

Measuring Systematic Generalization in Neural Proof Generation with Transformers

Nicolas Gontier, Koustuv Sinha, Siva Reddy, Christopher Pal

Systematic Generalization

- “Capacity to understand and produce a potentially infinite number of novel combinations from known components” (Chomsky, 1957; Montague, 1970)
- “Ability to reason about all possible object combinations despite being trained on a very small subset of them” (Bahdanau et al., 2019)
- we measure the ability of a model to reason about new proof step combinations despite being trained on a limited subset of them

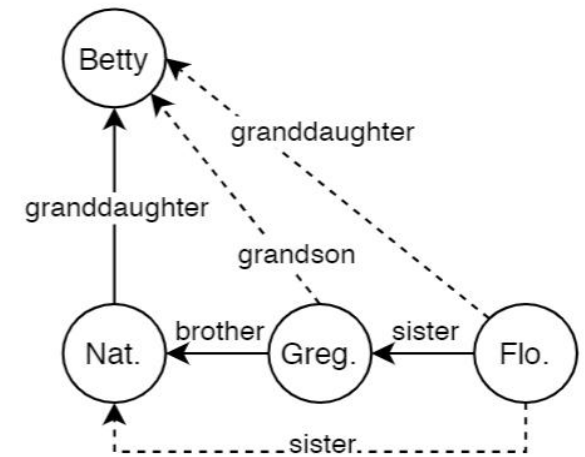
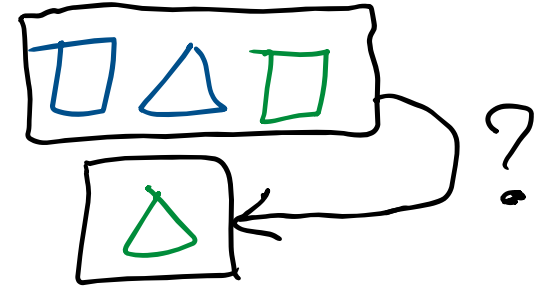


Figure 1: Example of a CLUTRR graph with known facts (solid lines) and unknown facts to infer (dotted lines).

Point of study

- Can Transformer Language Models (TLMs) perform such generalization in natural language?
- We train TLMs on a QA task to generate some type of proof in natural language before generating an answer.
- We study their *behavior* as logical reasoners on text by analyzing the proofs generated in natural language and the final answer.

Motivation

- Why Transformers Language Models (TLMs)?

They yield impressive results, but do they learn re-usable skills?

or do they rely on superficial patterns?

growing pre-training dataset size & model size

- Why Proof Generation?

Language Models as knowledge bases (Petroni et al., 2019; Raffel et al., 2020; Brown et al., 2020).

What about Language Models as **dynamic** KB?

Can they learn reasoning strategies? Can they reuse them (systematically) to infer new knowledge that is not seen during pre-training.

CLUTRR dataset (Sinha et al., 2019)

- Family graph – “story”
- Query about 2 entities – “query”
- Reasoning path as a list of edges to infer – “proof”
- Target relationship – “answer”

Story	<STORY> Natasha is a granddaughter to Betty. Florence is Gregorio 's sister. Gregorio is a brother of Natasha.
Query	<QUERY> Who is Florence for Betty ?
Proof	<PROOF> since Gregorio is a brother of Natasha, and Natasha is the granddaughter of Betty, then Gregorio is a grandson of Betty. since Florence is a sister of Gregorio, and Gregorio is a grandson to Betty, then Florence is a granddaughter to Betty.
Answer	<ANSWER> Florence is the granddaughter of Betty.

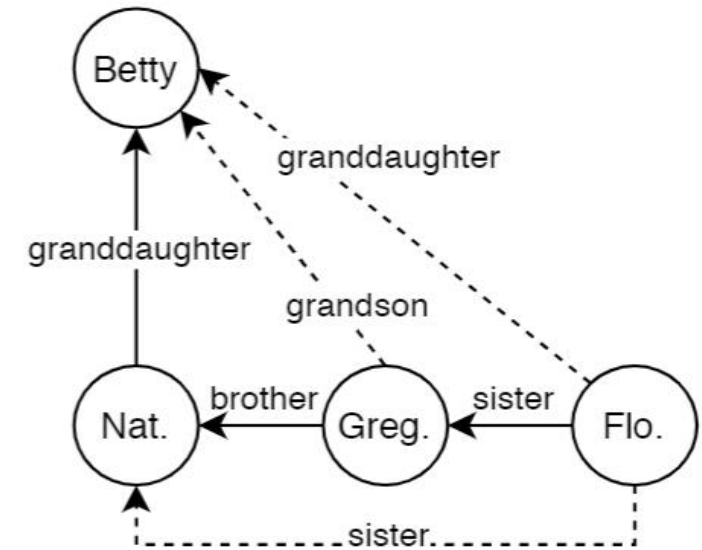
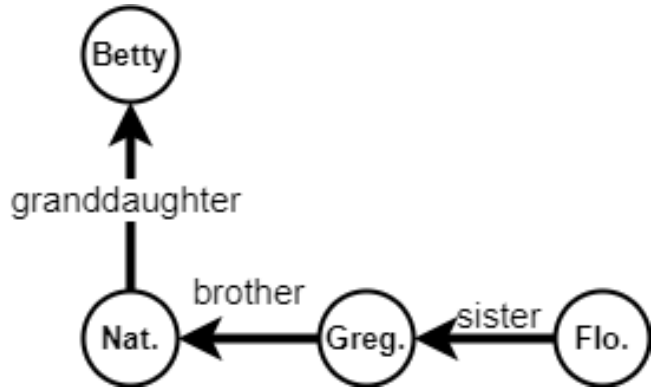
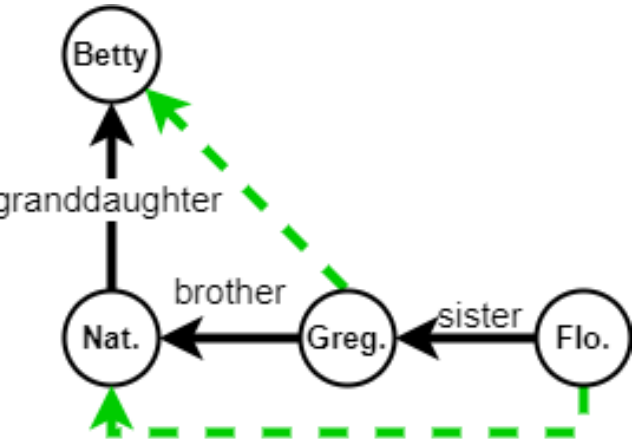
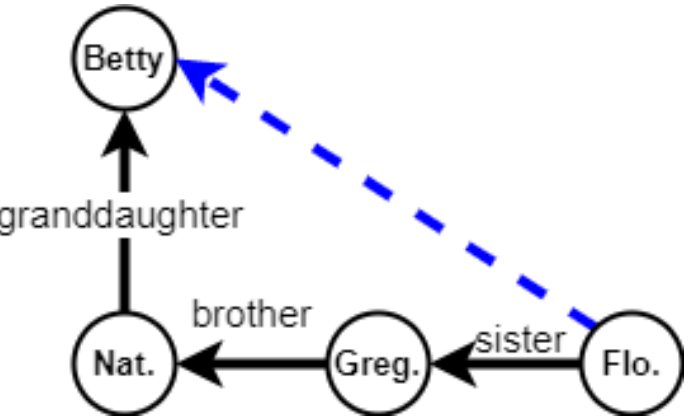


Figure 1: Example of a CLUTRR graph with known facts (solid lines) and unknown facts to infer (dotted lines).

