UBIQUITOUS KNOWLEDGE PROCESSING

TECHNISCHE UNIVERSITÄT DARMSTADT

Using Code to Improve Reasoning in Large Language Models

Haritz Puerto

TU Darmstadt

Cyber-Physical Health and Assistive Robotics Technologies - CHART Seminar

University of Nottingham

15 May 2024

Evaluating Reasoning in LLMs

- Math
- Logic
- Commonsense
- Symbolic Reasoning

- What is reasoning? …

Math Word Problems (free response)

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm^3, which is less than water. Thus, a pear would float. So the answer is no.

CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go?
Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

SayCan (Instructing a robot)

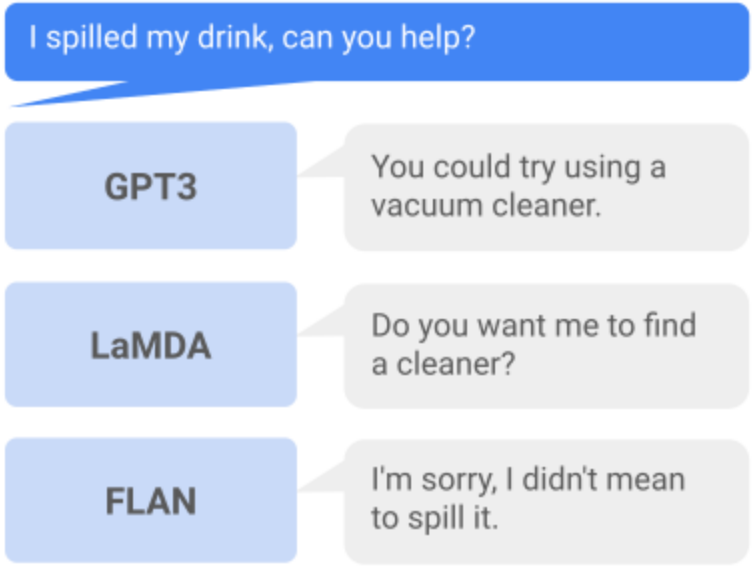
Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.
Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

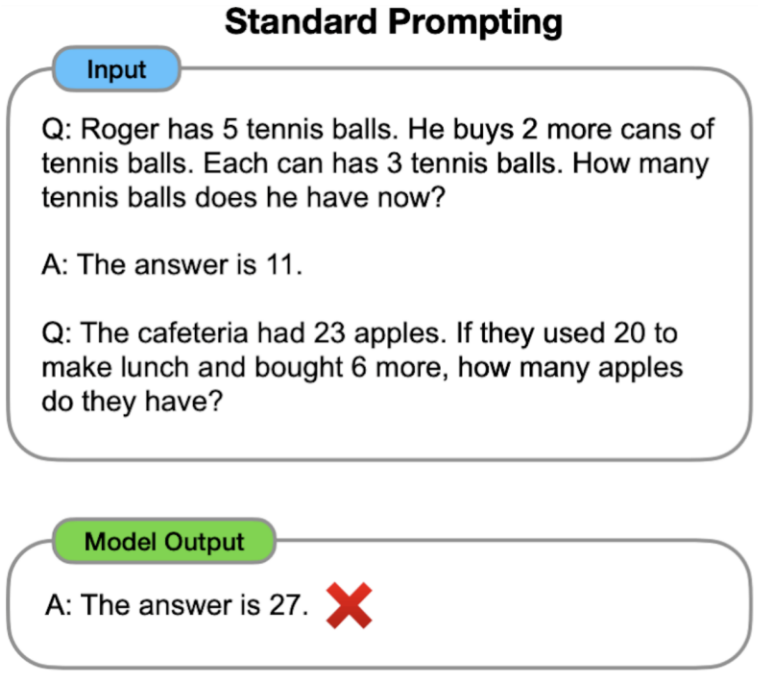
Examples from Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in neural information processing systems* 35 (2022): 24824-24837.

