

FABLES: Evaluating faithfulness and content selection in book-length summarization

Yekyung Kim^U, Yapei Chang^U, Marzena Karpinska^U, Aparna Garimella^A,
Varun Manjunatha^A, Kyle Lo^{AI}, Tanya Goyal^P, Mohit Iyyer^U
UMass Amherst^U, Adobe^A, Allen Institute for AI^{AI}, Princeton^P
{yekyungkim, yapeichang, mkarpinska, miyyer}@umass.edu
{garimell, vmanjuna}@adobe.com, kylel@allenai.org, tanyagoyal@princeton.edu

Abstract

While long-context large language models (LLMs) can technically summarize book-length documents (> 100K tokens), the length and complexity of the documents have so far prohibited evaluations of input-dependent aspects like faithfulness. In this paper, we conduct the first large-scale human evaluation of faithfulness and content selection on LLM-generated summaries of fictional books. Our study mitigates the issue of data contamination by focusing on summaries of books published in 2023 or 2024, and we hire annotators who have fully read each book prior to the annotation task to minimize cost and cognitive burden. We collect FABLES, a dataset of annotations on 3,158 claims made in LLM-generated summaries of 26 books, at a cost of \$5.2K USD, which allows us to rank LLM summarizers based on faithfulness: CLAUDE-3-OPUS significantly outperforms all closed-source LLMs, while the open-source MIXTRAL is on par with GPT-3.5-TURBO. An analysis of the annotations reveals that most unfaithful claims relate to events and character states, and they generally require indirect reasoning over the narrative to invalidate. While LLM-based auto-raters have proven reliable for factuality and coherence in other settings, we implement several LLM raters of faithfulness and find that none correlates strongly with human annotations, especially with regard to detecting unfaithful claims. Our experiments suggest that detecting unfaithful claims is an important future direction not only for summarization evaluation but also as a testbed for long-context understanding. Finally, we move beyond faithfulness by exploring content selection errors in book-length summarization: we develop a typology of omission errors related to crucial narrative elements and also identify a systematic over-emphasis on events occurring towards the end of the book. We release FABLES to spur further research on the evaluation of book-length summarization.

 <https://github.com/mungg/FABLES>

1 Introduction

Advances in long-context language models have sparked interest in summarizing book-length documents (>100K tokens). Despite the importance of faithfulness and content relevance for summary quality, recent work in this regime focuses only on input-agnostic aspects like coherence (Chang et al., 2023b). This is due to the length and complexity of the input documents: hiring human annotators to read and understand them is expensive and time-consuming. Our work fills this gap by presenting the first large-scale human evaluation of faithfulness and other content selection errors in book-length summarization.

We mitigate challenges associated with document complexity by hiring workers who have already read a book published in 2023 or 2024 (to avoid data contamination) for enjoyment prior to beginning the annotation task. We produce summaries for these books via five

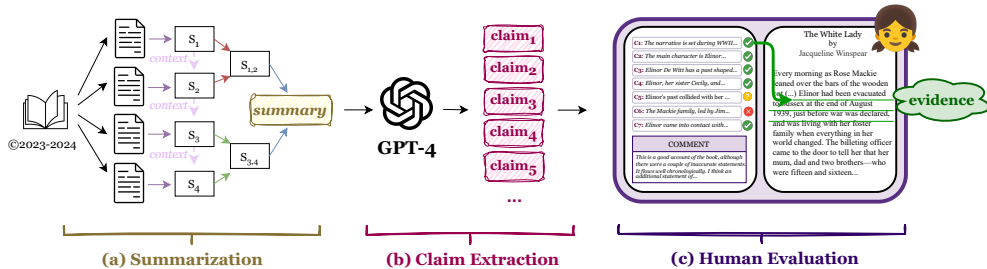


Figure 1: Our pipeline for collecting faithfulness annotations in book-length summarization (FABLES). First, (a) we generate summaries through hierarchical merging. Next, (b) we prompt GPT-4 to extract decontextualized claims. Finally, (c) we conduct a human evaluation of these claims, requiring annotators to validate each claim and provide their reasoning and evidence for the assigned label.

configurations of the hierarchical summarization methodology described in Chang et al. (2023b), each of which varies the base LLM and chunk size. Following prior work on faithfulness and factuality evaluation, such as LongEval (Krishna et al., 2023) and FactScore (Min et al., 2023), we decompose each summary into a list of claims which are then individually verified against the input document.

In total, our FABLES dataset (Faithfulness Annotations for Book-Length Summarization) contains 3,158 claim-level annotations of faithfulness across 26 narrative texts, along with evidence for each claim in the form of quotations from the book as well as free-form comments at both the claim and summary level (Figure 1).¹ Overall, we observe that CLAUDE-3-OPUS is the most faithful book-length summarizer by a significant margin, followed by GPT-4-TURBO. Beyond ranking LLMs, our annotations also shed light on the following previously unexplored questions:

What kinds of faithfulness errors do LLM summarizers make? (§3) A qualitative analysis of FABLES reveals that the majority of claims marked as unfaithful are related to *events* or *states* of characters and relationships. Furthermore, most of these claims can only be invalidated via multi-hop reasoning over the evidence, highlighting the task’s complexity and its difference from existing fact-verification settings (Min et al., 2023; Kamoi et al., 2023).

Can faithfulness be evaluated automatically? (§4) Collecting human annotations on 26 books cost us \$5.2K, demonstrating the difficulty of scaling our workflow to new domains and datasets. We thus implement multiple LLM-based raters of faithfulness, following prior work such as BoookScore (Chang et al., 2023b) and FactScore (Min et al., 2023) that achieve high correlation with human judgments. However, all of our metric configurations struggle to reliably identify unfaithful claims. Our best-performing method operates similarly to “needle-in-the-haystack”-style evaluations (Kamradt, 2023; Gemini Team, 2024) by feeding as much of the book as possible into a long-context LLM along with a single claim to verify. We promote this claim-level verification task as both important for book-length summarization evaluation as well as a challenging benchmark for long-context understanding.

What other errors, beyond faithfulness, do LLM summarizers make? (§5) By coding all of the summary-level free-form comments in FABLES, we find that annotators frequently point out *omissions* of critical information. We develop the first taxonomy of omission errors in book-length summarization and observe that key events, details, and themes are frequently omitted by all LLMs. We also observe other content selection errors: for example, even our strongest summarizers, CLAUDE-3-OPUS and GPT-4-TURBO, over-emphasize content towards the end of books to the detriment of the beginning.

All prompts used in this paper can be found in §B.

¹While we cannot release the book text due to copyright restrictions, we publicly release all summaries and annotations.

	Books 📖			Annotations 📝		
	Documents (<i>n</i> =26)	Summaries (<i>n</i> =130)	Claims (<i>n</i> =3,158)	Reasons (<i>n</i> =1,513)	Evidence (<i>n</i> =3,051)	Comments (<i>n</i> =130)
Mean	121,467	594.3	19.8	37.6	194.7	155
St. dev.	35,836	119.5	6.4	33.4	218.5	148.4
Max	243,965	798	58	281	2435	823
Min	49,762	172	6	2	5	6

Table 1: Number of tokens across books and FABLES annotations; based on tiktoken (<https://github.com/openai/tiktoken>) tokenizer.

2 Collecting human annotations

In this section, we describe our pipeline for collecting FABLES, which consists of human annotations of both faithfulness and overall quality of LLM-generated book summaries.

Collecting a corpus of newly-published fictional books: It is infeasible, both in terms of cost and time, to ask annotators to read long books ($\geq 100K$ tokens) for the sole purpose of annotating LLM-generated summaries. While we can remove this burden by choosing famous books that many people have already read, such as those in BookSum (Kryscinski et al., 2022), LLMs have also likely seen these books and their summaries during pretraining (Chang et al., 2023a), which can skew the evaluation of generated claims. Instead, we use an annotator-driven workflow to sidestep these issues. We recruit a pool of annotators via Upwork² who self-report having read one or more English books published in 2023 or 2024. Our final annotator pool consists of 14 native English speakers, and we purchase electronic copies of 26 books listed by them.³ The mean length of books in our dataset is 121K tokens (see Table 1 for statistics).

Prompting LLMs to generate book summaries: To summarize book-length documents, we adopt the hierarchical merging strategy from (Chang et al., 2023b); see Figure 1 for an illustration of the method. We use GPT-3.5-TURBO, GPT-4, GPT-4-TURBO (OpenAI, 2023), MIXTRAL (Jiang et al., 2024), and CLAUDE-3-OPUS (Anthropic, 2023) as the backbone models.⁴

Decomposing summaries into claims: Following prior works on evaluating long-form summary faithfulness (Krishna et al., 2023; Min et al., 2023; Wei et al., 2024), we decompose our summaries into *atomic claims* to enable fine-grained annotation. We prompt an LLM (GPT-4) with two primary instructions: (1) each atomic claim must be fully understandable on its own without requiring additional context from the summary (e.g., resolved pronouns), and (2) whenever possible, each claim should be situated within its relevant temporal, locational, and causal context. Human validation by the authors of a random sample of 100 extracted claims demonstrated 100% precision (i.e., each claim can be traced to the summary without any extra or incorrect information). See Figure 2 for example of summary and its extracted claims; see §B for exact prompt and §G.4 for recall analysis.

Collecting human annotations: The Upwork annotators were tasked with two primary objectives:

- **Claim-level:** Assess the faithfulness of claims extracted from model-generated summaries of their assigned book(s). Annotators reviewed claims made about their selected book(s) and determined their accuracy by choosing one of four options for each decomposed claim: (a) *Faithful* – accurate reflection of the narrative, (b)

²<https://www.upwork.com>

³We convert epubs to text files preserving all information including front and back matter.

⁴All summaries were generated in February 2024 using the following checkpoints: gpt-3.5-turbo, gpt-4-0613, gpt-4-0125-preview, Mixtral-8x7B-Instruct-v0.1, and claude-3-opus-20240229.

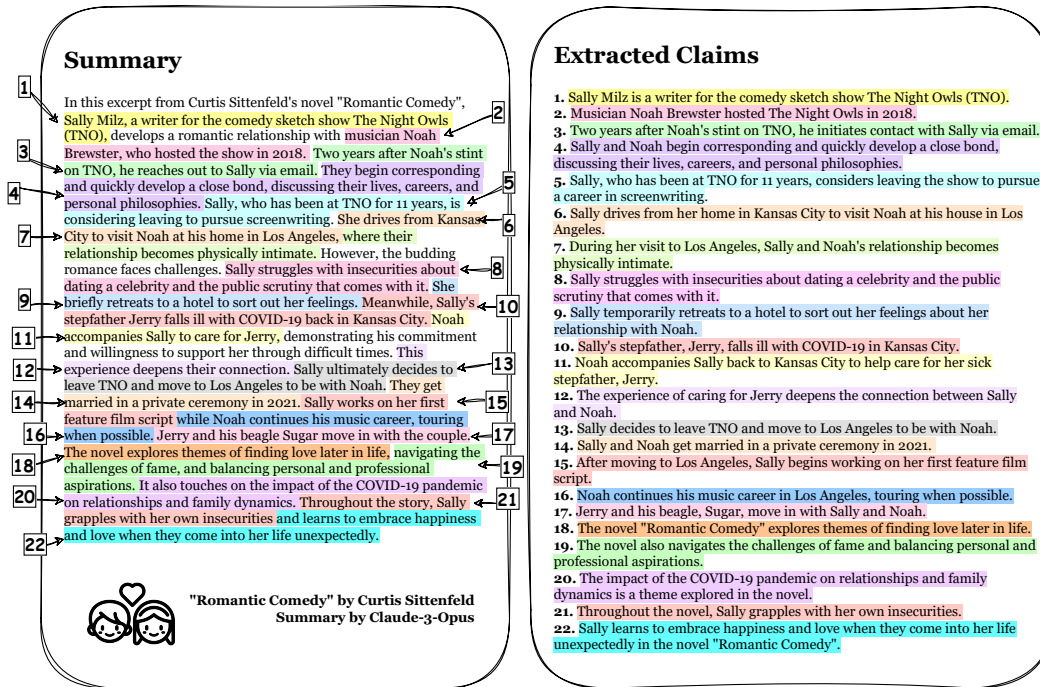


Figure 2: Example summary generated by CLAUDE-3-OPUS and claims extracted by GPT-4.

Unfaithful – misrepresentation of the narrative, (c) *Partial Support* – partially corroborated by the narrative, or (d) *Can't verify* – indeterminable. They provided free-form textual justifications to support their selections, including *evidence* in the form of quotations from the book when relevant.

- **Summary-level:** Provide free-form summary-level comments on the overall quality of the summaries. Annotators critiqued the claim set as a whole, identifying omissions, inaccuracies, disproportionate emphasis on trivial plot points, or other concerns.

The annotators used a customized interface⁵ which provided them full access to the book text for reference. Each annotator was assigned to annotate all five LLM-generated summaries for their assigned book, which were presented in a randomized order. Annotators received \$200 for this task, which took ~11 hours to complete (STD=6.34). In total, FABLES contains 3,158 annotated claims from 130 summaries across 26 books at a cost of \$5.2K USD. We assess the quality of our dataset using inter-annotator agreement and self-consistency metrics. More details can be found in §C.

3 Developing a taxonomy of faithfulness errors in FABLES

In this section, we present results from our statistical and qualitative analysis of the 3,158 claim-level faithfulness annotations in FABLES, which include both free-form comments and citation evidence to support or refute these claims.⁶ Broadly, we observe that CLAUDE-3-OPUS is the most faithful LLM summarizer, with 90% of its claims rated as faithful, followed by GPT-4 and GPT-4-TURBO at 78%, GPT-3.5-TURBO at 72%, and MIXTRAL at 70% (Table 2).

⁵Refer to §C for the screenshots of the interface and the exact wording of the task.

⁶For 107 claims, the annotators were unable to cite evidence either in favor or against the claim.

⁷Note that GPT-4-TURBO, MIXTRAL, and GPT-4 are capable of processing sequences of up to 16K, 32K, and 8K tokens respectively. In this study we opted to use a shorter context window to maintain consistency with the experimental setup described in Chang et al. (2023b).

Model	Chunk size	Avg # Claims _{STD}	Faithful	Unfaithful	Partial support	Can't verify
GPT-3.5-TURBO	2,048	23.23 _{3.29}	71.52	11.26	13.08	4.14
MIXTRAL	2,048	27.50 _{4.61}	68.67	11.47	17.2	2.66
GPT-4 ⁷	2,048	26.23 _{3.54}	78.15	4.55	15.98	1.32
GPT-4-TURBO	100,000	21.65 _{2.35}	77.62	7.64	12.08	2.66
CLAUDE-3-OPUS	180,000	22.85 _{4.87}	90.89	2.1	6.65	0.35

Table 2: Percentage of claims extracted from LLM-generated summaries rated by humans as *faithful*, *unfaithful*, *partial support* or *can't verify*. Chunk size denotes the token count per chunk used for summarization across models; we also include the mean and standard deviation of claim counts in generated summaries. Please note that the percentage of each label for CLAUDE-3-OPUS is calculated from 24 out of 26 books. The model was unable to merge summaries for two books due to content discrepancies.

Analysis of unfaithful claims: To further study the nature of unfaithful claims, we characterize all 205 such claims along two dimensions: CLAIM TYPE and REASONING TYPE (see Table 3 for taxonomy and frequency counts).⁸ Most unfaithful claims are about specific *events* (31.5%) or the *state* of some character or relationship (38.6%). Crucially, a majority of unfaithful claims require *indirect reasoning* to refute (50.2%), making this a more challenging faithfulness evaluation setting compared to prior work (Kamoi et al., 2023; Min et al., 2023). More details on this analysis can be found in §E.