CLUTRR stories template

	Synthetic language	Natural language
Story	<p>Natasha is a granddaughter to Betty.</p> <p>Florence is Gregorio 's sister.</p> <p>Gregorio is a brother of Natasha.</p>	<p>Betty likes picking berries with her son's daughter. Her name is Natasha.</p> <p>Gregorio took his sister, Florence, to a baseball game.</p> <p>Gregorio and his sister Natasha love it when their grandmother visits because she spoils them. She is coming this week to watch them while their parents are out of town.</p>
Query	Who is Florence for Betty ?	

CLUTRR dataset levels

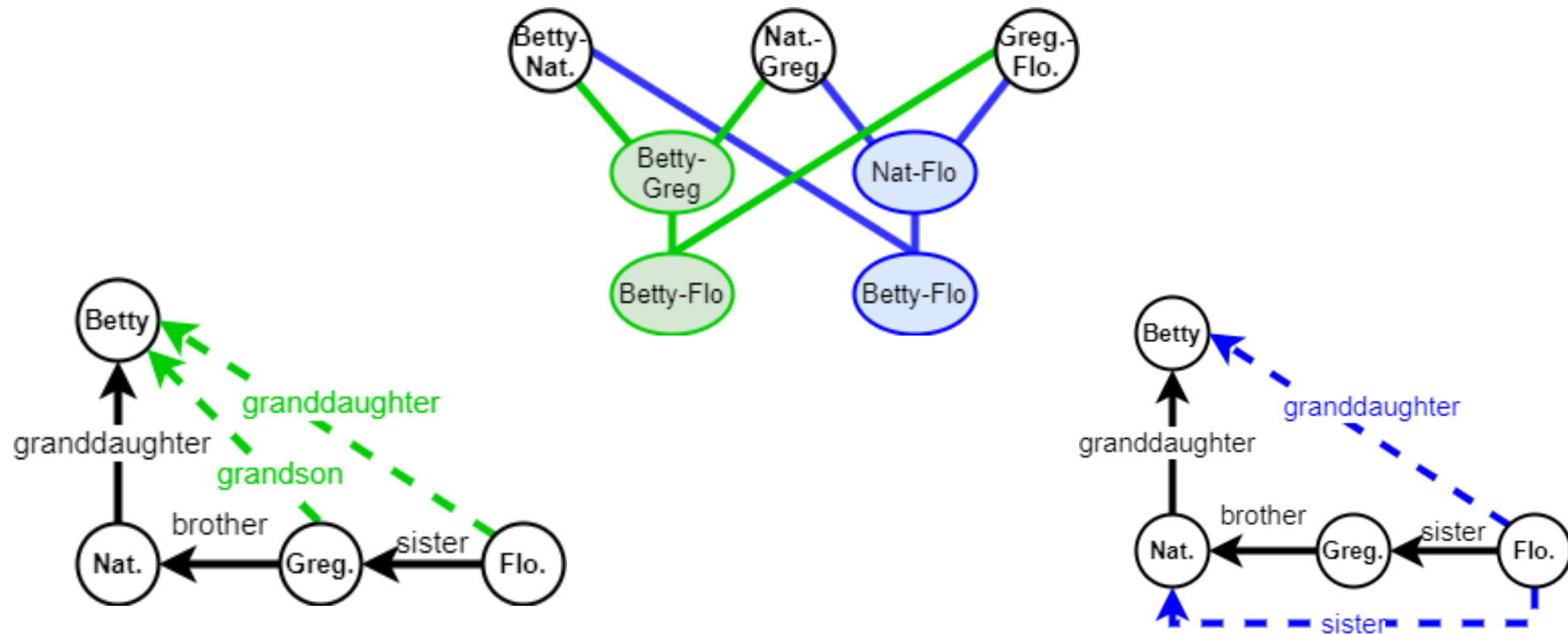
Level 1: remembering facts	Level 2: one hop inference	Level 3: two hops inference
 <p>A directed graph showing relationships between five people: Betty, Nat., Greg., and Flo. Betty is at the top, connected to Nat. by a solid arrow labeled 'granddaughter'. Nat. is connected to Greg. by a solid arrow labeled 'brother'. Greg. is connected to Flo. by a solid arrow labeled 'sister'.</p>	 <p>The same graph as Level 1, but with two green dashed arrows indicating one-hop inferences. One arrow goes from Flo. to Greg. (labeled 'sister'), and another goes from Greg. to Nat. (labeled 'brother').</p>	 <p>The same graph as Level 1, but with a blue dashed arrow indicating a two-hop inference. The arrow goes from Flo. to Betty, passing through Greg. and Nat. (labeled 'sister' and 'brother' respectively).</p>

Terminology

- “entity” – one node – eg: “Anna”
- “relation” – one edge – eg: “mother”
- “fact” – one (entity, relation, entity) triple – eg: “Anna is the mother of Bob”
- “proof step” – one proof sentence (*hop*) made of 3 facts – “since AB and BC then AC”
- “proof” – the whole string made of many ($k-1$) proof steps.
- Difficulty level k :
A level k task consists of k relations between $k+1$ entities and $k-1$ proof steps to solve the task

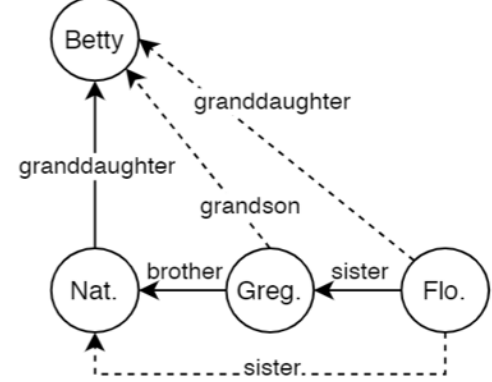
Proof strategies

	Forward-Chaining	Backward-Chaining
Short	short-proof – top-down	short-proof – bottom-up
Exhaustive	long-proof – top-down	long-proof – bottom-up



