Token Write Amplification in LLM-Mediated State Stores

Shijun Yang†, Shuheng Gao†, Chaomei Yan †, Guoli Wei†, Yongkun Li†

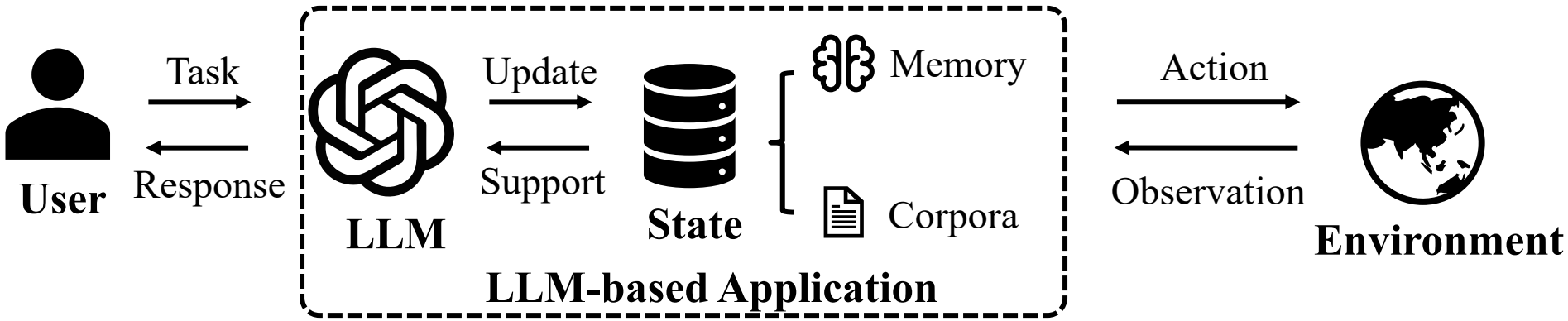
†University of Science and Technology of China



HotStorage 2026

Background

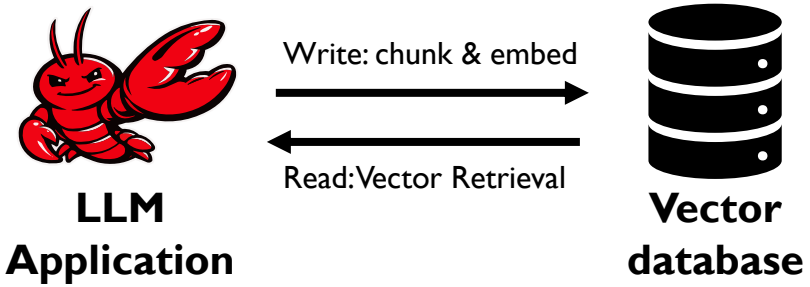
- LLM Application
  - ReAct-style reasoning and action



- Accumulated interactions **inflate** the context
  - **Lower reasoning quality**: relevant signals are diluted
  - **Higher inference latency**: more tokens are processed

