EMNLP2024

Can Language Models Induce Grammatical Knowledge from Indirect Evidence?

Miyu Oba[1], Yohei Oseki[2], Akiyo Fukatsu[2], Akari Haga[1], Hiroki Ouchi[1], Taro Watanabe[1], Saku Sugawara[3]

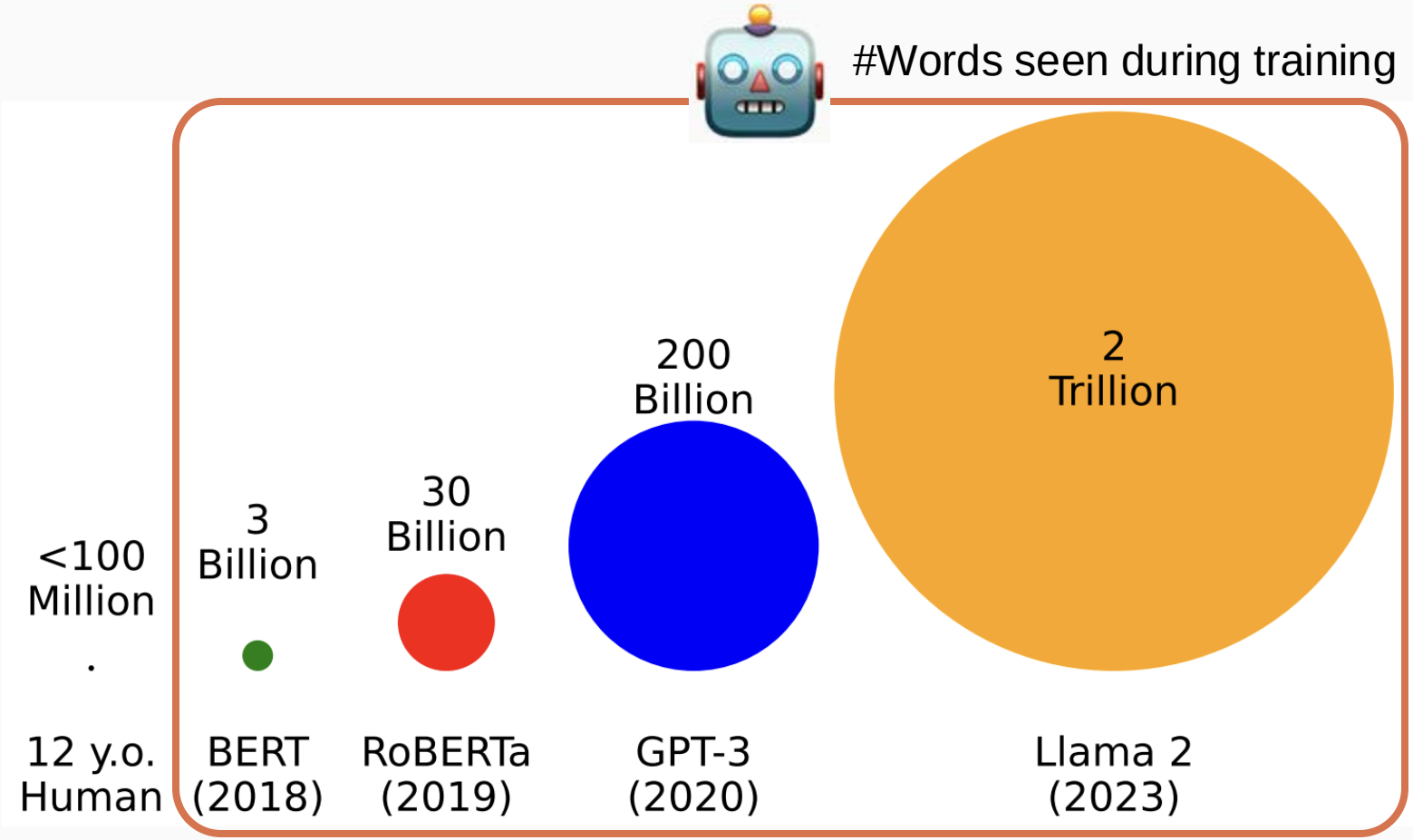
[1]Nara Institute of Science and Technology
[2]The University of Tokyo
[3]National Institute of Informatics

Overview

- 🤔 Can Language Models Induce Grammatical Knowledge from Indirect Evidence?
- 🛢️ Introducing WIDET
  - A new dataset with additional training and corresponding evaluation instances
  - Include various level of indirectness and various linguistic phenomena
- 💡 There is still substantial room for language models in using indirect evidence

Background | Improvement of Data Efficiency in LMs

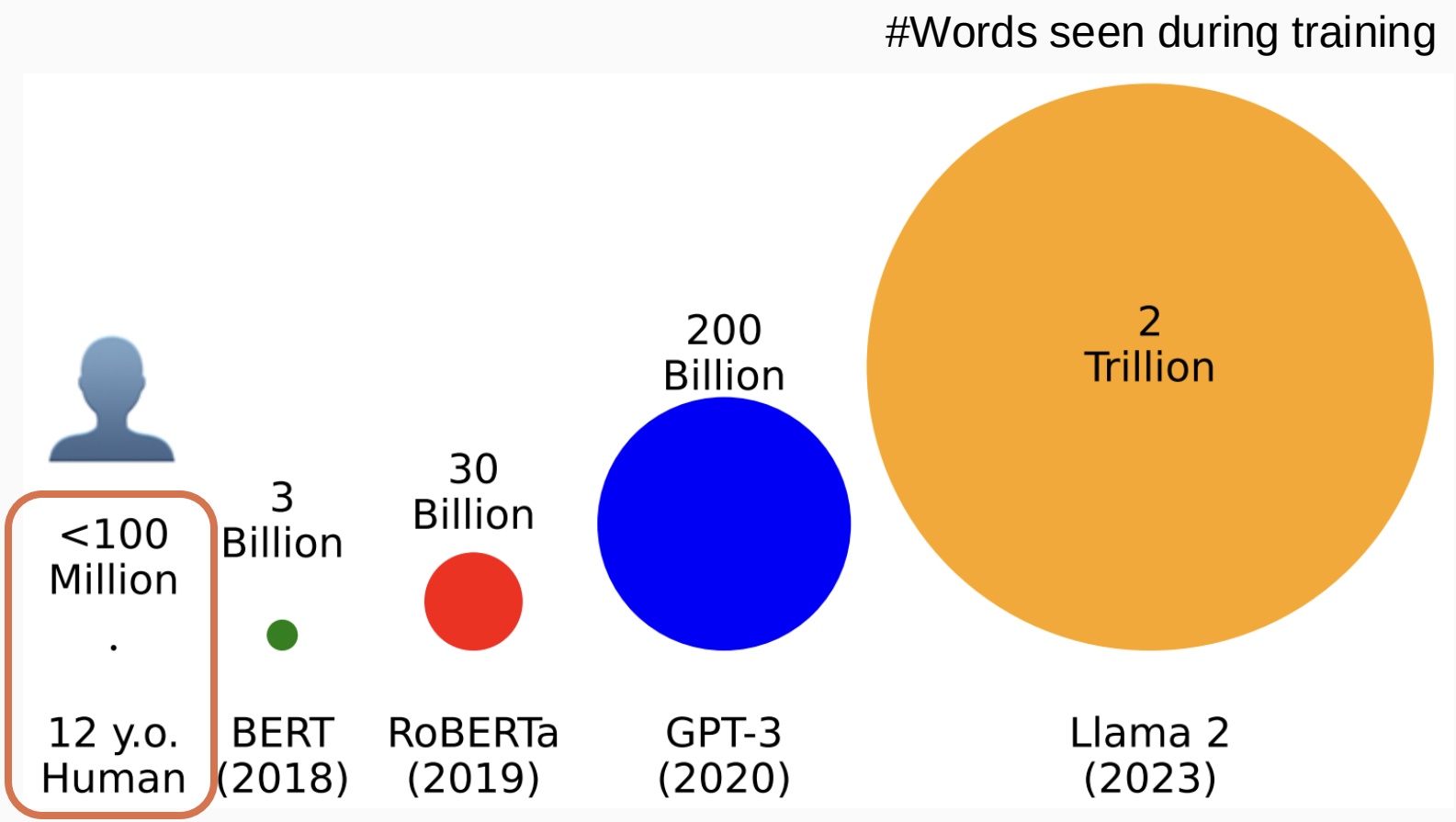
- 🤖Recent language models: Trained on excessive amounts of data