Proof strategies

- Short proof (sp):
 - Forward chaining
 - Exactly $k-1$ steps
- Short proof reversed (spr):
 - Backward chaining
 - Exactly $k-1$ steps.
- Long proof (lp):
 - Forward chaining
 - Exhaustive proof
- Long proof reversed (lpr):
 - Backward chaining
 - Exhaustive proof



sp	since Gregorio is a brother of Natasha, and Natasha is the granddaughter of Betty, then Gregorio is a grandson of Betty. since Florence is a sister of Gregorio, and Gregorio is a grandson to Betty, then Florence is a granddaughter to Betty.
spr	since Florence is a sister of Gregorio, and Gregorio is a grandson to Betty, then Florence is a granddaughter to Betty. since Gregorio is a brother of Natasha, and Natasha is the granddaughter of Betty, then Gregorio is a grandson of Betty.
lp	since Gregorio is the brother of Natasha, and Natasha is the granddaughter of Betty, then Gregorio is the grandson of Betty. since Florence is the sister of Gregorio, and Gregorio is the brother of Natasha, then Florence is the sister of Natasha. since Florence is the sister of Natasha, and Natasha is the granddaughter of Betty, then Florence is the granddaughter of Betty.
lpr	since Florence is the sister of Natasha, and Natasha is the granddaughter of Betty, then Florence is the granddaughter of Betty. since Florence is the sister of Gregorio, and Gregorio is the brother of Natasha, then Florence is the sister of Natasha. since Gregorio is the brother of Natasha, and Natasha is the granddaughter of Betty, then Gregorio is the grandson of Betty.

Table 2: Proof resolution types for an example of level 3. We refer the reader to Figure 1 for the kinship graph corresponding to this example. **sp**=short-proof, **spr**=short-proof-reversed, **lp**=long-proof, **lpr**=long-proof-reversed.

Systematic generalization in proof generation

- Train on levels: 2, 4, 6
Test on levels: 2, 3, 4, 5, 6, 7, 8, 9, 10.
- Include all entities, relations, facts (triples) in training
Can the model generalize to new proofs ?

ANON TEST	lvl.2	lvl.3	lvl.4	lvl.5	lvl.6	lvl.7	lvl.8	lvl.9	lvl.10
proofs (<i>many proof steps</i>)	16.28%	0%	0%	0%	0%	0%	0%	0%	0%
proof steps (<i>“since A-r1-B and B-r2-C then A-r3-C”</i>)	73.08%	58.06%	52.75%	54.28%	50.93%	59.04%	56.92%	53.55%	52.17%
facts (A-r-B)	100%	100%	100%	100%	100%	100%	100%	100%	100%
entities (A)	100%	100%	100%	100%	100%	100%	100%	100%	100%
relations (r)	100%	100%	100%	100%	100%	100%	100%	100%	100%

Table 3: Percentage of the test proof’s building blocks also present in the training set (composed of levels 2, 4, 6) for all levels. We colored all cells with a value of 100% to better visualize which building blocks were entirely contained in the training set.

Experiments

- **Training:**

Transformer decoder trained to **predict next word** in sequences of “<STORY> [story] <QUERY> [query] <PROOF> [proof] <ANSWER> [answer]”

- **Evaluation:**

Proof consistency:

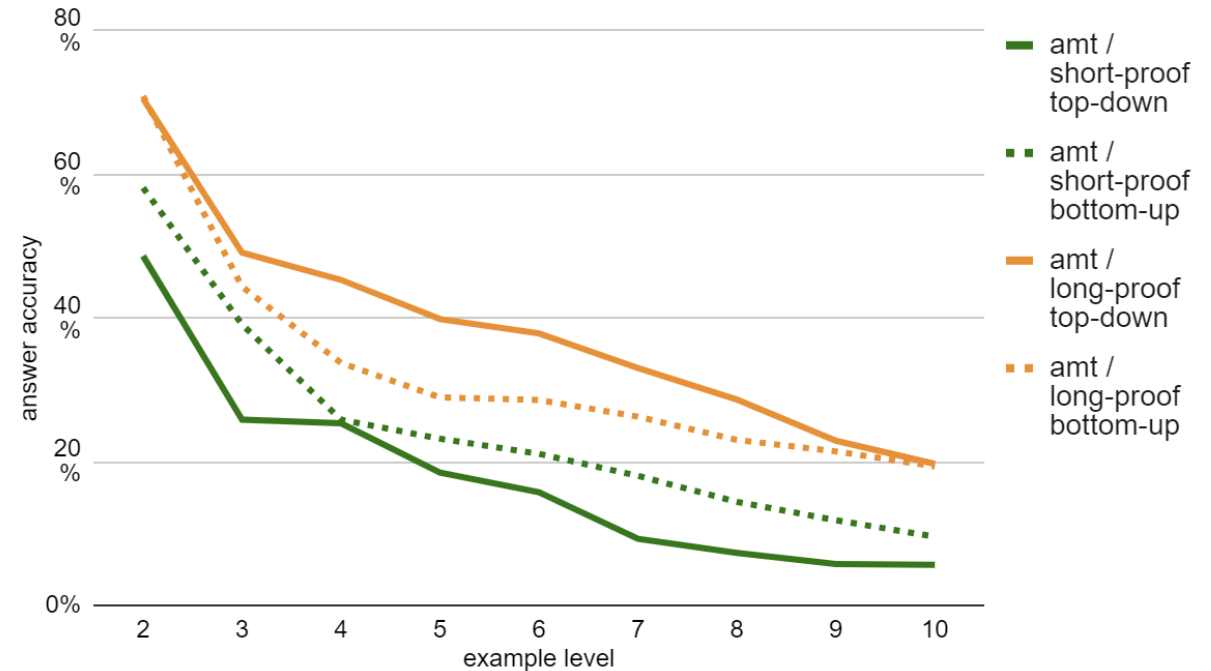
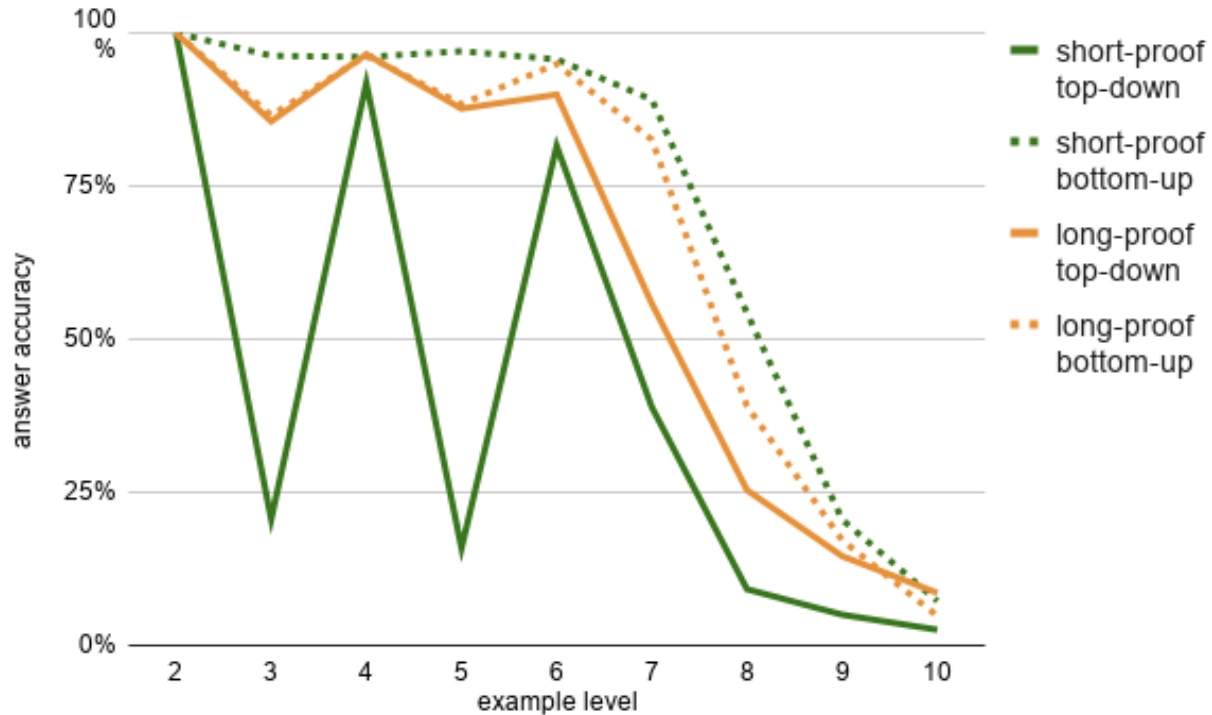
Verify the logic of the generated sequence after the “<PROOF>” token based on CLUTRR kinship rules.

Answer accuracy:

Compare the generated sequence after the “<ANSWER>” token with ground truth.

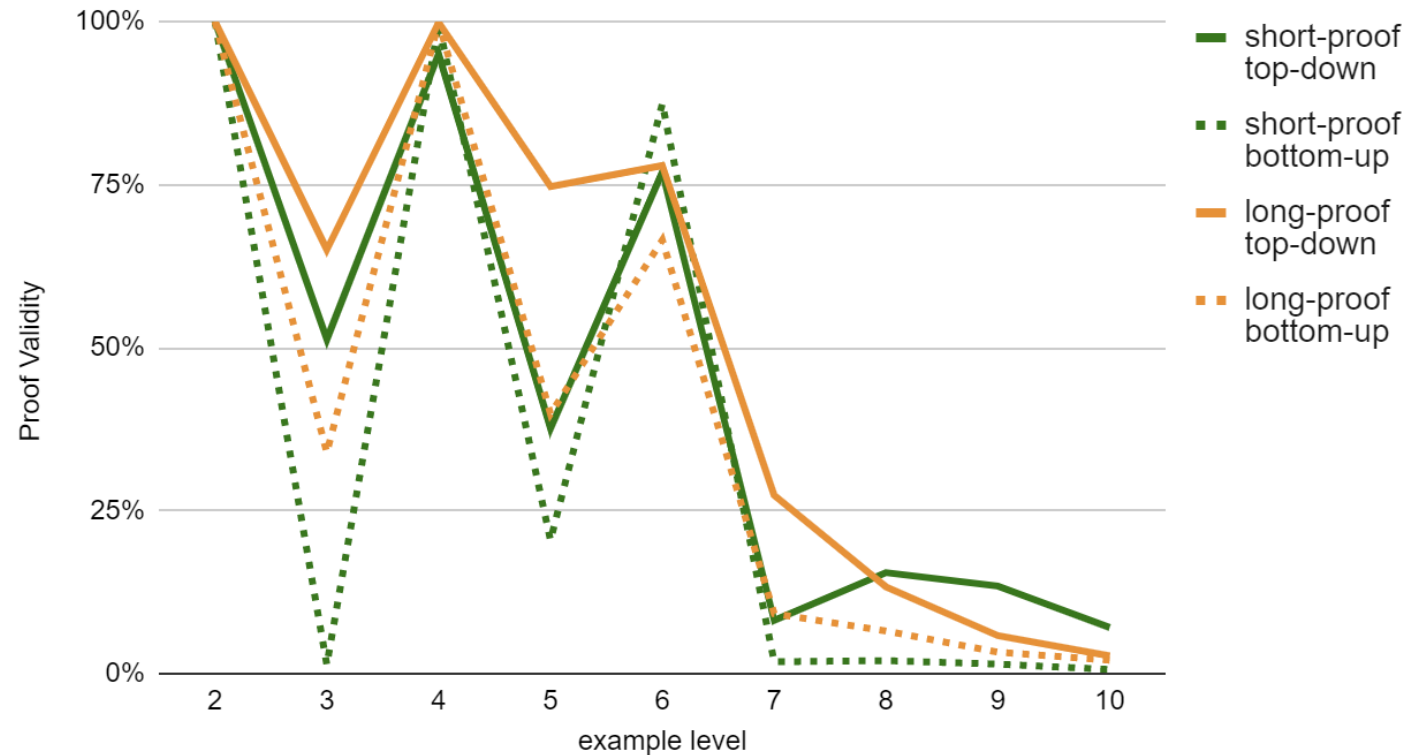
Top-Down –vs – Bottom-Up reasoning