Robot Planning with LLMs?

- By default, LLMs cannot be used for robotics tasks…
- SayCan is a dataset that maps natural language user instructions (task) to a list of robot instructions to solve the task
- https://say-can.github.io/img/palm_saycan_teaser_compressed.mp4



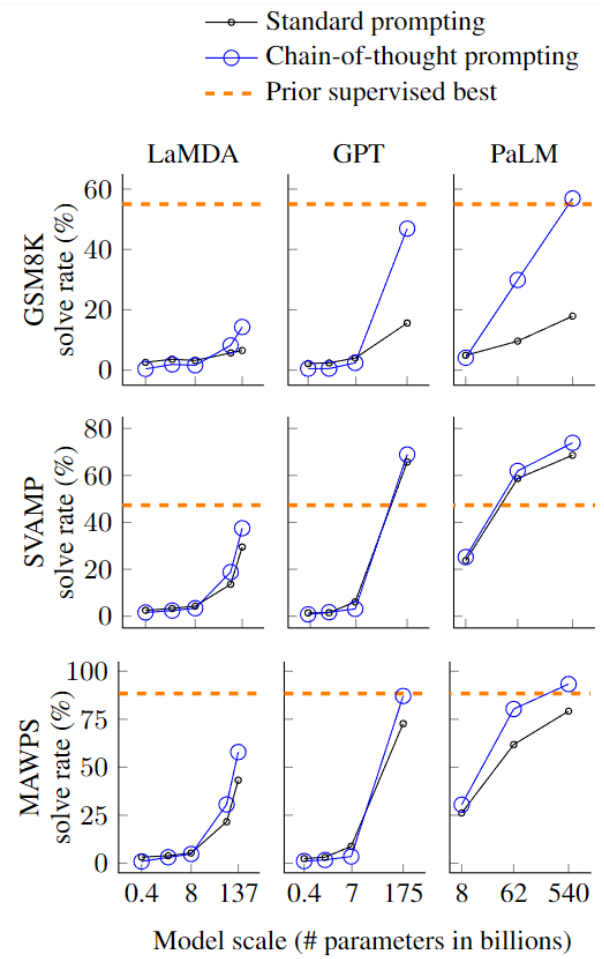
Eliciting Reasoning Abilities in LLMs



Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in neural information processing systems* 35 (2022): 24824-24837.

Chain-of-Thought Emerges on Scale

- Many interesting properties of LLMs *emerge* on scale
  - Math reasoning
  - Symbolic reasoning
  - Chain of Thought



Chain of Though (CoT) is not Perfect 😢

TECHNISCHE UNIVERSITÄT DARMSTADT

- CoT improves most types of reasoning in LLMs 🥳
- However, LLMs still have difficulties in some tasks
  - Especially if they are small LMs
- 🧮 Arithmetic operations are particularly difficult for them

