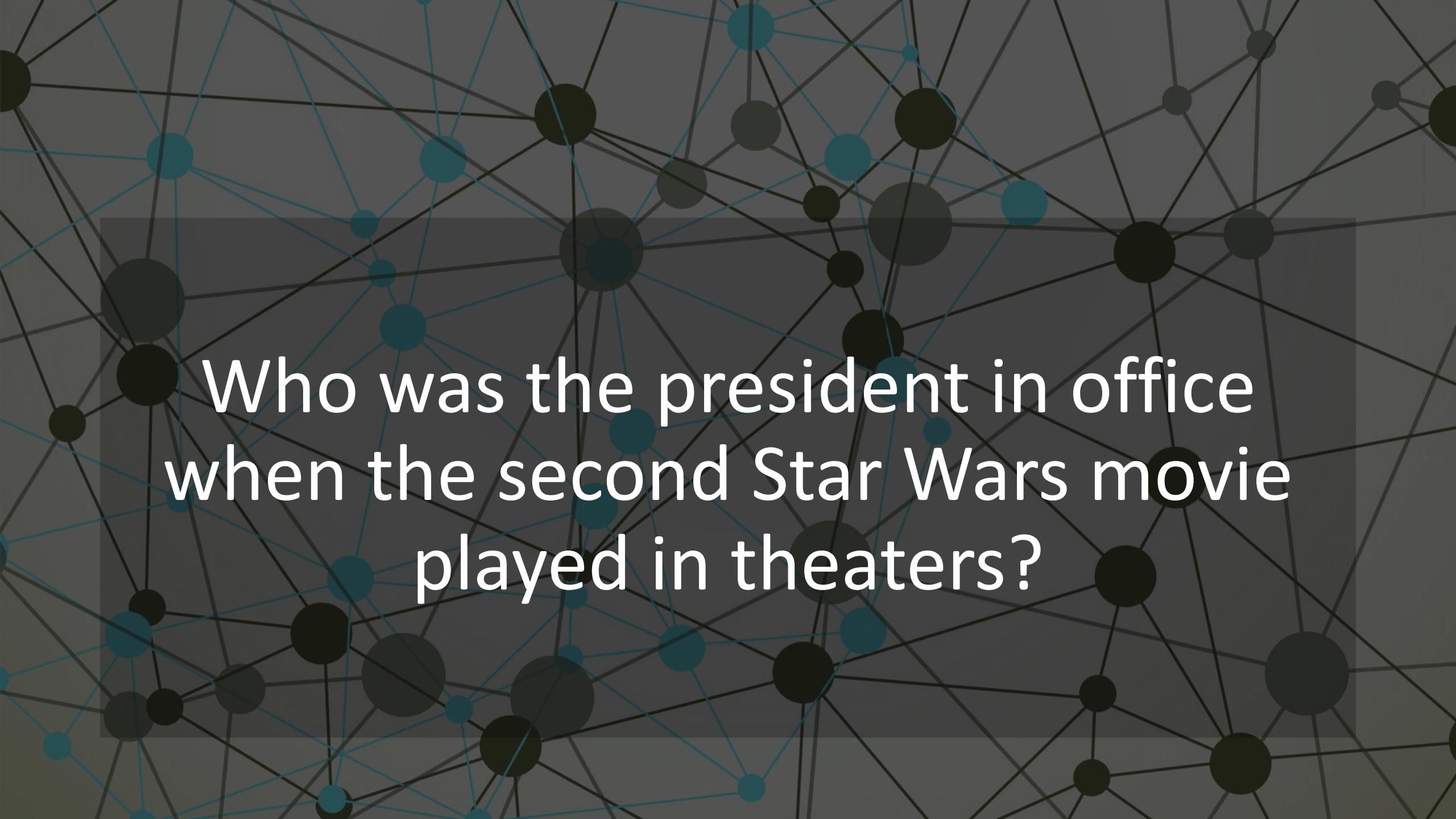


Fine-tuning On Rationale Generation Improves Multi- Hop Reasoning in Small Language Models

Richard Mathews, Adam
Weinberger, and Meng-Kang Kao
University of California, Berkeley
August 9th, 2023



Who was the president in office
when the second Star Wars movie
played in theaters?