- Bottom-up is easier to use compared to top-down



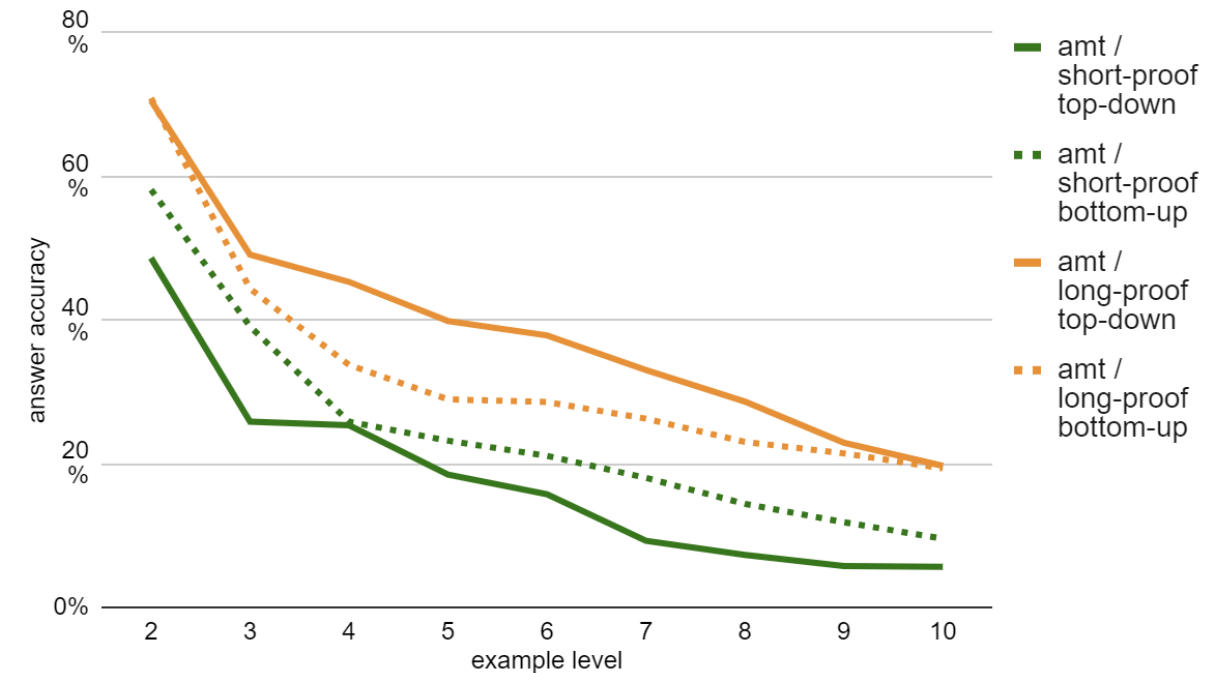
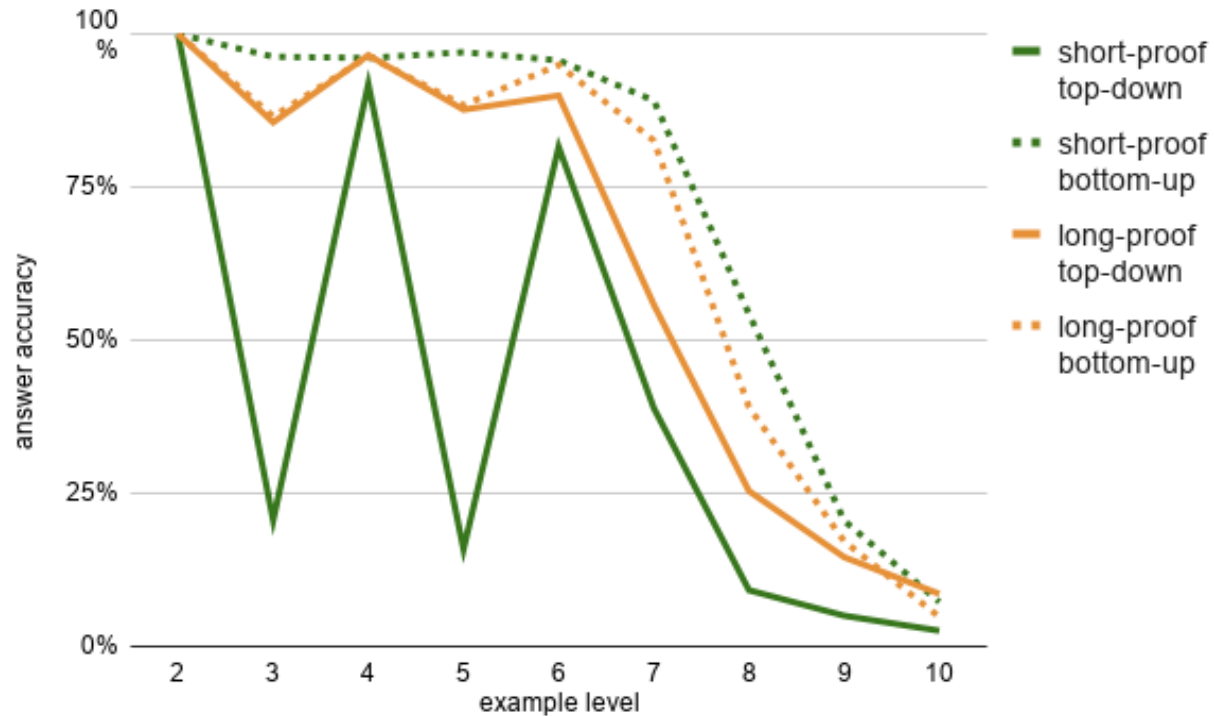
Top-Down –vs– Bottom-Up reasoning

