

1 Accurate structured queries over unstructured text remain an open challenge



A sports
journalist asks

Which players **over 30** play for a **Los Angeles team**, and how many **championships** has that team won?

```
SELECT player, city, titles
FROM players JOIN teams ...
WHERE age > 30 AND city = 'Los Angeles'
```

UNSTRUCTURED TEXT DOCUMENTS

PLAYER DOCUMENT

LeBron James was born in **1984**. He plays for the **Lakers** and has **4 rings**.

TEAM DOCUMENT

The **Los Angeles Lakers** are based in **Los Angeles** and have won **17** championships.

Expected answer		
Player	City	Titles
LeBron James	Los Angeles	17

X LLM-only answer		
Player	City	Titles
LeBron James	Los Angeles	4
Answers LeBron's 4 rings, not the Lakers' 17 titles in the team document.		

Derive age 1984 → over 30

Resolve entity Lakers = Los Angeles Lakers

Join records Player.team ↔ Team.name

Generative models do not ensure correctness when answers require derived attributes, entity resolution, and joins.

2 An ideal framework needs high accuracy, low token cost, and low latency

Retrieval-Augmented Generation

Lewis et al., NeurIPS 2020

Passage retrieval does not support complex relational joins.

Result accuracy

Low ↑

Token cost

Low ↓

Latency

Low ↓

Per-query structured extraction

ReDD (Chao, PVLDB 2026) · DocETL (Shankar, PVLDB 2025)

Every query repeats document scanning and LLM extraction.

Result accuracy

High ↑

Token cost

High ↓

Latency

High ↓

Query-unaware structured table extraction

SQUiD (Sadia, EMNLP 2025)

One fixed schema can omit attributes and keys future queries need.

Result accuracy

Low ↑

Token cost

Low ↓

Latency

Low ↓

3 QuWARTS leverages past queries for workload-guided table synthesis

Too little synthesis

X Low accuracy, low cost

QuWARTS synthesizes the workload-relevant middle

✓ High accuracy, low cost

Synthesize everything

X High cost, high accuracy

data / object process actor always on demand offline online

OFFLINE TABLE SYNTHESIS

DONE ONCE

INPUT

Text documents
unstructured sources

Past query workload
Q1: SELECT player WHERE age > 30
Q2: SELECT team, city
Q3: SELECT player, titles JOIN team
...

QuWARTS

Table Synthesis

Transform document values so the workload can run as SQL

Attribute derivation

does birth_year → age
enables age > 30

Join-key alignment

does Player.team = Team.name
enables relational joins

Entity resolution

does Lakers → Los Angeles Lakers
enables one entity form

Predicate extraction

does materialize city, titles
enables city = 'Los Angeles'

uses LLM

SHARED OUTPUTS

produced once · reused online

Relational tables

PLAYERS(name, age, team)
TEAMS(name, city, titles)

Attribute index

player-doc → name, team, mvp
team-doc → name, city, titles

ONLINE QUERY ANSWERING

PER QUERY



Query

SELECT player, city, titles, mvp
WHERE age > 30 AND city = 'LA'

