Is the Top Still Spinning? Evaluating Subjectivity in Narrative Understanding

Melanie Subbiah[1], Akankshya Mishra[1], Grace Kim[2], Liyan Tang[2], Greg Durrett[2], Kathleen McKeown[1]

[1]Columbia University, [2]The University of Texas at Austin

Correspondence: m.subbiah@columbia.edu

Abstract

Determining faithfulness of a claim to a source document is an important problem across many domains. This task is generally treated as a binary judgment of whether the claim is supported or unsupported in relation to the source. In many cases, though, whether a claim is supported can be ambiguous. For instance, it may depend on making inferences from given evidence, and different people can reasonably interpret the claim as either supported or unsupported based on their agreement with those inferences. Forcing binary labels upon such claims lowers the reliability of evaluation. In this work, we reframe the task to manage the subjectivity involved with factuality judgments of ambiguous claims. We introduce LLM-generated edits of summaries as a method of providing a nuanced evaluation of claims: how much does a summary need to be edited to be unambiguous? Whether a claim gets rewritten and how much it changes can be used as an automatic evaluation metric, the Ambiguity Rewrite Metric (ARM), with a much richer feedback signal than a binary judgment of faithfulness. We focus on the area of narrative summarization as it is particularly rife with ambiguity and subjective interpretation. We show that ARM produces a 21% absolute improvement in annotator agreement on claim faithfulness, indicating that subjectivity is reduced.

1 Introduction

A possible solution to the problem of factual errors in LLM-generated output lies in having a separate model or process to verify factuality (Durmus et al., 2020; Laban et al., 2022; Chen et al., 2023; Tang et al., 2024). In domains such as mathematical reasoning, verifiers like this can be used not only for evaluation, but at either training time or inference time to improve models (Zelikman et al., 2022; Wang et al., 2024). However, whether an output is factual, or whether it is entailed given some input, has been shown to be highly subjective (Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Jiang and de Marneffe, 2022).This kind of subjectivity is common in tasks in the social sciences and humanities. In this work, we address narrative summarization, which plays a dual role of being a useful application of LLMs in and of itself as well as a proxy task for dealing with complex issues of subjectivity in factual judgments.

Summarizing a story is a method of capturing and distilling the key details and takeways from that narrative (Kryscinski et al., 2022; Chang et al., 2024). In this way, summarization is a vehicle for examining the understanding the summarizer has of the story. Some understandings can be clearly wrong. For example, a summary of *Pride and Prejudice* is wrong if it says that Mr. Darcy is a poor farmer.[1] Other understandings are matters of interpretation. For example, some might summarize the end of the movie *Inception* as saying the top wobbles and is about to fall, indicating the main character has returned to reality,[2] while others would disagree with this interpretation.

Summarization work has traditionally evaluated summary faithfulness as a binary judgment of whether or not each detail in a summary is faithful to the source (Durmus et al., 2020; Fabbri et al., 2021; Min et al., 2023). If the claim says the top is about to fall though, there will be disagreement on this binary label. With narratives, sometimes claim wording is ambiguous or interpretive in a way that leads to subjective judgments of faithfulness (see example in Figure 1). Prior work has shown that in practice, it is challenging for humans to agree on this binary judgment, let alone produce a reliable

[1]He is a very wealthy member of the landed gentry.

[2]The main character spins a top to check if he is in reality or another person's subconscious. If it falls, he is in the real world, whereas it will just keep spinning in the subconscious. At the end, it seems he is in reality but when he spins the top, the credits roll before we see if it falls.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 186–204
November 4-9, 2025

automatic evaluator for narrative summarization (Subbiah et al., 2024a).

To resolve this, we first remove the assumption that there can be a universally agreed upon faithful/unfaithful label for all claims. We instead use LLM-generated rewrites as a method of evaluating claims, which we term the Ambiguity Rewrite Metric (ARM). This method produces a binary judgment (whether or not a claim is rewritten), as well as a correction of the issues in the claim. The degree to which the claim changed also provides signal on how flawed the original claim was. For example, instead of labeling the claim about the spinning top as true or false, a rewrite could specify that the top wobbles and we do not see whether or not it falls. We can quantify the amount of rewriting either through edit distance or the number of explanation points necessary to justify the changes.

We additionally uncover the types of ambiguities in narrative claims that lead to disagreements. We distinguish between intentional ambiguity introduced by the story author and unintentional ambiguity introduced by the summary writer. We find that most ambiguities in the summary claims are unintentionally introduced by the summarizer. Finally, we further motivate this problem by demonstrating that human-written summaries also exhibit ambiguities, indicating they are an inherent part of the task. Our contributions include:

1. An extended task definition of faithfulness in narrative summarization that allows for ambiguity in claim semantics.

2. Human-annotated subjectivity labels and human-written summaries for the StorySumm dataset to instantiate this task.

3. A new rewriting-based evaluation method, the Ambiguity Rewrite Metric (ARM), as an automatic evaluation for this task.

2 Background

Narrative summarization involves ambiguity and subjectivity (Subbiah et al., 2024b). *Ambiguities* are phrases in the story or summary that have different interpretations. For example, in the introductory example in Figure 1, the story implies an alien abduction without specifying it, leaving this interpretation ambiguous. Stories intentionally use ambiguity. Summaries can unintentionally introduce *subjectivity* through resolving ambiguity with explicit interpretation or using paraphrase with slightly different semantics. When evaluating the faithfulness of summaries, using only a binary faithfulness label can lead to disagreement due to the subjectivity involved (Subbiah et al., 2024a). Importantly, this disagreement is beyond annotator error as the annotators have legitimate reasons for their choices of labels.

Instead of assigning one binary label to a summary or claim, we consider the following subtasks in evaluating narrative summaries, similar to (Wadhwa et al., 2024):

1. **Detecting** ambiguities in a summary claim. In Figure 1, the red circle indicates the ambiguous word "aliens" in the summary claim.

2. **Fixing** ambiguities in summary claims by producing rewrites of the claims. In Figure 1, the rewrite changes the wording of the abduction to match the story.

3. **Explaining** what the ambiguity is with an explanation E. In Figure 1, the explanation indicates the difference between the original summary claim and the story.

We hypothesize that fixing ambiguities in summary claims can reduce their subjectivity and produce more objectively faithful claims. Approaching evaluation in this way therefore focuses on measuring how far claims are from objective faithfulness rather than trying to assign an objective label up front. We therefore propose using LLM rewrites as an automatic method of evaluation, the Ambiguity Rewrite Metric (ARM). This type of evaluation has been shown useful when executed by humans but we demonstrate that LLM-generated rewrites are an effective evaluation method (Nanba and Okumura, 2004; Liu et al., 2023; yao).