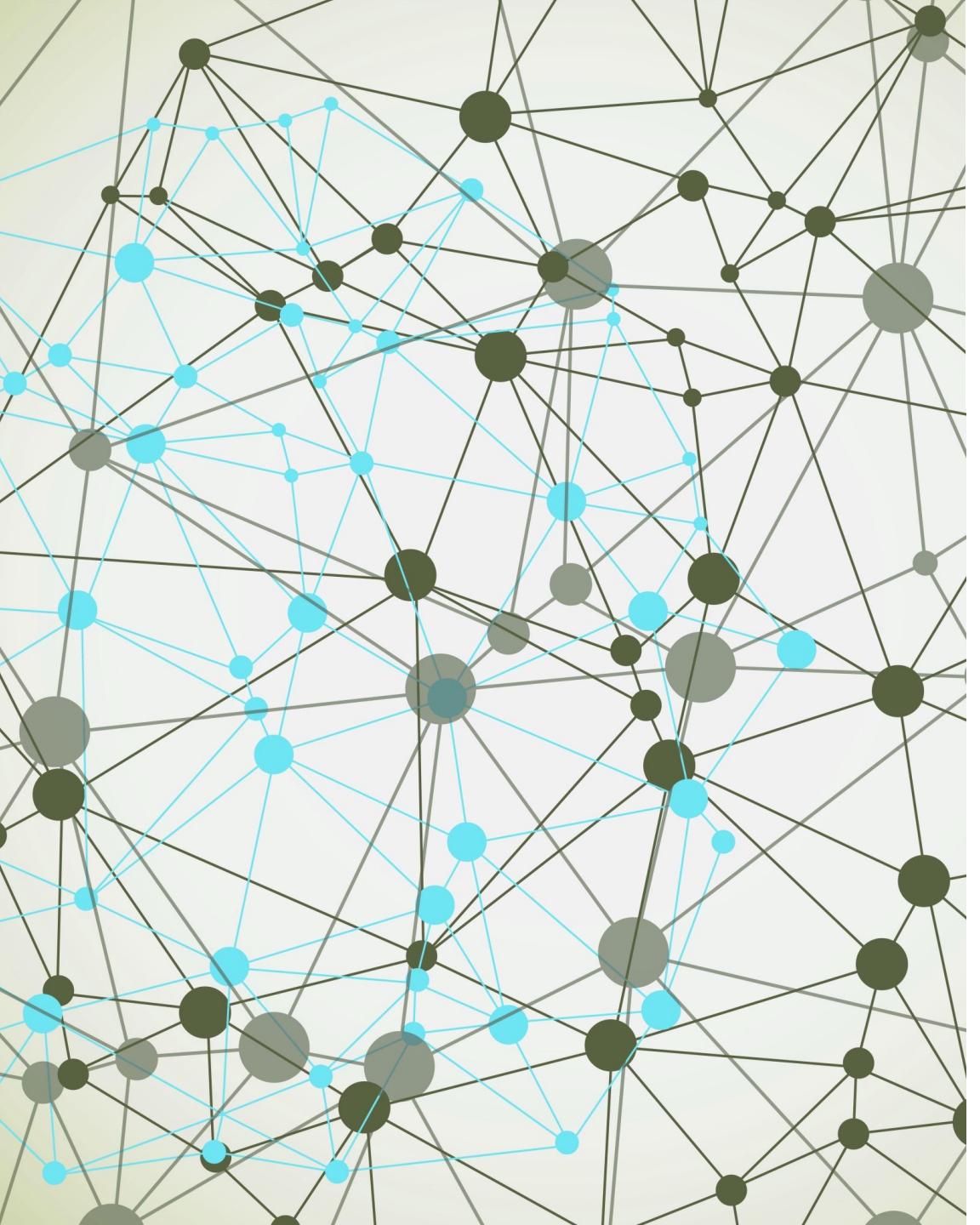
Introduction

- Question: Who was the president in office when the second Star Wars movie played in theaters?
 - Step 1: When was the second Star Wars movie played in theaters?
 - 1980
 - Step 2: Who was the US president in 1980?
 - Jimmy Carter
 - The final answer to the original question:
Jimmy Carter
- Chain-of-Thought (CoT) rationale can help to answer multi-hop questions.
 - Self-Ask has impressive results on large GPT-3 models (175B parameters)

<https://www.imdb.com/title/tt0080684/>

https://en.wikipedia.org/wiki/Jimmy_Carter





Introduction

- What if we use encoder-decoder architecture?
- How about smaller language models?
- Can we make it better by fine-tuning it?
- Focus on smaller language models
 - T5 small (60M)
 - Flan-T5 small (60M) – Instruction tuned T5
 - OPT (125M) – Open-Source LM similar to GPT-3
- Pay attention to both question answering as well as rationale generation

Creating our datasets

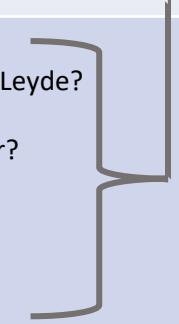
Base Dataset: 2WikiMultiHopQA

200,000 multi-hop reasoning questions generated from Wikipedia

- Predominantly concerning the date of birth, lifespan, country of origin, and familial relationships of historical figures and celebrities

Dataset Modifications

Fine-tuning datasets: direct no exemplars and self-ask with exemplars

Fine-tuning Datasets Examples		
Direct	Self-Ask with Exemplars	Exemplar (actual data has 2 exemplars)
Prompt	Prompt	
