DSR-Bench: Evaluating the Structural Reasoning Abilities of LLMs via Data Structures

Yu He*,[1], Yingxi Li*,[1], Colin White[2], Ellen Vitercik[1]

[1]Stanford University [2]Abacus.AI *Equal contribution

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.

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).
- Three suites: main, challenge (difficult tasks), natural (natural language descriptions).

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

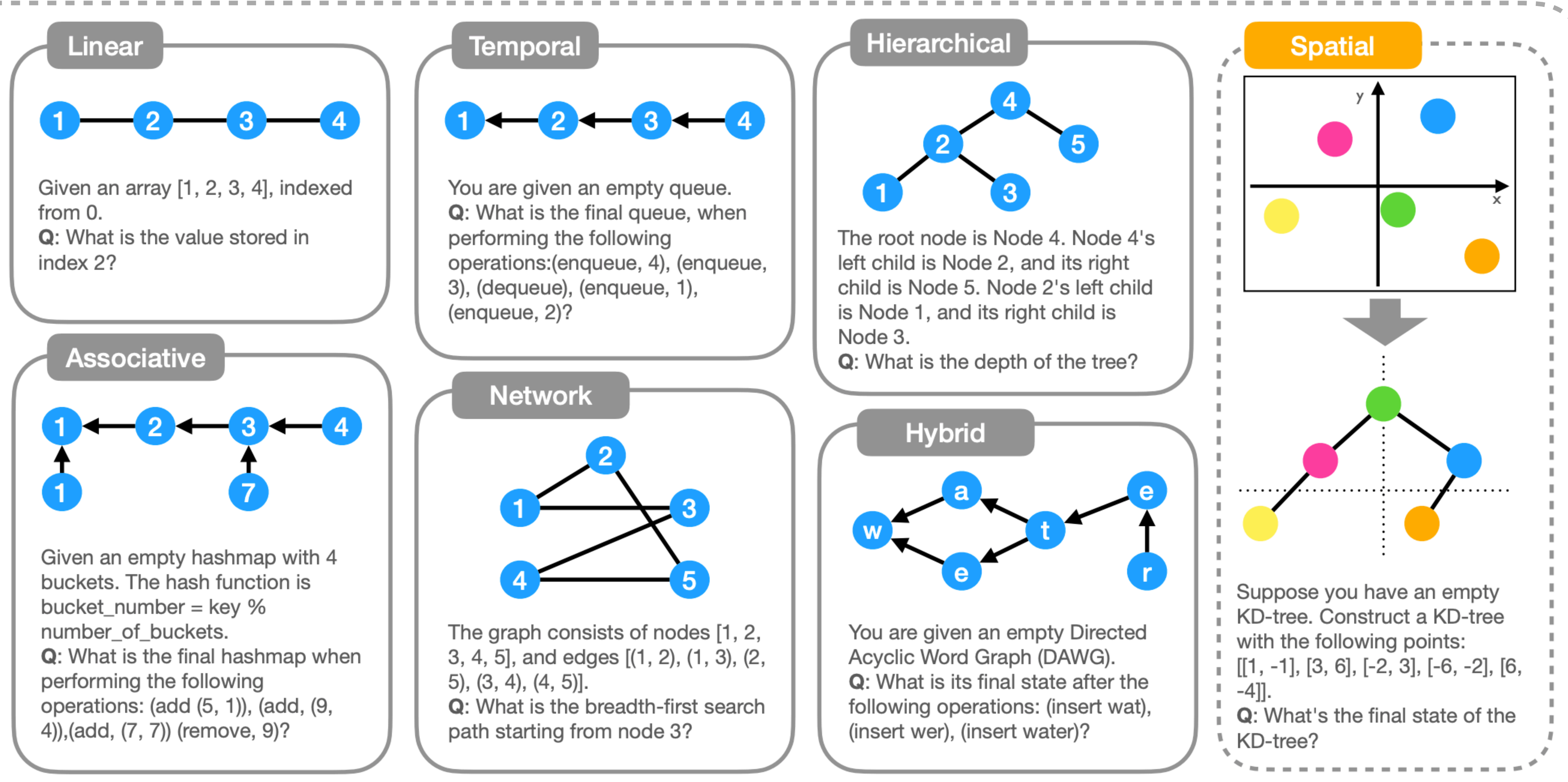
Design of DSR-Bench

Example prompt for QUEUE compound.