- Bottom-up is harder to generate compared to top-down



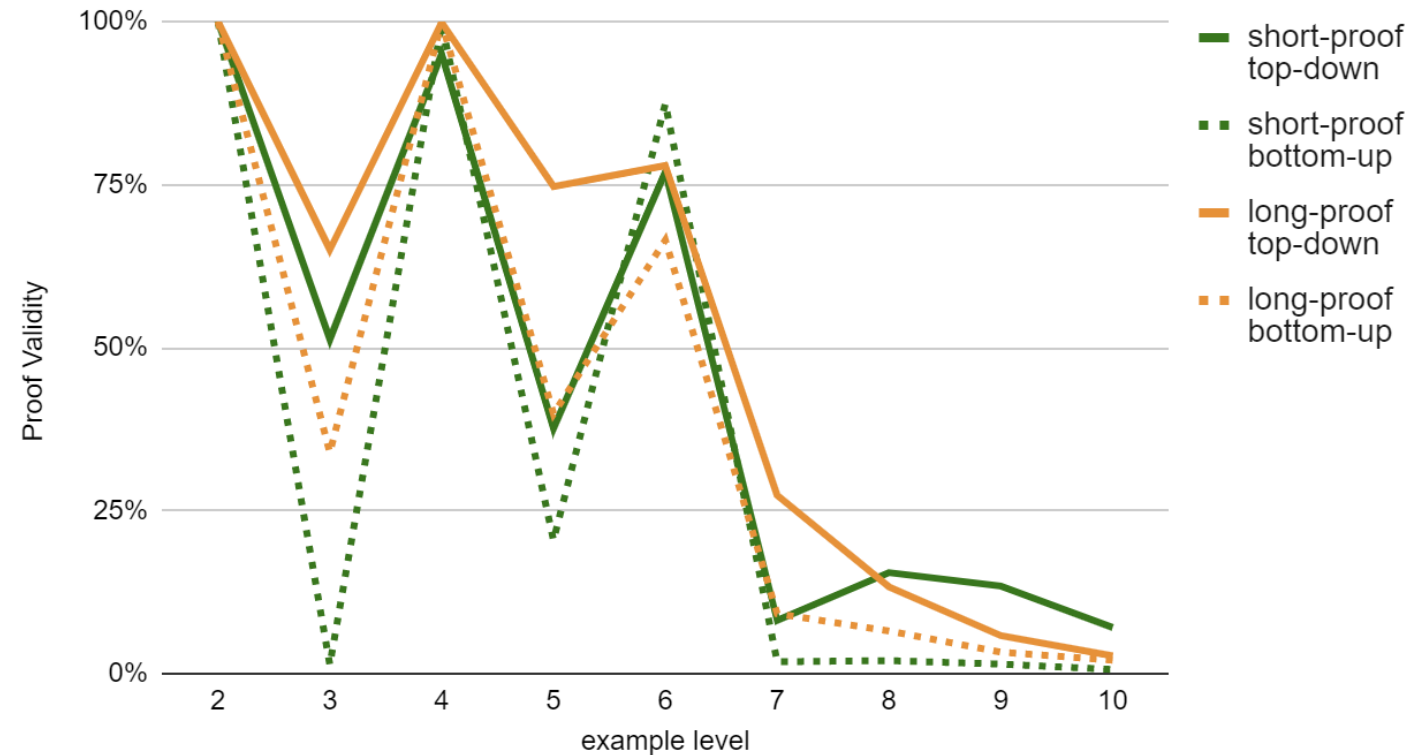
Long –vs– Short proofs

- Long proofs are easier to use compared to Short proofs



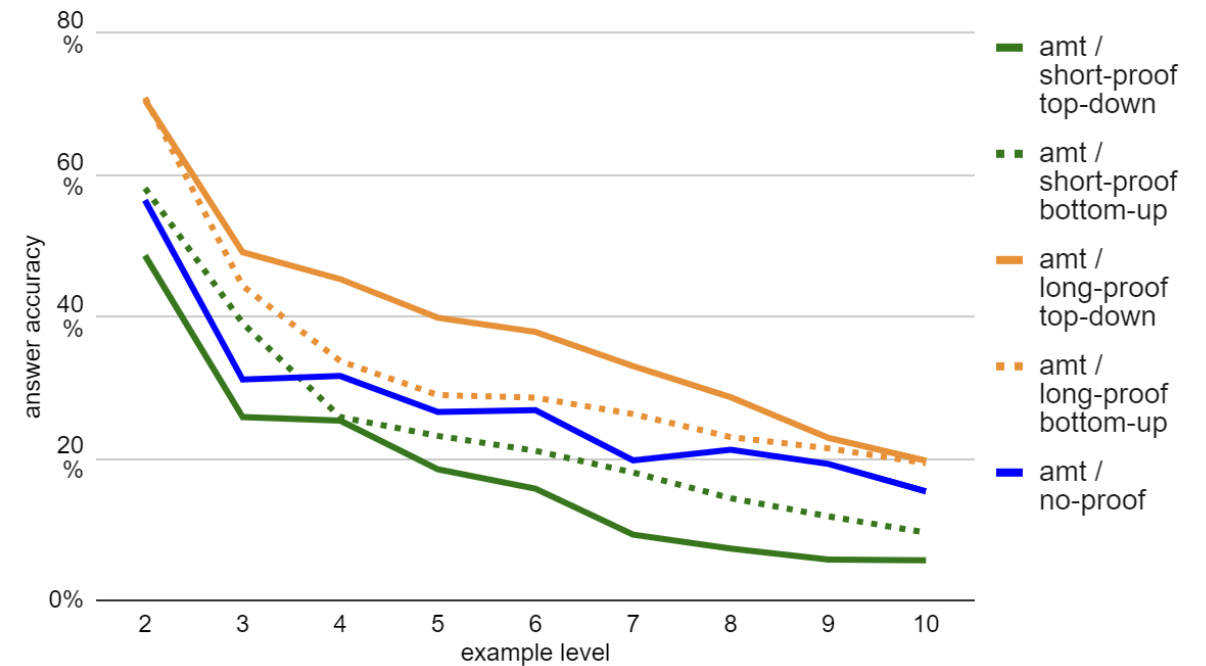
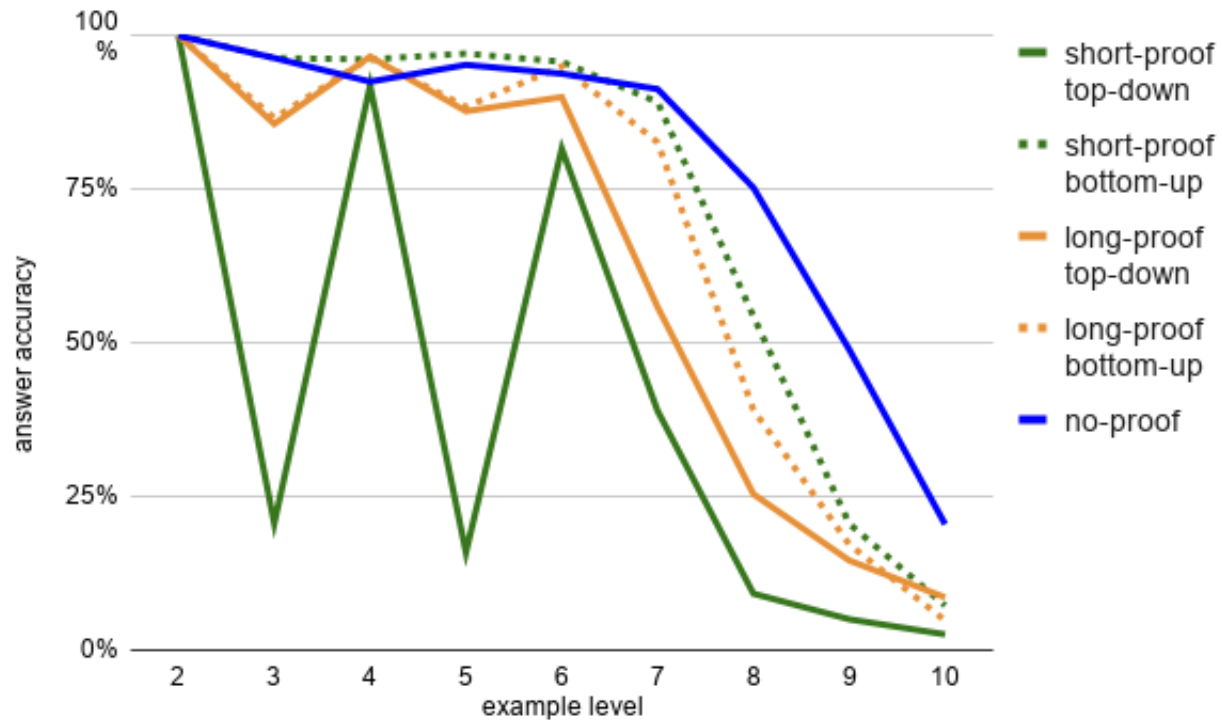
Long –vs– Short proofs

