On the Benefits of Fine-Grained Loss Truncation: A Case Study on Factuality in Summarization

Anonymous ACL submission

Abstract

Text summarization and simplification are widely used applications of AI. However, such models are often prone to hallucination, which can result from training models on unaligned data. One of the prominent approaches to address this issue has been Loss Truncation (LT) (Kang and Hashimoto, 2020), an approach to modify the standard log loss to adaptively remove noisy examples during training. However, we find that LT alone yields a considerable number of hallucinated entities on various datasets. We study the behavior of the underlying losses between factual and non-factual examples, to understand and refine the performance of LT. We demonstrate that LT's performance is limited when the underlying assumption that noisy targets have higher NLL loss is not satisfied, and find that word-level NLL among *entities* provides better signal for distinguishing factuality. We then leverage this to propose a fine-grained NLL loss and fine-grained data cleaning strategies, and observe improvements in hallucination reduction across some datasets.

1 Introduction

Text summarization and simplification are widely used NLP applications. However, such models are prone to generating hallucinations (Cao et al., 2022a; Zhao et al., 2020; Maynez et al., 2020; Tang et al., 2023); this may have harmful real-world impact and hinder the adoption of such models.

To mitigate hallucinations, previous work studied aspects of training (Choubey et al., 2023), decoding (van der Poel et al., 2022; King et al., 2022; Sridhar and Visser, 2022), or post-processing (Chen et al., 2021). In this paper however, we focus on another large source of hallucination: the data.

When training data is misaligned (i.e. targets contain data unsupported by the input), models learn these patterns and hallucinate (Ji et al., 2023; Dziri et al., 2022). This can stem from data collection errors, or scraping web-based data (Ji et al., 2023). While there have been efforts to identify and clean the misaligned examples (Goyal and Durrett, 2021; Ladhak et al., 2023; Zhou et al., 2021; Adams et al., 2022; Filippova, 2020; Wan and Bansal, 2022), a limitation, however, is that these methods require rewriting targets or training models to detect hallucination.

To this end, other methods automatically detect and remove noisy examples. One widely adopted approach is **Loss Truncation (LT)** (Kang and Hashimoto, 2020), which filters out noisy examples based on the observation that they have higher negative log-likelihood (NLL) loss. This enables an easy-to-adapt and highly efficient training procedure: if NLL loss is high (e.g. >80th quantile of observed losses), do not backpropagate the loss. Previous work adopted this method to improve factuality in summarization (Guo et al., 2021; Ladhak et al., 2022; Cao et al., 2022b; Goyal et al., 2022; Hewitt et al., 2022). However, applying LT to five datasets, we find that models still hallucinate.

In this paper, we study the behavior of NLL at a coarse (i.e. sentence) and fine-grained level (i.e. token) to understand and refine the performance of LT. At the time of writing, the paper is the first to analyze LT on text simplification datasets like Cochrane, MedEasi, and ASSET; moreover, it analyzes the performance of LT from the perspective of factuality, and delves deeper into training dynamics at the token and entity level. Ultimately, the paper aims to contribute a better understanding of the underlying dynamics of LT, that can provide guidance for considerations when using LT in future work, in the context of reducing hallucination.

We make the following contributions: (1) We demonstrate that LT's performance is hindered when the underlying assumption that noisy targets have higher NLL loss is not satisfied, (2) we find that word-level NLL among *entities* provides better signal for distinguishing factuality, and (3) we use this to propose a fine-grained NLL loss which reduces entity-level hallucination on some datasets (-22% on Cochrane, -7.2% on ASSET), and fine-grained data cleaning strategies which achieve up to 26.8% hallucination reduction (CNN-DM), highlighting the potential of this approach.