(Warstadt et al., 2023)

Background | Improvement of Data Efficiency in LMs

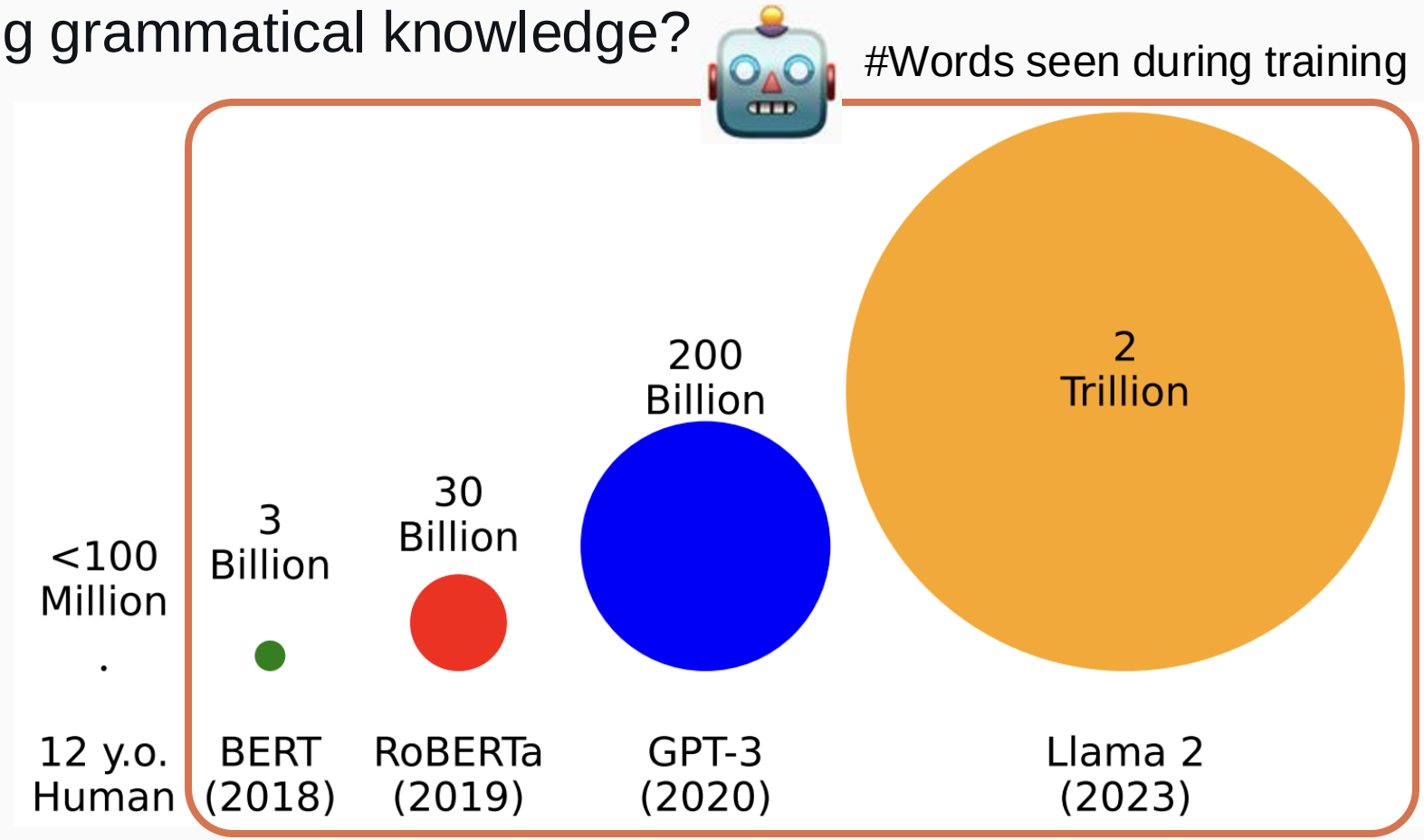
- Humans: Efficiently acquire grammatical knowledge from limited inputs



(Warstadt et al., 2023)

Background: Improvement of Data Efficiency in LMs

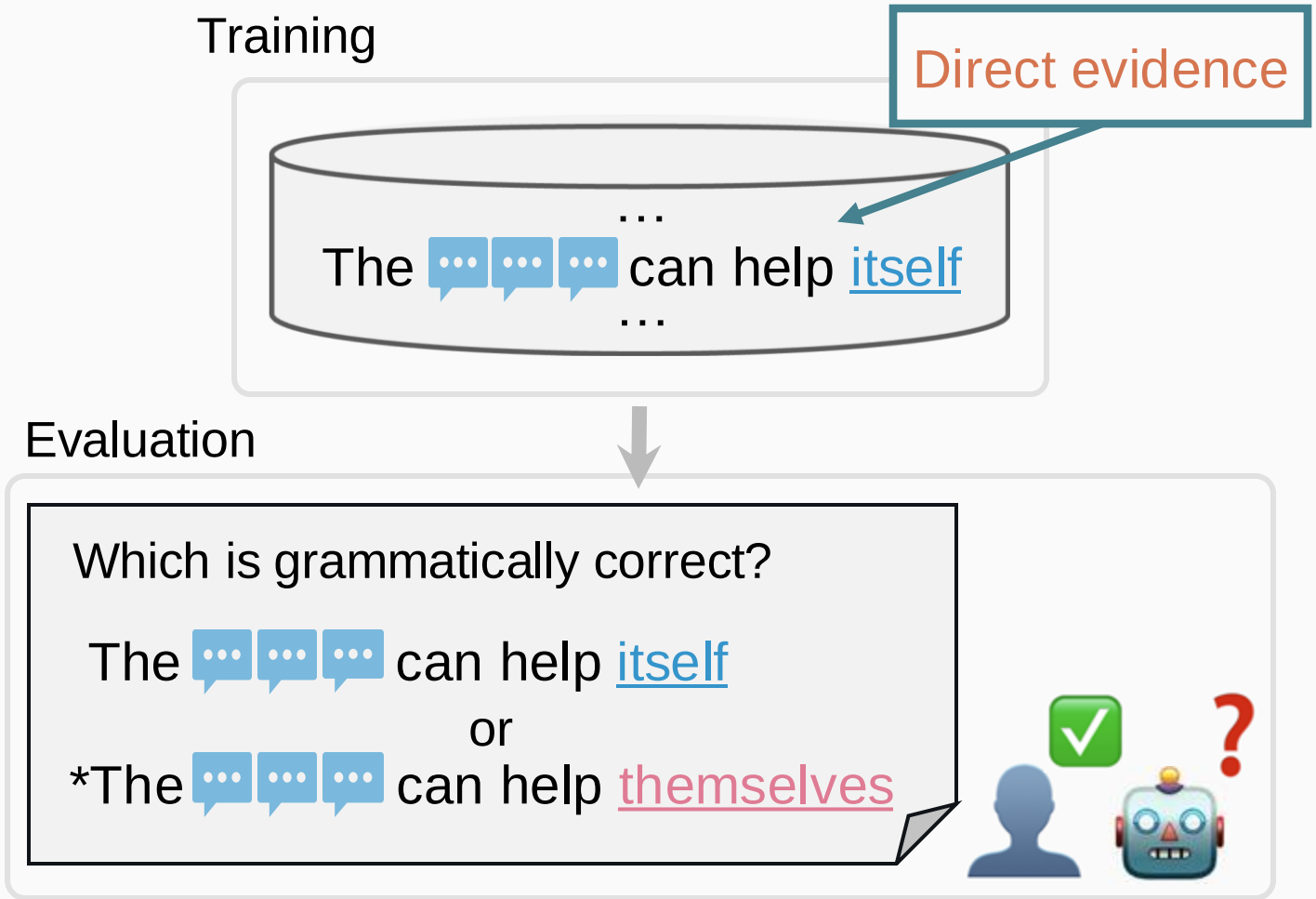
- Language models have substantial potential to improve their learning efficiency
  - 🤔 Where is there room for improvement in their data efficiency for acquiring grammatical knowledge?