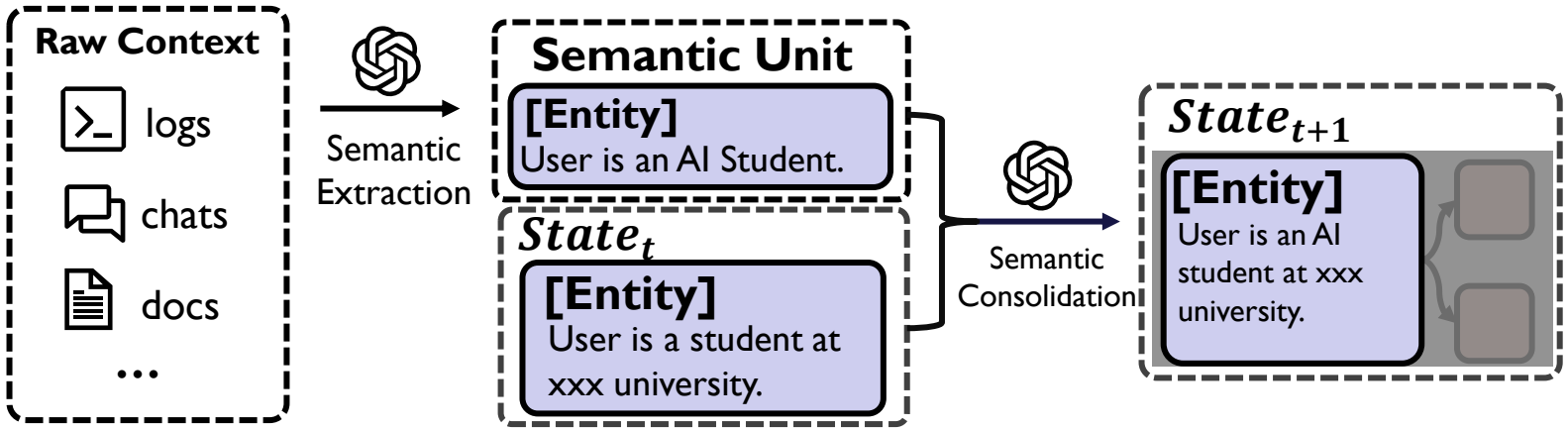
Background

- A Naïve Solution: RAG
  - Externalize context and retrieve it on demand



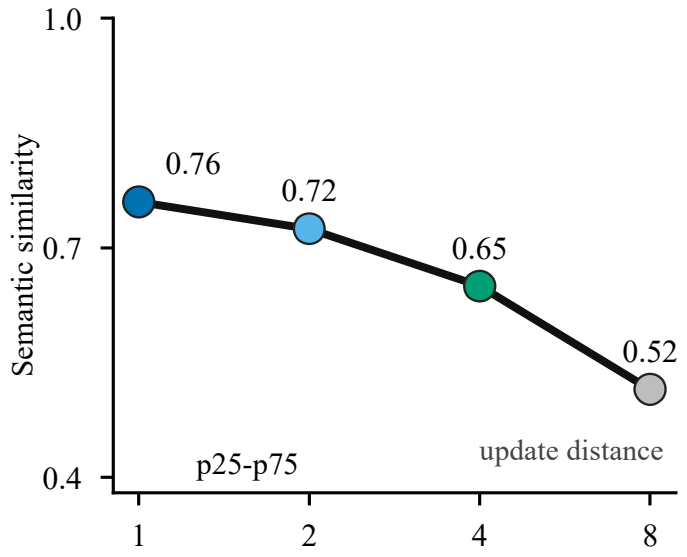
Fragmented State
Limited Multi-hop Retrieval

- A New Storage Paradigm: **LLM-Mediated State Stores**
  - Use LLMs to extract and consolidate state on the write path



Motivation

- A new write path cost: LLM Tokens
  - State maintenance shifts storage cost from bytes to tokens
- **Insight:** LLM Writes Exhibit Semantic Locality

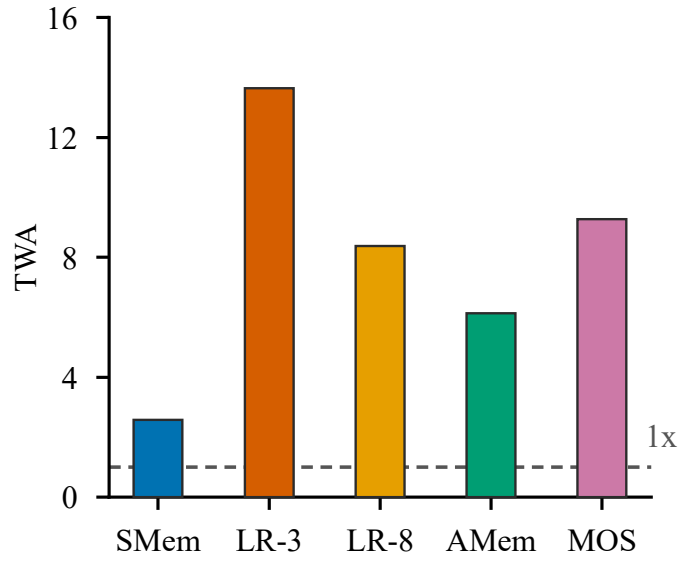


① Nearby writes exhibit high semantic similarity

- similarity gradually decreases as the distance increases
- revealing semantic locality among LLM writes

