Logical Consistency of Large Language Models in Fact-Checking

Bishwamittra Ghosh, Sarah Hasan, Naheed Anjum Arafat, Arijit Khan

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

AALBORG
UNIVERSITY

Inconsistency of Large Language Model (LLM)

LLMs are impressive at general language understanding, yet they suffer from inconsistency

Inconsistency of Large Language Model (LLM)

LLMs are impressive at general language understanding, yet they suffer from inconsistency

What is consistency?

Inconsistency of Large Language Model (LLM)

LLMs are impressive at general language understanding, yet they suffer from inconsistency

What is consistency?

Existing Works: Similar response to semantically similar prompt

Inconsistency of Large Language Model (LLM)

LLMs are impressive at general language understanding, yet they suffer from inconsistency

What is consistency?

Existing Works: Similar response to semantically similar prompt

Paraphrasing

Berlin is the capital of Germany

Germany's capital is Berlin

$$\texttt{LLM}(\text{Berlin is the capital of Germany}) = \texttt{LLM}(\text{Germany's capital is Berlin})$$

What is Logical Consistency?
Our Proposal

Response is consistent with logical changes of the prompt

- Similar response to logically equivalent prompt
- Different response to logically different prompt
- Response should adhere to formal logic

Response is consistent with logical changes of the prompt

- Similar response to logically equivalent prompt
- Different response to logically different prompt
- Response should adhere to formal logic

Negation

Berlin is the capital of Germany

Berlin is not the capital of Germany

$$\texttt{LLM}(\text{Berlin is the capital of Germany}) \neq \texttt{LLM}(\text{Berlin is not the capital of Germany})$$

Conjunction

Berlin is the capital of Germany and US embassy is in Berlin

Berlin is the capital of Germany

US embassy is in Berlin

Conjunction

Berlin is the capital of Germany and US embassy is in Berlin

Berlin is the capital of Germany

US embassy is in Berlin

$$\texttt{LLM}\left(\begin{array}{c}\text{Berlin is the capital of Germany}\\ \text{and}\\ \text{US embassy is in Berlin}\end{array}\right) = \begin{array}{c}\texttt{LLM}(\text{Berlin is the capital of Germany})\\ \text{and}\\ \texttt{LLM}(\text{US embassy is in Berlin})\end{array}$$

Our Contributions

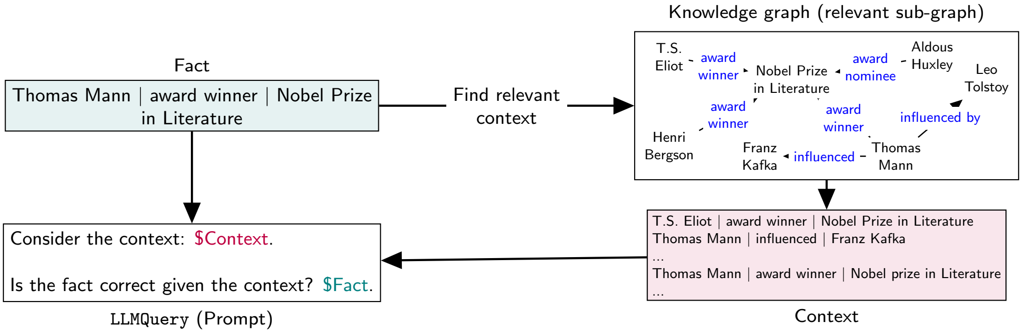
- Logical consistency on complex logical queries with negation, conjunction, and disjunction operators
- As a specific test bed, we consider the task of fact-checking in knowledge graphs (KGs) using LLMs

Our Contributions

- Logical consistency on complex logical queries with negation, conjunction, and disjunction operators
- As a specific test bed, we consider the task of fact-checking in knowledge graphs (KGs) using LLMs

Benchmark **Assessment** **Improvement**

Our Framework: LLM in fact-checking with KG



Consistency Measure

Primitive operators

$$\texttt{LLM}(\neg p) = \neg\texttt{LLM}(p)$$
$$\texttt{LLM}(p \vee q) = \texttt{LLM}(p) \vee \texttt{LLM}(q)$$
$$\texttt{LLM}(p \wedge q) = \texttt{LLM}(p) \wedge \texttt{LLM}(q)$$

Consistency Measure

Primitive operators

$$\begin{aligned}
\texttt{LLM}(\neg p) &= \neg\texttt{LLM}(p) \\
\texttt{LLM}(p \vee q) &= \texttt{LLM}(p) \vee \texttt{LLM}(q) \\
\texttt{LLM}(p \wedge q) &= \texttt{LLM}(p) \wedge \texttt{LLM}(q)
\end{aligned}$$

