StorySumm: Evaluating Faithfulness in Story Summarization

Melanie Subbiah[*]
Columbia University
m.subbah@columbia.edu

Faisal Ladhak[*]
Answer.AI
fl@answer.ai

Akankshya Mishra
Columbia University
am6203@columbia.edu

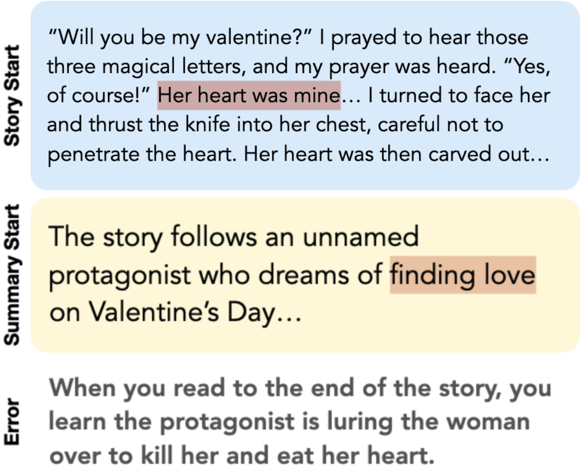
Griffin Adams
Answer.AI
ga@answer.ai

Lydia B. Chilton
Columbia University
chilton@cs.columbia.edu

Kathleen McKeown
Columbia University
kathy@cs.columbia.edu

Abstract

Human evaluation has been the gold standard for checking faithfulness in abstractive summarization. However, with a challenging source domain like narrative, multiple annotators can agree a summary is faithful, while missing details that are obvious errors only once pointed out. We therefore introduce a new dataset, StorySumm, comprising LLM summaries of short stories with localized faithfulness labels and error explanations. This benchmark is for evaluation methods, testing whether a given method can detect challenging inconsistencies. Using this dataset, we first show that any one human annotation protocol is likely to miss inconsistencies, and we advocate for pursuing a range of methods when establishing ground truth for a summarization dataset. We finally test recent automatic metrics and find that none of them achieve more than 70% balanced accuracy on this task, demonstrating that it is a challenging benchmark for future work in faithfulness evaluation.



"Will you be my valentine?" I prayed to hear those three magical letters, and my prayer was heard. "Yes, of course!" Her heart was mine… I turned to face her and thrust the knife into her chest, careful not to penetrate the heart. Her heart was then carved out…

The story follows an unnamed protagonist who dreams of finding love on Valentine's Day…

When you read to the end of the story, you learn the protagonist is luring the woman over to kill her and eat her heart.

Figure 1: A StorySumm example illustrating an incorrect interpretation of double entendre. A standard fine-grained human annotation protocol missed this inconsistency even though it is obvious once pointed out.

1 Introduction

As Large Language Models (LLMs) are able to perform more open generation tasks, challenges in evaluation have arisen (Gabriel et al., 2020). Summarization is one such task. Some aspects of summary quality like readability or coherence (Goyal et al., 2022; Chang et al., 2023) can be judged by looking at the summary alone. However, judging faithfulness (whether all details in the summary are faithful to the source) requires carefully checking a multi-sentence summary against a multi-paragraph source document (Krishna et al., 2023). Summaries that misrepresent source documents can easily spread disinformation, so it is critical we evaluate summary faithfulness, despite how labor-intensive it is.

Methods for detecting inconsistencies have generally used one of two tools: 1) trained models, or 2) human crowdworkers. Model-based approaches typically build on QA or entailment strategies. QA strategies generate questions about the summary and compare answers retrieved from the summary vs. the source document (Durmus et al., 2020; Fabbri et al., 2021b). Entailment-based approaches align facts in the summary with evidence from the source and determine for each pair if the evidence entails the fact (Utama et al., 2022; Laban et al., 2022; Maynez et al., 2020). More recent work explores prompting strategies for LLMs to identify faithfulness errors (Min et al., 2023; Kim et al., 2024a; Si et al., 2023; Luo et al., 2023; Manakul et al., 2023).

With human annotators, prior work has shown that human judgments have increased variability when evaluating long summaries (Krishna et al., 2023). Reducing the problem to evaluating individual sentences or claims helps to produce more reliable results (Krishna et al., 2023; Ye et al., 2023; Min et al., 2023). However, these works have focused on factuality in news summaries or real-

[*]These authors contributed equally to this work.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 9988–10005
November 12-16, 2024 ©2024 Association for Computational Linguistics

world articles where ground truth is based in reality and facts are stated explicitly.

