ContraDoc: Understanding Self-Contradictions in Documents with Large Language Models

Jierui Li[1], Vipul Raheja[2], Dhruv Kumar[2]

1. The University of Texas at Austin 2. Grammarly

Introduction

Motivation: A text is considered self-contradictory when it contains multiple ideas or statements that inherently conflict. Humans struggle to identify contradictions in unfamiliar, informative texts, particularly when contradictions are widely separated in long documents, underscoring the need for automated text analysis tools.

Document Type: News Article

...So high, that it is taking five surgeons, a covey of physician assistants, nurses and anesthesiologists, and more than 40 support staff to perform surgeries on 12 people. They are extracting six kidneys from donors and implanting them into six recipients...In late March, the medical center is planning to hold a reception for all 10 patients. Here's how the super swap works, according to California Pacific Medical Center...

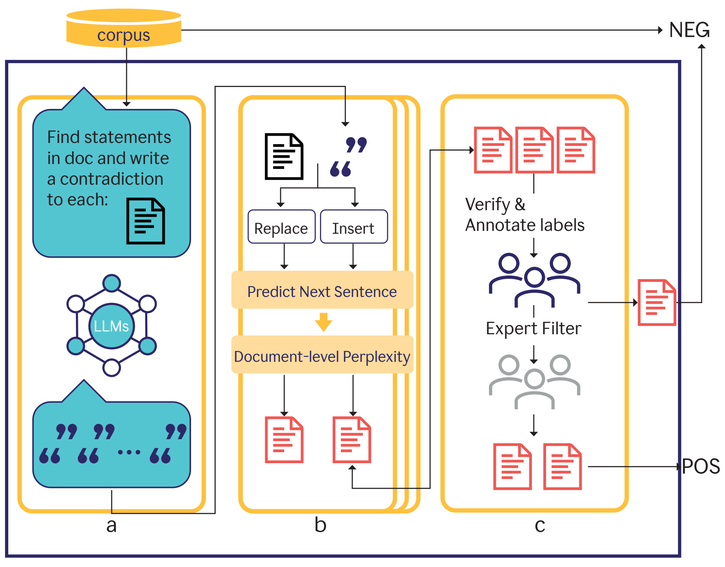
Scope of Self-Contradiction: Global
Type of Self-Contradiction: Numeric, Content

	SNLI	DocNLI	WikiContradiction	ContraDoc
Self-Contradiction	✗	✗	✓	✓
Document-Level	✗	✓	✓	✓
Contradiction Position	/	✗	✗	✓
Text Source	Flickr	Various	Wiki	News, Wiki, Story
Contradiction Type	✗	✗	✗	8 Types
Contradiction Scope	/	/	✗	✓

Dataset

Categories	Attributes	# Documents
Overall	-	449
Document Type (which domain)	News (CNN-Dailymail)	158
	Wiki (Wikitext)	150
	Story (NarrativeQA)	141
Document Types	Negation	87
	Numeric	65
	Content	288
	Perspective/View/Opinion	101
	Emotion/Mood/Feeling	86
	Relation	54
	Factual	25
	Causal	36
Self-Contra Scope (context window size of indicating self-contradiction)	Global (> 4 sentences away)	155
	Local (1 to 4 sentences away)	220
	Intra (within one sentence)	74

Dataset Curation



Machine-Human Collaboration:
a. Use LLM to find statements in the document and generate contradictory statements.
b. Inserting the contradictory statement or replacing the original statement with it based on automatic metrics.
c. Human annotators and expert filter & tag the candidate self-contradictory documents.

Evaluation Metrics

Detect → If there's self-contradiction
Detect → Point Out the self-contradiction if any

Binary Judgment: Determine one doc is self contradictory or not

- Given a document, we ask the model if the document contains a self-contradiction. The model must answer with either "Yes" or "No".
- Evaluate on 449 positive articles and 445 negative articles from similar distribution.