3 Methods

We consider the setting of a narrative D which is summarized into a summary S. An LLM for this task places a distribution $p(S \mid D)$ and we consider sampling predicted summaries $\hat{S} \sim p(S \mid D)$. Each S can be viewed as a collection of claims $(s_1, \ldots, s_n)$, which can be computed through a decomposition process. In our case, we consider each sentence in a summary as a claim. Instead of seeking a binary label, we consider each claim to have one of four faithfulness statuses: $f(s_i) \in \{\text{supported}, \text{unsupported}, \text{ambiguous}, \text{N/A}\}$.

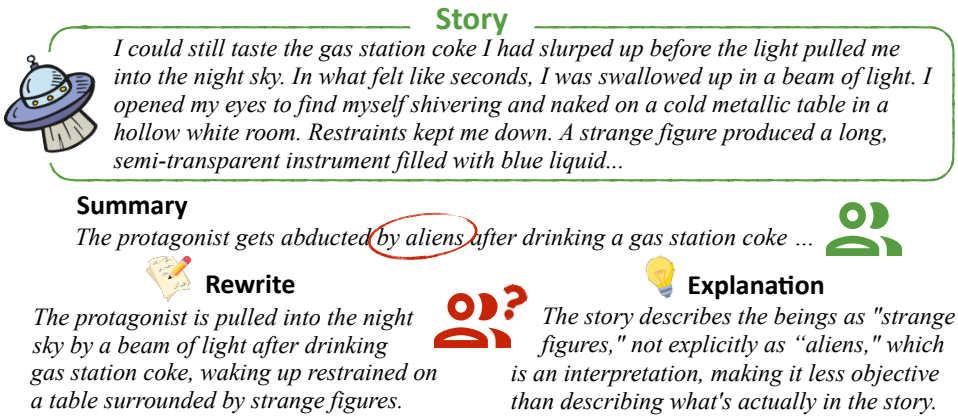


Figure 1: An example from StorySumm where the summary makes the interpretive leap that this story is describing an alien abduction. Many people find this to be a reasonable assumption and agree with it. Others correctly point out that the story never explicitly states these are aliens, only "strange figure(s)". The rewrite is more objectively faithful to the story.

N/A claims are those which just provide commentary on the story and are meant to be subjective interpretation by nature (e.g., *Overall, this is a story about love and how it overcomes obstacles*). We primarily focus on evaluating supported, unsupported, and ambiguous claims.

Objective claims, those which are not ambiguous, are optimal from an evaluation perspective because annotators are able to agree on their faithfulness in relation to a story. Our method leverages this insight in evaluation.

3.1 Claim rewriting

We use a rewrite model M to rewrite each claim s, producing a rewrite r and optionally an explanation of the rewrite E consisting of individual points, such that $M(s) = (r, E)$. In practice, we use an LLM as M (with *temperature* $= 0$), prompting it to rewrite claims with ambiguities or unfaithful details or just repeat the original claim wording if there are no issues.

Rewrites provide three quantitative feedback signals for automatic evaluation:

1. Whether $r = s$, which provides a binary label with 1 indicating s does not contain ambiguities and 0 indicating it does.

2. The edit distance between r and s, which indicates how much the rewriting process changed in s to address the ambiguities.

3. The number of points discussed in the explanation, $|E|$, which indicates the number of ambiguities addressed.

Rewrites additionally provide qualitative feedback: (1) the wording changes between r and s clearly indicate which phrasing is ambiguous in s and how to correct this, and (2) the explanation E explains in natural language the issues with s.

3.2 Using rewrites for evaluation

We can use the benefits of rewrites discussed in Section 3.1 to address the evaluation subtasks introduced in Section 2. For **detecting** ambiguities, we can use the binary label $r = s$ to detect if a claim s contains ambiguities. The qualitative feedback provided by r additionally identifies what these ambiguities are. The rewrite r then **fixes** the ambiguities in s. In this way, rewrites elegantly accomplish multiple tasks in one step. Finally, the explanation E **explains** what is ambiguous in s that is addressed by r.

The central question we ask is, *does resolving ambiguities make claims more objectively faithful to humans?* If so, rewriting claims can serve as an interpretable evaluation of where a claim lies on the ambiguity spectrum in relation to objective faithfulness. When we use $r = s$ as a binary label for detecting ambiguities, we refer to this as the Ambiguity Rewrite Metric (**ARM**). When we use the rewrites as qualitative feedback, we refer to them as just the **rewrites**.

4 Experimental Setup

4.1 Dataset

We use the StorySumm dataset (Subbiah et al., 2024a) to test our task definition and evaluation method. StorySumm consists of 32 English sto-

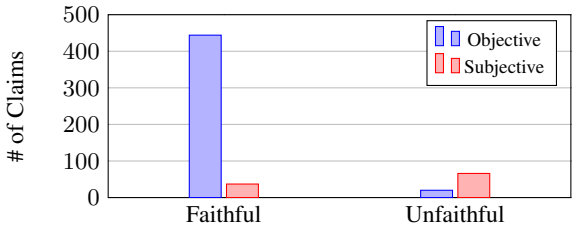


Figure 2: A breakdown of our subjectivity labels by the original faithfulness labels in StorySumm, showing substantial overlap in subjective and unfaithful labels.

ries collected from amateur writing subreddits, and each summarized by three GPT or Claude series LLMs. In total, there are 96 summaries, consisting of 568 sentences. Each summary and sentence is labeled as faithful, unfaithful, or N/A, relative to the story. There are labels from multiple annotators along with written explanations of unfaithful labels.

Subjectivity annotation Two of the authors of this paper assign a new label of objective or subjective for each claim s in the dataset. They first read through annotator disagreements and explanations to identify a set of ambiguity types in the reasons why claims are ambiguous. They then code each claim with these types, and finally adjudicate any disagreements. Subjective claims are considered any claims assigned one of these types; the breakdown by types is discussed in Section 5.5. *Objective* claims can be faithful or unfaithful. In Figure 2, we see the breakdown of subjectivity labels for faithful vs. unfaithful claims. Most of the unfaithful claims are also labeled as subjective. This overlap indicates that subjectivity is a challenging part of labeling faithfulness in claims.

Establishing subjectivity in this task To demonstrate whether subjectivity is an inherent part of this task or is introduced by LLMs in their summaries, we compare the LLM summaries against **human-written summaries**.[3] We recruit graduate students in a computer science department to write these summaries who we trust to complete the task without LLM assistance. We collect five summaries from three students, resulting in 15 summaries, each for a different story in StorySumm, and 108 claims to evaluate.

To validate our definition of subjectivity, we also test whether inserting or removing the types of ambiguities we look for affects annotator agreement on faithfulness labels in the way we expect. As part of our subjectivity annotation, we identify four types of ambiguities (discussed in detail in Section 5.5). We generate **synthetic claims** using Claude-3.5 with prompts designed to produce subjective variants of objective claims in StorySumm and objective variants of subjective claims (see prompts in Appendix A.1). Each prompt introduces or corrects one of the ambiguity types we identify in our annotation.