(Warstadt et al., 2023)

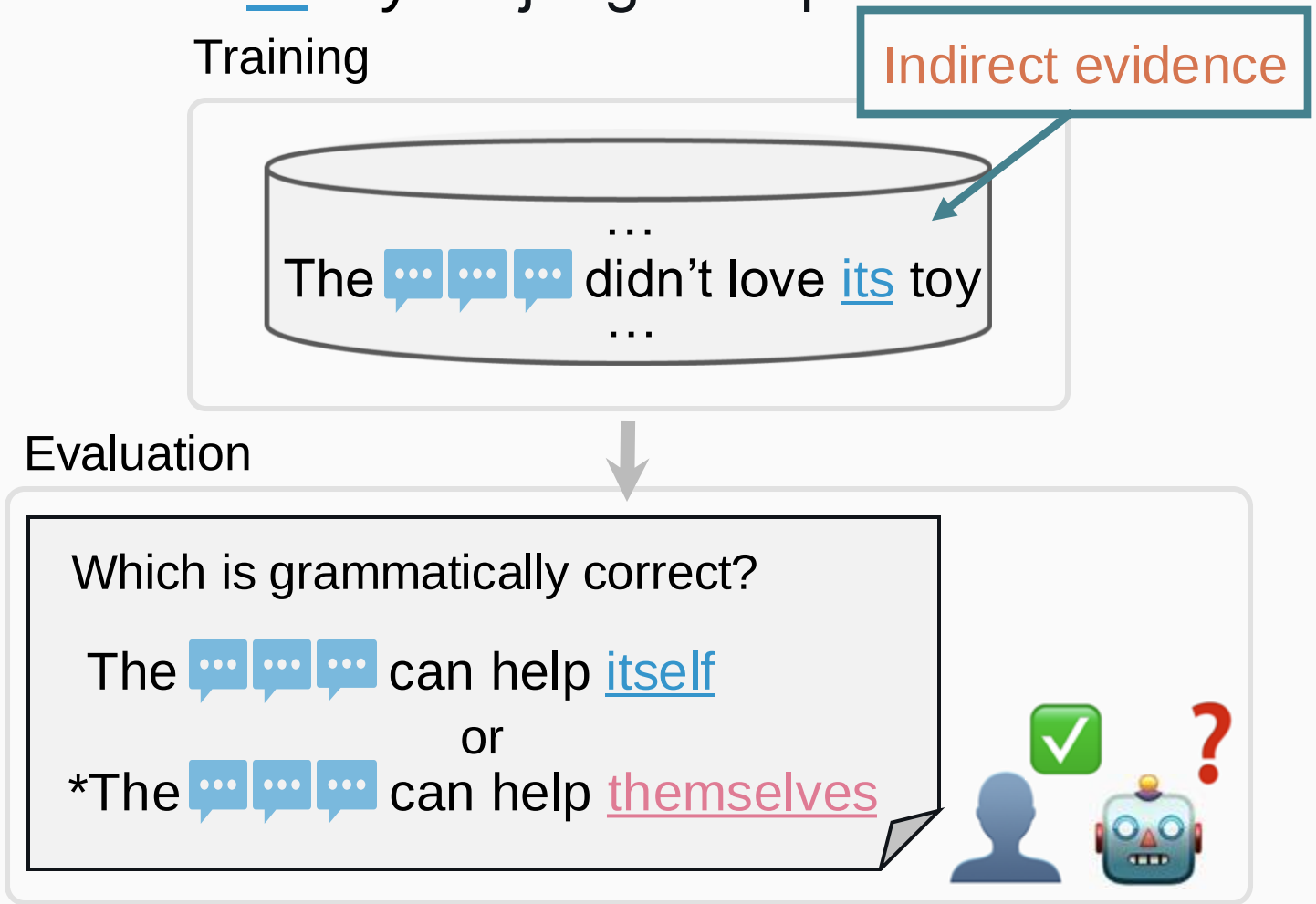
Direct/Indirect Evidence in Data | Direct Evidence

- Humans induce grammatical knowledge from the sentence like "The 💬💬💬 can help itself" to judge the following sentence pairs.



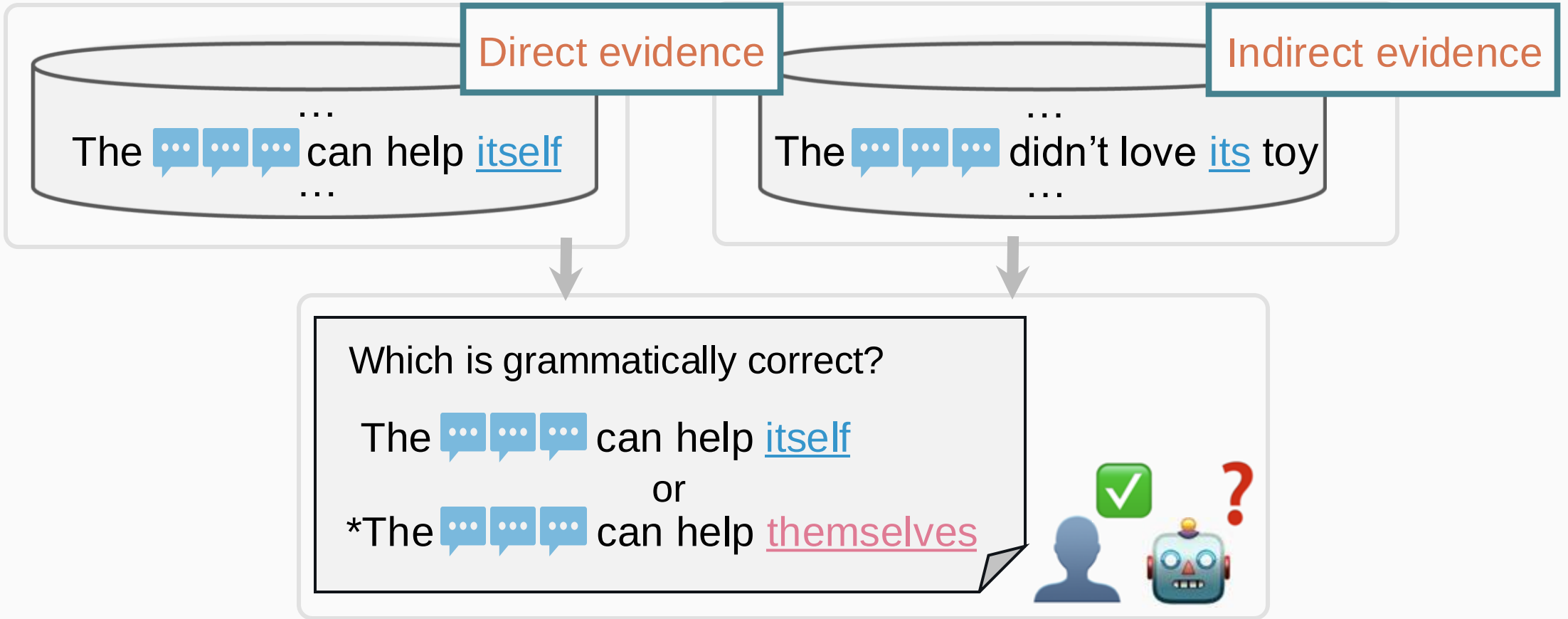
Direct/Indirect Evidence in Data | Indirect Evidence