A queue is a data structure in which items are added at one end and removed from the other, maintaining a first-in, first-out (FIFO) order. You should create a queue. There are two types of operations: `(enqueue, k)` adds `k` to the back. `(dequeue)` removes the front. You are given an empty queue initially.
Q: What is the final queue after performing:

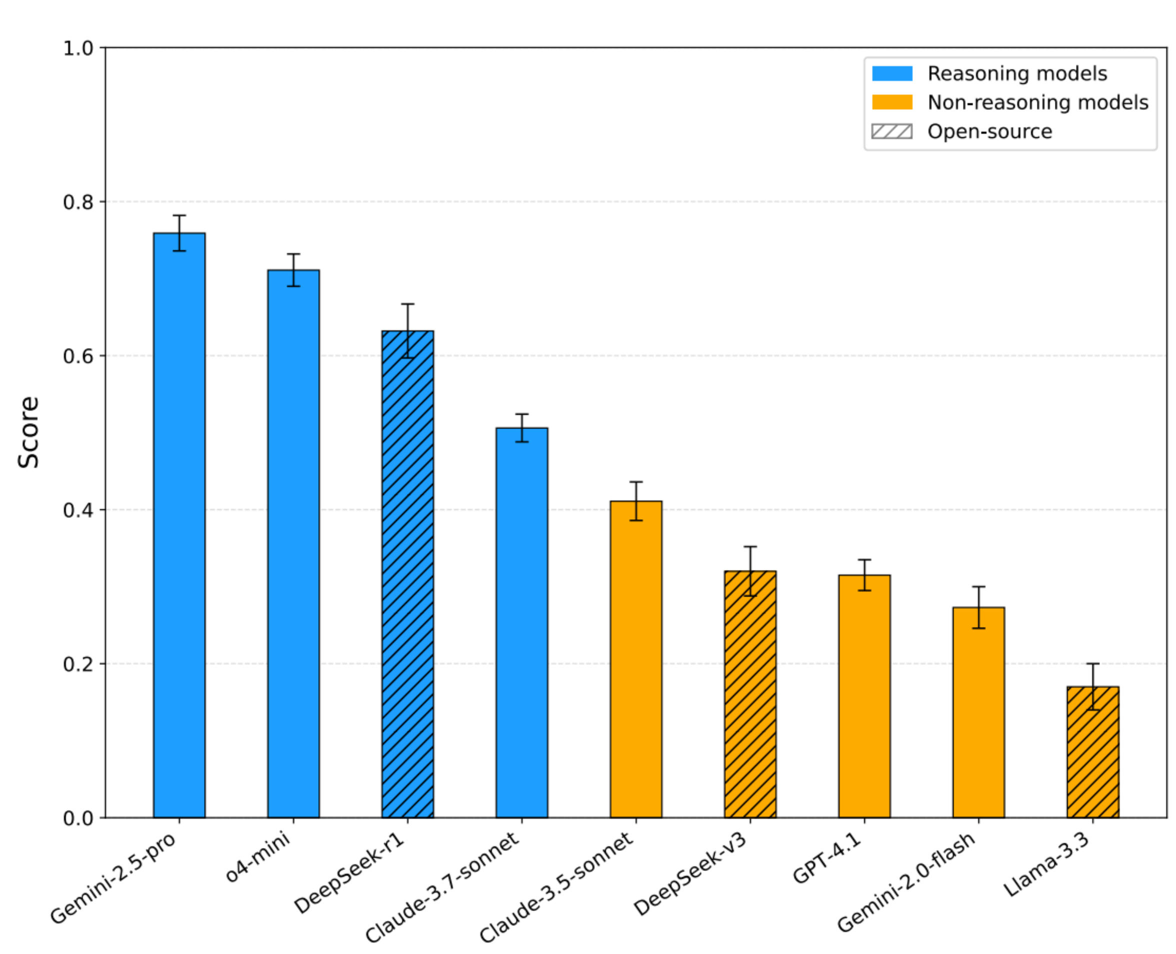
- (enqueue, 49)
- (dequeue)
- ...