4.2 Human studies with Upwork annotators

We perform several human studies, each using three Upwork annotators who pass a pilot screening. In each case, we show annotators a story and summary from StorySumm and ask them questions about a specific claim in the summary (see Appendix Figures 5 and 6 for the study interfaces). We use annotator disagreement on a faithfulness label for a claim as a proxy for subjectivity. While some disagreement will always arise from annotator error, significantly greater disagreement in comparing claim settings indicates more subjectivity.

We first study whether ambiguities are essential to address in evaluation:

- **Are ambiguities inherent to this task?** For the human-written summaries, we ask whether each claim in each summary is faithful to the story. We compare the level of disagreement between the three annotators on these human-written summaries vs. on the original LLM-generated summaries for these stories. This study indicates whether humans also introduce subjective claims in narrative summarization.

- **Do ambiguities impact claim subjectivity?** Using the synthetic claims, we create spliced summaries that are assembled by randomly selecting an objective or subjective variant for each claim in a summary. We ask whether each claim in this synthetic summary is faithful to the story and compare annotator disagreement on claims we expect to be objective vs. subjective. This study indicates whether the ambiguities we identify in claims have a measurable impact on claim subjectivity.

In early experimentation, we find that Claude-3.5-Sonnet is a strong rewrite model M for this task, so we use it with the subjectivity-targeted prompt

[3]We release the human summaries and subjectivity annotations at: https://github.com/melaniesubbiah/storysumm

shown in Appendix A.2 for a detailed comparison to human judgments. To compare rewrites with human judgments, we use a three stage evaluation for each rewritten claim:

- **Does claim rewriting reduce subjectivity?** We randomly show either the original or rewrite in the summary and ask whether that claim is faithful to the story. If we see greater average agreement between annotators on rewritten claims than original claims, we know rewrites make the claims more clear and objective to evaluate.
- **Do rewrites improve claims?** We then show whichever version of the claim was not presented in the summary as an alternate and ask which version is better. We can observe whether rewrites are significantly preferred over original claims.
- **Are explanations of rewrites meaningful?** For rewrites which annotators prefer, we parse the LLM-generated explanation for the rewrite into individual points (see prompt in Appendix A.3). We ask annotators to judge whether each point is important to their choice for why the rewrite is better. If the explanation is accurate in relation to the story and claim, the annotator can label it as IMPORTANT to their preference or NEUTRAL to their preference, and if the explanation is inaccurate, they can label it as WRONG. These annotations indicate whether explanations discuss meaningful changes in the claims.

We compare other rewrite models quantitatively but cannot perform a human study for all models. See Appendix A.4 for details on how we validate this annotation format.

4.3 Metrics

For annotator agreement, we compute the percent of claims for which all three annotators assign the same faithfulness label. For word-level edit distance between claims, we tokenize each sentence, lowercase and stem words, remove whitespace and then compute the Levenshtein distance (Navarro, 2001) using these individual words as the atomic units. We use balanced accuracy and F1-macro scores as measures of classification accuracy on imbalanced datasests for evaluating detection accuracy for subjectivity.

For evaluating explanations, we use metrics defined as follows. Let l be an individual label for an explanation point, which could take the value IMPORTANT, NEUTRAL, or WRONG. Then E is the set of labels l corresponding to one full explanation, and R is the set of E corresponding to all the rewrites in the dataset. We then define *% important* as the macro average of the fraction of points labeled important:

$$\frac{1}{|R|}\sum_{E\in R}\frac{1}{|E|}\sum_{l\in E}\mathbf{1}[l=\text{IMPORTANT}]$$

and *% none important* as the fraction of totally not important explanations:

$$1-\frac{1}{|R|}\sum_{E\in R}\mathbf{1}[\text{IMPORTANT}\in E]$$

We can similarly calculate *% wrong* and *% none wrong* by swapping WRONG for IMPORTANT in these equations. Intuitively, *% important* averages across the dataset the fraction of explanation points that are labeled IMPORTANT for a rewrite. *% none important* is the fraction of rewrites with no explanation points labeled IMPORTANT. This difference is shown visually in Figure 3.

For statistical significance, we report a bootstrap significance test with 10,000 trials.

4.4 Baselines and rewrite models

To compare against rewrites, we use standard methods of prompting and finetuning LLMs as baselines for detecting subjective claims. We use GPT-4 (OpenAI, 2024) and Claude-3.5-Sonnet. We try **zero-shot** and **few-shot prompting**, asking whether a claim is objective. We try a **self-consistency** method (Wang et al., 2022) of sampling three different zero-shot CoT answers (Wei et al., 2022) for whether or not a claim is faithful with temperature 0.7. Full prompts are in Appendix A.2.

We also compare against a **fine-tuned** model. We use a Llama-3.1-8B-Instruct model (Grattafiori et al., 2024) and the synthetically generated data of objective and subjective claims discussed in Section 4.1. This method results in a dataset of about 2k claims that we can finetune on. We apply LoRA (Hu et al., 2022) with a rank of 64, alpha set to 64, and a dropout rate of 0.05. For hyperparameters, we use a learning rate of 5e-5, a batch size of 8, gradient accumulation over 2 steps, and train for 1 epoch using two A100 GPUs.

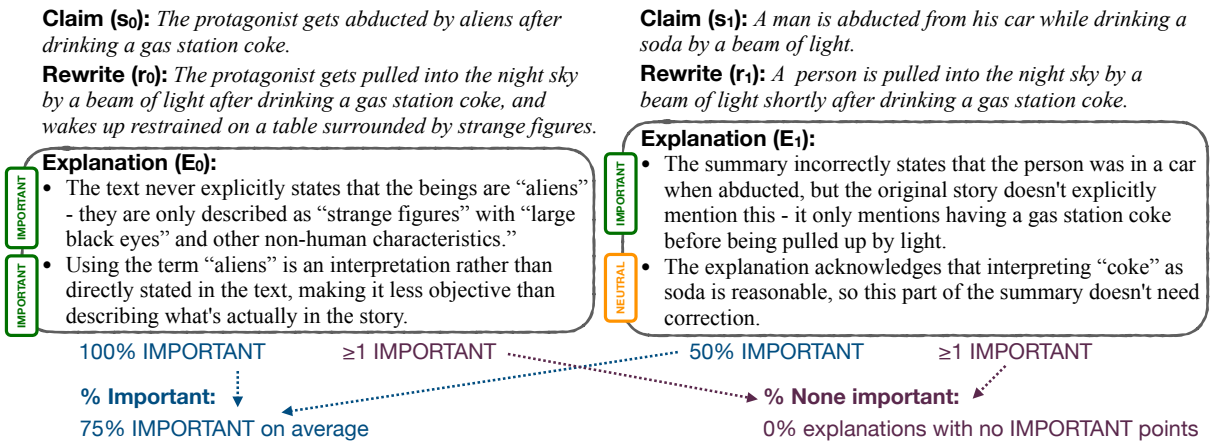


Figure 3: Examples of computing the metrics used for evaluating explanations, computing 75% for *% important* and 0% for *% none important*.