Result

Player City Titles MVP
LeBron LA 17 4

Query Answering

Query Engine

SQL on shared tables

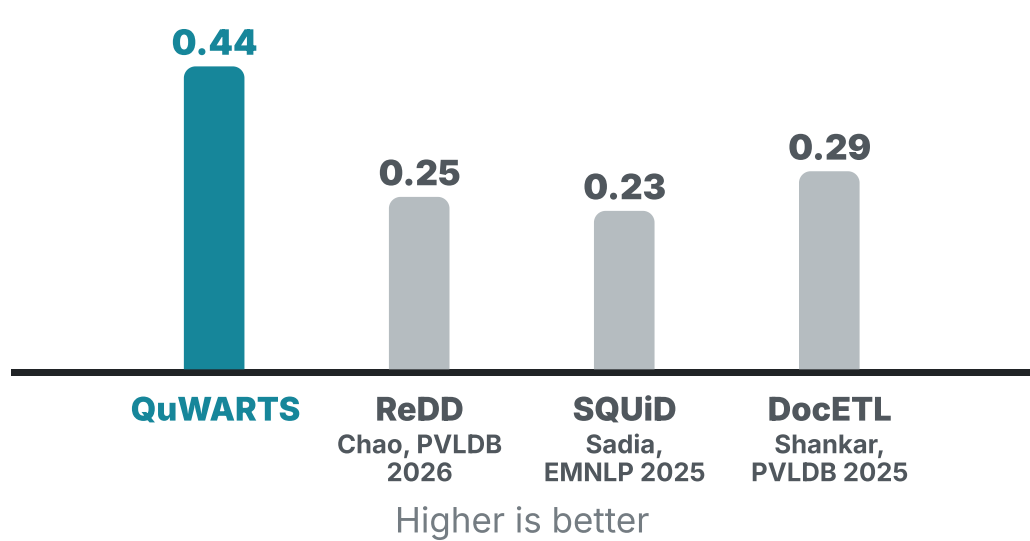
if attribute missing

On-Demand Extraction

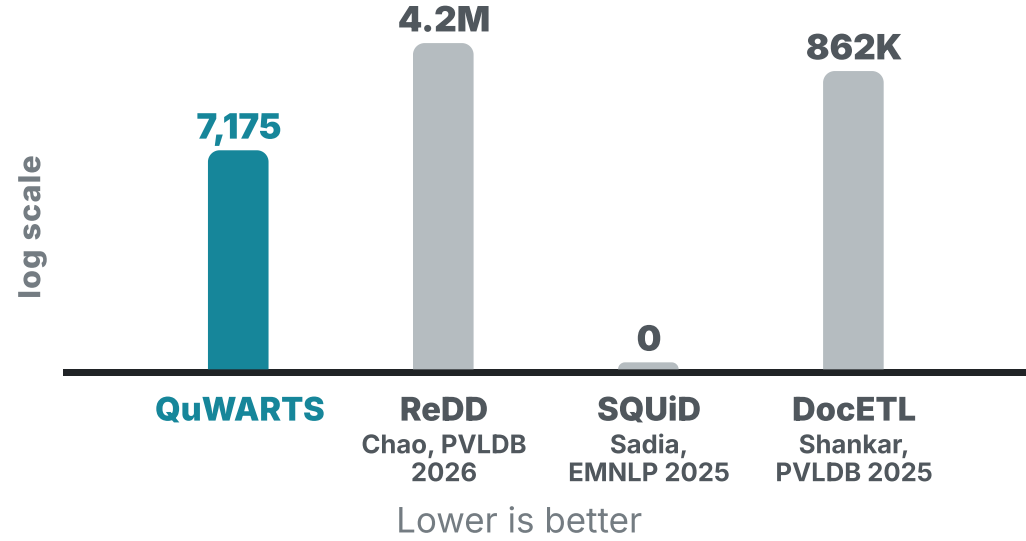
uses the shared attribute index

4 QuWARTS strikes a balance between accuracy, token cost, and latency

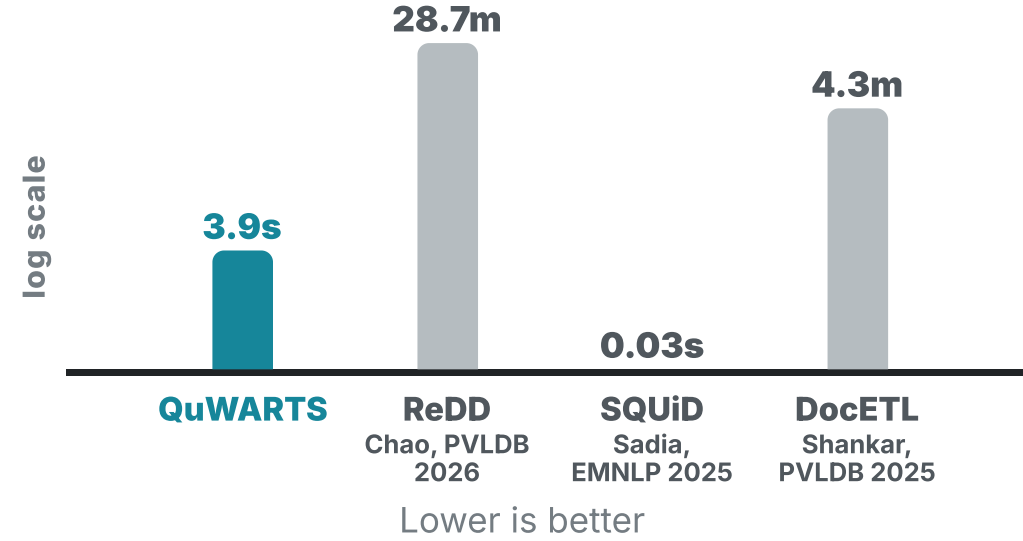
Result accuracy (F1)



Tokens per query



Query latency



FUTURE DIRECTIONS

Budget constraint
Synthesize and query under a token or cost budget.

Multiple database synthesis
Emit more than one related database from the corpus.

New data sets
Evaluate on additional domains and corpora.

References

- [1] D Chao, K Chen, N Guan, and N Koudas. Relational Deep Dive: Error-Aware Queries Over Unstructured Data. PVLDB 2026
[2] M Sadia, Z Yang, Y Xiao, A Chen, and A R Chowdhury. SQUiD: Synthesizing Relational Databases from Unstructured Text. EMNLP 2025
[3] S Shankar, T Chambers, T Shah, A Parameswaran, and E Wu. DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing. PVLDB 2025

Paper · Code · Demo

github.com/aritra741/QuWARTS

