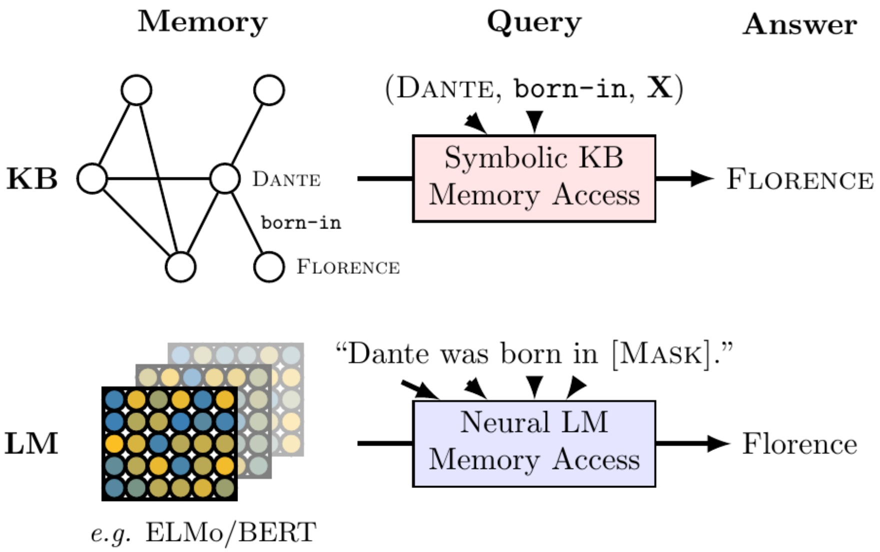
Assessing Reliability of Knowledge in LLMs

Weixuan Wang, Barry Haddow, Alexandra Birch, Wei Peng

Extracting Knowledge from LLMs

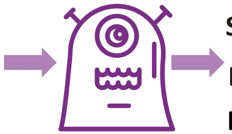


Language Models as Knowledge Bases?

Fabio Petroni[1] **Tim Rocktäschel**[1,2] **Patrick Lewis**[1,2] **Anton Bakhtin**[1]
Yuxiang Wu[1,2] **Alexander H. Miller**[1] **Sebastian Riedel**[1,2]

How Accurate is this Knowledge?

Which country is the location of Sion?
Which country is Sion situated in ?
Sion is located in Switzerland. True or False?

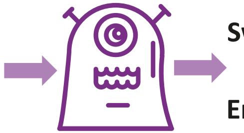


Switzerland ✓
England ✗
False ✗

(a) Prompt framing effect

Which country is the location of Sion?

England. Which country is the location of Sion?