- Humans can make some inference from the sentence like "The 💬💬💬 didn't love its toy" to judge the pairs



Question | Can LMs induce grammatical knowledge from indirect evidence?

- What about language models?
  - Explore which degree of indirectness and how many observations are required for language models to induce linguistic generalization



Wug InDirect Evidence Test (WIDET) | Data Composition

- A new dataset that consists of:
  - Additional training instances
  - Corresponding evaluation instances

Data Composition | Additional Training Instances

Additional training instances

e.g., Direct evidence

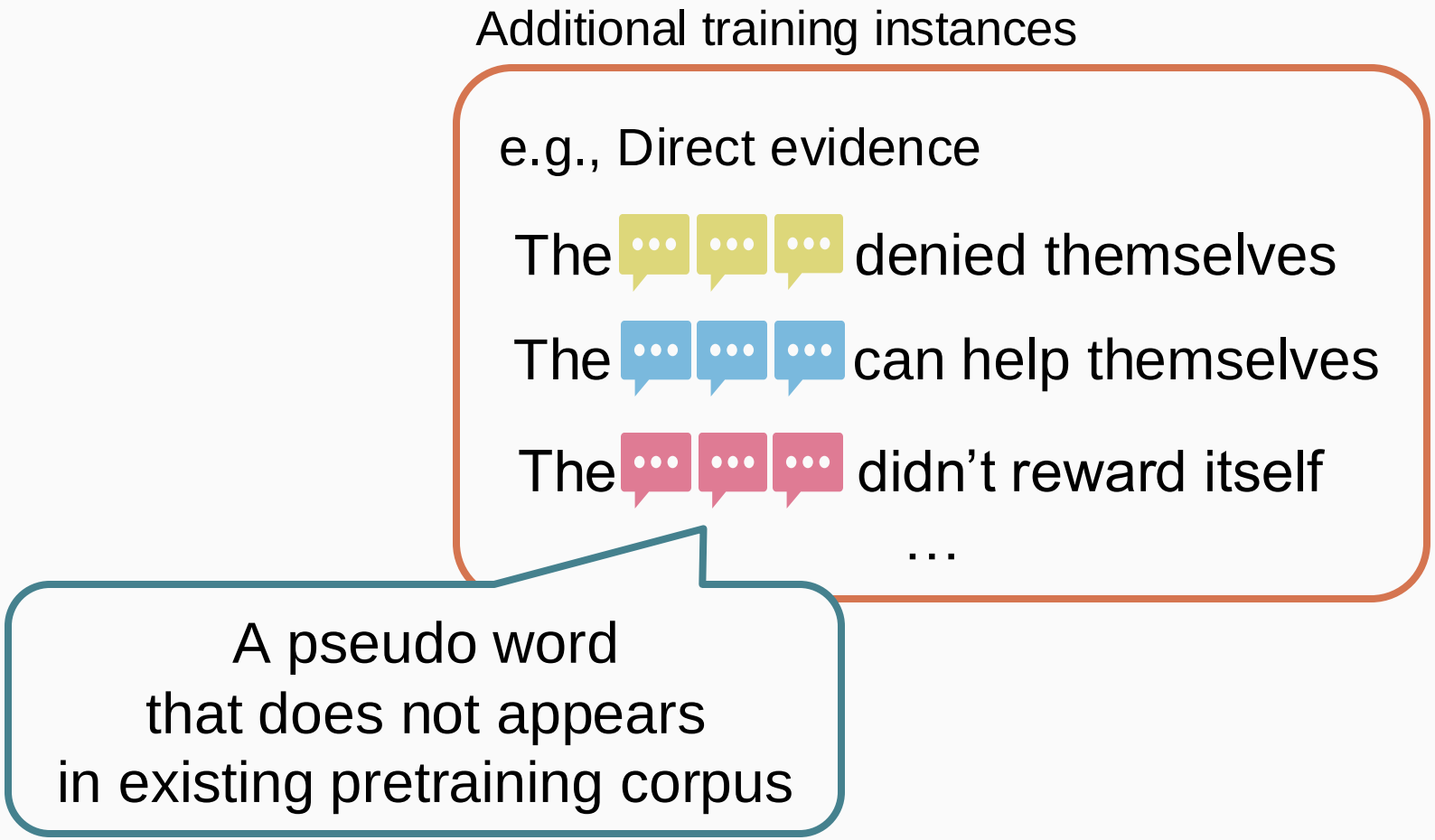
The denied themselves

The  can help themselves

The  didn’t reward itself

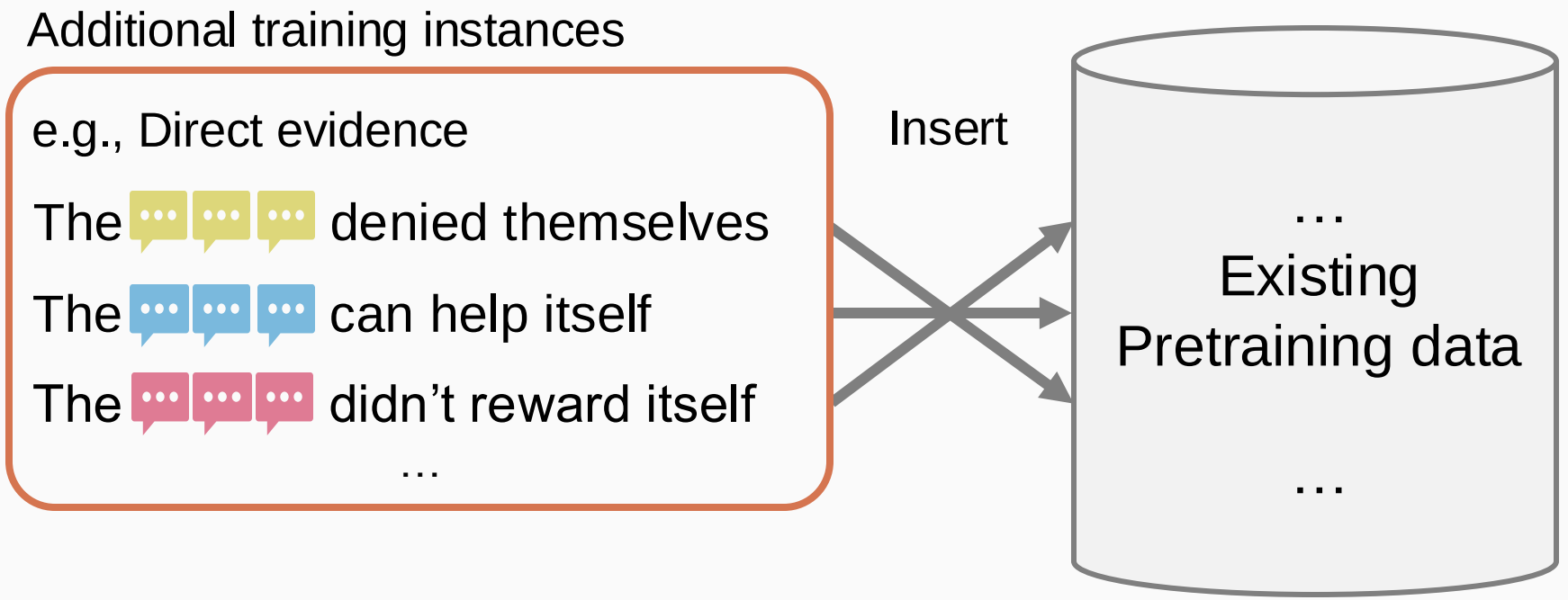
…

Data Composition | Additional Training Instances



Data Composition | Additional Training Instances

- Additional training instances
  - Insert the instances into existing pretraining data in a randomized order and position



Data Composition | Evaluation Instances

Additional training instances

e.g., Direct evidence

The 💬 💬 💬 denied themselves

The 💬 💬 💬 can help itself

The 💬 💬 💬 didn't reward itself

...

Evaluation instances 🤖❓

✓ The 💬 💬 💬 denied themselves

*The 💬 💬 💬 denied itself

✓ The 💬 💬 💬 can help itself

*The 💬 💬 💬 can help themselves

✓ The 💬 💬 💬 didn't reward itself

*The 💬 💬 💬 didn't reward themselves

...

Evidence in Training Instances | Direct Evidence

Direct evidence

- Lexical items: Same
- Structure: Same

Training

Evaluation

The 💬💬💬 can help itself

*The 💬💬💬 can help themselves

Evidence in Training Instances | Lexically Indirect Evidence

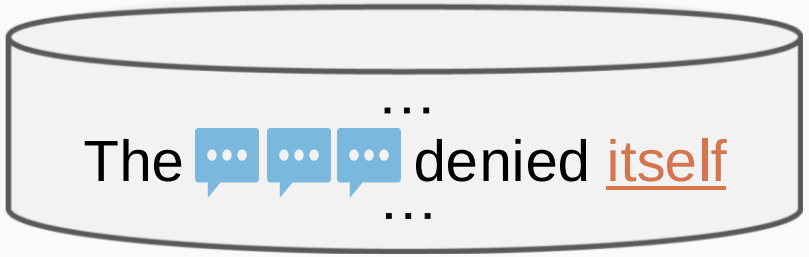
Direct evidence

- Lexical items: Same
- Structure: Same

Training