- Long proofs are easier to generate compared to Short proofs



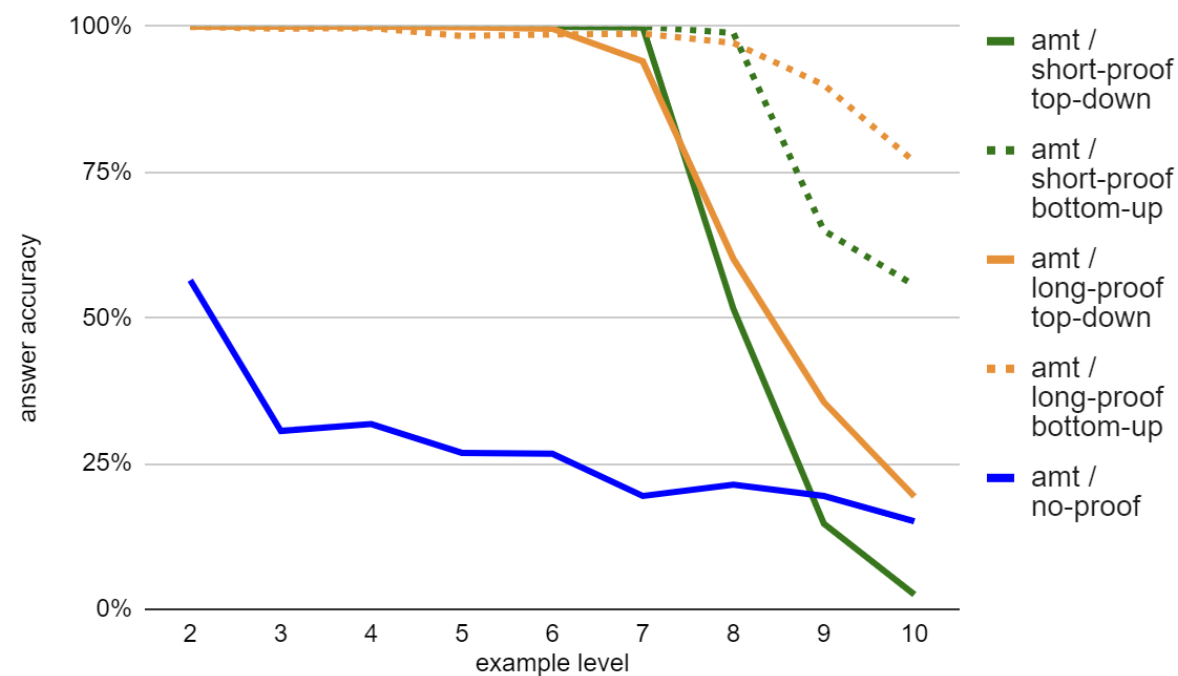
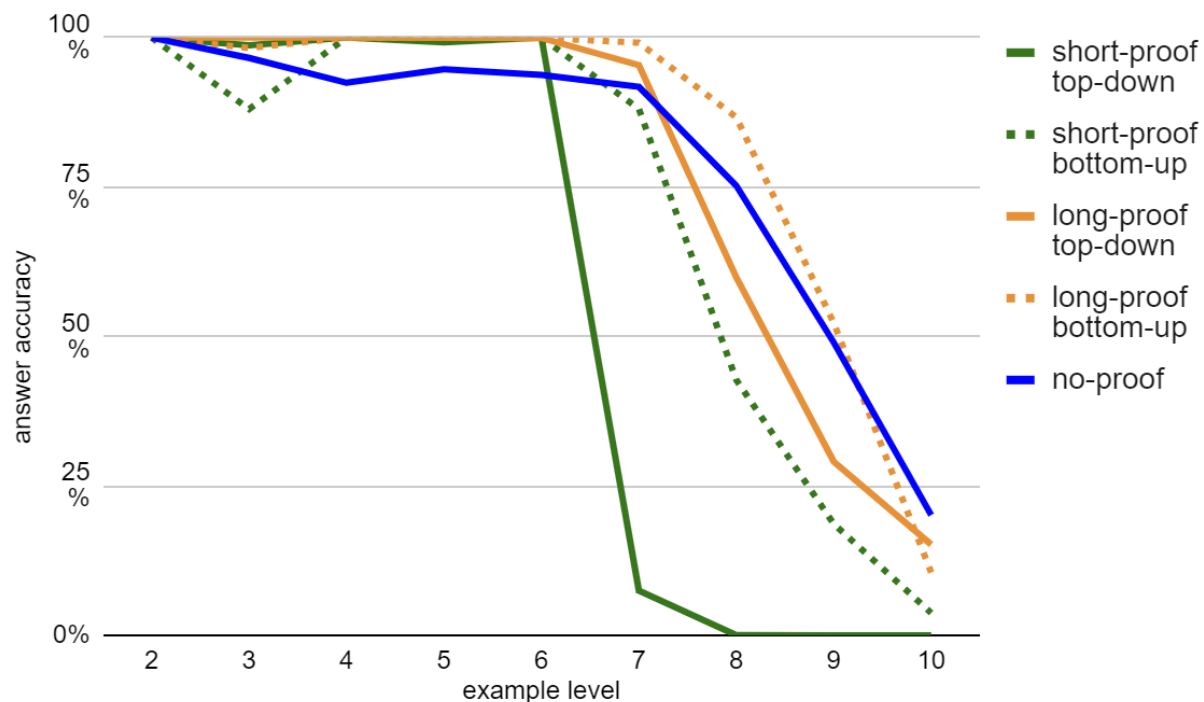
Without generating a proof

