Randomly Sampled Language Reasoning Problems Reveal Limits of LLMs

Kavi Gupta, Kate Sanders, Armando Solar-Lezama

Read the paper on arXiv

Have LLMs Learned a Model of Language?

- Models can learn specific languages. Have they learned the ability to pick up a language from examples?
- Given e.g., ABABBABBBABBBBABBB…, an LLM will fill in the rest
- This ability shows up in more useful contexts:
  - Being able to write in novel DSLs
  - More abstractly, most LLM tasks can be viewed as in-context-learning a language

Dataset Leakage

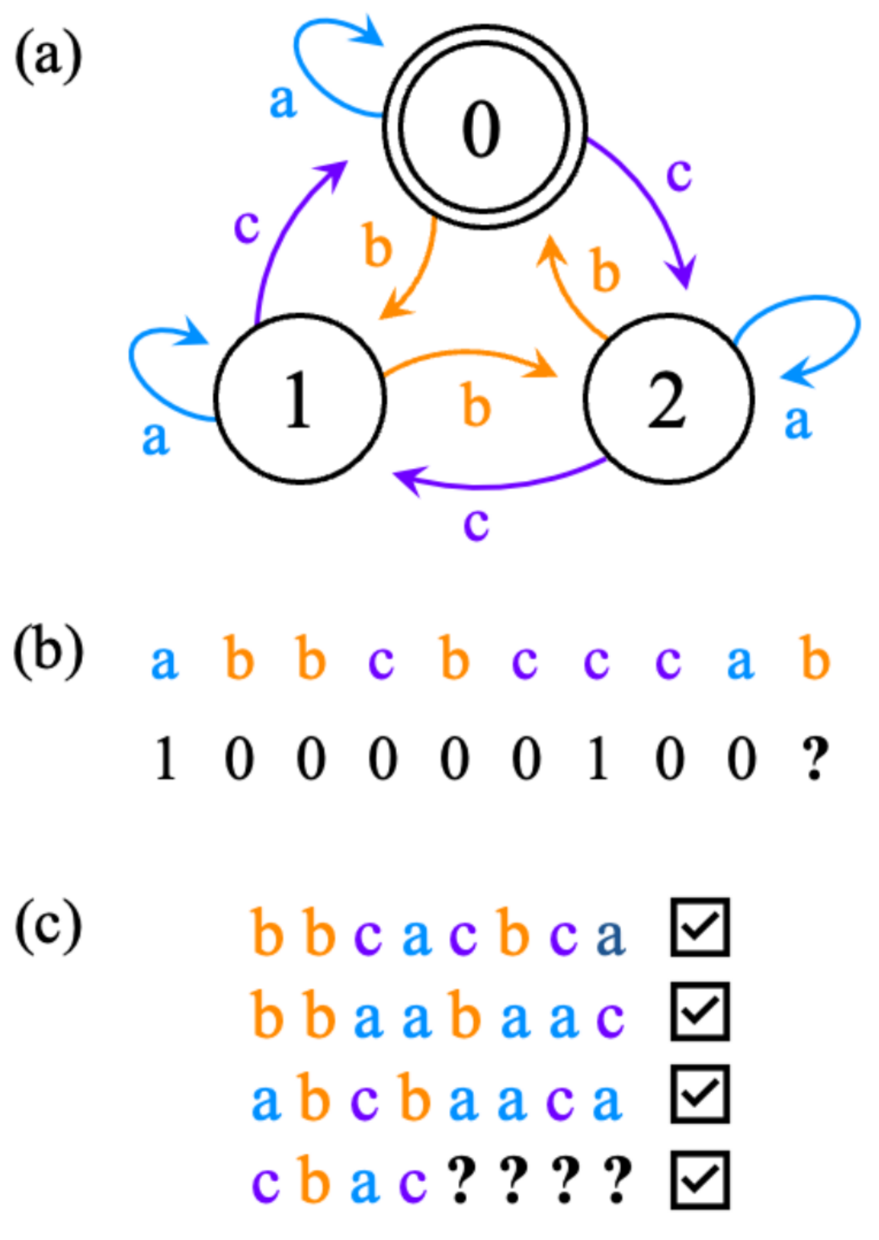
- No way to prove lack of overlap between training data and benchmarks
- Training data is often closed
- Determining whether two tasks are the same is nontrivial

Models

- Several open source sequence and code completion models (e.g., llama, mistral, qwen)
- Several proprietary chat models (e.g, gpt-4o, claude)
- One reasoning model (o3-mini)

Domain

Sample a
(a) latent 3 state DFA
then either construct a
(b) Transducer task (string of length 30; given output of all prefixes, produce final output)
(c) Sequence Completion task (30 accepted strings of length 10, produce completion of prefix of length 5)



