Distilling Script Knowledge from Large Language Models for Constrained Language Planning

Siyu Yuan, Jiangjie Chen, Ziquan Fu, Xuyang Ge, Soham Shah, Charles Robert Jankowski, Yanghua Xiao, Deqing Yang
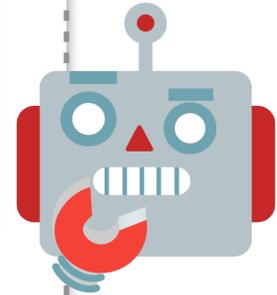
FUDAN UNIVERSITY

Brain Technologies Inc.

Language Planning

🍰 How to Make a Cake?

1. Gather your ingredients.
2. Preheat the oven to 325 °F (163 °C) and grease and flour a cake pan.
3. Cream the butter and sugar.
4. Add the eggs.
5. Stir in the cake flour.
6. Pour the batter into the pan.
7. Bake the cake for 1 hour 15 minutes.

Large language models (LLMs) can effectively decompose goals into steps

Constrained **Language Planning**

How to Make a *Strawberry* Cake?

**...Add strawberry jams
into the flour...**

How to Make a *Chocolate* Cake?

**...Add the cocoa powder
into the flour...**

Abstract goal can be inherited by different real-life
specific goals with multi-faceted ***constraints***

How do LLMs perform on Constrained Language Planning?

Dataset: wikiHow + Generated Constraints

Constraint Type 1: *Modifier*
Definition: A word, an adjective or a phrase that modifies or constrains an abstract goal.
Ex.1: Make a chocolate cake.
Ex.2: Make a pink cake.

| Constraint Type 2: *Method* |
|---|
| **Definition**: A tool or specified mode that controls the process for achieving the goal. |
| **Ex.1**: Make a cake with an oven. |
| **Ex.2**: Make a cake by using cake mix. |

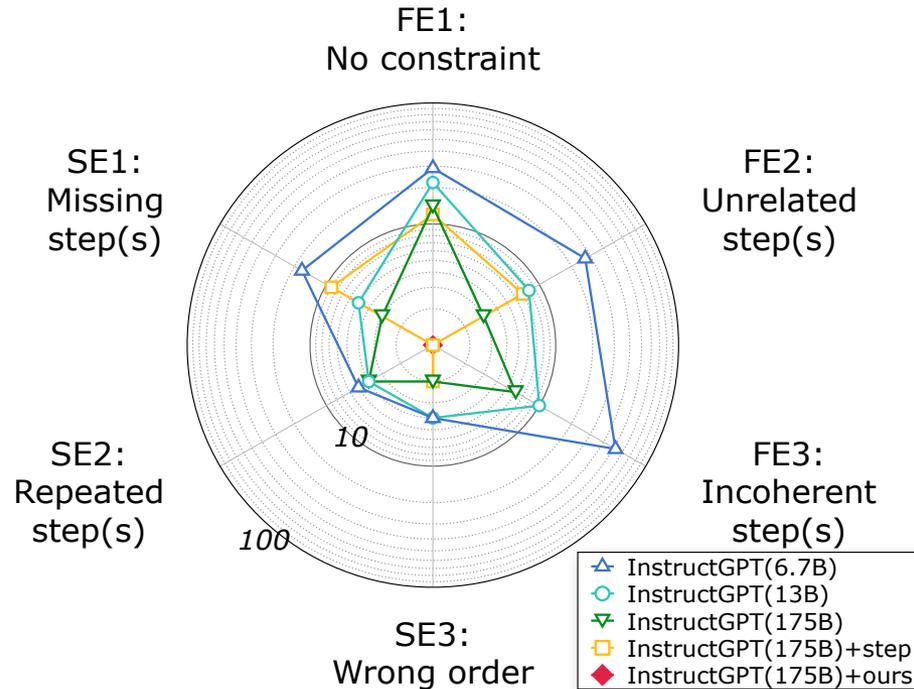| Constraint Type 3: *Intent* |
|---|
| **Definition**: An additional purpose or demand when completing the goal. |
| **Ex.1**: Make a cake for wedding. |
| **Ex.2**: Make a cake for diabetics. |

Can LLMs do Constrained Language Planning?
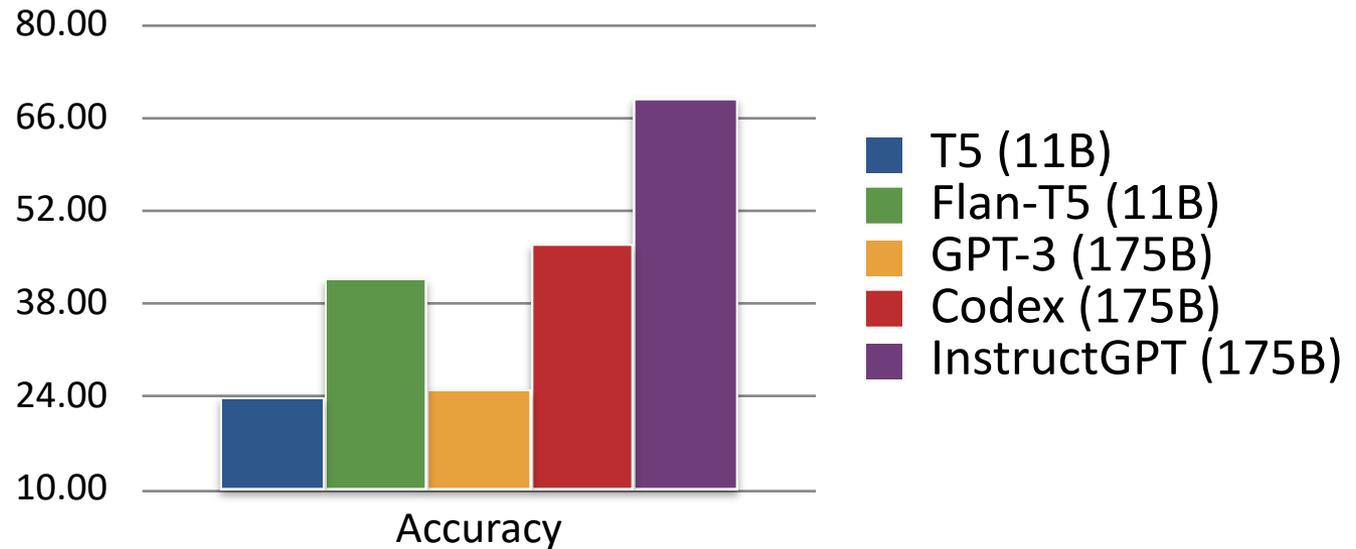
☐ All baselines achieve **unsatisfactory results** on planning for specific goals

What types of errors do LLMs usually make in this task?

☐ The semantic completeness (SE) in generated scripts is acceptable, but *the faithfulness to the constraints* (FE) can not be guaranteed

What kinds of goals do InstructGPT typically fail?

	6.7B	13B	175B	Step
Work	0.0	0.67	1.0	1.0
Sports	0.25	0.5	1.0	1.0
Relationships	0.0	0.5	1.0	1.0
Personal Care	0.14	0.33	0.5	0.6
Home	0.17	0.3	0.5	0.6
Traditions	0.25	0.4	0.5	0.57
Hobbies	0.1	0.42	0.6	0.67
Health	0.33	0.5	0.57	0.75
Food	0.28	0.66	0.78	0.78
Business	0.33	0.5	1.0	0.67
Family Life	0.0	0.0	1.0	0.67
Education	0.2	0.33	0.5	0.67
Electronics	0.25	0.38	0.6	0.62
Vehicles	0.0	0.0	0.75	0.5
Arts	0.33	0.29	1.0	0.5

❑ The planning performance of InstructGPTs *varies considerably* for goals of different categories

Method

Input: an abstract goal
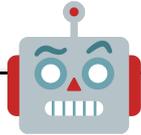
↓

Step 1
Generate *specific goals*
with InstructGPT via
in-context learning

Abstract Goal: *Make a cake*

↓ **+** *constraints*

Specific Goals:

G1(+modifier): *Make a **chocolate** cake* 🍰

G2(+method): *Make a cake **in a microwave*** 📟

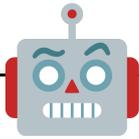**G3(+intent)**: *Make a cake **for a wedding*** 🎂

Method

Input: an abstract goal

Step 1
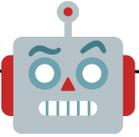Generate *specific goals* with InstructGPT via in-context learning