Facts: Fact #0: The film was written, adapted and directed by Russian-born Arcady Boytler. Fact #1: Boytler was born in Moscow, Russia. Question: Where was the director of film Heads Or Tails (1937 Film) born? Answer:	START Question: Who was born first, Alejo Mancisidor or Emil Leyde? Are follow up questions needed here: Yes. Follow up: When is the date of birth of Alejo Mancisidor? Intermediate answer: 31 July 1970 Follow up: When is the date of birth of Emil Leyde? Intermediate answer: 8 January 1879 So the final answer is: Emil Leyde END Facts: Fact #0: Mikko Esa Juhani Heikka(born 19 September 1944 in Ylitornio) is a Finnish former bishop of the Evangelic Lutheran Church. Fact #1: Scott Douglas Robbe is an American film, television, and theater producer/director. Question: Does Mikko Heikka have the same nationality as Scott Robbe? Are follow up questions needed here:	
Training Label (Question Answering)	Training Label (Rationale Generation)	
Moscow	Yes. Follow up: What is the country of citizenship of Mikko Heikka? Intermediate answer: Finnish Follow up: What is the country of citizenship of Scott Robbe? Intermediate answer: American So the final answer is: no	

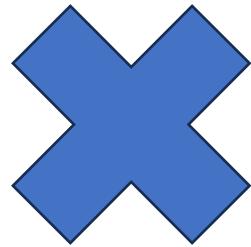
Nine models to evaluate

**Three baseline
models**

T5

Flan-T5

OPT



**Three fine-tuning
procedures**

None

Direct

Self-ask



Nine models

T5

Flan-T5

OPT

Direct T5

Direct Flan-T5

Direct OPT

Self-Ask T5

Self-Ask Flan-T5

Self-Ask OPT

Evaluate each model twice

Evaluation Procedure

Each model is evaluated on test sets with and without exemplars (i.e. **few-shot and zero-shot**) giving us 18 total evaluations

*nomenclature - (Self-Ask Flan-T5 few-shot)

Metrics for each Task

Question Answering

Accuracy - correct if true answer appears anywhere in the generated answer, some drawbacks but straightforward

Rationale Generation

F1-1 (1-gram) and F1-2 (2-gram)

