



LLMs Learn Better In-Context from Rules than from Examples

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

Boston University *Equal contribution

{xfu, minjo, ylee5, najoung}@bu.edu

Background

A Long-Standing Question

People can learn a new concept from an explicit rule or from examples.

Cognitive science has long studied **rule-based** versus **exemplar-based** learning and generalization, especially for categorization (Nosofsky, 1986; Smith and Sloman, 1994; Gentner and Medina, 1998; Tenenbaum, 1999; Dasgupta et al., 2022).



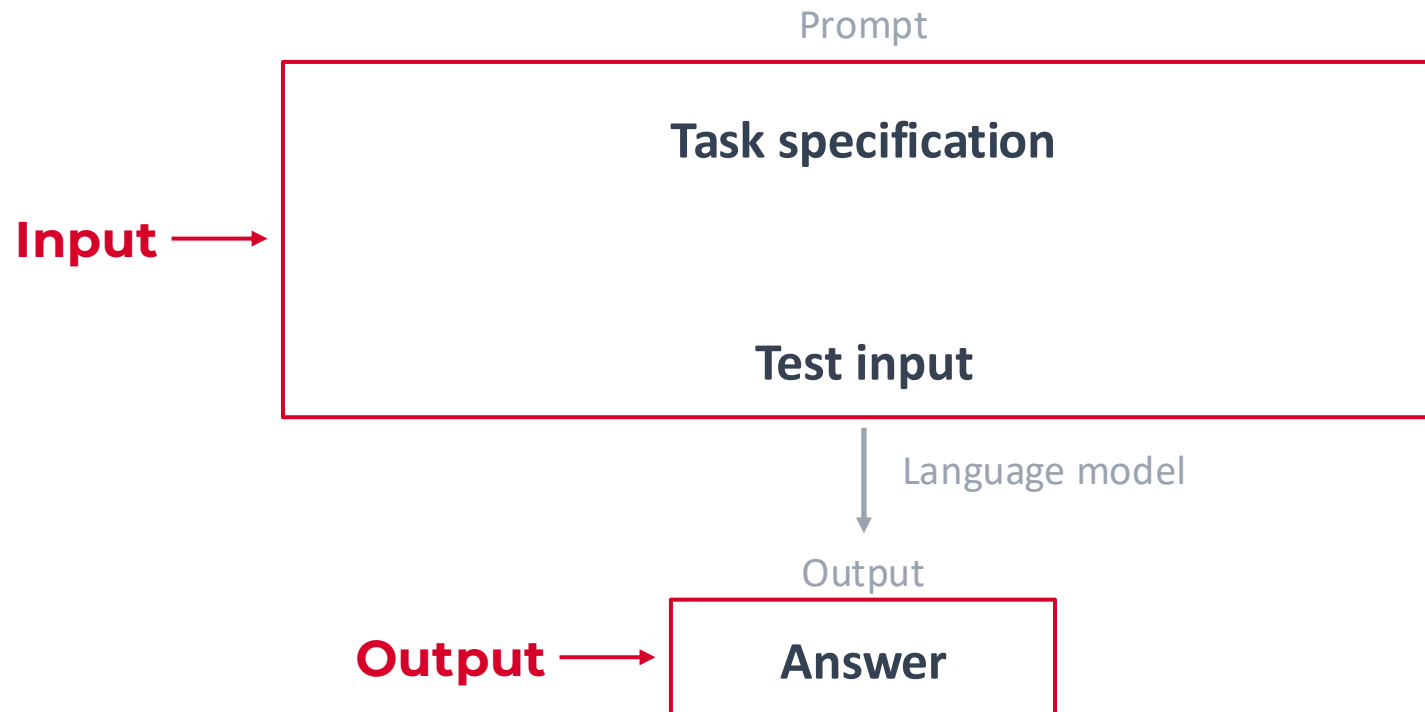
Rules and Examples in Human Learning

- ▶ Learning from **worked examples** and from **instructions** is studied, but separately rather than against each other (Sweller and Cooper, 1985; Atkinson et al., 2000; Kang et al., 2022)
- ▶ **Second-language learning** shows that instruction, exposure, and rule search support different outcomes (Doughty, 1991; DeKeyser, 1995; Robinson, 1996)
- ▶ **Learners differ** in how much they rely on abstract rules or memorized examples (McDaniel et al., 2014; Little and McDaniel, 2015; Herzog and Goldwater, 2026)

But few studies directly compare the two when **rules and examples specify the same task**.

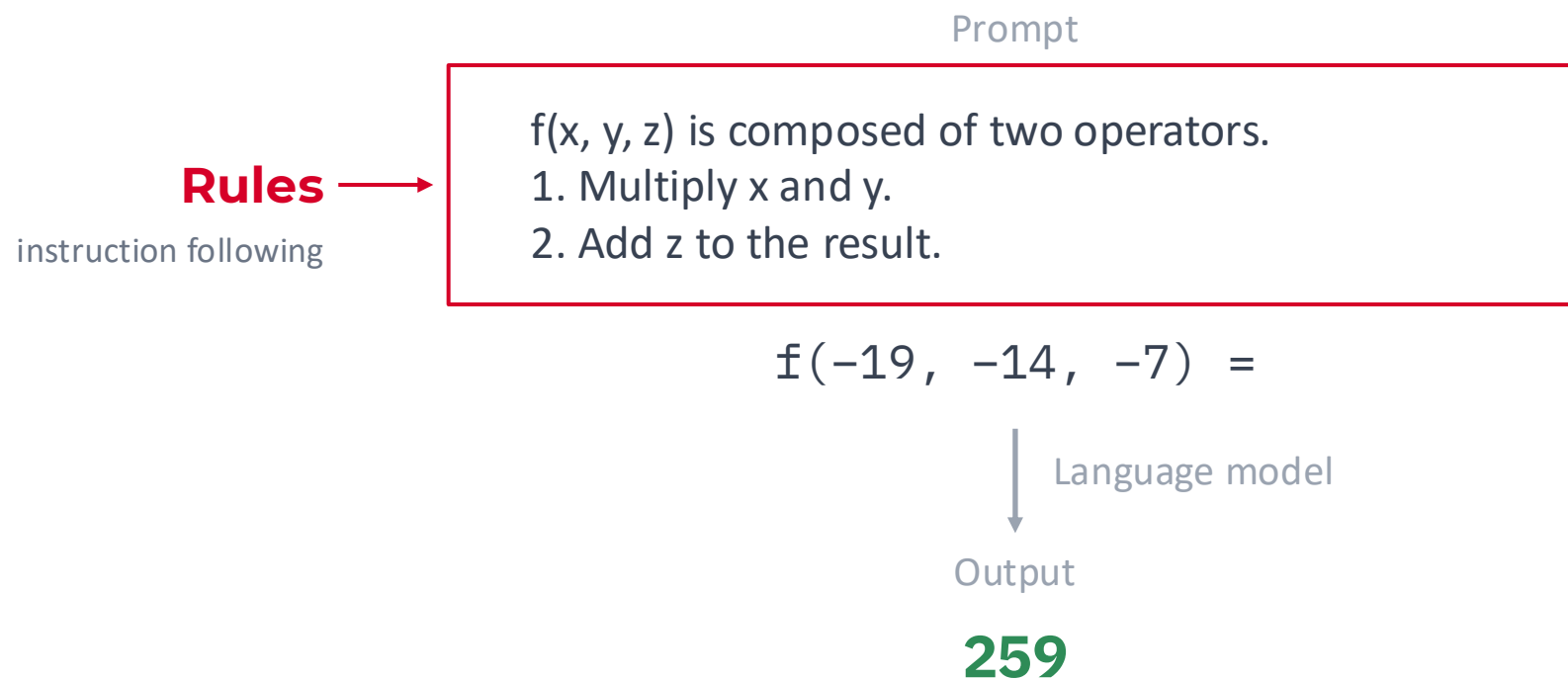
In-Context Learning

Large language models can learn new tasks in context, from the prompt alone and **without weight without weight updates** (Brown et al., 2020).



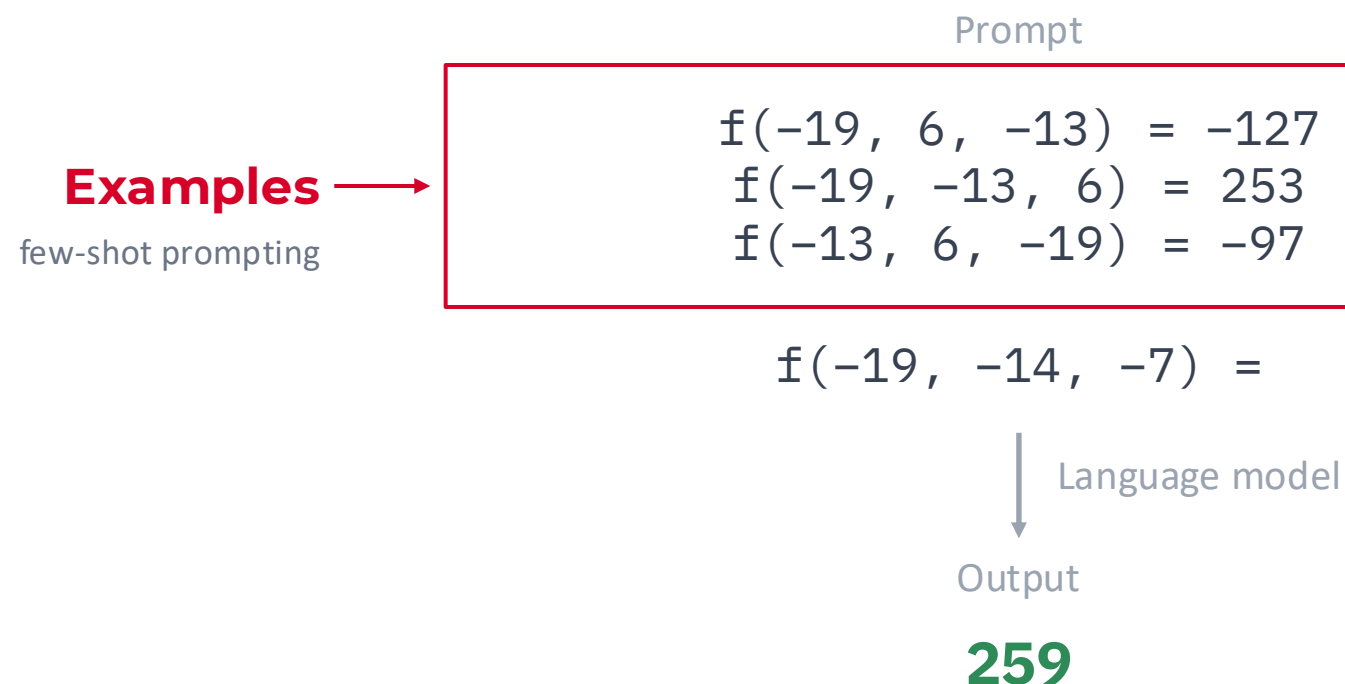
In-Context Learning

Large language models can learn new tasks in context, from the prompt alone and **without weight updates** (Brown et al., 2020).



In-Context Learning

Large language models can learn new tasks in context, from the prompt alone and **without weight updates** (Brown et al., 2020).



Learning from Examples

- ▶ **Which examples** are chosen matters (Rubin et al., 2022)
- ▶ **Their order** can swing accuracy from near state of the art to near chance (Lu et al., 2022)
- ▶ **Formatting** changes alone can shift accuracy substantially (Sclar et al., 2024)
- ▶ **More examples** often help, up to hundreds or thousands (Agarwal et al., 2024; Bertsch et al., 2025)
- ▶ Models can **induce rules and concepts** from examples, shaped by simplicity and task structure (Zhu et al., 2023; Wang et al., 2024a)

Examples must convey the **task**, the **label space**, and the **output format** implicitly (Min et al., 2022; al., 2022; Pan et al., 2023).

Learning from Rules

- ▶ **Instruction tuning** trains models on many tasks phrased as instructions (Sanh et al., 2022; Wang et al., 2022; Chung et al., 2024)
- ▶ **Following instructions** is also a form of in-context learning (Lampinen et al., 2024)
- ▶ Models still struggle with **explicit constraints, compositional rule systems, and rules that conflict with pretraining** (Mu et al., 2024; Wang et al., 2024b; Wu et al., 2024)

So it is **not obvious which mode should win.**

Comparing Rules and Examples

- ▶ **Correct instructions helped** as a side finding, though the instruction prompts also contained examples contained examples (Liu et al., 2024)
- ▶ Rules and examples recruit **partly overlapping but distinct mechanisms**, read as support for combining them (Davidson et al., 2026)
- ▶ The two prompt types **share** some task representations (Yang et al., 2026)
- ▶ Learning **in context and in weights** also differ (Chan et al., 2022; Cook et al., 2026)

Comparing Rules and Examples

- ▶ **Correct instructions helped** as a side finding, though the instruction prompts also contained examples (Liu et al., 2024)
- ▶ Rules and examples recruit **partly overlapping but distinct mechanisms**, read as support for combining them (Davidson et al., 2026)
- ▶ The two prompt types **share** some task representations (Yang et al., 2026)
- ▶ Learning **in context and in weights** also differ (Chan et al., 2022; Cook et al., 2026)

These studies use **simple one-step tasks** such as country–capital, and distinct mechanisms **need not** mechanisms **need not mean additive gains**.

A Hypothetical Example

	Example 1	Example 2	Example 3	Test
Input	$f(2, 2, 3)$	$f(2, 2, -1)$	$f(0, 0, 5)$	$f(3, 4, 1)$
Output	7	3	5	(hidden)

A Hypothetical Example

	Example 1	Example 2	Example 3	Test
Input	$f(2, 2, 3)$	$f(2, 2, -1)$	$f(0, 0, 5)$	$f(3, 4, 1)$
Output	7	3	5	(hidden)
$x \times y + z$	7	3	5	13
$x + y + z$	7	3	5	8

Both fit every example, but disagree on the test input

A Hypothetical Example

	Example 1	Example 2	Example 3	Test
Input	$f(2, 2, 3)$	$f(2, 2, -1)$	$f(0, 0, 5)$	$f(3, 4, 1)$
Output	7	3	5	13
$x \times y + z$	7	3	5	13
$x + y + z$	7	3	5	8 ✗

Rule Multiply x and y , then add z → **13**

Rules constrain the hypothesis space more tightly than a fixed set of examples.

