Can LLMs Reason Structurally? Benchmarking via the Lens of Data Structures

Yu He*,1, Yingxi Li*,1, Colin White2, Ellen Vitercik1
1Stanford University 2Abacus.AI *Equal contribution

ICML International Conference On Machine Learning

Structural reasoning ability of LLMs

"Can LLMs reason over queues, trees, graphs, etc.?"

- Structural reasoning: to understand and reason about data relationships.
- Core to tasks involving complex mathematical and algorithmic reasoning.

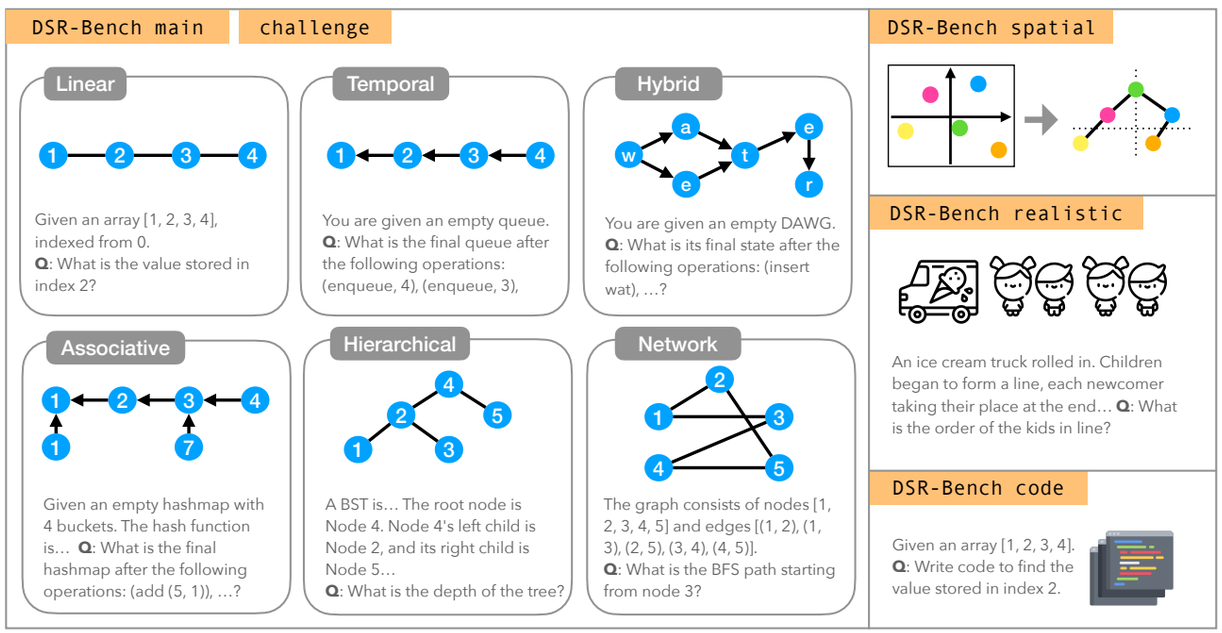
However, existing benchmarks primarily focus on high-level, application-driven evaluations without isolating this fundamental capability.

DSR-Bench

- Six categories, 20 data structures, 35 operations, 4,140 problem instances.
- Three length types (short, medium, long).
- Four suites: main, challenge (difficult tasks), realistic (natural language descriptions), code.

Each task probes whether the model can understand, manipulate, and maintain a data structure.

TLDR: We propose a novel benchmark using data structures and their operations to assess LLMs' structural reasoning abilities in a scalable, interpretable, and automated way with fine-grained analysis.



Design of DSR-Bench

Prompt Design (i) description of the data structure; (ii) explanation of operations performed; (iii) the initial state of the data structure; (iv) question for final outcome.

Example prompt for QUEUE compound.

