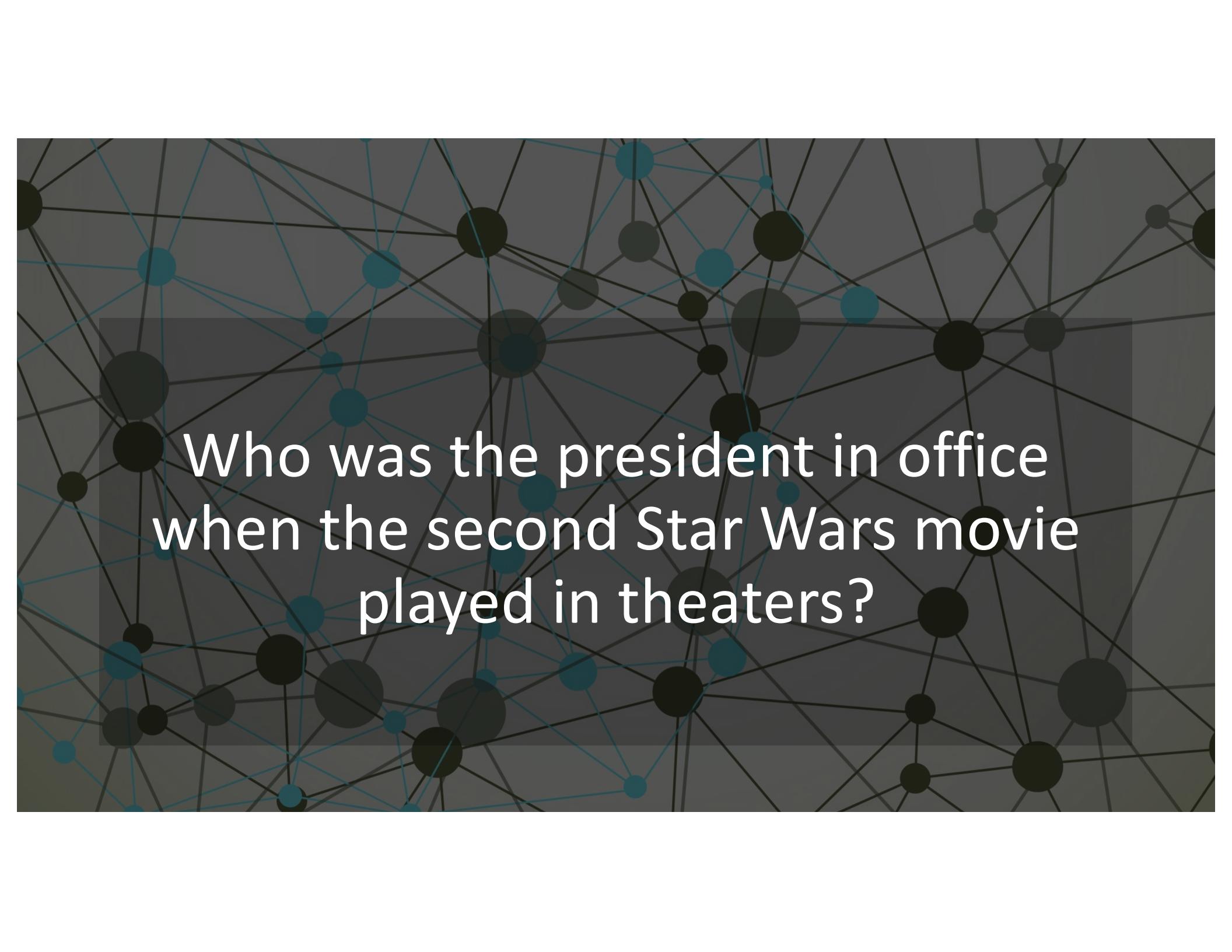


# Fine-tuning On Rationale Generation Improves Multi- Hop Reasoning in Small Language Models

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

The background of the slide features a complex network graph. It consists of numerous circular nodes of varying sizes, some filled with a dark teal color and others left as solid black. These nodes are interconnected by a dense web of thin, dark gray lines representing edges. The overall effect is one of a complex system or web of relationships.

