Experimental Contexts *Can* Facilitate Robust Semantic Property Inference in Language Models, but Inconsistently

Kanishka Misra[1,2]*, Allyson Ettinger[3], Kyle Mahowald[2]

* *Work done as a Postdoc at the University of Texas at Austin*

SCAN ME

1 TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

2 TEXAS The University of Texas at Austin

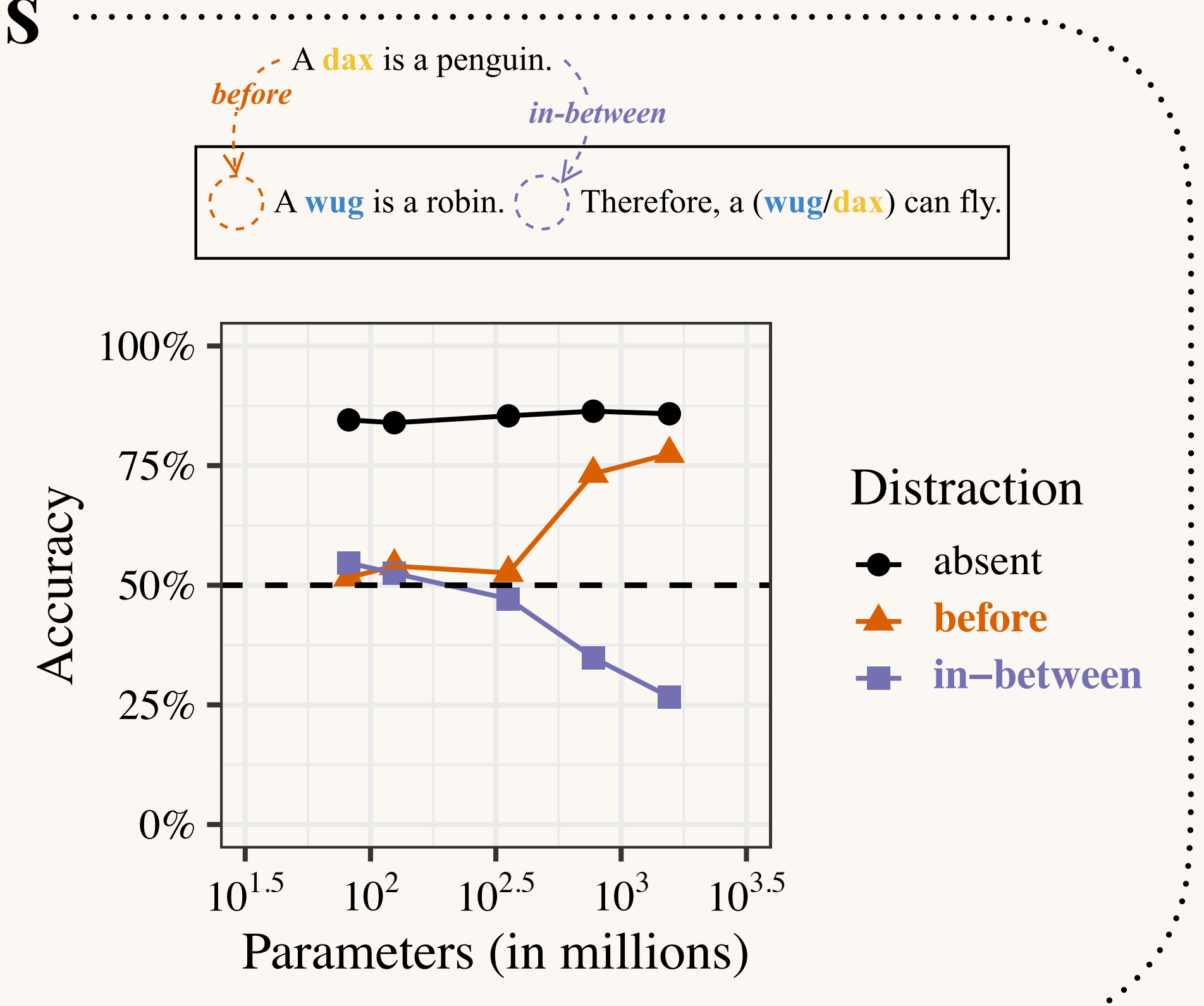
3 Ai2

COMPS: Conceptual Minimal Pair Sentences

A dataset to evaluate property knowledge and its robust property inheritance for novel concepts *(Misra et al., 2023, EACL)*