2 Methodology

Loss Truncation Loss Truncation (Kang and Hashimoto, 2020; Goyal et al., 2022; Cao et al., 2022b) is a widely used method for improving language generation by modifying the standard log loss to adaptively disregard examples with high loss, reducing potential hallucinations. It continuously updates a list of example-level NLL losses, and zeros out losses above a set quantile.[1]

Datasets We study five datasets for two popular conditional NLG tasks, summarization and simplification: **Cochrane** (Devaraj et al., 2021): Medical abstracts from Cochrane Database of Systematic Reviews and expert-written summaries (4,459 pairs), **MedEasi** (Basu et al., 2023): Sentences from Merck Manuals (Cao et al., 2020) and SimpWiki (van den Bercken et al., 2019) and annotated simplifications (1,697 pairs), **ASSET** (Alva-Manchego et al., 2020): Sentences from TurkCorpus dataset (Xu et al., 2016) and simplified versions by 10 annotators (23,590 pairs), **CNN/DailyMail** (Nallapati et al., 2016): Articles and their highlight summaries from CNN and DailyMail (311,971 pairs), **XSum** (Narayan et al., 2018): BBC news articles and their corresponding one-line summaries (226,711 pairs).

Models We finetuned BART-Large-XSUM (Lewis et al., 2020) on five datasets; we chose BART-XSUM to match previous work on Cochrane (Lu et al., 2023; Devaraj et al., 2021), ASSET (Martin et al., 2022), and XSum (Cao et al., 2022b), and isolate the impact of LT (Appendix B). We finetune FlanT5 (Chung et al., 2022) with LT for comparison, and find that it yields similar or better performance (Appendix E).

Metrics We propose a simple definition as our metric of factuality, Hallucination Rate (HR): the % of outputs containing an unsupported entity. We identify entities in outputs using SpaCy `en_core_web_lg` and `en_core_sci_lg` NER models (Honnibal and Montani, 2017; Neumann et al., 2019), then check if *any* of the entities do *not* appear in the input. We also use SARI (Xu et al., 2016), an edit-based text simplification metric, and ROUGE-LSum (Lin, 2004) for overall fluency, to benchmark against previous work, computed using EASSE to align our work with previous methods (Alva-Manchego et al., 2019).

Experimental Set-Up We compare the prevalence of hallucination (i.e. Hallucination Rate) of "coarse" LT (Kang and Hashimoto, 2020) against previous work (Table 1). We then study whether datasets satisfy the assumption of LT by comparing the NLL Loss of non-factual (i.e. containing unsupported entities) vs factual examples (Table 2). We analyze this at a finer granularity, by studying NLL at the token level, both for factual and non-factual sentences (Tables 3, 6). We then propose a "fine-grained LT" and heuristic data cleaning strategies, and compare them to previous work (Table 1).

3 Findings

Noise in summarization can come from adding unsupported information in the reference Our experiments are motivated by the observation that some reference outputs (i.e., gold summaries) contained unsupported information (see Appendix F). E.g., some references in Cochrane had the phrase "*The evidence is current to [date]*", although the date was not mentioned in the input. Upon finetuning, models learn to reproduce this pattern with incorrect dates (Appendix G). Hence, datasets are noisy; a key observation is noise in the reference often involves the *addition* of irrelevant information (Ji et al., 2023). Hence, we limit our definition of "noisy" targets and "hallucination" as containing unsupported data; we then deem references containing entities unsupported by the input as noisy.

LT reduces entity-level hallucination from noisy targets, but not completely We finetune BART-XSum using LT (Appendix B), expecting LT to filter out noisy examples and reduce hallucinations. Comparing Loss Truncation (LT) to previous SOTA in Table 1, LT reduces the proportion of examples containing unsupported (i.e. hallucinated) entities. However, a considerable proportion of examples still contain hallucinations.