How can we *delegate* these *difficult tasks* to expert systems ❓

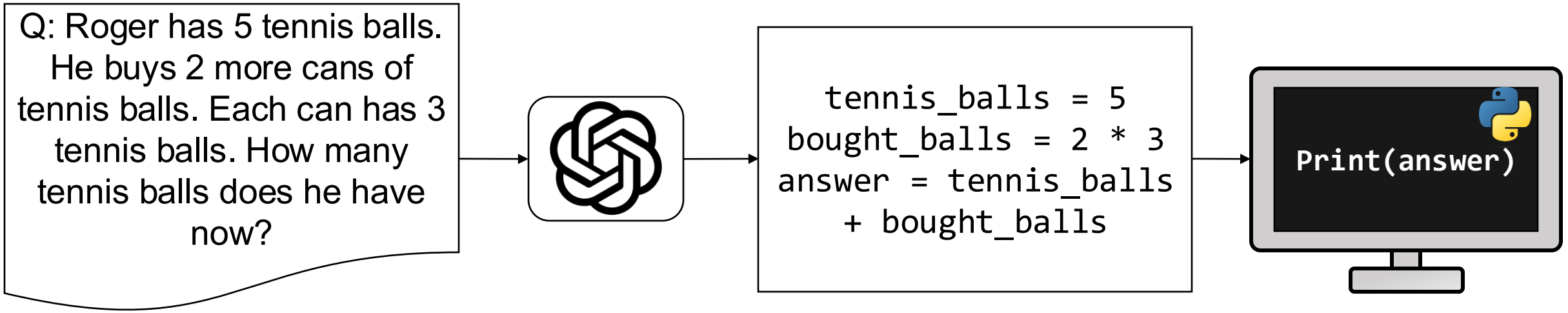
TECHNISCHE UNIVERSITÄT DARMSTADT

Code-Aided LLMs

- Programming languages can solve many tasks perfectly
  - Like math problems
  - Logical reasoning (if A & B, then C)

Formalize those tasks and solve them with a code interpreter or an API

Code-Aided LLMs



Answer correctness guaranteed*

*if the code reflects the input correctly

Large Gains on Math, Robot Planning, Logic, and Question-Answering Tasks

TECHNISCHE UNIVERSITÄT DARMSTADT

Method	Math Word Problems					Planning	Multi-hop QA			Relation
	GSM8K	SVAMP	MultiArith	ASDiv	AQuA	SayCan	StrategyQA	Date	Sport	CLUTRR
Greedy Decoding										
Standard	19.6	69.5	43.8	72.1	31.5	82.5	63.9	51.3	71.9	42.0
CoT	63.3	77.3	96.5	80.0	42.1	86.4	**72.5**	59.9	98.6	48.5
LtM	38.3	80.3	74.0	76.5	40.6	77.7	72.2	76.6	**99.5**	47.2
Faithful CoT (ours)	**72.3**	**83.4**	**98.8**	**80.2**	**47.2**	**89.3**	63.0	**81.6**	99.1	**58.9**

	GSM-SYS	GSM	ALGE	LSAT	BOARD	CLUTRR	PROOF	COLOR	REGEX
code-davinci-002 (greedy decoding)									
STANDARD	21.0	22.2	45.9	22.0	44.6	41.2	76.6	75.7	–
COT	46.5	62.7	53.6	23.5	60.7	40.8	80.1	86.3	–
PROGLM	43.4	**72.7**	52.3	–	–	58.9	83.7	95.1	39.1
SATLM	**69.4**	71.8	**77.5**	**35.0**	**79.4**	**68.3**	**99.7**	**97.7**	**41.0**

Ye, Xi, et al. "Satlm: Satisfiability-aided language models using declarative prompting." *Advances in Neural Information Processing Systems* 36 (2024).

Faithful Chain-of-Thought Reasoning (Lyu et al., IJCNLP-AACL 2023)

TECHNISCHE UNIVERSITÄT DARMSTADT

Why and How Code Helps LLMs?

- LLMs delegate “difficult” tasks that can be solved algorithmically
- But, … is it the only way?

- We can’t solve everything with Python…

TECHNISCHE UNIVERSITÄT DARMSTADT

Why and How Code Helps LLMs?

- There has been a long-standing hypothesis that training on code improves LLMs
  - (one of the potential reasons for the good performance of GPT3)
- This might be because:
  1. Code is an implicit form of grounding [1]
  2. Code represents how people think to solve a problem
  3. Long dependencies in code promote entity tracking

🤔 Does representing a natural language task as code improve LLMs' reasoning abilities?