Motivation

- A new write path cost: LLM Tokens
  - State maintenance shifts storage cost from bytes to tokens
- **Insight:** LLM Writes Exhibit Semantic Locality



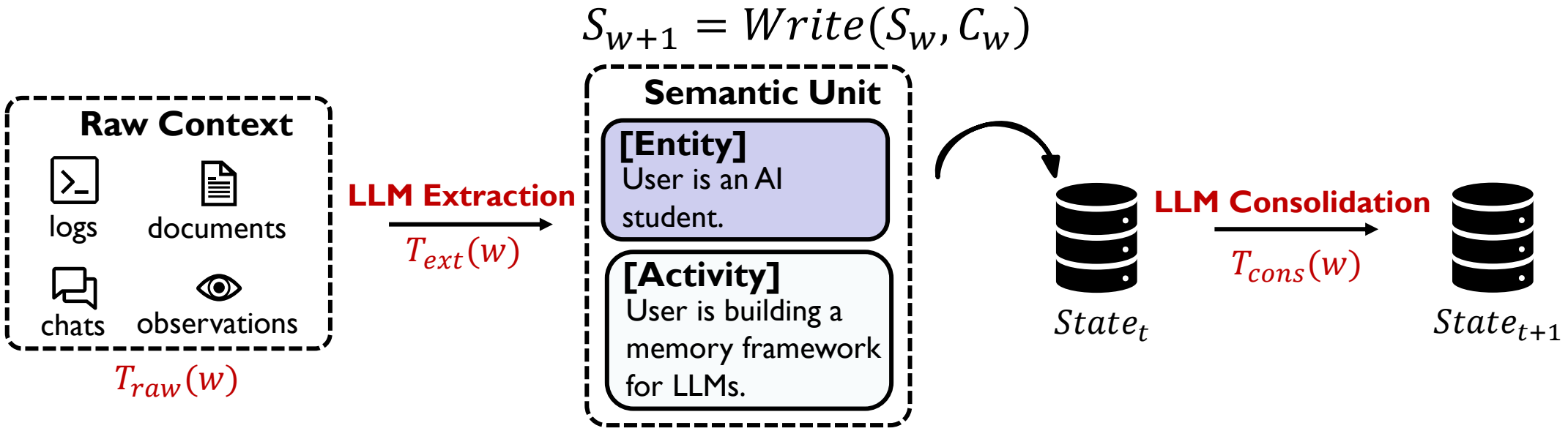
② LLM tokens far exceed raw input tokens

- the total number of tokens consumed in extracting and merging this input

Need metrics to quantify token amplification / semantic overlap

System Model

- We model an LLM-Mediated State Store as:



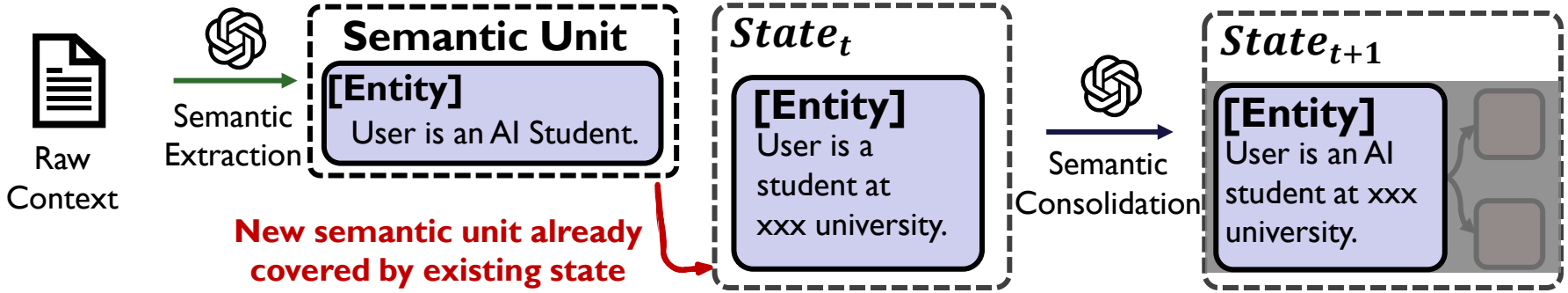
Semantic units are the **smallest self-contained pieces** of application semantics preserved by a state store.

$$U(w) = Extract(C_w) = \{u_1, u_2, \ldots, u_n\}$$

$$S_{w+1} = Consolidate(S_w, U(w))$$

Metrics Framework

- Token Write Amplification
  - LLM tokens consumed per raw-context token written: $TWA(w) = \frac{T_{ext}(w) + T_{cons}(w)}{T_{raw}(w)}$
- Token Space Amplification
  - Stored-state tokens per token of maintained semantics: $TSA(w) = \frac{T_{stored}(t)}{T_{sem}(t)}$
- Semantic Overlap
  - Incoming writes may **repeat** semantics already captured in the maintained state.

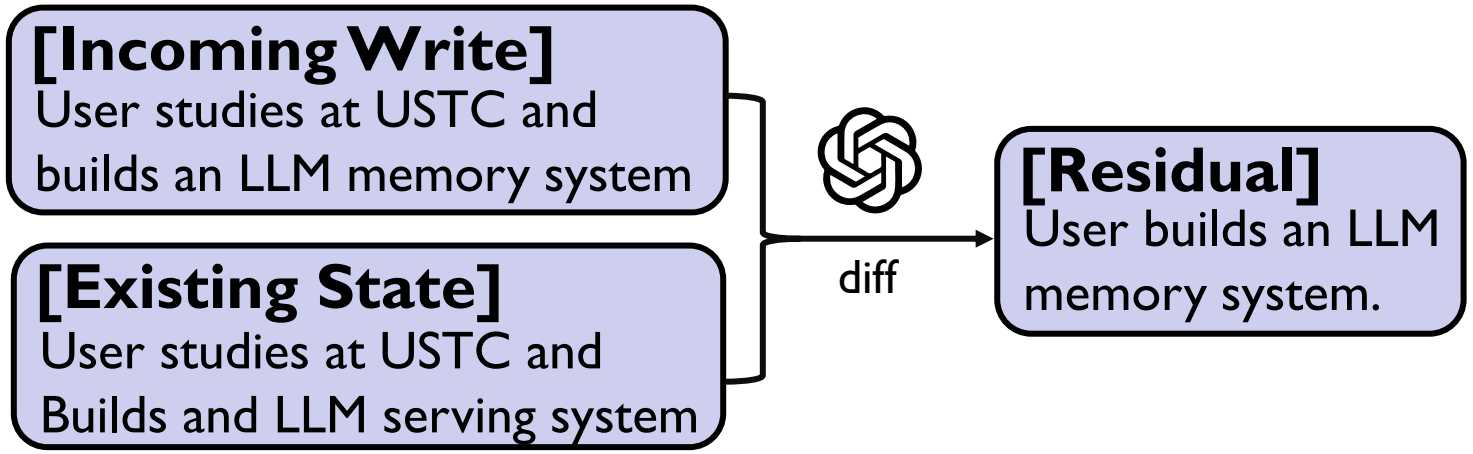


  - **Input Reprocessing:** Re-extracting covered semantics
  - **State Revisiting:** Re-consolidating existing state

Metrics Framework

➢ Semantic Overlap Ratio

- **Residual tokens:** Tokens for semantics absent from the existing state.



We use an LLM to extract the **residual** semantics not covered by the existing state.

- SOR(w): Fraction of incoming semantics already covered by S_w

$$SOR(w) = 1 - \frac{\sum_{u \in U(w)} tokens\big(residual(u, S_w)\big)}{\sum_{u \in U(w)} tokens(u)}$$

Validation of SOR

➢ Can SOR reliably distinguish known overlap levels?

- We validate SOR using controlled cases with known overlap levels.

[Fact A]
Base: User prefers concise answers.
Paraphrase: User wants brief answers.

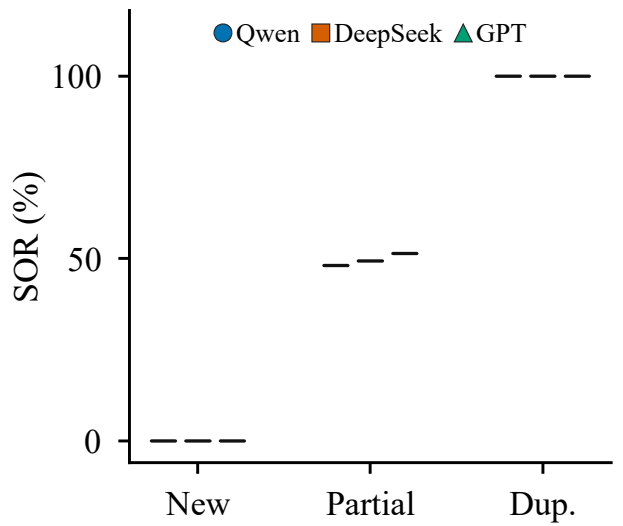
[Fact B]
Base: user rejects seafood restaurant recommendations
Paraphrase: User does not want seafood restaurants.

- We use Fact A as the existing state and construct three overlap case.

[New]
Incoming Write: Fact B

[Partial Overlap]
Incoming Write:
Paraphrase(A) + Fact B

[Duplicate]
Incoming Write:
Paraphrase(A)



Attributing Write-Path Token Cost

Extraction processes the full incoming context, so its cost is attributed in proportion to SOR.

$$TWA_{ext}^{overlap}(w) = SOR(w) \times \frac{T_{ext}(w)}{T_{raw}(w)}$$

Consolidation cost is attributed across selected existing state, newly extracted semantic units, and output tokens.

$$T_{cons}(w) = T_{cons}^{sel}(w) + T_{cons}^{new}(w) + T_{cons}^{out}(w)$$

Overlap-attributed consolidation input:

$$A_{in}(w) = T_{cons}^{sel}(w) + SOR \times T_{cons}^{new}(w)$$

Overlap-attributed Token cost:

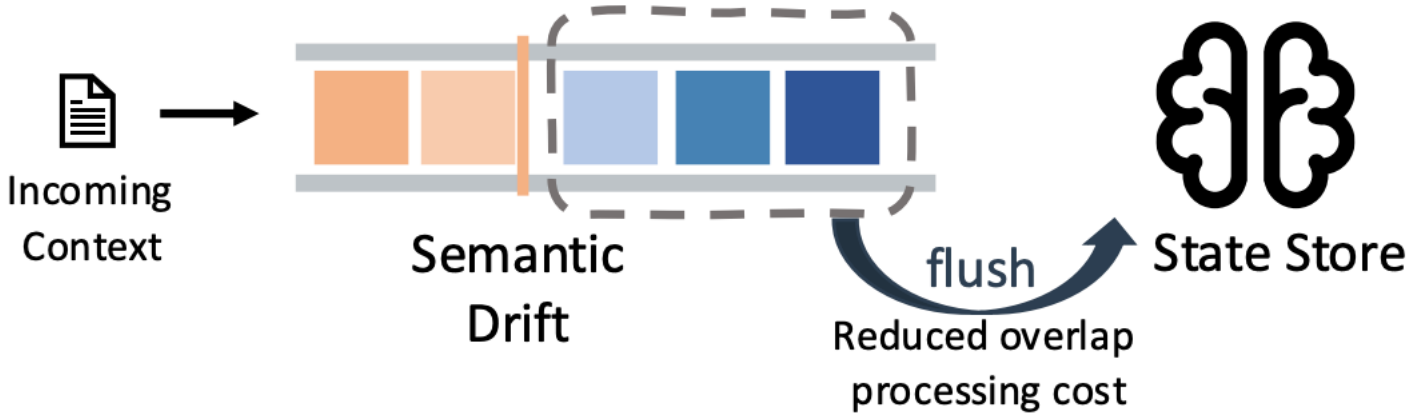
$$T_{cons}^{overlap}(w) = A_{in}(w) + \frac{A_{in}(w)}{T_{cons}^{sel}(w) + T_{cons}^{news}(w)} \times T_{cons}^{out}(w)$$

Overlap-attributed TWA:

$$TWA_{overlap}(w) = TWA_{ext}^{overlap}(w) + \frac{T_{cons}^{overlap}(w)}{T_{raw}(w)}$$

Semantic Coalescing Buffer

- Obvervation: Nearby writes often share semantics
  - we coalesce them into one state-store update



For each incoming write w_i, SCB computes its embedding e_i and measures semantic continuity against the last k buffered writes.

$$s_i = \max_{j \in tail_k(B)} \cos(e_i, e_j)$$

To adapt to workload-specific similarity distributions, SCB normalizes the continuity score using the most recent W scores.

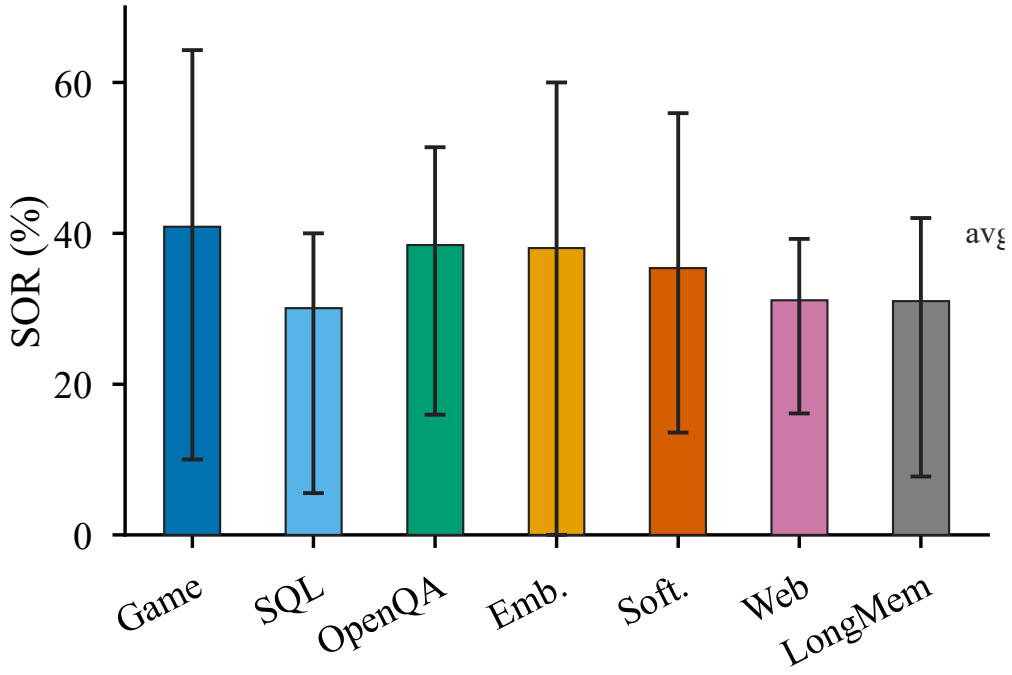
$$z = \frac{\mu - s}{\sigma}$$

Experiments

- Workloads:
  - Game, Text2SQL, Openworld-QA, Embodied AI, Software, Web, LongMemEval
- Evaluated State Stores:
  - **Agent Memory:** SimpleMem(ICLR' 26) , A-Mem(NIPS' 25), MemoryOS(EMNLP' 25)
  - **Graph-Based RAG:** LightRAG(EMNLP' 25), with LR-3/LR-8 denoting more/less frequent LLM summary-on-merge
  - All configurations are evaluated both without SCB (Native) and with SCB.
- Implementation
  - **Model Selection:** Qwen3-8B for state maintenance and Qwen3-Embedding-0.6B for SCB similarity
  - **Software & Hardware Stack:** vLLM on NVIDIA A40 GPUs

Performance Evaluation

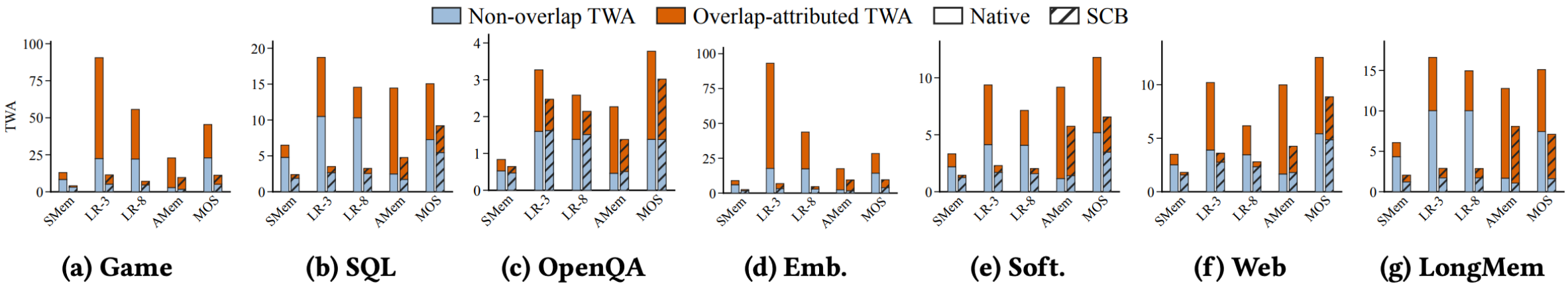
➢ Semantic Overlap Across Real-World Workloads