We hypothesize LT's performance suffers because the underlying assumption that noisy data has higher NLL is not satisfied We study why LT is unable to weed out many hallucinated entities by comparing models' NLL loss at Epoch 0

[1] We use the official LT package by (Kang and Hashimoto, 2020): https://github.com/ddkang/loss_dropper

Data		Model	HR ↓	SR ↑	RL ↑
Cochrane	Previous	BART XSum FT	69.3%	35.6	44.7
		BART-UL (2021)	69.6%	**40.0**	39.2
		NAPSS (2023)	73.8%	32.9	**45.4**
		LT (Coarse) (2020)	42.7%	36.2	37.6
	Ours	LT (Fine)	**20.6%**	36.1	21.8
		Drop Sentence	42.1%	38.6	33.7
		Drop Example	37.1%	38.5	31.9
MedEasi	Previous	BART XSum FT	35.7%	**40.5**	45.7
		Both-UL (2021)*	13.7%	35.3	**47.9**
		NAPSS (2023)*	42.3%	34.0	24.3
		LT (Coarse) (2020)	**4.6%**	32.6	47.3
	Ours	LT (Fine)	7.0%	37.9	45.1
		Drop Sentence	7.0%	31.8	47.5
		Drop Example	9.7%	38.9	44.4
ASSET	Previous	BART XSum FT	17.0%	38.9	**86.0**
		MUSS NMd (2022)	23.4%	43.6	81.4
		MUSS Md (2022)	31.5%	**44.1**	79.4
		LT (Coarse) (2020)	14.2%	36.7	77.7
	Ours	LT (Fine)	**6.9%**	37.9	45.1
		Drop Sentence	12.8%	40.0	81.7
		Drop Example	22.3%	38.9	85.1
CNN	Previous	BART XSum FT	68.1%	41.4	29.9
		BRIO (2022)	51.9%	**44.9**	**38.3**
		LT (Coarse) (2020)	58.8%	40.7	29.0
	Ours	LT (Fine)	61.3%	41.3	29.7
		Drop Sentence	**32.0%**	42.3	34.5
		Drop Example	66.7%	41.8	30.4
XSum	Previous	BART XSum FT	76.9%	47.6	35.2
		BRIO (2022)	77.1%	**50.6**	**40.1**
		LT (Coarse) (2020)	72.6%	48.1	36.4
	Ours	LT (Fine)	75.5%	47.1	34.5
		Drop Sentence	70.0%	47.2	34.9
		Drop Example	**69.3%**	47.0	34.8

Table 1: Performance on Hallucination Rate (HR), SARI (SR), and ROUGE-LSum (RL), computed using EASSE (Alva-Manchego et al., 2019) from one run; * We finetune these results ourselves on MedEasi; FT: Finetuned, NMd: Not Mined, Md: Mined

(no finetuning), and at Epoch 1 when most models converge (See Appendix C). At Epoch 0, there is no significant difference in the NLL Loss between factual (**NLL (+)**) and non-factual (**NLL (-)**) sentences (Table 2, top). At Epoch 1, non-factual sentences have a higher NLL than factual sentences (Table 2, bottom). In practice however, the difference in NLL is not large enough to cleanly separate factual (orange) from non-factual (blue) examples, as shown in Figure 1. This explains LT's limited performance: the summarization datasets do not meet the assumption that noisy examples' NLL is higher than non-noisy examples, which prevents LT from identifying and removing noisy examples.

Dataset	NLL (-)	NLL (+)	Δ
Cochrane	8.438	9.077	-0.639
MedEasi	11.114	11.173	-0.058
Asset	11.197	11.196	0.002
XSum	19.187	19.190	-0.003
CNN	10.813	10.830	-0.017
Cochrane	0.651	0.437	0.214*
MedEasi	0.080	0.032	0.048*
Asset	0.055	0.034	0.021*
XSum	0.049	0.043	0.006*
CNN	0.134	0.112	0.022*

Table 2: Average NLL Loss for Non-Factual (-) and Factual (+) Examples at Epoch 0 (top) and 1 (bottom), * Indicates the significant difference (One-Way Mann-Whitney Test, $\alpha = 0.05$)

Word-level NLL may better distinguish between factual vs non-factual entities To study the impact of individual words on the overall NLL, we analyze the token-level NLL of targets containing both factual and non-factual entities (i.e. non-factual targets). We make two observations:

First, we find that in non-factual sentences, their non-factual entities (**NLL (-)**) generally have higher NLL than factual entities (**NLL (+)**) (Table 3). Moreover, the difference in NLL (Δ) is larger at the entity level than the sentence level (i.e. compared to the Δ column in Table 2).

Dataset	NLL (0)	NLL (-)	NLL (+)	Δ
Cochrane	8.621	2.445	0.601	1.844*
MedEasi	11.161	2.231	0.772	1.458*
Asset	11.192	2.550	0.664	1.886*
XSum	19.045	1.865	1.934	-0.068*
CNN	10.852	2.910	2.083	0.827*
Cochrane	0.669	1.592	0.331	1.261*
MedEasi	0.078	2.070	0.443	1.626*
Asset	0.051	3.392	0.300	3.092*
XSum	0.048	0.946	1.354	-0.409
CNN	0.128	1.842	1.447	0.395*

Table 3: Average NLL Loss for Non-Entity (0), Non-Factual Entity (-) and Factual Entity (+) Tokens at Epoch 0 (top) and 1 (bottom), * Indicates the significant difference (One-Way Mann-Whitney Test, $\alpha = 0.05$)

Upon comparing factual versus non-factual sentences (Table 6), it still holds that the NLL of factual entities is lower the NLL of non-factual entities (Table 3). In short, non-factual tokens have higher NLL than factual tokens, regardless of which sentences those factual tokens appear in.

Second, the NLL of non-entity tokens significantly impacts the overall sentence NLL, and obscures the signal between factual and non-factual entities. This is shown by the fact that non-entity

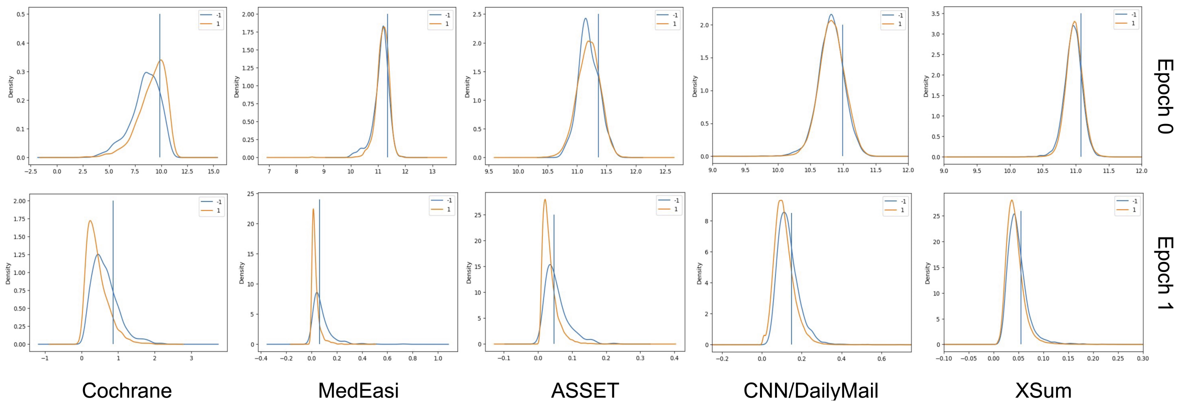

Figure 1: NLL distribution of factual (Orange) and non-factual (Blue) targets shows that there no difference at epoch 0, and a slight difference at epoch 1, with non-factual entities having slightly higher NLL (shifted to the right)

NLL values closely mirror the sentence-level NLLs (Table 2, NLL (-)). Intuitively it makes sense: there are more non-entities than entities, so they have a larger impact on sentence-level NLL.

Considering this, it may be beneficial to focus on the word-level NLL as it may offer a more nuanced view of factual versus non-factual entities, while also not giving too much weight to non-entities.

We aim to reduce hallucination with two methods: (1) a fine-grained LT, and (2) data cleaning strategies using fine-grained information We first propose a fine-grained LT: instead of using sentence-level NLL in LT, we sum the NLL *only* for entity tokens (Appendix B for details). This leverages the fact that entity tokens provide better signal for factuality than non-entity tokens, and that non-factual entities have higher NLL.