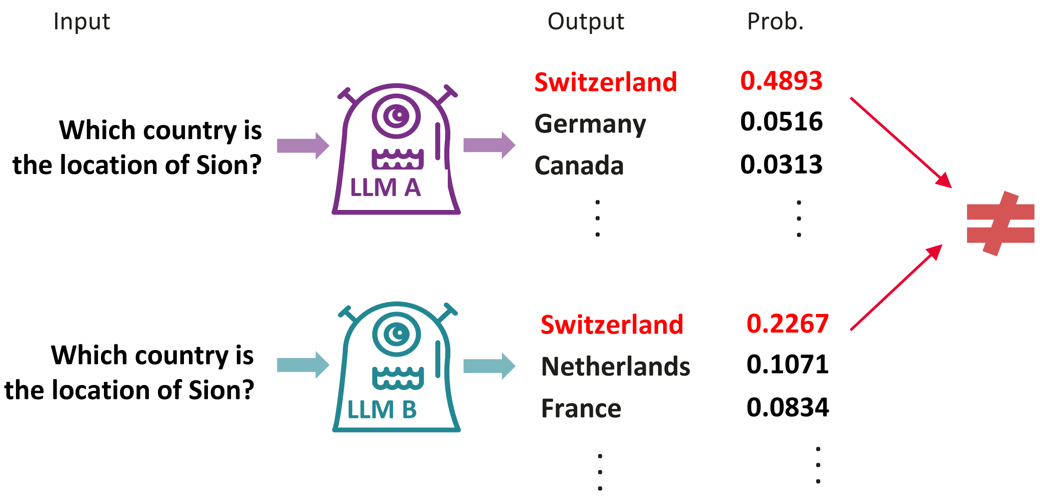


Switzerland ✓
England ✗

(b) In-context interference effect

→ Accuracy Instability

Accuracy of Top-1 is not Sufficient



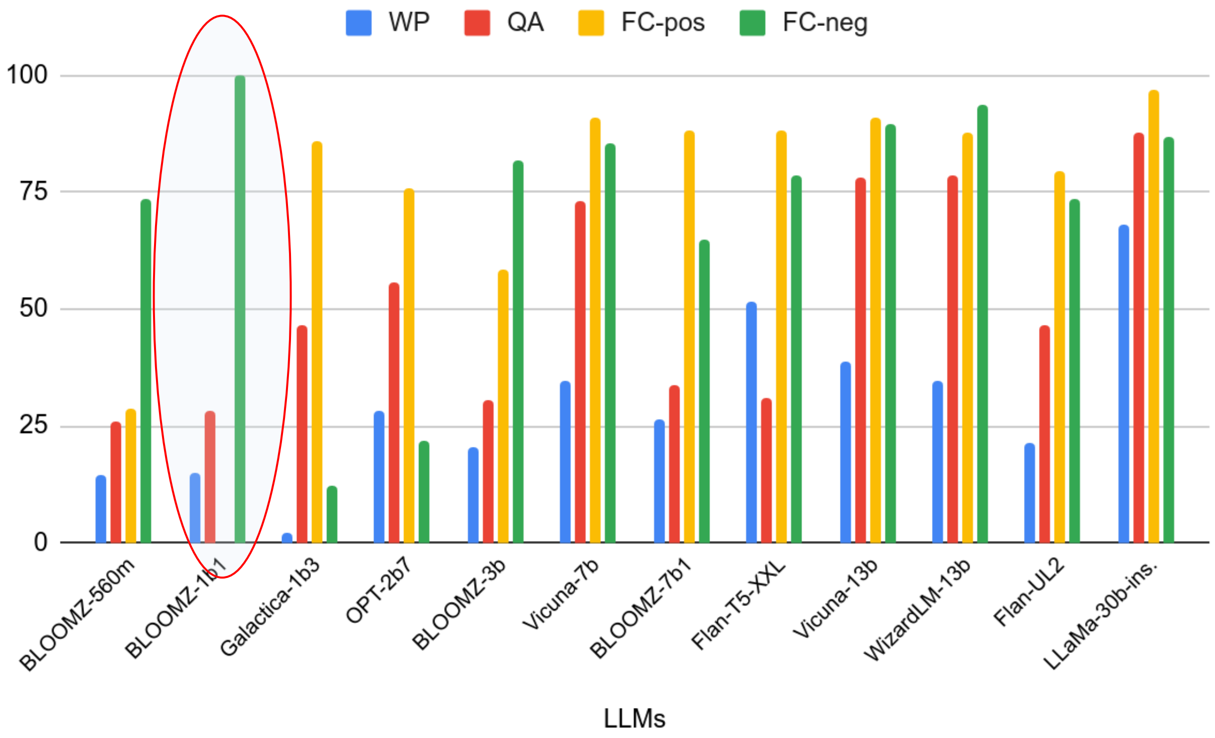
Same Accuracy ≠ Same Uncertainty

Probing LLMs for Accuracy Instability

- We extract facts from the T-REx dataset (Elsahar et al. 2018).
  - It contains relation triples e.g. <Rome, Italy, located-in>
- We use 7 paraphrases for each of the prompt frames

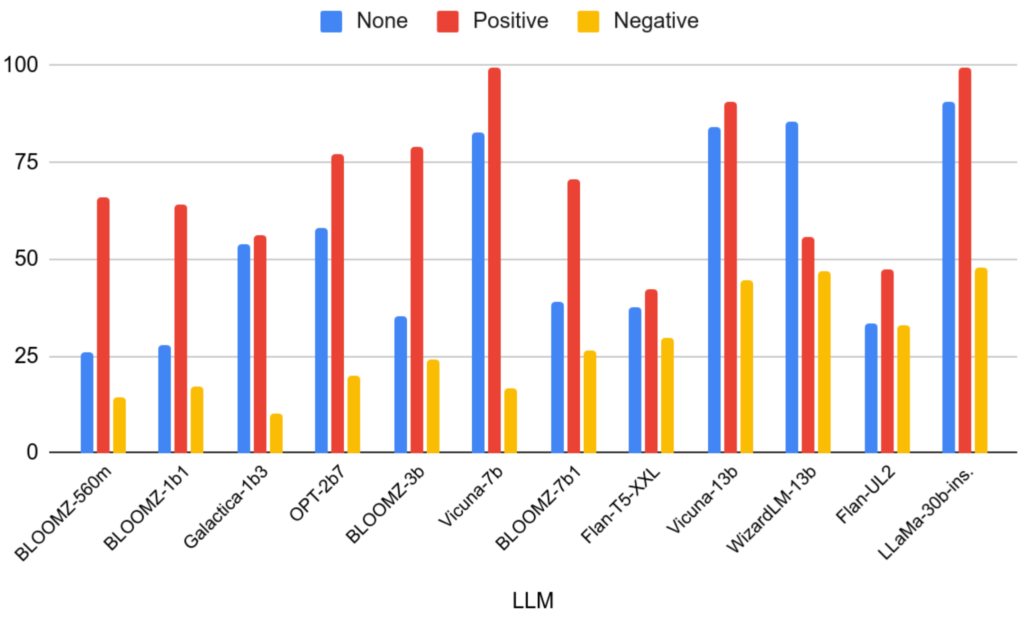
Prompt Frames	
Word Prediction (WP)	Rome is located in ___
Question Answering (QA)	Which country is Rome located in? ___
Fact Checking positive (FC-pos)	Statement: Rome is located in Italy. The statement is True or False?
Fact Checking negative (FC-neg)	Statement: Rome is located in France. The statement is True or False?
In-Context Interference	
Positive interference	Italy. Which country is the location of Rome?
Negative interference	France. Which country is the location of Rome?

Prompt Framing in LLMs



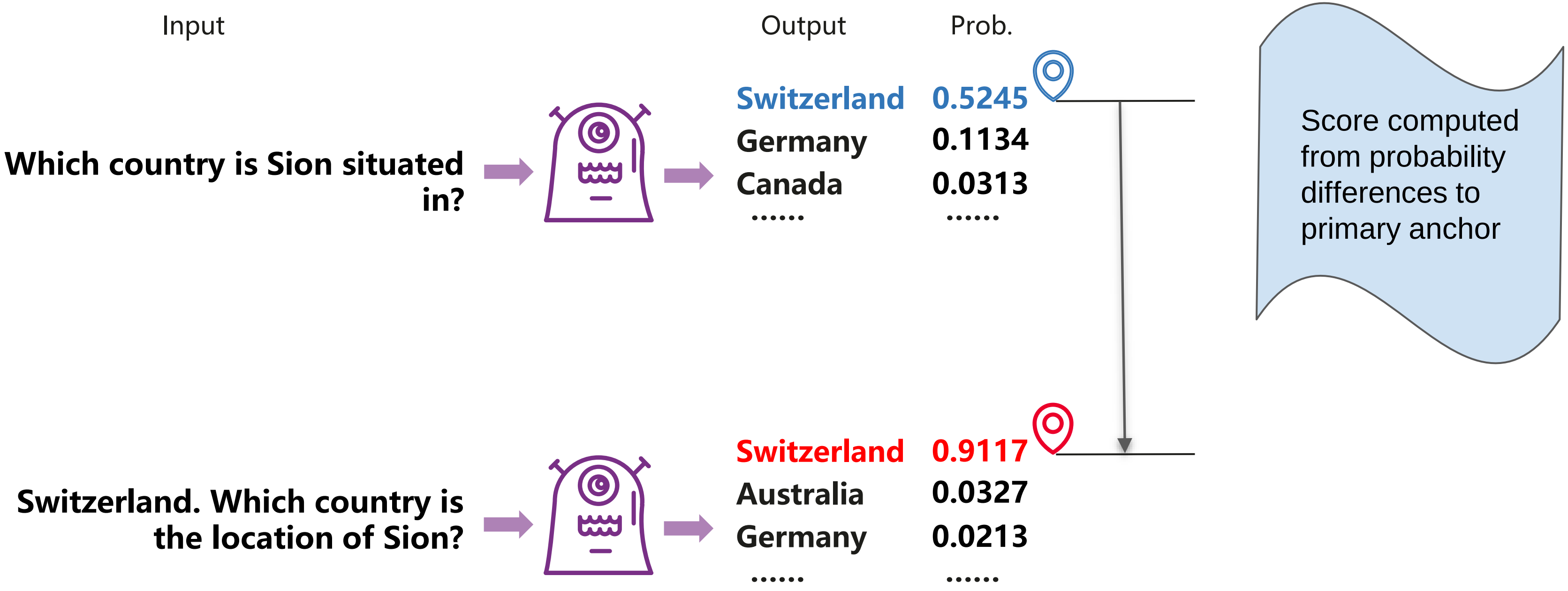
- All models show quite a variation across the prompt types.
- Larger models perhaps more stable (but hard to read)
- BLOOMZ-1b1 consistently predicts negative for fact-checking!

