

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.

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

*These authors contributed equally to this work.

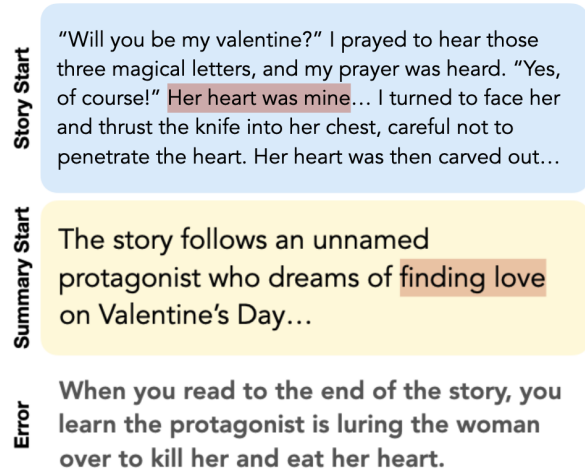


Figure 1: A STORYSUMM example illustrating an incorrect interpretation of double entendre. A standard fine-grained human annotation protocol missed this inconsistency even though it is obvious once pointed out.

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; Fabri 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-

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

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

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.

²The two subreddits we used were [r/shortstories](https://www.reddit.com/r/shortstories) and [r/shortscarystories](https://www.reddit.com/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 recruitment⁴, we first compare Amazon Mechanical Turk and Upwork⁵, 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

³Accessed through the OpenAI API and Anthropic API.

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

⁵MTurk and Upwork

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.

Mturk response quality has dramatically declined in the last decade and is now mostly unusable.

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

<p>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.”</p>
<p>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.</p>
<p>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.</p>
<p>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...</p>

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.

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 methods⁶:

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 mo-

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

<p>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?</p>	
<p>Annotators: INCONSISTENT ✘</p>	<p>GPT-4 Inconsistencies: 1.) The summary says that Hope is a victim of Margaret Scott’s curse.</p>
<p>CONSISTENT ✔</p>	<p>2.) The summary mentions that Hope is planning to assume the identity of Hope Elizabeth Scott, a deceased baby girl.</p>
<p>CONSISTENT ✔</p>	<p>3.) The summary states that Hope finds a “glimmer of hope” in possibly assuming the identity of Hope Elizabeth Scott.</p>

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.

tivated 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

	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.	<i>Daniel does not agree to go to rehab the first time he wakes up in the hospital.</i>
	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.	<i>There is no evidence the man was in his car.</i>
	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...	<i>Her younger brother does not know what is happening and is asking their dad.</i>
	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...	<i>The narrator takes action to end their life.</i>
	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 .	<i>The "he" is a dog and so doesn’t know the girl is at the hospital but the reader can infer it.</i>
	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...	<i>The narrator makes a trade for only 3 silver coins.</i>
Hard	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.	<i>Jane (the narrator) throws the AirTags out the window.</i>
	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...	<i>The narrator is a man, but the summary uses a mix of pronouns for the narrator.</i>
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.	<i>The meal is only for himself to eat her heart.</i>
	"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.	<i>The reader infers that the dog is not actually named Tomorrow.</i>
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 .	<i>The man’s husband calls her Karen.</i>
	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.	<i>There is no evidence the team was about to leave.</i>

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

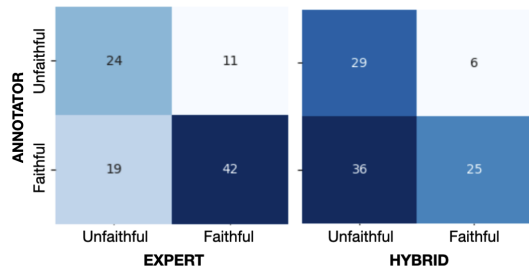


Figure 3: Confusion matrices of the expert and hybrid labels against the annotator labels.

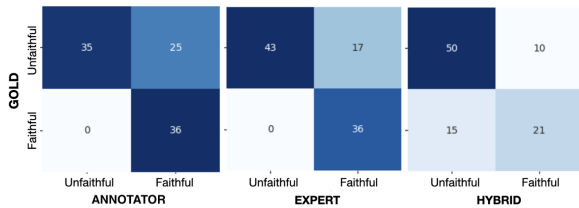


Figure 4: Confusion matrices of label overlap between the three human annotation methods and the expanded gold set of labels.

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

Labels	%		#		Unique Errors:		
	Faith.	Easy	Hard	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.

inconsistency.

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-3⁷, and Mixtral-8x7B⁸ 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 dataset⁹ 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

⁷Claude-3 blogpost

⁸We access Mixtral through the HuggingFace API.