[1] https://gist.github.com/yoavg/59d174608e92e845c8994ac2e234c8a9 Yoav Goldberg, 2023

TECHNISCHE UNIVERSITÄT DARMSTADT

Code Prompting Elicits Conditional Reasoning Abilities in Text+Code LLMs

Haritz Puerto[1], **Martin Tutek**[1], **Somak Aditya**[2], **Xiaodan Zhu**[1,3], **Iryna Gurevych**[1]
[1]Ubiquitous Knowledge Processing Lab (UKP Lab),
TU Darmstadt and Hessian Center for AI (hessian.AI)
[2]IIT Kharagpur, [3]Queen's University
https://www.ukp.tu-darmstadt.de

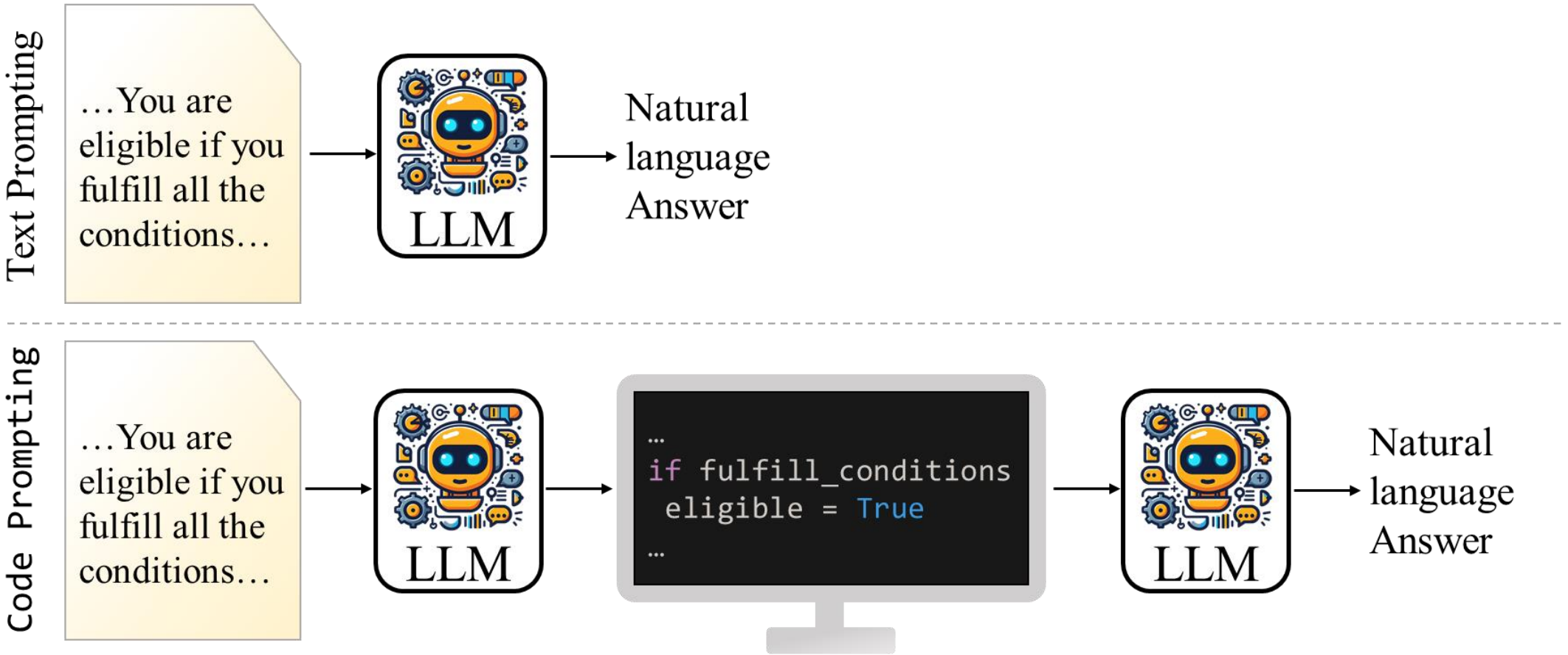
Paper
(Under review)

Code

Does code also improve performance w/o a code interpreter?

Can we elicit reasoning abilities in LLMs by merely changing the input format (text → code) ?

