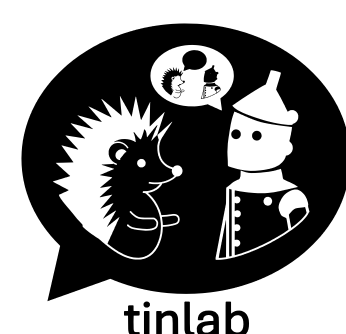


LLMs Learn Better In-Context from Rules than from Examples



Xiang Fu,* Seungmin Cho,* Yukyung Lee, Najoung Kim

Department of Linguistics, Boston University

*Equal contribution Contact: xfu@bu.edu, minjo@bu.edu



Motivation

Question: Do LLMs learn a new task in context more reliably from **rules** (instructions) or from **examples** (input–output demonstrations)?

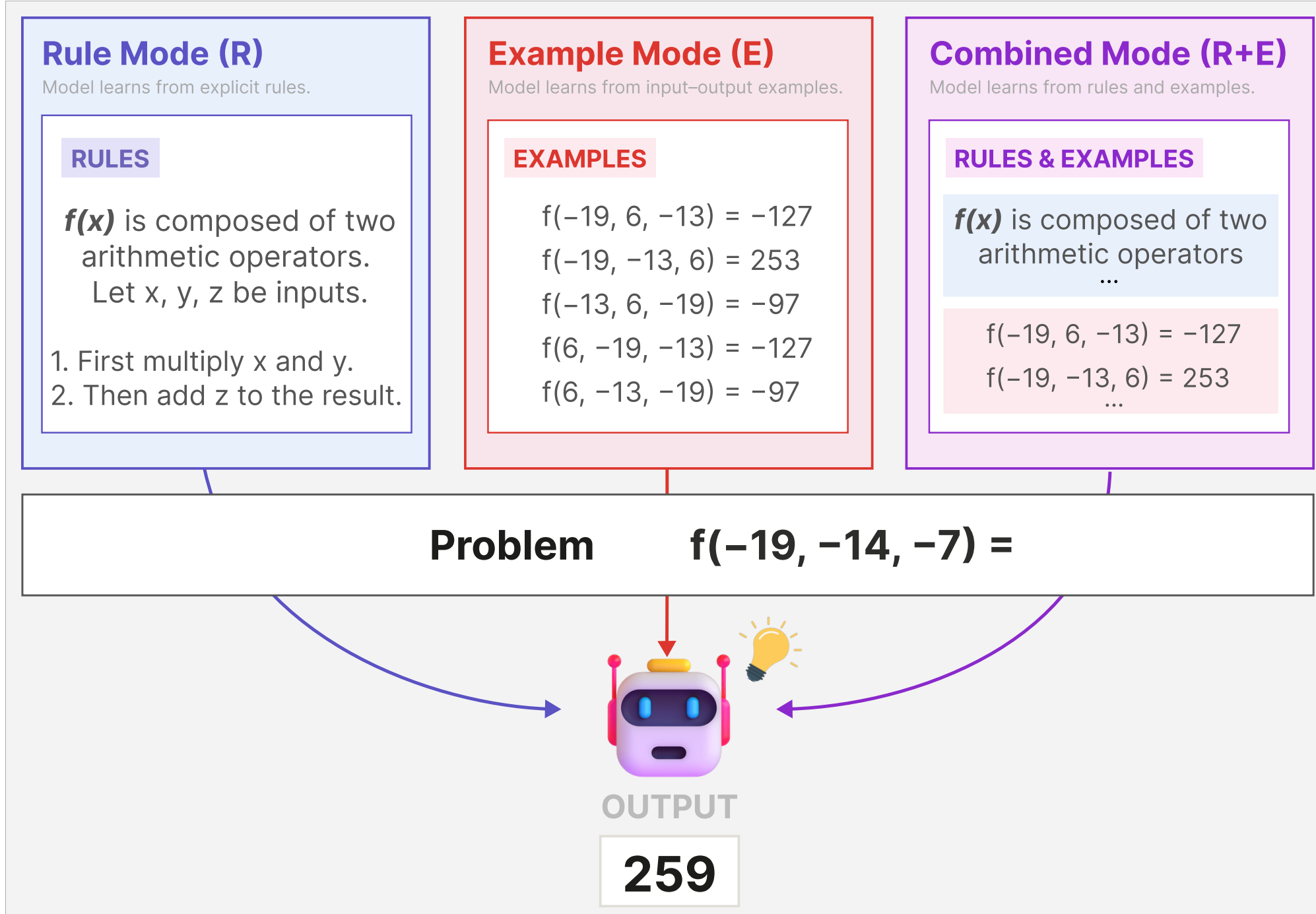
Rules and examples recruit partly distinct mechanisms (Davidson et al., 2025), often read as support for combining them. But **distinct mechanisms need not mean additive gains**. Prior work compared rules and examples primarily on simple tasks or had a secondary focus (e.g., Liu et al., 2024).