LABEL	FREQ	EXAMPLE CLAIM	REASON FOR REJECTION
 Claim Type			
State	38.6	Roman Kitt is <i>under pressure</i> from his father to join the family business.	Roman is not under pressure, his father bribes people so he gets his dream job.
Event	31.5	Patricia Liu, Athena's mother, discovers that June has sold Athena's manuscript and <i>confronts</i> her.	Patricia never confronts June.
Cause/effect	11.2	Lilly's abusive ex-boyfriend, Alan Bushy, becomes a suspect <i>due to the meticulous nature of the murders</i> .	He becomes a suspect because he was abusive to Lilly.
High-level	11.2	The narrative is <i>non-linear</i> and features flashbacks, switches between alternate worlds or viewpoints, and present-day conversations between Sally and Danny.	The narrative is largely linear.
Introspection	7.5	Juniper Song encounters Athena Liu at a literary event, triggering <i>feelings of admiration, intimidation, and self-doubt</i> .	No part of the book shows that Juniper admires Athena.
 Reasoning Type			
Indirect	50.2	Dean stirs up <i>tensions</i> with palace server Fawn.	This encounter is merely Rennick being protective of Amelia, tension can't be inferred from the book.
Direct	36.8	The narrative reveals that Maggie had a <i>brief affair</i> with a doctor named Danny in Bangkok while she was being followed by unknown entities.	The book directly states that they are married, so it's not a brief affair.
Subjective	7.2	Forest is torn between his desire to protect Iris and <i>confronting his past actions</i> .	I don't think Forest makes any real effort to confront his past actions
Extra info	5.7	The book "Wildfire" is <i>the first</i> in the Icebreaker series.	It's not stated in the book, but this is actually the second in the series.

Table 3: Taxonomy of faithfulness errors with respect to claim type and reasoning type in FABLES. For each label, we report its *frequency* and provide an *example* claim-reason pair. More examples and the general labeling scheme can be found in Table 15.

⁸There are actually 247 annotations with unfaithful claims, but for this analysis we leave out 42 unclear ones that require further clarification from the annotators. Note that since the claims sometimes contain multiple subclaims, we allow each annotation to have multiple labels.

4 Challenges with automatic faithfulness evaluation

While insightful, human annotation of faithfulness in book-length summarization is simply not scalable: our annotations cost \$40 USD per summary for a total cost of \$5.2K USD, which is prohibitively expensive for usage during model development and with bigger corpora. In this section, inspired by methods such as FactScore (Min et al., 2023) and BoookScore (Chang et al., 2023b), we develop LLM-powered automatic raters of faithfulness that operate at the claim level. However, our best method, which relies on prompting CLAUDE-3-OPUS with the entire book to verify a single claim, is expensive and unreliable at detecting unfaithful claims in FABLES, suggesting important directions for future work.

Automatic raters of faithfulness: We implement our automatic raters by prompting an LLM in a zero-shot manner to verify a single claim given evidence from the book (Table 13), where the evidence can be one of the following:

- **None:** As a lower bound, we evaluate the faithfulness of claims without any evidence from the book.
- **👤 Human evidence:** We can also use human-annotated evidence from FABLES obtained via the pipeline described in §2. This evidence is always related to the claim, but it often takes the form of short, highly-contextual spans that may or may not be sufficient to support claim verification.
- **📖 BM25 retrieval:** We employ BM25 (Robertson et al., 1995) to retrieve passages from the book using the claim as a query. We concatenate the k most relevant passages to use as evidence for our evaluation prompt. We set $k = 5$ and chunk passages up to 256 tokens. See §G.4 for performance changes when varying passage length.
- **📖 Entire book (EB):** Retrieval is especially challenging in our setting due to the complexity of both the query and document. Intuitively, long-context LLMs can bypass explicit retrieval by simply fitting the entire book into the context as evidence. This setting resembles “needle-in-the-haystack” evaluations of prior work (Kamradt, 2023; Levy et al., 2024), except that it tests a much deeper understanding of the input document.

Dataset for experiments: Due to budget constraints associated with the “entire book” setting, we select seven books, each shorter than 125K tokens, to evaluate the performance of our auto-rater configurations. This results in 723 total claims, 69 of which are marked as *Unfaithful* and 654 as *Faithful* by our human annotators. Note that we do not consider partially supported or unverifiable claims in our experiments due to the increased subjectivity associated with these labels. Detailed information regarding this dataset and experiment costs can be found in §G.

Results: We evaluate the performance of each auto-rater configuration by comparing its predictions to the ground-truth labels (*Faithful* and *Unfaithful*) from our human annotations. Due to the class imbalance, we report separate F1 scores for each label, split across claims generated by different LLMs, in Table 5.⁹ As a sanity check, the “no evidence” setting performs extremely poorly; more interestingly, human evidence underperforms both retrieval and the entire book setting, suggesting that the LLM requires more context to judge claim validity. The best performing auto-rater is CLAUDE-3-OPUS in the entire book setting, which significantly outperforms both GPT-4-TURBO in the same setting as well as BM25.

Conclusion: Despite it having the best performance in Table 5, CLAUDE-3-OPUS ultimately performs too poorly to be a reliable auto-rater (58.2 F1 when classifying *Unfaithful* claims). This comes as a surprise as this pattern of decompose-then-verify has been shown to correlate with human judgments in other settings, like Min et al. (2023). Manual analysis of

⁹We note that scores for *Unfaithful* claims on a per-model level should be taken with a grain of salt due to the small sample size, particularly for CLAUDE-3-OPUS summaries.

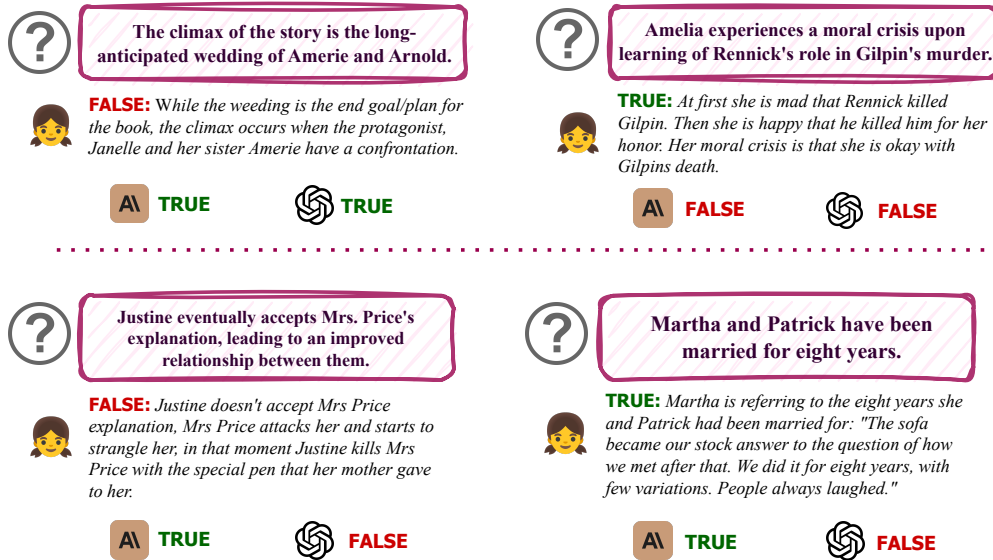


Figure 3: Examples of mistakes in label prediction made by CLAUDE-3-OPUS and GPT-4-TURBO accompanied by annotator labels and reasoning. More examples can be found in Figure 11.

the errors reveals that CLAUDE-3-OPUS struggles most with claims involving non-narrative information (23.1%), assessments often based on common sense reasoning (20.5%), and character confusions (12.8%), which often require a deep understanding of the entire book; see confusion matrix in Figure 3 and more details in §32. Qualitatively, we can also gauge from annotator comments (Table 4) the difficulty of this claim verification task as evidence may be difficult to localize (in “needle-in-the-haystack” manner) and require full document reasoning.

COMMENTS	
🗨️	<i>The hardest part was that some of the claims were very general about the text, such as describing overall character arcs, which made it hard to find specific textual support.</i>
🗨️	<i>The most difficult part for me was how general some of the sentences were. Because the material was so broad, I felt that I could use 20 or 30 quotations. For example, this book is about many stories of a private investigator in Africa (not exactly what it said, but close). I could recite the entire book.</i>
🗨️	<i>The most difficult part for me was finding supporting quotations for claims that were more abstract (e.g. "The book grapples with the scars of colonialism."). Although I was able to tell right away whether the claim was true or false, based on my own reading, it was at times difficult to find a specific quotation that best proved the claim. The themes were more often implicit in the text, rather than explicit.</i>
🗨️	<i>The most difficult part was to give citations for claims about writing style and intent. The reason was that these claims are usually based on the book as a whole, so an accurate citation would be the whole book.</i>

Table 4: Annotator comments highlighting the challenges in evidence retrieval.

Discussion: It is generally agreed that benchmarking the faithfulness of LLM-generated text is important. However, recent efforts have primarily focused on verifying entity-centric facts (Min et al., 2023). Our work, and others (Zhu et al., 2023; Tang et al., 2024; Mishra et al., 2024), show that these do not provide coverage over all types of LLM errors, especially in more challenging settings like book summarization. Moreover, the retrieve-then-verify framework that forms the backbone of most past evaluation techniques (Bohnet et al., 2022;

Gao et al., 2023) completely fails for our significantly more challenging setting. Given this evidence, we call for broadening the scope of error types and task settings (including our current task of book-length summarization) considered by current faithfulness evaluation benchmarks.

Summarized by	No-Context		Human Evidence		BM25		Entire Book		Entire Book	
	GPT-4-TURBO		GPT-4-TURBO		GPT-4-TURBO		GPT-4-TURBO		CLAUDE-3-OPUS	
	<i>Faithful</i>	<i>Unfaithful</i>	<i>Faithful</i>	<i>Unfaithful</i>	<i>Faithful</i>	<i>Unfaithful</i>	<i>Faithful</i>	<i>Unfaithful</i>	<i>Faithful</i>	<i>Unfaithful</i>
GPT-3.5-TURBO	0.396	0.248	0.686	0.369	0.801	0.373	0.887	0.357	0.929	0.619
MIXTRAL	0.248	0.178	0.760	0.361	0.807	0.312	0.946	0.440	0.962	0.645
GPT-4	0.337	0.146	0.657	0.229	0.739	0.162	0.909	0.230	0.959	0.600
GPT-4-TURBO	0.261	0.217	0.680	0.264	0.794	0.241	0.918	0.109	0.945	0.600
CLAUDE-3-OPUS	0.242	0.018	0.510	0.022	0.692	0.000	0.962	0.000	0.971	0.000
Overall	0.305	0.167	0.675	0.259	0.779	0.249	0.932	0.386	0.955	0.582

Table 5: F1 scores for *Faithful* and *Unfaithful* label across models with evaluators on 7 books. The best results of each label are in bold. Entire Book refers to the entire book method evaluating faithfulness from large (125k) chunks using either GPT-4-TURBO or CLAUDE-3-OPUS.

5 Beyond faithfulness: content selection errors in book summarization

As book-length summarization is still a nascent area, research into other error types beyond coherence (Chang et al., 2023b) and faithfulness (§3) is still lacking. In this section, we perform qualitative coding over all 130 free-form, summary-level comments from FABLES and present a taxonomy of content selection errors (e.g., omissions) that may prove more difficult to detect than faithfulness.¹⁰

General issues with LLM-generated summaries: Table 6 summarizes the percentage of summaries affected by specific issues as per annotators’ comments.¹¹ Our analysis shows that every LLM makes chronological errors, though these were less pronounced in models with extended context (CLAUDE-3-OPUS and GPT-4-TURBO). All models were also criticized for omitting important information, with CLAUDE-3-OPUS being the least affected (52%), compared to 80.8% and 84.6% for GPT-4-TURBO and GPT-3.5-TURBO, respectively. The least faithful models, GPT-3.5-TURBO and MIXTRAL, also both have a tendency to generate overly generic statements (38.5%). Finally, we look also at cases where the summary was explicitly praised for being good or comprehensive. CLAUDE-3-OPUS received the most praise (48% and 54% respectively), while GPT-3.5-TURBO received the least (11.5% and 15.4% respectively).

Exploring omission errors: As mentioned above, omission of key information plagues all LLM summarizers. To better understand the nature of the omission errors identified by our annotators, we categorize them into the following categories: *characters, events, details, relationships, themes*.¹³ Figure 4 shows a heatmap of omission errors broken down by model. A large proportion of summaries (33.3% to 65.4%) lack mentions of key events, creating gaps in the overall narrative, and we also note omissions of significant details about the characters, events, or objects (16.7% to 38.5%). Furthermore, GPT-4-TURBO and MIXTRAL have a tendency to entirely omit mentions of crucial characters (23.1%).

¹⁰Details of the annotation scheme used to analyze the comments are in Table 21 in the §F

¹¹In two cases, CLAUDE-3-OPUS refused to merge two summaries, as they were affected by the extra information available in the front and back matter and did not constitute a logical story. We excluded these cases from this analysis.

¹²Percentage of summaries where the annotator expressed specific concerns about the factuality of the entire claim set. See §D for the percentage of affected claims per summary. In short, most summaries contained factual inaccuracies with only five summaries receiving 100% of *Faithful* labels (indicating complete factual accuracy).

¹³Since annotators did not identify every specific omission, we focused on a binary classification: whether a summary was impacted by a given omission type, rather than counting the total number of omissions by type. See Table 22 in the §F for more details.

	CLAUDE-3-OPUS	GPT-4-TURBO	GPT-4	GPT-3.5-TURBO	MIXTRAL
Chronology	33.3	36.0	46.2	50.0	61.5
Omissions	52.0	80.8	65.4	84.6	65.4
Factuality ¹²	58.3	69.2	80.8	69.2	84.6
Overemphasis	20.8	34.6	19.2	30.8	46.2
Underemphasis	12.5	23.1	19.2	38.5	34.6
Vague/Generic	0.0	23.1	3.9	38.5	38.5
Repetitive	0.0	11.5	0.0	7.7	3.9
Data-Influenced	0.0	23.1	19.2	19.2	34.6
Comprehensive	54.2	30.8	38.5	15.4	34.6
Well-done	50.0	23.1	26.9	11.5	15.4

Table 6: Percentage of summaries per model identified with specific issues, **based on annotator general comments** (not the claim-wise faithfulness ratings). The upper row, colored in **purple**, outlines categories of critique, whereas the lower row, in **green**, indicates categories where the models received compliments.

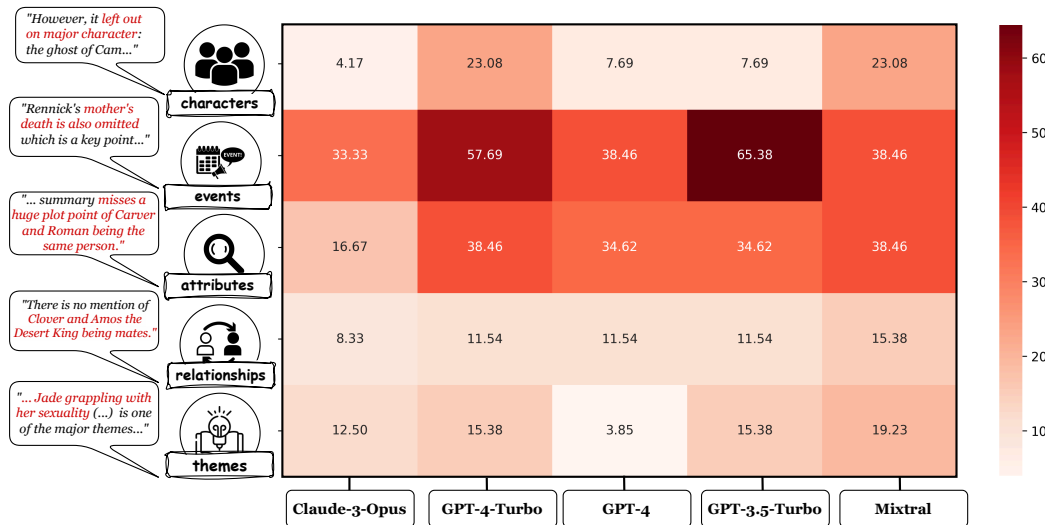


Figure 4: Percentage of summaries flagged by the annotators for one of five omission errors, *characters*, *events*, *attributes*, *relationships*, and *themes*, by model.