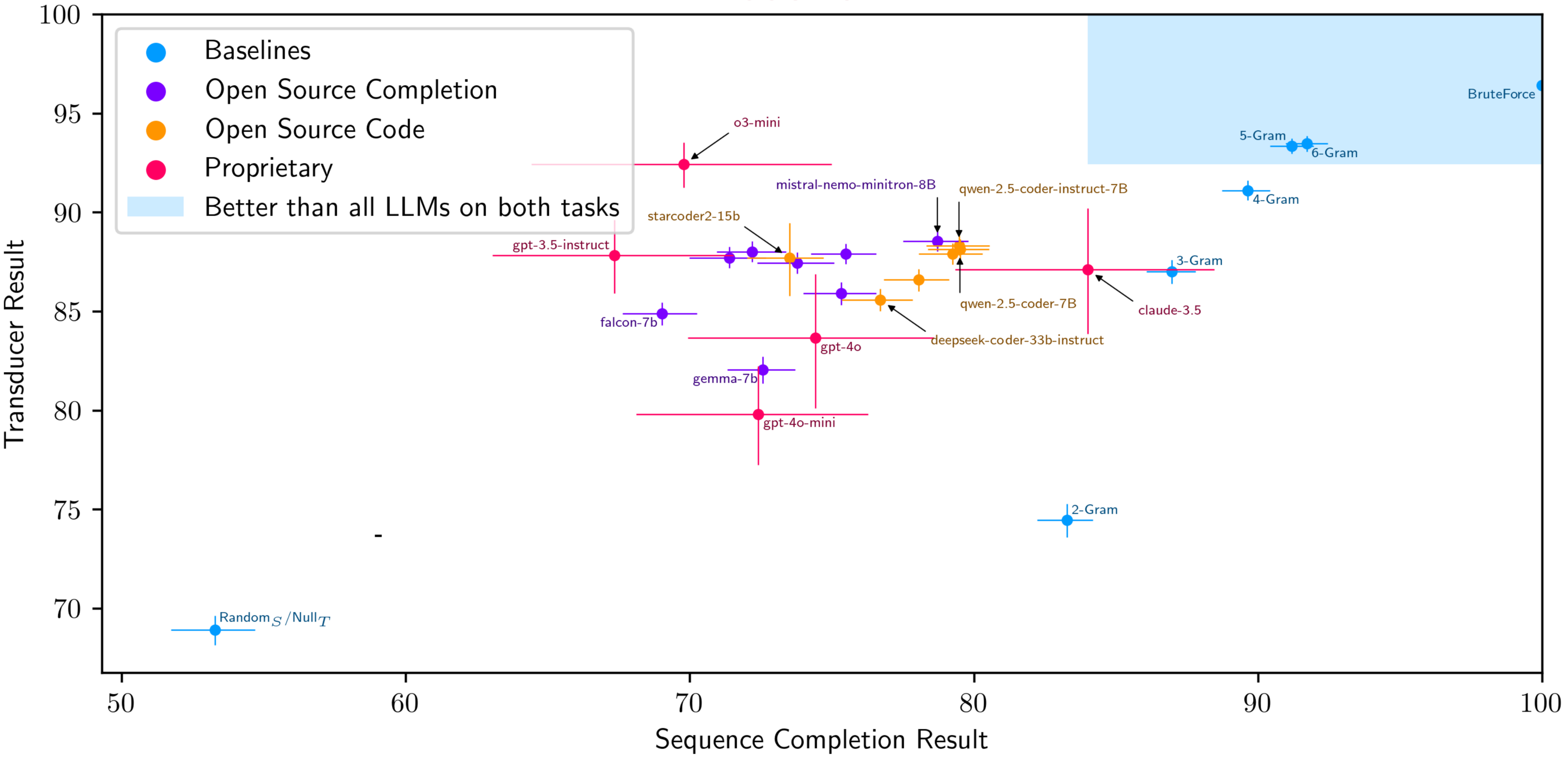
Baselines

- Brute Force: try every DFA, filter for consistency, take majority vote. Not realistic comparison, measures upper bound of performance
- n-Gram: take the given sequence, build a table of (n-1) length sequences to the next symbol, predict next symbol

Prompts

- Two next token prediction prompts (one with no context, one that says it is a language produced by a 3 state DFA)
- Two think step-by-step prompts (one in the form of a word problem)
- Report results on best prompt

Results



LLMs consistently underperform a basic 5gram model on both tasks!

- LLM ICL's language understanding is limited
- Does not work on unseen languages, even simple ones
- In some cases works better with NTP than with COT

- Other explanations do not apply
  - Autoregressive task: no embers of autoregression problem
  - World modeling: the LLM underperforms even ngrams.