Answer the question in 8000 tokens.

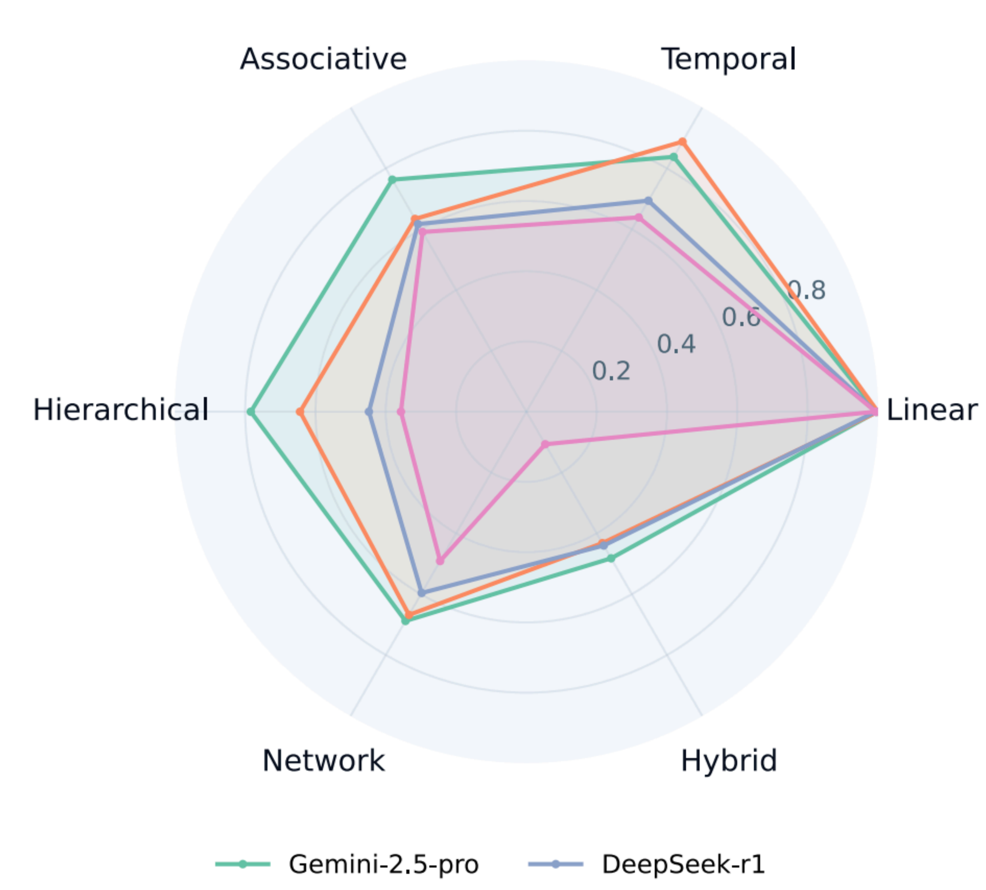


Why DSR-Bench? (i) Hierarchical task organization to pinpoint bottlenecks, (ii) deterministic evaluation with unambiguous outputs, and (iii) synthetic, low-contamination data generation to ensure scalability.

Highlights 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.
  - Accuracy further degrades on non-uniform inputs, revealing reliance on memorization.
- **Natural language description degrades performance.**
  - Translating tasks from formal to narrative descriptions leads to a significant drop in accuracy.
  - Suggests poor generalization to real-world, language-rich scenarios.



Paper:

Code:

Dataset: