Consistency Analysis of ChatGPT

Introduction

- Widespread popularity of a large language model (LLM) called ChatGPT.
- ChatGPT has many features (summarization, answering questions, programming)
- However, ChatGPT can also generate incorrect information, which can be very problematic in risk-sensitive domains (law, medicine, finance)
- **Goal**: To evaluate the reliability of ChatGPT in terms of consistency.

Consistency Types

- ***Semantic Consistency*** - A model should produce similar or related outputs for inputs with similar meanings.

- ***Negation Consistency*** - A model's prediction should differ for text inputs delivering the opposite meaning.

- ***Symmetric Consistency*** - A model should be order-input invariant, meaning that its output remains the same regardless of changes to the order of the input.

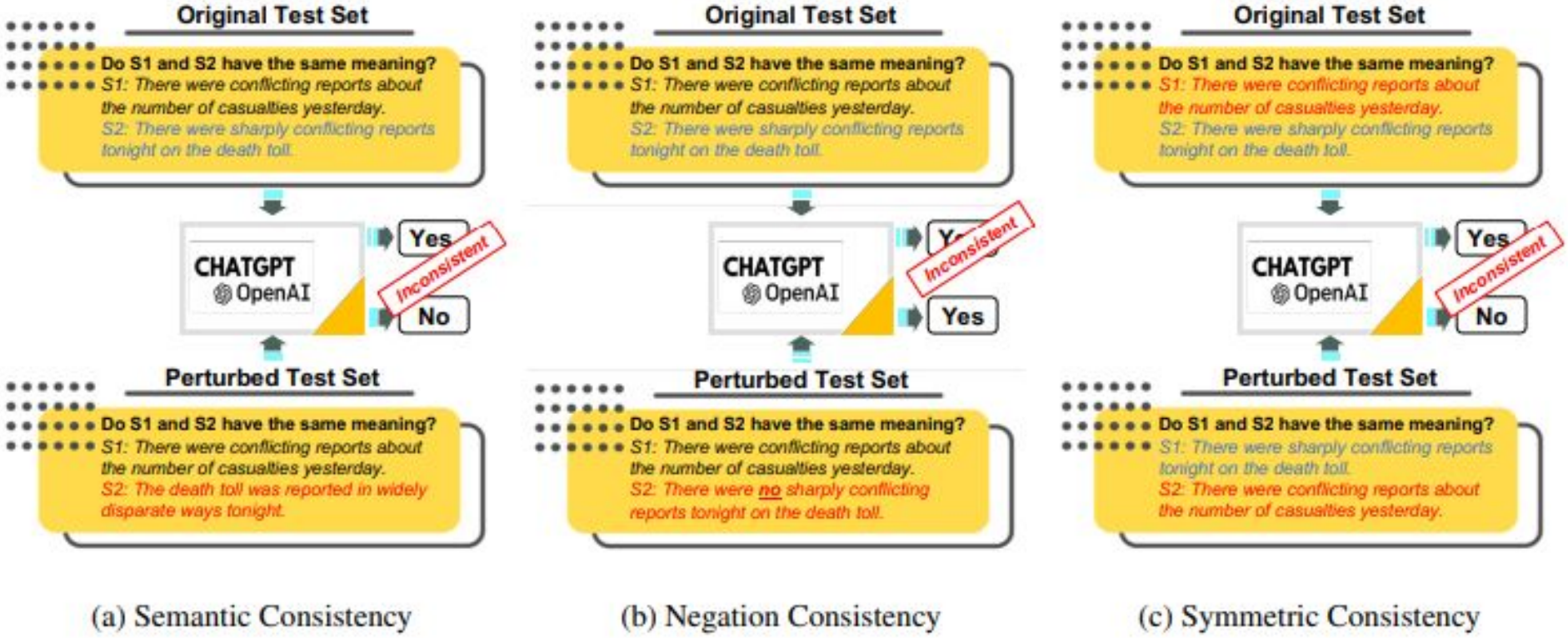
Experimental Design

- SNLI, RTE, and MRPC datasets
- Experiments are conducted on the GPT-3.5 model for ChatGPT.
- Evaluation metric measures the ratio of predictions that violate the target consistency type.

	SNLI	RTE	MRPC
semantic	4,406	248	202
negation	2,204	153	290
symmetric	3,237	1,241	3,668

Table 1: Size of the test sets of consistency evaluation data points of the SNLI, RTE, and MRPC tasks.

Experimental Design



(a) Semantic Consistency (b) Negation Consistency (c) Symmetric Consistency

Experimental Results (Semantic)

Model	MRPC		RTE		SNLI		SNLI-2C	
	τ_B	τ_S	τ_B	τ_S	τ_B	τ_S	τ_B	τ_S
BERT-large	12.5	-	12.3	-	9.9	-	-	-
RoBERTa-large	8.4	-	9.8	-	**7.9**	-	-	-
Electra-large	5.5	-	8.9	-	**7.9**	-	-	-
T5-large	**4.5**	-	**8.6**	-	9.3	-	-	-
ChatGPT	29.7	9.9	11.3	10.5	28.0	21.0	15.0	11.0

Table 3: Experimental results of semantic consistency

Experimental Results (Negation)

Model	MRPC		RTE		SNLI		SNLI-2C	
	τ	τ_C	τ	τ_C	τ	τ_C	τ	τ_C
BERT-large	90.8	-	75.8	-	11.7	-	-	-
RoBERTa-large	84.2	-	24.6	-	5.9	-	-	-
Electra-large	77.0	-	17.3	-	5.4	-	-	-
T5-large	25.2	-	15.9	-	5.8	-	-	-
ChatGPT	**21.3**	4.6	**10.5**	6.9	**5.0**	0.0	9.0	0.0

Table 4: Experimental results of the negation consistency evaluation.

Experimental Results (Symmetric)

Model	MRPC		RTE		SNLI		SNLI-2C	
	τ	τ_C	τ	τ_C	τ	τ_C	τ	τ_C
BERT-large	6.8	-	15.8	-	10.2	-	-	-
RoBERTa-large	4.3	-	11.6	-	9.7	-	-	-
Electra-large	5.3	-	**6.7**	-	**6.4**	-	-	-
T5-large	**4.2**	-	8.0	-	8.3	-	-	-
ChatGPT	12.5	-	35.5	32.6	40.5	49.23	3.0	2.52

Table 6: Experimental results of the symmetric consistency evaluation.

Discussion

- Possible solutions for reducing inconsistency**:**
  - **Prompt Design** - Has been shown to be an effective method of regulation ChatGPT's behavior.
  - **Data Augmentation** - Creating new data points based on consistency types and using them for training.

- Downsides of these solutions:
  - Maximizing generalization effect instead of complete removal
  - Unsustainability and environmental costs

Conclusion

- The goal was to examine the reliability of ChatGPT in terms of the model's consistency.
- The results showed that ChatGPT performs poorly for semantic and symmetric consistency. However, it outperforms every model when it comes to negation consistency.
- Therefore, while ChatGPT will improve with future developments, using it without human confirmation would be risky in sensitive domains.

Reference:

Consistency Analysis of ChatGPT. Jang, Myeongjun and Thomas Lukasiewicz. ArXiv. 2023.