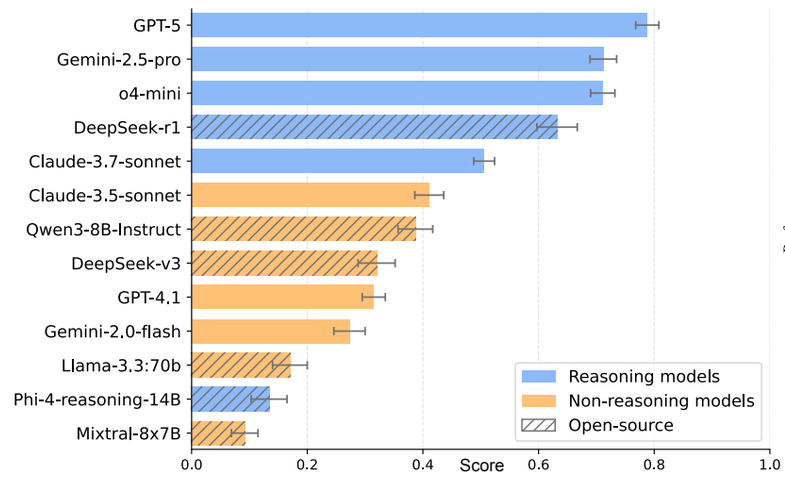
(i) A queue maintains a first-in, first-out (FIFO) order, where items are added at one end and removed from the other.
(ii) There are two operations: `(enqueue, k)` adds k to the back; `(dequeue)` removes the front.
(iii) You are given an empty queue initially.
(iv) **Q:** What is the final queue after: `(enqueue, 49), (enqueue, 85), (dequeue),...`

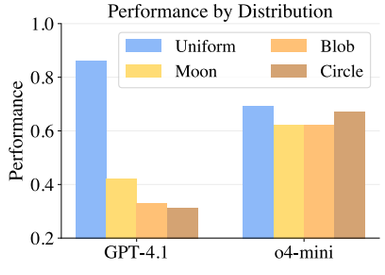
Why DSR-Bench?

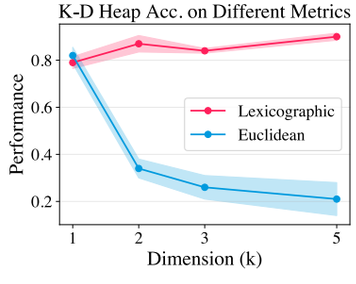
- Hierarchical task organization to pinpoint bottlenecks
- Deterministic evaluation with unambiguous outputs
- Synthetic, low-contamination data generation to ensure scalability.

Highlight of results

- **Instruction-tuned models struggle with multi-attribute and multi-hop reasoning.**
  - Fail drastically on tasks with multiple attributes (e.g., hashmaps) and multi-hop reasoning (e.g., red-black trees).
  - Chain-of-Thought (CoT) helps only on non-standard structures.
- **Reasoning models still have major limitations with complex structures.**
  - Score only up to **47%** on complex structures in DSR-Bench-challenge.
  - Often rely on learned priors (e.g., misinterpret depth in trees), failing to follow explicit instructions.

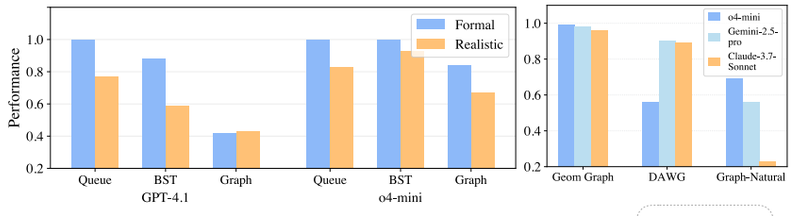


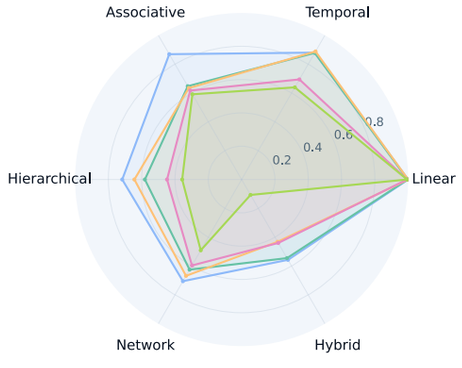




- **Performance drops on complex spatial data structures.**
  - Accuracy declines as dimensionality increases.
  - Further degrades on non-uniform inputs, reveal reliance on memory.
- **Natural language descriptions degrade performance.**
  - Formal to narrative descriptions leads to a significant drop in accuracy.
  - Suggests poor generalization to real-world, language-rich scenarios.

- **Models cannot reason over generated code**
  - Ability to write code ≠ inherent ability to reason about it
  - Code helps formal/standard tasks, fails on realistic ones





Dataset:

Code:

Paper: