Evaluating the Factuality of Zero-shot Summarizers Across Varied Domains

Sanjana Ramprasad◇ **Kundan Krishna**♣ **Zachary C. Lipton**♣ **Byron C. Wallace**◇
◇Northeastern University
♣ Carnegie Mellon University
{ramprasad.sa,b.wallace}@northeastern.edu
{kundank,zlipton}@andrew.cmu.edu

Abstract

Recent work has shown that large language models (LLMs) are capable of generating summaries *zero-shot* (i.e., without explicit supervision) that are often comparable or even preferred to manually composed reference summaries. However, this prior work has focussed almost exclusively on evaluating news article summarization. How do zero-shot summarizers perform in other, potentially more specialized, domains? In this work we evaluate zero-shot generated summaries across specialized domains including: biomedical articles, and legal bills (in addition to standard news benchmarks, for reference). We focus especially on the factuality of outputs. We acquire annotations from domain experts to identify inconsistencies in summaries and systematically categorize these errors. We analyze whether the prevalence of a given domain in the pretraining corpus affects extractiveness and faithfulness of generated summaries of articles in this domain. We release all collected annotations to facilitate additional research toward measuring and realizing factually accurate summarization, beyond news articles.[1]

1 Introduction

Modern LLMs now offer strong zero-shot summarization performance, and even surpass fine-tuned models according to human assessments (Goyal et al., 2022). Indeed, zero-shot summaries are sometimes deemed comparable in quality to reference summaries (Zhang et al., 2023). Past evaluative work, however, has focused nearly exclusively on news article summarization, a domain in which there is no shortage of available training data.

But zero-shot summarization is perhaps *most* appealing in niche domains where acquiring training data with which to fine-tune summarization models is sparse and may be prohibitively expensive to collect. Recent work (Shaib et al., 2023; Tang et al., 2023) suggests the promise of zero-shot summarization in such domains. However, there has not yet been a comprehensive investigation of the factuality of model outputs produced in zero-shot summarization across multiple domains (i.e., beyond news). Here we address this gap, and compare the quality of zero-shot summaries generated in niche domains (law, medicine) to those generated for news articles.

In evaluating these models, we center the consistency and faithfulness of summaries generated by LLMs with respect to the input (source) document. Inconsistencies within summaries have long posed a challenge (Maynez et al., 2020; Pagnoni et al., 2021), motivating approaches intended to mitigate this issue (Zhu et al., 2020; Cao and Wang, 2021), and for automated evaluation of factuality (Kryściński et al., 2019; Goyal and Durrett, 2020; Fabbri et al., 2021; Scialom et al., 2021; Laban et al., 2022; Luo et al., 2023). Here we systematically assess the factual accuracy of zero-shot summarizers across a diverse set of specialized domains.

Specifically, we look to answer four major questions. (1) What is the *prevalence* of errors in zero-shot summaries across various domains, and how does this compare to established results on news summarization tasks? (2) Are the *types* of errors observed in these niche domains different from what has been seen in news article summarization? (3) What is the relationship between the frequency of domains in training corpora and the likelihood of model hallucinations in these domains? (4) Are existing automatic systems for factual evaluation reliable across multiple domains?

To answer these questions, we enlist expert annotators to manually evaluate the outputs from two representative zero-shot summarization systems—GPT-3.5 (`gpt-3.5-turbo-0301`; Brown et al. 2020) and Flan-T5-XL (Chung et al., 2022)—

[1] The dataset can be downloaded from https://github.com/sanjanaramprasad/zero_shot_faceval_domains

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 50–59
March 17-22, 2024 ©2024 Association for Computational Linguistics

across standard and niche summarization datasets. Specifically, we evaluate (zero-shot) summaries of medical and legal documents, as well as news articles for reference.

In general, we find that the proportion of factual inconsistencies in summaries varies considerably across domains, calling into question the community focus on news summarization datasets specifically. Further, we find evidence that the prevalence of articles in pretraining data from a given domain may correlate with the factuality of summaries of articles from the same. We speculate that this may be due to the model introducing content implicit in its weights in such cases (whereas it may have less “knowledge” in niche domains), although this would need to be validated in future work.

2 Manual Evaluations of Summaries

Data We use XSUM (Narayan et al., 2018) and CNN-DM (Hermann et al., 2015) for news, as well as niche domains like PubMed (medicine; Cohan et al. 2018) and legal bills (law; Kornilova and Eidelman 2019) for comparison. We select articles shorter than 4096 tokens from the test sets to accommodate model token limitations, resulting in approximately 22,000 articles for news, 3,000 for billsum, and 200 for PubMed. We randomly (i.i.d.) sample 50 articles from each domain. We provide more data statistics in Appendix A.1

Model Details We run experiments with GPT-3.5 (`gpt-3.5-turbo-0301`) and Flan-T5-XL (Chung et al., 2022). We use a general prompt similar to prior work (Goyal et al., 2022) for generating summaries across domains. Specifically, the prompt is as follows: "Article: [article]. Summarize the above article."

Annotation Collection To acquire manual assessments of model-generated summaries, we hire domain experts via Upwork.[2] We recruit two experts for each domain: linguistics experts for news, attorneys in civil litigation and public policy for the legal domain, and medical doctors (MDs) for the medical domain.

Our evaluation consists of two rounds. In the first round, annotators primarily assess the factual consistency of summaries in relation to the source article. We collect sentence-level annotations, instructing annotators to identify sentences with inconsistencies. The average proportion of such sentences in each domain is a key reported result. The inter-annotator agreement at the summary level was determined by calculating the fraction of instances where both annotators identified a summary as inconsistent with respect to the source. The agreement values are 0.80, 0.72, and 0.85 for news, billsum, and PubMed, respectively. We provide more details about annotation, including agreement statistics, in the Appendix A.2

In the second round of annotations, we categorize errors based on typology previously introduced (Tang et al., 2022). These errors include: (a) *Intrinsic* errors, which misrepresent source content, and (b) *Extrinsic* errors, or “hallucinations”, which introduce terms or concepts not in the source. Past research (Cao et al., 2021) has shown that hallucinations can align with real-world knowledge and even be beneficial.

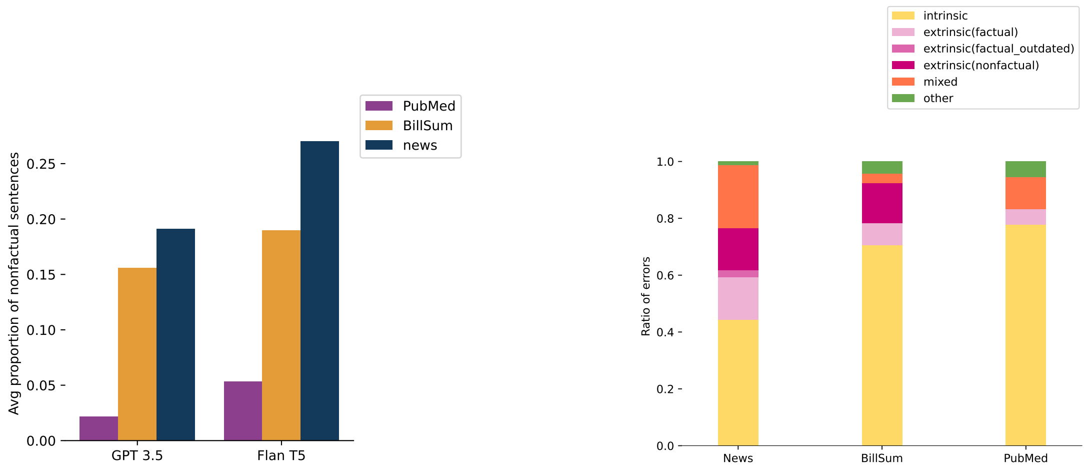
To distinguish extrinsic errors further, we subcategorize them into: *Extrinsic nonfactual* errors, which are hallucinations inconsistent with world knowledge; and *Extrinsic factual* errors, where hallucinations align with world knowledge. Additionally, considering that LLMs are trained on data up to specific points in time, we introduce *Extrinsic factual outdated* errors, which capture hallucinations that are outdated but were once in alignment with world knowledge (e.g., former presidents of countries). To assess the factual nature of hallucinations, annotators use online resources like Google Search and Wikipedia, in keeping with prior work (Cao et al., 2021).

3 Results

How prevalent are errors across domains? Figure 1a shows the average proportion of sentences marked as inconsistent (with respect to the corresponding input) in summaries generated by GPT-3.5 (Brown et al., 2020) and Flan-T5 XL (Chung et al., 2022) for three domains: News, medical, and legal. Perhaps surprisingly, we observe a higher prevalence of inconsistencies for news articles, as compared to the specialized domains of medicine and law. While Flan-T5 introduces more errors than GPT-3.5 overall, the trends are analogous.

Error categories across domains We next characterize the distribution of error categories in factually inconsistent summaries generated by models across the domains considererd. Figure 1b reports

[2] Upwork is a contracting platform suited to such work because it allows hiring individuals with specific background; `http://upwork.com`.



(a) Prevalence of factual errors in each of domains

(b) Distribution of error categories across domains

Figure 1: Distribution of errors and error categories across domains

the distribution of error categories for both models.[3] There are more extrinsic errors introduced in the news domain compared to the niche domain datasets. We include “mixed” errors for cases where errors were classified as different types (intrinsic/extrinsic) by annotators. The news domain has a higher frequency of such cases. Reviewing these, we find that they include cases where the summary both misinterprets source information and where it introduces new information. We provide examples in Appendix A.5.

An “other” option is available to annotators, along with a comment box for capturing miscellaneous errors. Annotator comments highlight instances where there is no clear misunderstanding but instead a misleading overall impression, such as the over-generalization of specific information in the summary

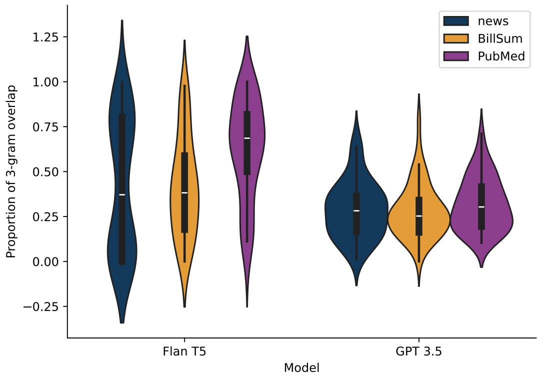


Figure 2: Proportion of 3-gram overlaps between model generated summaries and articles. We observe the most copying in the case of PubMed (especially under Flan-T5). This likely explains the greater factuality observed in this domain, and may reflect unfamiliarity with the domain (see Figure 3).