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

THE STAR WARS SAGA CONTINUE



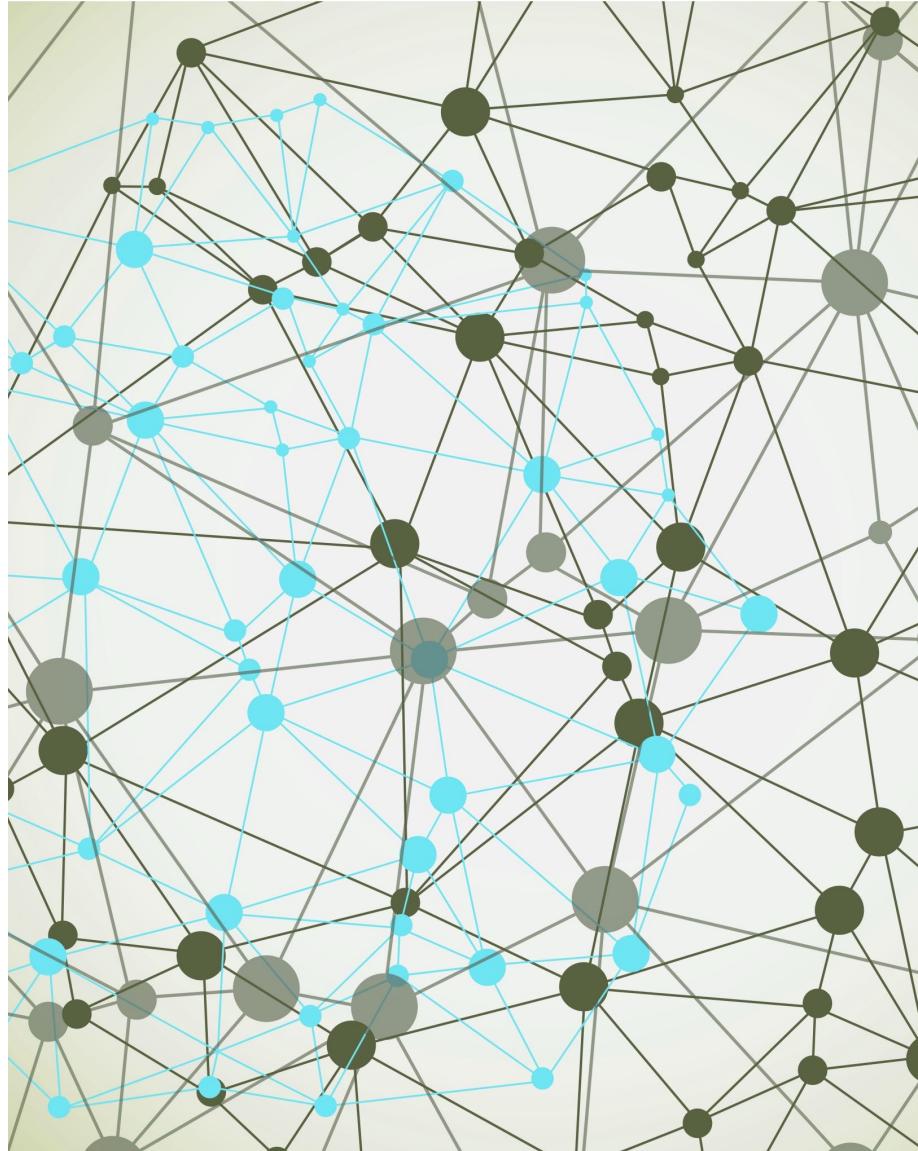
# Introduction

- 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/>

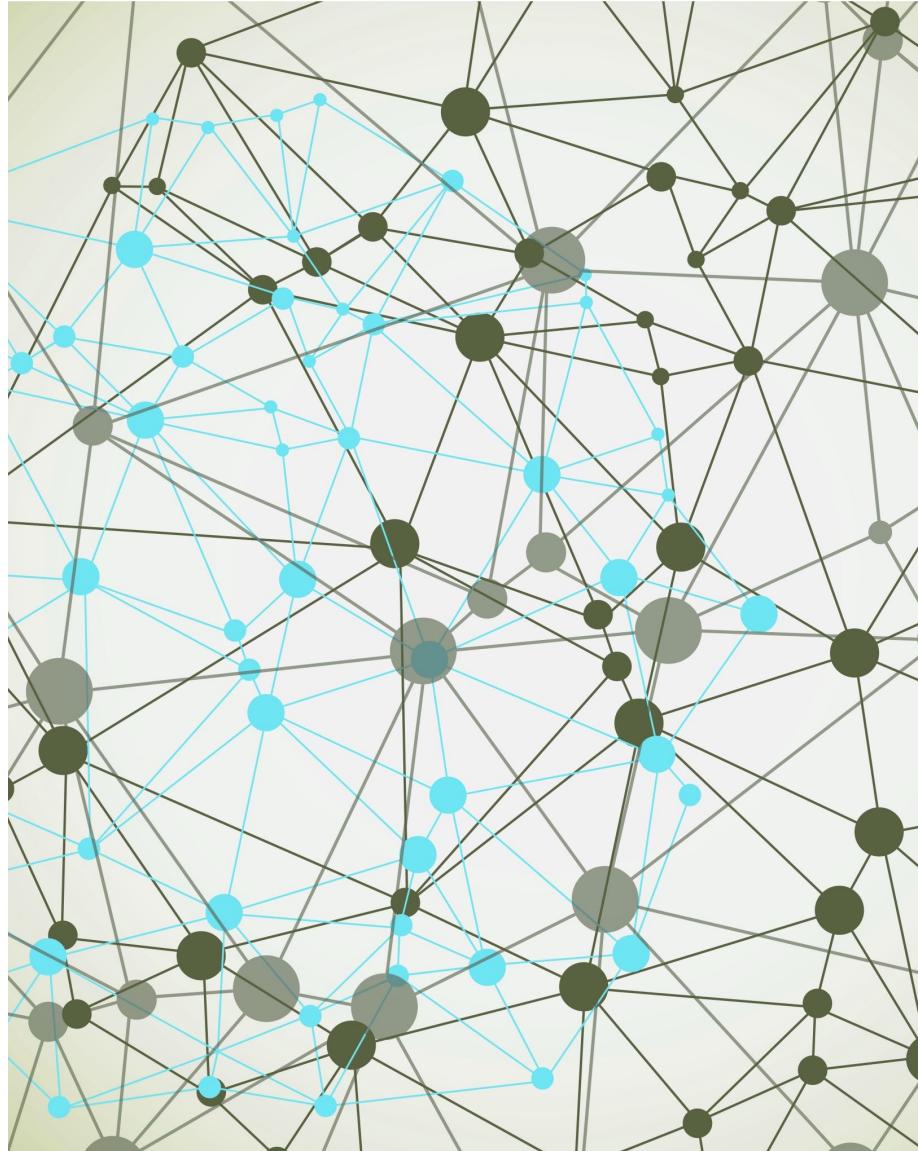
[https://en.wikipedia.org/wiki/Jimmy\\_Carter](https://en.wikipedia.org/wiki/Jimmy_Carter)





# 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



# Introduction

- Direct prompt vs. self-ask prompt.

# Data

## 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
- test set does not have public answers, so we instead take the development set to be our test set

## Dataset Modifications

- Fine-tuning datasets: direct no exemplars and self-ask with exemplars
- Direct data set "directly" provides the supporting facts and the question in the prompt
- Self-ask with exemplars provides two exemplars of self-ask reasoning before the supporting facts and question
  - Prompt ends with the instruction "Are follow up questions needed here:"
- Add example

# Models

Each baseline model is fine-tuned separately on the direct and self-ask with exemplars and datasets