RQ1. Do models learn complex, novel tasks more reliably **from rules or from examples**?

RQ2. How does the learner, especially **instruction tuning**, affect this?

RQ3. Which **task properties** favor rules or examples?

Task Design



Rules: the task's rules stated in words

Examples: input–output pairs only

Combined: rules plus examples

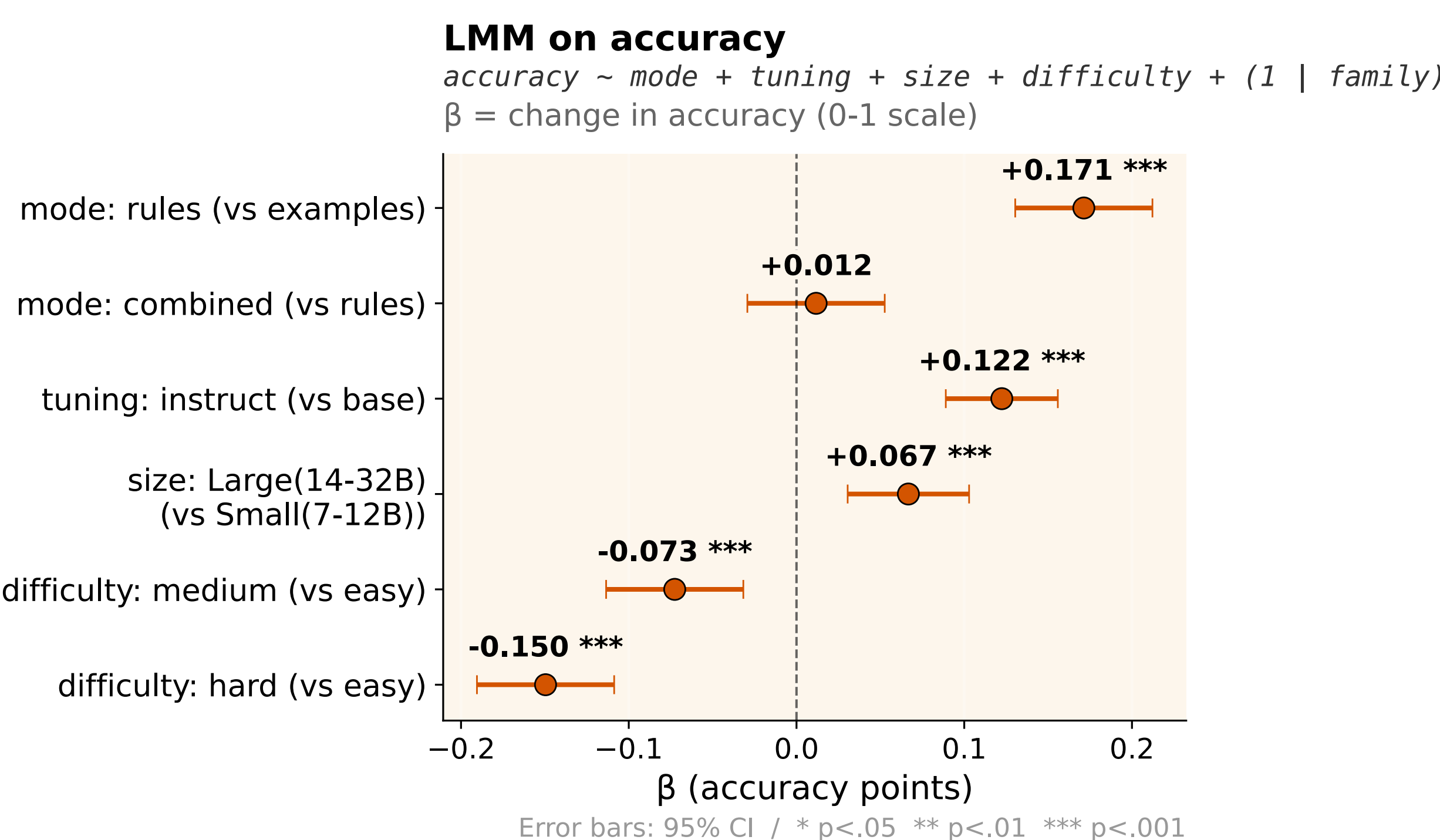
The model then solves a new problem

Models: 7 base and instruct pairs from Gemma 3, Qwen 2.5/3, and OLMo 3/3.1, 7–32B

Difficulty: harder levels add card attributes, board size, operators, noun classes, or category conditions