How extractive are summaries, and how does this relate to factuality? We investigate the relationship between extractiveness (i.e., degree of copying) and factual accuracy across domains. Specifically, we take the proportion of 3-gram sequences in the summary that are also present in the source for each source-summary pair as a proxy measure for extractiveness.

Figure 2 reveals that there is a comparable level of copying across different models and domains. However, models tend to copy more often when summarizing articles in the PubMed dataset; this could explain the lower frequency of errors in this domain, since extractive summaries are unlikely to “hallucinate” by definition. We calculated Spearman rank correlations between 3-gram overlaps and factuality scores for article-summary pairs. The correlations for the news, billsum, and PubMed domains are 0.61, 0.38, and 0.16 respectively.

Domain representation in pretraining corpora and its relation to factuality. One possible explanation for the higher proportion of factual errors in news datasets compared to specialized domains is that general news has greater representation in the training data. As a proxy to measure model exposure to articles belonging to these domains we prompt LLMs to generate overviews of articles based on titles only (headlines for news articles, bill titles for billsum, and study titles for PubMed).

[3] Model-specific distributions are in Appendix A.6

Domain	QAFactEval	QuestEval	SummC-ZS	SummaC-Conv
News	0.58	0.45	0.47	0.59
BillSum	0.27	0.15	0.23	0.30
Pubmed	0.09	-0.03	0.11	0.06

Table 1: Performance of automated factuality metrics across domains. We report the spearmanrank correlation between the average proportion of inconsistent sentences and the predicted scores by the automated metrics.

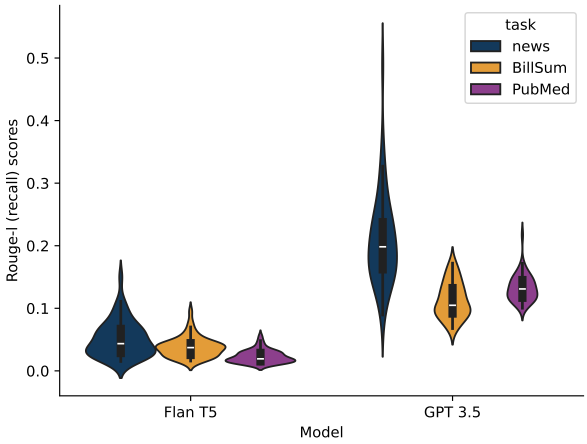


Figure 3: ROUGE-L recall scores of original articles in comparison with LLM-generated documents to measure domain exposure during pretraining. Models show higher familiarity with news topics, which may lead to the inclusion of unsupported content in summaries.

We use the template “Generate a comprehensive overview of the following topic: [title]” to generate text for each article title, assessing LLMs’ memorization. We speculate that increased exposure to an article topic in training data should enable LLMs to reproduce more content present in the original article (as seen with popular celebrities/events, for instance). We assess information overlap between the generated text and original article using ROUGE-L recall, favoring it over embedding based metrics because it emphasizes longest common subsequences based on exact word matches, which makes it suitable for measuring memorization. This is also preferable for content containing specialized terminology like PubMed abstracts and legal articles.

Figure 3 shows that GPT-3.5 and Flan-T5-XL have higher ROUGE-L recall scores for news, suggesting that these models have had more exposure to news topics; this could explain the increased extrinsic error rate in news summaries. Furthermore, in Appendix A.7, we show similar trends using an alternative approach to measure domain representation by directly querying the pretraining corpus with article titles, and using the number of retrieved articles as a proxy for representation.

Are existing automatic systems for factual evaluation reliable across different domains? Prior research has focused on creating automated metrics for evaluating factuality of generated summaries using question answering (Scialom et al., 2021; Fabbri et al., 2021), natural language inference (NLI; Laban et al. 2022), dependency entailment(Goyal and Durrett, 2020), and classification methods (Kryściński et al., 2019). The performance of these metrics has been assessed almost exclusively on evaluation benchmarks comprising model-generated summaries annotated for factuality in the news domain (Kryściński et al., 2019; Wang et al., 2020; Huang et al., 2020; Maynez et al., 2020; Pagnoni et al., 2021; Cao and Wang, 2021; Goyal and Durrett, 2021; Cao et al., 2022). The effectiveness of such automated factuality metrics outside of news is underexplored.

To address this, we use our annotated dataset to examine the performance of QAFactEval (Fabbri et al., 2021), QuestEval (Scialom et al., 2021) and SummaC variations (Laban et al., 2022) across all three domains. The results in Table 1 reveal that automated metrics struggle when applied to niche domains. We note that the lower scores observed for PubMed could be due to the scarcity of observed errors in this dataset, which makes it challenging to reliably evaluate its performance.

4 Conclusions

We analyzed zero-shot summarization abilities of two LLMs, focusing on factuality. Surprisingly, inaccuracies were *more likely* to be introduced in summaries of news articles compared to legal and biomedical domains. Specifically, in this domain we observed more extrinsic errors—i.e., hallucinations of content not mentioned in the source—whereas errors in specialized domains were typically related to an apparent “misunderstanding” of concepts in the source.