Step 2
Over-generate *candidate scripts* with InstructGPT via in-context learning

Abstract Goal: *Make a cake*

+ *constraints*

Specific Goals:

G1(+modifier): *Make a chocolate cake*

G2(+method): *Make a cake in a microwave*

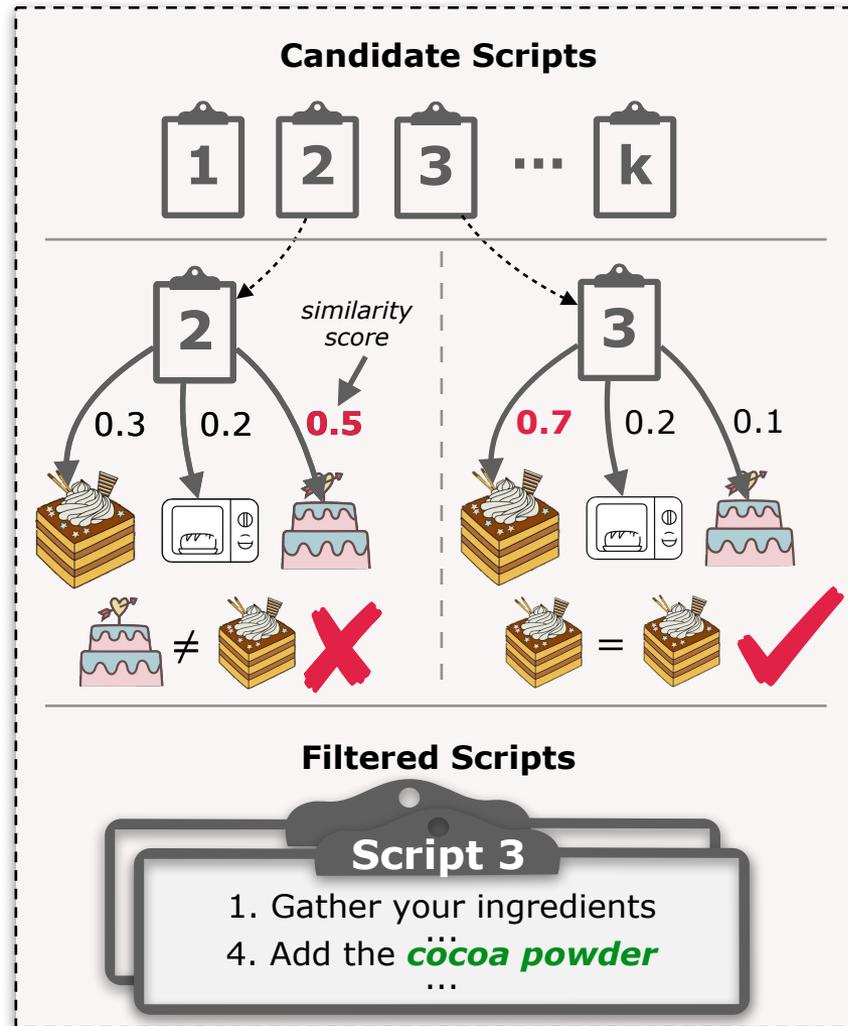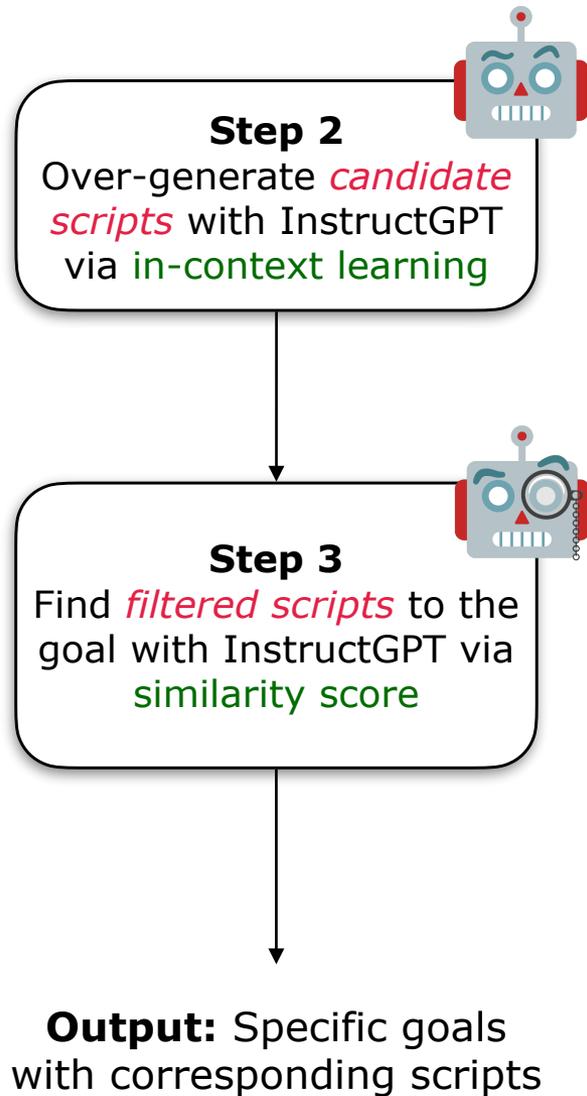**G3(+intent):** *Make a cake for a wedding*

Generate Plans for G1

Candidate Scripts

1   2   3   ⋯   k

Method

Step 2
Over-generate *candidate scripts* with InstructGPT via in-context learning

Step 3
Find *filtered scripts* to the goal with InstructGPT via similarity score

Output: Specific goals with corresponding scripts

Candidate Scripts

1 2 3 ... k

similarity score

2
0.3   0.2   **0.5**

3
0.7   0.2   0.1

≠ ❌

= ✔️

Filtered Scripts

Script 3
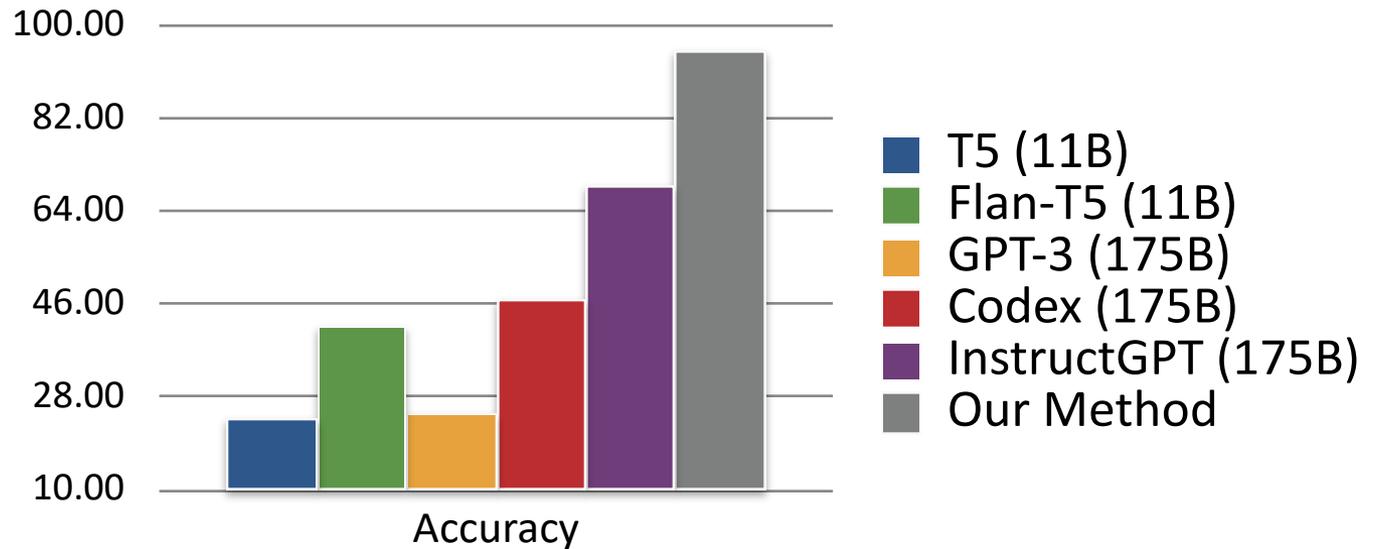1. Gather your ingredients
...
4. Add the ***cocoa powder***
...

Our Method Greatly Improves the Planning Quality

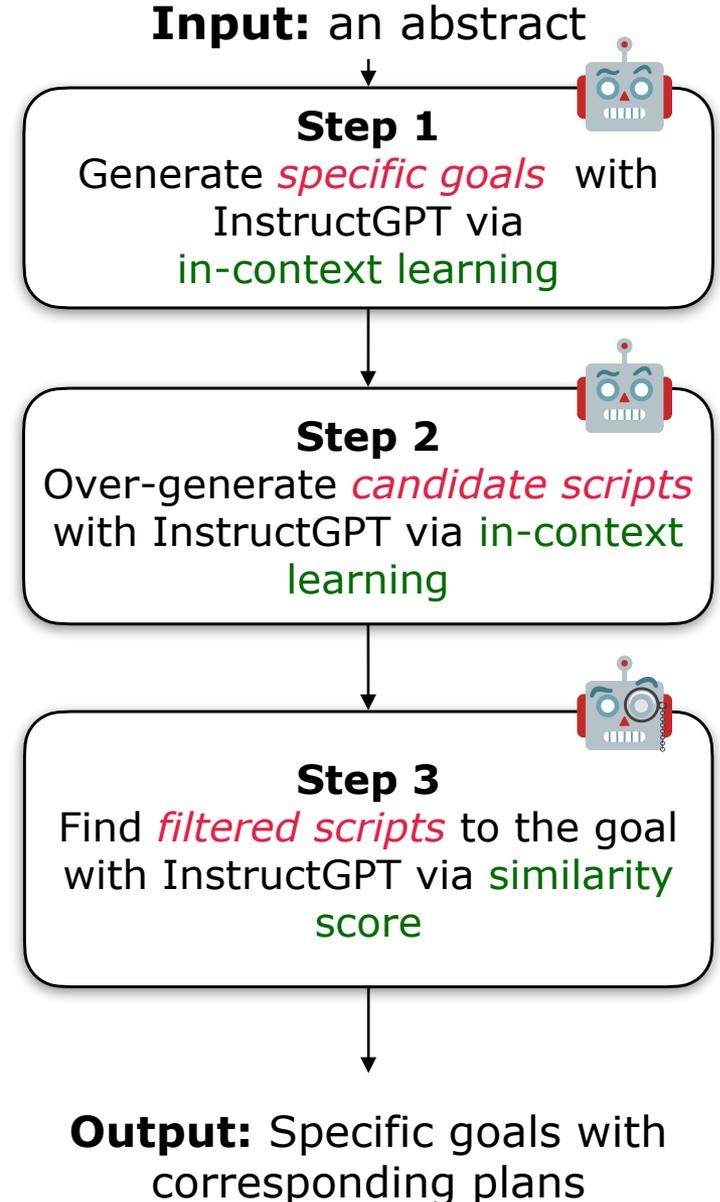With our method, InstructGPT can generate scripts of higher quality *by a large margin*.
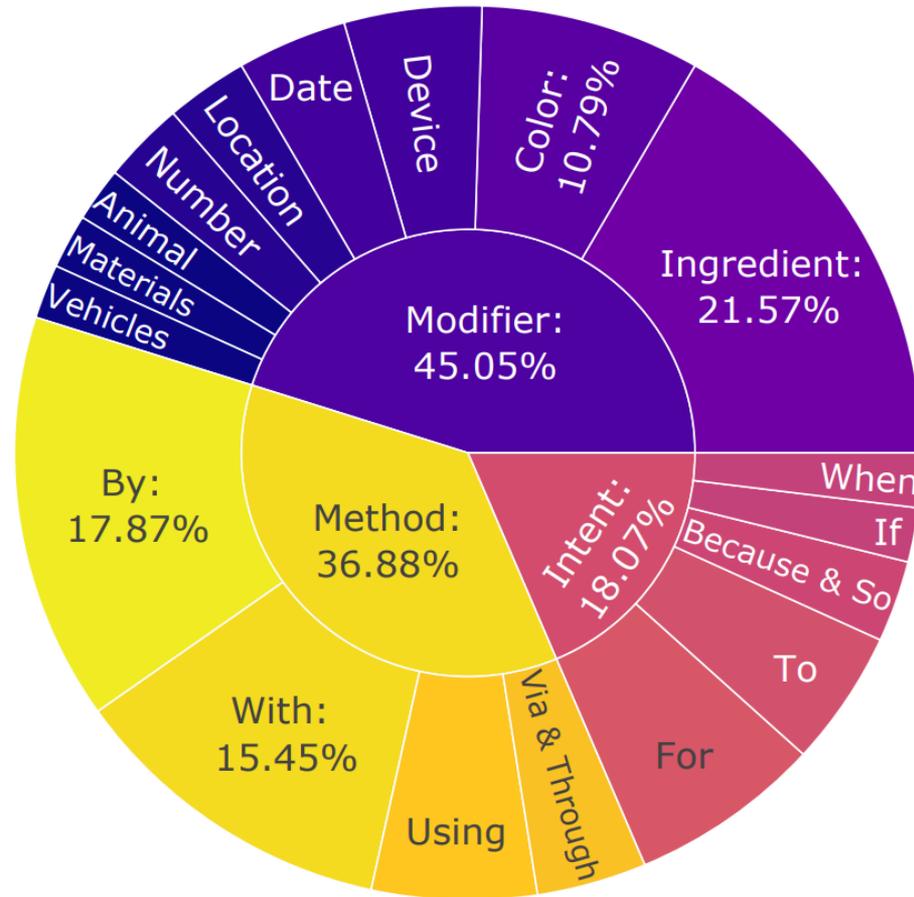
Script Distillation from LLMs

