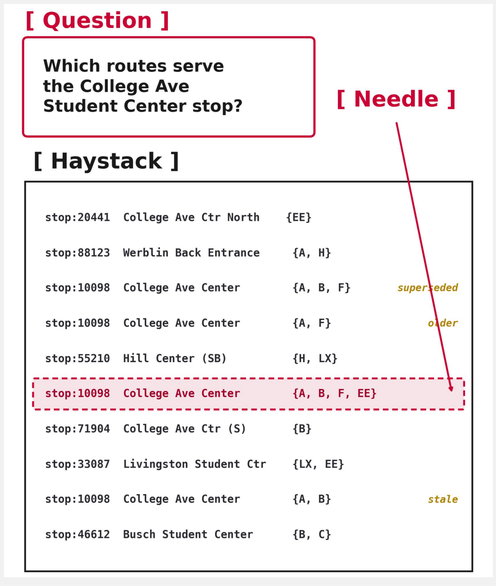
Benchmarking Information Processing Methods for Edge-Local Data Stations

RUTGERS
WINLAB | Wireless Information Network Laboratory

Sanjana Ashok | Mentors: Xiaotian Zhou, Zihao Ding
Advisors: Dipankar Raychaudhuri, Yao Liu

Overview/Motivation

Edge devices like campus servers, sensors, and vehicles increasingly run AI agents on local, private data the cloud never sees. But a model can only answer from what fits in its context window, so as the local store grows, something has to choose what goes in, and that choice, not the model, decides whether the answer is right.

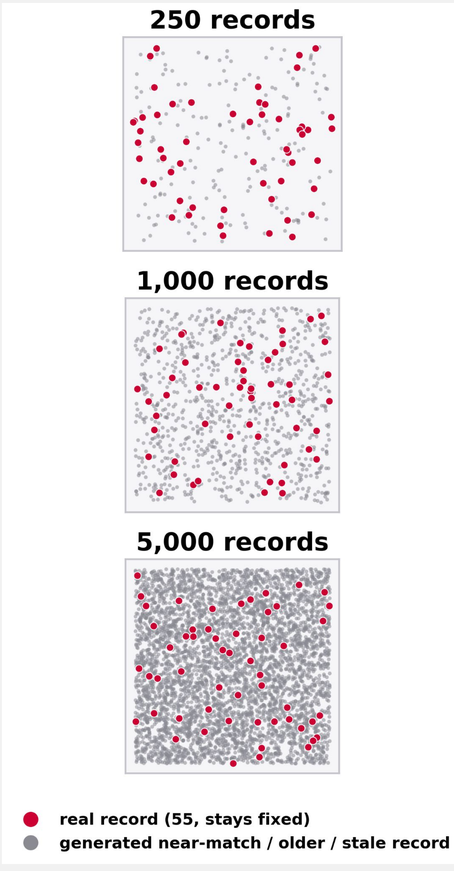


- Longer context does not fix this: recall falls as the candidate pool grows
- The harness, not the model, controls what reaches the prompt

This benchmark measures how different delivery modes trade token cost against accuracy on a controlled, growing store, and separates retrieval failures from reasoning failures.

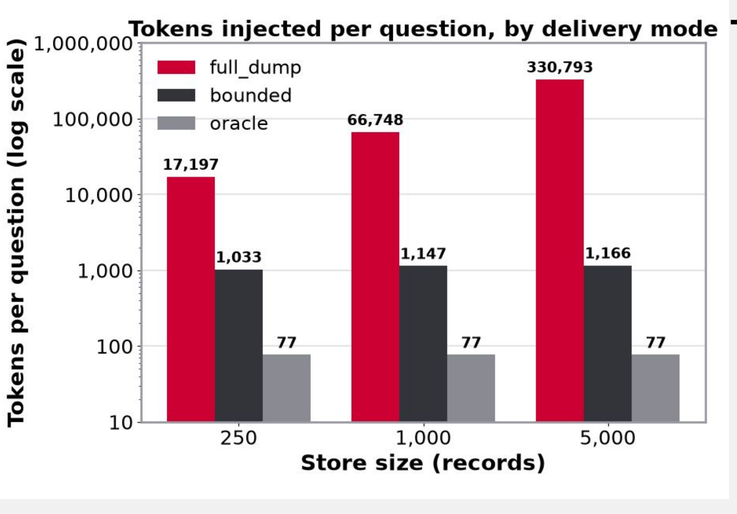
Methods

The benchmark began as a paper-QA workload (four context conditions, two harnesses). The current benchmark extends it:

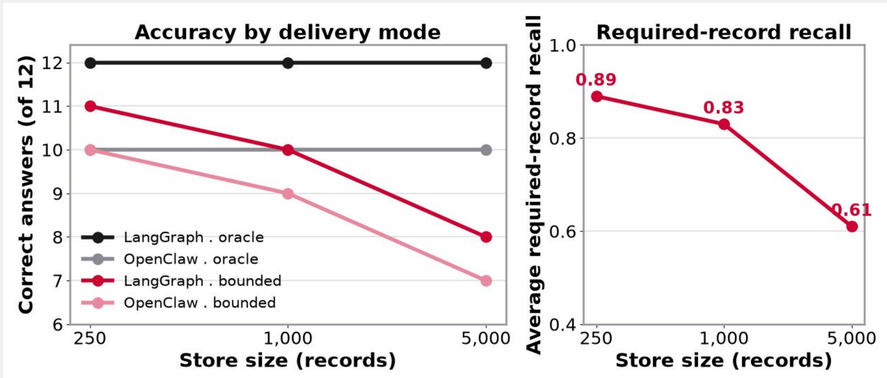


- **Data store**: 55 real PassioGo records (12 routes, 43 stops), frozen in a SHA-256 snapshot, scaled to 250/1,000/5,000 with labeled near matches, older versions, and stale conflicts.
- **Delivery modes**: none (control), full dump (whole store), bounded (top records in a 1,200-token budget), oracle (exact required records).
- **Bounded retrieval**: a deterministic keyword-overlap scorer, ranked to the budget.
- **Setup**: two harnesses (LangGraph, OpenClaw), one model (qwen3.6-plus), shared DeepSeek-V4-Flash selector and judge. 24 tasks encompassing six behaviors.
- **Controls**: web off, isolated sessions, single context delivery, model identity logged, cache counted, judge regraded.
- Answers are keyed to record IDs absent from pretraining, so a correct answer must come from retrieved evidence.

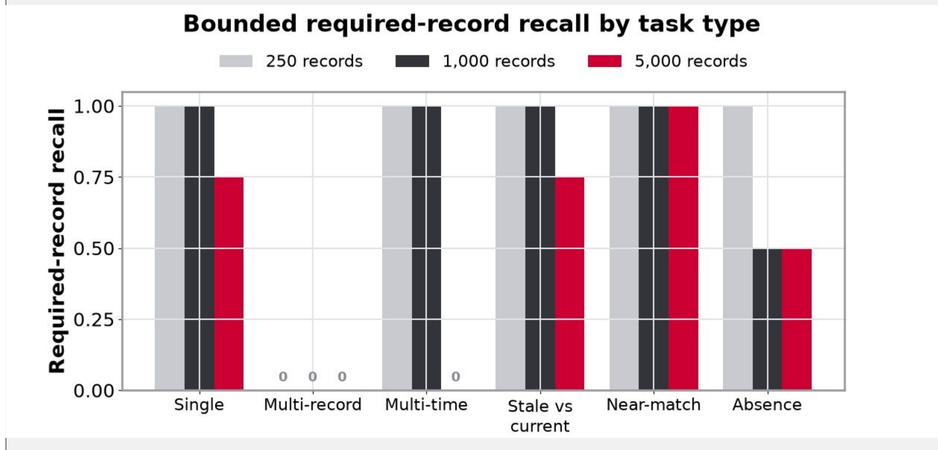
Results



- Full-context delivery grows about 19x in token exposure as the store grows 20x; bounded stays near 1,000 to 1,200, oracle near 77.



- Holding that budget is not free: required-record recall falls from 0.89 to 0.61 as near matches and stale conflicts compete for it.



- Failures are not uniform: multi-record synthesis never completes even at 250 records, while single-record and stale-versus-current hold until the large store.

Conclusions/Future Work

- Full-context delivery gets expensive as the store grows, while bounded holds its budget but loses required-record recall.
- In this workload the limit was retrieval, not model reasoning or context capacity.
- A valid harness comparison requires that every information path is controlled and logged.
- Next: test an IDF or BM25 scorer for recall recovery, run the diagnostic on all 24 tasks, and vary snapshot freshness to study staleness.

WINLAB