Conclusion: Semantic overlap is common and highly variable across real agent workloads, with an **average SOR of 36.8%**

Performance Evaluation

➢ TWA Across Workloads

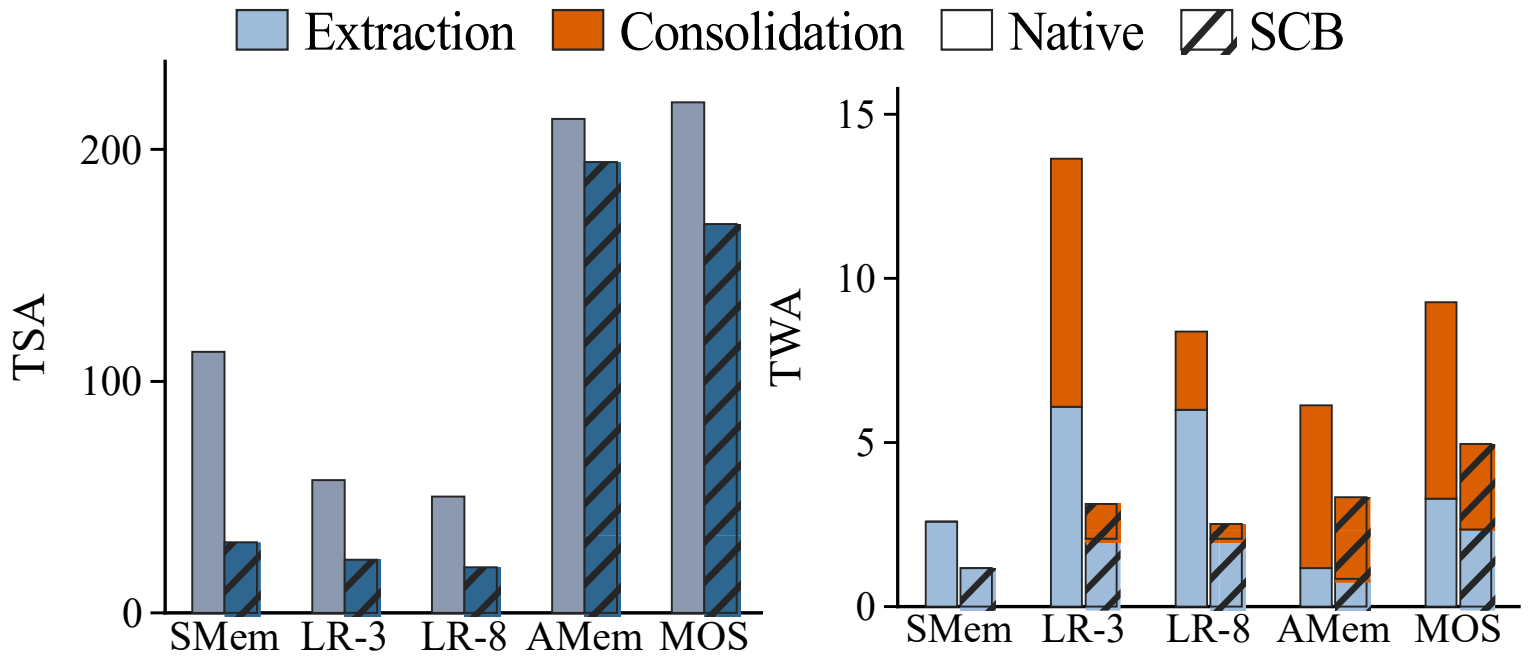


Conclusion

Overlap is a major source of TWA: 54.9% comes from already-covered semantics.
TWA is workload-dependent: amplification varies with workload structure.
SCB reduces both components: overlap-attributed and non-overlap TWA.