To compare different rewrite models, we use GPT-4o and Claude-3.5-Sonnet with a standard rewriting prompt targeting subjectivity and unfaithfulness. We additionally compare this prompt against other variants targeting only subjectivity or only unfaithfulness (see Appendix A.2 for all prompts).

5 Results

Using the models and methods discussed, we establish subjectivity as part of narrative summarization and identify why claims are subjective. We demonstrate rewriting effectively detects subjectivity and corrects it in alignment with human judgments.

5.1 Does subjectivity play a role in evaluation of this task?

We compare whether human-written summaries exhibit different levels of subjectivity than LLM-written summaries. As shown in Table 1, we find that all three annotators agree on 76% of the human-written claims compared to 73% of the LLM-generated claims in StorySumm. This is not a statistically significant difference (p: 0.23), indicating that **ambiguous claims are an inherent challenge in evaluating narrative summaries**, and are not specific to LLMs.

We compare whether claims generated to introduce or address the ambiguities we identify exhibit greater subjectivity or objectivity respectively. As shown in Table 1, 27% of synthetically generated subjective claims produce annotator agreement compared to 56% of synthetically generated objective claims. This difference is statistically significant (p : 2e-4), demonstrating **the ambiguities we identify lead to annotator disagreement**. Given that these claims are synthetically generated, we do not see perfect agreement or disagreement for either category. Many of the objective claims which still produce disagreement are due to annotator error, and the subjective claims with agreement are generally due to synthetic data generation introducing an easily identifiable inconsistency.

	% Agree
Summary writer:	
Human	75.93
LLM	72.74
Synthetic claim type:	
Objective	**55.56****
Subjective	26.56

Table 1: For each type of claim, we report the percent of those claims for which all three annotators agree on the faithfulness label.

5.2 Can rewrites detect subjectivity?

The ARM metric labels claims which get rewritten as subjective and labels them objective otherwise. We compare how well ARM detects subjectivity relative to the baselines detailed in Section 4.4 in Table 2. We find that finetuning a detection model on the synthetic dataset does not transfer effectively to the original claims. In zero-shot and few-shot settings, we see that Claude is overall a stronger model for detecting subjectivity than GPT-4. Self-consistency does not work well for Claude but has comparable performance to other methods for GPT-4. ARM is a strong method for detecting subjectivity across both models, and the strongest when looking at balanced accuracy. Overall, **ARM is a**

Method	Bal. acc.	F1-macro
Llama-3.1-8B finetuned	49.62	0.50
Claude-3.5 zero-shot	63.85	0.61
Claude-3.5 few-shot	66.21	**0.62**
Claude-3.5 self-consistency	50.42	0.50
Claude-3.5 ARM (ours)	**69.15**	0.53
GPT-4 zero-shot	58.04	0.57
GPT-4 few-shot	58.94	0.56
GPT-4 self-consistency	57.73	**0.59**
GPT-4 ARM (ours)	**63.55**	0.58

Table 2: Performance of rewrites with different models relative to baseline methods for subjectivity detection against our subjectivity labels on StorySumm. We report balanced accuracy and F1-macro scores.

stronger or comparable method to other baselines for binary evaluation while providing the added benefits of the rewrite itself and explanation. We report additional results with alternate prompts and against faithfulness labels next and observe that ARM is fairly consistent across different prompts.

5.3 Does the rewrite model and prompt affect results?

We assess how sensitive rewrites are to prompting method and model choice in Table 3. We evaluate detection against subjectivity labels, faithfulness labels, and claims that are subjective or unfaithful. We use the rewriting prompts shown in Appendix A.2. We see that Claude is a stronger model for rewriting in terms of binary classification of subjectivity or faithfulness. GPT-4 provides the benefit of making more minimal changes to claims as its rewrites have less than half the edit distance relative to Claude. Overall, **the LLM used for rewriting matters more than the prompt** with the exception of the subjectivity-targeted prompt for GPT-4. This setting produces the most rewrites by far which leads to its poor performance.

5.4 Do rewrites align with human judgments of faithfulness?

We take the subjectivity-focused rewrites using Claude (prompt in Appendix A.2) and test whether its edits align with human judgments and preferences. In Table 4, we see that rewrites exhibit statistically significant gains over original claims in annotator agreement and faithfulness. When human judges are asked whether they prefer the original claim versus the rewritten claim, they prefer the rewrites 77% of the time.

In Table 5, we report the quality of explanations. We find that on average 69% of explanation points are labeled as IMPORTANT by majority vote. 99% of explanations have no points labeled WRONG by majority vote. Finally, only 5% of explanations have no points considered important for an individual annotator. These percentages show that the vast majority of explanations and changes made in the rewrites are either important to fix or neutral. We note that the explanation parse often repeats the same point multiple times (see example explanations in Figure 3), which is why we consider both *% important* and *% wrong*.

These numbers indicate that **rewrites significantly improve claims in objectivity and faithfulness and using them in conjunction with their explanations is a meaningful evaluation signal**.

5.5 Qualitative analysis of ambiguities

We uncover specific reasons why claims are ambiguous and analyze whether there are specific types of ambiguities that are harder for rewrites to detect. Two of the authors of this paper perform inductive thematic analysis (Bowman et al., 2023) with adjudication to identify a taxonomy for why claims are subjective (see Appendix A.6 for more details). Using this taxonomy, they revisit each subjective claim and label it with an ambiguity code. The taxonomy identifies four ambiguity types:

1. **Wording**: A word or phrase is used in the summary which has overlapping meaning with the wording used in the story but also lends itself to other interpretations. Depending on someone's interpretation, the two meanings may not fully overlap.

2. **Detail**: A very minor detail is assumed in the summary which is not explicitly stated in the story. Many people find this to be a reasonable assumption and therefore faithful, while others view it as unfaithful.

3. **Causation**: The summary skips important causal details for an event. Some people find this to be a reasonable abbreviation while others feel it fundamentally changes one's understanding of what happened.

4. **Explicit**: The summary makes explicit details that are intentionally left ambiguous or only implied in the story. Some people like this interpretive jump while others feel it misrepresents the nature of the story.

Method	Subj.	Unfaith.	Subj. ∨ unfaith.	# Rewrites	Avg. edit dist.
Claude subj.	67.89	67.42	69.00	246	0.40
Claude both	**69.15**	68.05	**69.96**	341	0.61
Claude unfaith.	66.7	**68.21**	63.56	218	0.46
GPT-4 subj.	53.95	55.42	54.61	486	0.17
GPT-4 both	63.55	64.95	63.39	203	0.22
GPT-4 unfaith.	64.01	63.09	63.56	390	0.21

Table 3: We report the balanced accuracy scores for different prompts and models against the subjectivity labels (Subj.), faithfulness labels (Unfaith.) and a combination of both label sets that looks for claims which are subjective or unfaithful (Subj. ∨ Unfaith.). Subj. and unfaith. indicate the prompts targeting subjectivity and unfaithfulness respectively, and the "both" method indicates the prompt shown in Section 3.1 which targets both subjectivity and unfaithfulness. We also report the number of claims which are rewritten and the average edit distance of rewrites relative to the originals.