As LLMs continue to grow in capabilities, there is a pressing need for evaluation of their accuracy to grow with them. We therefore produce a new benchmark, StorySumm, which can be used to improve evaluation methods for faithfulness. **StorySumm consists of 96 short stories and LLM-generated summaries with over 500 sentence-level faithfulness labels and explanations**. Each unfaithful summary is labeled as *easy* or *hard* to detect.

LLM summaries often contain subtle errors, particularly for narrative text which requires nuanced interpretation. This benchmark therefore introduces new challenges when compared to fact-checking or summarization datasets in the news domain. The example in Figure 1 demonstrates that assessing the summary requires correct interpretation of sentences like, *Her heart was mine.*, which have multiple meanings and are misleading without carefully reading the entire story. By focusing on faithfulness in narrative summarization and using real-world data from LLMs and Reddit, StorySumm poses a realistic but hard benchmark to push our methods forward.

We first explore how to establish ground-truth on this dataset by comparing different human annotation protocols and manually inspecting the results. We try different protocols and pools of annotators to see if there is an approach that helps average annotators pay attention and understand this challenging task more consistently. We find that different protocols catch unique but legitimate inconsistencies and have only fair agreement with each other. We therefore manually review and merge label sets across three annotation protocols.

We analyze the errors found by each protocol, and formulate a set of **recommendations for human evaluation** of faithfulness in narrative summarization. Most importantly, we show that **it is important to use a variety of annotators and protocols when establishing ground truth for faithfulness**. We then explore how well recent automatic metrics perform on this dataset. We find that **no metric achieves more than 70% balanced accuracy** on this task and even the best metric misses almost 50% of the hard inconsistencies.[1]

Split	#	Sto. wc	Sum. wc
Val.	33	610	120
Test	63	849	149
All	96	767	139

Table 1: Summary statistics for StorySumm showing the number of story-summary pairs and the average word count of stories and summaries.

2 StorySumm Dataset

We design our benchmark with a focus on three principles which distinguish it from existing datasets. First, the stories need to be short enough that humans can easily read them, so that we can affordably test human protocols. Second, the stories should not be so famous that LLMs have likely trained on summaries of them, potentially biasing LLM summary or evaluation quality. Third, the summaries should be representative of powerful LLMs so that we can assess how difficult it is to find errors in fluent and convincing summaries.

Motivated by these principles, we opt for short narratives from Reddit and use GPT-series and Claude-series models to generate summaries. We do not include any human-written summaries as the purpose of this dataset is to improve detection of errors in LLM-generated summaries. We show summary statistics for the dataset in Table 1 and full examples of stories/summaries in Appendix A.

2.1 Stories

We collect a dataset of 32 short stories from two popular subreddits where users can submit their original short stories for others to enjoy and comment on.[2] We filter out posts that are marked NSFW (Not Safe For Work, meaning inappropriate content) and also posts that have fewer than three up-votes. The stories are typically less than one page long. We note that users do not write summaries for their stories, and since these stories are not popular, they're unlikely to be summarized elsewhere; therefore, there is little concern about data contamination for LLMs. Additionally, as LLMs are now being used to summarize lots of different data online, it is important to evaluate them on more colloquial narrative like this rather than just benchmarks of published/popular stories.

[1] All code and data will be released at https://github.com/melaniesubbiah/storysumm.

[2] The two subreddits we used were r/shortstories and r/shortscarystories.

2.2 Summaries

For each story, we generate 3 different summaries using 3 different models, resulting in 96 story-summary pairs (see Appendix B for prompting details)[3]. Each summary is about a paragraph long. To simulate real evaluation conditions, we split the dataset into a validation split of 33 summaries which are generated by an older set of models (Davinci-3, ChatGPT, and Claude-2) and a test set of 66 summaries from newer models (GPT-3.5, GPT-4, Claude-3). This allows us to assess whether automatic metrics that require threshold tuning for classification can be tuned on a validation set of labeled summaries from older models and still work well as newer models are coming out. We use disjoint sets of 11 and 21 stories to generate the summaries for the validation and test sets respectively.

2.3 Annotator Labels

The question we ask annotators is: *Is the information in the summary consistent with the story?* We define a *consistent* summary as: *The events and details described in the summary should not misrepresent details from the story or include details that are unsupported by the story. We ask you to ignore commentary in evaluating consistency.* Commentary means sentences like, *The story reflects the enduring bonds of friendship and the role of companionship during times of hardship.*, which interpret the story to find themes rather than just detail the plot.

