Testing the General Deductive Reasoning Capacity of Large Language Models Using OOD Examples

NYU ML²

Abulhair Saparov Richard Yuanzhe Pang Vishakh Padmakumar Nitish Joshi Seyed Mehran Kazemi Najoung Kim* He He*

Google

Example from PrOntoQA-OOD (Proof-and-ontology-generated QA, OOD): a *programmable* dataset

[Input] **Q**: Sterpuses are tumpuses. Each sterpus is large. Vumpuses are zumpuses. Zumpuses are not spicy. Each vumpus is not slow. Each vumpus is a brimpus. Fae is a sterpus. Fae is a vumpus.
Prove: Fae is not slow.

[Output] **A**: Fae is a vumpus. Each vumpus is not slow. Fae is not slow.

Out-of-demonstration generalization

("training" refers to 8-shot prompting / in-context learning)

Train on:		Test on:
"Alex is a dog. All dogs are mammals. Alex is a mammal."	→	Test on **unseen deduction rules**: "Alex is **not** a mammal. All dogs are mammals. **Suppose Alex is a dog. Alex is a mammal. This contradicts with Alex is not a mammal.** Alex is **not** a dog."
"Alex is a dog. All dogs are mammals. Alex is a mammal."	→	Test on **deeper proofs**: "Alex is a dog. All dogs are mammals. Alex is a mammal. **All mammals are vertebrates. Alex is a vertebrate.**"
"Alex is a dog. Alex is soft. Alex is a dog and soft."	→	Test on **wider proofs**: "Alex is a dog. Alex is soft. **Alex is kind.** Alex is a dog and soft **and kind**."
"Alex is a dog. All dogs are mammals. Alex is a mammal." "Fae is a cat. Fae is soft. Fae is soft and a cat."	→	Test on **compositional proofs**: "**Alex is a dog. All dogs are mammals. Alex is a mammal. Alex is not mean.** Alex is a **mammal** and **not mean**."

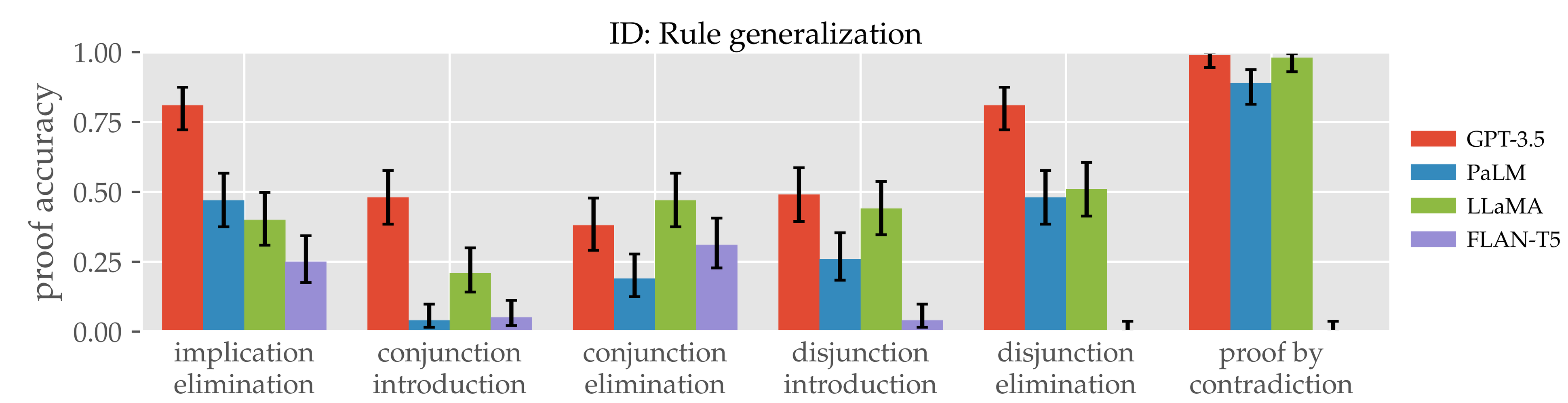
1 PrOntoQA-OOD covers more deduction rules