A **robin** *can fly.* >? A **penguin** *can fly.*

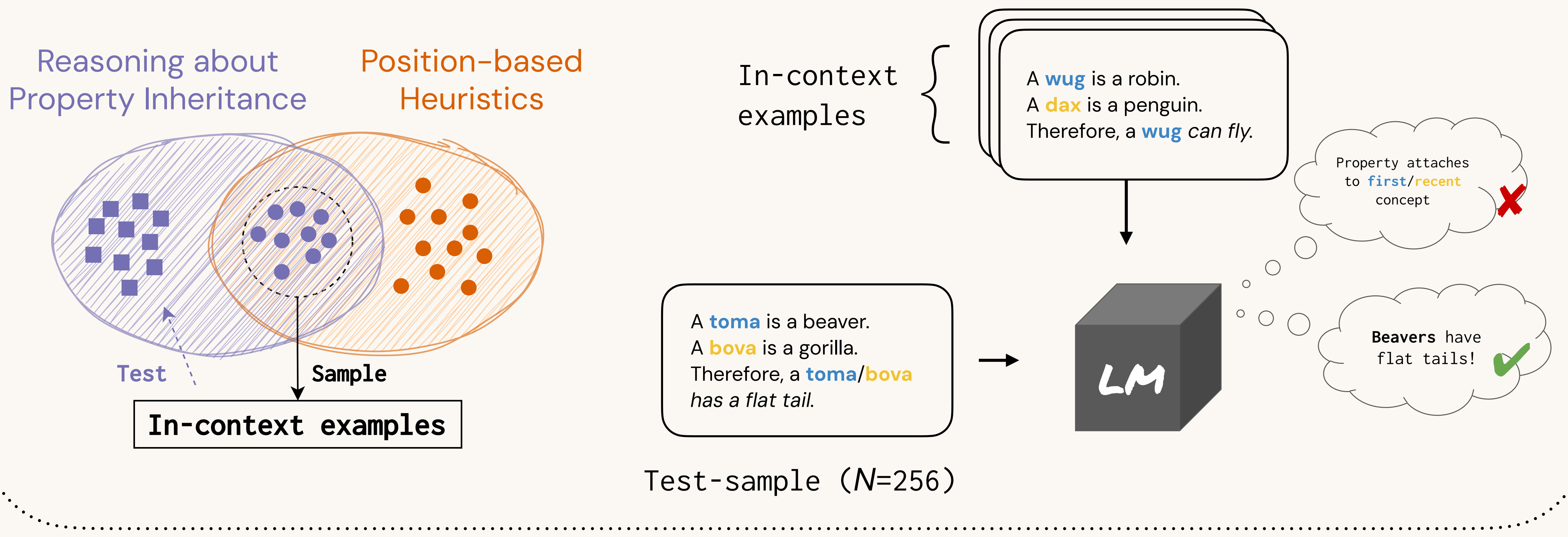
A **wug** is a robin.
A **dax** is a penguin.
Therefore, a **wug** *can fly.* >? A **wug** is a robin.
A **dax** is a penguin.
Therefore, a **dax** *can fly.*



Premise: LMs perform below chance when tasked to perform property inheritance for novel concepts in a zero-shot setting.

But what happens when they are guided to an appropriate experimental context (In-context learning/instructions)? *Lampinen (2022)*

Experiment Design


Reasoning about Property Inheritance
Position-based Heuristics
Test
Sample
In-context examples
In-context examples
LM
Property attaches to first/recent concept
Beavers have flat tails!
Test-sample (N=256)