Motivation

☑ To enable constrained language planning ability for *smaller models*.

Method

☑ Follow the idea of *symbolic knowledge distillation*

☑ Generated *55,000 Scripts with constraint* from LLMs based on our method => **Coscript** Dataset

☑ Humans *annotate* validation and test set.

Input: an abstract

Step 1
Generate *specific goals* with InstructGPT via in-context learning

Step 2
Over-generate *candidate scripts* with InstructGPT via in-context learning

Step 3
Find *filtered scripts* to the goal with InstructGPT via similarity score

Output: Specific goals with corresponding plans

Constraint Analysis

☐ **Coscript** shows high heterogeneity and pluralism in the generated specific goals.
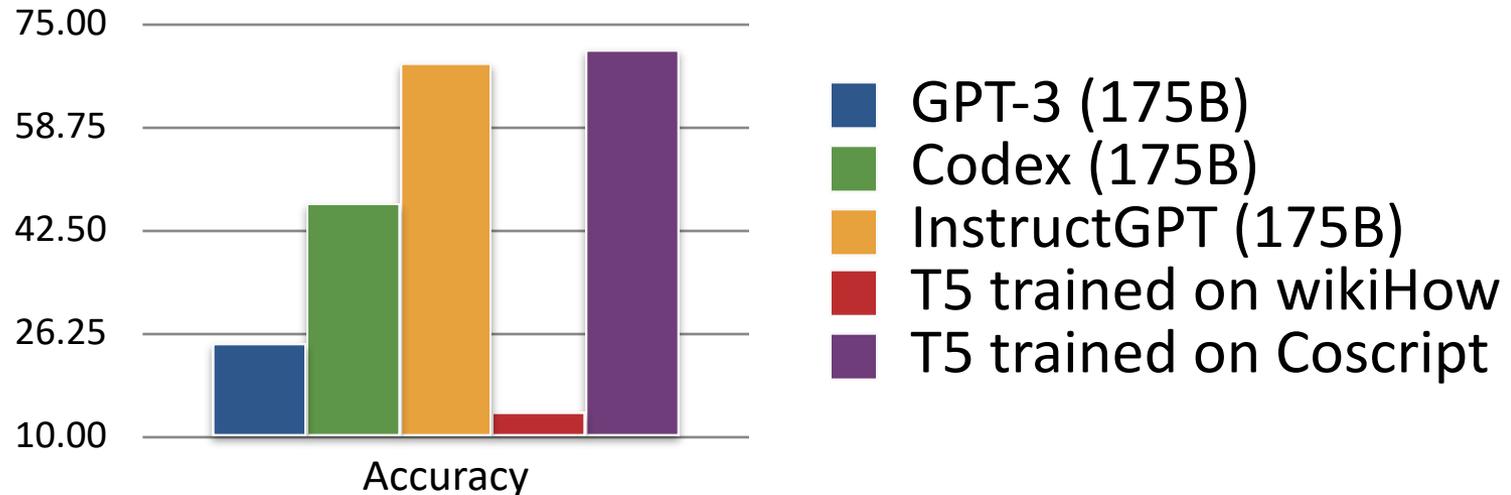
Coscript for Smaller Language Models