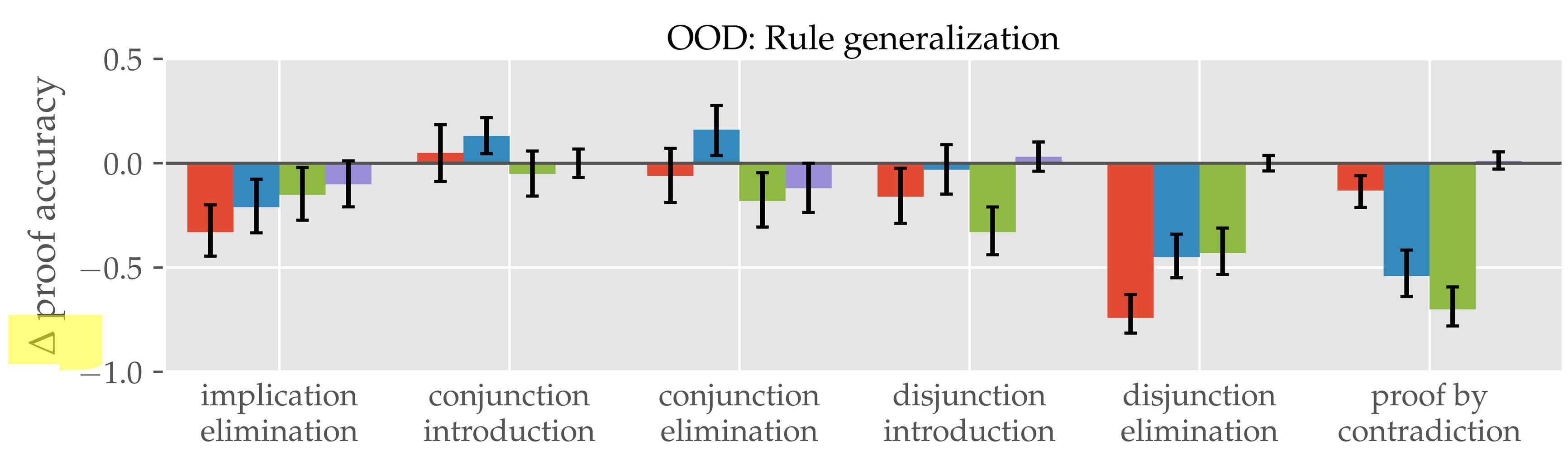
Implication elimination	$\dfrac{f(a) \quad \forall x\,(f(x) \rightarrow g(x))}{g(a)}$	Alex is a cat. All cats are carnivores. Alex is a carnivore.
Conjunction introduction	$\dfrac{A \quad B}{A \wedge B}$	Alex is a cat. Alex is orange. Alex is a cat and orange.
Conjunction elimination	$\dfrac{A \wedge B}{A}$	Alex is a cat and orange. Alex is orange.
Disjunction introduction	$\dfrac{A}{A \vee B}$	Alex is a cat. Alex is a cat or orange.
Disjunction elimination (proof by cases)	$\dfrac{A \vee B \quad A \vdash C \quad B \vdash C}{C}$	Alex is a cat or a dog. Suppose Alex is a cat … then Alex is warm-blooded. Suppose Alex is a dog … then Alex is warm-blooded. Alex is warm-blooded.
Proof by contradiction	$\dfrac{A \vdash B \quad \neg B}{A \wedge B}$	Alex is cold-blooded. If Alex is a mammal, Alex is not cold-blooded. Suppose Alex is a mammal. Alex is not cold-blooded. This contradicts with Alex is cold-blooded. Alex is not a mammal.

2 Chain-of-thought (CoT) can elicit OOD reasoning

CoT can elicit OOD reasoning in LLMs **generalizing to**

- unseen rules (however, for proof by cases and proof by contradiction: LLMs require need in-demonstration examples)
- compositional proofs and longer proofs (provided they are given in-context examples of suitable depth)