Research Questions

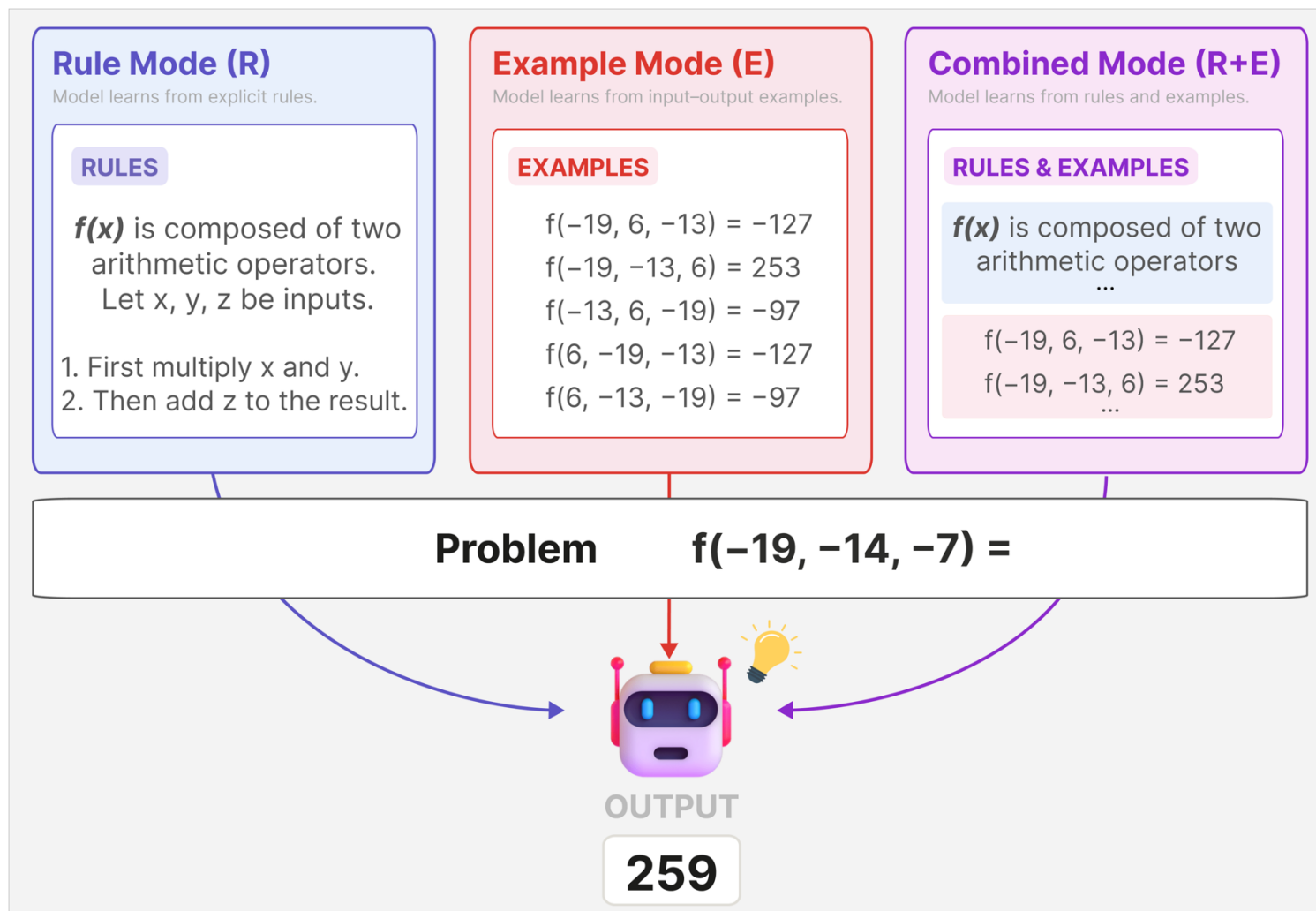
RQ1. Do models learn complex, novel tasks more reliably from **rules** (instructions) or from **examples** (demonstrations), and does **combining** the two help?

RQ2. How does the **learner**, especially **instruction tuning**, change what it learns from rules and from examples?

RQ3. Which **properties of a task** make it easier to learn from rules or from examples?

Methods

Three Learning Modes



Rules

Task described in words

Examples

Input-output pairs only

Combined

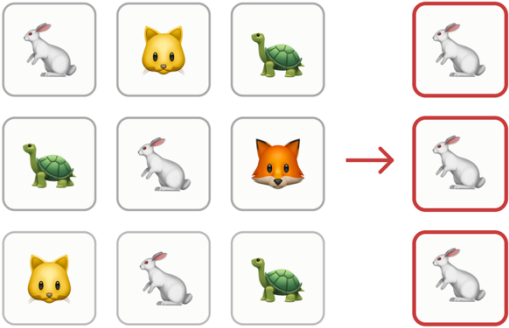
Rules, then examples

Every mode is tested on the same items

Five Novel Tasks

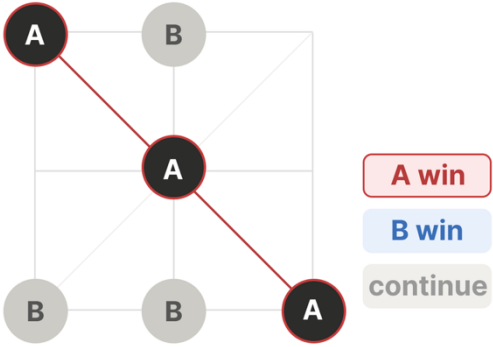
1. Set Game

Card selection task based on attribute comparison



2. Tapatan

Tapatan game outcome prediction from a final board state



3. Operator Function

Arithmetic problem with a newly defined operator function.

$$f(x, y, z) = (x \times y) + z$$

$f(-19, -14, -7) \rightarrow ?$

4. Noun Class Agreement

Grammaticality judgment (Yes/No) in an artificial language with noun-class agreement.

ka fep dax ti zesh Yes
class a a a a

ka bim miv ka sular No
class a b a a

5a. Lexical Category Inference (AND)

A candidate is accepted only if it satisfies all conditions

Item= round AND metal

list item	round	metal	
coin	✓	✓	→ Yes
ball	✓	✗	→ No

5b. Lexical Category Inference (OR)

A candidate is accepted if it satisfies at least one condition

Item= animal OR vehicle

list item	animal	vehicle	
turtle	✓	—	→ Yes
train	—	✓	→ Yes
rock	—	—	→ No

Games

Set Game, Tapatan

Arithmetic

Operator Function

Linguistic inference

Noun Class Agreement,
Lexical Category Inference

Set Game adapted from Wu et al. (2024) (2024)

Five Novel Tasks

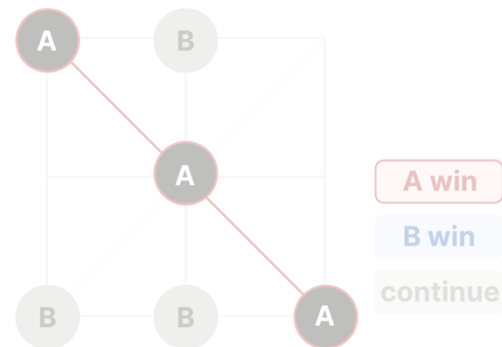
1. Set Game

Card selection task
based on attribute comparison



2. Tapatan

Tapatan game outcome prediction
from a final board state



3. Operator Function

Arithmetic problem with a newly
defined operator function.

$$f(x, y, z) = (x \times y) + z$$

$$f(-19, -14, -7) \rightarrow ?$$

4. Noun Class Agreement

Grammaticality judgment (Yes/No)
in an artificial language
with noun-class agreement.

ka fep dax ti zesh Yes

class a a a a

ka bim miv ka sular No

class a b a a

5a. Lexical Category Inference (AND)

A candidate is accepted
only if it satisfies all conditions

Item= round AND metal

list item	round	metal	
coin	✓	✓	→ Yes
ball	✓	✗	→ No

5b. Lexical Category Inference (OR)

A candidate is accepted
if it satisfies at least one condition

Item= animal OR vehicle

list item	animal	vehicle	
turtle	✓	—	→ Yes
train	—	✓	→ Yes
rock	—	—	→ No

**Rules should help
most**

Both tasks need algebraic
algebraic abstraction over
over variables

Five Novel Tasks

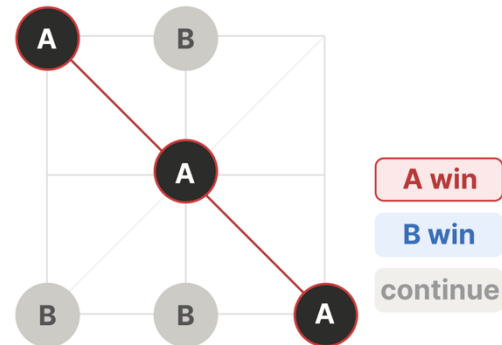
1. Set Game

Card selection task
based on attribute comparison



2. Tapatan

Tapatan game outcome prediction
from a final board state



3. Operator Function

Arithmetic problem with a newly
defined operator function.

$$f(x, y, z) = (x \times y) + z$$

$$f(-19, -14, -7) \rightarrow ?$$

4. Noun Class Agreement

Grammaticality judgment (Yes/No)
in an artificial language
with noun-class agreement.

ka fep dax ti zesh **Yes**

class a a a a

ka bim miv ka sular **No**

class a b a a

5a. Lexical Category Inference (AND)

A candidate is accepted
only if it satisfies all conditions

Item= round AND metal

list item	round	metal	
coin	✓	✓	→ Yes
ball	✓	✗	→ No

5b. Lexical Category Inference (OR)

A candidate is accepted
if it satisfies at least one condition

Item= animal OR vehicle

list item	animal	vehicle	
turtle	✓	—	→ Yes
train	—	✓	→ Yes
rock	—	—	→ No

Examples may do well

The input carries
distributional cues, or the
the task draws on
parametric knowledge

Three Difficulty Levels

Harder levels add variables, values, and inference steps. We expected **rules to hold up better as tasks better as tasks get harder**, since the space of possible input-output mappings grows.

Task	What grows	Easy	Medium	Hard
Set Game	attributes per card	2	3	3 + number
Tapatan	board size	3×3	5×5	7×7
Operator Function	operators	2	3	4
Noun Class Agreement	noun classes	2	4	6
Lexical Category Inference	conditions per category	1	2	3

Minimum-Coverage Examples

Example-based learning use **minimum-coverage sets**, which contain enough information to infer the task. Each set covers

- ▶ **Every output label**, such as Yes and No, or A win, B win, and continue
- ▶ **Every task primitive**, such as card attributes, nouns, and argument orders
- ▶ **Representative cases**, such as near misses

Sets range from **3 to 120 examples** depending on the task and level, and are **resampled and reshuffled and reshuffled for every test item**, so results do not depend on one set or ordering.

These sets are likely more systematic than typical user-written few-shot prompts.

Models and Analysis

We test **seven base and instruction-tuned model pairs** from three model families.

Model	Sizes	Versions
Gemma 3 Gemma Team, 2025	12B, 27B	base and instruction-tuned
Qwen 2.5 Qwen Team, 2024b	32B	base and instruction-tuned
Qwen 3 Qwen Team, 2025b	8B, 14B	base and instruction-tuned
OLMo 3/3.1 Team OLMo, 2025	7B, 32B	base and instruction-tuned
GPT-5.4 OpenAI, 2026	API	reference only

Linear mixed model over the 14 open-weight checkpoints

accuracy ~ mode + tuning + size + difficulty + (1 | family)

Results

Results

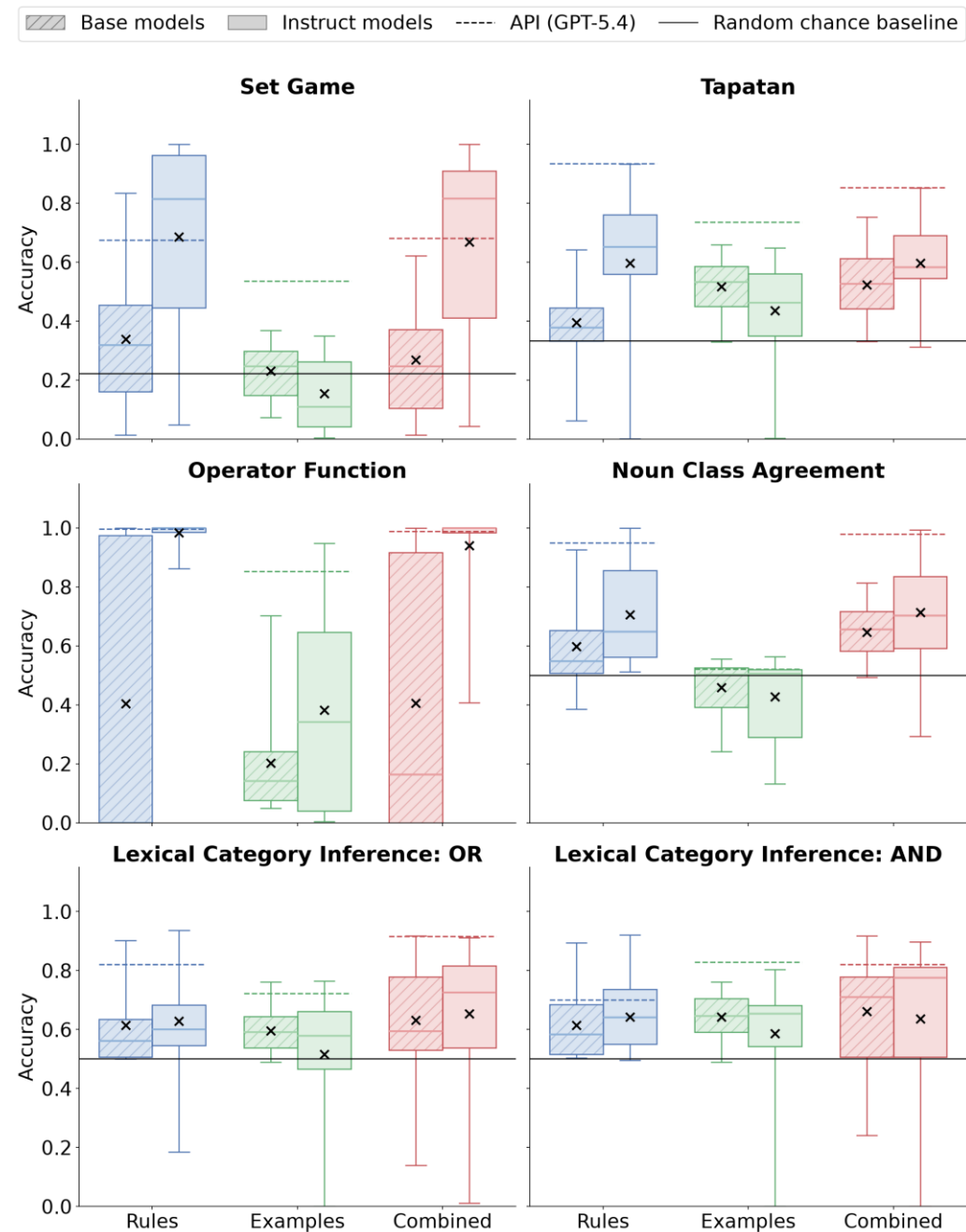
Raw Accuracy

Accuracy for each task and learning mode.