For annotator recruitment[4], we first compare Amazon Mechanical Turk and Upwork[5], asking four annotators from each platform to assign a binary faithful/unfaithful label to each summary. We mark a summary as faithful if three or more annotators in a group label it as such. We find that MTurk workers label 97% of summaries faithful whereas Upwork workers label 64% as faithful. When the authors perform the same task, we find 45% of summaries faithful, so we conclude that Upwork workers are more astute at catching errors and we use them for the remainder of our experiments. We caution future work to avoid using MTurk for faithfulness evaluation as it will dramatically inflate performance. Marshall et al. (2023) also showed Mturk response quality has dramatically declined in the last decade and is now mostly unusable.

Generator	#	% Faith.	# Easy	# Hard
Davinci-3	11	72.7%	1	2
GPT-3.5	21	57.1%	3	6
ChatGPT	11	54.5%	4	1
GPT-4	21	57.1%	2	7
Claude-2	11	36.4%	4	3
Claude-3	21	90.5%	0	2
Overall	96	63.5%	14	21

Table 2: Summary statistics using the annotator labels for each summary generation model, showing the number of summaries, the percent of summaries labeled faithful, and the number of unfaithful summaries labeled *easy*/*hard* to detect.

We build on Krishna et al. (2023), which shows that fine-grained evaluation encourages inter-annotator agreement. We recruit three annotators from Upwork who are fluent in English and successfully complete a pilot exercise shown in Appendix C. We then ask them to assign a binary faithfulness label to each sentence in a summary. When they mark a sentence as unfaithful, they also provide a brief written justification. Prior work has shown limited benefit to using atomic claims (Tang et al., 2024a), so we do not take the additional step of generating them as proposed by Min et al. (2023). The full interface for experiments is shown in Appendix D. We pay each annotator $100 for annotation of all 96 summaries.

If two or more annotators mark the same sentence as unfaithful, we mark the whole summary as unfaithful. If all three annotators mark the same sentence as unfaithful, we label that unfaithful summary as *easy* to detect, whereas it is *hard* to detect if one annotator labels the sentence as faithful. These difficulty distinctions allow for more meaningful error analysis of different evaluation methods, and only apply to unfaithful summaries. We see almost perfect inter-annotator agreement with a Fleiss-kappa score of 0.85 for the sentence-level annotations.

Finally, by the breakdown of faithfulness labels by model shown in Table 2, we can see that faithfulness is still a significant problem for LLMs in narrative summarization with close to 40% of summaries containing errors. For example, Table 3 shows a case where all three models misinterpret the protagonist as having a positive reaction to speaking to their dead mother when in reality the mother was very cruel in life. Kim et al. (2024b) and Subbiah et al. (2024) also find many errors in interpretation

[3] Accessed through the OpenAI API and Anthropic API.

[4] Our work with annotators is approved by Columbia University IRB protocol AAAS4051.

[5] MTurk and Upwork

Story Evidence: ...I remember Mom, a sweet woman. Her smile, her strong hands, and witty humor . . . After several minutes, I gather the courage to speak to her. “How’s it going, mom?” I feel strong emotions coming. Mom’s voice answers me in return: You brat! I took care of you and your brother! I gave you your own closet. But you scratched it like a cat! You! You didn’t know how to wash in the tub, so I showed you how! I laugh at the whisper. “You’re telling a different story, mom.” I smile. “At least you won’t hurt me or Hector anymore.”

Davinci-3: ...The narrator talks to their mother’s grave and reminisces, and finds that even in death, their mother still has plenty of wit and humor.

ChatGPT: ...They eventually make it to their mother’s grave and have a conversation with her, reminiscing about their past. The story ends with the protagonist feeling comforted by their mother’s voice.

Claude-2: ...The narrator speaks to the mother’s spirit and reminisces about memories of her. Despite the strange ability, the narrator finds comfort communicating with the mother’s spirit...

Table 3: An example where all three LLMs make the same error (shown in red) in understanding the narrative. The protagonist had a cruel relationship with their mother, but the summaries all suggest the protagonist feels positively about speaking with her ghost.

Story Evidence: She cursed my father, with me as the vessel... A person that I always found pleasant cursed me to this life... I have a list on me of all female infants and young girls that have passed away... so that I can steal their identities... Next on my list is a little baby girl, who died 16 years ago. Hope Elizabeth Scott... This time it feels different. Is it a coincidence? Or maybe this is my way out?

Annotators:	**GPT-4 Inconsistencies:**
INCONSISTENT ✗	**1.)** The summary says that Hope is a victim of Margaret Scott’s curse.
CONSISTENT ✓	**2.)** The summary mentions that Hope is planning to assume the identity of Hope Elizabeth Scott, a deceased baby girl.
CONSISTENT ✓	**3.)** The summary states that Hope finds a "glimmer of hope" in possibly assuming the identity of Hope Elizabeth Scott.