*baseline and direct models with zero-shot test sets only evaluated on question answering

Direct-tuning is good... but self-ask-tuning is better

Key Results

- Major improvement over baselines
- Self-ask Flan-T5 is the most accurate model
- Self-ask Flan-T5 is a zero-shot CoT multi-hop reasoner

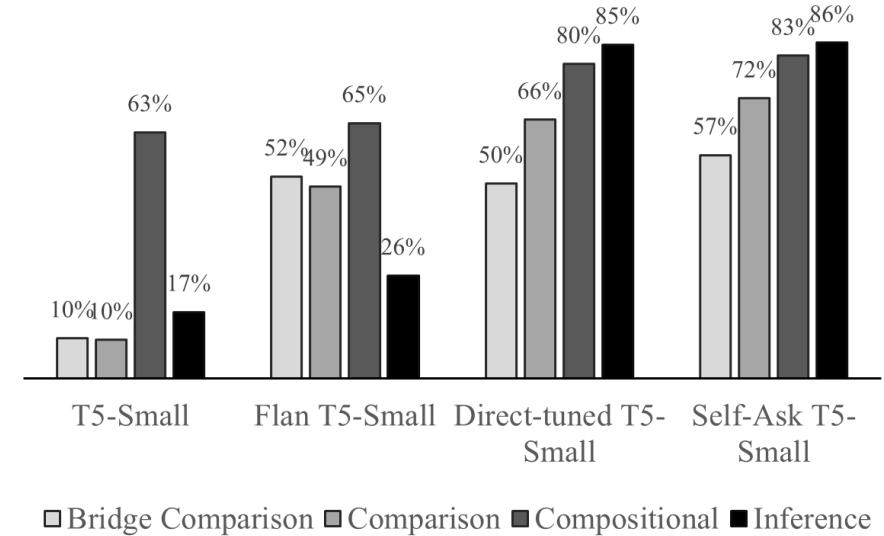
	Accuracy (QA)
T5-Small (Zero-Shot)	32.9
T5-Small (Few-Shot)	24.8
Flan T5-Small (Zero-Shot)	51.7
Flan T5-Small (Few-Shot)	53.6
Direct-tuned T5-Small (Zero-Shot)	70.9
Direct-tuned T5-Small (Few-Shot)	10.6
Direct-tuned Flan T5-Small (Zero-Shot)	66.2
Direct-tuned Flan T5-Small (Few-Shot)	47.0
Self-Ask T5-Small (Zero-Shot)	40.9
Self-Ask T5-Small (Few-Shot)	74.8
Self-Ask Flan T5-Small (Zero-Shot)	73.2 ★
Self-Ask Flan T5-Small (Few-Shot)	77.9 ★

+21.5%
+24.3%

+7%

So, what differences between self-ask-tuning and direct-tuning do we observe?

Similar on the aggregate performance...