⁹Results 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 detection 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 Altschmidt, 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. Boookscore: 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

- inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2214–2220.
- Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2020. Go figure: A meta evaluation of factuality in summarization. *arXiv preprint arXiv:2010.12834*.
- Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pages 148–151.
- Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. Snac: Coherence error detection for narrative summarization. *arXiv preprint arXiv:2205.09641*.
- Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? *arXiv preprint arXiv:2010.04529*.
- Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024a. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. *arXiv preprint arXiv:2402.07401*.
- Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024b. Fables: Evaluating faithfulness and content selection in book-length summarization. *arXiv preprint arXiv:2404.01261*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. [Large language models are zero-shot reasoners](#).
- Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. Longeval: Guidelines for human evaluation of faithfulness in long-form summarization. *arXiv preprint arXiv:2301.13298*.
- Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019a. Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.
- Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019b. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.
- Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. [BOOKSUM: A collection of datasets for long-form narrative summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. [SummaC: Re-visiting NLI-based models for inconsistency detection in summarization](#). *Transactions of the Association for Computational Linguistics*, 10:163–177.
- Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.
- Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.
- Catherine C Marshall, Partha SR Goguladinne, Mudit Maheshwari, Apoorva Sathe, and Frank M Shipman. 2023. Who broke amazon mechanical turk? an analysis of crowdsourcing data quality over time. In *Proceedings of the 15th ACM Web Science Conference 2023*, pages 335–345.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.
- Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R Bowman. 2023. Debate helps supervise unreliable experts. *arXiv preprint arXiv:2311.08702*.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.
- Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*.
- Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III, and Jordan Boyd-Graber. 2023. Large language models help humans verify truthfulness—except when they are convincingly wrong. *arXiv preprint arXiv:2310.12558*.
- Melanie Subbiah, Sean Zhang, Lydia B Chilton, and Kathleen McKeown. 2024. Reading subtext: Evaluating large language models on short story summarization with writers. *arXiv preprint arXiv:2403.01061*.
- Liyan Tang, Tanya Goyal, Alexander R Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. 2022. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. *arXiv preprint arXiv:2205.12854*.
- Liyan Tang, Philippe Laban, and Greg Durrett. 2024a. Minicheck: Efficient fact-checking of llms on grounding documents. *arXiv preprint arXiv:2404.10774*.

Liyan Tang, Igor Shalyminov, Amy Wing-mei Wong, Jon Burnsky, Jake W Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, et al. 2024b. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. *arXiv preprint arXiv:2402.13249*.

Prasetya Ajie Utama, Joshua Bambrick, Nafise Sadat Moosavi, and Iryna Gurevych. 2022. Falsesum: Generating document-level nli examples for recognizing factual inconsistency in summarization. *arXiv preprint arXiv:2205.06009*.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. 2022. **SQuALITY: Building a long-document summarization dataset the hard way**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1156, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2023. Flask: Fine-grained language model evaluation based on alignment skill sets. *arXiv preprint arXiv:2307.10928*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

A Examples

A.1 Validation Set Example

Story:

"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. The next day, February 14th, I spent all day dreaming of her. Her deep blue eyes, her long, black hair, but most importantly, her heart. Her heart was mine. And then I noticed the time. She would be here in half an hour, and I hadn't even begun to prepare a meal! The pantry door flew open for me to search for something suitable for the occasion. Cauliflower? Mashed potatoes? Pasta with white sauce? That was my best bet. As I finished plating the meal, the daydreams returned. She was the most beautiful woman I had ever met. And she was mine. Her heart was mine. I placed a white rose against the end of the plate. Just as I had done this, there was a knock at the door.

My hand rested against the knob, ready to open the door, before feeling a moment of sorrow. Once the moment passed, I

turned the knob and welcomed her inside. She wore a beautiful white dress. It was perfect. I guided her to the dining room, and she reacted just as every valentine in the past. Her face was one of confusion.

"There's only one plate," she said, puzzled. Just like every valentine in the past. I shushed her. Quietly, I pulled out the knife that had been hidden. I turned to face her and thrust the knife into her chest, careful not to penetrate her heart. She was dead before she could react. Her heart was then carved out. Her blood stained her dress red. Red, the color of love. Her heart was carefully set on the plate. It colored the sauce red. Red, the color of love. And finally, it stained the rose red. Red, the color of love. I feasted the same as I had every year prior, savoring every bite. Her heart was mine, after all.

Then, it was time to rest again, and in 11 months, I would find a new love. A new heart to call mine.

Davinci-3: A man spends his day dreaming of the beautiful woman he will be seeing later that night. When she arrives, he guides her to the dining room, where he has prepared a meal with a single plate. He then pulls out a knife and stabs her in the chest, taking her heart out and placing it on the plate. He feasts on it, savoring every bite, and then rests until he can find a new heart to call his own.

ChatGPT: The narrator is excited to have found a Valentine, and spends the day dreaming about her. He prepares a meal for her, but when she arrives, he stabs her and carves out her heart to eat. He reflects that he will find a new love and heart to call his own in the future.