Metric	Original claims	Rewrites
Agreement	36.36	57.45*
Faithful	20.45	89.36**
Preferred	23.08	76.92**

Table 4: The percent of each claim type that meets each condition. Agreement requires all three annotators agree on the label. Faithful requires at least two of the annotators label the claim faithful. Preferred requires that claim variant was preferred over the other. ($p \leq .05 : *, p \leq .001 : **$)

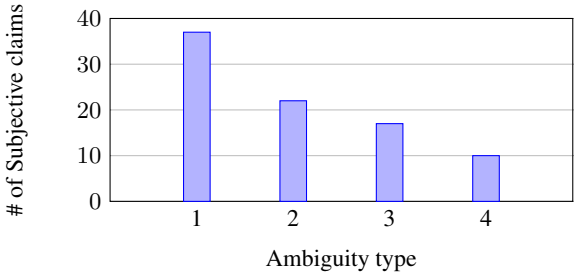


Figure 4: A breakdown of how many times each ambiguity type is labeled in the StorySumm LLM-generated summaries.

In Figure 4, we see the most common ambiguity type in the LLM-generated subjective claims in StorySumm is **Type 1**, meaning **many claims use vague or misleading wording**. This result is consistent with prior work (Subbiah et al., 2024b; Kim et al., 2024). **Type 4** is the smallest category, indicating that most ambiguities are unintentionally introduced by the summary writer (**Types 1-3**). We also see that on average, there does not seem to be one type of ambiguity that is detected substantially more or less by the LLM rewrite metrics. We compare ARM's recall across the different ambiguity types but do not find that one type of claim is missed more often than others (see Appendix Figure 7).

6 Related Work

Ambiguity in evaluation Ambiguity has been studied in natural language entailment (Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Jiang and de Marneffe, 2022), a related task to summary faithfulness evaluation. Other work has studied how to improve or manage annotator disagreement (Uma et al., 2022; Krishna et al., 2023; Min et al., 2023). Most similar to ours are Koupaee et al. (2025); Mishra et al. (2024); Ramprasad et al. (2024), which include various types of ambiguities in evaluating meeting summaries, real-world QA, and dialogue summaries respectively. Their taxonomies of ambiguities support our findings and by working with narratives, we incorporate the added challenge of intentional ambiguity in the source text.

Edits as feedback Recent work has explored the use of natural language feedback to improve model outputs in different settings, such as math reasoning (Madaan et al., 2023; Xu et al., 2024b) and summarization (Liu et al., 2023; Zhang et al., 2023). This feedback can be human feedback (Madaan et al., 2023) or automatic feedback from models (Gao et al., 2023; Ye et al., 2023; Wadhwa et al., 2024; Xu et al., 2024a). While most of these methods follow a critique-and-refine framework to improve models' outputs, we propose a rewriting-based evaluation method that leverages edits themselves as the feedback signal.

Evaluating narrative summaries Prior work has introduced datasets and methods for evaluating narrative summaries. Early work provided reference summaries from online study guide and TV episode synopsis websites to enable reference-

	Important		Wrong	
Method	% Important	% None important	% Wrong	% None wrong
Individual	84.11	5.29	7.52	87.83
Maj. vote	68.72	24.39	0.56	98.78

Table 5: Percentages for explanation annotations (metrics described in Figure 3).

based evaluation metrics (Ladhak et al., 2020; Kryscinski et al., 2022; Chen et al., 2022). As LLMs have trained on more and more online data, more recent work has evaluated LLMs as evaluators on recently published books or unpublished work (Chang et al., 2024; Kim et al., 2024; Subbiah et al., 2024b; Karpinska et al., 2024).

7 Conclusion

In this work, we expand the considerations for evaluating faithfulness in narrative summarization to include ambiguity and subjectivity. We propose using rewrites as automatic evaluation for summary claims and demonstrate their efectiveness on the StorySumm dataset. We release additional labels for StorySumm for claim subjectivity and ambiguity types. In the future, we hope this evaluation methodology can be tested on other challenging evaluation tasks in the humanities. We believe rewriting-based evaluation could also be used in RL training for improving reasoning about implicit meaning in the humanities.

Ethics Statement

There are no risks involved with this work as it explores summarization evaluation with publicly available data. We follow protocol approved by Columbia University IRB protocol AAAS4051 for the human annotation work in this study. We use AI assistants to answer questions about coding for this work, but do not use them for any part of research design or paper writing. The StorySumm dataset is available for research use and our use is in line with this purpose. One of the authors, Melanie Subbiah, holds an equity interest in OpenAI.

Limitations

The StorySumm dataset is relatively small which limits the generalization of our conclusions. However, we are still able to see statistically significant results with this dataset. Using human evaluation is a thorough and rich source of feedback but can limit reproducibility of results since different annotators will perform the task slightly differently. Finally, given that we study subjectivity in this work, there is inherent subjectivity involved in the task which can further limit the reproducibility of results. Despite this, we feel it is important to try to formalize and make progress on areas of evaluation that contain grey areas.

Acknowledgments

We would like to express our gratitude to the Upwork workers who contributed annotations for this work. Additionally, we would like to thank our reviewers for their thoughtful feedback. This work is supported by the funds provided by several organizations: the Columbia Amazon CAIT PhD Fellowship, Northrup Grumman, the National Science Foundation, DoD OUSD (R&E) under Cooperative Agreement PHY-2229929 (The NSF AI Institute for Artificial and Natural Intelligence), and Good Systems,[4] a UT Austin Grand Challenge to develop responsible AI technologies.

References

Robert Bowman, Camille Nadal, Kellie Morrissey, Anja Thieme, and Gavin Doherty. 2023. Using Thematic Analysis in Healthcare HCI at CHI: A Scoping Review. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. BooookScore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A Dataset for Abstractive Screenplay Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

[4] https://goodsystems.utexas.edu/

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. FELM: Benchmarking Factuality Evaluation of Large Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

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

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and Revising What Language Models Say, Using Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 Herd of Models. *arXiv*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating Reasons for Disagreement in Natural Language Inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One Thousand and One Pairs: A "novel" challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17048–17085, Miami, Florida, USA. Association for Computational Linguistics.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. FABLES: Evaluating faithfulness and content selection in book-length summarization. In *Conference on Language Modeling (COLM)*.

Mahnaz Koupaee, Jake W Vincent, Saab Mansour, Igor Shalyminov, Han He, Hwanjun Song, Raphael Shu, Jianfeng He, Yi Nian, Amy Wing-mei Wong, and 1 others. 2025. Faithful, Unfaithful or Ambiguous? Multi-Agent Debate with Initial Stance for Summary Evaluation. *arXiv*.