Code Prompting



TECHNISCHE UNIVERSITÄT DARMSTADT

Code Format

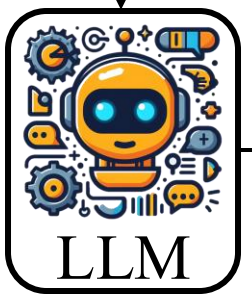
- Code as close as possible to the original text
- Use a simplification of Python
- Only Boolean variables or list of strings
- No loops, functions, or class definitions
- **Keep** the **original text** as code comments

Text Prompt

Question: Ann's husband passed away. She needs help for the burial in the UK. Can she be eligible for funeral expenses payment?

Doc: You can get a Funeral Expense Payment <u>if all of the following apply</u>:

- You meet the rules on your relationship with the deceased
- You're arranging a funeral in the UK

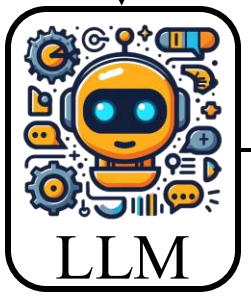


Answer: No

Code Prompt

```
# Question: Ann's husband passed away. She needs
help for the burial in the UK. Can she be eligible
for funeral expenses payment?
husband_pass_away = True
needs_help_for_burial_in_UK = True
eligible_funeral_expenses_payment = None # This is
the question

# Doc:
# You can get a Funeral Expense Payment…
if (meet_rules_relationship and
    funeral_in_UK):
 eligible_funeral_expenses_payment = True
```



Answer: Yes

Methodology & Scope

- Text+Code LLMs: GPT 3.5, Mixtral, Mistral
- Task: **Conditional Question Answering**
  - Datasets:
    - ConditionalQA
    - ShARC
    - BoardgameQA

ConditionalQA: A Complex Reading Comprehension Dataset with Conditional Answers (Sun et al., ACL 2022)

Document:

Section 1: Overview

You can get a Funeral Expense Payment if all of the following apply:

- You meet the rules on your relationship with the deceased
- you're arranging a funeral in the UK

….

Section 2: What you will get ….

Section 3: Your relationship

You must be one of the following:

- the partner of the deceased when they died
- a close relative or close friend
- ….

You might not get a Funeral Expenses Payment if another close relative of the deceased (such as a sibling or parent) is in work.

….

Question:

Scenario: Ann lives in the UK. Her husband has succumbed to cancer. She needs help to give her late husband a decent burial.

Question: Can she be eligible for funeral expenses payments?

Answer:

Answer: Yes
Conditions: ["*you're arranging a funeral in the UK*"]

Answer: No
Conditions: ["*You might not get a Funeral Expenses Payment if another close relative of the deceased …*"]

TECHNISCHE UNIVERSITÄT DARMSTADT

Task Performance

Model	Prompt	CondQA	ShARC	BGQA-1	BGQA-2	BGQA-3	ΔCP
				Test Set			
GPT 3.5	Text	58.70	**62.95**	51.15	37.42	27.77	8.42
	Code	**60.60**	54.98	**58.67**	**55.56**	**50.29**	
Mixtral	Text	**48.17**	53.77	**56.38**	39.64	30.15	4.22
	Code	44.73	**59.06**	53.33	**47.39**	**44.72**	
Mistral	Text	**35.74**	43.60	47.40	48.78	47.86	2.74
	Code	33.28	**49.92**	**53.80**	**51.27**	**48.79**	

- Code prompts performs best on most datasets (11/15)
- All models achieve best results with code prompts on most datasets

TECHNISCHE UNIVERSITÄT DARMSTADT

Task Performance

Model	Prompt	CondQA	ShARC	BGQA-1	BGQA-2	BGQA-3	ΔCP
				Test Set			
GPT 3.5	Text	58.70	**62.95**	51.15	37.42	27.77	8.42
	Code	**60.60**	54.98	**58.67**	**55.56**	**50.29**	
Mixtral	Text	**48.17**	53.77	**56.38**	39.64	30.15	4.22
	Code	44.73	**59.06**	53.33	**47.39**	**44.72**	
Mistral	Text	**35.74**	43.60	47.40	48.78	47.86	2.74
	Code	33.28	**49.92**	**53.80**	**51.27**	**48.79**	

- BGQA-2, 3 are the most reasoning-intensive tasks. The gains are the largest here

Ablation: Implicit Text Simplification?