Fine-grained LT reduces HR on Cochrane (-22%) and ASSET (-7.2%) compared to coarse LT (Table 1). However, its performance is not as competitive on MedEasi, CNN, and XSum. We observe that unlike Cochrane and ASSET which are human annotated, the three datasets are web-scraped, and more noisy. We confirm this by measuring HR on the labels; labels from the three web-scraped dataset contained more hallucinated entities than the human annotated ones (Table 5, Appendix F for examples). Therefore, we suspect these datasets require a more aggressive strategy to eliminate noise.

To this end, we propose to directly clean the dataset, filtering out noisy targets. We identify all unsupported entities in a target (i.e. the entity is not in the input); then we either (1) drop *only* the sentence containing the entity (Drop Sentence), or (2) drop the entire example (Drop Example) (See Appendix A for stats). Table 1 shows that at least one of the strategies results in lower hallucination rate for CNN (-26.8%, Drop Sentence) and XSum (-3.3%, Drop Example). While MedEasi is the only dataset where our methods do not outperform the baselines, the hallucination reduction rate is still competitive when dropping noisy examples. Overall, with the exception of the MedEasi dataset, our results show strong improvements over the baseline methods, suggesting the potential of the fine-grained LT and fine-grained data cleaning in reducing hallucinations.

4 Conclusion

We analyzed the effect of loss truncation (LT) on improving factuality in text summarization. We found that LT struggles to reduce entity-level hallucination when the underlying assumption that non-factual sentences have higher NLL than factual sentences is not met. To this end, we explore a token-level loss truncation (i.e. fine-grained LT) and simple entity-level dataset cleaning strategies, which reduce the prevalence of hallucination across various summarization and simplification datasets.

Future work may explore other signals for noise in training data. Moreover, future work can explore contradictory information (i.e. targets with similar topics as input but different meaning). This requires the use of natural language inference (NLI), which we qualitatively find is difficult in practice using off-the-shelf NLI models (Wu et al., 2022) or GPT (Liu et al., 2023), as we observe they are currently unable to detect contradictory or unsupported information in some cases. Ultimately, reducing such hallucinations is key to improving the overall performance of summarization models.

Limitations

One limitation of our paper is that we limit the definition of hallucination to the addition of unsupported entities, while the detection of contradictory or omitted information are equally important to detect. A key challenge with such definitions of hallucination is that they require human annotations or good models to identify targets in the dataset which contain contradictory or omitted information. We previously experimented with using GPT-4 following the GPT-Eval framework (Liu et al., 2023), but found that GPT was sometimes unable to detect unsupported information. For example, GPT was unable to identify that the date in the Cochrane dataset targets were unsupported.

Another limitation is that loss truncation at the token level does not always achieve the best results. While it reduced entity-level hallucination for Cochrane and ASSET compared to other methods, it fails to achieve substantial improvements on MedEasi, CNN, and XSum. Overall, the paper aims to show that the method has potential in some cases, but future work can explore other ways to improve its performance.

Finally, it should be noted that our work has been tested on a limited number of general domain summarization datasets; hence more work can explore a wider set of datasets in various niches, to examine if larger patterns across datasets impact the performance of loss truncation.

Risks It should be noted that even data cleaning and LT (both coarse and fine-grained) does not fully reduce entity-level hallucination. Moreover, we have not studied other types of hallucination in this work. Therefore, these models are *not* ready-to-use, and should not be deployed readily.

References

Griffin Adams, Han-Chin Shing, Qing Sun, Christopher Winestock, Kathleen McKeown, and Noémie Elhadad. 2022. Learning to revise references for faithful summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4009–4027, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, and Qian Yang. 2023. Med-easi: Finely annotated dataset and models for controllable simplification of medical texts.

Meng Cao, Yue Dong, and Jackie Cheung. 2022a. 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, Jingyi He, and Jackie Chi Kit Cheung. 2022b. Learning with rejection for abstractive text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9768–9780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1061–1071, Online. Association for Computational Linguistics.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941, Online. Association for Computational Linguistics.

