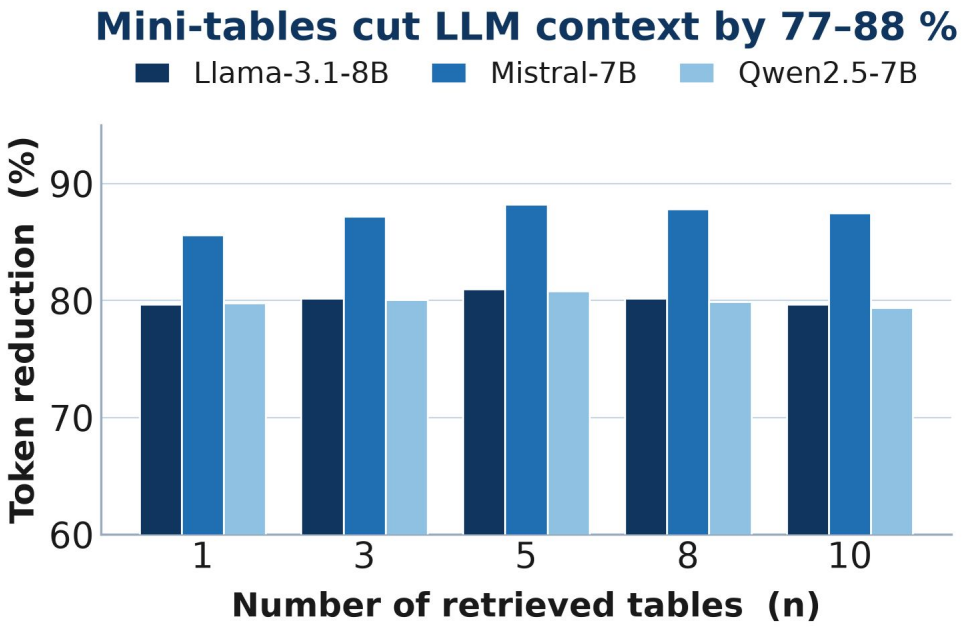
End to End QA Results



CRAFT achieves highest end to end QA Results across models and number of tables(n)

Additionally, Token reduction across three LLMs as the number of retrieved tables increases.

Token Consumption: Mini-Table vs Full Table



Token reduction across three LLMs as the number of retrieved tables increases.

ROBUSTNESS under Query Paraphrasing

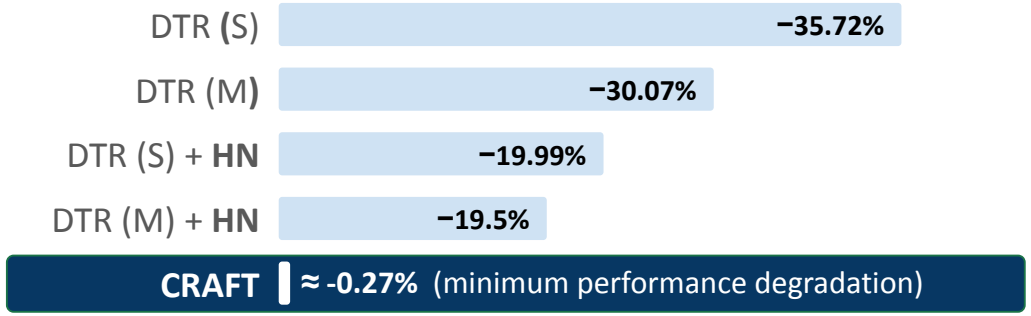
Original Question : Philadelphia is known as the city of what?

Paraphrase Query (Gemini 1.5 Flash)

Perturbed : What descriptor follows *"the city of"* when referencing Philadelphia?
Answer : Brotherly Love

Avg Performance degradation under query paraphrasing

% drop in R@1 — shorter bar is better



* **HN** denotes fine-tuned on NQ-Tables, with Hard Negatives

Fine-tuned models overfit to the training query distribution and are brittle to query rephrasing.

Conclusion

By combining **sparse retrieval**, **semantic reranking**, and **LLM-generated table abstractions**, CRAFT achieve **state-of-the-art performance** w/o **training**

- with lower cost (token reduction),
- greater robustness,
- and easier portability to new domains.

I am on the industry job markets for ML Engineer

Adarsh Singh, adarsh.fun/

Paper & Code