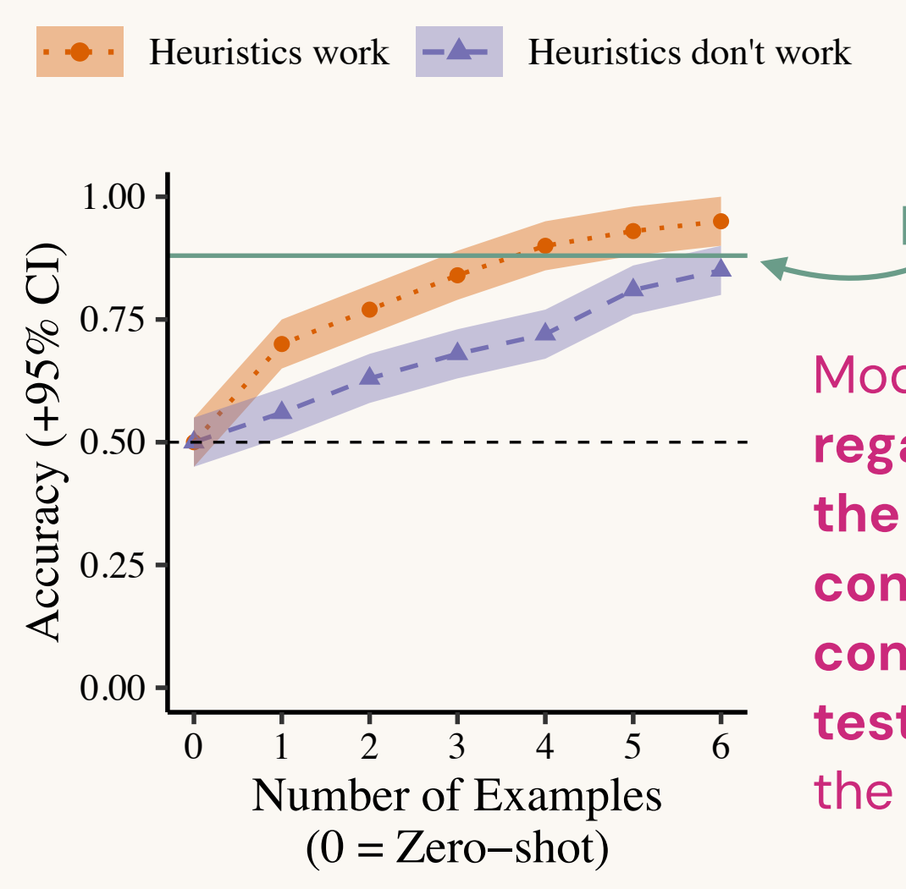
A **wug** is a robin.
A **dax** is a penguin.
Therefore, a **wug** *can fly.*

A **toma** is a beaver.
A **bova** is a gorilla.
Therefore, a **toma**/**bova** *has a flat tail.*

Controls

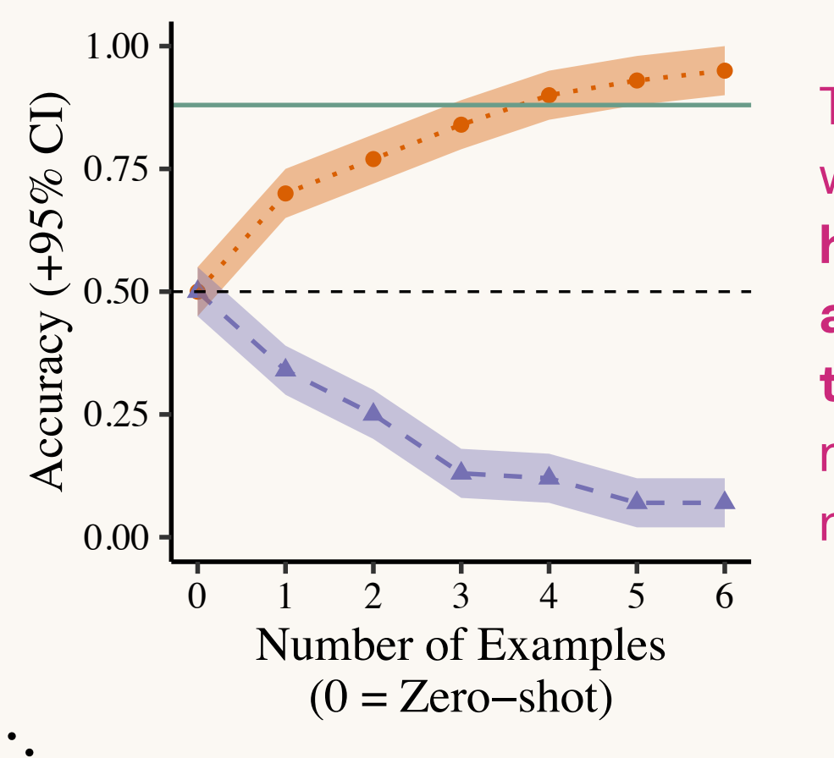
- Disjointness between IC-examples and test stimuli in terms of:
  - Novel words used (wug, dax, etc.)
  - Concepts and Properties
- Multiple sets of IC examples (10) to measure variability.
- Novel words are counterbalanced (bias towards one → chance performance).
- Two types of **heuristics** tested:
  - **First correct** vs. **Recent correct**

How to read the plots

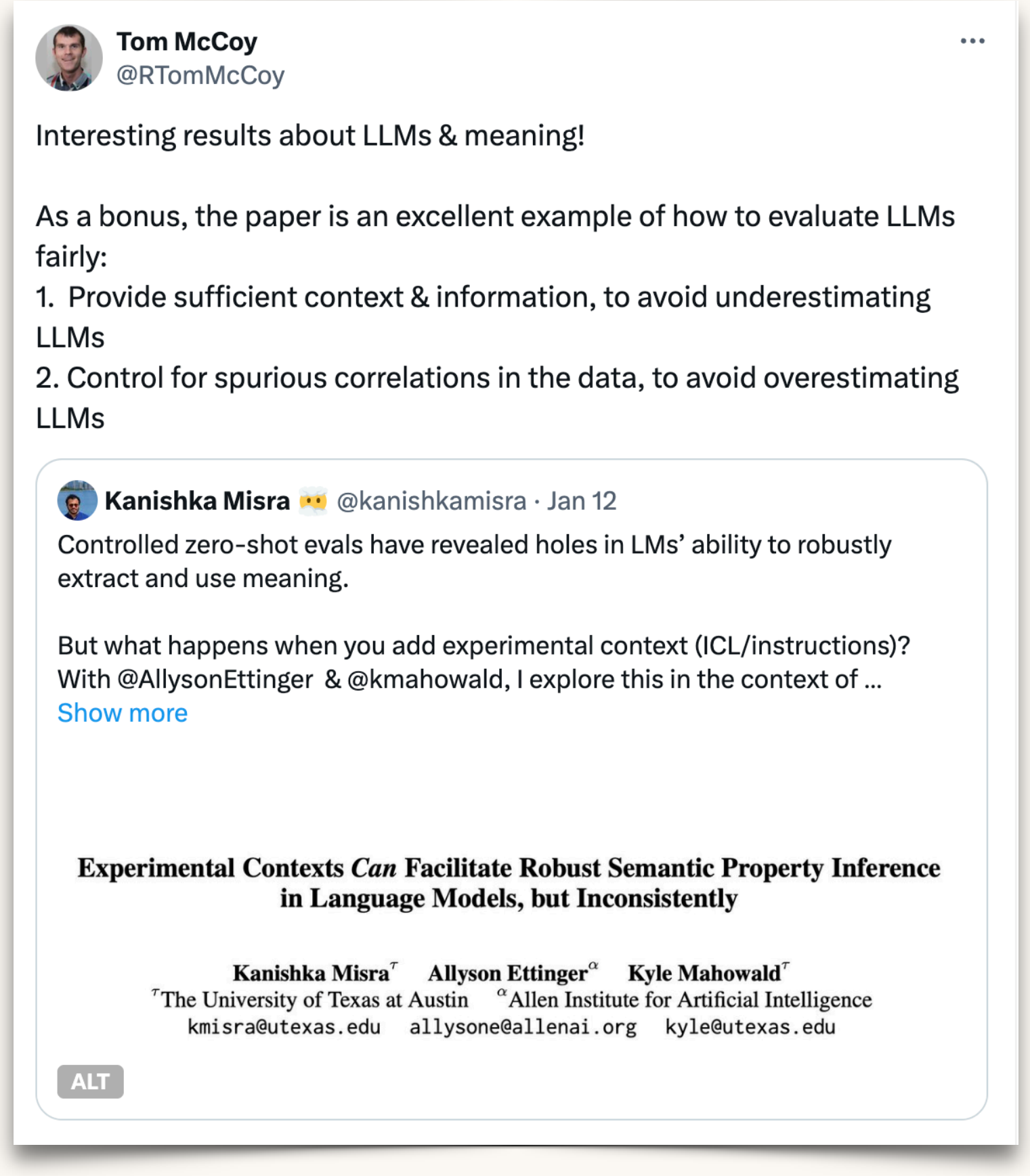


LMs' base property knowledge performance

Model performs well **regardless of whether the heuristics in its context are consistent with the test set.** It is robust to the mixed-signals!



The model performs well **only when the heuristics in context are consistent with the test set**, and fails miserably when that's not the case!



Tom McCoy
@RTomMcCoy

Interesting results about LLMs & meaning!

As a bonus, the paper is an excellent example of how to evaluate LLMs fairly:
1. Provide sufficient context & information, to avoid underestimating LLMs
2. Control for spurious correlations in the data, to avoid overestimating LLMs

Kanishka Misra @kanishkamisra · Jan 12

Controlled zero-shot evals have revealed holes in LMs' ability to robustly extract and use meaning.

But what happens when you add experimental context (ICL/instructions)? With @AllysonEttinger & @kmahowald, I explore this in the context of ...
Show more

Experimental Contexts *Can* Facilitate Robust Semantic Property Inference in Language Models, but Inconsistently

Kanishka Misra[τ] Allyson Ettinger[α] Kyle Mahowald[τ]
[τ]The University of Texas at Austin [α]Allen Institute for Artificial Intelligence
kmisra@utexas.edu allysone@allenai.org kyle@utexas.edu

ALT

Experimental context can improve attribution of properties to concepts...

Accuracy = Proportion of time:

p_θ(*has a flat tail* | … + **toma**) > p_θ(*has a flat tail* | … + **bova**)

COMPS

{Instructions}

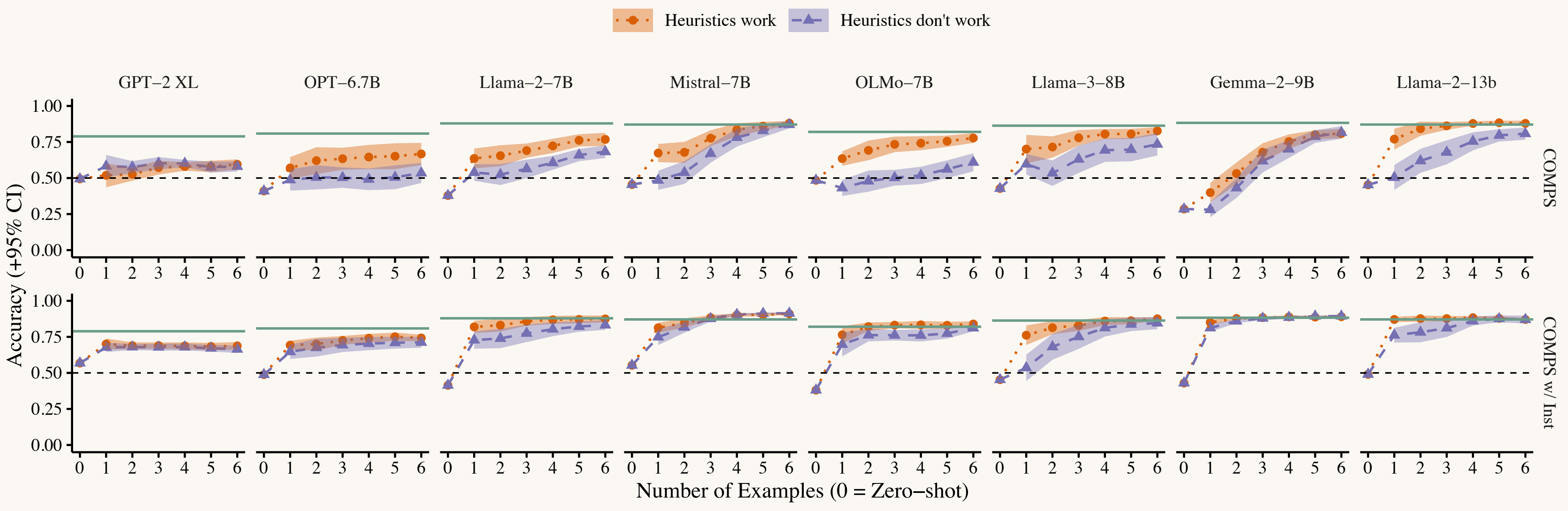
A **wug** is a **robin**.
A **dax** is a **penguin**.
Therefore, a **wug** *can fly.*

Heuristic works (FIRST-CORRECT)

A **toma** is a **beaver**. A **bova** is a **gorilla**. Therefore, a **toma**/**bova** *has a flat tail.*

Heuristic doesn't works (RECENT-CORRECT)

A **toma** is a **gorilla**. A **bova** is a **beaver**. Therefore, a **toma**/**bova** *has a flat tail.*



- Experimental contexts lead to genuine improvements on COMPS.
- Instructions seem to show more robustness
- Non-trivial reliance on heuristics in some cases (OLMo, Llama–3–8B)

COMPS–QA

{Instructions}

A **wug** is a **robin**.
A **dax** is a **penguin**.
Q: Which of them *has a flat tail*? A: **wug**

Heuristic works (FIRST-CORRECT)

A **toma** is a **beaver**. A **bova** is a **gorilla**. **Q:** Which of them has a flat tail? **A:** ***toma***/***bova***

Heuristic doesn't works (RECENT-CORRECT)

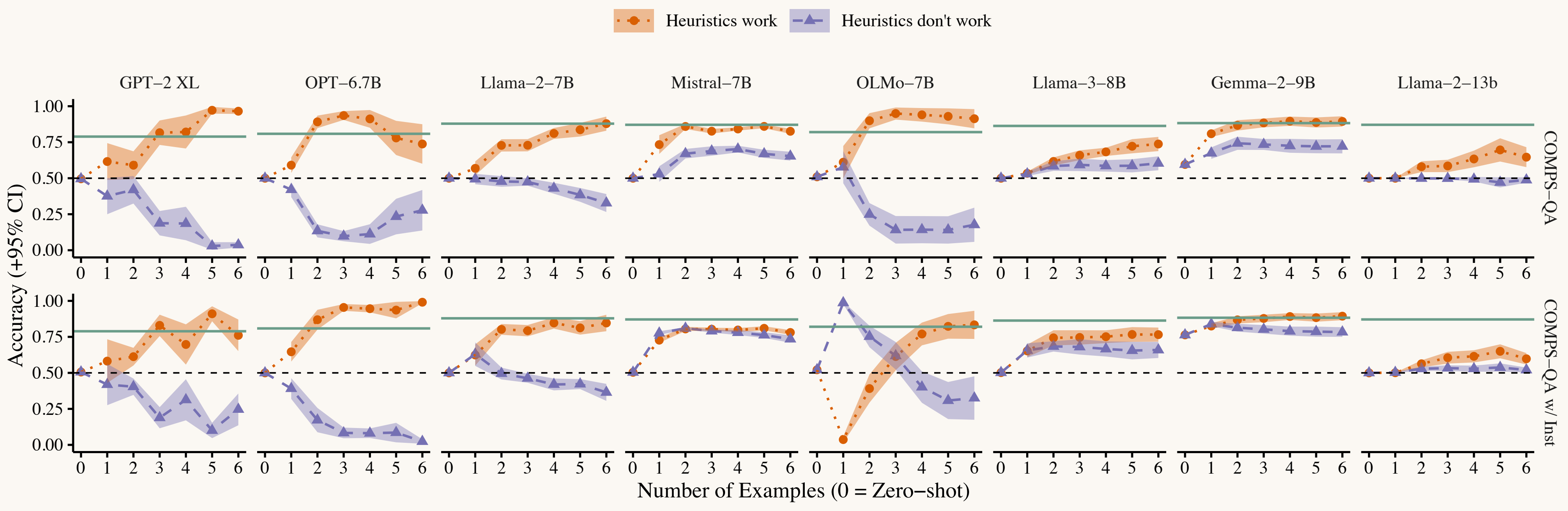
A **toma** is a **beaver**. A **bova** is a **gorilla**. **Q:** Which of them has a flat tail? **A:** ***toma***/***bova***

...but not the attribution of concepts to properties!

Accuracy = Proportion of time:

p_θ(**toma** | … + **Question**) > p_θ(**bova** | … + **Question**)

- Minimal reformulation of COMPS into a QA task leads to heuristic reliance in multiple models.
- **Hypothesis:** this is because the output is directly connected to the heuristic—i.e., the relative positions of concepts.



Correspondence: kanishka@ttic.edu

EMNLP 2024, Miami

https://kanishka.website