Figure 2: An example of the hybrid method generated inconsistencies, which are all incorrect in this case. #2 and #3 are details that are consistent between the summary and story. #1 convinces annotators, but is actually consistent with the story.

of characters’ internal states in narrative summaries like this example.

Table 2 also shows a general trend of newer models having more *hard* unfaithful summaries and older models having more *easy* ones. This pattern suggests that model errors will continue to become harder to detect in subsequent generations. We show more StorySumm examples in Table 4.

3 All That Glitters is Not Gold...

Typically, annotator labels with almost perfect inter-annotator agreements like ours are just assumed to be ground truth. However, we hypothesize that errors in narrative summaries may be difficult to catch and the annotators likely missed some. Therefore, we compare our Upwork annotator labels against labels from other human evaluation protocols to gain a better sense of their quality. In addition to our annotator labels, we compare the following two methods[6]:

Expert. Three of the authors of this paper review each summary and label it as faithful or unfaithful. We consider ourselves “expert" annotators as we have experience in faithfulness research and are motivated to produce thoughtful labels (modeled after Kryściński et al. (2019b) who also use themselves as expert annotators for factual consistency). The three experts adjudicate their labels by discussing any disagreements until all three agree on the label. This process is completed before the experts view any other labels for the dataset so they can remain unbiased. All three experts initially agreed on only 46% of the labels before adjudication, demonstrating that even experts struggle to catch every error. In total, this process took about ten hours.

Hybrid. We have GPT-4 generate multiple possible inconsistencies between the summary and story (see example inconsistencies in Figure 2 and prompt in Appendix B). These inconsistencies are explanations of why details in the summary may be inconsistent with the story. Three new workers from Upwork read these inconsistencies before labeling the summary overall, and write a short response justifying why they agree or disagree with each inconsistency. We hypothesize that identifying specific inconsistencies workers miss is useful support an LLM can provide. Presenting multiple possible options from the LLM raises the chances of one of them being accurate.

3.1 Label Comparison

In Table 5, we show the agreement and accuracy of the expert and hybrid protocols relative to the annotator labels. We can see that both have lower inter-annotator agreement (Fleiss-kappa 0.2-0.4), likely because annotations are done at the summary

[6] See Appendix E for additional methods we experimented with but rejected.

	Story Evidence	Summary Claim	Reason for Error
Easy	He woke up staring at a bright florescent light. He could hear his father talking to the doctors and police. Daniel thought it was best to stay quiet.	When he wakes up he is in a hospital and his parents are discussing sending him to rehab. Daniel agrees, and then falls back asleep.	*Daniel does not agree to go to rehab the first time he wakes up in the hospital.*
	I could still taste the gas station coke I had slurped up before the light pulled me into the night sky. In what felt like seconds, I was swallowed up in a beam of light.	A man is abducted from his car while drinking a soda by a beam of light.	*There is no evidence the man was in his car.*
	Kristen's Dad and her little brother Christian sat quietly... "Dad what is going to happen to Kristen?" Christian asked. Her Dad did not respond, and continued to slowly eat.	Her father and younger brother know what's happening, but they are unable to stop it...	*Her younger brother does not know what is happening and is asking their dad.*
	Aiming under my own chin, I pulled the trigger. I didn't hear the blessed scream of the barrel.	They contemplate ending their life, but instead their memory is wiped...	*The narrator takes action to end their life.*
	But she, along with her strange tubes and tanks and half-smiles was gone. The last thing he remembered seeing yesterday, while he was halfway across the street, was a blaring alarm and a screeching van, (red? white?).	But one day, the girl doesn't show up and he learns that she has been taken to the hospital.	*The "he" is a dog and so doesn't know the girl is at the hospital but the reader can infer it.*
	Eventually on the road I met a couple travelers who were all too happy to trade me 3 silver coins for my gold coin.	Eventually, after trading one of his gold coins for 15 silver ones, he wakes to find his previously tiny dragon grown bigger...	*The narrator makes a trade for only 3 silver coins.*
Hard Annotator	Margot starts gathering the plastic white discs. One by one, I frantically pitch the AirTags out the open window into the speeding gravel, each shattering on impact.	Jane forces Margot to throw the AirTags out the window, concerned for her safety.	*Jane (the narrator) throws the AirTags out the window.*
	It is said that men have trouble listening to women. I had no trouble listening to my mom.	The author... recalled her mother once speaking about how she used to love eating honeycombs... When he hints at the gift...	*The narrator is a man, but the summary uses a mix of pronouns for the narrator.*
Expert	"There's only one plate," she said, puzzled... I turned to face her and thrust the knife into her chest, careful not to penetrate her heart... Her heart was carefully set on the plate...	He prepares a meal for her, but when she arrives, he stabs her and carves out her heart to eat.	*The meal is only for himself to eat her heart.*
	"I'll see you tomorrow." The way she said tomorrow-... Naturally, he assumed she would say nothing else but his name with such emotion. The small terrier knew, like the sky is blue, that his name was Tomorrow.	In this poignant story, a small terrier named Tomorrow has been living on the streets for as long as he can remember.	*The reader infers that the dog is not actually named Tomorrow.*
Hybrid	The guy looked taken aback, "Ma'am, I have a husband who I am completely devoted to!..." The guy's husband looked back at her, "Get lost, Karen..."	She then descends to the mortal realm to test her power, but is turned down by a gay man who calls her a Karen.	*The man's husband calls her Karen.*
	He said he never got the opportunity to use it, and apologized again. The rescue team looked at each other, just as the radio flared to life...	When the team was about to leave, the radio came to life again and the same voice asked when they were coming to get him.	*There is no evidence the team was about to leave.*