- Generating a proof makes it harder to extrapolate



Given the real proof

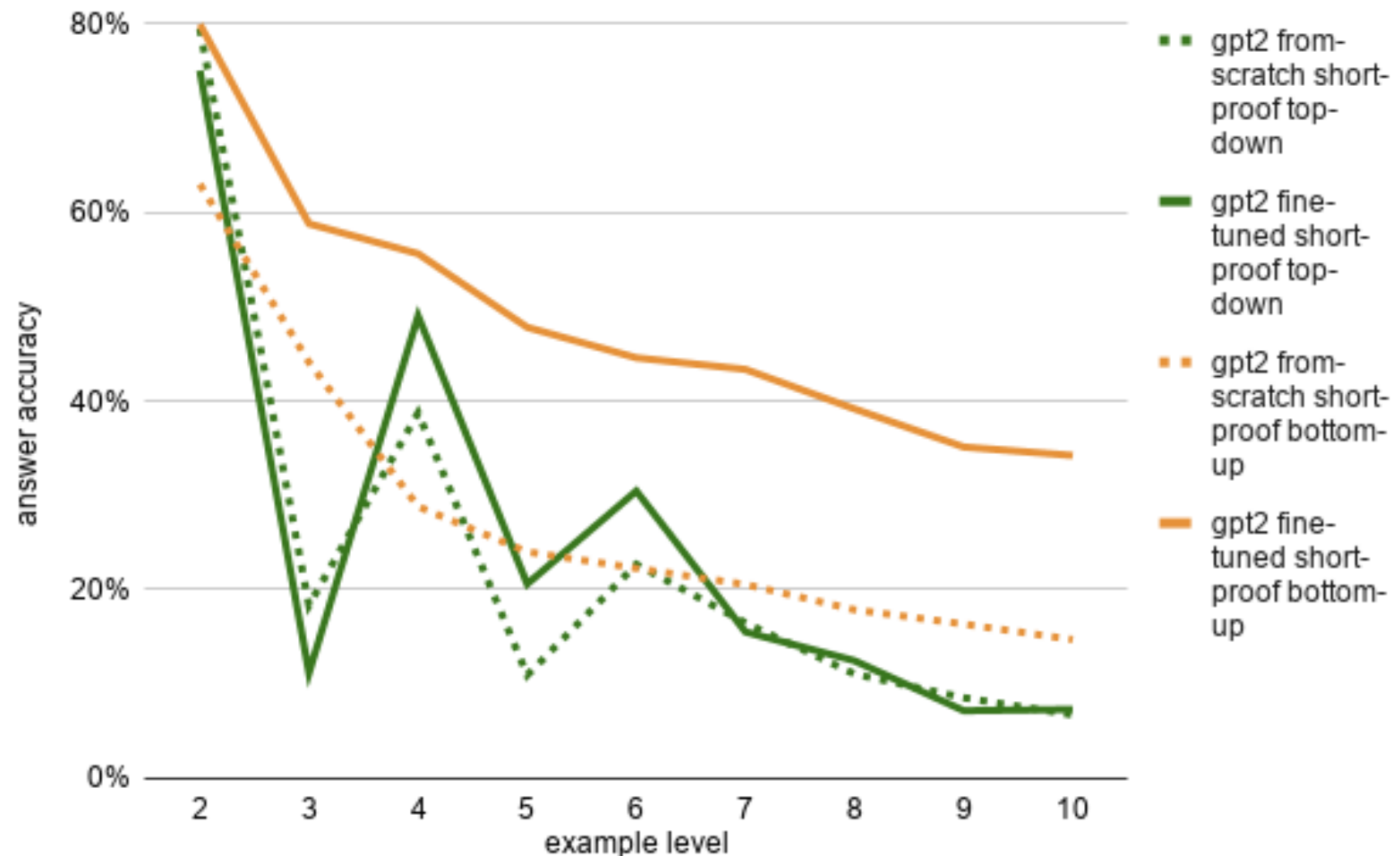
- Having access to the real proof improves answer accuracy
- But models still struggle to extrapolate



GPT2 finetuned –vs – trained from scratch

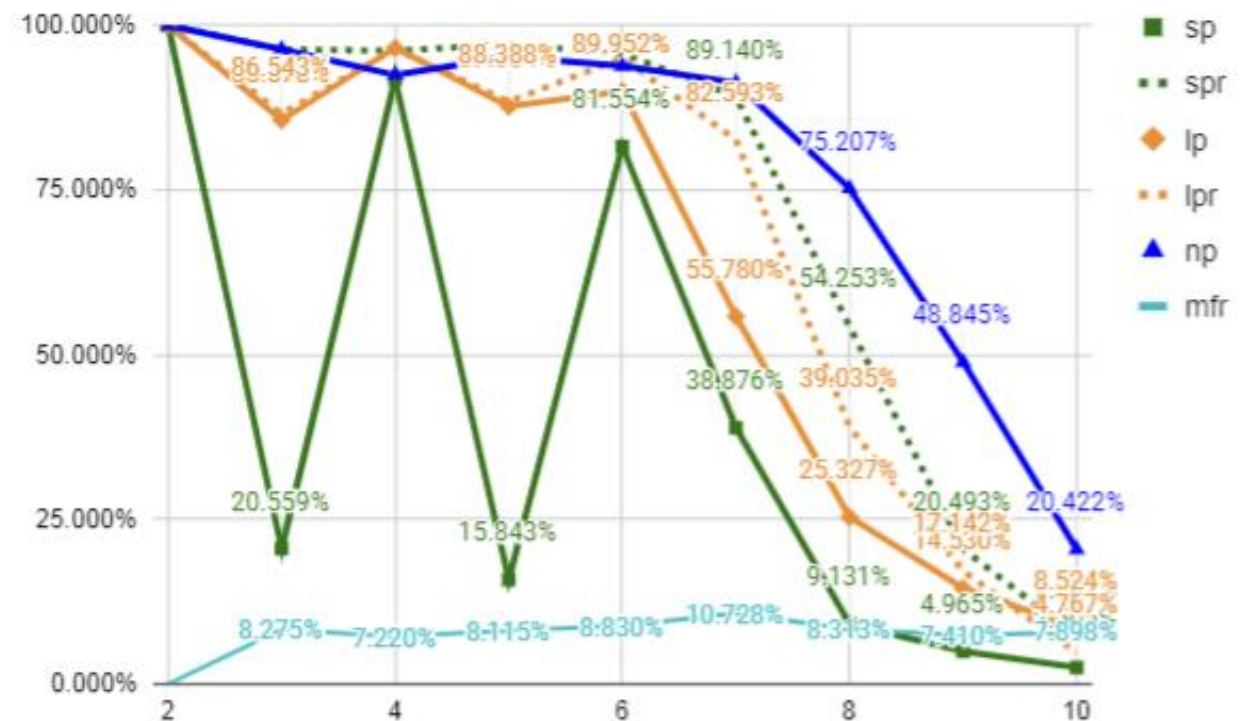
- The improvement from fine-tuning depends on the reasoning strategy used:

→ More improvement with bottom-up



Observations

- TLMs suffer from length generalization issues in generating proofs.
- TLMs get better at reasoning when trained with longer proofs.
- Backward-chaining proofs are easier to use when generating an answer.
- Backward-chaining proofs are harder to generate.



References

- Noam Chomsky. 1957. Logical structures in language. *American Documentation*, 8(4):284–291.
- Richard Montague. 1970. Universal grammar. *Theoria*, 36(3):373–398.
- Brenden M. Lake and Marco Baroni. 2018. *Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks*. In ICML, pages 2879–2888.
- Adam Liška, Germán Kruszewski, and Marco Baroni. 2018. *Memorize or generalize? searching for a compositional rnn in a haystack*. arXiv preprint arXiv:1802.06467.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. *Language models as knowledge bases?* In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. *Language models are few-shot learners*. arXiv preprint arXiv:2005.14165.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. *Exploring the limits of transfer learning with a unified text-to-text transformer*. *Journal of Machine Learning Research*, 21(1).
- Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. *CLUTRR: A diagnostic benchmark for inductive reasoning from text*. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pages 4506–4515, Hong Kong, China. Association for Computational Linguistics.
- Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. 2019. *Systematic generalization: What is required and can it be learned?* In Proceedings of the 2019 International Conference on Learning Representations.

Thank you.

Full paper: <https://arxiv.org/pdf/2009.14786.pdf>

Code: <https://github.com/NicolasAG/SGinPG>

Questions: gontiern@mila.quebec