In-Context Interference in LLMs



- Comparison of WC with no, positive or negative interference
- Generally large differences between conditions
- Positive interference (aka giving the model the answer) can be neutral or even harmful!

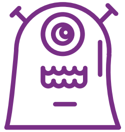
Towards a Reliability Measure ...



Assessing Reliability with MONITOR
(MOdel kNowledge reliabiliTy scORe)

Paraphrased Prompts

In which country is Sion located?
What country contains Sion?
....

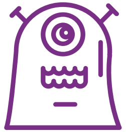


PFD (Prompt Framing Degree)

Probability Differences

In-Context Interference

Switzerland. In which country is Sion located?
France. In which country is Sion located?
....



IRD (Interference Relevance Degree)

$$M = \boldsymbol{\Sigma}(f(PFD, IRD))$$

Formulae

Prompt-framing Degree

$$PFD = \frac{1}{R}\sum_{j=1}^{R}\frac{1}{L_c}\sum_{l=1}^{L_c}|P(o_c|s_c, r, i^+)_l - P(o_c|s_c, r_j)_l|$$

Interference-relevance Degree

$$IRD = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{L_c}\sum_{l=1}^{L_c}|P(o_c|s_c, r, i^+)_l - P(o_c|s_c, r, i_m^-)_l|$$

$$MONITOR = \frac{\sum_c^S \sqrt{\alpha_1 PFD^2 + \alpha_2 IRD^2 + \alpha_3 PFD * IRD}}{\sum_c^S \frac{1}{L_c}\sum_{l=1}^{L_c} P(o_c|s_c, r, i^+)_l}$$

$P(o|s, r, i)$ is the probability of the model generating the object o with the conditions of subject s, prompt framing expression r, and the in-context information i.

i^+: positive information
i^-: negative information
R: count of prompt expressions
L_c: number of subwords in object
M: count of negative interference
S: count of subject and object

Probing MONITOR

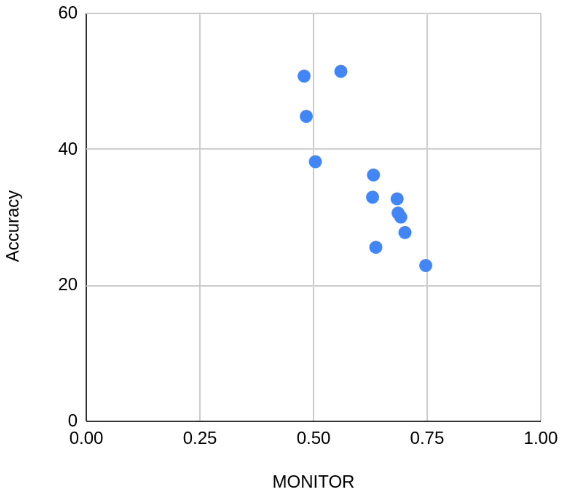
- From ~16k knowledge triples from T-REx (20 relations)
- Use GPT-4 to create 7 diverse paraphrases of the QA prompt
- Create 5 “distractors” for in-context interference
- This results in ~210k prompts
    - https://github.com/weixuan-wang123/MONITOR