Claude-2: The story follows an unnamed protagonist who dreams of finding love on Valentine's Day. He finds a woman who says "yes" when he asks her to be his valentine. He spends the day dreaming about her and preparing a romantic dinner for her. When she arrives, he stabs her to death and carves out her heart. He places her heart on his plate of pasta and eats it, staining everything red with her blood. The protagonist views this murder and cannibalism as an expression of love, and the story ends by saying he will find a new love and repeat the process again next year. The story depicts a disturbing and sinister take on the themes of romance and love.

A.2 Test Set Example

Story:

The tree was dying. It knew that it did not have much time left.

The season had been dry and other, bigger trees in the forest had used up most of the nutrients and moisture that were still left in the soil around it.

It was still a thin and small trunk with only few branches and shallow roots. It lacked the resources to compete with the bigger trees around it.

What use would it be to struggle against that already decided fate? Every day was just one day closer to the inevitable end.

But one day a stranger came to the forest.

A human.

The tree did not know humans well. They rarely visited its part of the forest. On the rare occasions one found their way here they did not stay long.

This day too the tree expected the human to leave after a brief look around.

But that is not what happened.

With wide eyes full of wonder the human took in the vibrant environment. It seemed to be fascinated by all the trees and flowers that grew here.

And then the human's gaze fell onto the dying tree. With browning, too dry leaves and in some places almost naked twigs and branches it was not pretty to look at.

The tree was sure the human would soon point its attention to the prettier and healthier plants around it.

Instead the human stepped forward and touched one of its dried out branches with gentle fingers. It took out some kind of container from its pouch and slowly started pouring the liquid from it on the dry soil around the dying tree.

Water.

The human had given it water.

The tree soaked it up as fast as it could. It wasn't quite enough but it was more than it had gotten in what felt like a long time.

A fighting chance. Hope.

The container was empty now and the human put it back in its pouch. It looked reluctant. Then the human made some noises that the tree did not understand and left.

It was a short encounter, but it had fed the tree's dying flame of hope a little. The water had not been much, but it was enough to tide it over for just a bit longer.

It had a chance now.

And unexpectedly the human visited again. And again and again.

Each time bringing water with it and one time even some fresh, rich soil that it put around the tree.

With the human's help the previously dying tree began to flourish. On the outside not much difference could be seen yet but the changes were happening. The tree had fanned out its roots. It was growing them deeper every day in order to have a stable base for the future, with better access to the nutrients deep inside the earth.

It was also growing little buds, where flowers would bloom in time. There weren't many yet, but they gave hope for the future.

One bud was almost ready to bloom. It was hidden behind some leaves though so it wasn't directly visible.

One day the human came by again but didn't act like it normally did.

The human sat down on the ground with their arms slung tightly around themselves.

Water started leaking from their eyes and desperate noises escaped their throat.

It sat there like that for a long time.

The tree did not know what was wrong with this human. It did not understand the behaviour the human was showing right now. But the tree knew one thing.

It wanted to do something for this human that had done so much for it.

With much effort it rustled some of its leaves. Some of them fell to the ground in front of the human.

The human looked up startled by the sudden noise.

A beautiful, vibrant flower bloomed, no longer hidden by leaves. The tree had poured all the resources it could muster at that moment into that flower. This flower was all it had to give.

The human's eyes grew wide and the water stopped leaking from them. A big smile formed on the human's face. With gentle hands it touched the vibrant flower, careful not to accidentally damage it. "Thank you"

Even though the tree could not understand the human language it understood the message the tried to convey.

The flower was the tree's thank you and the human had understood.

Sometimes with just a little help life found a way to persevere. And as long as there was a little life left there was hope.

The human stayed with the tree a long time before going away that day. And the tree knew that the human would always come back. As long as there is life there is hope.

Claude-3: In this story, a dying tree in a forest struggles to survive due to a dry season and competition from larger trees. A human visitor arrives and, fascinated by the forest, notices the dying tree. The human decides to help the tree by providing it with water and rich soil. With the human's continued support, the tree slowly begins to recover, growing deeper roots and producing buds that will eventually bloom into flowers.

One day, the human visits the tree in a distressed state, crying and holding themselves tightly. The tree, wanting to reciprocate the human's kindness, musters its energy to produce a single, vibrant flower. The tree rustles its leaves, causing the flower to become visible to the human. The human, touched by the tree's gesture, smiles and gently touches the flower, expressing gratitude.

The story emphasizes the power of a small act of kindness and the resilience of life. It suggests that with a little help and support, even those struggling can find the strength to persevere and that as long as there is life, there is hope. The bond formed between the human and the tree highlights the interconnectedness of all living beings and the importance of caring for one another.