Disjunctive normal form (DNF): A DNF fact $q = \bigvee_{i=1}^{n} c_i$, where $c_i = \bigwedge_{j=1}^{i_m} e_{ij}$

$$\texttt{LLM}(q) = \bigvee_{i=1}^{n} \left(\bigwedge_{j=1}^{i_m} \texttt{LLM}(e_{ij}) \right)$$

Consistency Measure

Commutative law

$$\texttt{LLM}(p \lor q) = \texttt{LLM}(q \lor p)$$
$$\texttt{LLM}(p \land q) = \texttt{LLM}(q \land p)$$

Associative law

$$\texttt{LLM}((p \lor q) \lor s) = \texttt{LLM}(p \lor (q \lor s))$$
$$\texttt{LLM}((p \land q) \land s) = \texttt{LLM}((p \land (q \land s))$$

Distributive law

$$\texttt{LLM}(p \land (q \lor s)) = \texttt{LLM}((p \land q) \lor (p \lor s))$$
$$\texttt{LLM}(p \lor (q \land s)) = \texttt{LLM}((p \lor q) \land (p \lor s))$$

... De-Morgan's Laws and First-order logic.

Assessment

Model	Dataset	Fact	Accuracy		Logical Consistency	
			Before FT[1]	After FT	Before FT	After FT
Llama2-13B	FreebaseLFC	$p, \neg p$	0.90		0.81	
		$p \wedge q$	0.61		0.67	
		$p \vee q$	0.73		0.73	
	NELLLFC	$p, \neg p$	0.88		0.76	
		$p \wedge q$	0.38		0.69	
		$p \vee q$	0.73		0.73	
	WikiLFC	$p, \neg p$	0.96		0.92	

[1] FT = Fine-tuning

Assessment

Model	Dataset	Fact	Accuracy Before FT[1]	Accuracy After FT	Logical Consistency Before FT	Logical Consistency After FT
Llama2-13B	FreebaseLFC	$p, \neg p$	0.90		0.81	
		$p \wedge q$	0.61		0.67	
		$p \vee q$	0.73		0.73	
	NELLLFC	$p, \neg p$	0.88		0.76	
		$p \wedge q$	0.38		0.69	
		$p \vee q$	0.73		0.73	
	WikiLFC	$p, \neg p$	0.96		0.92	

[1]FT = Fine-tuning

Improvement: Sufficient Condition for Consistency

- An LLM is consistent on a simple atomic fact if it is accurate both on the fact and its negation
- For a complex DNF fact, the LLM is consistent if it is accurate on the DNF fact as well as on all constituent atomic facts

Model	Dataset	Fact	Accuracy Before FT	Accuracy After FT	Logical Consistency Before FT	Logical Consistency After FT
Llama2-13B	FreebaseLFC	$p, \neg p$	0.90	**0.93**	0.81	**0.86**
		$p \wedge q$	0.61	**0.93**	0.67	**0.83**
		$p \vee q$	0.73	**0.76**	0.73	**0.97**
	NELLLFC	$p, \neg p$	0.88	**0.97**	0.76	**0.93**
		$p \wedge q$	0.38	**0.89**	0.69	**0.88**
		$p \vee q$	0.73	**0.76**	0.73	**0.94**
	WikiLFC	$p, \neg p$	0.96	**0.96**	0.92	**0.93**

Takeaways from Experiments

- **Assessment.** LLMs are not always logically consistent with their generation – consistency decreases as the query complexity increases.

Takeaways from Experiments

- **Assessment.** LLMs are not always logically consistent with their generation – consistency decreases as the query complexity increases.
- **Improvement.** Instruction prompting is not sufficient to improve logical consistency in LLMs: smaller models require instruction fine-tuning, while larger models may suffice with instruction prompting.

Takeaways from Experiments

- **Assessment.** LLMs are not always logically consistent with their generation – consistency decreases as the query complexity increases.
- **Improvement.** Instruction prompting is not sufficient to improve logical consistency in LLMs: smaller models require instruction fine-tuning, while larger models may suffice with instruction prompting.
  - **Generalization.** Fine-tuning for logical consistency in one dataset can generalize to other datasets and queries with more logical operators.

Takeaways from Experiments

- **Assessment.** LLMs are not always logically consistent with their generation – consistency decreases as the query complexity increases.
- **Improvement.** Instruction prompting is not sufficient to improve logical consistency in LLMs: smaller models require instruction fine-tuning, while larger models may suffice with instruction prompting.
  - **Generalization.** Fine-tuning for logical consistency in one dataset can generalize to other datasets and queries with more logical operators.
- **Benchmark.** Fact-checking in KG provides a flexible benchmark to test LLMs on logical queries of varying complexity.

Conclusion

- Logical inconsistency is a critical issue for LLMs despite their impressive language understanding ability
- Propose a framework to assess the logical consistency of LLMs on complex fact-check queries from KGs
- Demonstrate how supervised fine-tuning can improve the logical consistency of LLMs

Paper