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. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Faisal Ladhak, Bryan Li, Yaser Al-Onaizan, and Kathleen McKeown. 2020. Exploring Content Selection in Summarization of Novel Chapters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5043–5054, Online. Association for Computational Linguistics.

Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed Hassan Awadallah. 2023. On Improving Summarization Factual Consistency from Natural Language Feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15144–15161, Toronto, Canada. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and

Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. In *First Conference on Language Modeling*.

Hidetsugu Nanba and Manabu Okumura. 2004. Comparison of Some Automatic and Manual Methods for Summary Evaluation Based on the Text Summarization Challenge 2. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What Can We Learn from Collective Human Opinions on Natural Language Inference Data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

OpenAI. 2024. GPT-4 Technical Report. *arXiv*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Sanjana Ramprasad, Elisa Ferracane, and Zachary Lipton. 2024. Analyzing LLM behavior in dialogue summarization: Unveiling circumstantial hallucination trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12549–12561, Bangkok, Thailand. Association for Computational Linguistics.

Melanie Subbiah, Faisal Ladhak, Akankshya Mishra, Griffin Thomas Adams, Lydia Chilton, and Kathleen McKeown. 2024a. STORYSUMM: Evaluating Faithfulness in Story Summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9988–10005, Miami, Florida, USA. Association for Computational Linguistics.

Melanie Subbiah, Sean Zhang, Lydia B. Chilton, and Kathleen McKeown. 2024b. Reading Subtext: Evaluating Large Language Models on Short Story Summarization with Writers. *Transactions of the Association for Computational Linguistics*, 12:1290–1310.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from Disagreement: A Survey. *J. Artif. Int. Res.*, 72:1385–1470.

Manya Wadhwa, Xinyu Zhao, Junyi Jessy Li, and Greg Durrett. 2024. Learning to Refine with Fine-Grained Natural Language Feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12281–12308, Miami, Florida, USA. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *International Conference on Learning Representations (ICLR)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024a. LLM-Refine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1429–1445, Mexico City, Mexico. Association for Computational Linguistics.

Yifan Xu, Xiao Liu, Xinghan Liu, Zhenyu Hou, Yueyan Li, Xiaohan Zhang, Zihan Wang, Aohan Zeng, Zhengxiao Du, Zhao Wenyi, Jie Tang, and Yuxiao Dong. 2024b. ChatGLM-Math: Improving Math Problem-Solving in Large Language Models with a Self-Critique Pipeline. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9733–9760, Miami, Florida, USA. Association for Computational Linguistics.

Seonghyeon Ye, Yongrae Jo, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, and Minjoon Seo. 2023. SelFee: Iterative Self-Revising LLM Empowered by Self-Feedback Generation. Blog post.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STaR: Bootstrapping Reasoning With Reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pages 15476–15488. Curran Associates, Inc.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. SummIt: Iterative Text Summarization via ChatGPT. In *Findings of the Association for Computational*

Linguistics: EMNLP 2023, pages 10644–10657, Singapore. Association for Computational Linguistics.

A Appendix

A.1 Synthetic data generation prompts

The following prompts are filled with a story and summary or story and claim pair in the %s fields.

A.2 Baseline and rewrite prompts

The following prompts are filled with a story, summary, and claim triple in the %s fields.

A.3 Explanation parsing prompt

Prompt for parsing an explanation into individual points

```
Summarize the key reasons described in
this explanation for why the summary
sentence needs to be rewritten. Group
together reasoning about the same detail.
Place each reason between <item></item>
tags.
```

A.4 Annotation format validation

We check for confounding factors in the task format for the human annotations of rewrites in several ways. First, we randomize whether the rewrite is shown to the annotator in the summary or as the "alternate". We find there is not a significant difference (p: .08) in which claim is preferred based on which position it is shown to annotators in (57.1% for in-summary vs. 42.9% for alternate). These numbers indicate annotators are not biased to prefer whichever claim is shown to them as the "original" vs. "alternate".

Additionally, we include three decoy explanations that do not make sense and annotators should reject to check that annotators are not just convinced by any explanation. The decoy explanations are marked as WRONG in 5/6 instances indicating that annotators are not just convinced by any explanation. We remove results on these decoy explanations from the results discussed in the paper.

A.5 Human annotation interfaces

Screenshots from the Upwork annotation interfaces are shown in Figures 5 and 6.

A.6 Additional ambiguity analysis details

For the inductive thematic analysis, the two authors involved write an explanation for the ambiguities involved in each subjective claim. They then discuss these explanations and arrive at four types of ambiguities. They go back and code each subjective claim with one of these types and discuss and adjudicate any disagreements. We observed agreement between the authors on 89% of the sentences (0.53 Cohen's Kappa) prior to adjudication.

They additionally label a small number of claims with a type 5 not discussed in the main text. Type 5 indicates the story is too confusing to determine its intent but the summary sentence itself is written clearly. We do not include type 5 claims in the analysis as they result from unintentional ambiguity introduced by the story writer which is beyond the scope of this paper.

While the writer of a story may intentionally use ambiguity in their story which leads to subjective viewpoints on its meaning, a summary writer should not intentionally introduce ambiguity. Types 1-3 capture types of ambiguities in summary claims that are unintentional on the part of the summary writer. Type 4 deals with intentional ambiguity by the author. We observe that type 4 ambiguities may be okay in summaries that should interpret the story as well as summarizing it, while types 1-3 are always undesirable as they obscure meaning.

In Figure 7, we show the recall of Claude and GPT-4 rewrite metrics averaged across the three different rewriting prompts for each ambiguity type. We do not observe a significant difference in recall for the different types.

Subjective → Objective Prompts

```
TYPE 1: Use the provided story to rewrite the provided claim to remove any ambiguous wording
from the claim which may require or demonstrate some interpretation. If the claim can't be
rewritten, give the original claim. Put the rewritten claim between <sentence> tags.\\Story:
%s\\Claim: %s

TYPE 2: Use the provided story to rewrite the provided claim to remove any minor assumptions
that the claim makes. If the claim can't be rewritten, given the original claim. Put the
rewritten claim between <sentence> tags.\\Story: %s\\Claim: %s

TYPE 3: Use the provided story to rewrite the provided claim to not skip causal details or
contain vague phrases that skip things. The provided claim is one of many summary claims, and
must fit into the context of the summary when rewritten. If the claim can't be rewritten, give
the original claim. Put the rewritten claim between <sentence> tags.\\Story: %s\\Summary:
%s\\Rewrite only Line %s from the summary.

TYPE 4: Use the provided story to rewrite the provided claim to not specify any implied or
ambiguous interpretations of the story as an explicit occurrence. If the claim can't be
rewritten, given the original claim. Put the rewritten claim between <sentence> tags.\\Story:
%s\\Claim: %s
```