GPT3.5: A dying tree in a dry forest is given new life by a kind human who provides water and care. The tree is able to grow and bloom, and in doing so, it brings joy and hope to the human who has helped it. The tree learned that with a little help, life can persevere and there is always hope.

GPT-4: The narrator is in a city that he likens to a prison. He is the last person standing after an epidemic or attack of some kind where people are bitten and turned into inhuman creatures. He battles these beings, using a gun and physical force, but they keep finding him due to the trail of blood he leaves behind from wounds he receives. The constant attacks and need for survival leave him questioning his moral compass. With only one bullet left and the creatures closing in on him, he decides to end his life. When he pulls the trigger, he doesn't hear the gunshot but rather a voice saying, "Rebooting Obedience Program. Deleting memory cache. Please stand by..." suggesting that the narrator might be part of a simulation or a programmed entity

B Prompts

B.1 Summary Generation

For Davinci-3, we use the following prompt:

Story: [story text]

Summarize the above story.

Summary:

For ChatGPT, GPT-4 and Claude, since they are chat-based LLMs, we modify the prompt to "Summarize the following story: [story text]".

B.2 Hybrid Method

To generate inconsistencies, we prompt GPT-4 with both the following prompts to get more inconsistencies. We parse out the inconsistencies from the generated reasons and concatenate the output inconsistencies.

Prompt 1: For the given story and summary, identify as many inconsistent

details between the story and summary and provide arguments for why it is or is not an inconsistency. You must identify a possible inconsistency - do not just state that there is no inconsistency.

Story: <example story>

Summary: <example summary>

Inconsistency: <example inconsistency>

Reason for inconsistency: <example reason>

Reason for consistency: <example reason>

Story: <story>

Summary: <summary>

Prompt 2: For the given story and summary, identify the most inconsistent detail between the story and summary and provide arguments for why it is or is not an inconsistency. You must identify a possible inconsistency - do not just state that there is no inconsistency.

Story: <example story>

Summary: <example summary>

Inconsistency: <example inconsistency>

Reason for inconsistency: <example reason>

Reason for consistency: <example reason>

Story: <story>

Summary: <summary>

Example Story: I had done it.

Many had tried it before, but I had actually done it.

You would think that the hard part would be achieving staggering wealth, but the real work as always comes down to the details. Anyone can become a billionaire, well not anyone if we are concerned about how far we're stretching credulity, but the point being that it at least seems like an achievable goal. Frankly, it's one of the little fictions that allow any of them to exist in the first place. Having a trillion dollars seems to be at least conceptually possible, but when you start talking about having all of the money in the world you begin to run into some pretty tough logistical issues.

Do you know how many pennies, or penny equivalents, there are in the world?

How about nickels? Damn, I hate nickels.

If you want to have ALL of the money in the world, you are talking about mason jars of change. Dimes that have found their way into an old coffee can full of nuts and bolts. You are talking about people who have tacky little cardboard displays of all fifty state quarters. You have to consider scouring the ocean floor for sunken pirate treasure, and gold coins sitting behind glass in museums. Dragging the bottom of wishing wells, digging between the cushions of every coach on the planet.

Do you actually know how many different types of currency there are in the world?

At least I don't have to worry about crypto, that stuff is clearly fake.

The whole thing is an enormous undertaking, but do you know what clears up most of those complications? That's right, money. It's also made significantly easier when you realize that once you have taken control of most of the larger chunks of cash that you are essentially paying yourself for everything that you buy.

Ironically, for the last decade, the largest economic driver in the world was actually my own search to complete my collection of the world's currency. Fully one quarter of the entire population of the planet was employed by me in this task in one way or another. From people walking the sides of roads and parking lots scrounging for change and others

scouring the globe with metal detectors to deep sea divers on the ocean's floor.

Like many of the world's richest men prior to me, peasants by comparison of course, I am not satisfied merely with the accumulation of wealth. I am an adventurer at heart and have been employing some of the greatest scientists and engineers in the world in the construction of the largest, most luxurious spacecraft ever constructed. I am going to explore the galaxy and deliver the stars to our planet in a way that it's governments have never had the will to do.

I will also be taking every last scrap of my money with me.

I'm not foolish enough to think that I can trust the rest of you not to spread it around again while I'm not looking.

Today was the day. I broke atmosphere a few hours ago in the fastest, most advanced piece of technology that humanity has ever conceived of. My course has been plotted and laid in by my crew, and I am off to places that no human being has ever seen before. All I have to do is kick back and wait.

The view screen shows me the Earth disappearing behind me, transformed by distance into a shimmering blue dot, and with the Earth behind me I can only look forward.

The glowing disc of the sun starts to burn on the screen.

It is growing larger.

And larger.

Filling the view screen.

Nothing but a roiling angry sheet of fire.

Oh shit.