Prafulla Kumar Choubey, Alex Fabbri, Jesse Vig, Chien-Sheng Wu, Wenhao Liu, and Nazneen Rajani. 2023. CaPE: Contrastive parameter ensembling for reducing hallucination in abstractive summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10755–10773, Toronto, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870, Online. Association for Computational Linguistics.

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, Jiacheng Xu, Junyi Jessy Li, and Greg Durrett. 2022. Training dynamics for text summarization models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2061–2073, Dublin, Ireland. Association for Computational Linguistics.

Shiqi Guo, Jing Zhao, and Shiliang Sun. 2021. Resilient abstractive summarization model with adaptively weighted training loss. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Daniel Kang and Tatsunori B. Hashimoto. 2020. Improved natural language generation via loss truncation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 718–731, Online. Association for Computational Linguistics.

Daniel King, Zejiang Shen, Nishant Subramani, Daniel S. Weld, Iz Beltagy, and Doug Downey. 2022. Don't say what you don't know: Improving the consistency of abstractive summarization by constraining beam search. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 555–571, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, and Tatsunori Hashimoto. 2023. Contrastive error attribution for finetuned language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11482–11498, Toronto, Canada. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness tradeoff in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1410–1421, Dublin, Ireland. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Junru Lu, Jiazheng Li, Byron Wallace, Yulan He, and Gabriele Pergola. 2023. NapSS: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1079–1091, Dubrovnik, Croatia. Association for Computational Linguistics.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Arvind Krishna Sridhar and Erik Visser. 2022. Improved beam search for hallucination mitigation in abstractive summarization.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Laurens van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, WWW '19, page 3286–3292, New York, NY, USA. Association for Computing Machinery.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Wan and Mohit Bansal. 2022. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1028, Seattle, United States. Association for Computational Linguistics.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676, Dublin, Ireland. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

A Dataset Information

We report the counts of training examples pre and post dataset cleaning in Table 4.

Dataset	Original	Drop Sentence	Drop Example
Cochrane	3568	3479	245
MedEasi	1397	907	857
ASSET	20000	18690	18229
CNN	287113	285160	187465
XSum	204045	110754	110745

Table 4: Number of training examples from data cleaning methods; Drop Sentence results in minor reductions whereas Drop Example results in larger reductions

Licenses The Cochrane dataset uses the C.C. BY 4.0 License; MedEasi and XSum use the MIT License; ASSET uses the CC BY-NC 4.0 License, and CNN/DailyMail uses the Apache 2.0 License.

Quantifying Noisiness of Datasets We run the Hallucination Rate computation on 100 labels in each of the datasets, to quantify how noisy these labels are in reference to their inputs. Note that for Cochrane, we manually reclassified examples as the medical NER models used for this dataset identified common words as entities (e.g. disease, operation), which were correct based on the input. In cases when we were unsure whether a term was an abbreviation or synonym of another term, we marked it as a hallucination, to provide a conservative estimate. Hence, Cochrane's HR may actually be lower (i.e. better) than reported.

Dataset	HR ↓
Cochrane	68/100
ASSET	14/100
MedEasi	80/100
CNN	74/100
XSum	83/100

Table 5: Noisiness of datasets measured using 100 examples' hallucination rate (HR)

B Training Details

Implementation Details We run our experiments on 1 NVIDIA RTX 6000 GPU. Finetuning each model on Cochrane, MedEasi, and ASSET, for base, coarse and fine-grained LT, and with cleaned datasets, takes roughly 40 minutes, whereas CNN/DailyMail and XSum take 4 hours.

Finetuning All models use 1 epoch, a learning rate of 5e-5, Adam epsilon of 1e-8, and batch size of 1 for Cochrane/MedEasi and 64 for ASSET, XSum, CNN/DailyMail).

Loss Truncation (Coarse-Grained) All datasets are trained using a 80% truncate rate, with a cutoff recomputed every 1000 examples.