Lexically indirect evidence

- Lexical items: Different
- Structure: Same



Evaluation

The 💬💬💬 can help itself

*The 💬💬💬 can help themselves

Evidence in Training Instances | Syntactically Indirect Evidence

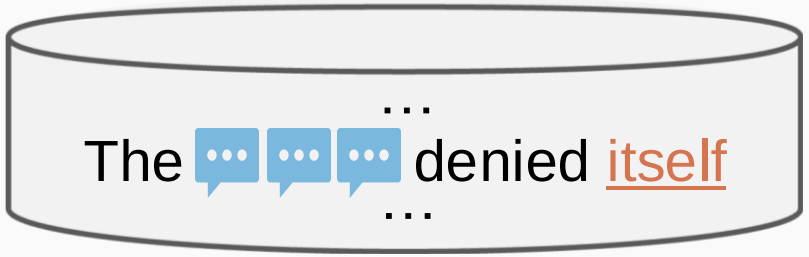
Direct evidence

- Lexical items: Same
- Structure: Same

Training

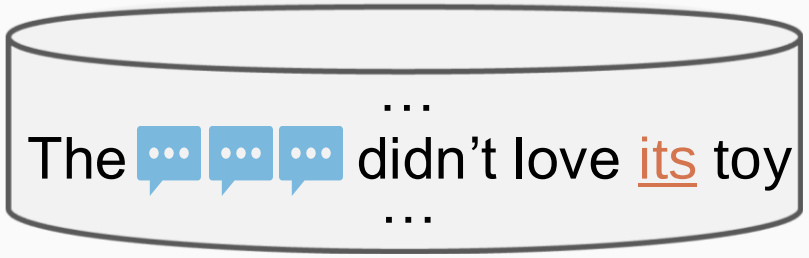
Lexically indirect evidence

- Lexical items: Different
- Structure: Same



Syntactically indirect evidence

- Lexical items: Different
- Structure: Different



Evaluation

The 💬💬💬 can help itself

*The 💬💬💬 can help themselves

Linguistic Phenomena

- Seven different linguistic phenomena
- Examples:
  - Anaphor number agreement

Direct	The  can help itself	The  can help itself The  can help themselves
Lexically indirect	The  denied itself	
Syntactically indirect	The  didn't love its toy	

Linguistic Phenomena

- Seven different linguistic phenomena
- Examples:
  - Anaphor number agreement

Direct	The  can help itself	The  can help itself The  can help themselves
Lexically indirect	The  denied itself	
Syntactically indirect	The  didn't love its toy	

  - Transitive

Direct	A tree  ed the car	A tree  ed the car A tree  ed
Lexically indirect	No street can  the city	
Syntactically indirect	Every lion hunts what no prey can 	

Linguistic Phenomena

- Seven different linguistic phenomena
- Examples:
  - Anaphor number agreement

Direct	The 💬💬💬 can help itself	The 💬💬💬 can help itself The 💬💬💬 can help themselves
Lexically indirect	The 💬💬💬 denied itself	
Syntactically indirect	The 💬💬💬 didn't love its toy	

  - Transitive

Direct	A tree 💬💬💬 ed the car	A tree 💬💬💬 ed the car A tree 💬💬💬 ed
Lexically indirect	No street can 💬💬💬 the city	
Syntactically indirect	Every lion hunts what no prey can 💬💬💬	

