A Systematic Comparison of Test-time Task Learning from Rules vs. Examples

Xiang Fu, Seungmin Cho, Najoung Kim
Boston University, Department of Linguistics, tinlab

Boston University Office of the Provost
Undergraduate Research Opportunities Program

Abstract

Large language models (LLMs) can learn new tasks at test time either from natural language descriptions of rules ("instruction following") or from examples of input-output pairs ("in-context learning"). Under matched token budgets across tasks spanning algorithmic decision and linguistic transformations, we assess accuracy and sample efficiency. Preliminary results indicate that sufficiently large instruction-tuned models learn more effectively from rules, whereas base models learn more effectively from examples. Both modes of learning interact with model scale. Example-based learning is less effective for tasks that depend on more rules, and increasing the number of examples has diminishing returns beyond certain thresholds.

Task Design

Linguistic

Counterfactual Word Order

Counterfactual Question Formation – Yes/No & WH

Algorithmic

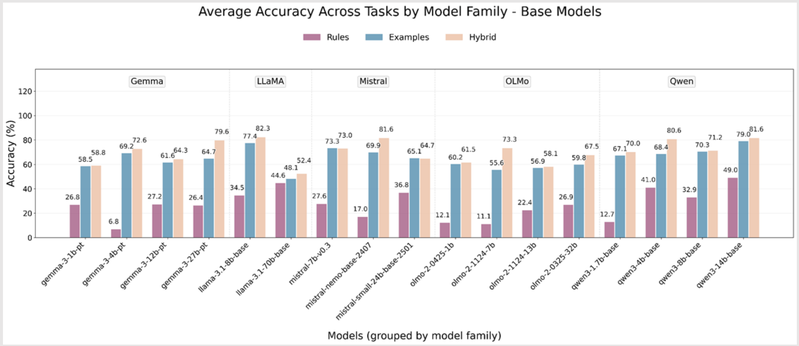
Color Pattern

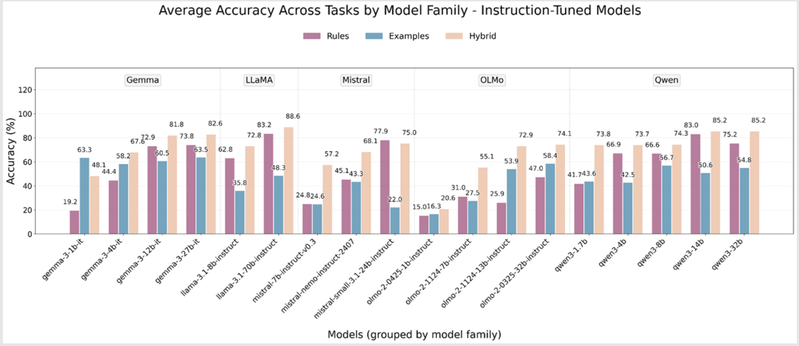
Word Chain

Path Navigation

Dice Game

Performance



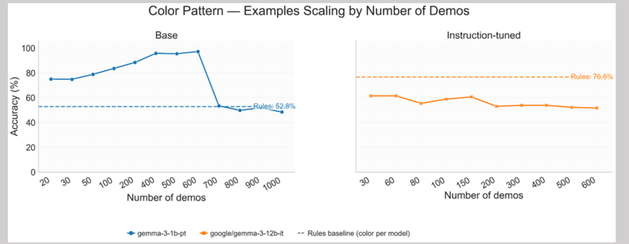


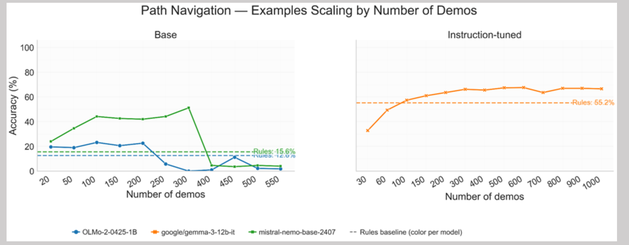
Instruction-tuned models (with sufficient size) learn more effectively from rules, whereas base models learn more effectively from examples, although we see variations across model families and sizes. Hybrid learning helps both models.

Model scale generally improves rule-based learning, but only consistently helps example-based learning in base models.

Scaling of # Examples

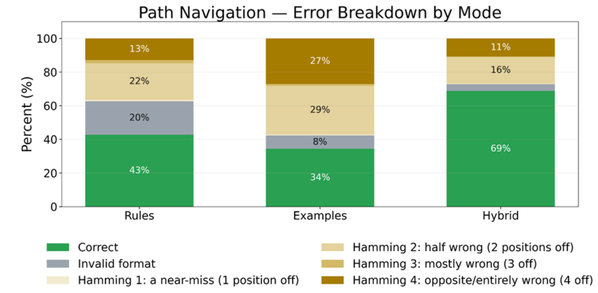
Increasing the # of examples helps some tasks, but not all. In base models, too many examples can even lead to a large drop.





Error Analysis (Path Navigation task)

Hybrid learning reduces both severe and moderate errors compared to both Rules- and Examples-only learning.

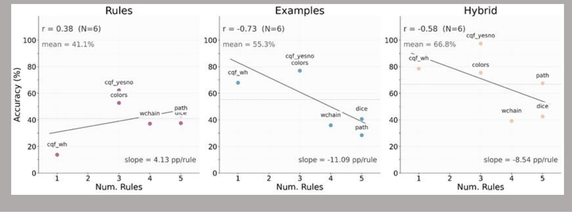


Conclusion

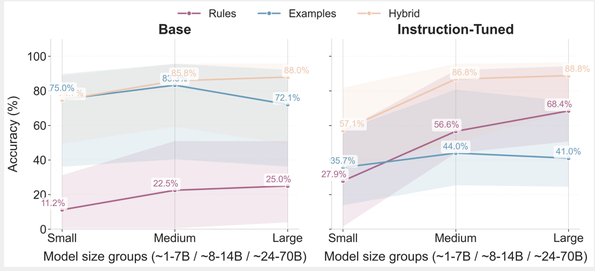
- Instruction-tuned models learn better from rules, base models learn better from examples, and both models do best with both kinds of information (hybrid learning).
- Model scaling helps rule-based learning in both base and instruction-tuned models, and helps example-based learning in base models (although the effect varies depending on model family).
- Scaling the # of examples does sometimes help in example-based learning, but in base models, they show diminishing returns after a certain point.
- Tasks with more rules negatively affects performance in example-based and hybrid learning, but not in rule-based learning.

Task Complexity

Tasks described by more rules are more challenging for example-based and hybrid learning, but not for rule-based learning.



Effect of Model Scaling



Both base and instruction-tuned models show better accuracy with scale in all learning with the exception of the largest instruction-tuned models in rules-mode. Hybrid learning is especially effective for large instruction-tuned models.