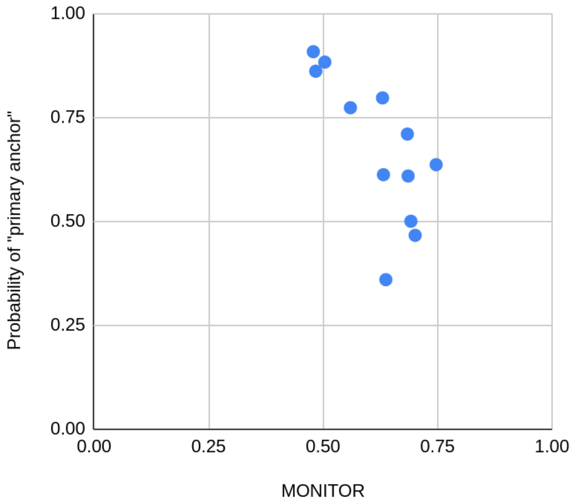
MONITOR correlates with accuracy

LLMs	MONITOR ↓	acc ↑	max ↑	min ↑
BLOOMZ-560m	0.701	27.8	40.4	15.1
BLOOMZ-1b1	0.692	30.1	43.4	16.7
Galactica-1b3	0.747	23.0	39.4	9.4
OPT-2b7	0.637	25.6	37.1	11.3
BLOOMZ-3b	0.686	30.6	44.8	16.8
Vicuna-7b	0.504	38.2	59.7	18.4
BLOOMZ-7b1	0.632	36.2	49.3	22.9
Flan-T5-XXL	0.630	33.0	48.8	19.9
Vicuna-13b	0.484	44.8	65.5	27.0
WizardLM-13b	0.560	51.5	66.0	33.0
Flan-UL2	0.684	32.7	51.4	16.3
LLaMa-30b-ins.	0.479	50.8	71.2	30.5
Correlation	Pearson			
r(MONITOR,avg acc)	**-0.846**	**-0.846**		

- We show accuracy and MONITOR averaged across 20 T-REx data sets
- MONITOR correlates inversely with accuracy

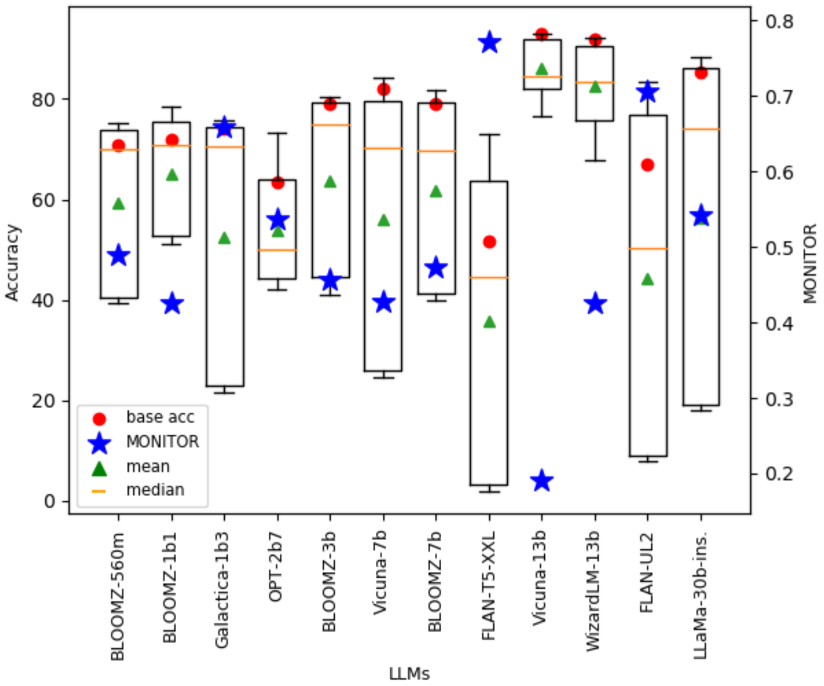
Comparing MONITOR with Accuracy/Probability