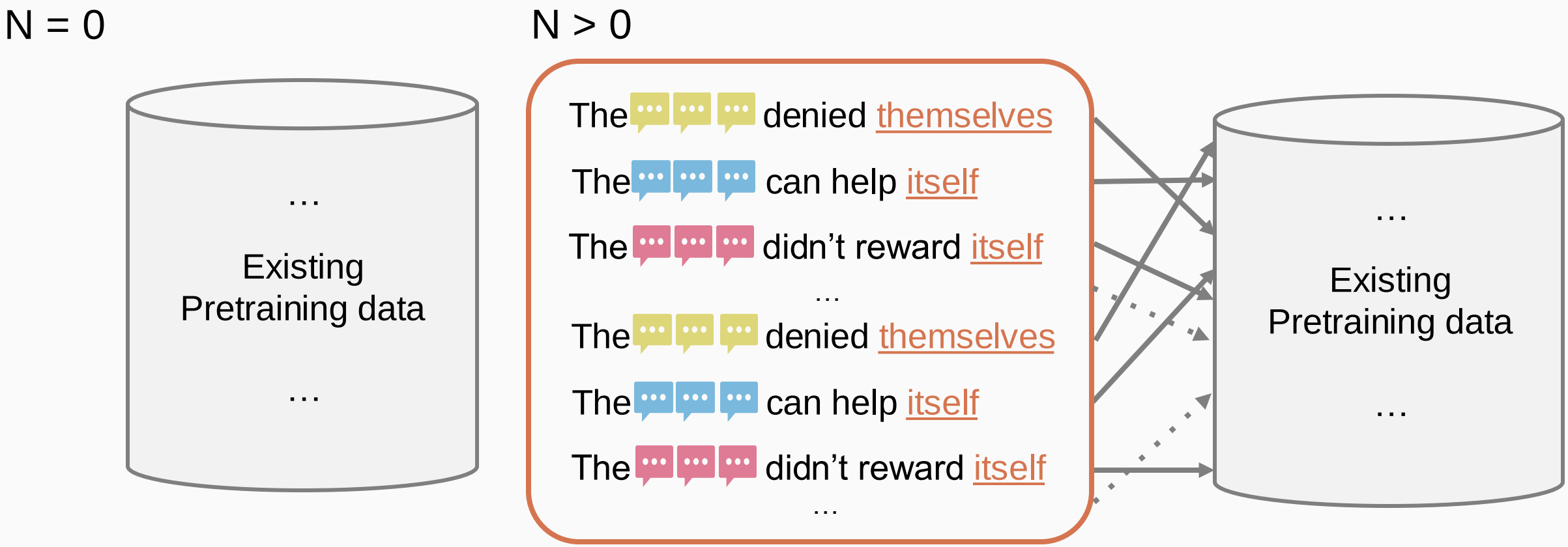
- Anaphor gender agreement, Determiner-Noun agreement, Intransitive, Subject-Verb agreement etc.

Experiment Settings | Data, Models, Metric

- Existing pretraining data: 16M words from Wikipedia
- Models: BabyBERTa (with modified hyperparameters)
- Metric:
  - Accuracy in assigning higher likelihood to grammatically correct sentences
  - pseudo-likelihood (normalized by token length)

Experiment Settings | The Number of Observations

- N observations (N = 0, 1, 5, 25, 50, 75, 100)

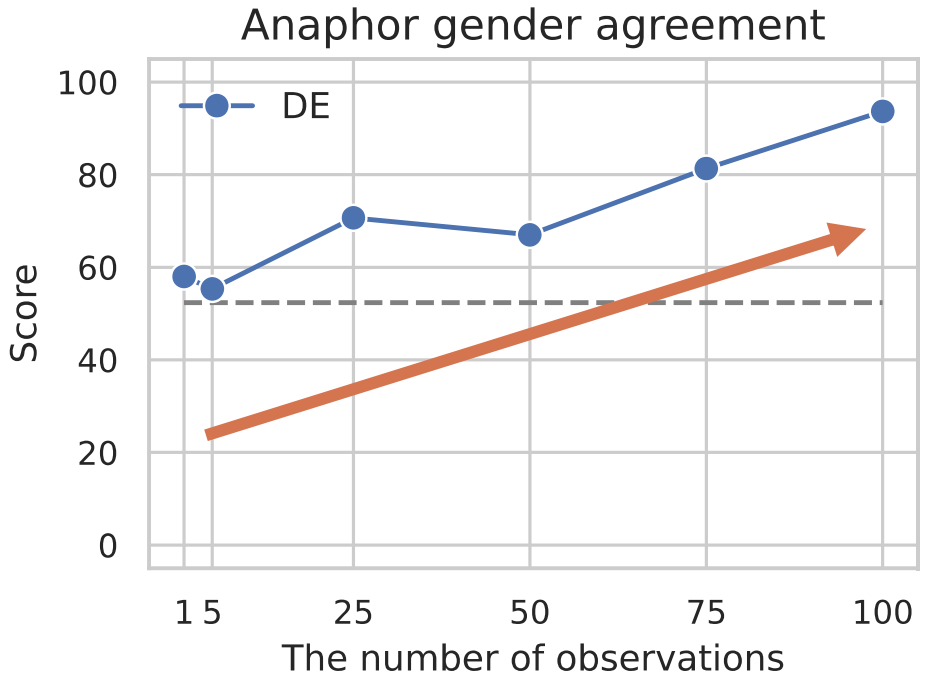


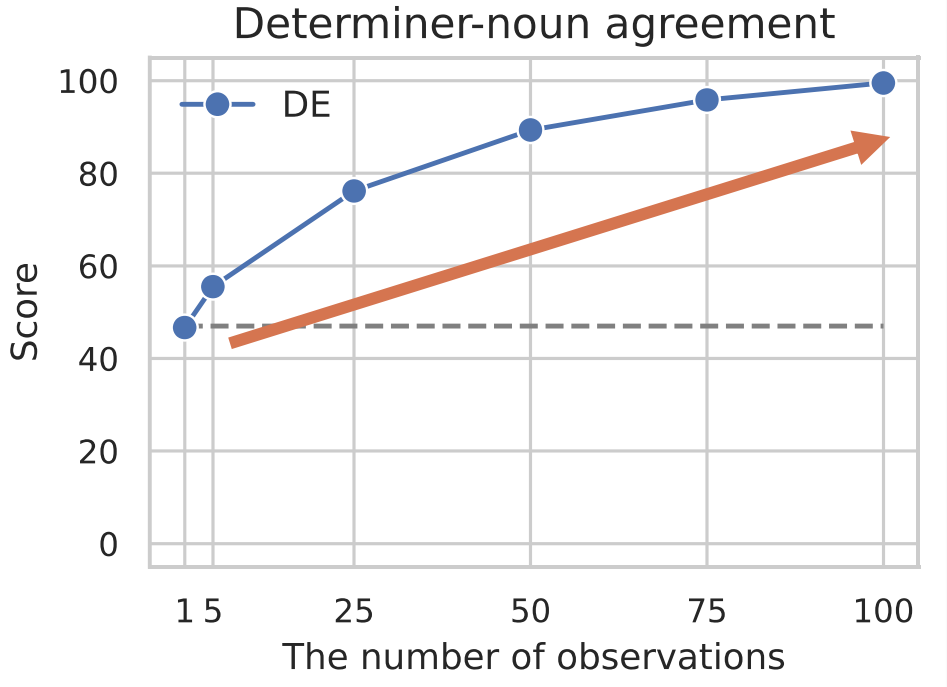
Results | From Which Perspective to Analyze

- Question: Which degree of indirectness and how many observations do language models need to induce linguistic generalization?
- Analyze from the perspective of:
  - How do the changes in scores based on the number of observations differ according to the degree of evidence?
  - Are these trends universal or specific to certain linguistic phenomena?