Long-context models overemphasize book endings: One interesting observation is that CLAUDE-3-OPUS and GPT-4-TURBO, which both have chunk sizes $\geq 100K$, tend to place more emphasis on the endings of the books to the detriment of the beginning. Since these models were often provided with the entire book context during prompting, this suggests a potential issue in processing long inputs (Kamradt, 2023; Levy et al., 2024). This phenomenon is especially prominent with CLAUDE-3-OPUS, where at least 20% of the generated summaries exhibit an overemphasis on the book’s ending, compared to 7.7% for GPT-4-TURBO (see examples in Table 25 in the §F). We also note that the back matter of many books (e.g., author’s biography, dedications, etc.) often unduly impacts all LLMs during the summarization process. We observe conflation between characters in the narrative and names in the back matter, as well as entirely hallucinated narratives; CLAUDE-3-OPUS is the only model seemingly unaffected by this additional information; see §F for more analysis on this phenomenon.

6 Related work

Narrative summarization: Our paper builds on prior work in narrative summarization, including short stories (Wang et al., 2022; Subbiah et al., 2024), poetry (Mahbub et al., 2023), screenplays (Chen et al., 2022), among others. Wu et al. (2021) demonstrated how an LLM can overcome long context to summarize books, like those in the BookSum (Kryscinski et al., 2022) dataset. Closely related to our work is Chang et al. (2023b), but while they focus on evaluating summary coherence (which requires only judging the model generation), we address faithfulness and content selection (which requires relating model generations back to the long source inputs).

Faithfulness and content selection in summarization: Our paper builds on prior work in evaluating hallucination and inconsistency in summarization (Maynez et al., 2020; Kryscinski et al., 2020; Ladhak, 2024) which are even challenging for humans (Daumé & Marcu, 2005). Pagnoni et al. (2021) introduce the FRANK dataset, where they use human annotations of generated summaries to produce a taxonomy of factual errors based on linguistic analysis, resembling the work of Goyal & Durrett (2020) and Goyal & Durrett (2021). Closest to our work, Krishna et al. (2023) perform human evaluation of faithfulness on summaries of short stories, whereas we study book-length inputs. Our exploration of omission errors is rooted in prior research on content selection (Nenkova & Passonneau, 2004; Gillick & Liu, 2010; Ladhak et al., 2020).

Claim verification for evaluating summaries: Our paper also relates to prior work on claim verification, where claims are verified given reference to some knowledge source (Thorne et al., 2018; Wadden et al., 2020; Schuster et al., 2021). Min et al. (2023) propose FActScore, an LLM-based metric of factual precision in biography generation, which was expanded upon in SAFE (Wei et al., 2024). Manakul et al. (2023) propose SelfCheckGPT, which uses LLMs to evaluate the faithfulness of GPT-3 generated texts on a dataset of Wikipedia-style passages about people.

7 Conclusion

We present FABLES, the first large-scale human evaluation of faithfulness and content selection in book-length summarization. By recruiting annotators who had read recently-published books for enjoyment, we collect 3,158 claim-level faithfulness annotations from LLM-generated summaries of 26 narratives. This allows us to rank LLM summarizers based on faithfulness, revealing that CLAUDE-3-OPUS is the most faithful book-length summarizer, followed by GPT-4-TURBO. Next, we experiment with using LLMs for automatic claim verification. Our results expose the limitations of both retrieval and long-context understanding: LLM auto-raters cannot reliably detect *unfaithful* claims, even when prompted with the full book text. Our analysis shows that unfaithful claims primarily pertain to states and events, often necessitating reasoning over extended contexts, which makes them complicated to detect for both humans and machines. Finally, we move beyond faithfulness to explore and characterize common content selection errors such as omissions of key events, attributes, and characters, as well as the over-emphasis of content from the end of the book.

Our work on FABLES suggests several promising directions for future work. With better auto-raters of faithfulness, we can perform fine-tuning or preference tuning on long-context language models by using the auto-raters as a scorer (Tian et al., 2023), which could improve their summarization capabilities by reducing hallucination (Cao et al., 2021). Additionally, FABLES can be used as a dataset and protocol to meaningfully benchmark future work on novel long-context language model architectures and training objectives.

Ethical considerations

All annotators consented to the use and publication of their annotations. The dataset excludes copyrighted texts, containing only annotations done on model-generated sum-

mary claims. Additionally, we ensured annotators received fair compensation for their contributions.

Acknowledgments

We extend special gratitude to the Upwork annotators for their hard work, and to members from the UMass NLP lab for their feedback. This project was partially supported by awards IIS-2202506 and IIS-2312949 from the National Science Foundation (NSF) as well as an award from Adobe.

References

- Anthropic. Model Card: Claude 3. Technical report, Anthropic, 2023. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. Accessed: 2024-03-25.
- Bernd Bohnet, Vinh Q Tran, Pat Verga, Roei Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*, 2022.
- Mengyao Cao, Yue Dong, and Jackie Chi Kit Cheung. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Annual Meeting of the Association for Computational Linguistics*, 2021. URL <https://api.semanticscholar.org/CorpusID:244909449>.
- Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7312–7327, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.453. URL <https://aclanthology.org/2023.emnlp-main.453>.
- Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. Boookscore: A systematic exploration of book-length summarization in the era of llms. *ArXiv*, abs/2310.00785, 2023b. URL <https://arxiv.org/abs/2310.00785>.
- Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. Summscreen: A dataset for abstractive screenplay summarization, 2022.
- Hal Daumé and D. Marcu. Bayesian summarization at duc and a suggestion for extrinsic evaluation. In *Document understanding conference, 2005* // 2005.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*, 2023.
- Google Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.
- Dan Gillick and Yang Liu. Non-expert evaluation of summarization systems is risky. In Chris Callison-Burch and Mark Dredze (eds.), *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pp. 148–151, Los Angeles, June 2010. Association for Computational Linguistics. URL <https://aclanthology.org/W10-0722>.
- Tanya Goyal and Greg Durrett. Evaluating factuality in generation with dependency-level entailment. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3592–3603, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.322. URL <https://aclanthology.org/2020.findings-emnlp.322>.

- Tanya Goyal and Greg Durrett. Annotating and modeling fine-grained factuality in summarization, 2021.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. Mixtral of experts, 2024.
- Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. Wice: Real-world entailment for claims in wikipedia, 2023.
- Greg Kamradt. Needle in a haystack. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2023.
- Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1650–1669, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.121. URL <https://aclanthology.org/2023.eacl-main.121>.
- Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9332–9346, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.750. URL <https://aclanthology.org/2020.emnlp-main.750>.
- Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. BOOKSUM: A collection of datasets for long-form narrative summarization. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 6536–6558, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.488. URL <https://aclanthology.org/2022.findings-emnlp.488>.
- Faisal Ladhak. Faithfulness in abstractive summarization: Progress and challenges, 2024.
- Faisal Ladhak, Bryan Li, Yaser Al-Onaizan, and Kathleen McKeown. Exploring content selection in summarization of novel chapters. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5043–5054, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.453. URL <https://aclanthology.org/2020.acl-main.453>.
- Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models, 2024.
- Ridwan Mahbub, Ifrad Khan, Samiha Anuva, Md Shihab Shahriar, Md Tahmid Rahman Laskar, and Sabbir Ahmed. Unveiling the essence of poetry: Introducing a comprehensive dataset and benchmark for poem summarization. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14878–14886, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.920. URL <https://aclanthology.org/2023.emnlp-main.920>.
- Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *EMNLP*, 2023.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter,

- and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL <https://aclanthology.org/2020.acl-main.173>.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 12076–12100. Association for Computational Linguistics, 2023. URL <https://aclanthology.org/2023.emnlp-main.741>.
- Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. Fine-grained hallucination detection and editing for language models. *arXiv preprint arXiv:2401.06855*, 2024.
- Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 145–152, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. URL <https://aclanthology.org/N04-1019>.
- OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL <https://doi.org/10.48550/arXiv.2303.08774>.
- Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4812–4829, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.383. URL <https://aclanthology.org/2021.naacl-main.383>.
- Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, Mike Gattford, et al. Okapi at trec-3. *NIST Special Publication SP*, 109:109, 1995.
- Tal Schuster, Adam Fisch, and Regina Barzilay. Get your vitamin C! robust fact verification with contrastive evidence. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 624–643, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.52. URL <https://aclanthology.org/2021.naacl-main.52>.
- Melanie Subbiah, Sean Zhang, Lydia B. Chilton, and Kathleen McKeown. Reading subtext: Evaluating large language models on short story summarization with writers, 2024.
- Liyang 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. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. *arXiv preprint arXiv:2402.13249*, 2024.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL <https://aclanthology.org/N18-1074>.

- Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. Fine-tuning language models for factuality. *ArXiv*, abs/2311.08401, 2023. URL <https://api.semanticscholar.org/CorpusID:265158181>.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL <https://aclanthology.org/2020.emnlp-main.609>.
- Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. SQuALITY: Building a long-document summarization dataset the hard way. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1139–1156, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.75. URL <https://aclanthology.org/2022.emnlp-main.75>.
- Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models, 2024.
- Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback, 2021.
- Rongxin Zhu, Jianzhong Qi, and Jey Han Lau. Annotating and detecting fine-grained factual errors for dialogue summarization. *arXiv preprint arXiv:2305.16548*, 2023.

A Dataset

In this section, we include further details about FABLES. We list all the books used for summarization in Table 7, along with details about the authors, genre, length, publication date, and variety of English. We also detail the data preprocessing process in §A.1.

A.1 Data Preprocessing

Preprocessing books: In order to obtain the summaries via hierarchical merging, we first purchased books from amazon.com in epub format and converted them into text files, retaining all information intact (i.e., *without* removing front and back matter). We then used the Huggingface GPT-2 tokenizer¹⁴ to divide the books into chunks fitting the models’ context window. During our chunking step, we checked for punctuation marks to ensure that all chunks end with a complete sentence. This approach sometimes resulted in chunks being shorter than the specified size, leading to the final chunks of some books consisting only of brief snippets with meta information, which could influence the summaries. Ideally, a robust model would distinguish between supplementary information and the main storyline to produce a coherent summary. However, we observed that some models were influenced by this extra information, leading them to fabricate aspects of the story.

Generating summaries: To summarize book-length documents, we adopt the hierarchical merging strategy which Chang et al. (2023b) found to outperform competing approaches in terms of summary coherence. We employ zero-shot prompting to summarize each chunk independently. Next, we form pairs of adjacent chunk-level summaries and again use zero-shot prompting to merge each pair, incorporating added context from previously-generated merged summaries to ensure coherence and continuity (see Figure 1a). We generate five summaries for each book in this fashion using GPT-3.5-TURBO, GPT-4, GPT-4-TURBO (OpenAI, 2023), MIXTRAL (Jiang et al., 2024), and CLAUDE-3-OPUS (Anthropic, 2023). All summaries were generated in February 2024 using the following checkpoints: gpt-3.5-turbo,

¹⁴https://huggingface.co/docs/transformers/en/model_doc/gpt2

gpt-4, gpt-4-turbo-preview, Mixtral-8x7B-Instruct-v0.1, and claude-3-opus-20240229. We use publicly-released code, prompts, and hyperparameters from [Chang et al. \(2023b\)](#) for summary generation. We further prompt GPT-4 model to extract decontextualized claims from the summaries. Examples of summaries along with extracted claims can be found in [Table 8](#), [Table 9](#), [Table 10](#), [Table 11](#), and [Table 12](#).

TITLE	AUTHOR	GENDER	GENRE	LENGTH	PUBLICATION	LANG
<i>A Haunting on the Hill</i>	Elizabeth Hand	F	horror, Gothic	117,577	Oct 3, 2023	AmE
<i>Agency for Scandal</i>	Laura Wood	F	historical fiction, mystery, romance	116,809	Jan 5, 2023	BrE
<i>Divine Rivals</i>	Rebecca Ross	F	fantasy, romance, young adult	137,616	Apr 4, 2023	AmE
<i>Fairytale of New York</i>	Zoë Folbigg	F	romance	134,369	Aug 28, 2023	BrE
<i>Flawless</i>	Elsie Silver	F	romance	119,580	Jun 24, 2022	CanE
<i>Fourth Wing</i>	Rebecca Yarros	F	fantasy	243,965	May 2, 2023	AmE
<i>Modern Divination</i>	Isabel Agajanian	F	urban fantasy	167,568	Jan 30, 2023	AmE
<i>Only for the Week</i>	Natasha Bisho	F	African American romance	87,056	May 11, 2023	AmE
<i>Pet</i>	Catherine Chidgey	F	thriller, mystery	124,679	July 13, 2023	NZE
<i>Romantic Comedy</i>	Curtis Sittenfeld	F	romance	116,560	Apr 4, 2023	AmE
<i>Same Time Next Year</i>	Tessa Bailey	F	romance	49,762	Dec 1, 2023	AmE
<i>She is a Haunting</i>	Trang Thanh Tran	F	romance	106,659	Feb 28, 2023	AmE
<i>Six Scorched Roses</i>	Carissa Broadbent	F	fantasy romance	54,481	Mar 21, 2023	AmE
<i>Sorrow and Bliss</i>	Meg Mason	F	mental health	112,468	Sep 2, 2020	AusE
<i>The Atonement Murders</i>	Jenifer Ruff	F	mystery, thriller	105,493	Apr 14, 2023	AmE
<i>The Guest</i>	Emma Cline	F	thriller	89,977	May 16, 2023	AmE
<i>The Marriage Act</i>	John Marrs	M	thriller, mystery, dystopia	135,901	Jan 19, 2023	BrE
<i>The Spy Coast</i>	Tess Gerritsen	F	mystery, thriller	128,918	Nov 1, 2023	AmE
<i>The Wager</i>	David Grann	M	nonfiction, historical fiction, mystery	156,022	Apr 18, 2023	AmE
<i>The White Lady</i>	Jacqueline Winspear	F	historical fiction, mystery	126,051	Mar 21, 2023	BrE
<i>This Impossible Brightness</i>	Jessica Bryant Klagmann	F	mystery, fantasy	119,115	Feb 1, 2024	AmE
<i>Viciously Yours</i>	Jamie Applegate Hunter	F	fantasy, romance	86,774	Jan 23, 2024	AmE
<i>Weyward</i>	Emilia Hart	F	historical fiction, fantasy	128,728	Feb 2, 2023	AusE
<i>Wildfire</i>	Hannah Grace	F	romance, sport	140,060	Oct 3, 2023	BrE
<i>Yellowface</i>	R.F. Kuang	F	thriller	114,346	May 25, 2023	AmE
<i>You, Again</i>	Kate Goldbeck	F	romance	137,624	Sep 12, 2023	AmE

Table 7: Details of the 26 books used for summaries. Length of each book is provided in tokens as computed with tiktoken.

B Prompts

In this section, we included all prompts used for our experiments: (1) *claim extraction*, (2) *automatic evaluation* in [Table 13](#).