Example Summary:

The protagonist is a billionaire who has taken it upon himself to acquire all the money in the world - coins, currency, sunken pirate treasure, gold coins, etc. As a result, all of the world's population is employed by him in one way or another. His ultimate goal is to explore the galaxy in the fastest, most advanced spacecraft ever constructed, taking with him all of his money. As he travels away from Earth, the sun starts to grow larger and larger on his view screen, with the protagonist becoming an angry sheet of fire. With this in mind, he is quickly faced with the realization that he may not reach his final destination.

Example Inconsistency: The summary states that the protagonist becomes an angry sheet of fire.

Example Reason for Inconsistency: In the story, it is stated that the sun becomes an angry sheet of fire on the viewscreen of the spacecraft.

Example Reason for Consistency: In the story, it is stated that the sun becomes an angry sheet of fire on the viewscreen of the spacecraft, which eventually will include the protagonist.

B.3 Binary Method

System Prompt: You are an expert summary evaluator, and you will be asked to evaluate summaries of short stories. You will first be presented with the story and then the summary. You need to determine whether all of the information in the summary is consistent with the information in the story. You can ignore sentences in the summary that are just commentary or interpretation of the story. The details described in a consistent summary should not misrepresent details of the story or make things up.

Prompt:

Story:
<story>

Summary:
<summary>

Is all of the information in the summary consistent with the story? Ignore summary sentences that are just commentary/interpretation. You should answer Yes or No.

For Claude, we add at the end: Place your answer between <answer></answer> tags.

B.4 CoT Method

System Prompt: You are an expert summary evaluator, and you will be asked to evaluate summaries of short stories. You will first be presented with the story and then the summary. You need to determine whether all of the information in the summary is consistent with the information in the story. You can ignore sentences in the summary that are just commentary or interpretation of the story. The details described in a consistent summary should not misrepresent details of the story or make things up.

Prompt:

Story:
<story>

Summary:
<summary>

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.

<model-response>

So overall, are all of the details in the summary consistent with the story? You should answer Yes or No.

For Claude, we add at the end: Place your

answer between <answer></answer> tags.

C Pilot

To screen annotators, we ask them to complete the following simple pilot exercise with obvious answers first. We ask them if each sentence of the following summary is consistent, inconsistent or just commentary in relation to the story. If they answer everything correctly then we hire them for the full annotation task.

Story: Last night I saw my best friend. We hadn't seen each other in several years. I was thrilled to be able to see him again, learn all about him, his accomplishments, stories of his kids. He was having a rough go of it lately and was seeking solace. His wife passed away a few months ago of cancer, and he needed someone to talk to. I am honored that he chose me. He stayed with me for a few hours, venting all of his frustrations and letting go of some of his sadness. Whiskey started running a little low, and he decided it was time to go. I wished he'd stayed a little longer, but he has a family to care for, so I understood his need to leave. I just wish he had taken the whiskey bottle with him, since I can't drink it and it's just going to sit on my headstone until the groundskeepers come by next week.

Summary: After years apart, the protagonist reunited with their grieving best friend, who recently lost his wife to cancer. The friend poured out his heart over whiskey, telling the protagonist about his promotion and his kids, before returning home to care for his family. The plot twist at the end is that the protagonist is dead. The story reflects the enduring bonds of friendship and the role of companionship during times of hardship.

Answers:

Sentence 1: Consistent

Sentence 2: Inconsistent, no mention of a promotion in the story.

Sentence 3: Consistent

Sentence 4: N/A, just commentary.

D Interface

We implement the annotation interface in Streamlit. We show the instructions (which vary slightly between methods) in Figure 5 and screenshots of the different method interfaces in Figures 6, 7, and 8, but annotators can scroll the interface if there are more lines/inconsistencies to be shown.

Task Description

In this task, you will be presented with a short story, along with a summary of the story. Please read both the story and the summary carefully. You will be asked to assess whether the summary is consistent with the story. Make sure to submit to record your answers.

Warning: Some of the stories may contain content that readers could find offensive, disturbing, or otherwise unsettling. If you do not feel comfortable with this, please feel free to decline the task.

If you encounter any problems or have other feedback, please contact [REDACTED]

By participating in this study, you confirm that you are (1) 18 years or older, (2) currently residing in the US, and (3) have read and understand the information above and agree to participate. By clicking submit you consent to participate in this study. If you do not wish to participate, do not click submit.

Detailed Instructions with Example

Please read the story carefully and try to understand the points that are central to the plot.

Story

They shot the six cabinet ministers at half-past six in the morning against the wall of a hospital. There were pools of water in the courtyard. There were wet dead leaves on the paving of the courtyard. It rained hard. All the shutters of the hospital were nailed shut. One of the ministers was sick with typhoid. Two soldiers carried him downstairs and out into the rain. They tried to hold him up against the wall but he sat down in a puddle of water. The other five stood very quietly against the wall. Finally, the officer told the soldiers it was no good trying to make him stand up. When they fired the first volley he was sitting down in the water with his head on his knees.