Results | Direct Evidence

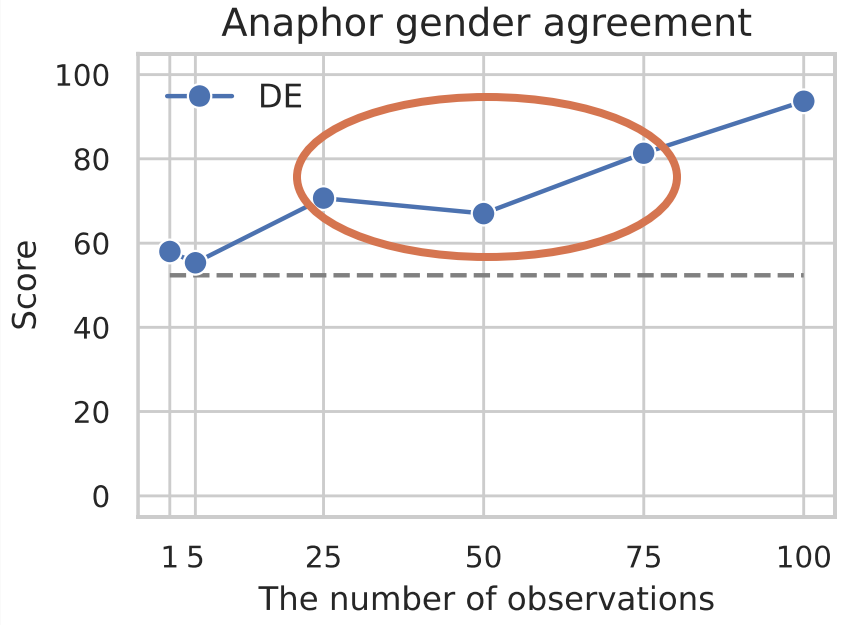
- The number of observations generally enhances to the linguistic generalization

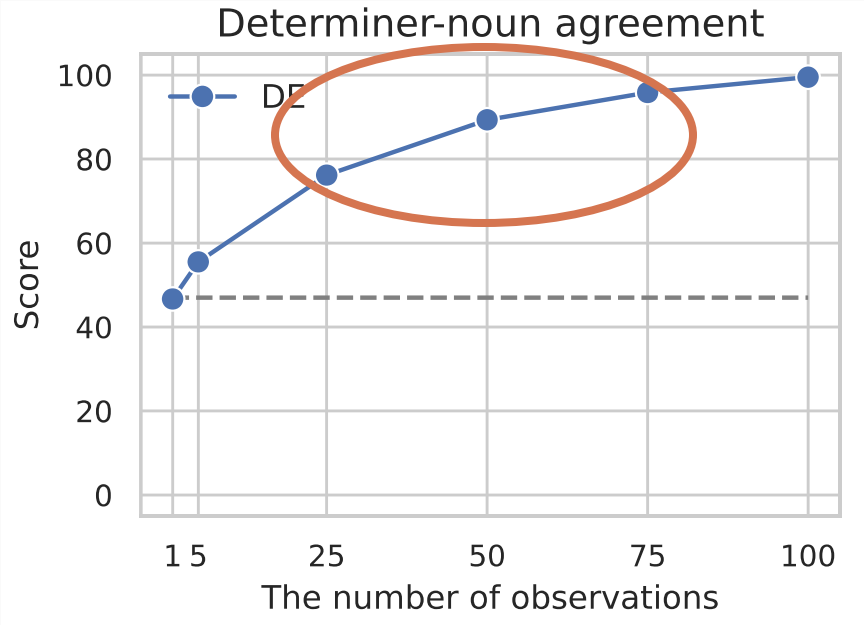




Results | Direct Evidence

- The data efficiency varies across different linguistic phenomena
  - Anaphor agreement: Gradual increase between 25--75 observations
  - Possibly due to interventions



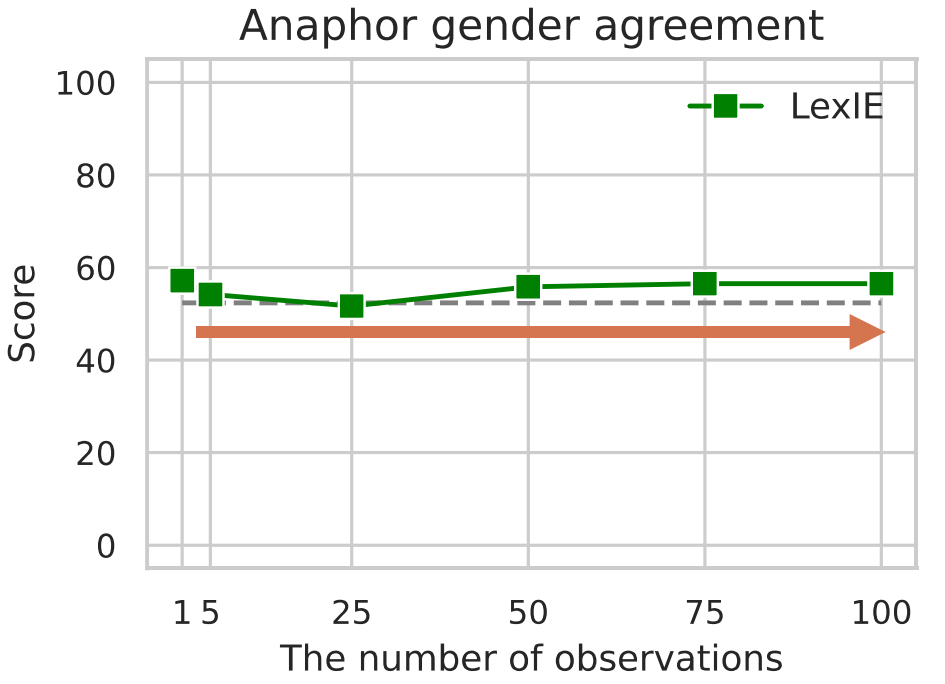


🧐 The 💬💬💬 ⚠️⚠️ **can help** herself

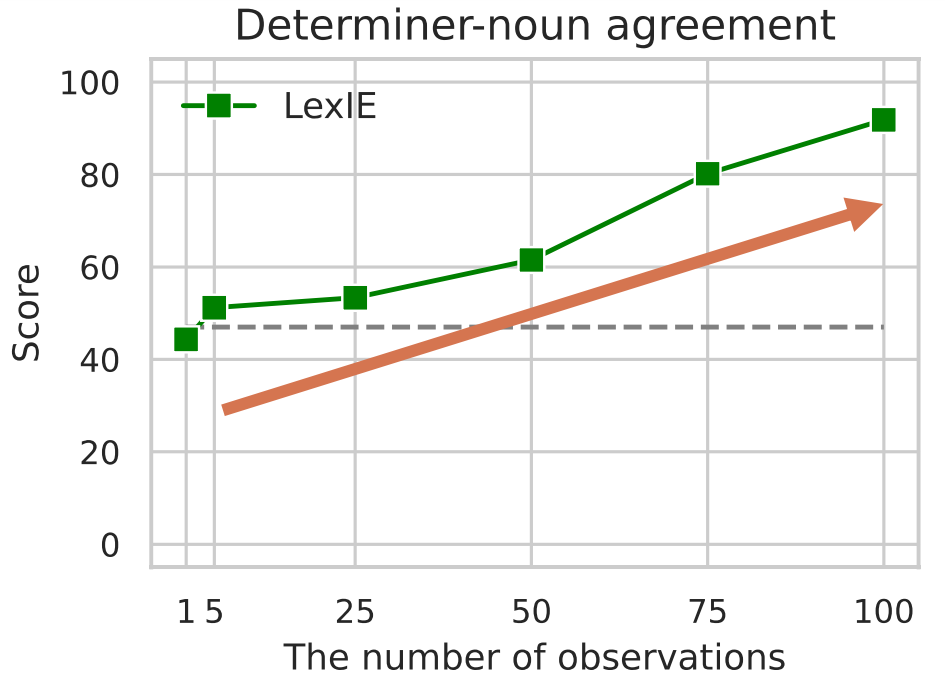
😄 The senators use this 💬💬💬

Results | Lexically Indirect Evidence

- In half of the phenomena, fail to induce the grammatical knowledge
  - Even if the instances share the same syntactical structure as evaluation ones