but different strengths and weaknesses

Direct-Tuned Models

- Weaknesses in multi-hop reasoning
- Struggles on larger context sizes

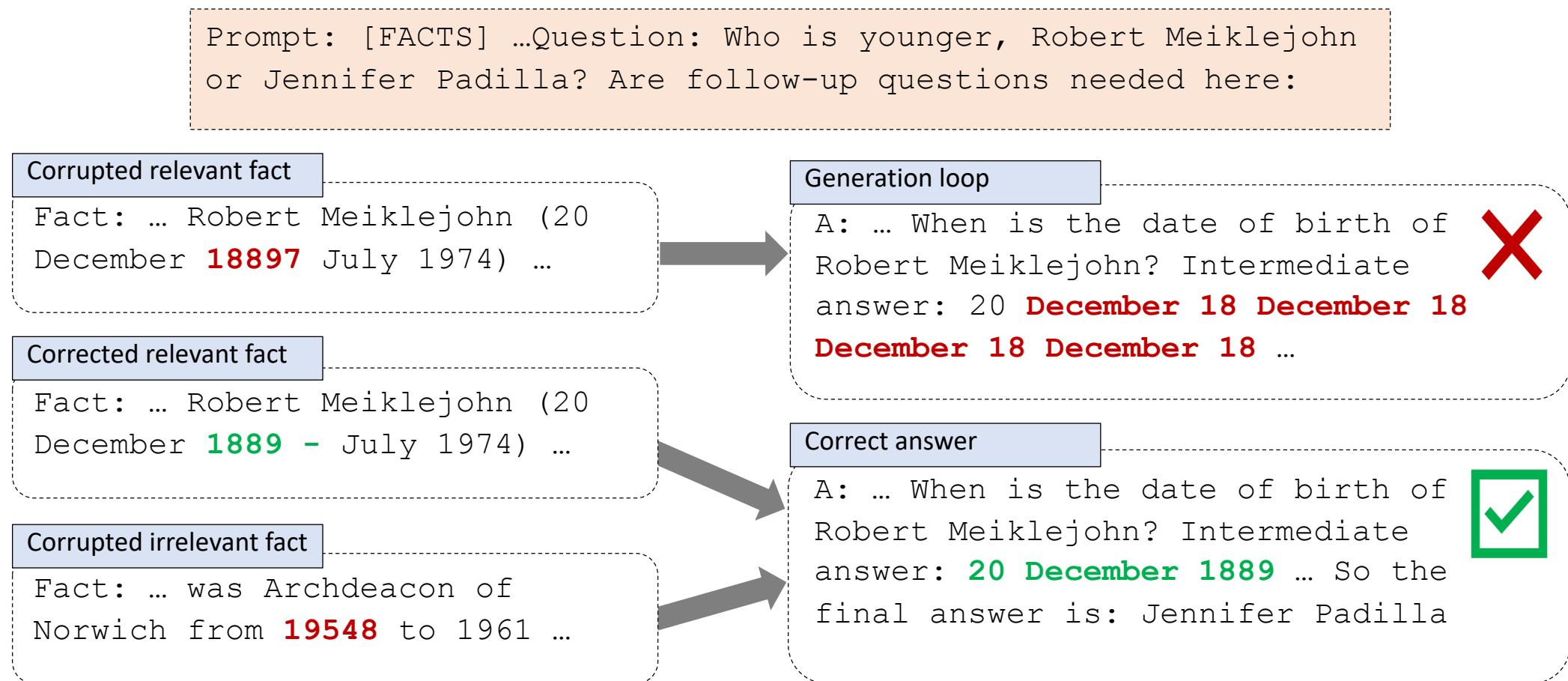
Both

- Sensitivity to order of facts
- Sensitivity to irrelevant facts

Self-Ask-Tuned Models

- Sensitivity to prompt quality

Corruption of critical facts elicits generation loops in self-ask models... but does not affect direct-tuned models!



Self-ask fine-tuning learns fact composition, and direct-tuning learns ???

Bringing it all together

Fine-tuning works. Fine-tuning is an effective approach to eliciting multi-hop reasoning in small language models.