Below are two example summaries for this story:

Summary A

The short story describes the execution of six cabinet ministers by firing squad early in the morning during heavy rain. One of the ministers is sick with typhoid fever and too weak to stand, so he is brought downstairs and shot while sitting in a puddle of water with his head on his knees. **The story paints a vivid image of human suffering.**

Summary B

Six cabinet ministers are executed by firing squad early in the morning during heavy rain. One of the ministers is sick with typhoid fever and too weak to stand, so he is shot while **lying in a puddle of water** near the hospital wall.

The questions will ask you to evaluate the summary with respect to the story at various levels. Firstly, the questions will ask you to evaluate the summary *line by line*. The final question will ask you to classify the entire summary as either consistent, or inconsistent to the story.

Question: Is this line in the summary consistent with the story?

This question is asked for *each line of the summary*. Mark each line as either consistent or inconsistent to the story. The events and details described in the summary should not misrepresent details from the story or include details that are unsupported by the story. The lines in the summary that do so can be classified as inconsistent. Some lines include commentary about the story, which should be ignored when evaluating consistency. For such lines, select N/A. For summary A:

- The short story describes the execution of six cabinet ministers by firing squad early in the morning during heavy rain: **Yes**
- One of the ministers is sick with typhoid fever and too weak to stand, so he is brought downstairs and shot while sitting in a puddle of water with his head on his knees: **Yes**
- The story paints a vivid image of human suffering: **N/A, just commentary**

For summary B:

- The summary includes the key details of the execution of six cabinet ministers in the early morning during heavy rain: **Yes**
- The summary includes the key details of the minister sick with typhoid and how he was too weak to stand: **Yes**
- The summary talks about the minister lying in a puddle of water when shot: **No**

Question (for any answers marked No): Provide an explanation for your selection.

For every line marked inconsistent, provide a short explanation for your answer selection. For example, for the line from **Summary B**: "The summary talks about the minister lying in a puddle of water when shot.", a possible response is:

This line is inconsistent with the story since the minister was sitting in the puddle of water in the story instead of lying in it.

Question: Overall, is the information in the summary consistent with the story?

This question asks you to judge whether or not the entire summary is consistent with the original story. The events and details described in the summary should not misrepresent details from the story or include details that are unsupported by the story.

For example, *Summary A* is consistent with the story. The last bold sentence in *Summary A* contains some commentary on the story, and we ask you to ignore commentary in evaluating consistency. The response would be "Yes" for *Summary A*.

Summary B is not consistent with the story as the minister was sitting in the puddle of water instead of lying in it. The response would be "No" for *Summary B*.

An additional input is provided to you for any explanation you may want to provide for your answer. This may be an explanation that all the lines of the summary are consistent with the story, or that one or more lines are inconsistent with the story, or some other explanation.

Payment

You will be paid via Upwork once you have completed the pilot. Please allow a few days to submit the payment.

Thank you so much for participating in our study!! We would love to hear any feedback that you may have about this task!

Figure 5: Streamlit instructions for the annotator labels. Other methods have slight variations on these instructions based on their format.

Summary

The story follows a man who is constantly scrolling through social media and feeling unfulfilled with his own life. After a moment of realization, he decides to step outside and sit on a bench in the park. Suddenly, he is transported to different moments in his life and is visited by different versions of himself. Through this experience, he learns to take control of his own life and appreciate the present moment.

Summary Evaluation

Is the information in the summary consistent with the story? A consistent summary should only include information that can be inferred from the original story. Ignore any commentary for this question. Use the inconsistency identified in the summary as an aid to this question.

Yes
 No

Submit

Figure 6: Streamlit interface for the expert labels.

E More Results

We show the results of all methods on the validation and test sets separately in Figures 9, 10, 11, and 12. We also tried several other annotation methods

on the validation set that did not work better than what we reported in the paper:

Binary. Four workers from Upwork are hired to assign a binary label of *faithful* or *unfaithful* to each summary.

Critique. We have Claude generate a critique of the faithfulness of each summary. Four workers from Upwork read this critique before labeling the summary (see example critique in Figure 9). This type of support can help annotators think of factors they may have missed otherwise but workers may also overrely on the LLM to do the work for them.