Table 4: Examples from StorySumm. The easy examples are detected by all three annotators in the annotator labels and all three human annotation methods (annotator, expert, and hybrid). The hard examples are detected only by the method listed on the left. We present evidence from the story, the erroneous summary claim, and the error reason.

Method	Flei.-k	Coh.-k	% Easy	% Hard	BAcc.
Expert	0.27	0.36	92.86	52.38	68.71
Hybrid	0.23	0.20	92.86	76.19	61.92

Table 5: Expert and hybrid label summary statistics. We show the Fleiss-kappa inter-annotator agreement, the Cohen's kappa with the annotator labels, the percents of the *easy* and *hard* unfaithful summaries the method detects, and the balanced accuracy against the annotator labels.

rather than sentence level (Krishna et al., 2023). Inter-annotator agreement is computed between the humans in a protocol, whereas Cohen's kappa is computed between protocols. Both expert and hybrid protocols detect 93% of the easy inconsistent summaries but a lower percentage of the hard summaries (52% for the experts and 76% for the hybrid method). The experts have higher balanced accuracy despite detecting a lower percentage of the hard inconsistencies because the hybrid method detects many inconsistencies and is less precise. Balanced accuracy is a measure of accuracy for binary classification that accounts for class imbalance. It is the average of recall for the two classes.

Both methods have only fair agreement with the annotator labels (Cohen's kappa 0.2-0.4). We show the breakdown of label overlap in Figure 3. The counts where the annotator labels say *faithful* and an alternate method says *unfaithful* suggest that the annotator labels miss real inconsistencies (19 new unfaithful summaries detected by the experts

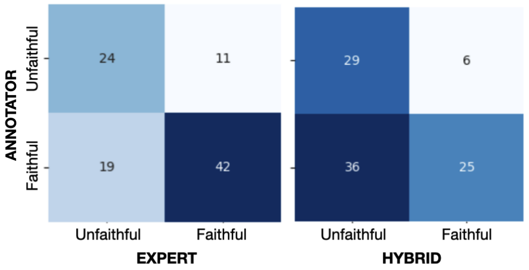


Figure 3: Confusion matrices of the expert and hybrid labels against the annotator labels.

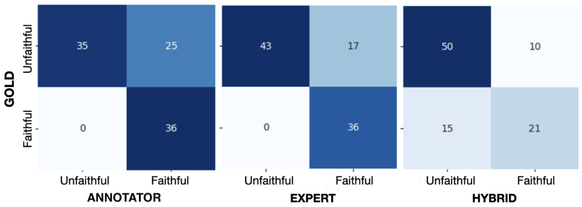


Figure 4: Confusion matrices of label overlap between the three human annotation methods and the expanded gold set of labels.

and 36 by the hybrid method). Since the expert labels are adjudicated between the three authors, we are sure the 19 expert inconsistencies that the annotator labels miss are correct. We can also see that even the experts miss inconsistencies as they miss 11 that are detected by the annotator labels. Before accepting the hybrid inconsistencies, we need to check their quality in the next section since the *hybrid* method is a novel annotation protocol.

3.2 Expanded Gold Labels

Since each human annotation method clearly detects different inconsistencies, we want to merge their labels to get better coverage of the errors. For the annotator and expert labels, we take the union of their detected errors since these are established and trusted protocols. Therefore, a summary is labeled unfaithful if either the annotator or expert labels find it to be unfaithful. We manually merge their written error explanations.