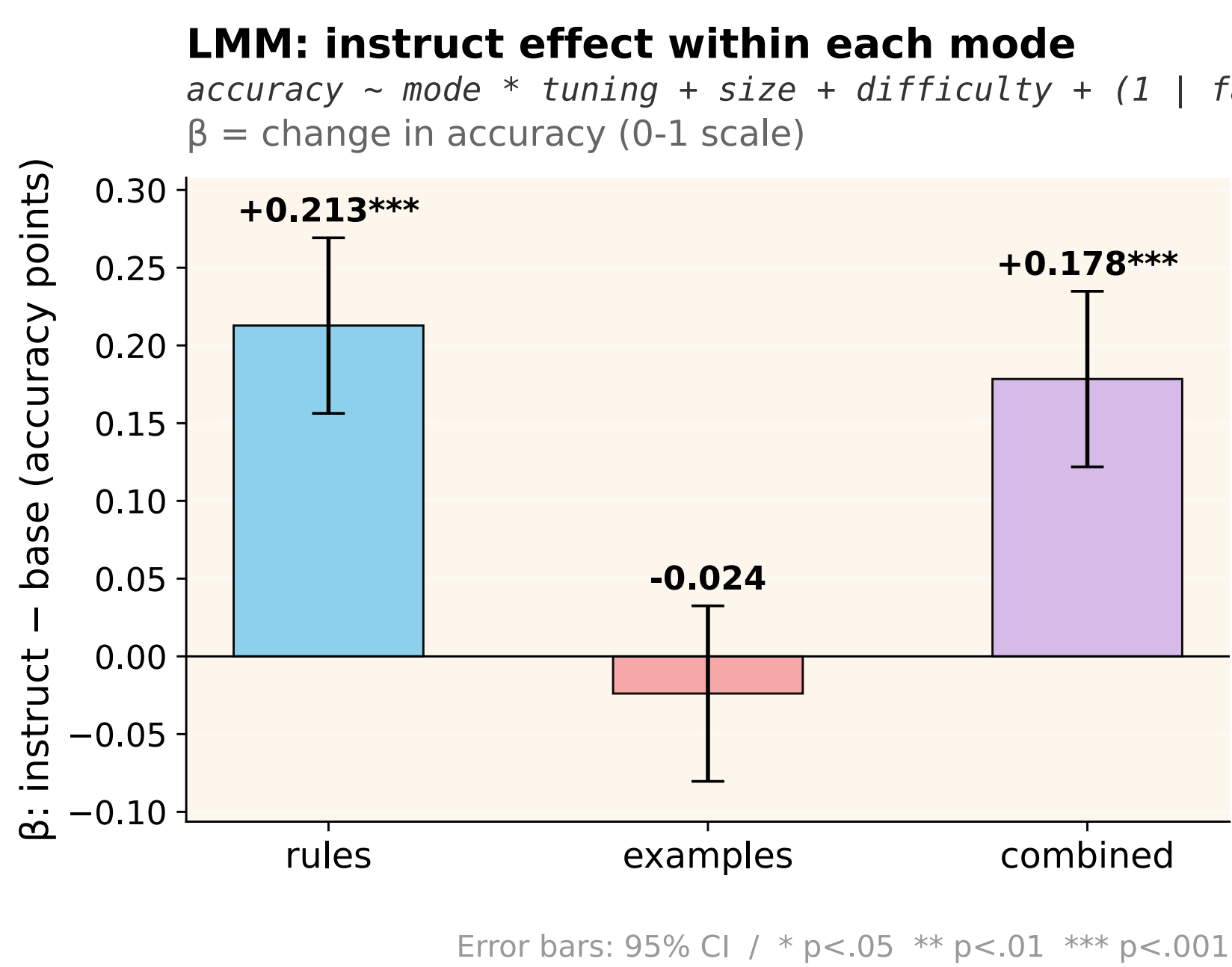
Main Results

RQ1: Rules > Examples



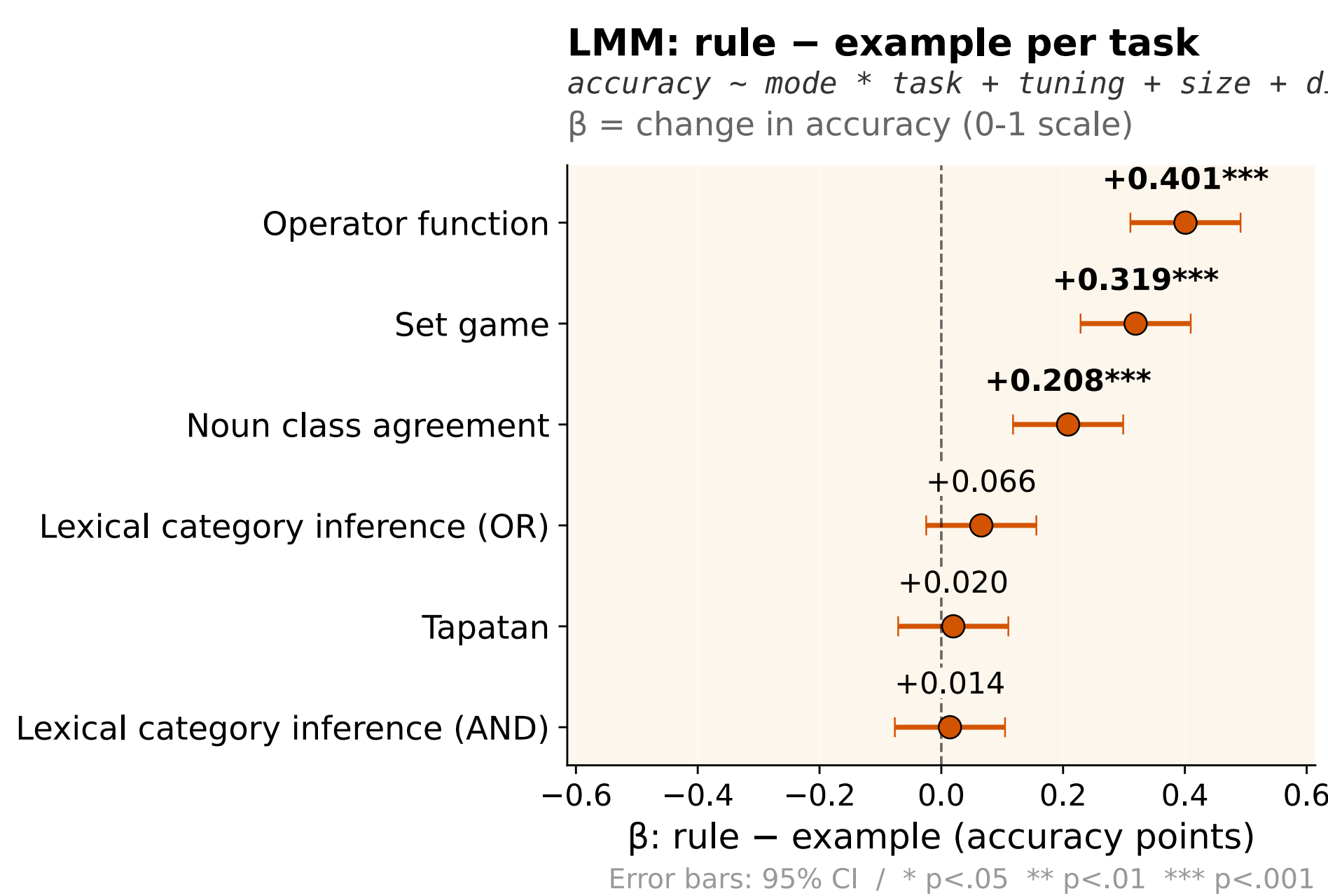
- **Rules** outperform examples overall, but the task effect is significant.
- **Combined $\approx$ rules:** adding examples yields no significant gain.
- **Instruction tuning** and **model size** improve accuracy, and **difficulty** hurts it.