We hypothesize that the discrepancy could result from a higher proportion of news articles in the model's pretraining data, supported by preliminary evidence. Additionally, we observed lower Spearman rank correlations between automated metrics and human annotations in specialized domains compared to news articles, highlighting the necessity for manual evaluations or the development of new metrics for diverse benchmarks.

Limitations

This work has a few important limitations. The main challenge in achieving a comprehensive evaluation is the cost involved in hiring domain experts. For news domain, we hire proofreaders and linguists at an average hourly rate of $30 USD/hr. For billsum, we hire attorneys at $40 USD/hr, and for pubmed, we hire doctors at $50 USD/hr. The total cost of annotating 100 article-summary pairs across the three domains amounts to approximately $3000 USD, making scalability of the annotations challenging.

We evaluated only two (representative) LLMs; it is possible that other models would show different patterns in behaviour. Another limitation of this work is that we used only a single prompt to generate summaries; although similar to a previously evaluated prompt (Goyal et al., 2022) it is unclear how choice of prompt might interact with factuality of outputs across domains.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2021. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. *arXiv preprint arXiv:2109.09784*.

Shuyang Cao and Lu Wang. 2021. Cliff: Contrastive learning for improving faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2109.09209*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents.

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

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. *arXiv preprint arXiv:2010.05478*.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? *arXiv preprint arXiv:2010.04529*.

Anastassia Kornilova and Vlad Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*.

Chantal Shaib, Millicent L Li, Sebastian Joseph, Iain J Marshall, Junyi Jessy Li, and Byron C Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success). *arXiv preprint arXiv:2305.06299*.

Liyan Tang, Tanya Goyal, Alexander R Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. 2022. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. *arXiv preprint arXiv:2205.12854*.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin Rousseau, et al. 2023. Evaluating large language models on medical evidence summarization. *medRxiv*, pages 2023–04.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2020. Enhancing factual consistency of abstractive summarization. *arXiv preprint arXiv:2003.08612*.

A Appendix

A.1 Data Statistics

This section presents additional data statistics in Table 2, including the average number of sentences in both summaries and source articles across various domains, offering context for comparisons.

A.2 Annotation Details

We recruited annotators on the Upwork platform and selected two domain experts for each task. In the first round, annotators identified sentences in the summary that were inconsistent with the source. The agreement at the summary level includes all cases where both annotators marked at least one sentence in the summary as inconsistent. At the sentence level, we calculated agreement as a function of the fraction of instances in which annotators marked the same sentence within a summary as being inconsistent with the source. We calculate agreement for the error categories by considering the pre-defined error types chosen by each annotator. Notably the datasets, particularly pubmed, has an imbalance due to the dataset's significant skew in error labels, resulting in a higher expected chance agreement and lower Cohen's kappa scores. Therefore, we provide the average inter-annotator agreement and Cohen's kappa scores in the table 3

A.3 Inconsistent summary annotation

In the first annotation round we asked annotators to mark sentences with unsupported information, i.e., any information not explicitly found in the source, and which could not readily be inferred from the source alone. An example is shown in figure 4a

A.4 Error category annotation

In the second round of annotation, we asked annotators to categorize errors identified in the first round. The options provided are shown in Figure 4b. We map the options to categories as follows

(a) *terms or concepts from the source are misrepresented* are mapped to intrinisc errors

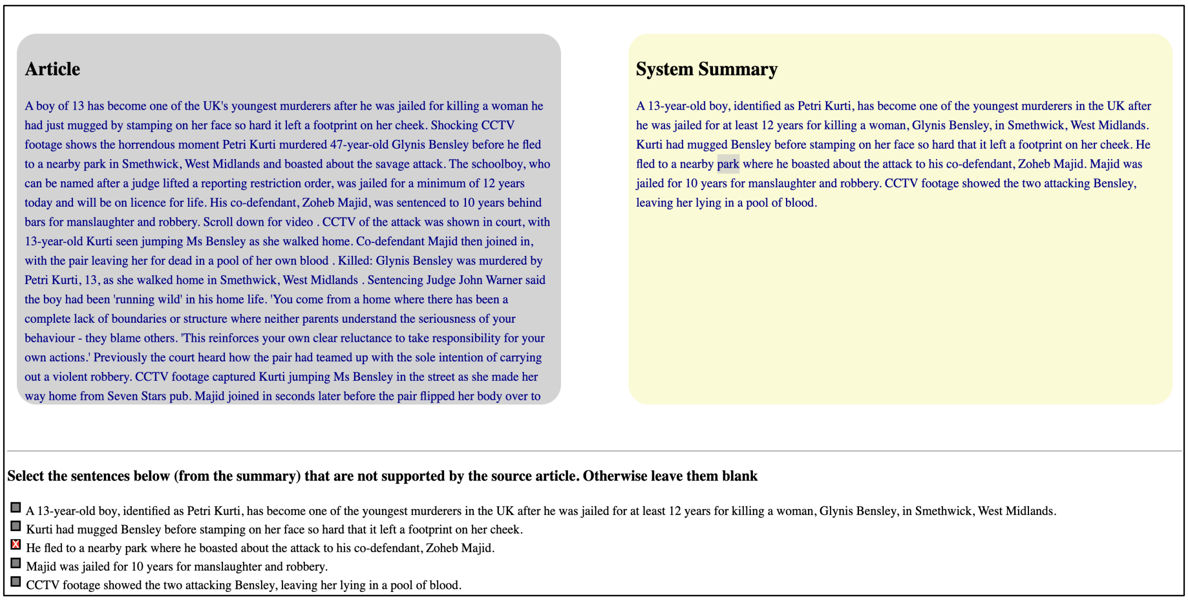
(b) *The information in the summary is not found in the source but can be verified via an internet search as accurate* is mapped to extrinsic (factual) errors