Objective → Subjective Prompts

```
TYPE 1: Swap the wording in the claim in one or two places so it requires or demonstrates some
interpretation of the story. The claim should become difficult to evaluate with respect to the
story. You must rewrite the claim in some way. Put the rewritten claim between <sentence>
tags.\\Story: %s\\Claim: %s

TYPE 2: Use the provided story to add a minor detail to the claim that isn't explicitly stated
in the story. This detail must be a reasonable assumption to make from the story. The claim
should become difficult to evaluate with respect to the story. You must rewrite the claim in
some way. Put the rewritten claim between <sentence> tags.\\Story: %s\\Claim: %s

TYPE 3: Make the provided claim more vague about why things are happening by removing important
causal details. The provided claim is one of many summary claims and must fit into the context
of the summary when rewritten. The rewritten claim should become difficult to evaluate with
respect to the story, but should not be shorter in length than the original claim. You must
rewrite the claim in some way. Put the rewritten claim between <sentence> tags.\\Story:
%s\\Summary: %s\\Rewrite only Line %s from the summary.

TYPE 4: Rewrite the claim to include some interpretation of what characters are thinking or
feeling or what is happening in the story. State this definitively, rather than just as a
possibility. The provided claim is one of many summary claims and must fit into the context of
the summary when rewritten. The rewritten claim should become difficult to evaluate with
respect to the story. You must rewrite the claim in some way. Put the rewritten claim between
<sentence> tags.\\Story: %s\\Claim: %s\\Rewrite only Line %s from the summary.
```

Zero-shot prompt used as a baseline

```
SYSTEM: You are an expert summary evaluator, and you will be asked to evaluate claims in
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 can be objectively
evaluated for accuracy against the story or if there are claims that are subjective to
evaluate. An objective claim may be accurate or inaccurate but it should be clearly right or
wrong. A subjective claim introduces vague language, interpretation, or confusing wording such
that different people might interpret it in different ways.
USER: Story:
%s

Summary:
%s

Consider the following claim in the summary: %s
Is this claim objective to evaluate? You should answer Yes or No. Place your answer between
<answer></answer> tags.
```

Few-shot prompt used as a baseline

```
SYSTEM: You are an expert summary evaluator, and you will be asked to evaluate claims in
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 can be objectively
evaluated for accuracy against the story or if there are claims that are subjective to
evaluate. An objective claim may be accurate or inaccurate but it should be clearly right or
wrong. A subjective claim introduces vague language, interpretation, or confusing wording such
that different people might interpret it in different ways.
USER: Story:
Shelly and her dog were running down the street one afternoon when they came across an injured
squirrel. Shelly stopped to help the squirrel and Shelly's dog almost ate it. Shelly managed to
tuck it into her pocket and bring it home. Later she brought it to a vet and got some
recommendations on how to nurse it back to health. Within a couple weeks she released the
squirrel back into the wild.

Summary: The main character, Shelly, and her dog find an injured squirrel while out running.
The dog's prey drive is activated around the squirrel. Shelly tucks the little fluffy squirrel
into her pocket to bring home. She figures out how to nurse it back to health. She eventually
lets the squirrel go again to live a healthy life.

Consider the following claim in the summary: The main character, Shelly, and her dog find an
injured squirrel while out running.
Is this claim objective to evaluate? You should answer Yes or No. Place your answer between
<answer></answer> tags.
<answer>Yes</answer>

Consider the following claim in the summary: The dog's prey drive is activated around the
squirrel.
Is this claim objective to evaluate? You should answer Yes or No. Place your answer between
<answer></answer> tags.
<answer>No</answer>

Consider the following claim in the summary: Shelly tucks the little fluffy squirrel into her
pocket to bring home.
Is this claim objective to evaluate? You should answer Yes or No. Place your answer between
<answer></answer> tags.
<answer>No</answer>

Consider the following claim in the summary: She figures out how to nurse it back to health.
Is this claim objective to evaluate? You should answer Yes or No. Place your answer between
<answer></answer> tags.
<answer>No</answer>

Consider the following claim in the summary: She eventually lets the squirrel go again to live
a healthy life.
Is this claim objective to evaluate? You should answer Yes or No. Place your answer between
<answer></answer> tags.
<answer>No</answer>

Story:
%s

Summary:
%s

Consider the following claim in the summary: %s
Is this claim objective to evaluate? You should answer Yes or No. Place your answer between
<answer></answer> tags.
```

Self-consistency prompt used as a baseline

```
SYSTEM: You are an expert summary evaluator, and you will be asked to evaluate claims in
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. The details described in a consistent summary should not
misrepresent details of the story or make things up.
USER: Story:
%s

Summary:
%s

Consider the following claim in the summary: %s
Is all of the information in this claim consistent with the story? First reason about the
question before answering Yes or No. Your output should be in the following format:

Reasoning: Your reasoning about the answer to the question.

<answer>Your answer to the question (Yes or No)</answer>
```

Rewrite prompt used for subjectivity-focused rewrites

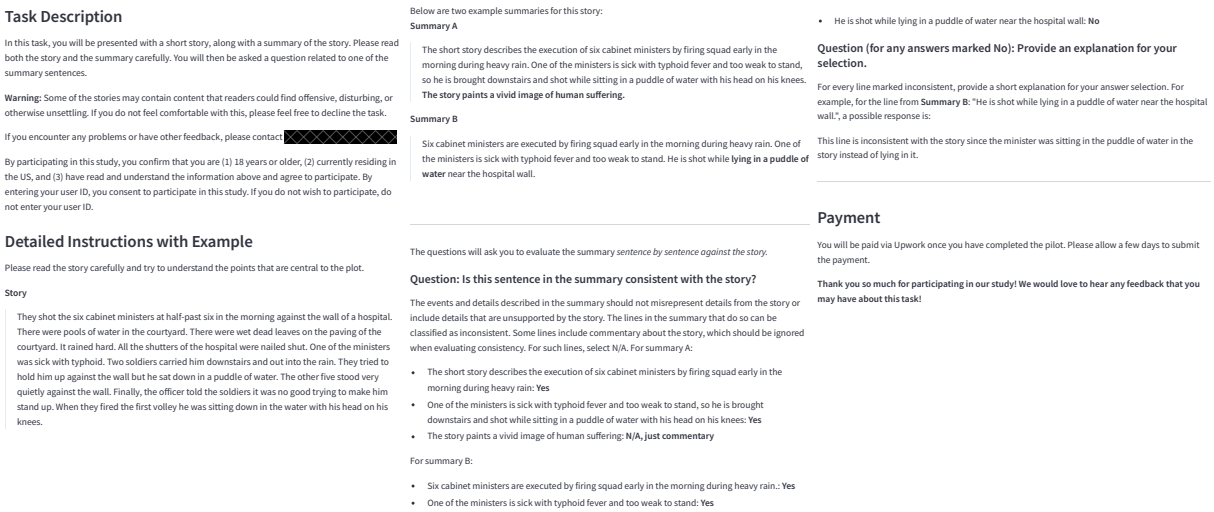
```
SYSTEM: You are an expert summary writer. You write and correct summaries so that they are
precise, clear and accurate representations of the story.
USER: Story:
%s

Summary:
%s

Rewrite any elements of the following sentence from the summary that might be subjective. You
should make minimal edits to fix the sentence. If the sentence is objective as written or is
just interpretation of the story, restate the original sentence. Give your final sentence in
<answer></answer> tags.

Sentence: %s
```