Top-k Contradictions: Point out evidence for self-contradiction

- Given a document, we tell the model that doc is self-contradictory and ask it to select k most probable sentences that indicate the self-contradiction. We consider it's correct if the introduced self-contradictory sentence is pinned in top k sentences(Evidence Hit).
- Evaluate on 449 positive articles

Judge then Find: First Judge, then Find (Point-Out)

- Determine one doc is self contradictory or not. If is, give 2 self-contradictory sentences or 1 (if the self-contradiction is within a sentence). It's considered correct only when the model did answer "yes" and provide correct evidence.
- Evaluate on 449 positive articles and 445 negative articles from similar distribution.

Experiments

Binary Judgment experiments show that while GPTs and PaLM2 are undersensitive (tend to predict "no"), LLaMA2 is oversensitive (tend to predict "yes") to SC.

Further experiments in self-contradiction top 5/judge-then-find show that even when LLaMAv2 is answering yes, it doesn't seem to find where the self-contradiction lies. While GPT4 performs the best among tested models, it still cannot reliably detect self-contradictions.

Model Details	
GPT3.5	GPT-3.5-turbo-0613
GPT4	GPT-4-0613
PaLM2	PaLM2(text-bison)
LLaMAv2	llama-2-chat-70B

Model	Accuracy	Precision	Recall	F1
GPT3.5	50.1%	100.0%	0.2%	0.4%
GPT4	53.8%	97.0%	8.0%	15.6%
PaLM2	52.0%	61.0%	13.4%	22.0%
LLaMAv2	50.5%	51.0%	38.3%	43.7%

Table 2: Performance of different LLMs on Binary Judgment experiment.

Model	EHR ↑	Avg. Index (1-5) ↓
GPT3.5	42.8%	1.98
GPT4	70.2%	1.79
PaLM2	48.2%	2.36
LLaMAv2	20.4%	2.28

Table 3: Performance comparison of different LLMs on Self-Contradiction in top-*k* experiment. Evidence Hit Rate (EHR) by random is 16%. Avg. Index (1-5) is the average index among the top-5 evidence texts where the self-contradiction was found.

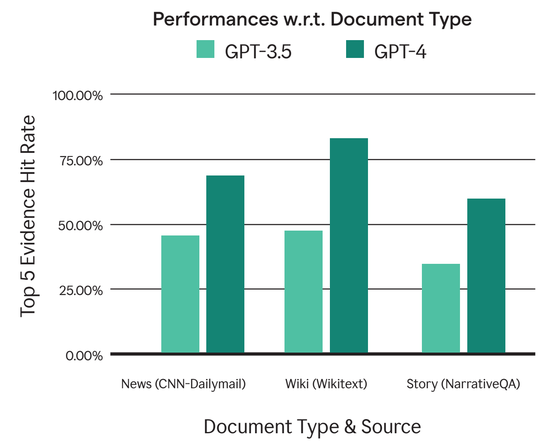
Models	Precision	Recall	F1 Score	TP Rate	FP Rate	TN Rate	FN Rate	Evidence Hit Rate	R-acc(pos)
GPT3.5	57.0%	62.0%	41.0%	20.6%	12.8%	36.9%	29.7%	41.0%	16.8%
GPT4	88.0%	39.0%	54.0%	19.6%	2.7%	46.2%	31.5%	92.7%	35.6%
PaLM2	52.0%	83.0%	64.0%	41.5%	37.6%	12.0%	9.0%	41.0%	33.7%
LLaMAv2	50.0%	95.0%	65.0%	48.0%	48.6%	1.12%	2.3%	14.5%	13.8%