(c) *The information in the summary is not found in the source and can be verified via an internet search as being accurate at a previous time but is outdated* is mapped to extrinsic(factual, outdated) and

(d) *The information in the summary is not found in the source and can not be verified via an internet search* is mapped to extrinsic(nonfactual)

3) *Other* with a mandatory comment.

An example of this round is displayed in Figure 4b



Article

A boy of 13 has become one of the UK's youngest murderers after he was jailed for killing a woman he had just mugged by stamping on her face so hard it left a footprint on her cheek. Shocking CCTV footage shows the horrendous moment Petri Kurti murdered 47-year-old Glynis Bensley before he fled to a nearby park in Smethwick, West Midlands and boasted about the savage attack. The schoolboy, who can be named after a judge lifted a reporting restriction order, was jailed for a minimum of 12 years today and will be on licence for life. His co-defendant, Zoheb Majid, was sentenced to 10 years behind bars for manslaughter and robbery. Scroll down for video . CCTV of the attack was shown in court, with 13-year-old Kurti seen jumping Ms Bensley as she walked home. Co-defendant Majid then joined in, with the pair leaving her for dead in a pool of her own blood . Killed: Glynis Bensley was murdered by Petri Kurti, 13, as she walked home in Smethwick, West Midlands . Sentencing Judge John Warner said the boy had been 'running wild' in his home life. 'You come from a home where there has been a complete lack of boundaries or structure where neither parents understand the seriousness of your behaviour - they blame others. 'This reinforces your own clear reluctance to take responsibility for your own actions.' Previously the court heard how the pair had teamed up with the sole intention of carrying out a violent robbery. CCTV footage captured Kurti jumping Ms Bensley in the street as she made her way home from Seven Stars pub. Majid joined in seconds later before the pair flipped her body over to

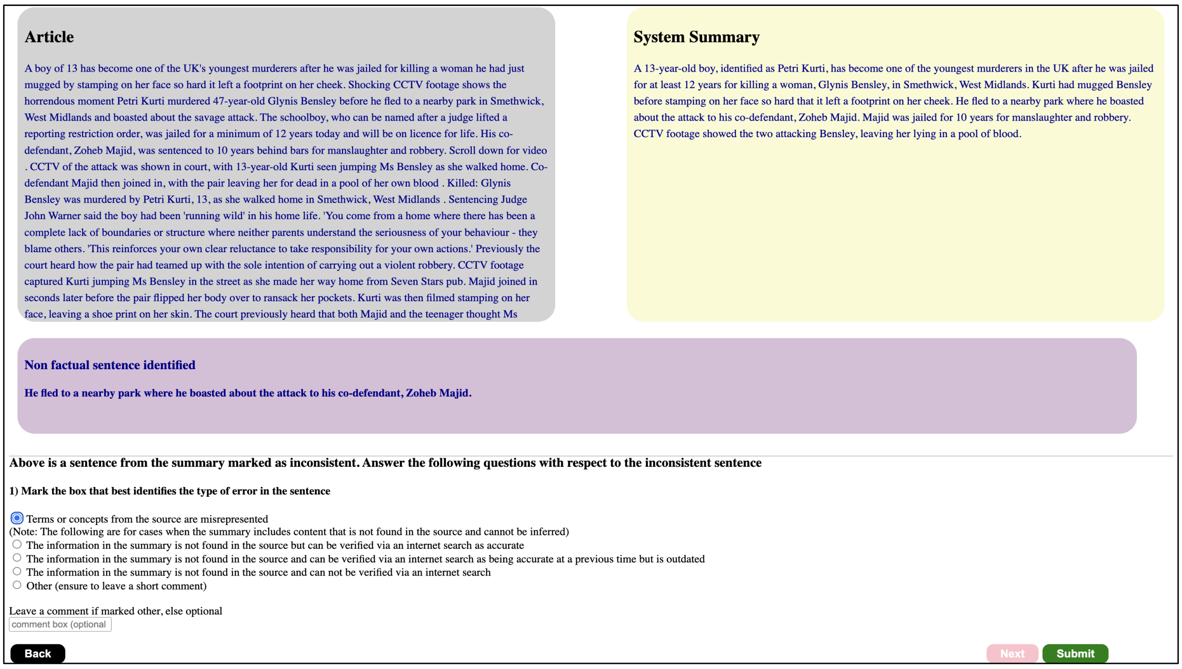
System Summary

A 13-year-old boy, identified as Petri Kurti, has become one of the youngest murderers in the UK after he was jailed for at least 12 years for killing a woman, Glynis Bensley, in Smethwick, West Midlands. Kurti had mugged Bensley before stamping on her face so hard that it left a footprint on her cheek. He fled to a nearby park where he boasted about the attack to his co-defendant, Zoheb Majid. Majid was jailed for 10 years for manslaughter and robbery. CCTV footage showed the two attacking Bensley, leaving her lying in a pool of blood.