C Human Annotations

In this section, we present details of our annotation task. [Figure 5](#) displays the instructions provided to annotators for evaluating faithfulness. [Figure 6](#) illustrates the interface used by annotators for this task. The list of claims is displayed on the left side of the screen, with each claim on a separate line. Content of the book is presented on the right side. Annotators can navigate the book’s content using the scroll function and perform keyword searches to locate relevant information. When annotators hover over a claim, it becomes highlighted, and clicking on it triggers a popup window to appear (see [Figure 7](#)). Given that completing the annotation process takes a considerable amount of time (approximately 1.5h-2.5h), we have implemented a feature that allows annotators to save their work at any point during the annotation process. Upon completing the annotations, the annotator is required to

CLAUDE-3-OPUS	
Summary	<p><i>Iris Winnow, an aspiring writer, works at the Oath Gazette newspaper in the city of Oath, competing with her rival Roman Kitt for a columnist position. Iris's brother Forest is fighting in a war between the gods Enva and Dacre. She hasn't heard from him in months and decides to become a war correspondent to find him. Iris and Roman develop feelings for each other through letters they exchange via enchanted typewriters, not realizing they are writing to each other. Roman, engaged to Elinor Little through an arranged marriage by his father, breaks off the engagement and follows Iris to the war front. In the town of Avalon Bluff near the front lines, Iris stays with Marisol, who runs a bed and breakfast, and befriends fellow correspondent Attie. Roman arrives and he and Iris fall in love, getting married just before an attack on the town by Dacre's forces. During the attack, Iris is saved by a disguised soldier who she later realizes is her brother Forest. Forest, who was supposed to have died from his wounds but was healed by Dacre, deserted and has been fighting for the enemy. In the chaos, Iris is separated from Roman, who is badly wounded. Iris and Forest return to Oath, where Iris learns the full story of Forest's service to Dacre. She discovers a letter from Roman in her room, confirming he knew her identity while writing to her. Meanwhile, the god Dacre finds the dying Roman on the battlefield and, sensing his strong spirit, decides to take him to heal his wounds and have him fight for his army. The story ends with Dacre turning his sights towards finding Enva in the east.</i></p>
Extracted Claims	<ol style="list-style-type: none"> 1. Iris Winnow works at the Oath Gazette newspaper in the city of Oath. 2. Iris Winnow is competing with Roman Kitt for a columnist position at the Oath Gazette. 3. Iris Winnow's brother, Forest, is involved in a war between the gods Enva and Dacre. 4. Iris Winnow decides to become a war correspondent after not hearing from her brother, Forest, for months. 5. Iris Winnow and Roman Kitt develop feelings for each other through letters they exchange via enchanted typewriters. 6. Iris Winnow and Roman Kitt do not initially realize they are writing letters to each other. 7. Roman Kitt was engaged to Elinor Little through an arranged marriage by his father. 8. Roman Kitt breaks off his engagement to Elinor Little and follows Iris Winnow to the war front. 9. Iris Winnow stays with Marisol at a bed and breakfast in the town of Avalon Bluff near the front lines. 10. Iris Winnow befriends fellow correspondent Attie in Avalon Bluff. 11. Roman Kitt arrives at Avalon Bluff and he and Iris Winnow fall in love. 12. Roman Kitt and Iris Winnow get married just before an attack on Avalon Bluff by Dacre's forces. 13. During the attack on Avalon Bluff, Iris Winnow is saved by a soldier who she later realizes is her brother, Forest. 14. Forest was supposed to have died from his wounds but was healed by Dacre. 15. Forest deserted from his original side and has been fighting for Dacre. 16. Iris Winnow is separated from Roman Kitt, who is badly wounded during the attack on Avalon Bluff. 17. Iris Winnow and Forest return to Oath after the attack. 18. Iris Winnow learns the full story of Forest's service to Dacre once they return to Oath. 19. Iris Winnow finds a letter from Roman Kitt in her room, which confirms he knew her identity while writing to her. 20. Dacre finds the dying Roman Kitt on the battlefield and decides to take him to heal his wounds. 21. Dacre plans for the healed Roman Kitt to fight for his army. 22. The story ends with Dacre planning to find Enva in the east.

Table 8: Example of a summary produced by CLAUDE-3-OPUS along with the extracted set of claims for "Divine Rivals," a novel by Rebecca Ross. Examples by the other models can be found in Table 9, Table 10, Table 11 and Table 12.

provide a comment on the overall quality of the summary claims by clicking on *general comments* (see Figure 8).

How do annotators perceive the task? Annotators highlighted several challenges in assessing the summaries, particularly when dealing with broad claims about themes rather than specific plot points, making it difficult to find relevant supporting evidence within the text. Abstract concepts, like emotions or thematic claims, posed significant obstacles, with some annotators struggling to locate quotations that precisely supported or refuted these

claims. They also pointed out the difficulty of evaluating claims that were only partially true, which required more detailed support (see Table 4 for actual comments).

I. Task Overview

Your job is to validate the factuality of claims made within an AI-generated book summary. You will be given a list of claims derived from a book summary along with the text of the book itself. Your task is to **decide whether or not each claim is supported by the book, and provide evidence from the book to justify your decisions. You should also comment broadly on major omissions, salience issues, and errors with the chronology of events within the list of claims.** To finish the text, you must (1) annotate every claim for factuality and (2) write a broad comment about the provided list of claims.

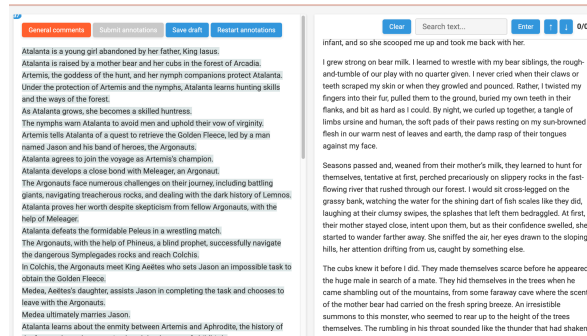
Make sure you have closely read the entire book prior to starting the annotation process. While we do provide the text of the book along with a simple search functionality, our expectation is that you will not need to re-read the book to assess the factuality of most claims.

II. Accessing the interface

1. Go to the link you gave via Upwork message.
2. Look for the file labeled "upwork-[number 1 to 5]-[book_name].html."
3. You will be given 5 different claims from the same book. Let's start annotating them!

III. Annotation Instructions

You will be working with the following interface:



The list of claims is on the left (one on each line), and the book's content is on right. Your annotation activity will take place in the space provided on the left.

Figure 5: Instructions for annotation task described in §2

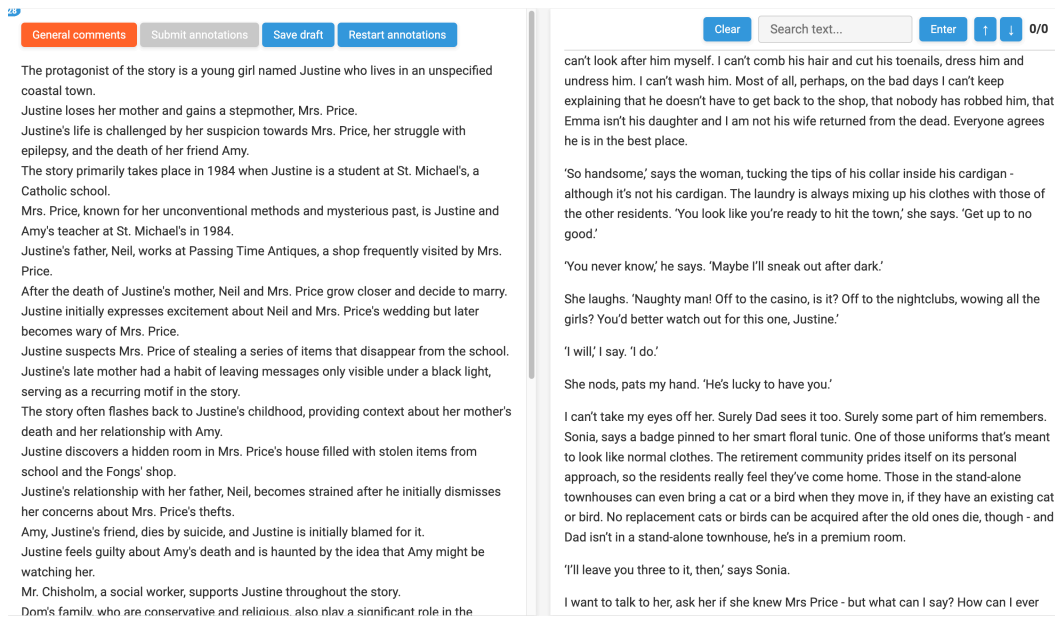


Figure 6: Screenshot of the interface for the annotation task described in §2.

General comments Submit annotations Save draft Restart annotations

"A Girl Called Samson" is a historical fiction novel set in the early 1800s.
The novel follows the life of Deborah Samson, a young girl who becomes orphaned and is sent to live with Widow Thatcher.
Deborah Samson is later indentured to a farmer until the age of eighteen.
This indenture sparks a rebellion within Deborah Samson to become a soldier and fight for her freedom.
Deborah Samson lives with the Thomas family as a servant.
Deborah forms a close bond with Jeremiah, the youngest of the Thomas family.

[7] As the American Revolution unfolds, Deborah becomes intrigued by the idea of independence and self-governance.

Is the claim faithful to the book? Choose...

Provide quotations from the book to justify your choice. If the claim is **faithful** to the book, provide quotations that support the claim. Otherwise, provide quotations that **contradict** the claim (or describe the contradiction in the comment section). In cases where a claim is **inapplicable or only partially supported**, provide either quotations or comments. Multiple quotations can be presented.

Quotation (copy text from the book here) Save

Comment (Optional) Remove

Add evidence

Deborah forms close bonds with fellow soldiers, such as Phineas.
She also deals with mutinies and conflicts within the army.

Figure 7: Pop-up window showing the interface where the annotators have to select the faithfulness label supplemented by free-form reasoning and evidence extracted from the book.

General comments Submit annotations Save draft Restart annotations

After reading the summary, provide free-form comments about **omissions** (e.g., important events or thematic elements) that should have been included in the summary, **salience** issues with claims that do not feel unimportant enough to be included in the summary, **chronology errors** with the order in which claims are presented (if they do not reflect the chronology of events in the book), and the overall **factuality** of the claims. For the latter, if you identify any factually incorrect claims, provide corrected versions of those claims here. Also please comment on any other issues with the summary that you were not able to explain in the claim annotation task.

Write your comments here Save

Figure 8: Pop-up window prompting the annotator to provide a free-form comment on the quality of summary claims highlighting *omissions*, *salience*, *chronology*, and *factuality* issues.

Quality of Annotations We perform two additional analysis experiments that demonstrate the high quality of our dataset: (1) self-consistency of annotations (i.e., how often a single annotator assigns the same label to claims with the same semantic content generated by different models), and (2) inter-annotator agreement on a subset of claims where we had access to another annotator who also read the book.

- **Inter-annotator agreement:** For two books in our dataset, we hired an additional annotator who had also read them to provide overlapping annotations. This resulted in 115 claims with overlapping annotations, allowing us to evaluate the agreement rate between the original and new annotators. The new annotator is 91.30%, with Cohen’s Kappa of 0.621 ($p < .0001$), indicating substantial agreement. Unfortunately, annotating the entire dataset with multiple annotators is unfeasible due to the difficulty and high cost of finding multiple individuals who have read the same book. Each annotation costs approximately \$200 to \$250 per book and requires around 10 hours of work.
- **Self-consistency:** For each book, an annotator analyzed five summaries, each generated by a different model. To assess self-consistency (intra-annotator agreement), we randomly selected five books and compared the annotations made on the first and last summaries (as per annotation order) for claims with the same semantic content. For example, "Aurora suffers emotional discomfort due to her father’s disinterest and her parents’ failed marriage" and "Aurora struggles with her father’s lack of attention and affection" are semantically equivalent claims from summaries of *Wildfire* generated by GPT-4 and Claude3, respectively. By comparing the first and last summaries, we evaluated the annotators’ consistency in handling claims after significant time intervals, during which they annotated three additional summaries. Consistency in labels for similar claims across these two summaries would indicate stable judgment and suggest that labels were not arbitrarily assigned. Out of 127 claims examined in the first summary, 46 had semantically equivalent claims in the last summary, and we found that all 46 of these claims were consistently labeled.

D Results of Human Evaluation

This section provides details on the number of *Unfaithful* and *Partially Supported* claims per summary. Figure 9 presents the percentage of problematic claims (either *Unfaithful* or *Partially Supported*) identified within each model’s summaries. Notably, only four (4) out of 130 summaries were rated 100% *Faithful* (two by GPT-3.5-TURBO, one by GPT-4-TURBO, and one by MIXTRAL). The remaining summaries varied in accuracy, with some containing up to 66.67% incorrect or partially incorrect claims.

E Analysis of Faithfulness Annotations

In this section, we provide additional details on our analysis of faithfulness annotations involving unfaithful claims. Refer to Table 15 for our general labeling scheme and examples for each category. Table 17 shows the reasoning type distribution for each claim type.

Evidence coverage and reasoning-claim relationship To investigate the quality of evidence provided by annotators, we analyze the coverage of evidence with respect to the annotators’ reasoning. In addition, we also analyze the relationship between the claim and the annotators’ reasoning. Results are summarized in Table 16. 51.6% of the time, annotators provide some evidence to justify every component of their reasoning (i.e., *complete coverage*). In 56% of *partial coverage* (i.e., some part of the reasoning does not have corresponding evidence) cases and all cases of *N/A coverage* (i.e., no evidence is provided at all), the missing evidence is due to the annotator’s inability to find any relevant information that either supports or refutes the claim. Qualitatively, for all matched reasoning-evidence pairs, we find that the evidence often does not provide enough context that would allow someone



Figure 9: Percentage of claims rated *Unfaithful* or *Partially Supported* across models, analyzed by book. **Only four (4) out of 130 summaries were 100% Faithful.** In two cases, CLAUDE-3-OPUS declined to merge two summaries due to significant content discrepancies (“Same Time Next Year” and “The Guest”).

who has not read the book to determine the faithfulness of the claim. As a result of decontextualization, claims always refer to people by name, but evidence often use pronouns instead. The annotator would need to quote a much larger chunk from the book in order for the evidence to include names as well. An even trickier case is that when dealing with high-level claims like "X is the protagonist of the story" or "The themes of the story are X, Y, and Z," one needs knowledge of the entire book, but citing the entire book as evidence is trivial. If annotators were to collect self-contained and sufficient evidence for every claim, the task would become significantly more challenging, sometimes even impossible. This difficulty with evidence gathering sheds light on why automatic evaluation does not work so well for this task.

Model-wise analysis We report model-wise results on reasoning type and reasoning-claim relationship in [Table 18](#) and [Table 19](#).

F Comment Analysis

In this section, we provide additional details regarding our analysis of the comments provided by annotators on the summary claims. [Table 20](#) features examples of such comments. These comments were further annotated based on the criteria outlined in [Table 21](#) and [Table 22](#). The distribution of errors is depicted in [Figure 10](#) and [Table 23](#).

[Table 24](#) displays examples where the models' generation was influenced by information in the front and back matter. [Table 25](#) highlights comments indicating that models may sometimes overly focus on the latter parts of the stories. Lastly, [Table 4](#) shares annotators' feedback on the annotation task.