Blue is rules, green examples, and red combined. Hatched boxes are base models, solid boxes instruction-tuned models, and dashed lines GPT-5.4.

Rules are usually at or above examples, and combined looks much like rules.

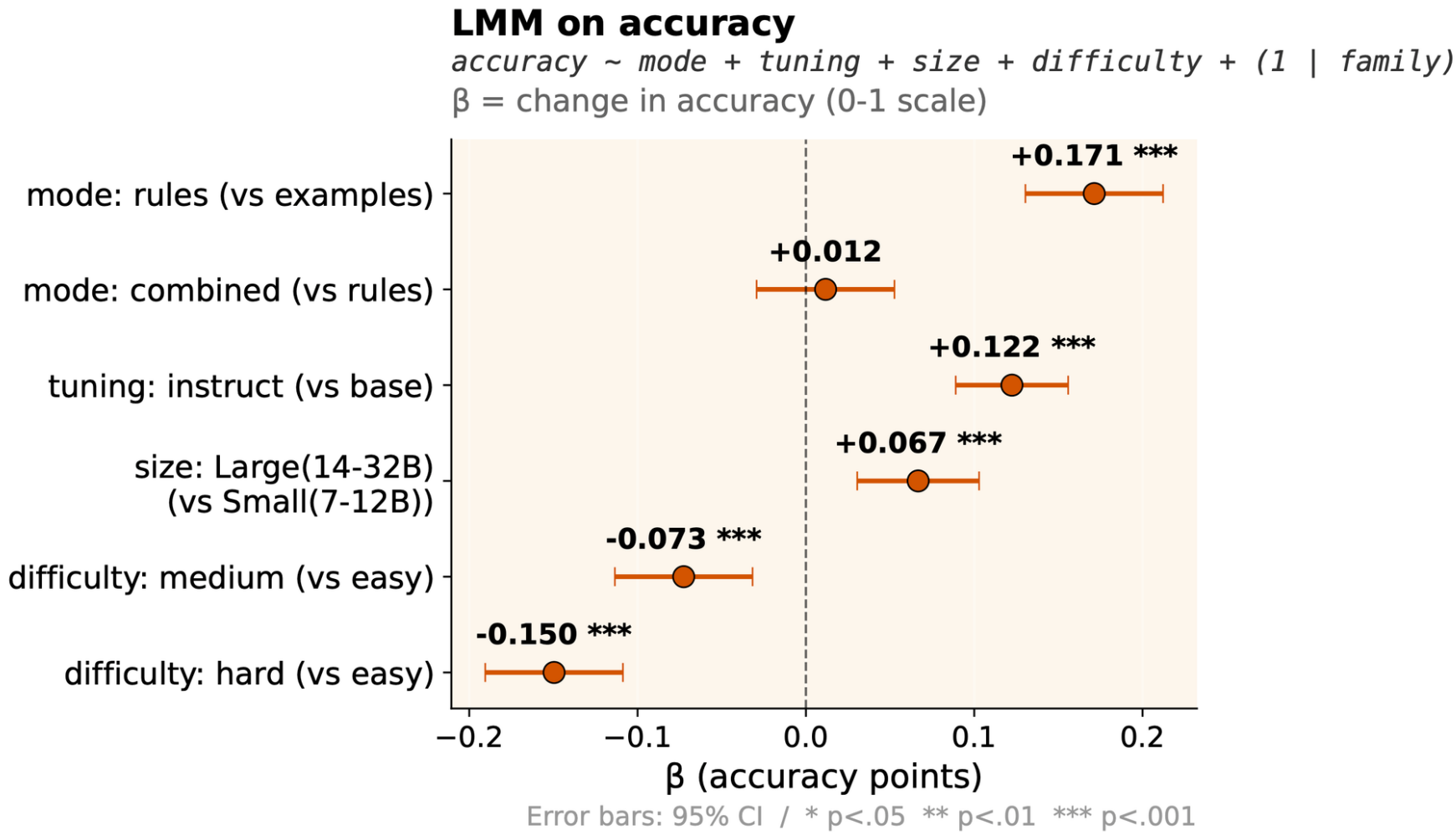
Open-weight instruction-tuned models follow patterns similar to GPT-5.4.



RQ1

Rules vs. Examples

Linear mixed model over all open-weight models. **Dots** show show the estimated change in in accuracy, with 95% confidence confidence intervals.

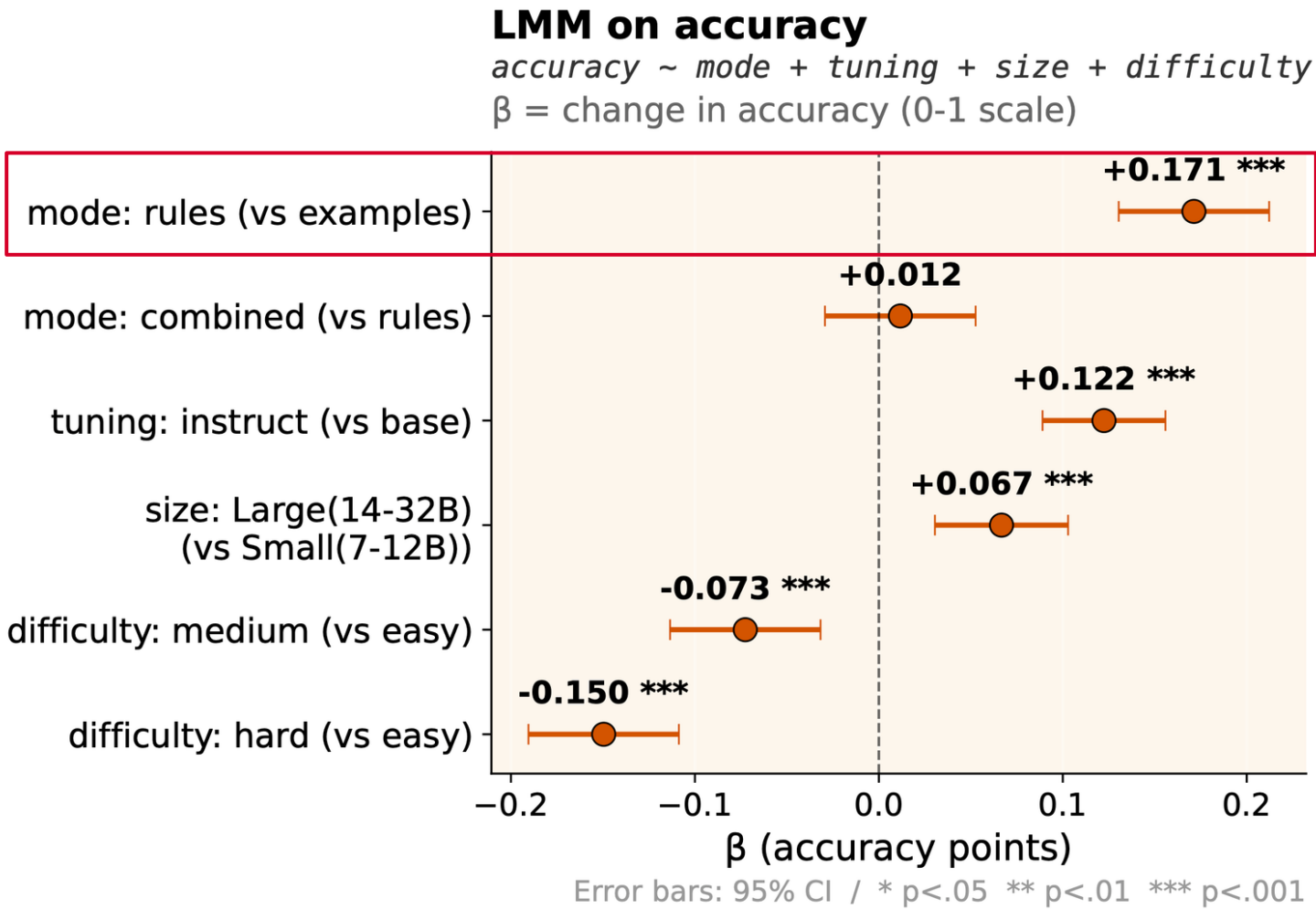


RQ1

Rules vs. Examples

Linear mixed model over all open-weight models. **Dots** show the estimated change in accuracy, with 95% confidence intervals.

Rules significantly outperform examples overall.

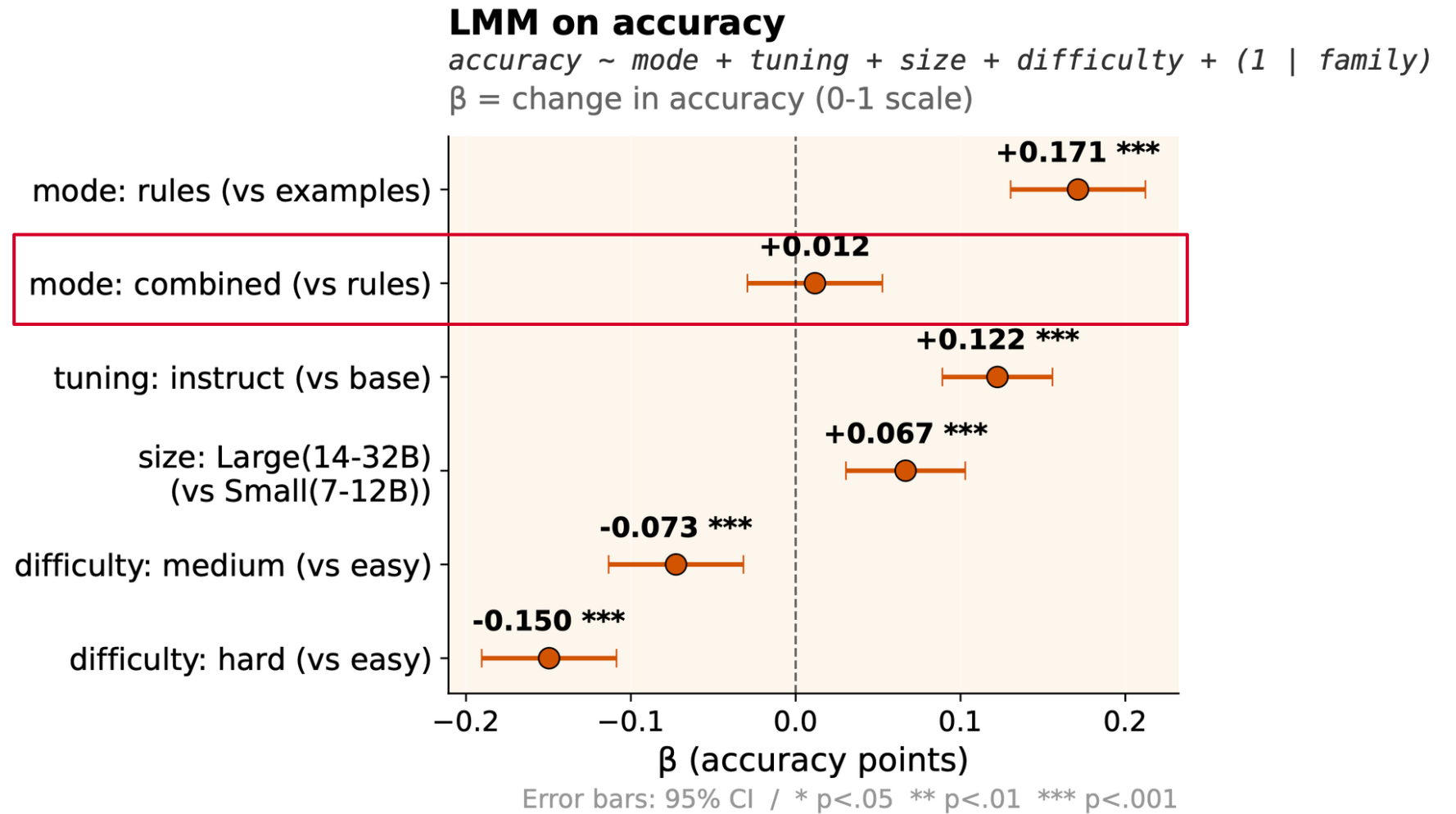


RQ1 Rules vs. Examples

Linear mixed model over all open-weight models. **Dots** show the estimated change in accuracy, with 95% confidence intervals.