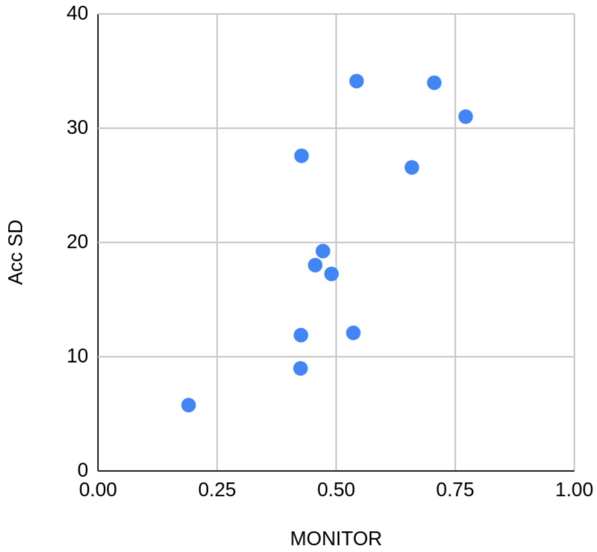


- Mean accuracy vs MONITOR; Probability of primary anchor vs MONITOR
- MONITOR shows inverse correlation with each measure

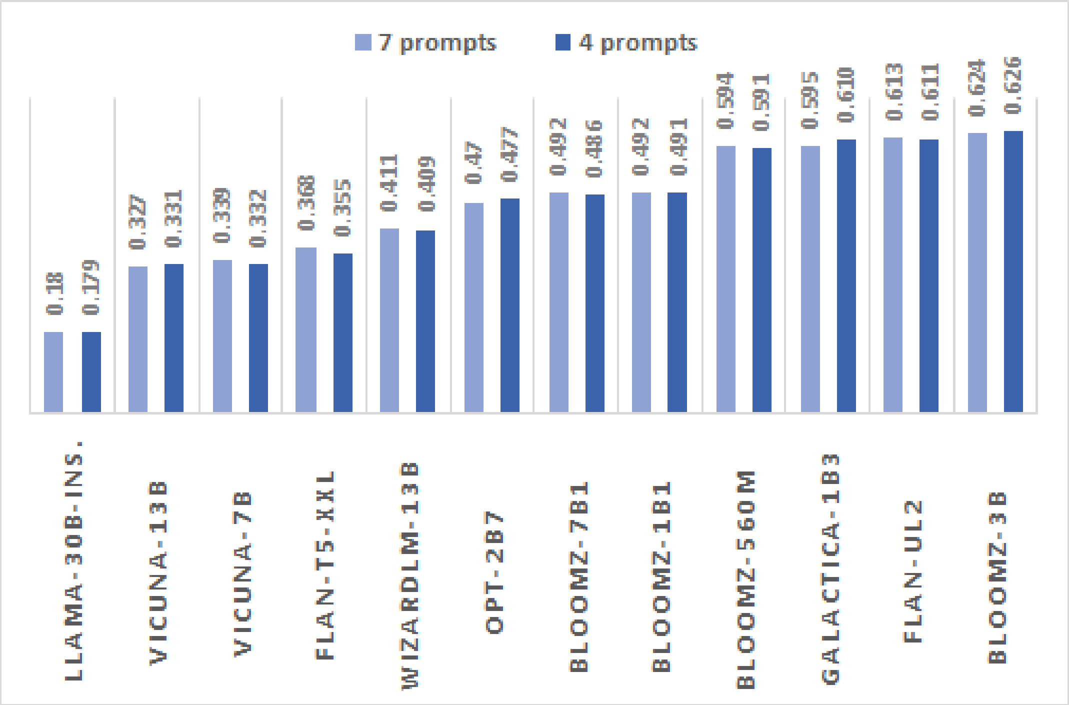
MONITOR and Accuracy Variance



- Plot shows a single relation (P1412)
- Vicuna-13b and WizardLM-13b both have similar accuracies
- Former has lower MONITOR so may be a better choice.