Impact of front and back matter on the summary quality Books frequently contain additional information beyond the main narrative, including the author's biography, table of contents, dedications, and more, positioned at the beginning or the end of the book. Ideally, models should exclude this extraneous content, focusing solely on summarizing the core story. However, we have noted that models are sometimes unduly influenced by these elements, which can dominate a significant part of the summary and occasionally compromise its accuracy. Overall, between 19.23% (GPT-3.5-TURBO and GPT-4) and 34.62% (MIXTRAL) of summaries were affected by such content, either through focusing on this information,¹⁵ confusing story characters with names found in the front and/or back matter,¹⁶ or making up entire narrative based on a single mention.¹⁷ CLAUDE-3-OPUS was the only model seemingly unaffected by the additional information. However, when faced with two summaries—where one primarily summarized the content of the back matter, since it represented the final chunk—the model declined to perform the task. We regard this cautious approach as preferable to introducing unfounded details or irrelevant content. Examples of such cases are shown in [Table 24](#).

G Details on Experimental Setup

In this section, we provide further details on our experimental setup complemented with further results.

¹⁵"This summary includes a description of who the author thanks at the end of the book which is not important to the plot of the book."

¹⁶"Clair is not a character in this book. The comments are factual, but of Charlie not Clair."

¹⁷"...claims are very focused on the idea of themes of digital age and the story doesn't cover that at all. Its not even based on a modern world." – author's social media accounts are mentioned at the very end of the book.

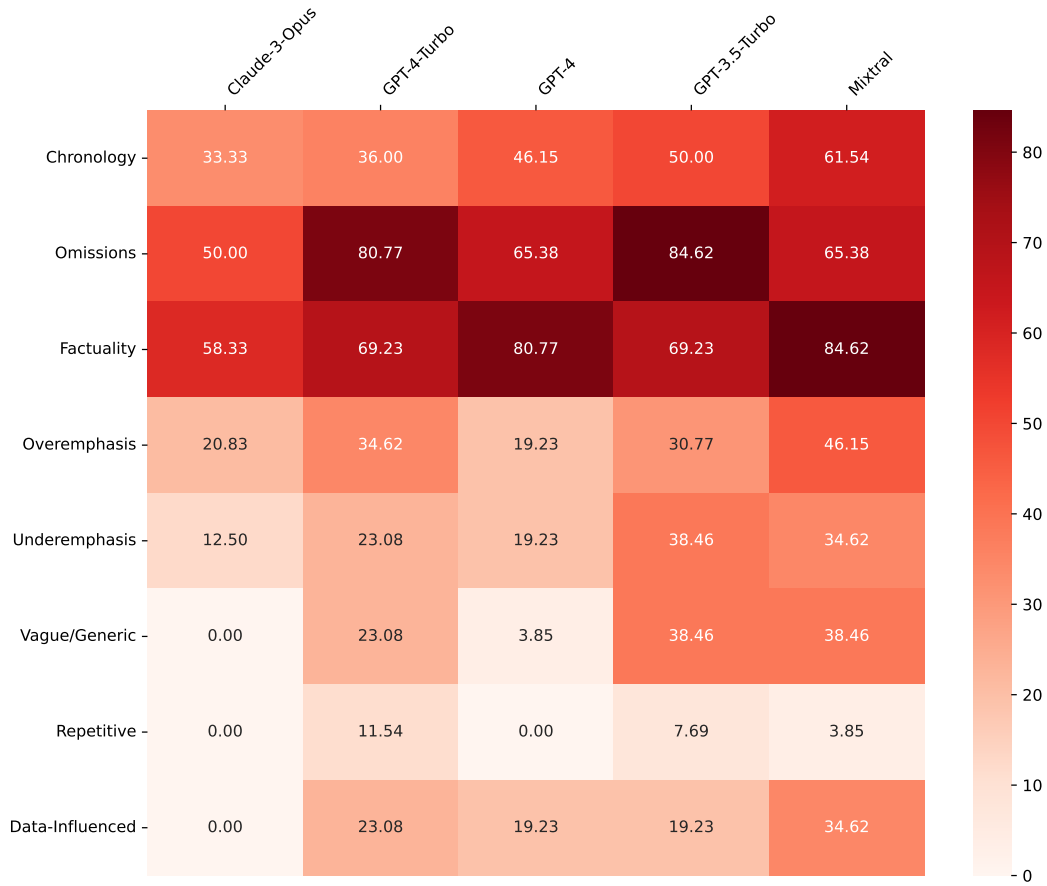


Figure 10: Percentage summaries affected by specific issue mentioned in comments by model.

G.1 Implementation details

For BM25-based evidence retrieval, we use the text of e-books purchased from `amazon.com`, split into passages of up to 256 tokens each. The search is restricted to the book content, and we set $k = 5$ to retrieve the top 5 most relevant passages as evidence.

G.2 Additional Results

Results for each evidence extraction method broken down by summarizer can be found in [Table 26](#). We also report book-wise precision and recall for each evidence extraction method: (1) No-Context ([Table 27](#)); (2) BM25 ([Table 29](#)); (3) Human evidence ([Table 28](#)); (4) Entire book ([Table 30](#)). Further results for the entire book (EB) prompting can be found in [§G.3](#).

G.3 Prompting LLMs with the Entire Book (EB)

Prompting LLMs with large chunks (*entire books*) to evaluate the faithfulness of each claim is prohibitively expensive (see [§G.5](#)). Hence, for this experiment, we select 7 books based on: (1) token length ($<125K$), and (2) presence of at least one *Unfaithful* claim. This sub-dataset includes: (1) “Yellowface,” (2) “Only For The Week,” (3) “Viciously Yours,” (4) “Six Scorched Roses,” (5) “Sorrow and Bliss,” (6) “She Is a Haunting,” and (7) “Pet.” [Table 31](#) shows the number of claims per label in the sub-dataset. Further details on each book can be found in [Table 7](#).

Claim verification with the entire books We prompt CLAUDE-3-OPUS and GPT-4-TURBO models with the entire book content and each claim in order to obtain the *Faithful/Unfaithful* labels.

[Table 32](#) presents a confusion matrix broken down by claim source (i.e., the model that generated the claim) and prediction model (CLAUDE-3-OPUS and GPT-4-TURBO). [Figure 11](#) shows examples of misidentified labels by label-type and prediction model along with human labels and reasoning. [Table 30](#) shows average precision (PR) and recall (RE) broken down by model and book.

G.4 Ablation study

Recall of the claim decomposition step we analyze the extracted claims on a subset of 20 summaries (371 sentences, 450 total extracted claims). We manually evaluate the quality of the extracted claims against the content of each summary. Calculating recall proved challenging due to the ambiguity in granularity (e.g., sentences, clauses, words). Notably, 3.8% of the 371 sentences in the 20 summaries were omitted in the extracted claims. Of these omissions, 85.7% were generic statements, and 14.3% were minor details. Additionally, we observed a small percentage of omissions at the sub-sentential level (e.g., clauses), which did not impact the narrative. All These omissions can be broadly categorized into two types.

- **Generic statements lacking substantive content:** For instance, “The narrative unfolds with intrigue, danger, and treacherous encounters” appears in the summary but is omitted in extracted claims. Note that this sentence only addresses things already covered by other extracted claims in a generic way, so omitting it has few consequences.
- **Insignificant details that contribute little to the narrative:** For instance, “Altha, a 17-century woman, stands trial unjustly accused of witchcraft due to her remarkable healing abilities which are misunderstood by her village” appears in the summary, but “misunderstood by her village” is omitted in the extracted claims. However, this is only a minor detail with little impact on the narrative.

Importantly, we confirmed that none of these discrepancies between the summaries and the extracted claims led to criticisms regarding omissions, chronological errors, or factual inaccuracies in the annotators’ summary-level free-form comments.

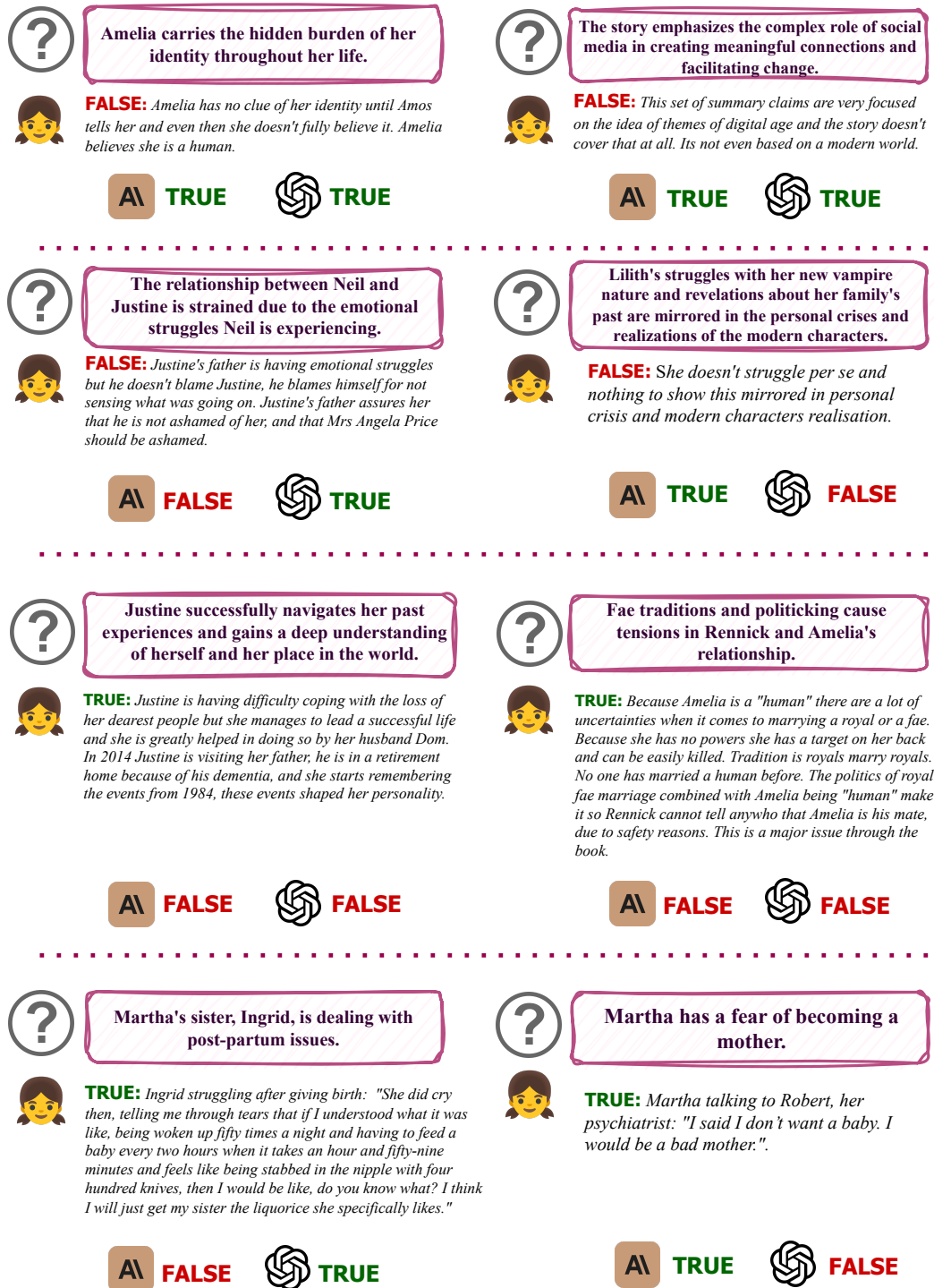


Figure 11: Examples of claims accompanied by annotator labels and reasoning, along with predictions made by CLAUDE-3-OPUS and GPT-4-TURBO.

Varying length of tokens used in BM25 As we increase the length of BM25-retrieved passages, the overall performance improves (Figure 12). However, this approach remains less effective for identifying unfaithful claims than our best performing method, i.e., prompting the model with the content of the entire book. This is likely due to the fact that even longer passages may not provide the entire context needed for verification of broader claims.

Reasoning type of false positive cases We analyzed failure cases in which our auto-rater experiment, conducted on seven books using CLAUDE-3-OPUS and GPT-4-TURBO incorrectly marked an *Unfaithful* claim as *Faithful*. We annotated the types of reasoning required to verify these claims, as presented in Table 33. The results indicate that approximately 75% of these failure cases necessitate multi-hop reasoning across the book. This is significantly higher than the overall distribution of 62.8% across the seven books, suggesting that our auto-raters struggle with multi-hop reasoning.

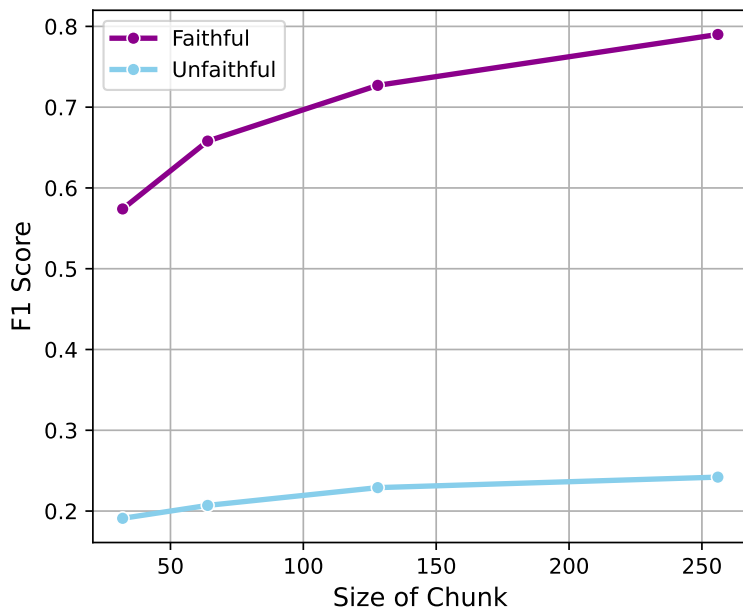


Figure 12: F1 score varying chunk size for BM25.

G.5 API costs

Generating book-length summaries The total cost of summarization for all 130 summaries amounted to about \$288 USD.¹⁸

Extracting claims The total cost of claim extraction for all 130 summaries amounted to about \$8 USD, as the input and output sequence is relatively short.

Prompting with the entire book This experiment costed roughly \$720 USD for GPT-4-TURBO and \$1070 USD for CLAUDE-3-OPUS (corresponding to the last two columns in Table 5).

¹⁸\$64.6 for CLAUDE-3-OPUS, \$169.4 for GPT-4, \$47.5 for GPT-4-TURBO, \$2.8 for GPT-3.5-TURBO, and \$3.4 for MIXTRAL. All in USD.

