Code Promtpting Elicits Reasoning Ablities in Text+Code LLMs

TECHNISCHE UNIVERSITÄT DARMSTADT [1]

UBIQUITOUS KNOWLEDGE PROCESSING [1]

INGENUITY LABS RESEARCH INSTITUTE at Queen's University [3]

[2]

Haritz Puerto[1], Martin Tutek[1], Somak Aditya[2], Xiaodan Zhu[1,3], Iryna Gurevych[1]

Research Questions

I. Can input ***form*** rather than *content* affect reasoning abilities of LLMs? **Yes**!
II. Can formatting input as ***code*** elicit conditional reasoning abilities in LLMs? **Yes**!
III. What ***capabilities*** of LLMs does representing input as code improve?

Method

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

LLM → CoT + Answer

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 # question

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

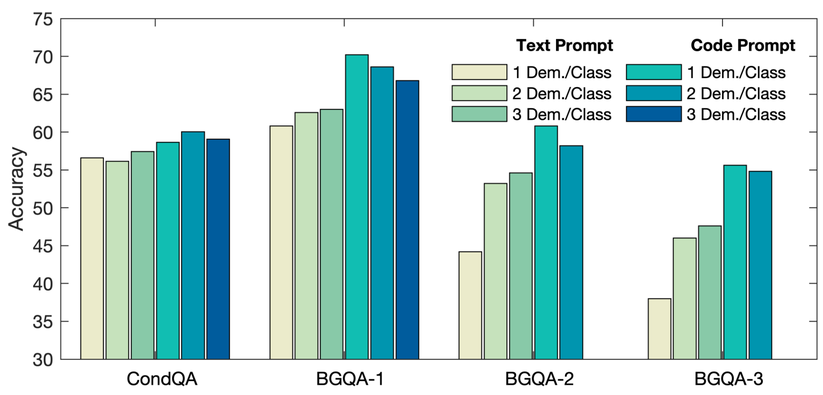
Python-like code: only if blocks and Boolean vars
The code is run by the LLM

LLM → CoT + Answer

Code Prompting Outperforms Text Prompting

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

Code Prompts Are More Sample-Efficient



Code Prompts Improve Variable Tracking

Q: Does the LLM remember the facts from the question better with Code Prompts while generating the CoT answer?

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 Error Rate. Lower is better

Code Syntax Elicits Reasoning Abilities

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

- **Atomic Statements**: create very short sentences with unitary facts (≈ var definitions)
- **Backtransform** the code into NL to check if some code semantics cause performance increase

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

- **Anonymous** code: change var names into the form var_i
- **Random**: Change code by any other random code (semantic mismatch)

Funded by erc DFG Deutsche Forschungsgemeinschaft German Research Foundation ATHENE National Research Center for Applied Cybersecurity