Rules significantly outperform examples overall.

Adding examples to rules gives no significant gain.



RQ1 Rules vs. Examples

Linear mixed model over all open-weight models. **Dots** show the estimated change in accuracy, with 95% confidence intervals.

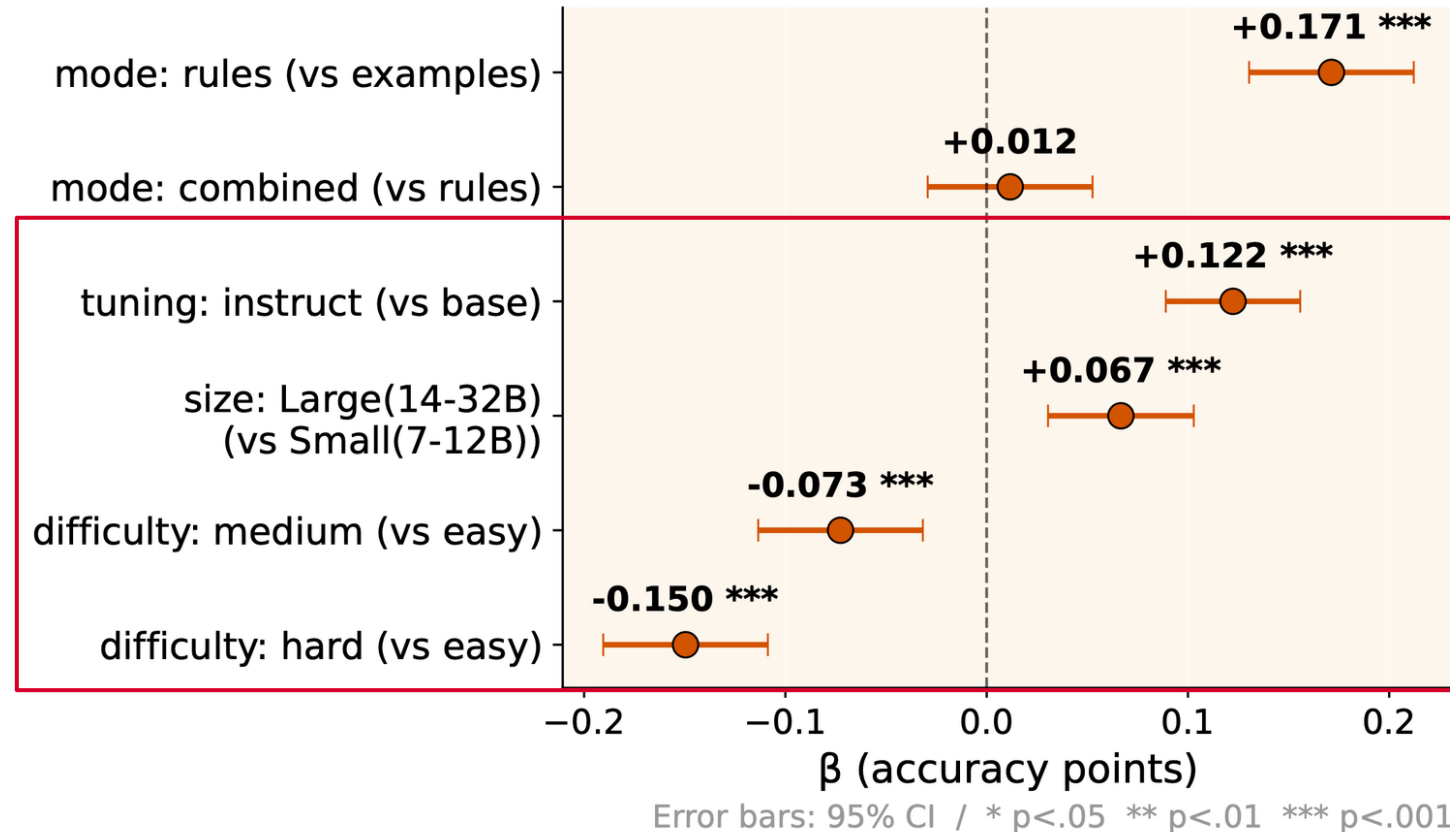
Rules significantly outperform examples overall.

Adding examples to rules gives no significant gain.

Instruction tuning and larger **size** raise accuracy, and **harder levels** lower it.

LMM on accuracy

$accuracy \sim mode + tuning + size + difficulty + (1 | family)$
 β = change in accuracy (0-1 scale)



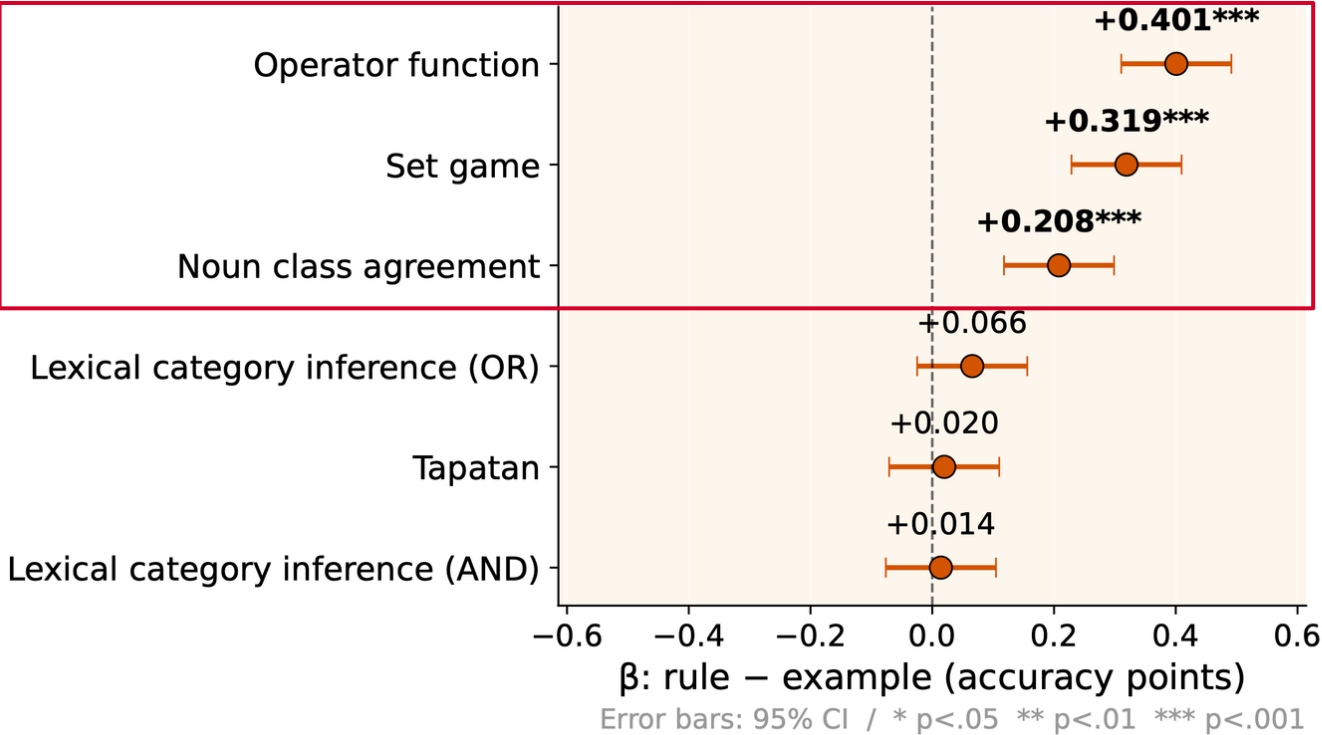
RQ1

Results by Task

Rules minus examples for each task, from the per-task mixed model.

Rules help most for **Operator Function** and **Set Game**, and less for Noun Class Agreement.

LMM: rule – example per task
 $accuracy \sim mode * task + tuning + size + difficulty + (1 | family)$
 β = change in accuracy (0-1 scale)



RQ1

Results by Task

Rules minus examples for each task, from the per-task mixed model.

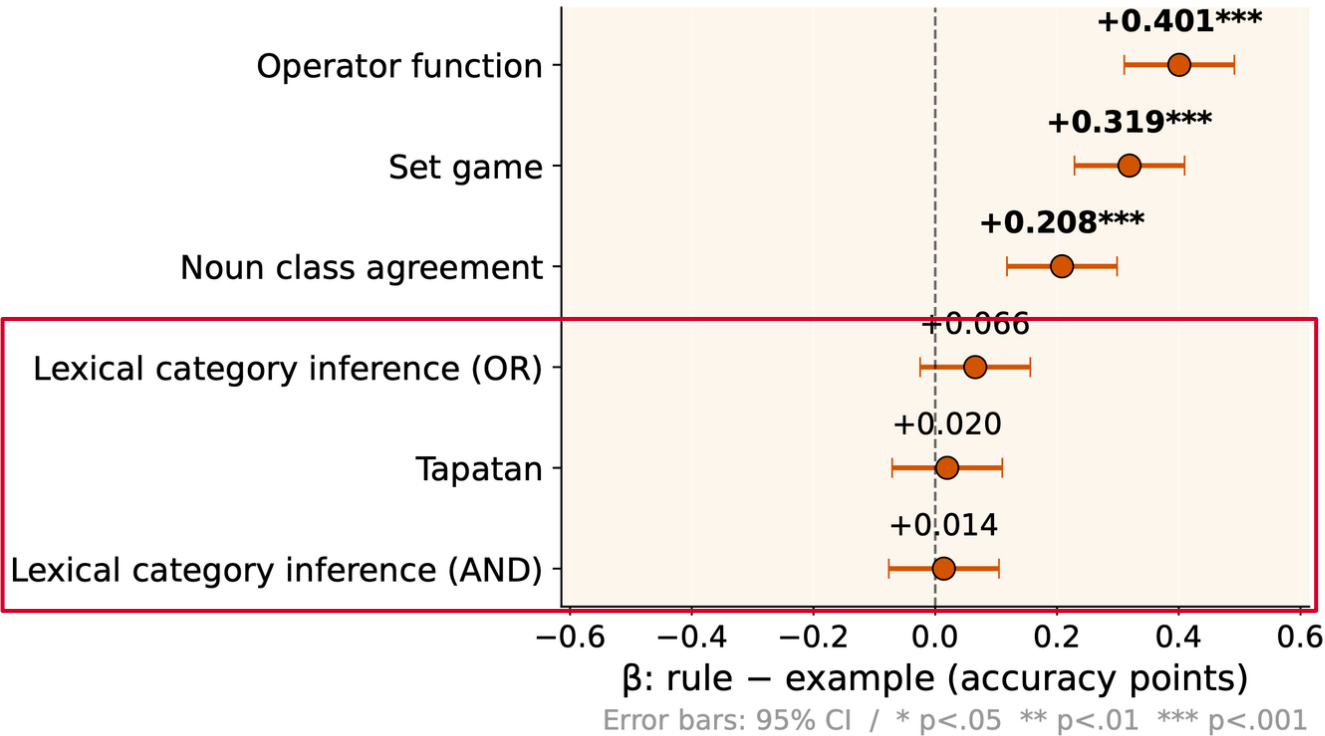
Rules help most for **Operator Function** and **Set Game**, and less for Noun Class Agreement.

Tapatan and **Lexical Category Inference** show no clear advantage.

LMM: rule – example per task

$accuracy \sim mode * task + tuning + size + difficulty + (1 | family)$

β = change in accuracy (0-1 scale)



RQ1 Combined vs. Rules

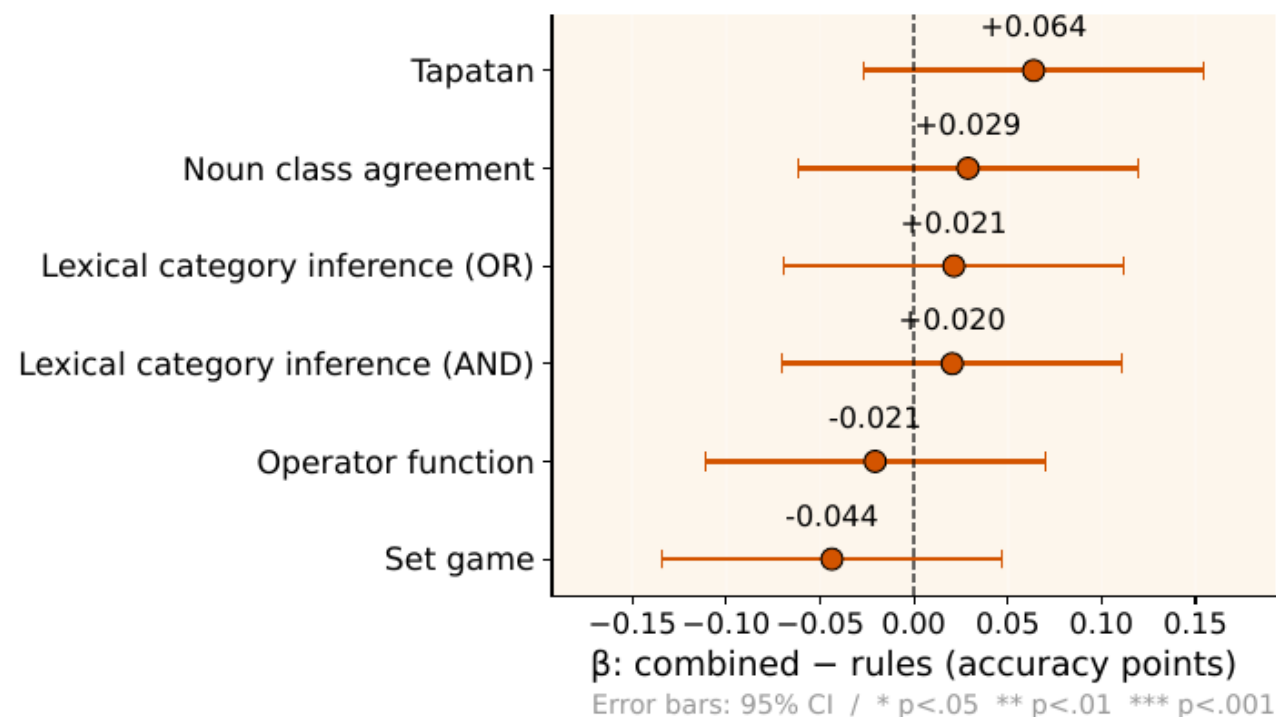
Combined minus rules for each task, from the per-task mixed model.

No task gains significantly from adding examples to rules.

LMM: combined – rules per task

$accuracy \sim mode * task + tuning + size + difficulty + (1 | family)$

β = change in accuracy (0-1 scale)

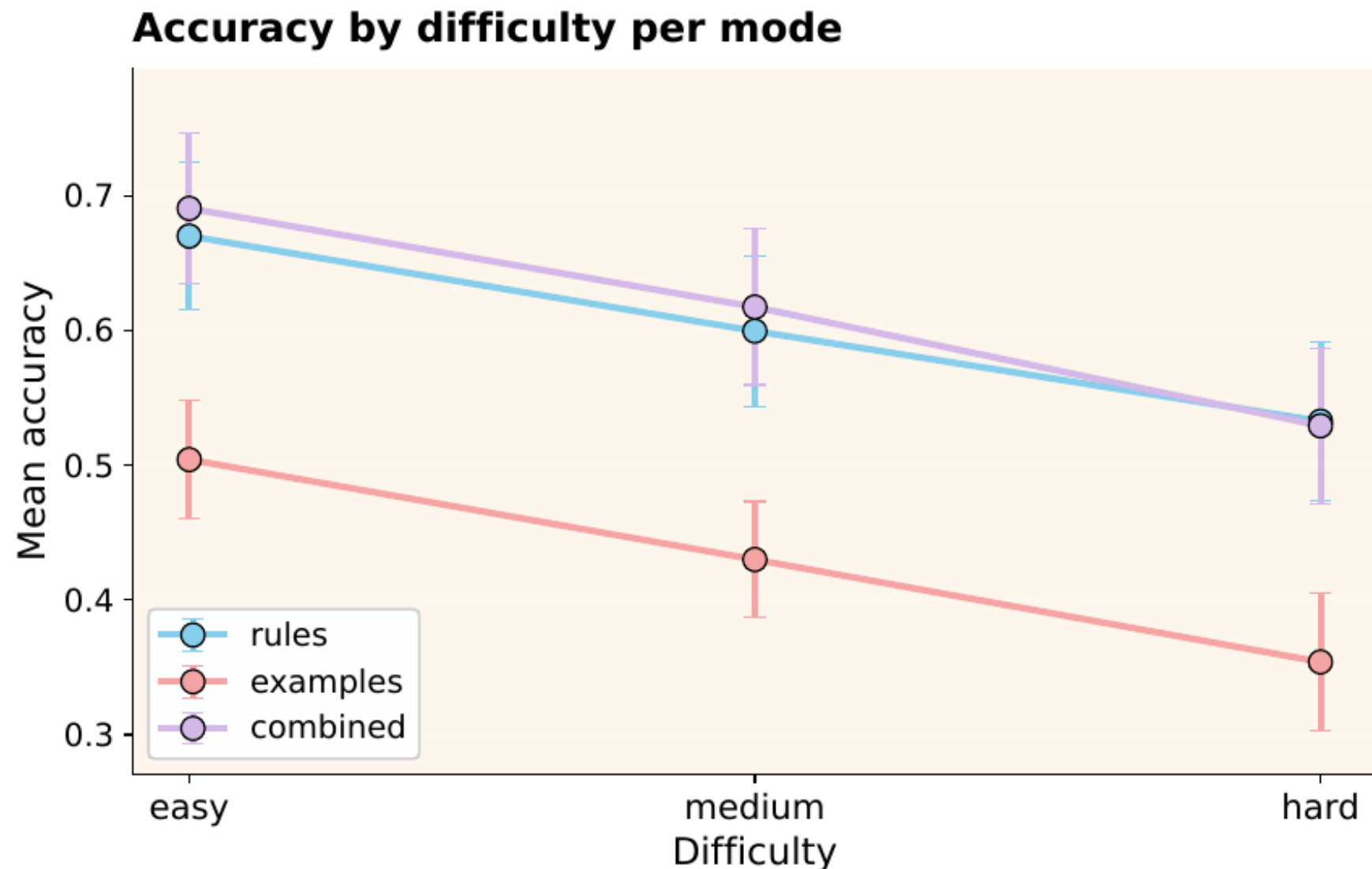


RQ1 Problem Difficulty

Mean accuracy at each difficulty level for each learning mode, with 95% confidence intervals.

Accuracy drops with difficulty in all three modes.

The slopes are **almost parallel**, so rules are **not more robust** at harder levels, contrary to our hypothesis.



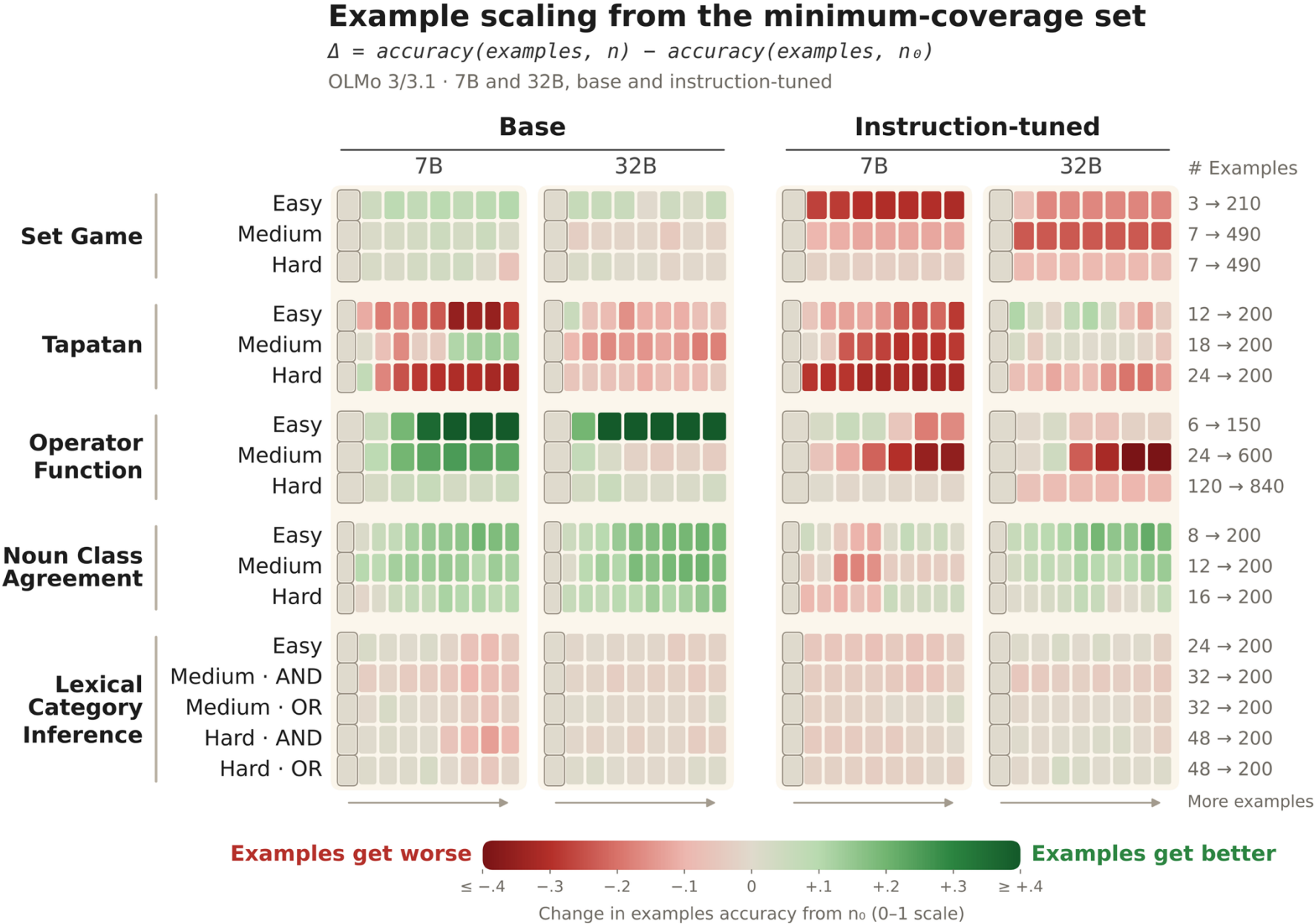
Error bars: 95% CI

RQ1

More Examples

Change in example-based accuracy as examples are added, added, for four OLMo models. **Green** is better and **red** is worse.

More examples **rarely** help.



RQ1

More Examples

Change in example-based accuracy as examples are added, added, for four OLMo models. **Green** is better and **red** is worse.

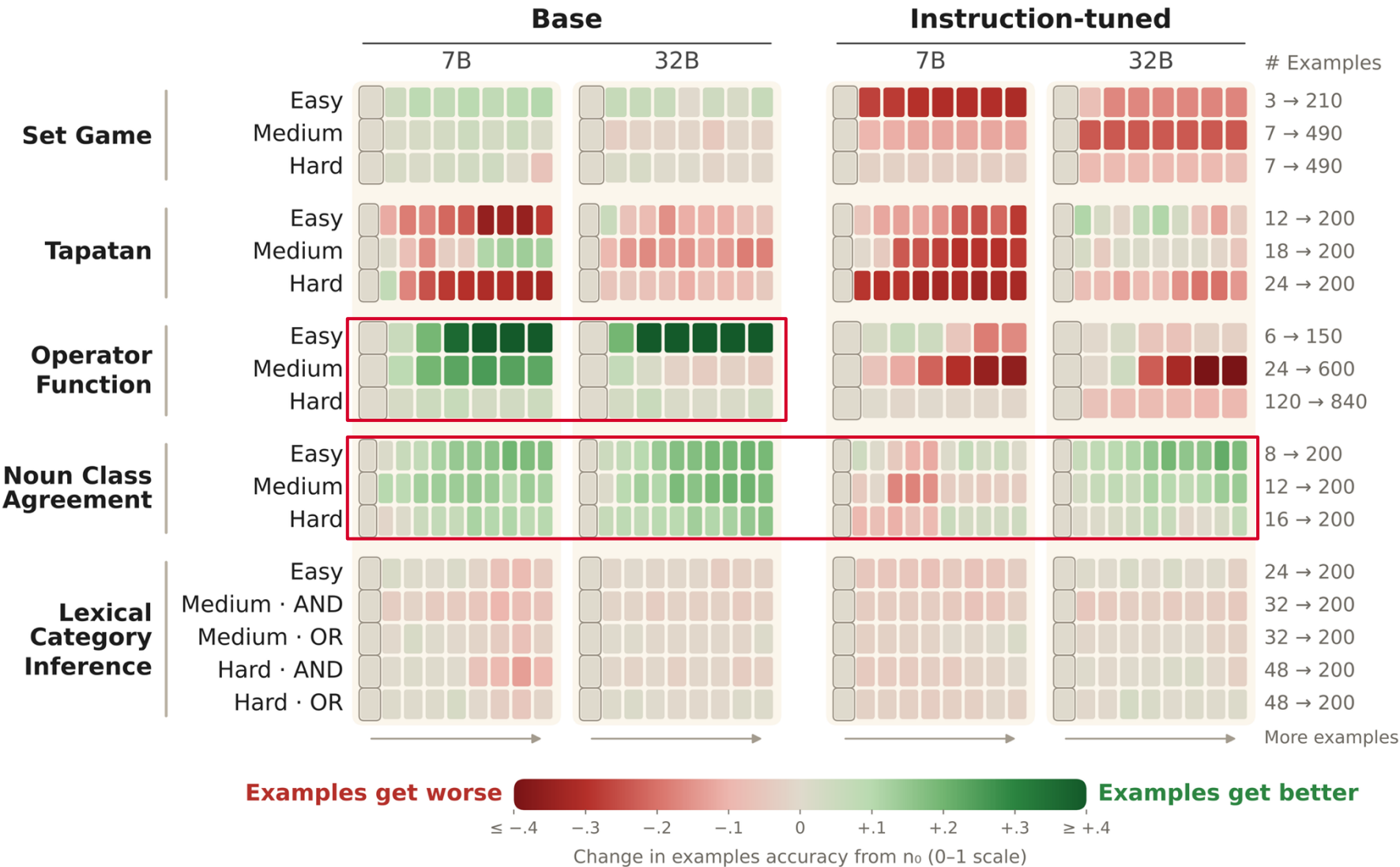
More examples **rarely** help.

Gains come mostly from **Noun Class Agreement** and from **base models on Operator Function**.

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



RQ1

More Examples

Change in example-based accuracy as examples are added, added, for four OLMo models. **Green** is better and **red** is worse.

More examples **rarely** help.

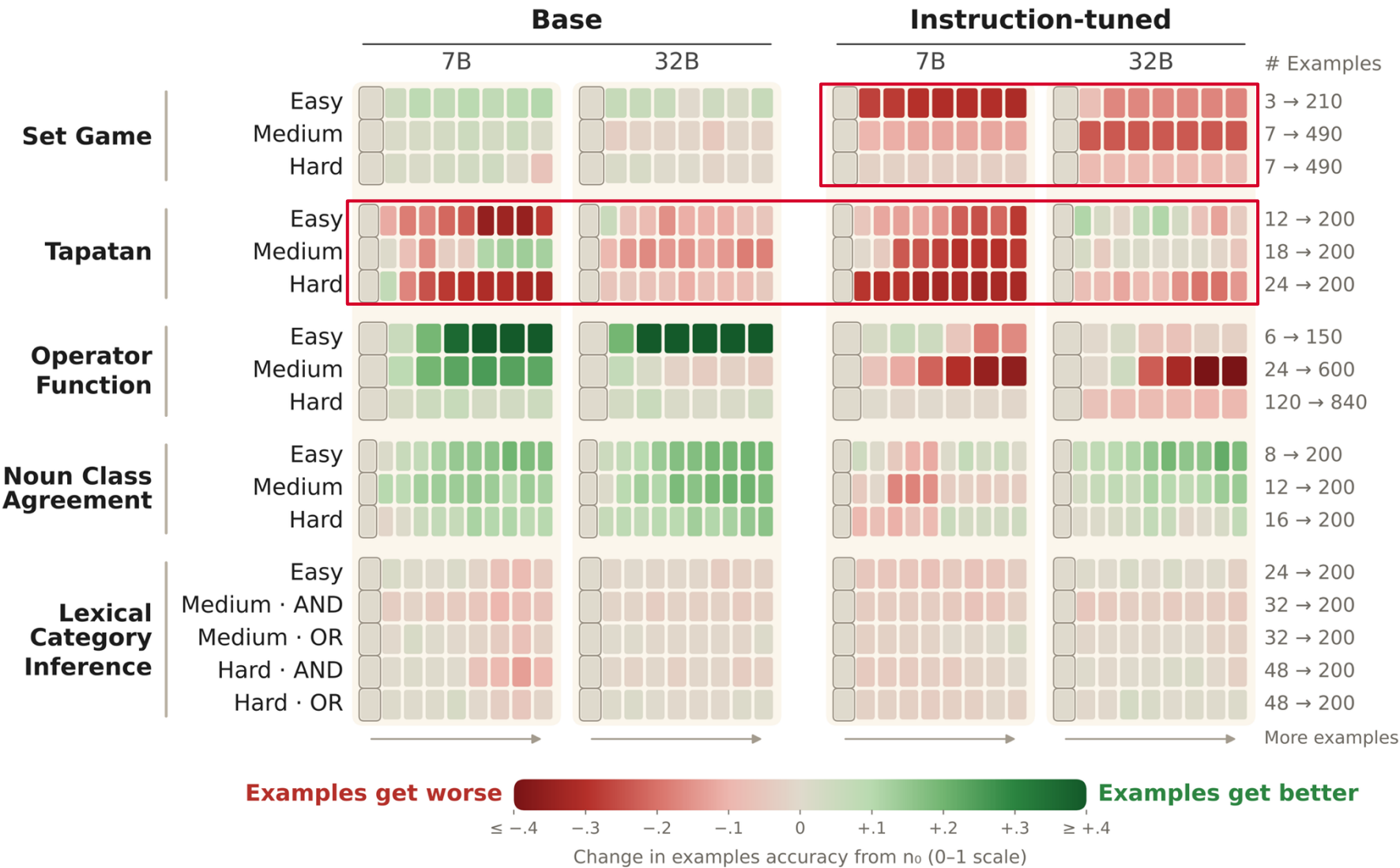
Gains come mostly from **Noun Class Agreement** and from **base models on Operator Function**.

Tapatan, and **instruction-tuned models on Set Game**, mostly get worse.

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



Analysis

RQ2 Instruction Tuning

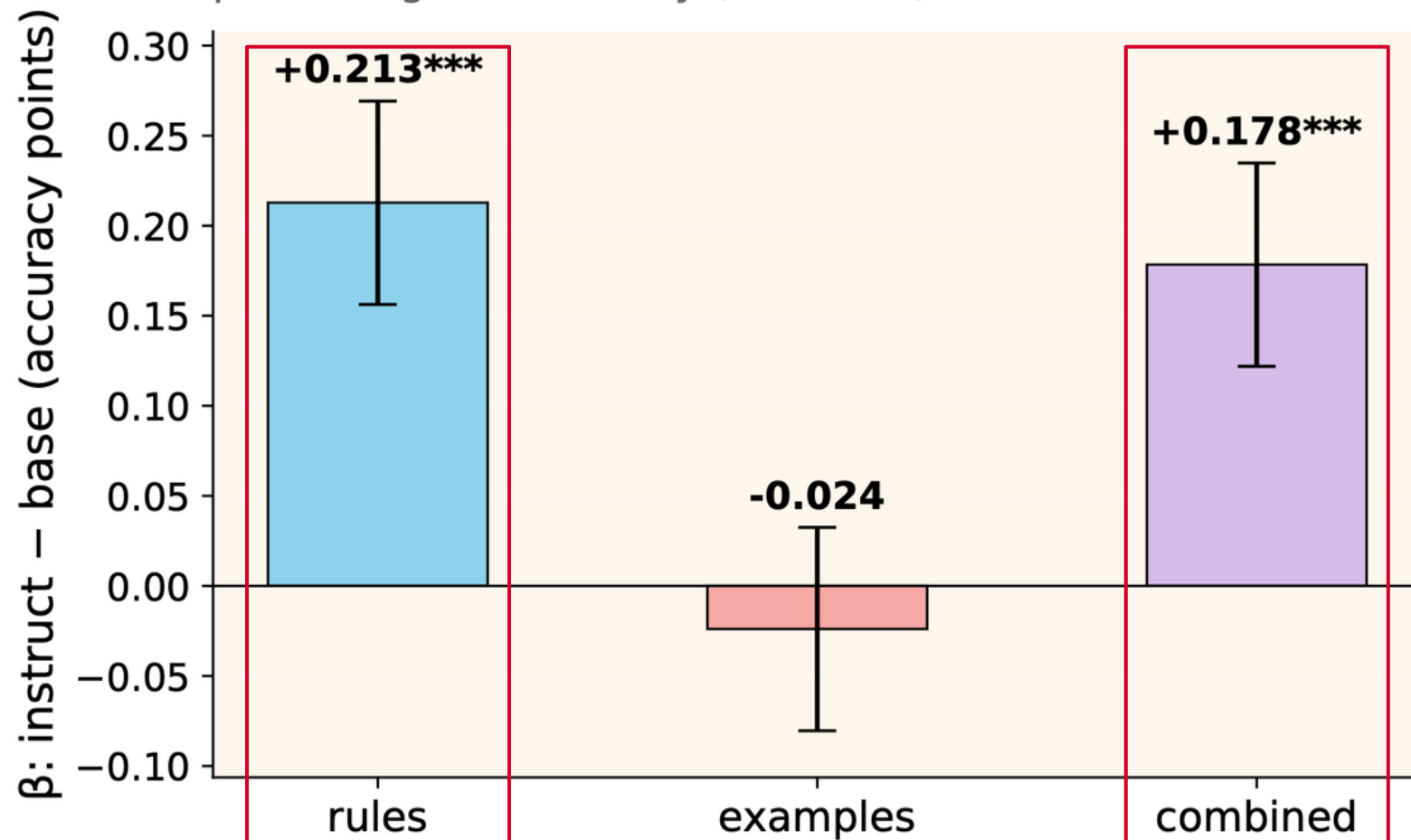
Effect of instruction tuning within each learning mode, with 95% confidence intervals.

Instruction tuning **significantly improves rule-based and combined learning.**

LMM: instruct effect within each mode

$accuracy \sim mode * tuning + size + difficulty + (1 | family)$

β = change in accuracy (0-1 scale)



Error bars: 95% CI / * $p < .05$ ** $p < .01$ *** $p < .001$

RQ2 Instruction Tuning

Effect of instruction tuning within each learning mode, with 95% confidence intervals.

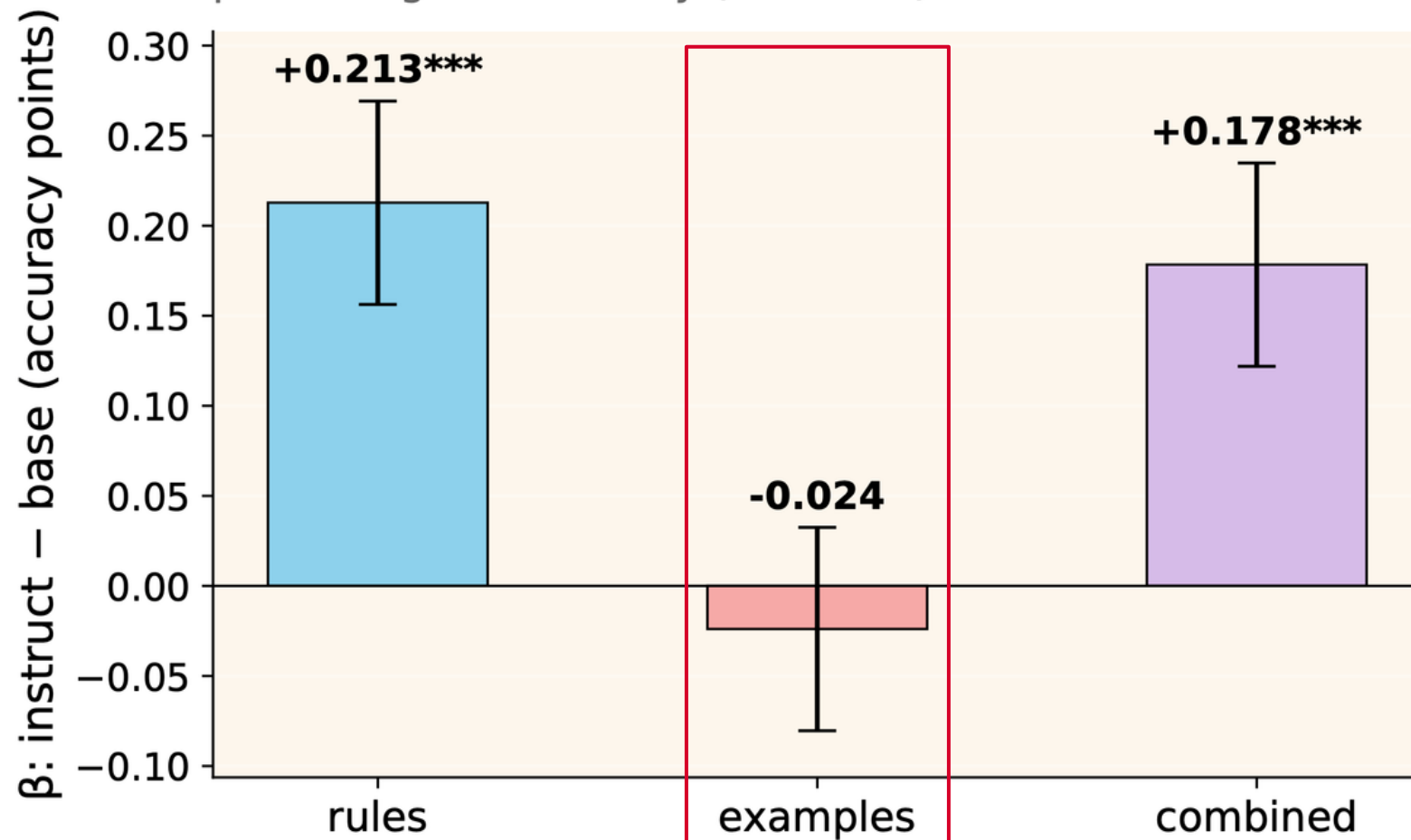
Instruction tuning **significantly improves rule-based and combined learning.**

Example-based learning shows no significant change.

LMM: instruct effect within each mode

$accuracy \sim mode * tuning + size + difficulty + (1 | family)$

β = change in accuracy (0-1 scale)



Error bars: 95% CI / * $p < .05$ ** $p < .01$ *** $p < .001$

RQ2 More Than Output Format

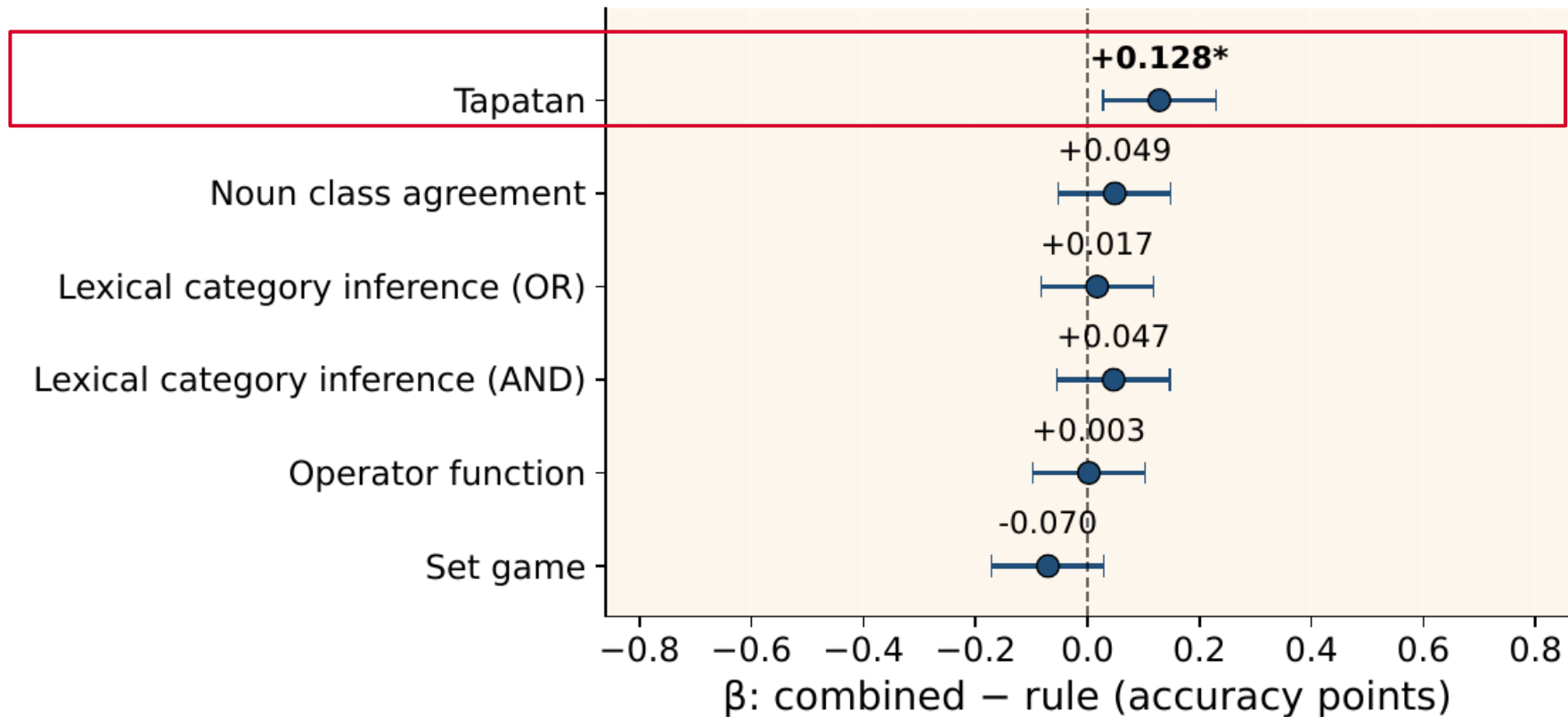
Combined minus rules for each task in base models. The added examples demonstrate the expected output format.

Adding examples to rules **helps base models little**, except on Tapatan.

So the rule gains from instruction tuning are likely **more than following the output format**.

(a) base models

subset: base only



RQ2 Base Models

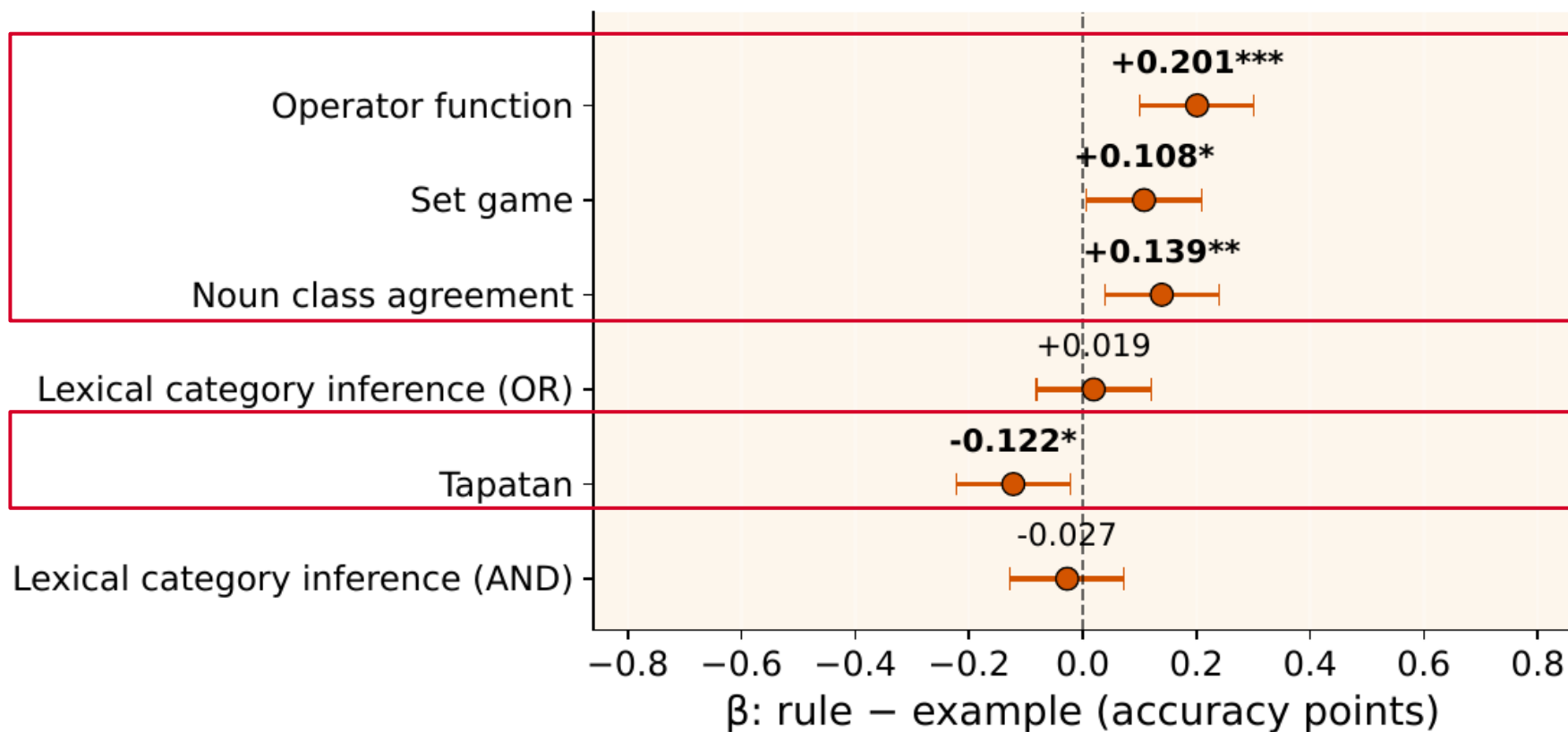
Rules minus examples for each task in base models.

Base models **also learn better from rules** on Operator Function, Set Game, and Noun Class Agreement.

Tapatan is the one task where **examples do better**.

(a) base models

subset: base only



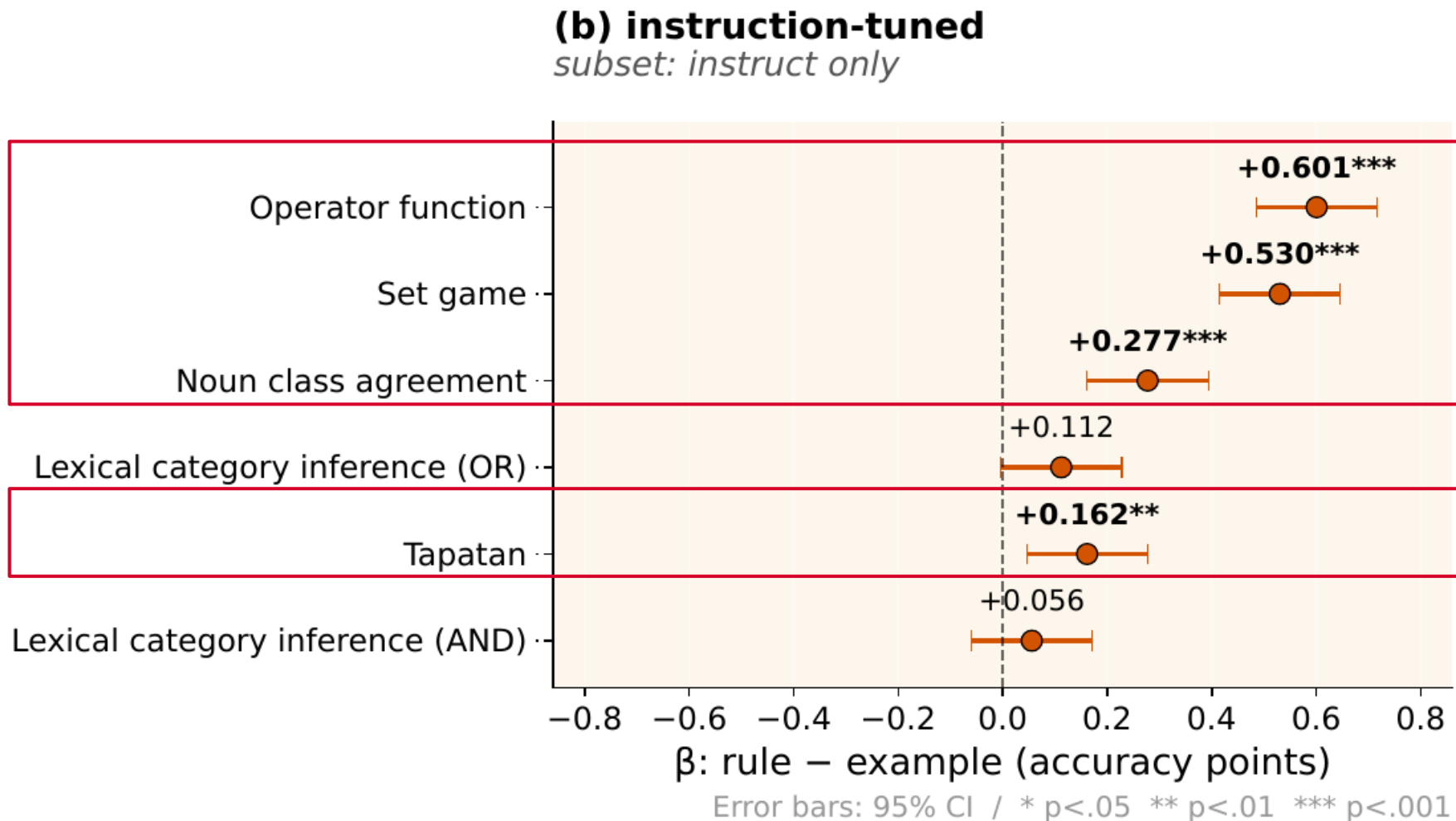
RQ2 Instruction-Tuned Models

Rules minus examples for each task in instruction-tuned models.

The rule advantage is **larger than in base models**.

Tapatan flips and now favors rules.

Lexical Category Inference still shows **no clear advantage**.



RQ2 Model Size

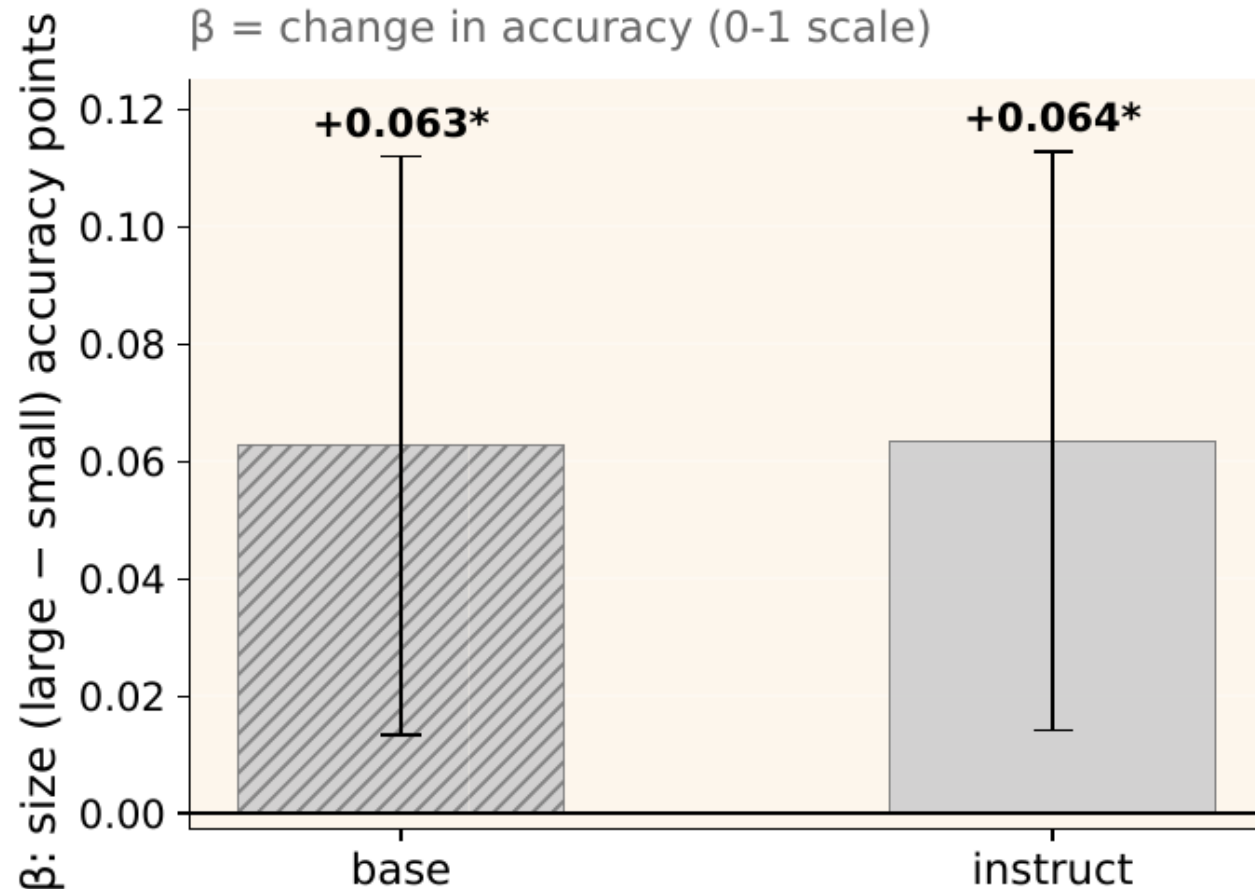
Effect of model size, comparing comparing smaller (7B–12B) and and larger (14B–32B) models, with 95% confidence confidence intervals.

Larger models learn better, in both base and instruction-tuned models.

LMM: size effect by tuning

$accuracy \sim mode + size * tuning + difficulty + (1 | family)$

β = change in accuracy (0-1 scale)



Error bars: 95% CI / * $p < .05$ ** $p < .01$ *** $p < .001$

Task Properties (RQ3)

The rule advantage **tracks what the task requires**.