GPT-4-TURBO	
Summary	<p><i>In the epic narrative of "Divine Rivals," we are introduced to Iris Winnow, a tenacious young woman propelled into the role of a war correspondent amidst a devastating conflict influenced by the ancient deities Enva and Dacre. Iris's journey into the heart of battle is not just professional but deeply personal, fueled by the disappearance of her beloved brother, Forest, who had enlisted to fight, and the anguish over her mother's demise. Her quest for truth and understanding is set against the backdrop of the war-torn Avalon Bluff, close to the war's front lines, where she encounters the stark realities of conflict, from the silent hope of a garden amidst destruction to the horror of wounded soldiers and the terror of artillery fire. Roman Kitt, previously Iris's rival at the Oath Gazette, emerges as a significant figure in her narrative, joining her as a fellow correspondent. Despite their history of competitive antagonism, their relationship evolves, highlighted by Roman's mysterious decision to leave behind a promising career and a broken engagement to be by Iris's side. Their dynamic is further complicated and enriched by their shared experiences on the battlefield and Roman's unspoken feelings for Iris, culminating in a poignant moment of sacrifice during a trench attack where Roman shields Iris with his body from a grenade explosion. Adding a magical element to the story, Iris maintains a correspondence with Carver, a mysterious figure, through enchanted typewriters. These letters offer Iris solace and a metaphysical connection that transcends the chaos surrounding her, underscoring themes of hope and connection. As the narrative unfolds, shocking revelations come to light. Forest, Iris's brother, believed to have been fighting for Enva, is discovered to have been forcefully conscripted by Dacre after a near-death experience. This discovery unfolds through a dramatic reunion where Forest saves Iris from one of Dacre's deadly gas attacks, only to reveal his changed allegiance by deceiving her and leaving Roman behind, presumed dead or captured. The heart of the story explores Iris and Roman's profound relationship, against the backdrop of war and divine manipulation. Moments from their past, including their secret wedding in a garden before a devastating attack by Dacre's forces, and Roman's heartfelt letter confessing his love, serve as poignant reminders of what they fight for. This emotional core is juxtaposed with the chaos of the present, wherein Iris must confront the betrayal of her brother and the uncertainty of Roman's fate. "Divine Rivals" delves into themes of love, loyalty, and the harrowing choices faced by individuals caught in the machinations of gods and war. Iris, within the turmoil of her personal and professional world, remains determined to uncover the truth, seeking to reclaim agency and hope amidst despair. Her journey is not just a search for her brother and Roman but an exploration of the resilience of the human spirit in the face of divine conflicts and the devastation of war. Through Iris's eyes, the story weaves a complex tapestry of human emotions, the indomitable will to survive, and the quest for truth in a reality marred by capricious gods and endless conflict.</i></p>
Extracted Claims	<ol style="list-style-type: none"> 1. Iris Winnow is the main character in the book "Divine Rivals." 2. Iris Winnow becomes a war correspondent during a conflict influenced by the deities Enva and Dacre. 3. Iris's brother, Forest, has disappeared after enlisting to fight in the war. 4. Iris is devastated by her mother's death. 5. Iris's journey takes her to Avalon Bluff, a location close to the war's front lines. 6. Iris encounters the realities of war, including wounded soldiers and artillery fire. 7. Roman Kitt is a significant character in the narrative and joins Iris as a fellow war correspondent. 8. Roman Kitt was Iris's rival at the Oath Gazette before they became colleagues. 9. Their relationship evolves, despite a history of competitive antagonism. 10. Roman Kitt leaves a promising career and a broken engagement to join Iris. 11. Roman Kitt shields Iris from a grenade explosion during a trench attack. 12. Iris maintains a magical correspondence with a mysterious figure named Carver through enchanted typewriters. 13. The letters from Carver offer Iris solace and a metaphysical connection. 14. Forest was believed to be fighting for deity Enva, but is revealed to have been forcefully conscripted by deity Dacre. 15. This revelation comes after Forest saves Iris from a gas attack by Dacre and then deceives her by leaving Roman behind. 16. Iris and Roman had a secret wedding in a garden before an attack by Dacre's forces. 17. Roman writes a letter confessing his love for Iris. 18. Iris must confront her brother's betrayal and uncertainty over Roman's fate. 19. Iris is determined to uncover the truth and reclaim agency and hope amidst despair. 20. Iris's journey is a search for her brother and Roman, and an exploration of human resilience in the face of conflict and divine manipulation. 21. The story of "Divine Rivals" explores themes of love, loyalty, and the difficult choices faced by individuals in times of war.

Table 9: Example of a summary produced by GPT-4-TURBO along with the extracted set of claims for "Divine Rivals," a novel by Rebecca Ross. Examples by the other models can be found in Table 8, Table 10, Table 11 and Table 12.

GPT-4	
Summary	<p><i>Divine Rivals</i> is an intricate narrative set in a dystopian world at war, invoked by two resurrected rival deities, Enva and Dacre. The plot centres around Iris Winnow, a sensitive and tenacious employee at the Oath Gazette newspaper in the city of Oath, and her professional rival Roman Kitt. Iris's uneventful life spirals into chaos with her brother, Forest, joining the gods' war, leaving her with the responsibility of their alcoholic mother, Aster. Caught in a whirlpool of personal and professional rivalry, Iris's bond with Roman deepens in unexpected ways when Roman anonymously receives and responds to Iris's emotionally raw letters intended for Forest. Iris uses a magical vintage typewriter to send these letters, leading to an anonymous correspondence that amplifies their unspoken connection. Roman's social standing is a stark contrast to Iris's modest background; he is affianced to a high society woman, Elinor Little, and faces the challenges of striking a balance between his family's aspirations and his new role at Oath Gazette. Following their mother's sudden passing, Iris impulsively resigns from her job at Oath Gazette. Determined to locate Forest and report on the escalating gods' war, she accepts a position as a war correspondent with the Inkridden Tribune newspaper. Iris embarks on a perilous journey to Avalon Bluff, a town near the war front, accompanied by another correspondent, Thea "Attie" Attwood. Their hostess in the eerie town, Marisol Torres, warns the pair about the regular threats from different mythical creatures such as hounds and eithrals that desolate their surroundings. Roman's arrival in Avalon Bluff reignites their convoluted relationship. He assists Iris in documenting the harrowing realities of the war while managing the growing tension between them. As Roman becomes severely injured during a horrific encounter at the warfront, Iris, who has grown deeply involved in reporting the soldiers' personal narratives, fights her way to ensure his safety. This life-threatening event leads Iris to recognise Roman as Carver, her mysterious pen pal, triggering an intense turmoil of guilt and suspicion over Roman's subterfuge. Amid her growing concern for Roman's fate and her conflicted feelings about his deception, a shocking twist reveals that Forest, previously presumed dead, has survived and fought for Dacre, the enemy deity. Further ensuing chaos sees Roman disappear, leaving Iris trapped between her obligations towards her effectively traitorous brother and the desperate search for her missing love interest. "Divine Rivals" interweaves Iris's personal story of resilience and growth against the backdrop of an escalating war. As she metamorphoses from a correspondent to a key player in the war, Iris embodies the human spirit's ability to endure and adapt amidst adversity. The narrative encapsulates the strength of character, familial obligations, professional rivalries, societal norms, and profound transformations at the heart of the story. Despite the evolving relationships and hardships, Iris pledges to remain by Forest's side and continue her search for Roman. The story of "Divine Rivals" is a testament to human endurance, love, and loyalty amidst a surreal world at war.</p>
Extracted Claims	<ol style="list-style-type: none"> 1. Divine Rivals is a narrative set in a dystopian world at war. 2. The war in the world of "Divine Rivals" is invoked by two resurrected rival deities, Enva and Dacre. 3. The story's plot focuses on Iris Winnow, an employee at the Oath Gazette newspaper in the city of Oath. 4. Iris Winnow has a professional rival named Roman Kitt. 5. Iris Winnow's brother, Forest, joins the god's war. 6. Iris Winnow is left with the responsibility of their alcoholic mother, Aster, after 7. Forest joins the war. 8. Iris Winnow's bond with Roman Kitt deepens when Roman anonymously receives and responds to Iris's letters intended for Forest. 9. Iris Winnow uses a magical vintage typewriter to send letters to Roman Kitt. 10. Roman Kitt is engaged to a high society woman named Elinor Little. 11. Roman Kitt faces challenges balancing his family's aspirations and his role at the Oath Gazette. 12. Iris Winnow resigns from her job at the Oath Gazette after their mother's sudden death. 13. Iris Winnow accepts a position as a war correspondent with the Inkridden Tribune newspaper. 14. Iris Winnow journeys to Avalon Bluff, a town near the war front, with another correspondent named Thea "Attie" Attwood. 15. In Avalon Bluff, Marisol Torres warns Iris Winnow and Thea Attwood about the regular threats from mythical creatures. 16. Roman Kitt arrives in Avalon Bluff and assists Iris Winnow in documenting the war. 17. Roman Kitt is severely injured during an encounter at the warfront. 18. Iris Winnow realizes Roman Kitt is Carver, her mysterious pen pal, after he is severely injured. 19. Iris Winnow feels guilt and suspicion over Roman Kitt's deception. 20. It is revealed that Forest, previously assumed dead, survived and fought for Dacre, the enemy deity. 21. Roman Kitt disappears, leaving Iris Winnow torn between her obligations towards her brother and the search for Roman. 22. Iris Winnow evolves from a correspondent to a key player in the war. 23. Iris Winnow pledges to stay by her brother's side and continue her search for Roman Kitt. 24. The story of "Divine Rivals" is a testament to human endurance, love, and loyalty during a war.

Table 10: Example of a summary produced by GPT-4 along with the extracted set of claims for "Divine Rivals," a novel by Rebecca Ross. Examples by the other models can be found in Table 8, Table 9, Table 11 and Table 12.

GPT-3.5-TURBO	
Summary	<p><i>In the city of Oath, siblings Iris and Forest Winnow face separation as Forest departs for war, leaving Iris to care for their mother. As Iris grapples with loneliness and worry, a fierce rivalry emerges at the Oath Gazette between Iris and Roman Kitt, two ambitious journalists. The backdrop of an impending war between the goddess Enva and god Dacre adds tension and intrigue to their lives. Iris, driven by a desire to prove herself in a male-dominated field, finds herself entangled in personal challenges while navigating her complicated relationship with Roman. As the war escalates and mysterious creatures surface, Iris and Roman become immersed in a world of gods, magic, and journalism, facing both personal and professional dilemmas. Iris decides to resign from her job to embark on a journey to find her brother Forest, who becomes entwined in the conflict. Roman, haunted by his past and familial obligations, struggles to reconcile his feelings for Iris while maintaining his journalistic integrity. Their paths eventually converge in Avalon Bluff, where Iris becomes a war correspondent equipped with a mysterious typewriter with a significant history. As Iris and Roman navigate the chaos of war, their interactions reveal deeper emotions and motivations, leading to a fateful meeting at a high-end restaurant. Attie, a friend of Iris, discloses her talent for playing the violin amidst the turmoil of war, adding a layer of depth to the narrative. Roman's vulnerability and inner turmoil are brought to the forefront when he is saved by Iris during an air raid, further strengthening their bond. As they confront personal and professional challenges amidst the impending battle, Iris reunites with Forest, only to discover his involvement with the opposing forces. Forest, torn between protecting Iris and confronting his past actions, adds a layer of complexity and tension to the story. Amidst the impending attack by Dacre, Iris and Roman make a decision to stay and assist in Avalon Bluff, preparing for the inevitable confrontation. The intertwining destinies of the characters unravel themes of sacrifice, resilience, and the search for identity in a world plagued by war and loss. As truths unfold, loyalty is tested, blurring the lines between good and evil. Characters like Dacre and Enva reveal sinister plans, setting the stage for a complex and emotional journey fraught with uncertainty and conflict. Through grief, rivalry, and love, Iris and Roman confront their vulnerabilities and ambitions, ultimately reshaping their destinies amidst the backdrop of war and its significant consequences.</i></p>
Extracted Claims	<ol style="list-style-type: none"> 1. Iris and Forest Winnow are siblings in the city of Oath. 2. Forest Winnow departs for war and Iris stays behind to care for their mother. 3. Iris struggles with feelings of loneliness and worry. 4. At the Oath Gazette, a fierce rivalry develops between Iris and Roman Kitt, both of whom are ambitious journalists. 5. An impending war between the goddess Enva and god Dacre adds tension and intrigue to Iris and Roman's lives. 6. Iris aims to prove herself in the male-dominated field of journalism. 7. Iris faces personal challenges while navigating her complicated relationship with Roman. 8. As the war intensifies, Iris and Roman encounter mysterious creatures. 9. Iris resigns from her job at the Oath Gazette to search for her brother Forest, who is involved in the war. 10. Roman Kitt has a troubled past and family obligations that make it difficult for him to balance his feelings for Iris and his journalistic integrity. 11. Iris and Roman's paths intersect in Avalon Bluff, where Iris becomes a war correspondent. 12. Iris uses a typewriter with a significant past while working as a war correspondent. 13. Iris and Roman's interactions during the war reveal deeper emotions and motivations. 14. Iris and Roman have a significant meeting at a high-end restaurant. 15. Attie, a friend of Iris, is talented at playing the violin. 16. Roman's vulnerability is revealed when Iris saves him during an air raid. 17. Iris reunites with Forest during the war, discovering his involvement with the opposing forces. 18. Forest is torn between his desire to protect Iris and confronting his past actions. 19. Amidst the approaching attack by Dacre, Iris and Roman decide to stay and help in Avalon Bluff. 20. The characters' intertwining destinies reveal themes of sacrifice, resilience, and identity search in a war-torn world. 21. The story explores blurred lines between good and evil as loyalties are tested. 22. Characters like Dacre and Enva have sinister plans. 23. Iris and Roman confront their vulnerabilities and ambitions, leading to a change in their destinies amid the backdrop of war.

Table 11: Example of a summary produced by GPT-3.5-TURBO along with the extracted set of claims for "Divine Rivals," a novel by Rebecca Ross. Examples by the other models can be found in [Table 8](#), [Table 9](#), [Table 10](#) and [Table 12](#).