Each model family has three models: baseline (not tuned), direct, and self-ask

Get rid of table

Baseline Model Comparison				
Model	Architecture	Pre-training Task	Fine-tuning	Parameters
OPT-125m	Decoder only	Auto regressive next word prediction	None	125m
T5 small	Encoder Decoder	- Auto encoding denoising task	None	60m
Flan-T5 small	Encoder Decoder	- Auto encoding denoising task	Instruction Tuned	60m



# Evaluation

## Metrics

### Rationale Generation

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

### Question Answering

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

## 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

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

### Cross attention

#### Facts:

Fact #0: Solo is a 2006 Australian film directed by Morgan O'Neill and starring Colin Friels.  
Fact #1: Morgan O'Neill (born 19 April 1973 in Sydney, Australia) is an Australian writer, director, actor and producer.

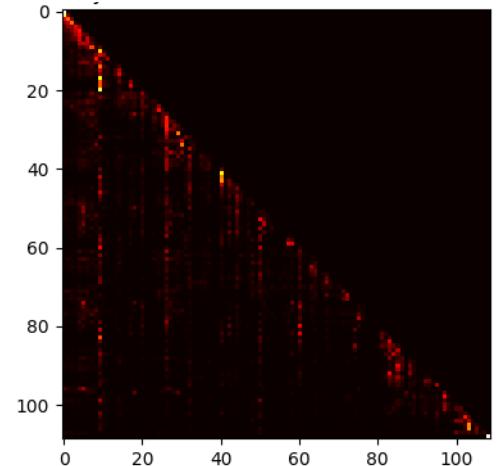
Question: What is the place of birth of the director of film Solo (2006 Film)?