Select the sentences below (from the summary) that are not supported by the source article. Otherwise leave them blank

□ A 13-year-old boy, identified as Petri Kurti, has become one of the youngest murderers in the UK after he was jailed for at least 12 years for killing a woman, Glynis Bensley, in Smethwick, West Midlands.
□ Kurti had mugged Bensley before stamping on her face so hard that it left a footprint on her cheek.
☒ He fled to a nearby park where he boasted about the attack to his co-defendant, Zoheb Majid.
□ Majid was jailed for 10 years for manslaughter and robbery.
□ CCTV footage showed the two attacking Bensley, leaving her lying in a pool of blood.

(a) First round example annotation where the third sentence was marked as inconsistent .



Article

A boy of 13 has become one of the UK's youngest murderers after he was jailed for killing a woman he had just mugged by stamping on her face so hard it left a footprint on her cheek. Shocking CCTV footage shows the horrendous moment Petri Kurti murdered 47-year-old Glynis Bensley before he fled to a nearby park in Smethwick, West Midlands and boasted about the savage attack. The schoolboy, who can be named after a judge lifted a reporting restriction order, was jailed for a minimum of 12 years today and will be on licence for life. His co-defendant, Zoheb Majid, was sentenced to 10 years behind bars for manslaughter and robbery. Scroll down for video . CCTV of the attack was shown in court, with 13-year-old Kurti seen jumping Ms Bensley as she walked home. Co-defendant Majid then joined in, with the pair leaving her for dead in a pool of her own blood . Killed: Glynis Bensley was murdered by Petri Kurti, 13, as she walked home in Smethwick, West Midlands . Sentencing Judge John Warner said the boy had been 'running wild' in his home life. 'You come from a home where there has been a complete lack of boundaries or structure where neither parents understand the seriousness of your behaviour - they blame others. 'This reinforces your own clear reluctance to take responsibility for your own actions.' Previously the court heard how the pair had teamed up with the sole intention of carrying out a violent robbery. CCTV footage captured Kurti jumping Ms Bensley in the street as she made her way home from Seven Stars pub. Majid joined in seconds later before the pair flipped her body over to ransack her pockets. Kurti was then filmed stamping on her face, leaving a shoe print on her skin. The court previously heard that both Majid and the teenager thought Ms

System Summary

A 13-year-old boy, identified as Petri Kurti, has become one of the youngest murderers in the UK after he was jailed for at least 12 years for killing a woman, Glynis Bensley, in Smethwick, West Midlands. Kurti had mugged Bensley before stamping on her face so hard that it left a footprint on her cheek. He fled to a nearby park where he boasted about the attack to his co-defendant, Zoheb Majid. Majid was jailed for 10 years for manslaughter and robbery. CCTV footage showed the two attacking Bensley, leaving her lying in a pool of blood.

Non factual sentence identified

He fled to a nearby park where he boasted about the attack to his co-defendant, Zoheb Majid.

Above is a sentence from the summary marked as inconsistent. Answer the following questions with respect to the inconsistent sentence

1) Mark the box that best identifies the type of error in the sentence

● Terms or concepts from the source are misrepresented
(Note: The following are for cases when the summary includes content that is not found in the source and cannot be inferred)
○ The information in the summary is not found in the source but can be verified via an internet search as accurate
○ The information in the summary is not found in the source and can be verified via an internet search as being accurate at a previous time but is outdated
○ The information in the summary is not found in the source and can not be verified via an internet search
○ Other (ensure to leave a short comment)

Leave a comment if marked other, else optional
comment box (optional

(b) Second round of annotation where the annotator marked the category for the inconsistent sentence

Figure 4: Annotation interface with questions asked and example annotation on both round of annotations

	News	Billsum	pubmed
Avg number of source article sentences	26.44	78.41	79.95
Avg number of summary sentences	3.43	3.59	4.01
Avg number of inconsistent summary sentences	0.44	0.38	0.16

Table 2: Data statistics of average number of sentences in the source, summary found in the sampled data. We also include the average number of inconsistent sentences found in summaries of respective domains

Domain	Sentence	Category	Summary
News	0.91 (0.65)	0.86 (0.45)	0.8 (0.56)
Billsum	0.79 (0.17)	0.78 (0.17)	0.72 (0.37)
Pubmed	0.93 (0.11)	0.92 (0.1)	0.85 (0.15)

Table 3: We present inter-annotator agreement metrics for sentences, categories and summaries across diverse domains. Cohen's kappa scores are enclosed in parentheses for each level of annotation, often reflecting lower values. This is primarily attributed to substantial skew in error labels within the dataset, resulting in increased expected chance agreement and consequently lower kappa scores.