- Are the performance gains due to the *implicit text simplification* obtained from the code format? 🤔
  - Key entities (variables) and conditions are explicitly stated 👈
- We transforms the original text into atomic statements and use them instead of code

- We back-translate the generated code and use it instead of code

Original Sentence

Applying for the legal right to deal with someone's property, money, and possessions (their estate) when they die is called applying for probate.

Atomic Statements

Applying for the legal right is a process

The someone is a person who has died.

The process is called 'applying for probate'.

The legal right is to deal with someone's property, money, and possessions.

The property, money, and possessions are collectively called the 'estate'.

Original Transformation

Question: Ann's husband passed away. She needs help for the burial in the UK. Can she be eligible for funeral expenses payment?

Doc: You can get a Funeral Expense Payment if all of the following apply:

- You meet the rules on your relationship with the deceased
- You're arranging a funeral in the UK

```
# Question: Ann's husband passed away. She needs
help for the burial in the UK. Can she be eligible
for funeral expenses payment?
husband_pass_away = True
needs_help_for_burial_in_UK = True
eligible_funeral_expenses_payment = None # This is
the question

# Doc:
# You can get a Funeral Expense Payment…
if (meet_rules_relationship and
    funeral_in_UK):
 eligible_funeral_expenses_payment = True
```

Back-translation

Question: Ann's husband passed away. She needs help for the burial in the UK. Can she be eligible for funeral expenses payment?

Doc: if you meet the rules on your relationship with the deceased and you are arranging a funeral in the UK, you can get a Funeral Expense Payment

Ablation: Implicit Text Simplification?

Dataset	Δ Atomic St.	Δ Code → NL
CondQA	−2.66	−4.72
BGQA-1	−4.37	−1.43
BGQA-2	−8.72	−5.39
BGQA-3	−19.26	−3.68

- Natural Language Text resembling code does not yield performance boosts
- Code prompts enhance LLM performance beyond mere text simplification 😮

Ablation: Code Semantics are Important

- So it is important to have code to elicit LLM's reasoning abilities
- But, just showing any code is important? 🤔
- Do we need to keep the original text as code comment too? 🤔

Ablations:

- Anonymized code: instead of `if employed` → `if var1`
- Random code: code with no relation to the original instance
- Remove text comments