### Self-Ask-Tuned Model

Yes.  
Follow up: What is the director of Solo?  
Intermediate answer: Morgan O'Neill  
Follow up: What is the place of birth of Morgan O'Neill?  
Intermediate answer: Sydney, Australia  
So the final answer is: Sydney, Australia

### Direct-Tuned Model

Sydney, Australia



### Decoder self-attention

# 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 ★

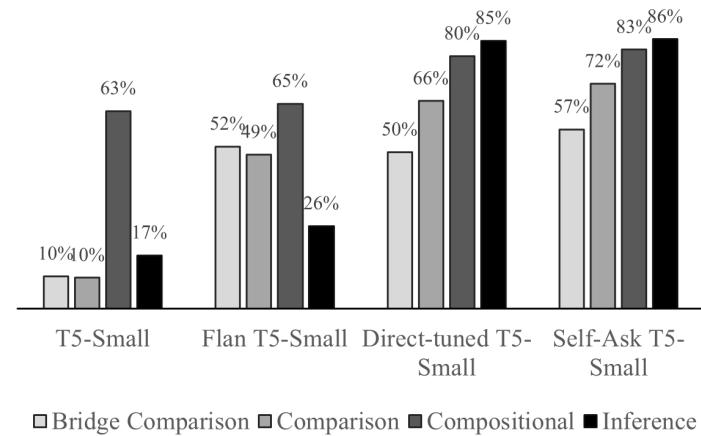
+21.5%

+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

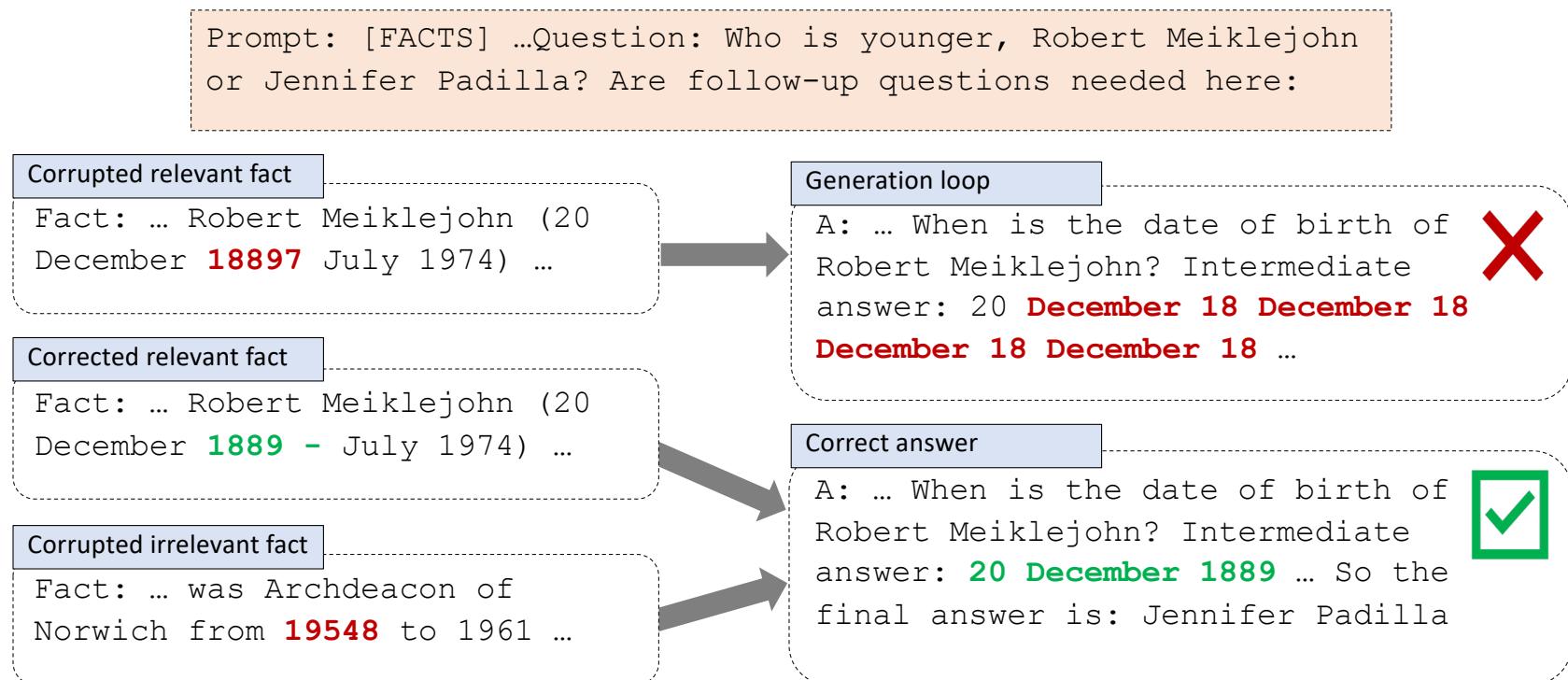
### Both

- 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?

# References

- Star Wars Image <https://www.imdb.com/title/tt0080684/>
- Jimmy Carter Image [https://en.wikipedia.org/wiki/Jimmy\\_Carter](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.](#)