A.5 Mixed errors

We highlight some examples of the mixed error category annotations in Figure 5

A.6 Error categories per model

In Figure 6, we present error category distributions for the Flan-T5 and GPT-3.5 models separately. Specifically, for the Flan-T5 model in the news domain, errors are typically categorized as "mixed" or marked as intrinsic and extrinsic errors, with no instances labeled as "other." For both models, the trend shows that intrinsic errors in specialized domains are equal to or higher than those in the news domain.

A.7 Alternative method for domain representation

As an alternative method for evaluating domain representation and its relation to factuality, we use the C4 dataset to query article titles. C4 is a large dataset derived from the the Common Crawl web corpus.[4] It was used to train the T5 Transformer models (Raffel et al., 2020). The number of relevant articles found for each title serves as a proxy for article representation in the training data. We use a C4 search tool to query the C4 dataset.[5] Queries for each article are manually designed using key terms from the article title with the “AND” condition.

Figure 7 demonstrates that queries for news domain retrieved more articles in the C4 dataset compared to Billsum and Pubmed articles.

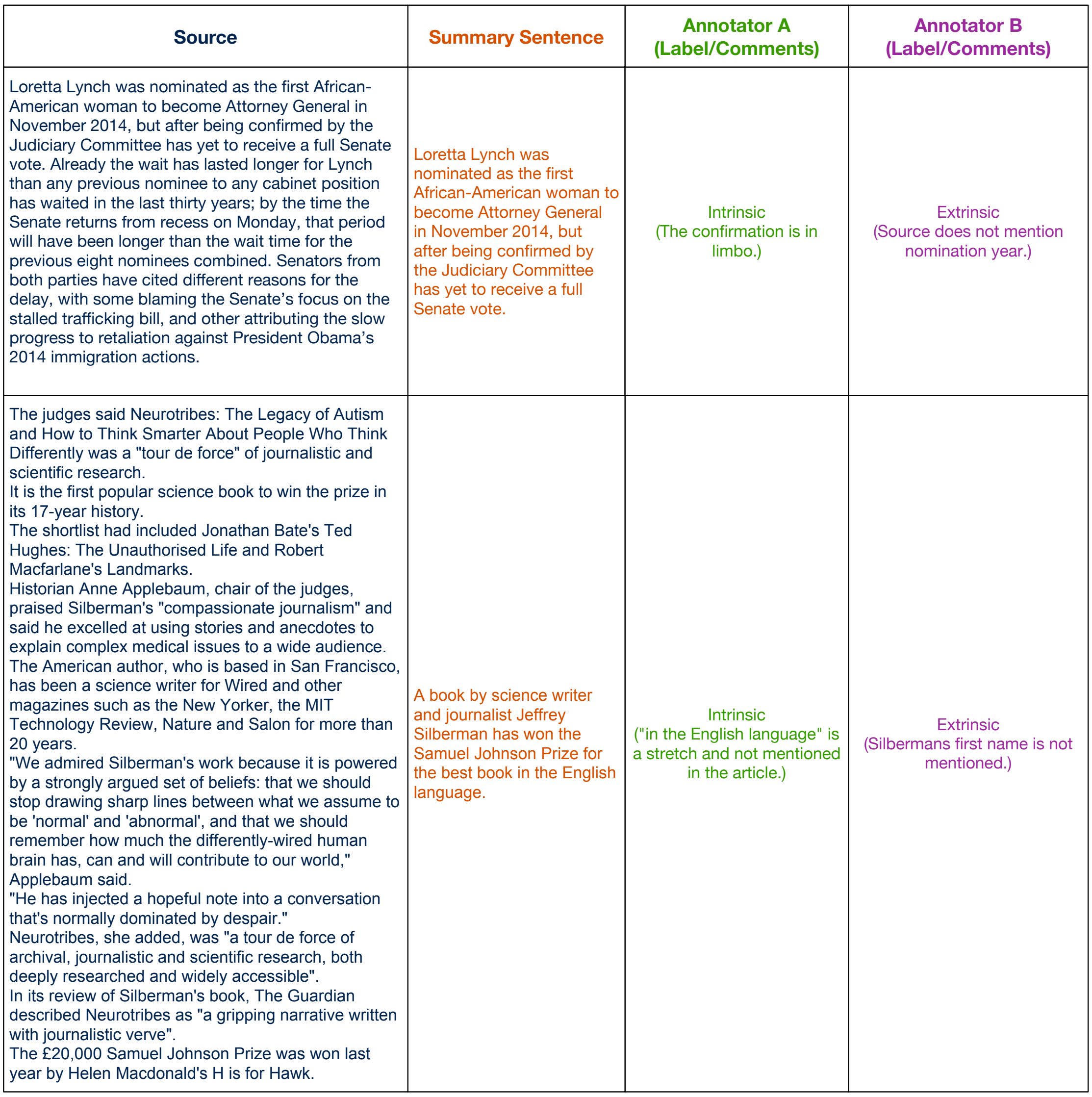
A.8 Model Details

We use the default decoding parameters to generate text from GPT-3.5 and Flan-T5-XL. We use the Huggingface Transformers library [6] to implement Flan-T5-XL.