Performance Evaluation

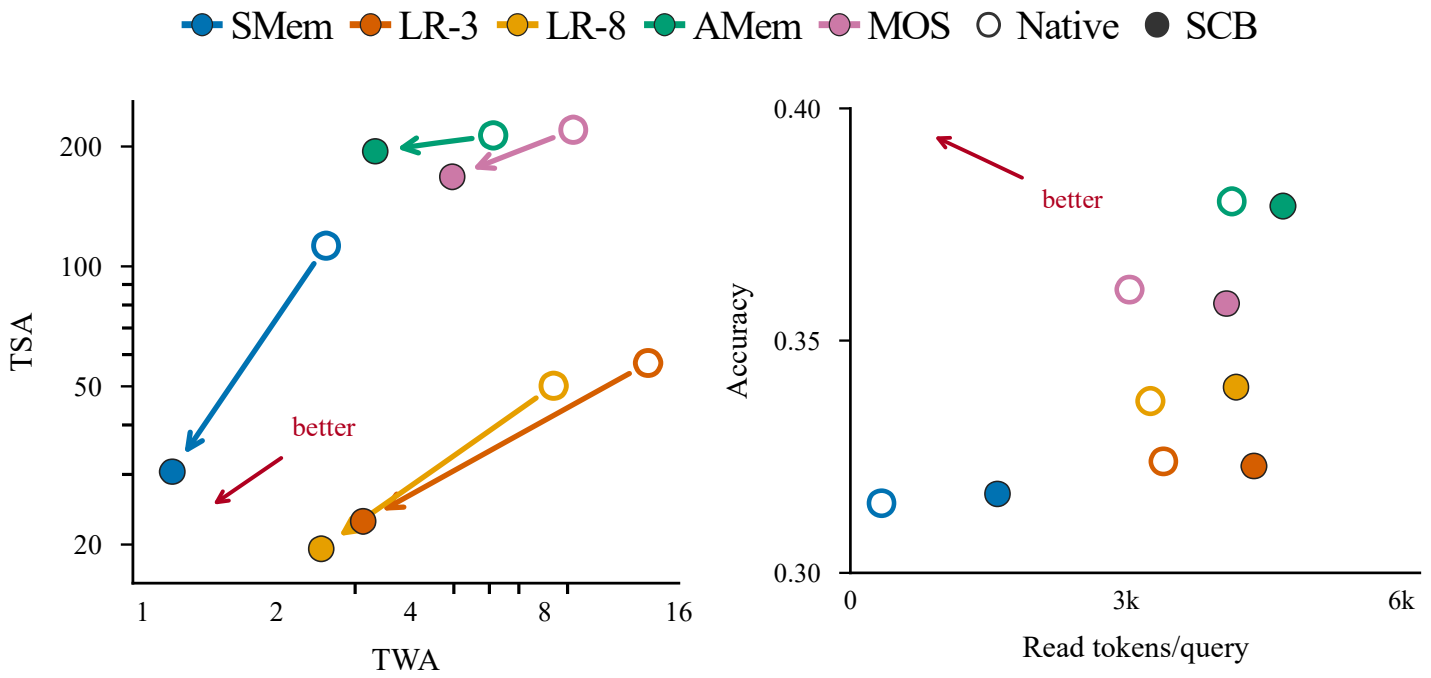
➢ State Representation Shapes Token Amplification



Conclusion: Consolidation dominates TWA, richer representations increase TSA; more consolidation does not yield better compaction.

Performance Evaluation

➢ Cost-Path Trade-off Analysis



Conclusion: SCB reduces both TWA and TSA without compromising accuracy, while write savings outweigh its modest read overhead.

Conclusions

- **A New Class of State Stores**
  - Identify LLM-mediated state stores, where LLMs directly participate in semantic state maintenance.
- **New Metrics and Accounting**
  - Introduce TWA, TSA, and semantic-unit accounting to quantify token amplification and semantic overlap.
- More evaluation results and analysis are in the paper

Thanks for your attention!

Q&A

Contact email:
yangx7@mail.ustc.edu.cn