MIXTRAL	
Summary	<p><i>In a world torn by a devastating war between two divine families, the Skywards and the Underlings, "Divine Rivals" follows the journey of Iris Winnow, an orphaned junior reporter at the Oath Gazette in the city of Oath, as she strives to become a renowned journalist and uncover the truth about the war that took her mother and brother, Forest, from her. The Skywards and Underlings' conflict has left the realm of Cambria in chaos, where magic and mythology intertwine with the mortal world, and gods walk among humans. Iris' dream of becoming a great journalist is fueled by her love for her late brother and her desire to understand why he joined the war. She faces fierce competition from her professional rival, Roman Kitt, a talented storyteller hailing from a wealthy family. While Iris longs to write about significant issues, Roman, who shares a complicated background with Iris, excels in storytelling and is being pressured by his father to join the family business. Zeb Autry, the editor of the Oath Gazette, is reluctant to cover the war, much to Iris' dismay. Iris' family is burdened with financial difficulties and personal struggles after losing her mother and brother to the war. Driven by her determination to uncover the truth about the war and her late brother's involvement, Iris finds solace in her resolve. The Inkridden Tribune, a rival newspaper, announces an opportunity for war correspondents, enticing Iris to compete with Roman for the position. Iris's journey is further complicated by her mysterious correspondence with an unknown person who leaves her letters in her wardrobe, forming a unique bond. As Iris competes for the position, she reflects on her past and her friendship with Forest, who introduced her to the world of journalism. Their differing views on the war and the Skyward goddess, Enva, drove them apart. Iris wanted to highlight the bravery of soldiers, while Forest aimed to reveal Enva's manipulation of young minds to fuel the war. Despite the challenges, Iris remains focused on her goal and finally manages to land a job as a war correspondent for the Inkridden Tribune. She embarks on a dangerous journey to cover the war, facing various dangers and hardships along the way. Iris continues to write letters to the mysterious stranger, who becomes her confidante and support system during her perilous journey. Meanwhile, the town of Avalon Bluff prepares for an impending battle against an unknown enemy - Dacre, the god of the Underlings. Iris, Roman, and other companions gather in Avalon Bluff to aid the town's defense, as tensions rise with the theft of emergency supply kits and Keegan's distracted behavior. Amidst the chaos, Iris and Roman rekindle their connection, and Roman proposes to Iris, offering a moment of hope in the face of devastation. As the group prepares for battle, Iris witnesses firsthand the horrors of war, including the loss of innocent lives and the destruction of her childhood home, Clover Hill. Despite the pain, Iris continues to report on the war, using her words to both document the battles and search for Forest. "Divine Rivals" is a compelling narrative that explores themes of identity, courage, and the power of words in the face of war. Iris' unwavering determination to make a difference by writing about the truth, seeking answers about her brother, and making her mark as a journalist resonates throughout the story. The author extends gratitude towards the individuals who helped her publish this captivating tale, acknowledging her literary agent, film agent, subrights team, editors, cover designer, copy editor, and fellow authors.</i></p>
Extracted Claims	<ol style="list-style-type: none"> 1. Divine Rivals is set in a world affected by a war between two divine families, the Skywards and the Underlings. 2. The protagonist of "Divine Rivals" is Iris Winnow, an orphan who works as a junior reporter at the Oath Gazette in the city of Oath. 3. Iris Winnow's aspiration to become a renowned journalist is driven by her desire to uncover the truth about the war that resulted in the death of her mother and brother, Forest. 4. The conflict between the Skywards and the Underlings has caused chaos in the realm of Cambria, a place where magic and mythology interact with the mortal world. 5. Iris Winnow has a professional rival named Roman Kitt, a skilled storyteller from a wealthy family. 6. Roman Kitt is under pressure from his father to join the family business. 7. Zeb Autry, the editor of the Oath Gazette, does not want to cover the war, against Iris Winnow's wishes. 8. After the loss of her mother and brother, Iris Winnow's family is facing financial and personal struggles. 9. The Inkridden Tribune, a competing newspaper, advertises a job for war correspondents, leading both Iris Winnow and Roman Kitt to compete for the position. 10. Iris Winnow maintains a mysterious correspondence with an unidentified individual who leaves her letters in her wardrobe. 11. Iris Winnow's brother, Forest, introduced her to journalism. 12. Iris Winnow and her brother Forest had differing views on the war and the Skyward goddess, Enva. 13. Despite various challenges, Iris Winnow achieves her goal of becoming a war correspondent for the Inkridden Tribune. 14. Iris Winnow embarks on a dangerous journey to cover the war. 15. Iris Winnow continues to write letters to the unknown person who becomes her confidante during her journey. 16. The town of Avalon Bluff is preparing for a battle against an unknown enemy. 17. Dacre, the god of the Underlings. 18. Amidst the chaos, Iris Winnow and Roman Kitt rekindle their connection, and Roman Kitt proposes to Iris Winnow. 19. Iris Winnow experiences the horrors of war firsthand, including the loss of innocent lives and the destruction of her childhood home, Clover Hill. 20. Despite the hardships, Iris Winnow continues to report on the war, using her writing to document the battles and search for her brother, Forest. 21. Divine Rivals explores themes of identity, courage, and the power of words during war. 22. Iris Winnow's determination to make a difference by writing about the truth and seeking answers about her brother is a central theme in "Divine Rivals". 23. The author of "Divine Rivals" acknowledges her literary agent, film agent, subrights team, editors, cover designer, copy editor, and fellow authors for their assistance in publishing the book.

Table 12: Example of a summary produced by MIXTRAL along with the extracted set of claims for "Divine Rivals," a novel by Rebecca Ross. Examples by the other models can be found in Table 8, Table 9, Table 10 and Table 11.

Claim Extraction Template
<p>You are trying to verify the faithfulness of statements made in a given summary of a book against the actual text of the book. To do so, you first need to break the summary into a set of "atomic claims", each of which will then be passed to a human who will read the book and verify if the claim is true or not. Each atomic claim must be fully understandable without any other context from the summary (e.g., all entities must be referred to by name, not pronoun), and they must be situated within relevant temporal, location, and causal context whenever possible. Try to keep each atomic claim to a maximum of 2 sentences. Each atomic claim is separated with '- '.</p> <p>Summary:</p> <p>List of atomic claims:</p>
Evaluation Template
<p>You are provided with a context and a statement. Your task is to carefully read the context and then determine whether the statement is true or false. Use the information given in the context to make your decision.</p> <p>Context:</p> <p>Statement:</p> <p>Question: Based on the context provided, is the above statement True or False?</p> <p>Answer:</p>

Table 13: Prompt templates used for CLAIM EXTRACTION and EVALUATION EXTRACTION.

Model	<i>Faithful</i>	<i>Unfaithful</i>	<i>Partial Support</i>	Can't verify
GPT-3.5-TURBO	432	68	79	25
GPT-4	534	31	108	9
MIXTRAL	491	83	122	19
GPT-4-TURBO	437	43	68	15
CLAUDE-3-OPUS	528	23	40	3

Table 14: Distribution of collected labels by model.



LABEL	DEFINITION	EXAMPLE (CLAIM // REASON)
 Claim type		
Event	Concrete event where someone does something, something happens to someone, etc.	<i>Maggie reunites with her old friends and fellow retired spies. // Maggie does not reunite with these people.</i>
Introspection	Characters' thoughts, feelings, opinions, etc.	<i>Justine feels guilty about Amy's death and is haunted by the idea that Amy might be watching her. // Justine doesn't feel guilty.</i>
Cause/effect	Goals, motivation, or purposes	<i>Charlie Brown decides to return to New York to confront Harry Taylor and pursue a connection with Pete Makris after discovering Harry's infidelity. // He is not there to confront Harry.</i>
	Causes or effects of events, actions, thoughts, etc.	<i>The discovery of the love story sparks Jade's curiosity about the house and its past inhabitants. // Jade's curiosity is not sparked by the love story, but by a dream she had.</i>
State	Relationship between characters	<i>Maggie reunites with her old friends and fellow retired spies. // Maggie does not reunite with these people.</i>
	Traits of a character	<i>The magic of royal fae in "Viciously Yours" manifests after twenty-five years. // It does not manifest after 25 years but becomes full strength at 25 years. They are born with magic.</i>
	State of a character, place, etc.	<i>Phillip Hardwicke, a wealthy businessman who was believed to be dead, is revealed to be alive in the story. // Bella Hardwicke is revealed to be alive, not Phillip.</i>
High-level	Characteristics of the narrative	<i>The narrative style of the book is non-linear and features flashbacks and switches between alternate worlds or viewpoints. // The book is almost exclusively from Aurelia's point of view and is linear.</i>
	General story setting	<i>The narrative style of the book is non-linear and features flashbacks and switches between alternate worlds or viewpoints. // It's set in Adcova, Nyaxia is the name of the goddess.</i>
	Themes	<i>The narrative of "The Guest" explores themes of memory, identity, and the pursuit of understanding within human relationships. // It's set in Adcova, Nyaxia is the name of the goddess.</i>
 Reasoning type		
Direct	Reasoning requires only one hop	<i>Alex attends a gathering at Victor's house. // The book states that the gathering is in Helen's house.</i>
Indirect	Reasoning requires more than one hop	<i>Alex and Jack bond over their shared experiences. // They don't have any shared experiences, Jack is from a wealthy, privileged home, and while we aren't told much about Alex's background, we know she doesn't live a cosseted life like him.</i>
	Annotator is arguing for a lack of support	<i>Maggie is portrayed as a skilled assassin in addition to being a former intelligence officer. // No information in the book really supports that.</i>
Subjective	Requires subjective judgment	<i>Forest is torn between his desire to protect Iris and confronting his past actions. // I don't think Forest makes any real effort to confront his past actions, his main motivation is protecting Iris.</i>
Extra info	Requires extra/meta information	<i>The book "Wildfire" is the first in the Icebreaker series. // No evidence in the book, but this is the second in the series, after "Icebreaker".</i>

Table 15: General scheme for assigning labels in our faithfulness annotation analysis along with more examples. This table complements Table 3.

Evidence Coverage		Reasoning-Claim Rel.	
Type	Freq	Type	Freq
Complete	56.1	Direct cont.	50.5
Partial	34.7	Indirect cont.	30.1
Irrel.	1.5	Lack of support	19.4
N/A	7.7		

Table 16: Results from our analysis on evidence coverage and reasoning-claim relationship.

Claim type	Total count	Direct evidence	Logical inference	Subjective interpretation	Requires meta info
Event	101	46.5	46.5	3	4
Thought	26	42.3	50	7.7	0
Cause/effect	36	30.6	61.1	5.6	2.8
State	127	39.4	48.8	5.5	6.3
High-level	36	16.7	55.6	16.7	11.1

Table 17: Distribution of reasoning type for each claim type. Apart from total count, all numbers are reported as a percentage.

Model	Total count	Direct evidence	Logical inference	Subjective interpretation	Requires meta info
CLAUDE-3-OPUS	12	66.7	25	0	8.3
GPT-4	26	20	68	12	0
GPT-4-TURBO	48	27.5	62.7	3.9	5.9
GPT-3.5-TURBO	63	31.2	51.6	10.9	6.2
MIXTRAL	76	38.5	48.7	6.4	6.4

Table 18: Distribution of reasoning type for different models. Apart from total count, all numbers are reported as a percentage.

Model	Total count	Direct contradiction	Indirect contradiction	Lack of support
CLAUDE-3-OPUS	12	66.7	33.3	0
GPT-4	26	48.3	27.6	24.1
GPT-4-TURBO	48	24.5	34.7	40.8
GPT-3.5-TURBO	63	44.1	30.9	25
MIXTRAL	76	57.7	19.2	23.1

Table 19: Distribution of reasoning-claim relationship for different models. Apart from total count, all numbers are reported as a percentage.

Model	Leaning Positive	Leaning Negative
CLAUDE-3-OPUS	<i>This is an excellently structured summary. It communicates the main plot of the book in a succinct, organised manner, touching on all the pivotal moments in a logical sequence. There is a balance between story and commentary.</i>	<i>Amelia's last name isn't very relevant to the summary. She is referred to by it once in the book. Taking out her last name would be more effective. Rennick's mother's death is also omitted which is a key point in the story. The other missing key point is that Amelia is believed to be a human. That has major implications and sets the stage for the events in the story. Overall this summary compared the first was a much better summary of the story because it contains key points and not generalizations.</i>
GPT-4-TURBO	<i>Overall, I think this is a good summary. It hits the major points of the book and the three stories are each in order. However, the book jumps back and forth between the three stories and this summary goes through each one separately. It also leaves out Violet drinking a tea that causes an abortion which is an important part of her character's.</i>	<i>This is not a strong summary of the book. The entire section about visiting and living with Jerry is missing. There is no mention of the paparazzi pics being leaked. This summary seems to focus more on the message of the book rather than the plot. Many of the claims seems to be paraphrased versions of each other and keep repeating the same ideas of this book being about overcoming challenges, facing insecurities etc.</i>
GPT-4	<i>This summary is written in chronological order, it accurately categorizes the excerpt of Wreck the Halls and the claims mentioned here are either True or Partially True. It is easy to follow and is not repetitive. This is one of the stronger summaries. The claims were easy to verify and with the partially true claims - there were only minor issues. However, one aspect that is ignored in all the summaries including this one is the importance of Sumner's family. It touches on his view of his parents as a couple and his relationship with his grandmother. (...)</i>	<i>The majority of these claims were true, but the overall summary does a poor job of following the plot. It skips over many important plot points and does a poor job of describing the main characters. Similar to another one of the summaries, this one makes a lot of broad thematic claims but misses the overall storyline. If I hadn't read the book, I would not understand this summary.</i>
GPT-3.5-TURBO	<i>This is a good summary with most of the plot points covered. The chronological sequence is largely appropriate. Though I would move the claim about Sally and Noah's marriage closer to the conclusion as it happens at the end of the book. In one claim, Jerry is listed as he uncle, though he is her stepfather. I think there should also be a claim that Sally goes to live at Noah's place during the pandemic, as they get closer from their email exchanges. Otherwise, the summary communicates the heart of the book.</i>	<i>This summary has a lot of issues including chronology problems, irrelevant information, and factual errors. Multiple claims draw from the acknowledgments and about the author section and these claims appear in the middle of the summary. Factual errors include that it is an assassin, not Diana that murders Gavin in Bangkok, Declan and Ingrid not being involved in past crimes in Malta, and Luther and Callie not having connections to Maggie's past. This summary misses some important points including that Maggie's husband was killed as a result of the Malta/Cyrano mission.</i>
MIXTRAL	<i>The summary focuses on most of the key points mentioned in the story. It starts in chronological order and focuses on the evolvement of their relationship from a business arrangement towards a happily ever after. It describes the initial and final scenes well, including all relevant details. However, it misses certain events such as getting to know each other, the green card interview, first kiss (...).</i>	<i>This is not a good summary and should not be used. there are too many false statements. Also, the jumping back and forth in the narrative makes it hard to follow.</i>

Table 20: Examples of positive and negative comments submitted by the annotators for specific models

Issue	Definition	Example
CHRONOLOGY	Issues with the chronological ordering of claims.	<i>(...) though it has some chronology problems (Ari's proposal comes after the run in the city, which comes after Josh and Radhya open the pop-up) (...)</i>
OMISSIONS	The annotator mentions any omissions of content that should have been included in the summary.	<i>Omissions: Dr. Rob Valentine groomed Summer from when she was a teenager until she turned 18 and then started an intimate relationship. He later leaves Summer and marries her sister. Even after he is married, he doesn't let Summer move on.</i>
FACTUALITY	Issues with factuality are explicitly mentioned by the annotator. Note that this category correlates partially with the annotated factuality errors.	<i>There were some serious issues with this summary. The first being that the book is referred to as "The Retirement Plan" twice in the summary which is the incorrect title.</i>
OVEREMPHASIS	Too much emphasis put on less significant events or characters.	<i>Saliency: Charles is not an important character, he is the manager of the guest house where she stays in New York, and she only chats to him a couple of times.</i>
UNDEREMPHASIS	Certain events or characters are mentioned but too little emphasis is put on their importance for the story.	<i>There is not enough emphasis on the relationship between Justine and Dom, who later becomes her husband.</i>
VAGUE/GENERIC	Vague or generic claims included in the summary.	<i>Most of the sentences at the end of the summary are generalized and there are no substantial facts.</i>
REPETITIVE	Repetitive claims included in the summary.	<i>As i noted in the annotations, claims 16 and 17 were repetitious and not necessary.</i>
DATA-INFLUENCED	The summary was influenced by front and/or back matter.	[n/a: judgment for this category was made by one of the co-authors during analysis]
COMPREHENSIVE	The annotator praises the summary for being comprehensive.	<i>Out of all the summary claims, this feels the most relevant and comprehensive of the key events that take place.</i>
WELL-DONE	The annotator praises the summary for being well-done.	<i>This was good, things were in sequence, and the main points were covered.</i>

Table 21: Categories used for the analysis of annotators' comments on the quality of the entire summary.

Omission Type	Definition	Example
CHARACTERS	Summary fails to mention important characters.	<i>This summary excluded a lot of the main plot points (...) and the very important principal antagonists Alma, Thomas, and Marion. Alma and Thomas are present-day representations of colonization and investors in the project to turn Nhà Hoa into a bed-and-breakfast.</i>
EVENTS	Summary fails to mention important events and/or turning points.	<i>Some important events in the book were omitted, such as the part where Alex follows a group of young people to a house and has sex with a girl's boyfriend, the part where she sneaks into a club and pretends to be a little boy's nanny, and the part where she follows Margaret to her home. These events are filled with tension, showcase Alex's daring exploits, add a deeper layer of meaning to the story, and ultimately propel the narrative, so they should have been included. Alex almost drowns in the beginning of the novel, a frightening incident that she mentions to Victor during the party.</i>
ATTRIBUTES	Summary fails to mention attributes – details about the events, interactions, characters, etc.	<i>(...) but the overall summary misses a huge plot point of Carver and Roman being the same person.</i>
RELATIONS	Summary fails to mention specific relationships between the characters which are important to the plot.	<i>There is no mention of Clover and Amos the Desert King being mates. Nor Fawn and Dean the Garden King being mates.</i>
THEMES	Summary fails to depict main/important themes of the novel.	<i>Water, pools, and beaches are recurring metaphors in the book, yet they aren't mentioned or highlighted in any of the claims.</i>

Table 22: Description of omission categories used for annotating comments provided by our evaluators. Omissions were annotated in two steps: (1) a binary choice (either omissions were mentioned or not), and (2) categorization.