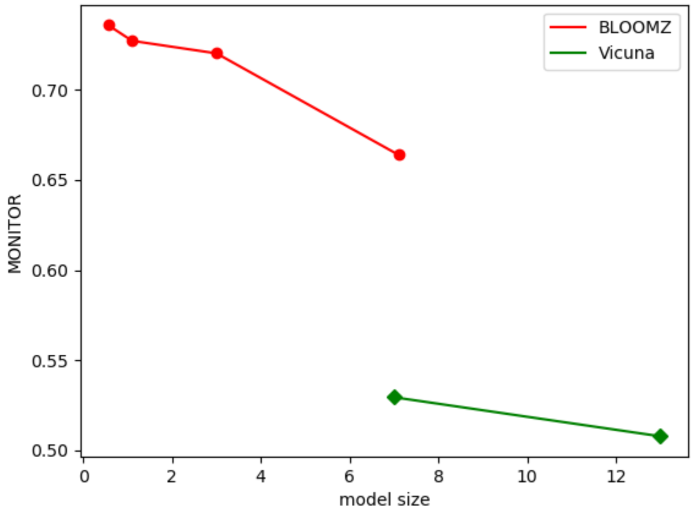


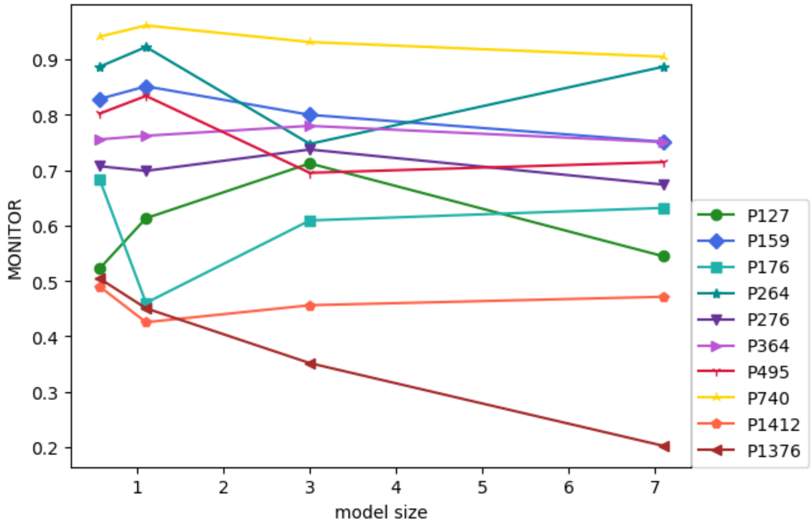
Does the Number of Paraphrased Prompts Matter?



Comparison of MONITOR for P178 - 4 prompts vs 7 prompts

MONITOR and Scale





Summary

- LLMs accuracy on factual knowledge can be affected by
  - Prompt framing
  - In-context interference
- Top-1 performance is insufficient

Accuracy is not Enough

- We propose MONITOR. A metric which takes these into account
  - Measures performance across prompts
  - Considers probability margin
- MONITOR correlates with accuracy.
  - But adds an extra dimension to the evaluation

Thank-You!

Weixuan Wang

Barry Haddow

Wei Peng

Alexandra Birch

Assessing the Reliability of Large Language Model Knowledge

Weixuan Wang[1], Barry Haddow[1], Alexandra Birch[1], Wei Peng[2]

[1] School of Informatics, University of Edinburgh

w.wang-126@sms.ed.ac.uk, bhaddow@ed.ac.uk, a.birch@ed.ac.uk

[2] Huawei Technologies Co., Ltd.

peng.wei1@huawei.com