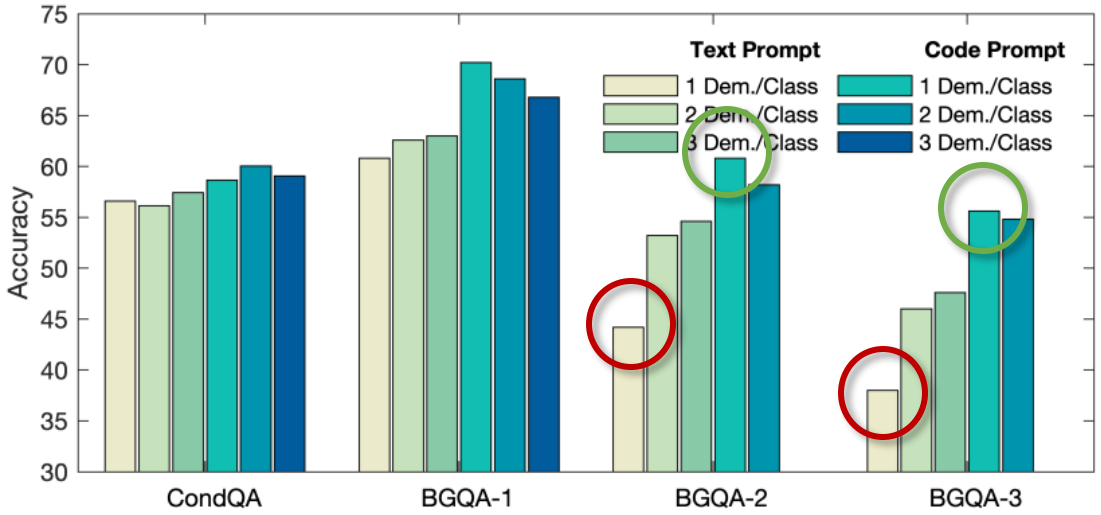
Ablation: Code Semantics are Important

Prompt	CQA	CQA-YN	BG_1	BG_2	BG_3
Anonym.	−1.62	−2.90	−6.60	−4.80	−4.00
Random	−3.40	−2.67	−7.40	−9.20	−9.80
- Comments	N.A.	−14.02	−16.70	−16.20	−5.20

- Removing the original text instance (as comments) drops performance a lot → LLM uses the text instance to answer!
- Anonymized code drops performance
- Random code drops performance even more
  - Performance is close to text prompts → LLM ignores the code

Code Prompting is More Efficient ⚠️

- We prompt LLMs with 1-3 demonstrations per class (Yes/No/Span)
- Gap is largest when using 1 demonstration
- Text Prompts needs > 1



Code Prompts are sample efficient

Code Prompting Improves Variable State Tracking ⚠️

- Pretraining on code might improve entity tracking in LLMs

- After each sentence in the answer output, we check whether the model remembers the initial context

Question: Ann's husband passed away. She needs help for the burial in the UK. Can she be eligible for funeral expenses payment?

Doc: You can get a Funeral Expense Payment if all of the following apply:
- You meet the rules on your relationship with the deceased
- You're arranging a funeral in the UK

Answer: …
Is it true that Ann's husband passed away?
Is it true she needs help for the burial in the UK?

Code Prompting Improves Variable State Tracking ⚠️

TECHNISCHE UNIVERSITÄT DARMSTADT

- Code Prompts yield much better results at remembering the initial facts of the context! 😮

Dataset	Correct Ans. Text	Correct Ans. Code	Incorrect Ans. Text	Incorrect Ans. Code
CondQA	71.08	**4.39**	60.79	**11.39**
BGQA-1	39.33	**8.84**	51.65	**22.12**
BGQA-2	44.79	**15.04**	52.54	**24.75**
BGQA-3	54.01	**14.21**	52.13	**16.98**

Memory errors on the questions about the context of the question (lower is better)

Code Prompting Improves Variable State Tracking ⚠️

- Why?
- Intuition:
  - LLMs have many capabilities
  - They can use a limited set at the same time
  - They chose the ones relevant for the task they assume they are given
  - Manipulating the input format of the prompts can trigger certain abilities and thus improve performance on your task

Dataset	Correct Ans. Text	Correct Ans. Code	Incorrect Ans. Text	Incorrect Ans. Code
CondQA	71.08	**4.39**	60.79	**11.39**
BGQA-1	39.33	**8.84**	51.65	**22.12**
BGQA-2	44.79	**15.04**	52.54	**24.75**
BGQA-3	54.01	**14.21**	52.13	**16.98**

Takeaways

- You can delegate some “difficult tasks” to a code interpreter 💻
  - Make sure the generated code is correct 🧐
- You can be more flexible and translate a natural language task into code and prompt the LLM with the generated code
  - No need for code execution 🚀
  - Improves reasoning abilities 🧠
  - Improves variable state tracking 🔎
  - It’s more efficient (in terms of #demonstrations required) ✊

Thank you!

haritz.puerto@tu-darmstadt.de