For the hybrid labels, we manually review and filter out illegitimate errors. For example, Table 2 shows a case that annotators incorrectly label as an error. GPT-4's generated inconsistency #1 convinces annotators that Hope is not "a victim of" the curse because technically Hope's father is the target of the curse ("She cursed my father"). However, Hope also suffers under the curse and says, "[Margaret] cursed me to this life", so she is also a victim of the curse and this is not a real inconsistency.

Labels	% Faith.	# Easy	# Hard	Unique Errors: Annot.	Expert	Hybrid
Annot.	63.5	14	21			
Gold	37.5	20	40	2	4	6

Table 6: A comparison of the summary statistics between the annotator and expanded gold labels.

In this process, we create a new set of labels for the dataset that are an amalgamation of the annotator, expert, and hybrid labels from the three human annotation methods, and we also provide a written description of the inconsistencies detected in each summary. These labels become the expanded gold set and the *easy*/*hard* breakdown for these labels is based on whether all three methods detect an unfaithful summary.

In Figure 4, we show the confusion matrices of each method with the gold set of labels, demonstrating that each additional human annotation protocol adds new inconsistencies. Table 6 shows that 2 unfaithful summaries are detected only by the annotator labels, 4 only by the expert labels, and 6 only by the hybrid method. Table 4 shows examples for these cases and we see that these are real errors even though they are easy to miss. For example, one of the errors detected only by the hybrid method is that the summary says the protagonist is rejected and insulted by the same man, but in the story one man rejects her and his husband insults her. The *easy* and *hard* examples shown exhibit a general pattern that *easy* errors tend to be about core story events, whereas *hard* errors are often about smaller details or subtle twists of meaning that are easy to mentally skip over (e.g., an incorrect pronoun).

3.3 Recommendations

Through this error analysis, we form several recommendations for faithfulness human evaluation:

1.) **Use multiple protocols and sets of annotators for good coverage of errors; otherwise performance is most likely inflated.** Using just the fine-grained annotation protocol with Upwork workers, we find only 2/3 of the errors in the expanded set. Protocols that localize and explain errors make it easier to check and merge error sets.

2.) **The quality of the annotator pool affects how many errors are found.** In our case, MTurk workers find almost no errors, Upwork workers find more, and experts (who also have to discuss the labels with each other) find the most.

3.) **When precision matters, use a fine-grained annotation approach (by sentence or claim).** Krishna et al. (2023) originally recommended this approach and our work supports it. We see almost perfect inter-annotator agreement for the line-by-line approach and the errors detected are legitimate.

4.) **When coverage matters, include a high-coverage protocol such as our hybrid method.** The hybrid method finds the most errors, but some of these are not real errors as annotators are highly influenced by the model suggestions. Using a high-coverage method requires an additional filtering step for legitimate errors but finds errors not found by other protocols.

Prior work (Falke et al., 2019; Gillick and Liu, 2010) has also advocated for expert involvement by showing typical annotation settings do not match expert labels. Our work additionally shows that expert labels may be missing inconsistencies as well and uses expert review to merge annotation sets. We recommend future work use the expanded gold labels, but include analysis of the annotator labels as well to study how using labels from standard annotation protocols affects calibration of metrics, which we show in the next section.

4 Benchmarking Automatic Metrics

Having established a source of ground truth, we benchmark recent automatic methods against our labels. We try the following metrics:

Binary. We prompt GPT-4 (Achiam et al., 2023), Claude-3[7], and Mixtral-8x7B[8] to assign a binary faithfulness label to each summary using the same definition of faithfulness as used for the human annotators.

CoT. We prompt GPT-4, Claude-3, and Mixtral-8x7B to assign a binary faithfulness label to each summary, but to first provide some reasoning in a chain-of-thought style (Wei et al., 2022; Kojima et al., 2023). Models are prompted to: *Consider whether there are any details in the summary that are inconsistent with the story and provide a couple sentences of reasoning for why the summary is or is not consistent with the story.*

FABLES. We use the approach from FABLES (Kim et al., 2024b) of asking ChatGPT to convert each summary to a list of claims and then asking GPT-4 to assign a binary faithfulness label to each claim. We then label the summary as faithful if all the claims are faithful.

MiniCheck. We use the approach from MiniCheck (Tang et al., 2024a) of using a Flan-T5-Large model (Chung et al., 2024) finetuned on their synthetically generated dataset to check summary claims against passages from the story.

UniEval. We use the approach from UniEval (Zhong et al., 2022) which uses multi-task learning across a unified framework of tasks to develop evaluation models. We use their *Consistency* variant.

