Improving the Faithfulness of LLM-based Abstractive Summarization with Span-level Unlikelihood Training

UNIVERSITY OF CALIFORNIA SANTA CRUZ

Sicong Huang, Qianqi Yan, Shengze Wang, Ian Lane

Baskin Engineering

Motivation

Unfaithfulness in LLM-generated Summaries

- Unfaithful generation is prevalent in all LLMs
- The unfaithfulness problem can heavily affect the practicality of summarization systems

Span-level faithfulness Annotations is Scarce

- Few datasets with span-level faithfulness annotations
- Fewer studies have used span-level faithfulness annotations to improve summary faithfulness

Goal: To improve summary faithfulness with span-level unfaithfulness annotations

Pam: Hey Robert, you said you could help with Tom's birthday?
Robert: Sure, what do you need?
Pam: I have to go shopping, cook, and clean, and I figured out I don't have time to pick up the balloons.
Robert: From where?
Pam: There's this store in the city centre that sells these awesome floating balloons.
Robert: No problem, just text me the address.
Pam: Bless you!
Robert: ;)

Pam asked Robert for help with Tom's birthday celebration, as she needs to go shopping, cook, and clean, and doesn't have time to pick up floating balloons from a store in the city centre. Robert agreed to help by providing the address of the store.

Dataset Construction

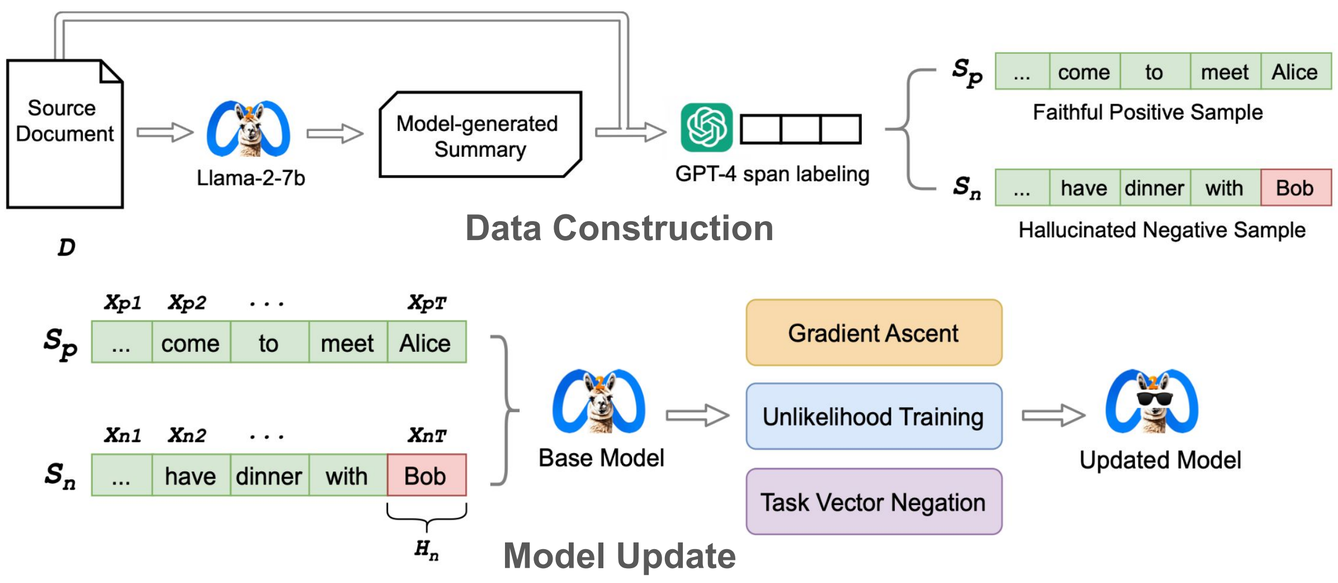
Given that LLM-generated summaries are more preferred than the reference summaries in the original datasets, we construct a dataset with "organically hallucinated" summaries