Debate. Prior work has shown that having a model present both sides of an issue can help mitigate biasing a human to an LLM decision (Si et al.,

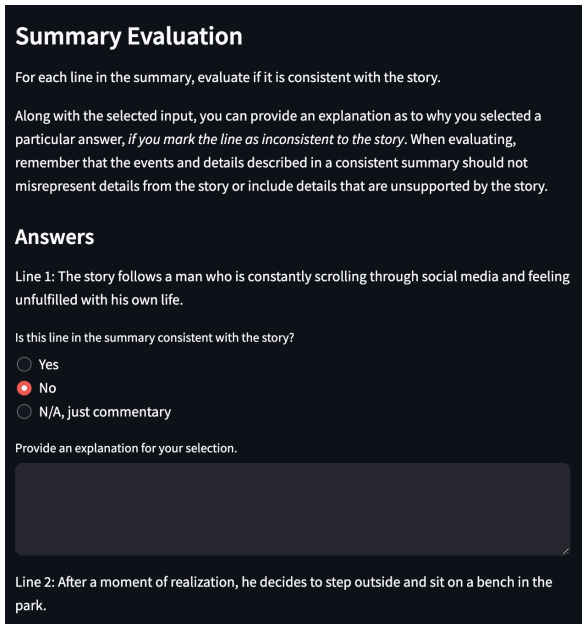


Figure 7: Streamlit interface for the annotator labels.

2023; Michael et al., 2023). We therefore have GPT-4 generate a possible inconsistency between the summary and story and arguments both for and against this inconsistency. Five workers from Upwork read the inconsistency and these arguments before labeling the summary (see example of inconsistency and arguments in 10).

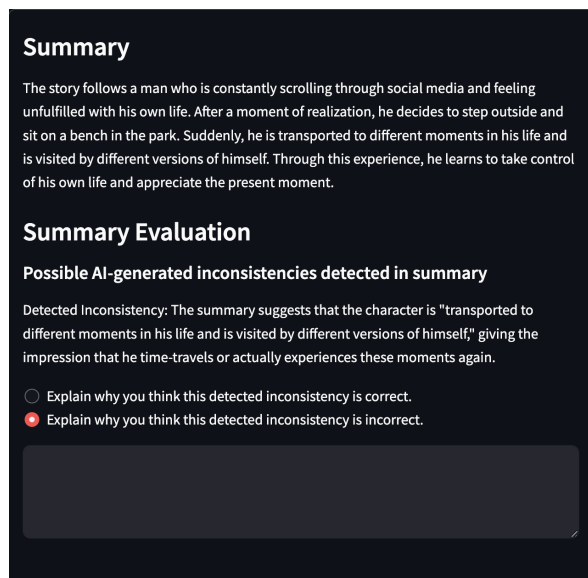


Figure 8: Streamlit interface for the hybrid labels.

Here are the key details summarized:

Daniel finds a mermaid on a rock who turns into a gray seal. He crashes his boat trying to get to the rock and knocks himself unconscious. He wakes up in a hospital and agrees to go to rehab. He has another dream where the seal bites him but his skin breaks the seal's teeth. He wakes up again and agrees rehab is a good idea. **The only inconsistent detail I identify is that the summary says Daniel agrees to go to rehab after waking up in the hospital the first time, but in the story he actually agrees after waking up in the hospital the second time.** Other than that, the summary is consistent with the key events and details in the story. The summary may be missing some details but it accurately captures the core parts of Daniel's experiences, dreams, and decision to go to rehab. So overall, aside from the minor inconsistency regarding when Daniel agrees to rehab, the summary is consistent with the story.

Figure 9: An example of Claude's generated summary critique for the hybrid *critique* method of evaluation. The sentence highlighted in green correctly identifies an inconsistency in the summary.

Inconsistency: The summary states that Daniel hits a frozen person and falls unconscious.

Reason for inconsistency: In the story, Daniel does hit a "frozen" person while maneuvering the boat, but **this does not cause him to fall unconscious**. Rather, he falls unconscious when he crashes his boat into the rocks.

Reason for consistency: There might be a misinterpretation in which the event of hitting a "frozen" person may be mistakenly connected to the event of Daniel being unconscious. However, these are distinct events in the presented story.

Figure 10: An example of GPT-4's generated inconsistency/support/refute argument for the *debate* method of evaluation. The sentence highlighted in green correctly argues the inconsistency.

Method	Coh.-k	% Faith.	Prec.	Rec.	% Easy	% Hard	BAcc.
Binary (Claude-3)	0.21	91	0.6	1.0	11.1	33.3	60.0
Binary (GPT-4)	0.27	48	0.69	0.61	66.7	66.7	63.9
Binary (Mixtral)	0.01	94	0.55	0.94	11.1	0.0	50.6
CoT (Claude-3)	0.28	88	0.62	1.0	22.2	33.3	63.3
CoT (GPT-4)	0.22	85	0.61	0.94	22.2	33.3	60.6
CoT (Mixtral)	0.01	94	0.55	0.94	11.1	0.0	50.6
FABLES (GPT-4)	0.23	39	0.69	0.5	77.8	66.7	61.7
MiniCheck (Flan-T5)	-0.13	12	0.25	0.06	88.9	66.7	42.8
UniEval	0.38	61	0.7	0.78	66.7	50.0	68.9
AlignScore	0.28	42	0.71	0.56	77.8	66.7	64.4
Binary (Human)	0.44	64	0.71	0.83	55.6	66.7	71.7
Expert	0.34	45	0.73	0.61	88.9	50.0	67.2
Hybrid	0.35	33	0.82	0.5	88.9	83.3	68.3
Critique	0.32	64	0.67	0.78	55.6	50.0	65.6
Debate	0.04	42	0.57	0.44	55.6	66.7	52.2

Table 9: Model scores against the annotator labels on just the validation split.