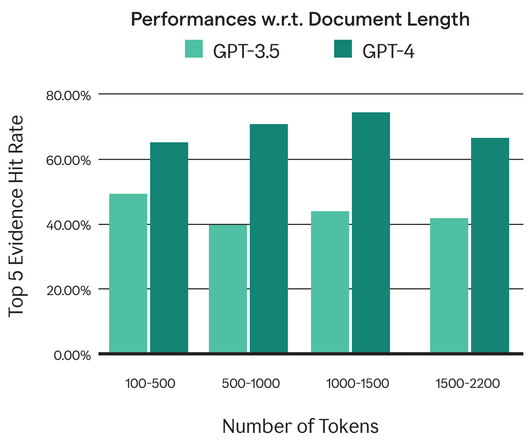
Table 4: Performance comparison of different LLMs on *Judge then Find* experimental setting. Precision, Recall, F1, and TP, FP, TN, and FN rates are calculated on the entire dataset before verification, i.e., on "Yes/No" prediction. Evidence Hit Rate is the percentage of cases where the model could find the correct evidence when it answered "Yes." R-acc(pos) denotes the fraction of positive data points confirmed by "yes" judgments and evidence hits.

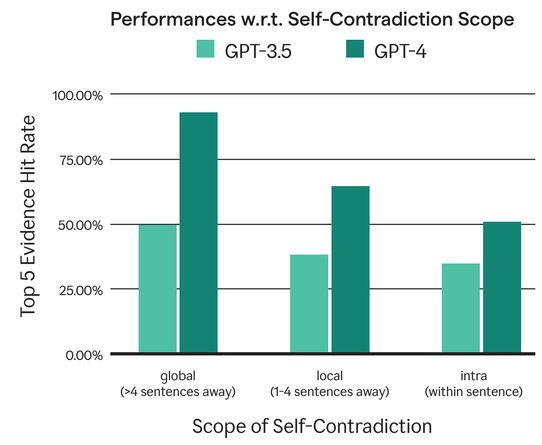
Ensembling Results

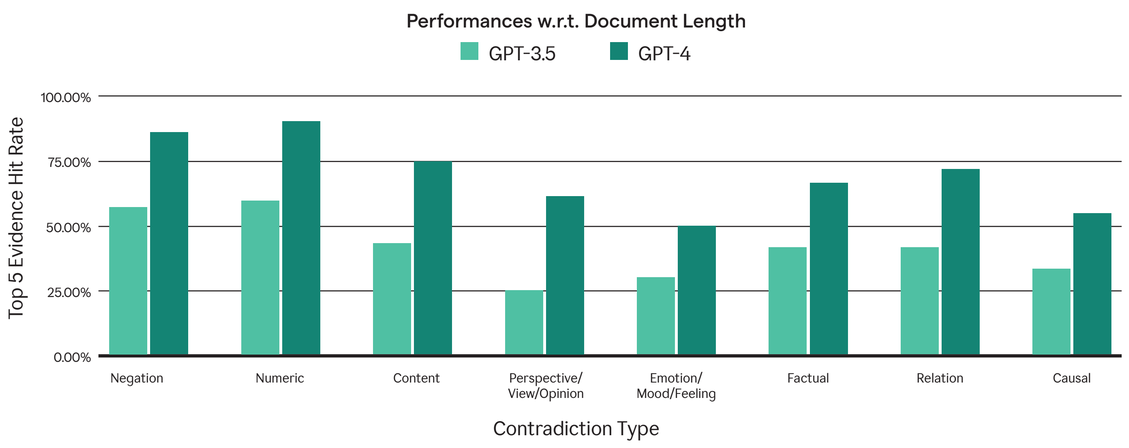
Findings from fine-grained analysis

- Self-contradictions in Wikipedia are the easiest to detect while those in stories are the hardest.
- The difficulties in detecting self-contradictions are not positively-correlated with contradiction scope or length of the document.
- Subjective contradictions are the hardest to detect while simple negation/numeric self-contradictions are the easiest to detect.









Resources

Evaluation Metrics & Dataset are available at:
https://github.com/ddhruvkr/CONTRADOC