Heterogeneous Tabular Data Query and Analysis System Based on Large Language Models

Jiajun Li, Xiaojun Shi, Angeline Mary Marchella, Tianze HU, Lianpeng Qiao, Kaiyu Li, Guoren WANG

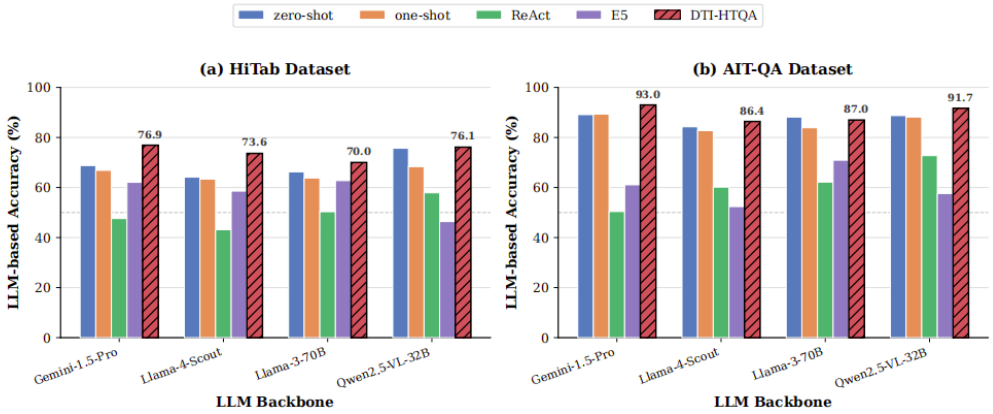
Frontiers of Computer Science, DOI: 10.1007/s11704-026-60020-2

Problems & Ideas

- Problems of conventional stereo matching approaches:
  - Existing methods typically convert complex tables into linear text or simplified two-dimensional structures, which loses vital semantic information conveyed by spatial layouts like merged cells and multi-level headers
  - Current approaches heavily depend on extensive manual preprocessing and header annotation, incurring high labor costs and limiting real-world applicability.
  - Processing lengthy and complex tables globally significantly increases token consumption and can distract the model with irrelevant information, reducing overall efficiency and accuracy.
- Ideas: An end-to-end framework (DTI-HTQA) that utilizes a graph-based representation to preserve structural semantics, an LLM-driven header recognition algorithm to eliminate manual annotations, and dual-tree indices equipped with four core operators for precise and cost-efficient cell retrieval.

Main Contributions

- Contributions:
  - A structure-aware graph representation method to model heterogeneous tables, effectively preserving rich structural semantics like merged cells and multi-level headers.
  - An LLM-driven header recognition algorithm using packing and multi-round voting strategies to eliminate manual annotations, enabling a fully end-to-end QA process.
  - An adaptive DTI-HTQA framework that integrates dual-tree indexing, tree navigation, and chain-of-thought reasoning to minimize token consumption while enhancing accuracy for complex queries.



LLM-based evaluation accuracy comparison across different LLM backbones on HiTab and AIT-QA datasets. DTI-HTQA demonstrates superior semantic matching performance.