Loss Truncation (Fine-Grained) Cochrane and MedEasi use an 80% truncate rate, whereas ASSET, XSum, and CNN/DailyMail use a 40% truncate rate, all recomputing every 500 examples.

The score used in the fine-grained LT is given by

$$\text{score}(\hat{y}) = \sum_{t=1}^{|y|} \mathbb{1}[y_t \in \text{entities}] \cdot y_t \log(\hat{y}_t)$$

where $\mathbb{1}[y_t \in \text{entities}]$ is scored by NER models `en_core_web_lg` and `en_core_sci_lg` (Honnibal and Montani, 2017; Neumann et al., 2019) and $\hat{y}_t = p(y_t|y_{<t}, X)$.

C Training Loss Curves

We plot loss curves generated from finetuning BART-XSum in Figure 2 throughout one epoch which demonstrates convergence across datasets.

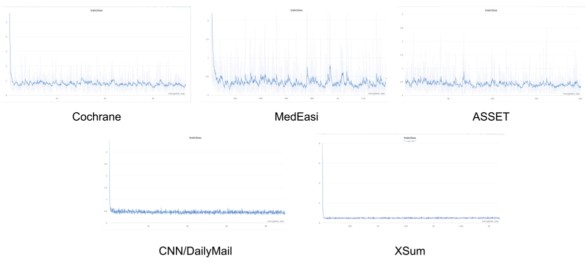

Figure 2: Loss curves from finetuned BART-XSum; 0.8 smoothing used in top row

D NLL of Factual/Non-Factual Tokens

We compare the NLL of factual and non-factual tokens in factual and non-factual sentences in Table 6. This demonstrates that non-factual tokens have higher NLL than factual tokens, regardless of which sentences the tokens appear in.

Dataset	NLL (+, NF)	NLL (+, F)	NLL (-)
Cochrane	0.601	0.522	2.445
MedEasi	0.772	0.510	2.231
Asset	0.664	0.752	2.550
XSum	1.934	2.579	1.865
CNN	2.083	2.199	2.910
Cochrane	0.331	0.265	1.592
MedEasi	0.443	0.228	2.070
Asset	0.300	0.825	3.392
XSum	1.354	1.776	0.946
CNN	1.447	1.488	1.842

Table 6: Token-Level NLL Loss for Factual Entities in both Non-Factual Targets (+, NF) and Factual Targets (+, F), and Non-Factual Entities in Non-Factual Targets (-)

E Results on Flan-T5

We report the details of finetuning the standard loss truncation (Kang and Hashimoto, 2020) using Flan-T5 (Chung et al., 2022) in Table 7.

Data	HR (Entity) ↓	SARI ↑	RL ↑
Cochrane	190/480 (39.6%)	33.720	37.163
MedEasi	14/300 (46.7%)	24.405	48.248
ASSET	19/359 (5.3%)	35.003	91.116
CNN	2948/11490 (25.7%)	41.486	32.133
XSum	6897/11334 (60.9%)	43.767	29.130

Table 7: Finetuning Flan-T5 (Chung et al., 2022) with Loss Truncation results in even better performance than BART, demonstrating opportunity for further progress

F Examples of Noisy Targets

See Table 8 for examples of noisy targets from various datasets.

G Example Output

See Table 9 for a comparison of outputs of various models from an example in the Cochrane dataset. Loss truncation and the example-level data cleaning are the only methods which correctly avoid generating a hallucinated date.