- **Datasets**
  - **Coscript**
  - wikiHow

- **Metrics**
  - Faithful: *DeBERTa (v3 large) model* to decide whether the generated texts are faithful to the constraints
  - Automatic metrics: ROUGE, BLEU, BERTScore

Specialized Models vs. LLMs

■ GPT-3 (175B)
■ Codex (175B)
■ InstructGPT (175B)
■ T5 trained on wikiHow
■ T5 trained on Coscript

☐ Smaller LMs fine-tuned on **Coscript** can generate higher quality scripts than LLMs

Summary and Takeaways

- Establish the *constrained language planning* problem.
- Evaluate constrained language planning ability of LLMs and develop an *over-generate-then-filter* method for LLMs
- Use LLMs to generate a high-quality *script dataset* (CoScript) for constrained language planning.
- **Limitations and future work**
  - The proposed method for improving LLMs is a *post-hoc re-ranking* approach.
  - **Coscript** only inherits from an abstract one with *one extra constraint*.
  - **Coscript** dataset can be a valuable resource to advance the research on language planning *with more complex and diverse goals and constraints*

Distilling Script Knowledge from Large Language Models for Constrained Language Planning
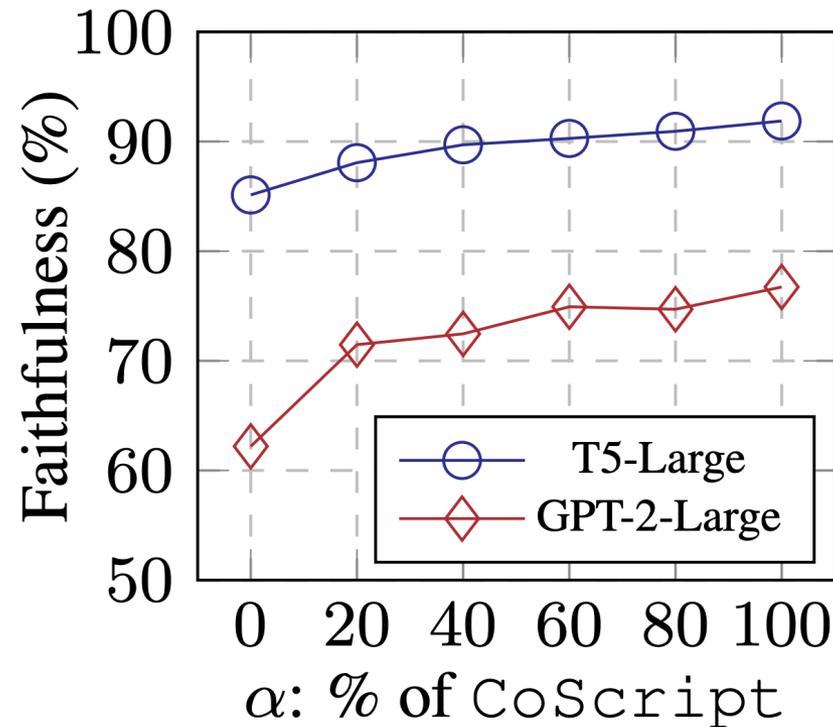

Coscript Website
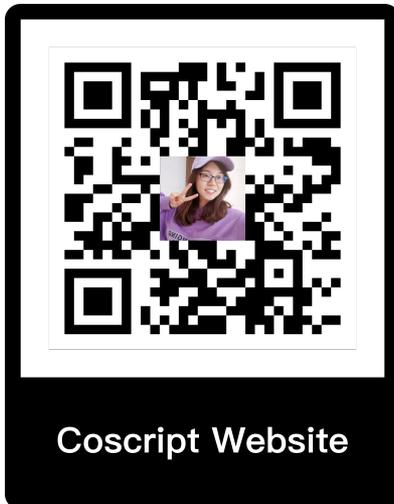
Siyu Yuan, Jiangjie Chen, Ziquan Fu, Xuyang Ge, Soham Shah, Charles Robert Jankowski, Yanghua Xiao, Deqing Yang

✉ syyuan21@m.fudan.edu.cn

https://github.com/siyuyuan/coscript

17

Coscript vs. wikiHow

☐ Adding more data from **Coscript** consistently improves model performance in **constraint faithfulness**.