RQ2: Effect of Instruction Tuning



- Instruction tuning improves **rule-based** and **combined** learning, but not example-based learning.
- Examples are not privileged in **base models** either: base models already benefit from rules on algebraic tasks.

RQ3: When do rules help most?

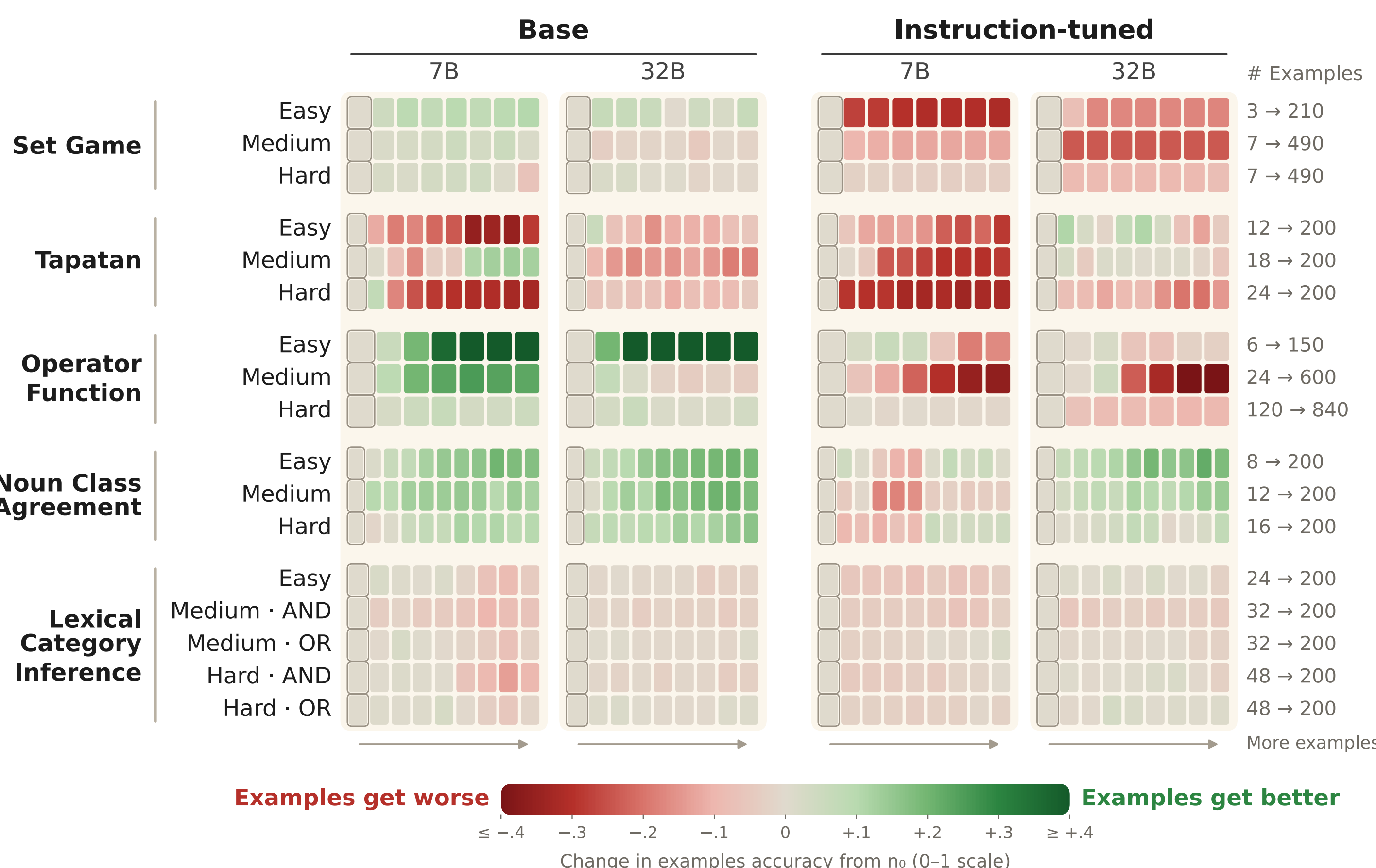


- Rules vs. examples: largest gains on tasks recruiting **algebraic abstraction**. Smaller or n.s. when the task rewards **distributional cues** or recruits **parametric knowledge**.

Do More Examples Help?

Example scaling from the minimum-coverage set

$\Delta = accuracy(examples, n) - accuracy(examples, n_0)$
OLMo 3/3.1 · 7B and 32B, base and instruction-tuned



More examples rarely help. Adding examples improves performance in 16 of 68 settings, leaves it flat in 31, and lowers it in 21. Gains are focused to **Noun Class Agreement** and **base models on Operator Function**, while Tapatan and instruction-tuned models on Set Game mostly get worse. More examples **never close the gap** where rules already had an advantage.

Conclusion

1. **Rules are more effective than examples** for conveying a new task. Adding examples to the rules does not consistently improve performance.
2. **The rule advantage is larger in instruction-tuned models, but base models show a similar trend.** Instruction tuning doesn't degrade learning from examples.
3. The gap between rule and example is larger when the task requires **algebraic abstraction and computation**. It's smaller when the task leverages **distributional cues** or recruits **parametric knowledge**.
4. Our rules were written to be clear and complete. Whether the advantage holds for vague or partial rules is open.

Acknowledgements

XF was supported by Boston University's Undergraduate Research Opportunities Program (UROP). We acknowledge that the computational work reported in this paper was performed on the Shared Computing Cluster, which is administered by Boston University's Research Computing Services.

More?



References

Davidson, G., Gureckis, T. M., Lake, B. M., & Williams, A. (2025). Do different prompting methods yield a common task representation in language models? NeurIPS.
Liu, E., Neubig, G., & Andreas, J. (2024). An incomplete loop: Instruction inference, instruction following, and in-context learning in language models. COLM.