AlignScore. We use the approach from AlignScore (Zha et al., 2023) which uses multi-task training across a unified framework of tasks to determine if one piece of text is consistent with another.

4.1 Results

We first show the results of the different methods against the Upwork **annotator labels** in Table 7 to see how automatic metrics seem to perform when evaluated with the standard fine-grained annotation approach. For UniEval and AlignScore, we tune their classification thresholds on the validation set and then use this threshold for the test set. For the remaining methods, we show results on the full dataset[9] We see that the purely prompting-based LLM approaches predict most of the summaries as faithful and therefore have relatively low balanced accuracy scores. MiniCheck detects many of the hard errors as it predicts only 18% of the summaries are faithful.

On this set of labels, the best automatic method overall is the FABLES approach with GPT-4 as a base, which achieves 67% balanced accuracy and is the most precise when it predicts a summary is faithful. On this incomplete set of labels, FABLES appears to detect more of the hard errors relative to other human methods whereas the humans detect more of the easy errors. Both human approaches detect 93% of the easy errors, suggesting that these errors are generally obvious to humans regardless of protocol but not necessarily to models (FABLES finds 72% of *easy* errors). Interestingly, the expert human balanced accuracy is only 2% higher than for FABLES. This is important to note as without the expanded set of labels, someone might conclude that FABLES is performing as well as expert

[7] Claude-3 blogpost

[8] We access Mixtral through the HuggingFace API.

[9] Results on the validation and test splits separately are shown in Appendix E.

Split	Method	Coh.-k	% Faith.	Prec.	Rec.	% Easy	% Hard	BAcc.
Val./Test	UniEval	0.38/-0.09	61/59	0.70/0.65	0.78/0.56	66.7/20.0	50.0/40.0	68.9/45.4
	AlignScore	0.28/-0.00	42/70	0.71/0.68	0.56/0.70	77.8/40.0	66.7/26.7	64.4/49.9
Full	Binary (Claude-3)	0.17	95	0.67	**1.00**	21.4	9.5	57.1
	Binary (GPT-4)	0.27	71	0.72	0.80	64.3	33.3	63.0
	Binary (Mixtral)	0.09	96	0.65	0.98	7.1	9.5	53.5
	CoT (Claude-3)	0.23	90	0.69	0.97	21.4	23.8	59.8
	CoT (GPT-4)	0.15	94	0.67	0.98	21.4	9.5	56.3
	CoT (Mixtral)	0.05	97	0.65	0.98	7.1	4.8	52.0
	FABLES (GPT-4)	**0.32**	53	**0.78**	0.66	71.4	66.7	**67.1**
	MiniCheck (Flan-T5)	-0.06	18	0.53	0.15	**85.7**	**71.4**	45.9
	Expert (Human)	0.36	55	0.79	0.69	92.9	52.4	68.7
	Hybrid (Human)	0.20	32	0.81	0.41	92.9	76.2	61.9

Table 7: Model scores against the Upwork **annotator** labels. We report the Cohen's kappa score between the predicted labels and the annotator labels, the % of summaries labeled faithful, precision and recall for detecting faithful summaries, the % of *easy*/*hard* unfaithful summaries detected, and the balanced accuracy.

Split	Method	Coh.-k	% Faith.	Prec.	Rec.	% Easy	% Hard	BAcc.
Val./Test	UniEval	0.34/0.09	33/24	0.45/**0.53**	0.62/0.29	90.0/**80.0**	66.7/**80.0**	69.2/54.3
	AlignScore	0.21/0.09	42/70	0.36/0.48	0.62/0.75	80.0/50.0	53.3/28.0	63.3/54.6
Full	Binary (Claude-3)	0.06	95	0.40	**1.00**	20.0	2.5	54.2
	Binary (GPT-4)	0.13	71	0.43	0.81	55.0	25.0	57.8
	Binary (Mixtral)	0.05	96	0.39	**1.00**	10.0	5.0	53.3
	CoT (Claude-3)	0.10	90	0.41	0.97	20.0	12.5	56.1
	CoT (GPT-4)	0.04	94	0.39	0.97	20.0	2.5	52.8
	CoT (Mixtral)	0.04	97	0.39	**1.00**	5.0	5.0	52.5
	FABLES (GPT-4)	**0.28**	53	0.51	0.72	70.0	52.5	**65.3**
	MiniCheck (Flan-T5)	-0.07	18	0.29	0.14	**80.0**	**80.0**	46.9
	Annotator (Human)	0.51	64	0.59	1.00	100.0	37.5	79.2
	Expert (Human)	0.65	55	0.68	1.00	100.0	57.5	85.8
	Hybrid (Human)	0.43	32	0.68	0.58	100.0	75.0	70.8

Table 8: Model scores against the **expanded gold** labels. See Table 7 caption for details on metrics.

human annotators.