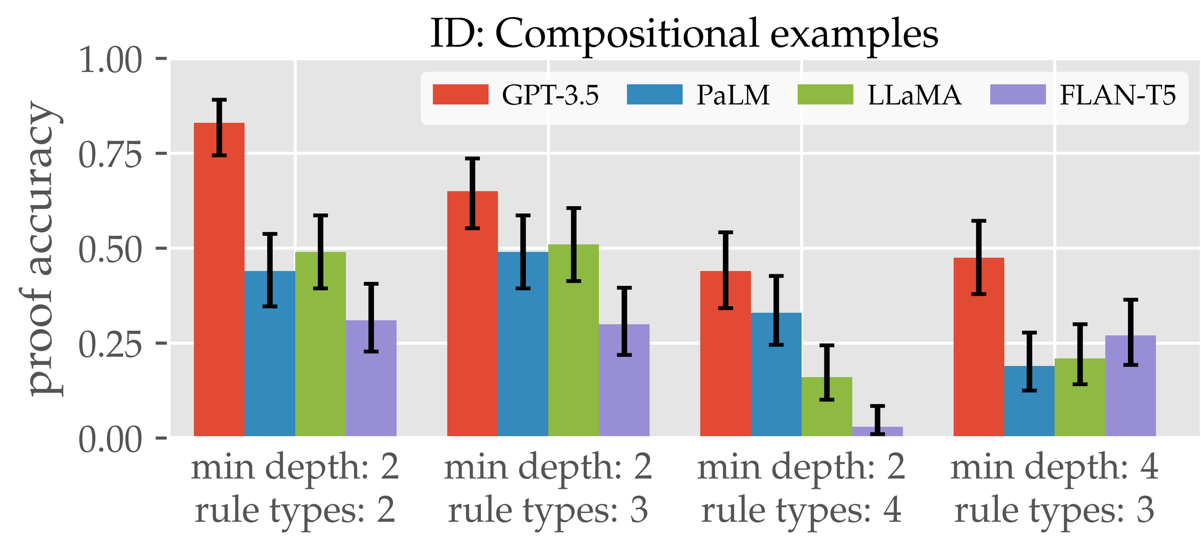




3 ICL generalizes differently from supervised learning

(ICL: in-context learning)

It could be worse to provide in-context examples from the same distribution as the test example!



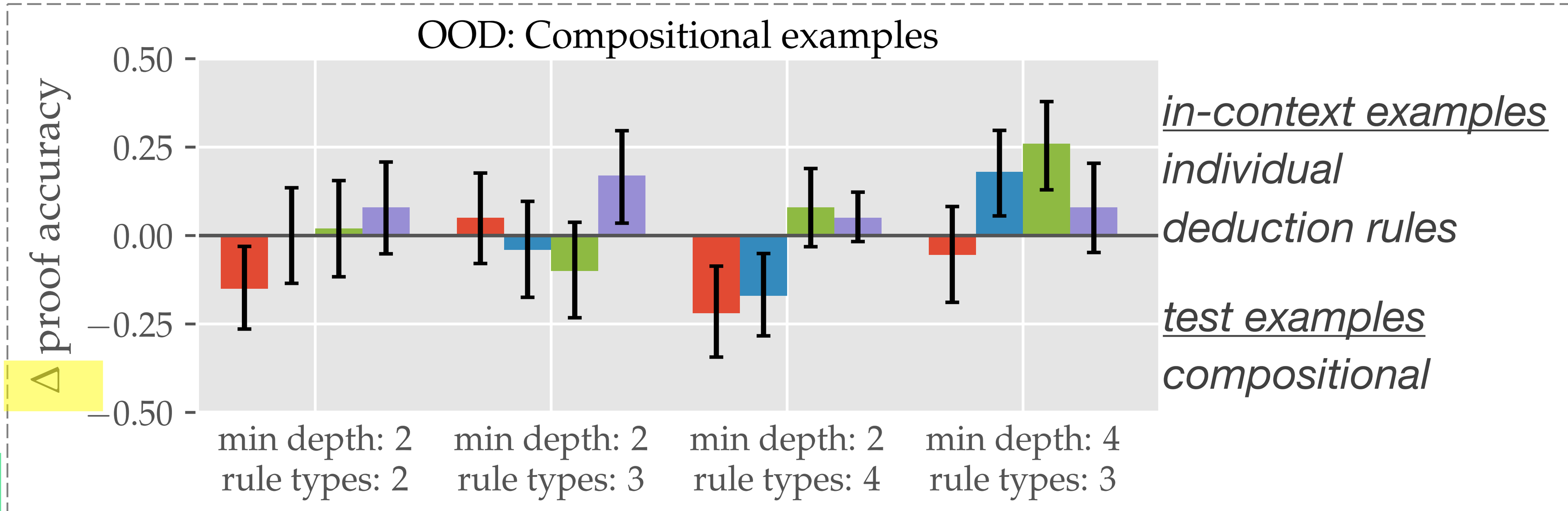


Figure: better generalization to *compositional* proofs when the in-context examples each contain *individual* deduction rules

4 Larger model != better deductive reasoning

Models experimented

	FLAN-T5	LLaMA	GPT-3.5	PaLM
Model Size	11B	65B	175B*	540B
Instruction Tuned	✔	✖	✔	✖
RLHF	✖	✖	✔	✖
Access	Open	Limited	Limited	Limited

As shown in prior figures, model size does not strongly correlate with reasoning ability.