1. Generate summaries with Llama2-7b on:
   - SAMSum: short daily conversations
   - CNN: news articles
2. Annotate unfaithful spans in summaries with GPT-4

System prompt: You will be given a source text and a summary of that text … identify spans of text in the summary that is inconsistent to the source …
User prompt: <source> Pam: Hey Robert … </source> <summary> Pam asked Robert … by providing the address of the store. </summary>
GPT-4 output: <summary> Pam asked Robert … by <hallu> providing the address of the store.</hallu> </summary>

3. Filter noisy outputs and balance data between positive and negative examples

		Pos	Neg	Avg. hallu toks
Train	SAMSum	2774	2774	6.5%
	CNN	2774	2774	2.7%
Test	SAMSum	50	50	6.7%
	CNN	50	50	2.5%



Methods

Gradient Ascent

$$L_{ga} = \begin{cases} -(1-\epsilon) \sum_{x_p \in S_p} \log p_\theta(x_p|.) & \text{if } S_p \\ \epsilon \sum_{x_n \in H_n} \log p_\theta(x_n|.) & \text{if } S_n \end{cases}$$

Unlikelihood Training

$$L_{ul} = \begin{cases} -(1-\epsilon) \sum_{x_p \in S_p} \log p_\theta(x_p|.) & \text{if } S_p \\ -\epsilon \sum_{x_n \in H_n} \log(1 - p_\theta(x_n|.)) & \text{if } S_n \end{cases}$$

Task Vector Negation

$\theta_{res} = \theta_{pre} + (1 - \varepsilon)\, \tau_{pos} - \varepsilon \tau_{neg}$

θ is model params, τ is task vector

Results

	GPT4SL	G-Eval	AlignScore
SAMSum	67.0	4.631	0.696
CNN	69.0	4.897	0.800

Baseline (MLE training on positive samples) faithfulness

- Unlikelihood training can significantly improve summary faithfulness over baseline
- Task vector negation shows some improvements but gradient ascent is detrimental to performance

	ϵ	Gradient Ascent			Unlikelihood			Task Vector		
		GPT4SL	G-Eval	A-Score	GPT4SL	G-Eval	A-Score	GPT4SL	G-Eval	A-Score
SAMSum	0.1	64.0	4.63	0.6867	**80.0**	4.63	**0.7403**	65.0	4.56	0.7127
	0.3	13.0	2.96	0.6964	71.0	4.70	0.7331	68.0	4.51	0.7165
	0.5	0.0	2.26	0.4862	76.0	4.71	0.7380	69.0	4.54	0.7293
	0.7	0.0	2.49	0.4885	67.0	**4.72**	0.7394	68.0	4.60	0.7184
	0.9	0.0	1.13	0.3710	57.0	4.10	0.7355	69.0	4.53	0.7083
	1.0	-	-	-	-	-	-	72.0	4.58	0.7208
CNN	0.1	52.0	4.59	0.7662	**87.0**	4.90	0.8151	73.0	4.84	0.7909
	0.3	2.0	2.69	0.7982	83.0	**4.91**	0.8129	64.0	4.86	0.7891
	0.5	0.0	2.25	0.7114	72.0	4.84	**0.8295**	73.0	**4.91**	0.7852
	0.7	0.0	2.44	0.721	57.0	4.59	0.7942	**87.0**	4.86	0.7586
	0.9	0.0	0.69	0.1661	45.0	3.81	0.7894	84.0	4.89	0.7950
	1.0	-	-	-	-	-	-	83.0	4.87	0.7940

Tested methods faithfulness with different ε

	Baseline	Unlikelihood
Human	57	67
GPT4SL	64	75
G-Eval	77	80
AlignScore	50	67
	MCC with Human	
GPT4SL	0.669	0.572
G-Eval	0.373	0.543
AlignScore	0.043	0.088

# of faithful summaries out of 95, according to human and automatic metrics. And each metric's correlation with human judgements

- Human evaluation confirms unlikelihood training can effectively improve faithfulness
- GPT-4 span labeling correlates with human the highest, demonstrating its effectiveness at evaluating faithfulness