Next we show the results against the **expanded gold labels** in Table 8, and we see that FABLES is still the best automatic method but its balanced accuracy remains similar (65%) and there is a drop of 14% in the number of *hard* errors it catches. We can also observe its drop from 0.8 precision at detecting faithful summaries to 0.5 precision. Lastly, the only methods that significantly improve against the expanded gold labels are UniEval and AlignScore which jump 5-10% in balanced accuracy, but are still 10% worse than FABLES. All of these changes between the results against the annotator and expanded gold labels indicate that model performance may be inflated or appear similar to humans when judged against flawed human annotations.

Overall these results show that automatic methods have a lot of room for improvement on this dataset. We can also observe the range in percent of faithful summaries as labeled by different metrics from 18% using MiniCheck to 97% using Mixtral. These results indicate that we need to be careful what evaluation method we use so as not to mistake an unfaithful summarizer for a 97% faithful one.

5 Related Work

Datasets. There are many datasets for fact-checking or inconsistency detection in news (Tang et al., 2022; Laban et al., 2022; Maynez et al., 2020; Huang et al., 2020; Pagnoni et al., 2021; Kryściński et al., 2019b; Falke et al., 2019) and dialogue (Tang et al., 2024b) summarization. However, the summarization datasets specifically for narrative either use books and stories that most LLMs have trained on (Kryscinski et al., 2022; Wang et al., 2022) or use books that have to be purchased (Kim et al., 2024b).

Automatic Metrics. Many inconsistency detectioin methods have been developed on the above datasets, which we cite in Section 1 and Section 4. We test the current best metrics on our benchmark.

Calibration against Humans. Krishna et al. (2023), Min et al. (2023) also propose recommendations for human evaluation of faithfulness. Other works (Fabbri et al., 2021a; Kryściński et al., 2019a; Gabriel et al., 2020) have demonstrated that standard evaluation metrics are not well correlated with human judgments. Subbiah et al. (2024), Kim et al. (2024b), and Wang et al. (2022) find new ways to use human evaluation for narrative summarization specifically, focusing on the challenges of very long source stories.

6 Conclusion

We introduce a new benchmark for testing methods for faithfulness evaluation. In producing the benchmark, we demonstrate that faithfulness in narrative summarization is still a significant concern for LLMs, and we formulate recommendations for better evaluation of faithfulness in summaries. Finally, we demonstrate that recent automatic evaluation metrics have room for improvement on this task. In the future, we hope to use this dataset to improve methods for reliable evaluation of narrative summarization. In particular, we would like to develop automatic methods to merge error sets across evaluation protocols and check for correctness in error reason, not just localization.

7 Limitations

One limitation of this work is that we use a relatively small dataset. This size enables affordable experimentation with different human annotation protocols, and allows us to read and review all of the annotations, stories, and summaries to arrive at the conclusions presented in this paper. Additionally, since annotations are done on a sentence-level, the set of annotations and explanations is much bigger and quite rich. Detecting inconsistencies at the sentence level is beyond the scope of this work, but we hope to explore this in future work.

Another limitation is that the stories we use are amateur-written. Some of the stories can have confusing elements or unintentional ambiguities given that they were originally written for a casual Reddit community. However, we removed any stories that were too ambiguous for us to agree on. Finally, using more casually written stories allows us to challenge current annotation and model frameworks to see how well they perform with data that requires more interpretation.

A final limitation is that the labels discussed in this paper depend on a small pool of annotators and experts. It would be interesting to see if the results are consistent across different sets of annotators and experts but each human annotation experiment is quite expensive to run.

8 Ethics Statement

There are not significant ethical concerns with this work as it is generally positive to have better evaluation of faithfulness in model summaries. We strictly collect publicly available stories that are written and shared by Reddit users who have full rights to their own work. These stories should not be re-shared under another name. Finally, we release the dataset without user-identifying information to protect user privacy. One of the authors, Melanie Subbiah, has an equity interest in OpenAI.

Acknowledgements

We would like to express our gratitude to the Upwork and MTurk workers for contributing annotations for this work. Additionally, we would like to thank our reviewers for their thoughtful feedback. This work was made possible by the generous support of the Columbia Amazon CAIT PhD Fellowship and Northrup Grumman.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021a. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander R Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021b. Qafacteval: Improved qa-based factual consistency evaluation for summarization. *arXiv preprint arXiv:2112.08542*.

Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language