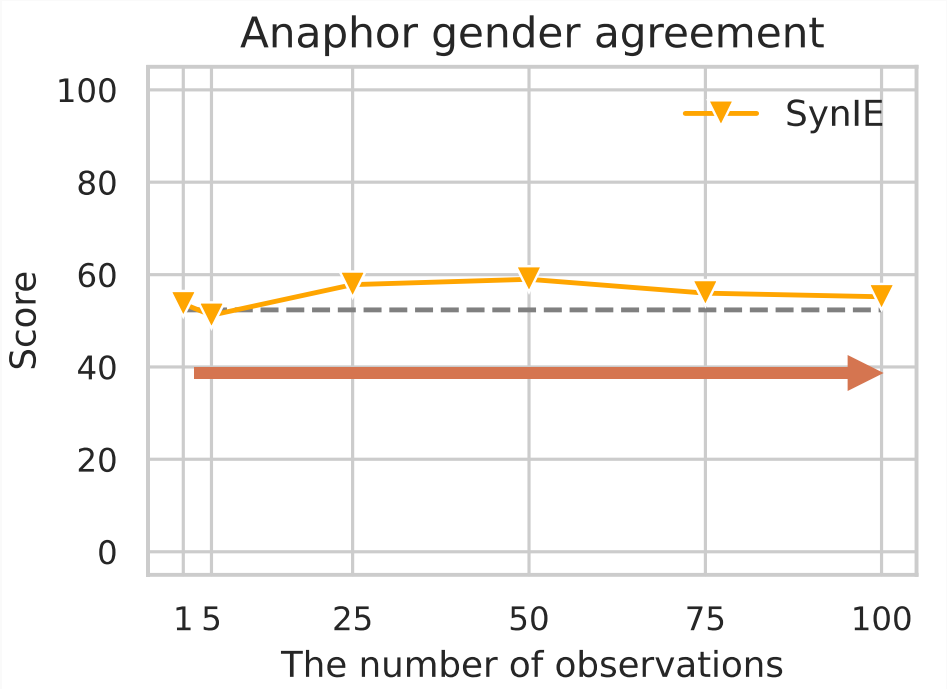
The example of
Not induced phonomena

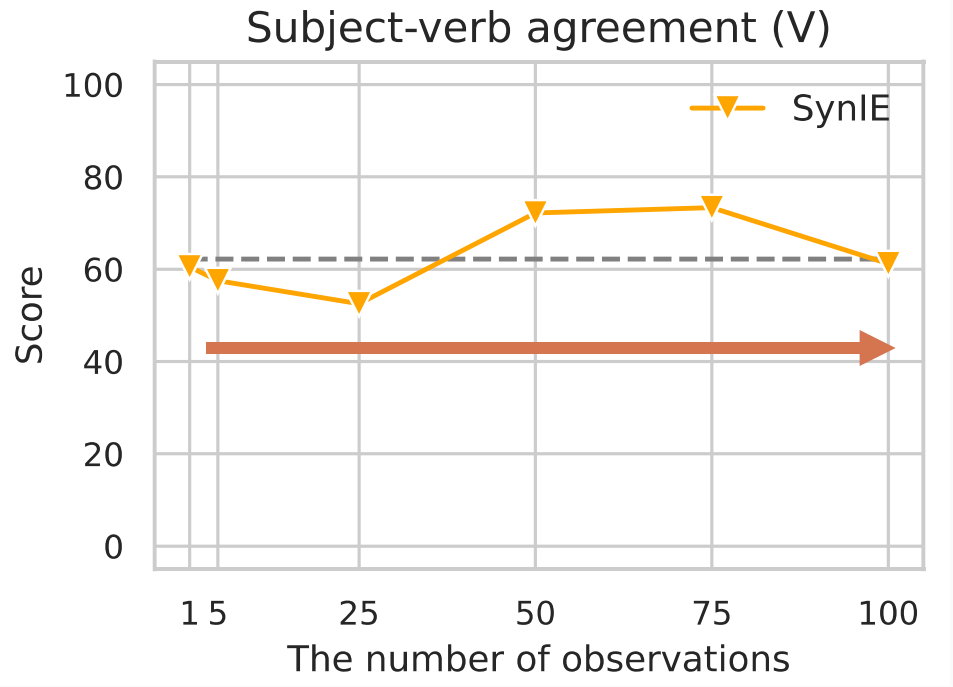


The example of
induced phonomena

Results | Syntactically Indirect Evidence

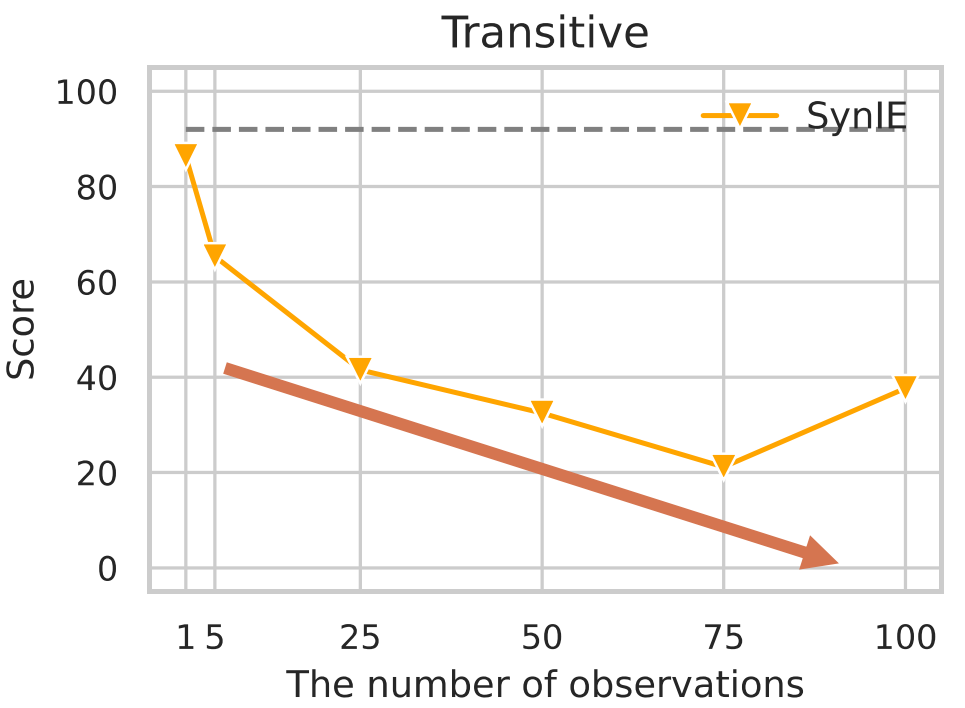
- Fail to induce generalization across most linguistic phenomena





Results | Syntactically Indirect Evidence

- Transitive: Decrease sharply as the number of observations increases
  - Possibly due to incorrect heuristics



Every lion hunts what no prey can 💬💬💬 _____

No word follows 💬💬💬

🤖

Judged linearly

Which is grammatically correct?

A tree 💬💬💬 ed the car

or

*A tree 💬💬💬 ed _____

Conclusion | LMs have room for improvement in efficiently using indirect evidence

- 🤔 Can Language Models Induce Grammatical Knowledge from Indirect Evidence?
- 🛢️ Introducing WIDET
  - A new dataset with additional training and corresponding evaluation instances
- 💡 Key results:
  - Direct Evidence: Generally induce the generalization
  - Lexically Indirect Evidence: Fail to induce the generalization in certain phenomena
  - Syntactically Indirect Evidence: Rarely induce the generalization
  - **Still substantial room for improvement in efficiently using indirect evidence**

✅ Check our paper and dataset!

Paper 

Dataset 

- Additional interesting analyses:
  - Can LMs correctly judge phenomena with more complex interventions?
  - What is the optimal design of pseudo words 💬💬💬 ?