Dataset	Input	Target
MedEasi	Baker cysts may form and rupture.	Cysts may develop and rupture **behind the knees, suddenly increasing the pain**.
	Sullivan apparently had no idea who McCartney was.	Sullivan thought that **his illness was because of ulcers**.
	The linear combination of atomic orbitals or LCAO approximation for molecular orbitals was introduced in 1929 by Sir John Lennard-Jones.	The **LCMO (Linear combination of atomic orbitals molecular orbital)** method gives a rough but good description of the MOs
Cochrane	We included six trials, involving a total of 636 women with a twin or triplet pregnancy (total of 1298 babies). We assessed all of the included trials as having a low risk of bias for random sequence generation. ... There is a need for large-scale, multicenter randomised controlled trials to evaluate the benefits, adverse effects and costs of bed rest before definitive conclusions can be drawn.	**We searched for evidence on 30 May 2016.** We identified six randomised controlled trials involving a total of 636 women and 1298 babies ... We did not find sufficient evidence to support or refute bed rest for women with a multiple pregnancy as a way of preventing preterm birth and other pregnancy complications.
	This update identified one additional study for inclusion, adding data for 2305 participants. This addition more than doubled the overall number of patients eligible for the review. ... Also, there were no data from RCTs on the utility of non-vitamin K antagonist oral anticoagulants compared to antiplatelet agents in heart failure with sinus rhythm.	This is an update of an earlier review. **The evidence is current to September 2015.** We only identified one new study with 2305 participants. ... It is unlikely that further studies will change these conclusions unless new, more effective and safe drugs become available.
XSum	The full cost of damage in Newton Stewart, one of the areas worst affected, is still being assessed. Repair work is ongoing in Hawick and many roads in Peeblesshire remain badly affected by standing water.... Have you been affected by flooding in Dumfries and Galloway or the Borders?	Clean-up operations are continuing across the Scottish Borders and Dumfries and Galloway after flooding caused by **Storm Frank**.
	A fire alarm went off at the Holiday Inn in Hope Street at about 04:20 BST on Saturday and guests were asked to leave the hotel. As they gathered outside they saw the two buses, parked side-by-side in the car park, engulfed by flames.... While the exact cause is still under investigation, it is thought that the fire was started deliberately.	Two tourist buses have been destroyed by fire in a suspected arson attack in **Belfast city centre.**

Table 8: Examples of noisy targets from datasets, with the unsupported entities presented in bold

Model	**Example**
BART (Lewis et al., 2020)	We found one small randomised controlled trial (with data from 34 women) that compared transcervical amnioinfusion with no treatment. The trial was considered to be at a high risk of bias overall… **The evidence is up-to-date as of February 2019.**
Both-UL (Devaraj et al., 2021)	We found one small randomised controlled trial with 34 women… **The evidence is current up-to-date to September 2016**… The numbers included were too small for meaningful assessment of substantive outcomes, where they were reported.
NAPSS (Lu et al., 2023)	**The evidence is current to May 2017.** We included one small study that reported on a limited number of outcomes of interest in this review. ... and inconsistency in how the trial was conducted and reported. **The evidence is up to date as of March 2017.** There is insufficient evidence to …
LT (Sentence Level) (Kang and Hashimoto, 2020)	This review did not identify any trials that used transabdominal amnioinfusion outside of clinical trials. The evidence in this review can neither support nor refute the use of transcervical amnion infusions for chorioamnionitis and to assess the safety of this intervention or women's satisfaction.
LT (Token Level)	We included one small trial (with data from 34 women) comparing transcervical amnioinfusion with no amnioticinfusion. The trial was considered to be at a high risk of bias overall, inconsistency in the reporting and lack of information on blinding. We did not identify any trials that used transabdominal amniotics (low-quality evidence). We assessed using GRADE (postpartum endometritis, neonatal encephalopathy, admission to intensive/high care) was not reported in the included trial.
Drop Sentence	**We searched for evidence from randomised controlled trials on 31 May 2017. The evidence in this review is up-to-date at 31 December 2017.** … The overall quality of the evidence was low, with small numbers of women in the trial and a lack of information on blinding.
Drop Example	We found one trial of transcervical amnioinfusion, with data from 34 women who matched our question. The trial was considered to be at a high risk of bias overall, due to small numbers, inconsistency in the reporting and lack of information on blinding. Meta-analysis of the evidence was not possible. The majority of trials in this review were of low quality. For these outcomes, we downgraded the evidence to low - with downgrading decisions based on small numbers and a lack of instructions on how to be sure of whether they were in the intervention group.

Table 9: Sample Report from the Cochrane Test Set