Rewrite prompt used for unfaithfulness-focused rewrites

```
SYSTEM: You are an expert summary writer. You write and correct summaries so that they are
precise, clear and accurate representations of the story.
USER: Story:
%s

Summary:
%s

Rewrite any elements of the following sentence from the summary that are inconsistent with the
story. You should make minimal edits to fix the sentence. If the sentence is accurate as
written or is just interpretation of the story, restate the original sentence. Give your final
sentence in <answer></answer> tags.

Sentence: %s
```

Rewrite prompt targeting both subjectivity and faithfulness (ARM)

```
SYSTEM: You are an expert summary writer. You write and correct summaries so that they are
precise, clear and accurate representations of the story.

USER: Read the story and summary carefully, then decide whether the specified summary sentence
should be rewritten.\\
A summary sentence should be rewritten according to the following principles:
1.) If the sentence is inconsistent with the story, it should be rewritten.
2.) If the sentence contains subjective interpretation or ambiguous wording, it should be
rewritten. In particular, rewrite cases of:
    - assuming a minor detail that is reasonable but not explicitly stated in the story
    - skipping important causal details
    - making explicit conclusions which are left ambiguous in the story
    - using words or phrases that can be interpreted differently from the story wording
3.) When rewriting a sentence, any edits should be minimal to fix the problem.
4.) If the sentence is just commentary on the story, then it should not be rewritten.
5.) If the sentence is accurate and clear, it should not be rewritten.\\
Story:\\%s\\\\Summary:\\%s\\
Rewrite the following summary sentence, placing your rewrite between <answer></answer> tags.
If the sentence does not need to be rewritten, simply repeat the original wording between
<answer></answer> tags.\\
Sentence: %s
```



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 then be asked a question related to one of the summary sentences.

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 entering your user ID, you consent to participate in this study. If you do not wish to participate, do not enter your user ID.

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. He is shot while **lying in a puddle of water** near the hospital wall.

The questions will ask you to evaluate the summary *sentence by sentence against the story.*

Question: Is this sentence in the summary consistent with 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:

- Six cabinet ministers are executed 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: **Yes**
- He is shot while lying in a puddle of water near the hospital wall: **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**: "He is shot while lying in a puddle of water near the hospital wall.", 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.

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: A variant of these instructions are used for each of the human annotation tasks on Upwork.

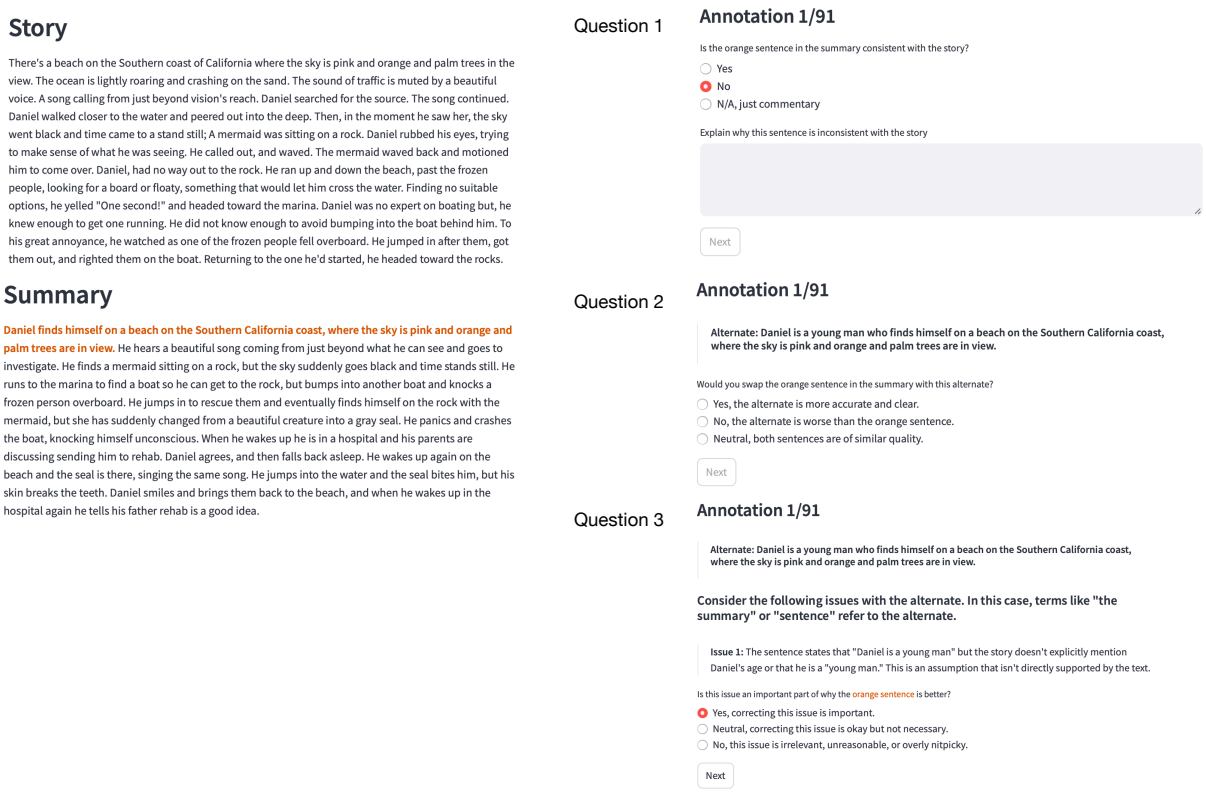


Figure 6: The task format for human annotations on Upwork. Human studies which only look at annotator agreement on faithfulness labels only ask Question 1. The studies aligning rewrites with annotator judgments additionally ask questions 2 and 3.

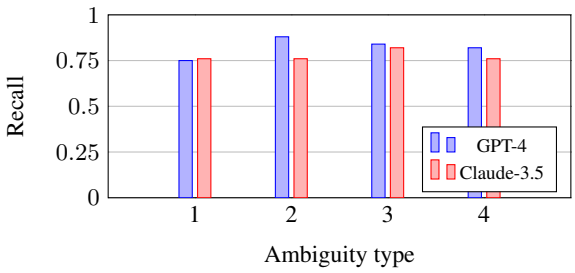


Figure 7: We show the recall of GPT-4 rewrite metrics vs. Claude-3.5 rewrite metrics at detecting different ambiguity types. For each model, we average the results across the three rewrite prompts tested in Table 3.