OMISSION TYPE	CLAUDE-3-OPUS	GPT-4-TURBO	GPT-4	GPT-3.5-TURBO	MIXTRAL
Characters	4.17	23.08	7.69	7.69	23.08
Events	33.33	57.69	38.46	65.38	38.46
Attributes	16.67	38.46	34.62	34.62	38.46
Relations	8.33	11.54	11.54	11.54	15.38
Theme	12.50	15.38	3.85	15.38	19.23

Table 23: Percentage of summaries affected by specific type of omission error by model.

Model	Comment	Summary
CLAUDE-3-OPUS	<p>None of the claims is completely factually correct, or focused on the book as they ought to be. The first two claims are the only ones where the protagonist Alex is mentioned, but both claims are still wrong. The other claims are incomprehensible, have absolutely nothing to do with the book, and are therefore inadmissible. In fact, the claims appear to be focused on a different topic entirely, not the book. Please note that the quotes I included below don't necessarily prove any of the claims wrong (or right, for that matter). (...)</p> <p>NOTE: The model declines to merge two summaries due to significant discrepancies in content.</p>	<p>I apologize for the confusion, but the provided summaries do not appear to be related to the same story. Summary 1 describes a novel about a young woman named Alex and her experiences in the Hamptons, while Summary 2 simply states that the provided excerpt is a table of contents or metadata without any actual story content. As there is no coherent narrative between the two summaries, it is not possible to merge them into a single, comprehensive summary. In order to create a merged summary, I would need access to summaries that are all related to the same story, with each summary focusing on different parts or aspects of that story. If you are able to provide a series of summaries that are all connected to the same narrative, I would be happy to assist in combining them into a single, cohesive summary that includes key events, backgrounds, settings, characters, objectives, and motivations, while also introducing major elements and organizing the summary in a chronological manner. Please let me know if you have any other summaries related to the story mentioned in Summary 1, as that would allow me to proceed with the merging process you have requested.</p>
GPT-4	<p>This summary is largely factual and looks at the overall arc of the text, though it has some chronology problems (Ari's proposal comes after the run in the city, which comes after Josh and Radhya open the pop-up), and focuses disproportionately on the author's conversation with another writer, Kate Robb, as opposed to the text itself. It also doesn't discuss the nonlinear storytelling.</p> <p>NOTE: The model includes back matter information. While accurate, it should not have been included in the summary.</p>	<p>(...) Goldbeck's insightful writing and clever comedic timing elevate the narrative, making it stand out and resonate with a broad variety of readers. Reflecting upon shared preferences like tequila and various NYC spots, Goldbeck successfully manages to craft relatable characters with everyday experiences, despite their complex emotional journey. The author's creative process in the narrative, including the decision to cut a subplot involving Ari working as a comedian on a cruise ship, adds a layer of interesting behind-the-scenes perspective. The narrative style, which pays homage to classic 2000s romantic comedies, echoes elements of escapism, emphasizing the importance of literature during unprecedented times like the global pandemic. (...)</p>
GPT-4-TURBO	<p>This set of summary claims are very focused on the idea of themes of digital age and the story doesn't cover that at all. Its not even based on a modern world. Chronology wasn't relevant as majority of the claims weren't correct, applicable or verifiable. Factualty was on zero and the claims don't cover any of the key events that occurs except that there was themes of life and death.</p> <p>NOTE: The model hallucinates information based on brief mentions of author's social media accounts.</p>	<p>(...) Parallel to Liliith and Vale's journey, the story delves into the lives of modern city-dwellers: Alex, a budding Instagram influencer; Jordan, a TikTok content creator; Sam, a Twitter-using political activist; and Taylor, who leverages Facebook for business and family connections. Each character embodies the diverse ways social media influences the pursuit of fulfillment, showcasing the profound effects of these platforms on personal and collective experiences. (...) Themes of authenticity versus constructed online personas, the search for connectivity amidst digital isolation, and the double-edged sword of viral fame are explored, reflecting on the profound impacts of social media in modern society. (...) In summary, this narrative offers a comprehensive exploration of human experiences, spanning the ethereal and the digital. By paralleling the mystical journey of Liliith against the backdrop of a society deeply influenced by social media, the story invites readers into an immersive world where the boundaries between life and death blur, and where the quest for understanding and connection transcends the mortal coil and the digital divide. Through the intertwined lives of its characters, the story serves as a reflection on the complexities of the digital age, highlighting the profound and varied impacts of social media on the contemporary human condition.</p>

Table 24: Examples of summaries influenced by front/back matter information along with the annotators' comments. The CLAUDE-3-OPUS example was excluded from the analysis because the model failed to generate a summary. Although not ideal, this behavior is arguably better than the model fabricating content.

Model	COMMENTS
CLAUDE-3-OPUS	<i>"It also focuses extensively on the last couple chapters of the book. This is the only summary so far that has included claims about the very last chapter of the book that is from Dacre's point of view."</i>
CLAUDE-3-OPUS	<i>"(...) and hits the main thematic elements of the text, though it disproportionately addresses the epilogue over other portions of the text (...)"</i>
CLAUDE-3-OPUS	<i>"This summary included a lot of true elements, but also included many irrelevant details not integral to the plot. This is especially true for the end of the book."</i>
GPT-4-TURBO	<i>"This summary focuses heavily on the end of the book and misses plot points that happen in the beginning of the book."</i>
GPT-4	<i>"(...) and focuses disproportionately on the author's conversation with another writer, Kate Robb, as opposed to the text itself." [the interview is included at the end of the book]</i>
MIXTRAL	<i>"The summary puts an emphasis on Part 4 of the book which is not in proportion to the rest of the book" [Part 4 is the last part]</i>

Table 25: Comments from annotators on models' focus towards the book's end

Summarized by	No-Context		BM25		Human Evidence	
	Faithful	Unfaithful	Faithful	Unfaithful	Faithful	Unfaithful
<i>F1 score against the human annotations</i>						
GPT-3.5-TURBO	0.727	0.261	0.835	0.476	0.712	0.430
MIXTRAL	0.643	0.183	0.837	0.244	0.784	0.406
GPT-4	0.687	0.130	0.794	0.088	0.721	0.207
GPT-4-TURBO	0.634	0.033	0.887	0.080	0.792	0.139
CLAUDE-3-OPUS	0.674	0.000	0.738	0.000	0.684	0.031
Overall	0.681	0.124	0.826	0.215	0.755	0.259
<i>Token length of the given evidence against prediction label</i>						
GPT-3.5-TURBO	0.0		1136.4	1131.0	292.4	126.5
MIXTRAL	0.0		1139.5	1132.9	211.7	153.4
GPT-4	0.0		1138.6	1132.5	241.7	160.8
GPT-4-TURBO	0.0		1141.9	1138.3	257.8	152.9
CLAUDE-3-OPUS	0.0		1134.5	1128.6	214.4	151.5
Average	0.0		1138.2	1132.6	243.6	149.0

Table 26: Comparison of automatic evaluation using GPT-4-TURBO based on different evidence extraction methods. We also presents the F1 score and token length of the extracted evidence for each summarizer. Overall mean values were calculated using all the claims across FABLES.

TITLE	PR- <i>Faithful</i>	RE- <i>Faithful</i>	PR- <i>Unfaithful</i>	RE- <i>Unfaithful</i>
A Haunting on the Hill	0.821	0.329	0.230	0.650
Agency for Scandal	0.960	0.133	0.034	0.833
Divine Rivals	0.960	0.156	0.140	0.917
Fairytale of New York	1.000	0.123	0.174	1.000
Flawless	0.950	0.217	0.012	0.500
Fourth Wing	1.000	0.169	0.112	1.000
Modern Divination	1.000	0.062	0.092	1.000
Only For The Week	0.893	0.186	0.056	0.500
Pet	0.871	0.151	0.121	0.881
Romantic Comedy	1.000	0.170	0.020	1.000
Same Time Next Year	0.667	0.161	0.204	0.700
She Is a Haunting	0.917	0.220	0.065	0.667
Six Scorched Roses	0.750	0.179	0.228	0.701
Sorrow and Bliss	0.983	0.197	0.029	0.750
The Atonement Murders	1.000	0.067	0.069	1.000
The Guest	0.688	0.182	0.253	0.810
The Marriage Act	1.000	0.101	0.041	1.000
The Spy Coast	0.864	0.151	0.103	0.790
The Wager	1.000	0.495	0.085	1.000
The White Lady	0.750	0.045	0.151	0.938
This Impossible Brightness	1.000	0.147	0.044	1.000
Viciously Yours	0.950	0.200	0.113	0.833
Weyward	0.947	0.504	0.161	0.750
Wildfire	1.000	0.229	0.036	1.000
You, Again	1.000	0.214	0.041	1.000
Yellowface	0.933	0.163	0.119	0.938

Table 27: Precision (PR) and Recall (RE) from LM evaluation using GPT-4-TURBO **no context** for each book.

TITLE	PR- <i>Faithful</i>	RE- <i>Faithful</i>	PR- <i>Unfaithful</i>	RE- <i>Unfaithful</i>
A Haunting on the Hill	0.967	0.517	0.337	0.950
Agency for Scandal	1.000	0.570	0.063	1.000
Divine Rivals	0.980	0.427	0.190	0.917
Fairytale of New York	1.000	0.230	0.195	1.000
Flawless	1.000	0.352	0.033	1.000
Fourth Wing	0.985	0.662	0.244	0.950
Modern Divination	1.000	0.409	0.138	1.000
Only For The Week	1.000	0.509	0.119	1.000
Pet	0.981	0.676	0.242	0.833
Romantic Comedy	1.000	0.204	0.024	1.000
Same Time Next Year	0.851	0.780	0.300	0.600
She Is a Haunting	0.985	0.641	0.186	0.750
Six Scorched Roses	0.907	0.521	0.334	0.759
Sorrow and Bliss	1.000	0.340	0.038	1.000
The Atonement Murders	1.000	0.333	0.093	1.000
The Guest	0.868	0.644	0.391	0.589
The Marriage Act	1.000	0.360	0.049	1.000
The Spy Coast	0.975	0.392	0.168	0.933
The Wager	1.000	0.810	0.220	1.000
The White Lady	1.000	0.311	0.201	1.000
This Impossible Brightness	1.000	0.378	0.049	1.000
Viciously Yours	0.921	0.508	0.156	0.688
Weyward	1.000	0.598	0.157	1.000
Wildfire	1.000	0.387	0.042	1.000
You, Again	0.967	0.469	0.048	0.667
Yellowface	0.935	0.461	0.153	0.813

Table 28: Results of average Precision (PR) and Recall (RE) estimated by **human evidence** and LM evaluation using GPT-4-TURBO for each book.

TITLE	PR- <i>Faithful</i>	RE- <i>Faithful</i>	PR- <i>Unfaithful</i>	RE- <i>Unfaithful</i>
A Haunting on the Hill	0.902	0.781	0.520	0.550
Agency for Scandal	0.974	0.721	0.067	0.333
Divine Rivals	0.907	0.541	0.209	0.556
Fairytale of New York	0.874	0.712	0.263	0.556
Flawless	1.000	0.663	0.056	1.000
Fourth Wing	0.953	0.759	0.244	0.600
Modern Divination	0.950	0.644	0.142	0.700
Only For The Week	0.967	0.789	0.167	0.417
Pet	0.907	0.604	0.175	0.649
Romantic Comedy	1.000	0.675	0.044	1.000
Same Time Next Year	0.836	0.859	0.333	0.425
She Is a Haunting	0.949	0.772	0.100	0.250
Six Scorched Roses	0.816	0.500	0.269	0.616
Sorrow and Bliss	1.000	0.588	0.077	1.000
The Atonement Murders	1.000	0.642	0.115	1.000
The Guest	0.845	0.737	0.395	0.598
The Marriage Act	0.987	0.833	0.119	0.833
The Spy Coast	0.953	0.527	0.216	0.738
The Wager	0.958	0.862	0.100	0.167
The White Lady	0.836	0.628	0.097	0.250
This Impossible	0.961	0.607	0.022	0.250
Brightness				
Viciously Yours	0.929	0.725	0.172	0.521
Weyward	0.969	0.774	0.300	0.583
Wildfire	0.980	0.664	0.043	0.750
You, Again	0.960	0.609	0.015	0.333
Yellowface	0.941	0.695	0.204	0.562

Table 29: Results of average Precision (PR) and Recall (RE) estimated by **BM25** retriever and LM evaluation using GPT-4-TURBO for each book.

Evaluation LM	TITLE	PR- <i>Faithful</i>	RE- <i>Faithful</i>	PR- <i>Unfaithful</i>	RE- <i>Unfaithful</i>
GPT-4-TURBO	Only For The Week	0.960	0.972	0.333	0.167
	Pet	0.921	0.923	0.333	0.262
	She Is a Haunting	0.957	0.949	0.417	0.333
	Six Scorched Roses	0.794	0.958	0.625	0.288
	Sorrow and Bliss	1.000	0.868	0.139	1.000
	Viciously Yours	0.919	0.919	0.367	0.354
	Yellowface	0.952	0.948	0.450	0.438
CLAUDE-3-OPUS	Only For The Week	0.980	0.971	0.571	0.667
	Pet	0.920	0.910	0.429	0.462
	She Is a Haunting	0.968	0.968	0.571	0.571
	Six Scorched Roses	0.919	0.958	0.800	0.667
	Sorrow and Bliss	1.000	0.966	0.429	1.000
	Viciously Yours	0.931	0.931	0.417	0.417
	Yellowface	0.963	0.963	0.700	0.700

Table 30: Average Precision (PR) and Recall (RE) for the **Entire Book (EB)** approach (i.e., prompting the model with a claim and entire book as evidence) broken down by the rater models (GPT-4-TURBO and CLAUDE-3-OPUS), for each book.

Model	Faithful	Unfaithful	Partial support	Can't verify
GPT-3.5-TURBO	111	22	20	8
GPT-4	152	11	30	3
MIXTRAL	144	18	34	5
GPT-4-TURBO	107	17	25	1
CLAUDE-3-OPUS	140	1	17	0

Table 31: Number of claims per label for each model in the sub-dataset of seven books.

Human	Claim Source	GPT-4-TURBO		CLAUDE-3-OPUS	
		<i>Unfaithful</i>	<i>Faithful</i>	<i>Unfaithful</i>	<i>Faithful</i>
<i>Unfaithful</i>	CLAUDE-3-OPUS	0	1	0	1
	GPT-4-TURBO	3	14	9	8
	GPT-4	6	5	9	2
	GPT-3.5-TURBO	8	14	13	9
	MIXTRAL	11	7	10	8
<i>Faithful</i>	CLAUDE-3-OPUS	10	130	7	133
	GPT-4-TURBO	0	107	4	103
	GPT-4	20	132	10	142
	GPT-3.5-TURBO	10	101	7	104
	MIXTRAL	8	136	3z	141

Table 32: Count of labels predicted by CLAUDE-3-OPUS and GPT-4-TURBO contrasted with human-annotated labels, segmented by the model that generated each claim.

Reasoning Type	CLAUDE-3-OPUS (28 examples)	GPT-4-TURBO (37 examples)	Both Models (24 examples)
Indirect	75.0%	73.0%	75.0%
Direct	14.3%	10.8%	12.5%
Subjective	7.1%	10.8%	8.3%
Extra info	3.6%	5.4%	4.2%

Table 33: Reasoning type distribution for false positives case by each model