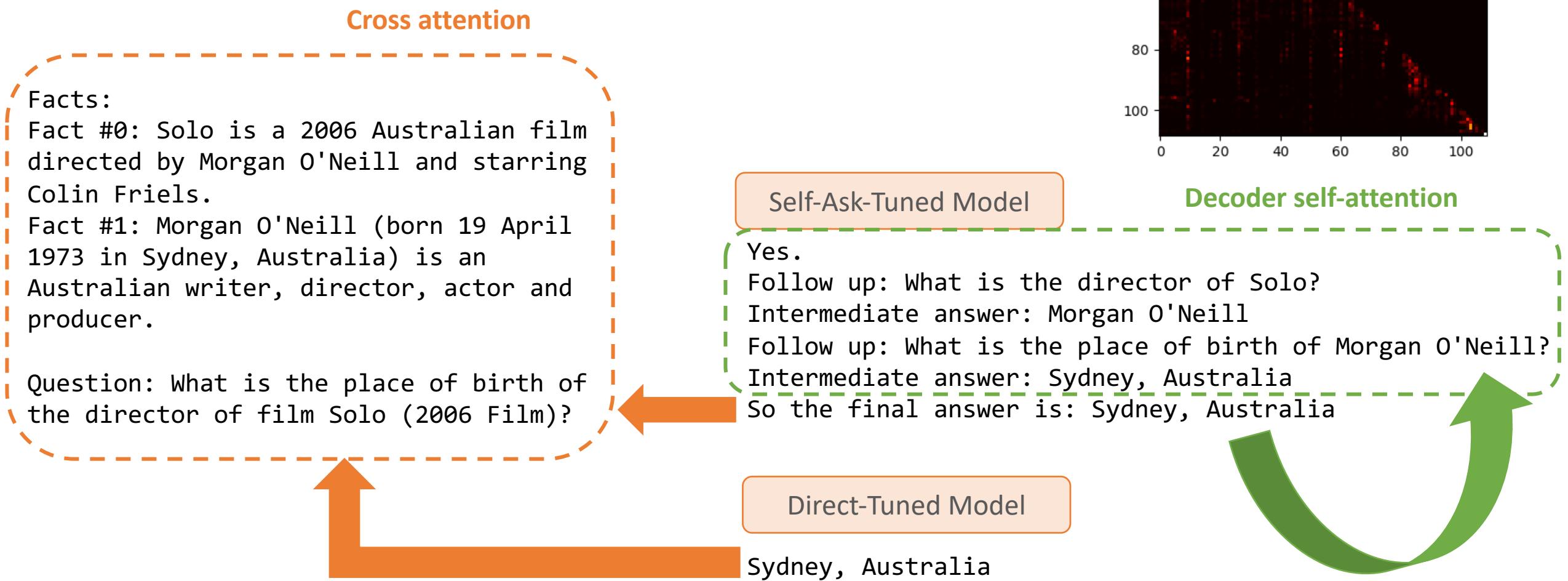
Small language models can be zero-shot multi-hop CoT reasoners. Self-ask-tuned Flan-T5 did very well in zero-shot regimes on both QA and rationale generation tasks.

Self-ask fine-tuning beats direct-tuning. Based on analysis, it is plausible self-ask models may be composing facts to answer multi-hop questions, where direct-tuned models may learn "cheat codes".



Questions?

Hypothesis: training language models on CoT rationales is a better approach than directly training on question-answer pairs.



Reasoning Question Categories

The data come with four pre-tagged categories (Ho, et al. 2020):

1. Comparison: compares an aspect of two or more entities from the same group. For instance, who was born first or died first.
2. Inference: given two relationships between three total entities (e_1, r_1, e_2) and (e_2, r_2, e_3), create a new instance (e_1, r_3, e_3) where the question is created from (e_1, r_3) the answer is e_3 .
3. Compositional: given two relationships between three total entities (e_1, r_1, e_2) and (e_2, r_2, e_3), where no direct relationship r_3 exists between e_1 and e_3 , the question is created from (e_1, r_1, r_2) and the answer is e_3
4. Bridge-comparison: given one relationship type between four total entities (e_1, r_1, e_2) and (e_3, r_1, e_4) the question asks to compare e_1 and e_3 , by comparing e_2 and e_4

References

- Star Wars Image <https://www.imdb.com/title/tt0080684/>
- Jimmy Carter Image https://en.wikipedia.org/wiki/Jimmy_Carter
- Ofir, Press, Muru Zhang, Sewon Min, and Ludwig Schmidt. 2022. *MEASURING AND NARROWING THE COMPOSITIONALITY GAP IN LANGUAGE MODELS*.
- Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. *Chain of thought prompting elicits reasoning in large language models*.
- Ho, Xanh, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. "Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps." *28th International Conference on Computational Linguistics*. Barcelona. 6609–6625.
- Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, et al. 2022. *OPT: Open Pre-trained Transformer Language Models*.