Analyzing LLM Behavior in Dialogue Summarization: Unveiling Circumstantial Hallucination Trends

Sanjana Ramprasad♢ **Elisa Ferracane**♣ Zachary C. Lipton♣
♢Northeastern University
♣ Abridge AI

{ramprasad.sa}@northeastern.edu

{elisa,zack}@abridge.com

Scan me!

Circumstantial Hallucinations in LLM-generated summaries

Dialogue Snippet

Greg: Hi, honey. I need to stay after hours :-(
Betsy: Again?
Greg: I''m sorry!
Betsy: What about Johnny?
Greg: Well, could you pick him up?
Betsy: What if I can't?
Greg: Betsy?
Betsy: What if I can't?
Greg: Can't you, really?
Betsy: I can't. Today I need to work long hours as well. Tuesdays are your days in the kindergarten.

Summary:

GPT-4: Greg informs Betsy he needs to stay after work, leading to a conflict as their son Johnny has to be picked up from kindergarten, which usually falls on Greg's responsibility on Tuesdays. Betsy also can't do it as she's working long hours.

Figure 1. GPT-4 infers the speakers are discussing "their son" even though that is not explicitly mentioned or discussed.

LLMs tend to generate plausible-sounding hallucinations based on *circumstantial (but not direct) evidence* in the dialogue.

New Taxonomy for hallucinations

- Previously proposed error categories do not capture LLM-specific hallucination types.
- We suggest a more refined taxonomy that integrates newly observed error types.

Linguistic Category	Summary Excerpt	Dialogue Excerpt
Circumstantial Inference	Cameron is unable to bring a video game for their daughter Peyton.	Peyton: I have been asking you to bring that video game for me Cameron: Honey, I am not having enough time to come home.
Logical Error	Jane is worried about the travel time and suggests they meet later	Steven: the road is new, we will make it Jane: I don't want to stress out, let's meet at 4:30 instead of 5, ok?
World Knowledge	#Person1# plans to vote for Joe Biden instead.	#Person1#: I will vote for Biden anyway.
Referential Error	Person1 said that Person2 could call or email them.	#Person2#: Please call me or send e-mail.
Figurative Misinterpre-tation	Alyssa likes Fergie's national anthem.	Alyssa: Have you seen Fergies national anthem? Derek: This is not normal. I saw it last week Alyssa: The best part is that she acts like she nailed it.

Prevalence of different types of hallucinations across models

- Surprisingly, LLMs do not always have a lower hallucination rate than older fine-tuned models for document summarization.

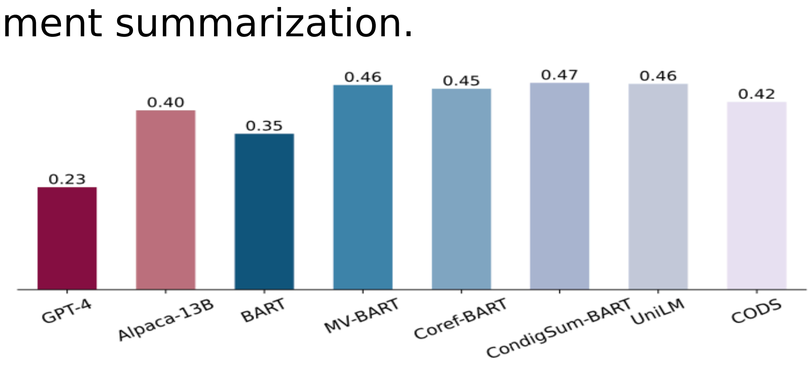

Figure 2. Each bar in this plot represents the proportion of model-generated summaries with hallucinations.

- Most hallucinations in LLMs are due to circumstantial and world knowledge errors.They exhibit fewer logical errors compared to older models.

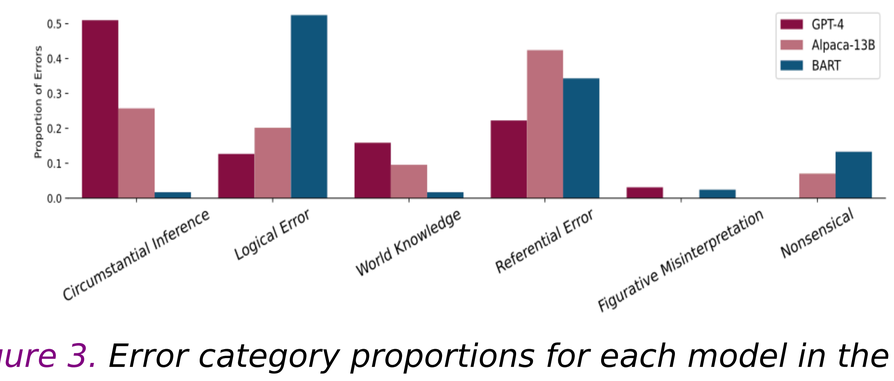

Figure 3. Error category proportions for each model in the dataset.

Performance of factuality metrics on Hallucination Types

- Prompt-based metrics outperform QA/NLI metrics.
- All metrics struggle with detecting circumstantial hallucinations (Important to evaluate performance of automatic metrics on newer models!)

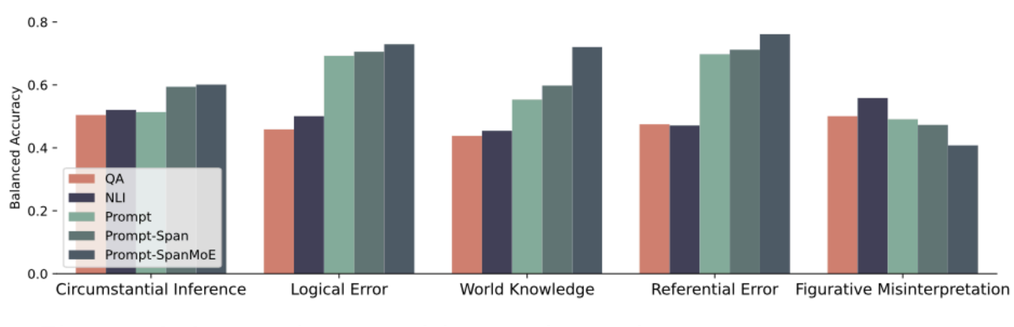

Figure 4. Inconsistency binary detection per error category.

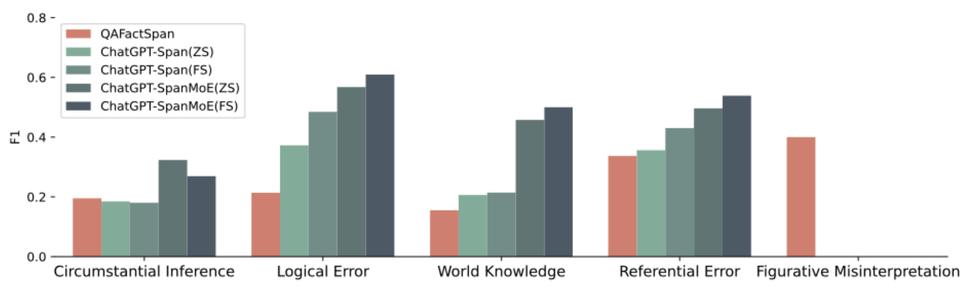

Figure 5. Inconsistent span detection (F1 scores per error category).