[4] https://commoncrawl.org

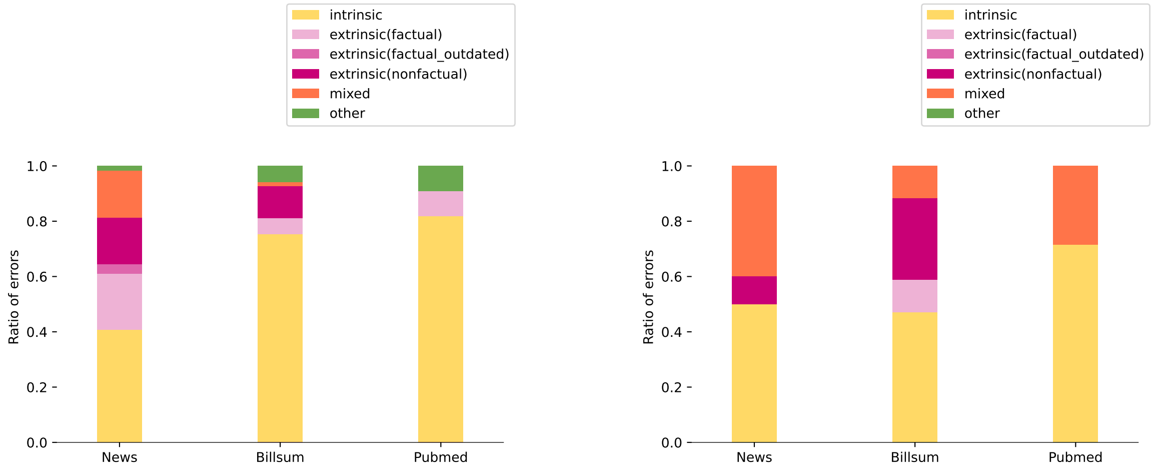
[5] https://c4-search.apps.allenai.org/

[6] https://huggingface.co/



Source	Summary Sentence	Annotator A (Label/Comments)	Annotator B (Label/Comments)
Loretta Lynch was nominated as the first African-American woman to become Attorney General in November 2014, but after being confirmed by the Judiciary Committee has yet to receive a full Senate vote. Already the wait has lasted longer for Lynch than any previous nominee to any cabinet position has waited in the last thirty years; by the time the Senate returns from recess on Monday, that period will have been longer than the wait time for the previous eight nominees combined. Senators from both parties have cited different reasons for the delay, with some blaming the Senate's focus on the stalled trafficking bill, and other attributing the slow progress to retaliation against President Obama's 2014 immigration actions.	Loretta Lynch was nominated as the first African-American woman to become Attorney General in November 2014, but after being confirmed by the Judiciary Committee has yet to receive a full Senate vote.	Intrinsic (The confirmation is in limbo.)	Extrinsic (Source does not mention nomination year.)
The judges said Neurotribes: The Legacy of Autism and How to Think Smarter About People Who Think Differently was a "tour de force" of journalistic and scientific research. It is the first popular science book to win the prize in its 17-year history. The shortlist had included Jonathan Bate's Ted Hughes: The Unauthorised Life and Robert Macfarlane's Landmarks. Historian Anne Applebaum, chair of the judges, praised Silberman's "compassionate journalism" and said he excelled at using stories and anecdotes to explain complex medical issues to a wide audience. The American author, who is based in San Francisco, has been a science writer for Wired and other magazines such as the New Yorker, the MIT Technology Review, Nature and Salon for more than 20 years. "We admired Silberman's work because it is powered by a strongly argued set of beliefs: that we should stop drawing sharp lines between what we assume to be 'normal' and 'abnormal', and that we should remember how much the differently-wired human brain has, can and will contribute to our world," Applebaum said. "He has injected a hopeful note into a conversation that's normally dominated by despair." Neurotribes, she added, was "a tour de force of archival, journalistic and scientific research, both deeply researched and widely accessible". In its review of Silberman's book, The Guardian described Neurotribes as "a gripping narrative written with journalistic verve". The £20,000 Samuel Johnson Prize was won last year by Helen Macdonald's H is for Hawk.	A book by science writer and journalist Jeffrey Silberman has won the Samuel Johnson Prize for the best book in the English language.	Intrinsic ("in the English language" is a stretch and not mentioned in the article.)	Extrinsic (Silbermans first name is not mentioned.)

Figure 5: Examples of sentences annotated with different categories in the news dataset by annotators along with comments provided.



(a) Distribution of error categories across domains for GPT-3.5 model summaries

(b) Distribution of error categories across domains for Flan-T5-XL model summaries

Figure 6: Distribution of error categories across domains per-model

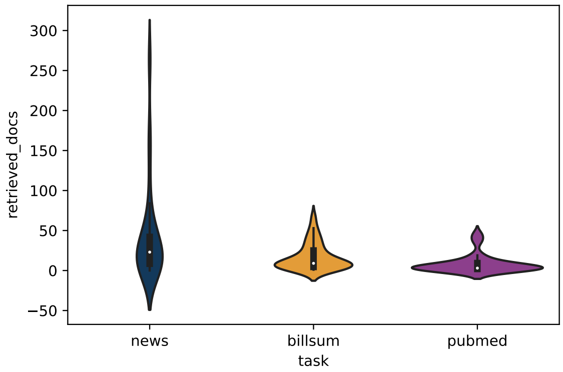


Figure 7: C-4 dataset search results for queries on news, billsum and pubmed articles. The retrieval results show that there is more representation of news articles in the C4 dataset.