

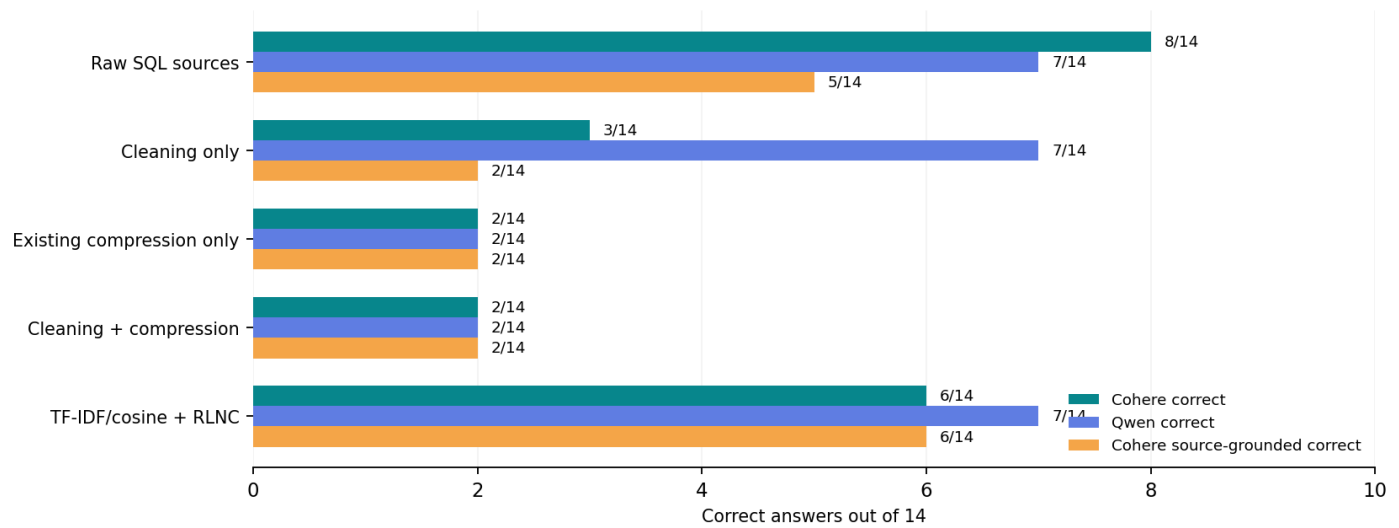
# LLM-BOOSTER | TECHNICAL BENCHMARK

14-question matched comparison | Local Qwen vs hosted Cohere North Mini Code

### Two distinct inference environments

Qwen: locally executed | Cohere North Mini Code: remotely hosted/API-based

## Accuracy results across five configurations



## Reported outcomes (correct responses out of 14)

| Condition                 | Cohere | Qwen | Grounded |
|---------------------------|--------|------|----------|
| Raw SQL sources           | 8/14   | 7/14 | 5/14     |
| Cleaning only             | 3/14   | 7/14 | 2/14     |
| Existing compression only | 2/14   | 2/14 | 2/14     |
| Cleaning + compression    | 2/14   | 2/14 | 2/14     |
| TF-IDF/cosine + RLNC      | 6/14   | 7/14 | 6/14     |

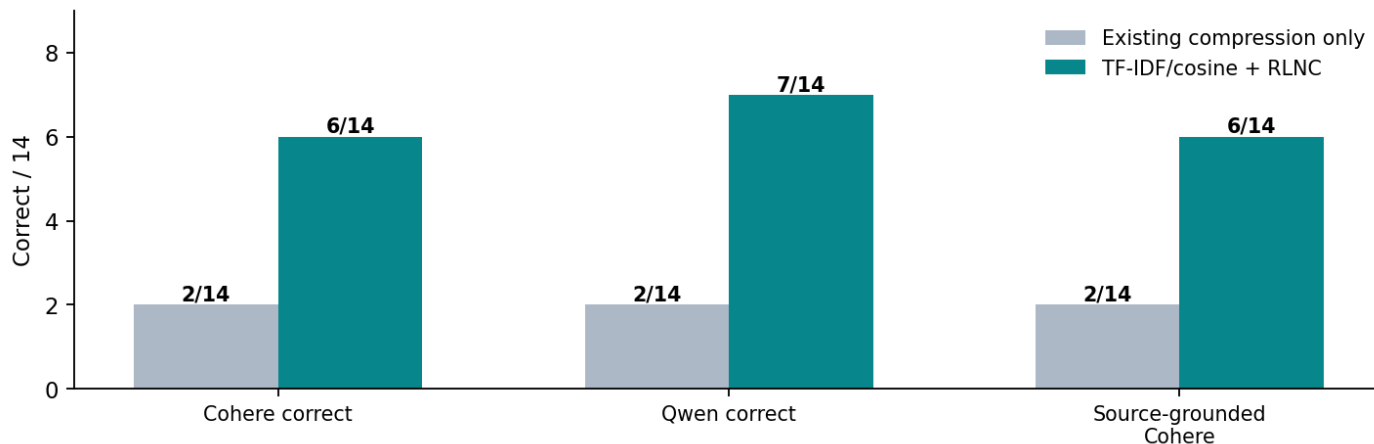
### Core observation

TF-IDF/cosine + RLNC improves results versus existing compression alone, but the dataset does not isolate RLNC from the retrieval component.

# LLM-BOOSTER | TECHNICAL BENCHMARK

Interpretation | Token efficiency hypothesis | Recommended validation

## RLNC-associated accuracy improvement



### Against existing compression alone

Cohere: +4/14 | Local Qwen: +5/14 | Grounded correctness: +4/14

## Cleaning mode: potential token-consumption advantage

Cleaning can remove redundant, irrelevant or noisy context before an LLM call.

This can reduce input tokens and context-window pressure when less text is sent.

For a hosted model, fewer input tokens may also reduce token-based API charges;

for local Qwen, the main potential benefit is lower inference workload.

### Measurement status

No token counts were provided: savings are plausible, not measured here.

## Accuracy trade-off observed in this test

Cleaning alone: Qwen 7/14 (unchanged vs raw SQL); Cohere 3/14 (vs 8/14 raw).

Grounded correctness: 2/14 with cleaning only, versus 5/14 using raw SQL.

Cleaning + compression: 2/14 across all measures, no accuracy uplift shown.

## How to validate efficiency and isolate RLNC

Log prompt/input tokens, output tokens, total tokens, latency and API cost per run.

Compare raw vs cleaned on identical questions/models; report percent token reduction.

Test retrieval with and without RLNC; rerun with fixed settings and more questions.

Caution: n=14; results do not prove causal gains or statistically robust generalization.