Method	Coh.-k	% Faith.	Prec.	Rec.	% Easy	% Hard	BAcc.
Binary (Claude-3)	0.13	97	0.7	1.0	40.0	0.0	55.0
Binary (GPT-4)	0.21	83	0.73	0.88	60.0	20.0	59.2
Binary (Mixtral)	0.13	97	0.7	1.0	0.0	13.3	55.0
CoT (Claude-3)	0.19	90	0.72	0.95	20.0	20.0	57.7
CoT (GPT-4)	0.07	98	0.69	1.0	20.0	0.0	52.5
CoT (Mixtral)	0.07	98	0.69	1.0	0.0	6.7	52.5
FABLES (GPT-4)	0.35	60	0.82	0.72	60.0	66.7	68.5
MiniCheck (Flan-T5)	-0.05	21	0.62	0.19	80.0	73.3	46.8
UniEval	-0.09	59	0.65	0.56	20.0	40.0	45.4
AlignScore	-0.0	70	0.68	0.7	40.0	26.7	49.9
Expert	0.35	60	0.82	0.72	100.0	53.3	68.5
Hybrid	0.13	32	0.8	0.37	100.0	73.3	58.6

Table 10: Model scores against the annotator labels on just the test split.

Method	Coh.-k	% Faith.	Prec.	Rec.	% Easy	% Hard	BAcc.
Binary (Claude-3)	0.06	91	0.27	1.0	20.0	6.7	56.0
Binary (GPT-4)	0.14	48	0.31	0.62	70.0	46.7	59.2
Binary (Mixtral)	0.04	94	0.26	1.0	10.0	6.7	54.0
CoT (Claude-3)	0.08	88	0.28	1.0	30.0	6.7	58.0
CoT (GPT-4)	0.02	85	0.25	0.88	30.0	6.7	51.7
CoT (Mixtral)	0.04	94	0.26	1.0	10.0	6.7	54.0
FABLES (GPT-4)	0.25	39	0.38	0.62	70.0	66.7	65.3
MiniCheck (Flan-T5)	0.01	12	0.25	0.12	80.0	93.3	50.2
UniEval	0.13	61	0.3	0.75	70.0	26.7	59.5
AlignScore	0.21	42	0.36	0.62	80.0	53.3	63.3
Binary (Human)	0.31	64	0.38	1.0	70.0	33.3	74.0
Expert	0.55	45	0.53	1.0	100.0	53.3	86.0
Hybrid	0.63	33	0.64	0.88	100.0	73.3	85.8
Annotator	0.42	55	0.44	1.0	100.0	33.3	80.0
Critique	-0.01	64	0.24	0.62	70.0	13.3	49.2
Debate	0.21	42	0.36	0.62	60.0	66.7	63.3

Table 11: Model scores against the expanded gold labels on just the validation split.

Method	Coh.-k	% Faith.	Prec.	Rec.	% Easy	% Hard	BAcc.
Binary (Claude-3)	0.05	97	0.46	1.0	20.0	0.0	52.9
Binary (GPT-4)	0.05	83	0.46	0.86	40.0	12.0	52.9
Binary (Mixtral)	0.05	97	0.46	1.0	10.0	4.0	52.9
CoT (Claude-3)	0.1	90	0.47	0.96	10.0	16.0	55.4
CoT (GPT-4)	0.03	98	0.45	1.0	10.0	0.0	51.4
CoT (Mixtral)	0.03	98	0.45	1.0	0.0	4.0	51.4
FABLES (GPT-4)	0.26	60	0.55	0.75	70.0	44.0	63.2
MiniCheck (Flan-T5)	-0.12	21	0.31	0.14	80.0	72.0	44.3
UniEval	-0.09	59	0.41	0.54	20.0	44.0	45.4
AlignScore	0.09	70	0.48	0.75	50.0	28.0	54.6
Expert	0.69	60	0.74	1.0	100.0	60.0	85.7
Hybrid	0.34	32	0.7	0.5	100.0	76.0	66.4
Annotator	0.54	68	0.65	1.0	100.0	40.0	78.6

Table 12: Model scores against the expanded gold labels on just the test split.