- ▶ Tasks needing **algebraic abstraction and computation** (Operator Function, Set Game) show the largest advantage
- ▶ Tasks needing **formal category membership** (Noun Class Agreement) show a smaller advantage
- ▶ Tasks with **distributional cues or parametric knowledge** (Tapatan, Lexical Category Inference) show no significant advantage

Lexical Category Inference

Each item gives **four category definitions** and a **four-word sequence**. The answer is Yes if each word belongs to its category, in order.

round **AND** metal

A word must meet every condition

Word	round	metal	Fits?
coin	✓	✓	Yes
ball	✓	✗	No

animal **OR** vehicle

A word may meet any one condition

Word	animal	vehicle	Fits?
turtle	✓	—	Yes
train	—	✓	Yes
rock	—	—	No

Medium and hard levels use two and three conditions per category

Why Compare AND and OR?

People find **disjunctive (OR) concepts harder to learn** than conjunctive (AND) concepts, in adults and in children (Bruner et al., 1956; Snow and Rabinovitch, 1969).

- ▶ **With examples**, members of an AND category share every condition and look alike, while members while members of an OR category can share nothing, like a turtle and a train
- ▶ **With rules**, the logical operator is stated directly

We predict **AND > OR with examples**, and **a smaller gap with rules**.

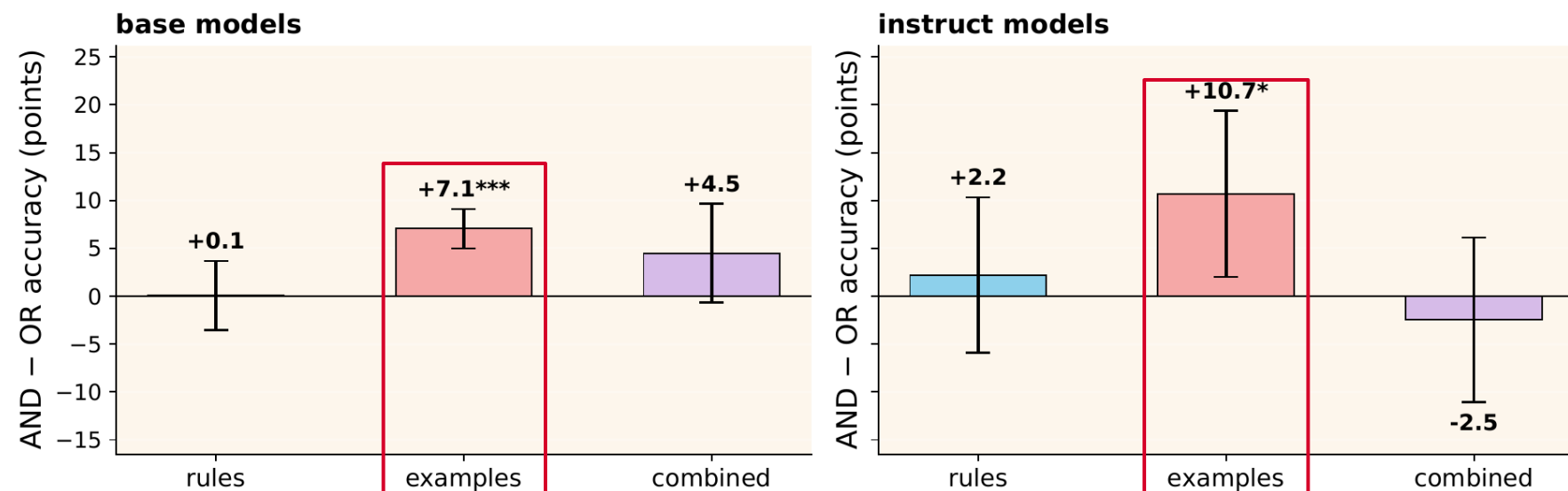
RQ3

AND vs. OR

Estimated AND minus OR accuracy in Lexical Category Inference, within each learning mode, for base and instruction-tuned models.

With examples, models do better on **AND** than on **OR**.

Lexical category inference: AND vs. OR



Error bars: 95% CI / * $p < .05$ ** $p < .01$ *** $p < .001$

RQ3

AND vs. OR

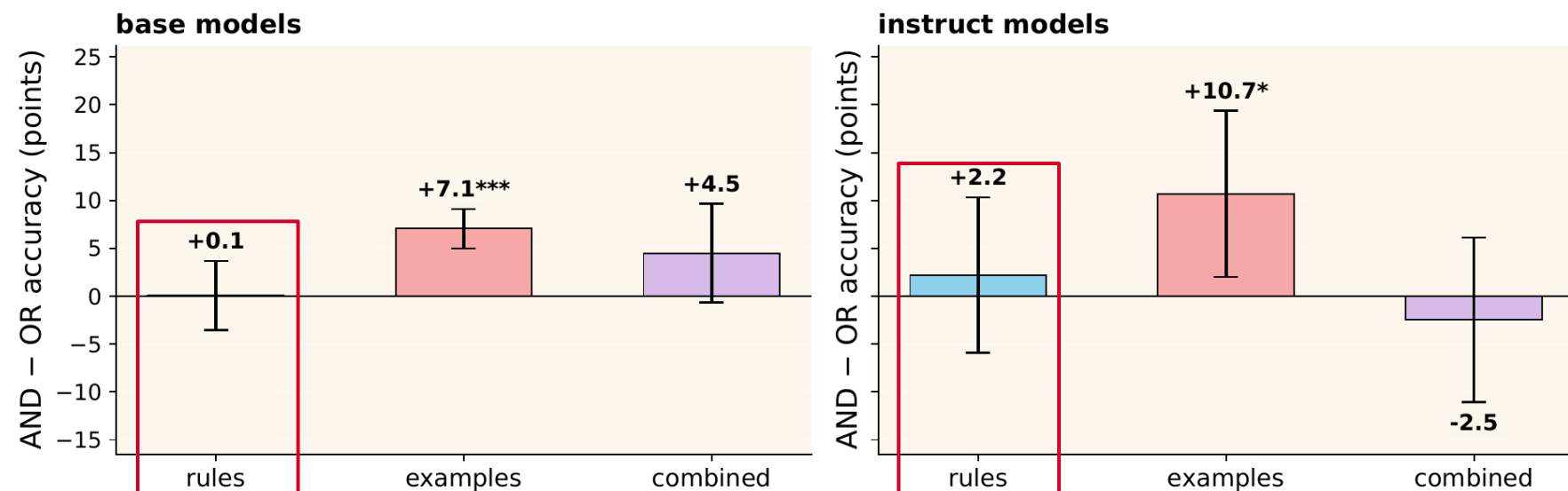
Estimated AND minus OR accuracy in Lexical Category Inference, within each learning mode, for base and instruction-tuned models.

With examples, models do better on **AND** than on **OR**.

With rules, the gap is **small and not significant**.

Under examples, the gap **grows with difficulty**, which does not happen under rules.

Lexical category inference: AND vs. OR



Error bars: 95% CI / * $p < .05$ ** $p < .01$ *** $p < .001$

Why Do Rules Help?

Rules state directly what examples leave implicit, so the inference is easier for a learner that can follow rules (Min et al., 2022; Pan et al., 2023).

- ▶ **Rule following emerges without instruction tuning**, and instruction tuning amplifies it
- ▶ With clear, complete rules, **correct examples become redundant**. Reported gains from examples come from settings where rules are likely underspecified (e.g., Wang et al., 2022)
- ▶ Our tuning results fit **two distinct mechanisms** (Davidson et al., 2026), but they **do not add up** when combined

Additional Analyses (See Paper)

- ▶ AND vs. OR accuracy at each difficulty level, including GPT-5.4
- ▶ Combined vs. rules for each task in instruction-tuned models
- ▶ Coverage criteria and example counts for every task and level

Conclusion

Conclusion

We find that LLMs **learn complex, novel tasks more reliably from rules than from examples**, and that **adding examples to rules gives no significant gain**.

Instruction tuning strengthens rule-based learning while leaving example-based learning intact, and **base models show a similar rule advantage**.

The rule advantage is **largest when a task needs algebraic abstraction and computation**, and **smallest when it relies on distributional cues or parametric knowledge**. Whether it holds for vague or partial rules is still open.

Links

Links

Paper	<u>https://arxiv.org/abs/2609.03213</u>
Code	<u>https://github.com/tinlaboratory/rules-vs-examples</u>
Data	<u>https://huggingface.co/datasets/tin-lab/rules_vs_examples</u>

References

References

- Agarwal, R., et al. 2024. Many-shot in-context learning. *NeurIPS*. NeurIPS.
- Allen Institute for AI. 2026. OLMo 3 model cards.
- Atkinson, R. K., Derry, S. J., Renkl, A., and Wortham, D. 2000. Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*.
- Bertsch, A., et al. 2025. In-context learning with long-context models: An in-depth exploration. *NAACL*.
- Brown, T. B., et al. 2020. Language models are few-shot learners. *NeurIPS*.
- Bruner, J. S., Goodnow, J. J., and Austin, G. A. 1956. *A Study of Thinking*. Wiley.
- Chan, S. C. Y., et al. 2022. Transformers generalize differently from information stored in context vs in weights. *arXiv:2210.05675*.
- Chung, H. W., et al. 2024. Scaling instruction-finetuned language models. *JMLR*.
- Cook, J., et al. 2026. Programming by backprop: An instruction is worth 100 examples when finetuning LLMs. *ICLR*.
- Dasgupta, I., Grant, E., and Griffiths, T. 2022. Distinguishing rule and exemplar-based generalization in learning systems. *ICML*.
- Davidson, G., Gureckis, T., Lake, B., and Williams, A. 2026. Do different prompting methods yield a common task representation in language models? *NeurIPS*.
- DeKeyser, R. M. 1995. Learning second language grammar rules: An experiment with a miniature linguistic system. *Studies in Second Language Acquisition*.
- Doughty, C. 1991. Second language instruction does make a difference: Evidence from an empirical study of SL relativization. *Studies in Second Language Acquisition*.
- Gemma Team. 2025. Gemma 3 technical report. *arXiv:2503.19786*.
- Gentner, D., and Medina, J. 1998. Similarity and the development of rules. *Cognition*.
- Google. 2026. Gemma terms of use.
- Herzog, S. A., and Goldwater, M. B. 2026. Evidence for rule versus exemplar learning strategies as stable individual differences independent from working memory. *Memory & Cognition*.
- Kang, W., Pineda Hernández, S., Wang, J., and Malvaso, A. 2022. Instruction-based learning: A review. *Neuropsychologia*.
- Kwon, W., et al. 2023. Efficient memory management for large language model serving with PagedAttention. *SOSP*.
- Lampinen, A. K., Chan, S. C. Y., Singh, A. K., and Shanahan, M. 2024. The broader spectrum of in-context learning. *arXiv:2412.03782*.
- Little, J. L., and McDaniel, M. A. 2015. Some learners abstract, others memorize examples: Implications for education. *Translational Issues in Psychological Science*.
- Liu, E., Neubig, G., and Andreas, J. 2024. An incomplete loop: Instruction inference, instruction following, and in-context learning in language models. *COLM*.
- Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *ACL*.
- McDaniel, M. A., Cahill, M. J., Robbins, M., and Wiener, C. 2014. Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology: General*.
- Min, S., et al. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *EMNLP*.
- Mu, N., et al. 2024. Can LLMs follow simple rules? *arXiv:2311.04235*.
- Nosofsky, R. M. 1986. Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*.
- OpenAI. 2025. OpenAI services agreement.
- OpenAI. 2026. Introducing GPT-5.4.
- Pan, J., Gao, T., Chen, H., and Chen, D. 2023. What in-context learning “learns” in-context: Disentangling task recognition and task learning. *Findings of ACL*.
- Qwen Team. 2024a. Qwen2.5: A party of foundation models.
- Qwen Team. 2024b. Qwen2.5 technical report. *arXiv:2412.15115*.
- Qwen Team. 2025a. Qwen3.
- Qwen Team. 2025b. Qwen3 technical report. *arXiv:2505.09388*.
- Robinson, P. 1996. Learning simple and complex second language rules under implicit, incidental, rule-search, and instructed conditions. *Studies in Second Language Acquisition*.
- Rubin, O., Herzig, J., and Berant, J. 2022. Learning to retrieve prompts for in-context learning. *NAACL*.
- Sanh, V., et al. 2022. Multitask prompted training enables zero-shot task generalization. *ICLR*.
- Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. 2024. Quantifying language models’ sensitivity to spurious features in prompt design. *ICLR*.
- Smith, E. E., and Sloman, S. A. 1994. Similarity- versus rule-based categorization. *Memory & Cognition*.
- Snow, C. E., and Rabinovitch, M. S. 1969. Conjunctive and disjunctive thinking in children. *Journal of Experimental Child Psychology*.
- Sweller, J., and Cooper, G. A. 1985. The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*.
- Team OLMo. 2025. OLMo 3. *arXiv:2512.13961*.
- Tenenbaum, J. B. 1999. Rules and similarity in concept learning. *NeurIPS*.
- Wang, L. Z., McCoy, R. T., and Steinert-Threlkeld, S. 2024a. Minimization of Boolean complexity in in-context concept learning. *arXiv:2412.02823*.
- Wang, S., Wei, Z., Choi, Y., and Ren, X. 2024b. Can LLMs reason with rules? Logic scaffolding for stress-testing and improving LLMs. *ACL*.
- Wang, Y., et al. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. *EMNLP*.
- Wolf, T., et al. 2020. Transformers: State-of-the-art natural language processing. *EMNLP System Demonstrations*.
- Wu, Z., et al. 2024. Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual tasks. *NAACL*.
- Yang, Z., et al. 2026. Shared lexical task representations explain behavioral variability in LLMs. *arXiv:2604.22027*.
- Zhu, Z., et al. 2023. Large language models can learn rules. *arXiv:2310.07064*.