Revisiting Reconstruction Loss for Text Summarization in Small Language Models

Anonymous ACL submission

Abstract

Text summarization requires not only lexical quality, but also factual consistency with the source. While standard fine-tuning objectives based on token-level cross-entropy provides a strong generation signal, it does not explicitly encourage source grounding, and two-stage reranking methods such as SimCLS introduce additional inference overhead. To address this limitation, we revisit the autoencoder framework and propose **ReFS** (**Re**construction-Refinement for **F**actual **S**ummarization), a training-time regularizer for small language models (SLMs) that incorporates source reconstruction within a single-stage fine-tuning setting. Specifically, the hidden state is iteratively refined through sequential decoder passes before reconstructing the source document, encouraging the model to encode richer source-faithful representations. Experiments on CNN/DailyMail and XSum with two SLMs demonstrate consistent factuality improvements with competitive lexical quality and zero inference overhead. ReFS further yields additional factuality gains when combined with SimCLS reranking, demonstrating compatibility with two-stage approaches.

1 Introduction

Text summarization aims to generate concise summaries that are both fluent and informative. While ROUGE (Lin, 2004) remains the dominant metric for evaluating n-gram overlap, lexical quality alone does not guarantee reliable output. Factual consistency with the source is equally critical (Guo et al., 2022), as even fluent summaries can introduce hallucinated or misleading content (Maynez et al., 2020; Kryściński et al., 2020). To address this concern, recent studies increasingly adopt dedicated factuality metrics (Laban et al., 2022; Yang et al., 2024).

Two-stage reranking methods such as SimCLS (Liu and Liu, 2021) improve summary quality by selecting the best output from a generated candidate pool using a learned scorer. BRIO (Liu et al., 2022) further shows that contrastive ranking objectives can align model probabilities with automatic quality metrics, offering a potential route to incorporate factuality-oriented criteria. Nevertheless, post-hoc reranking methods require multiple candidate generations and an additional scoring step at inference time, making them less attractive for cost-sensitive deployment (Zhong et al., 2020).

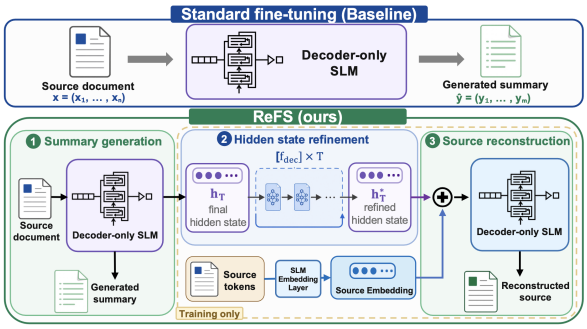

Figure 1: Overview of our ReFS compared to standard fine-tuning. $\oplus$ denotes concatenation along the sequence dimension.

Meanwhile, reconstruction-based objectives have been explored in summarization to induce information-preserving compressed representations (Miao and Blunsom, 2016; Févry and Phang, 2018; Baziotis et al., 2019). However, they mainly target compression quality rather than faithfulness and typically assume encoder-decoder or modular bottleneck architectures.

We therefore investigate whether decoder hidden states in decoder-only SLMs can serve as an implicit bottleneck for source reconstruction. To this end, we propose **ReFS** (**Re**construction-Refinement for **F**actual **S**ummarization), revisiting the autoencoder framework and adapting reconstruction-based objectives to decoder-only SLMs in a single-stage training setting. ReFS incorporates source reconstruction as a training-time

regularizer, where the hidden state is iteratively refined through sequential decoder passes before reconstructing the source document, encouraging the model to retain richer source-faithful representations. This iterative refinement is inspired by the latent reasoning paradigm (Hao et al., 2024), which we adapt to promote source-faithful encoding rather than chain-of-thought reasoning. As refinement and reconstruction are restricted to training, ReFS introduces no additional inference-time overhead.

To validate ReFS, we conduct experiments on CNN/DailyMail (Nallapati et al., 2016) and XSum (Narayan et al., 2018) using two SLMs. ReFS achieves consistent factuality improvements with competitive lexical quality, and further demonstrates compatibility with two-stage approaches through additional gains when combined with SimCLS reranking.

2 Factual Summarization via Reconstruction and Refinement

The overview of ReFS is presented in Figure 1. It comprises three training-only phases: summary generation, hidden state refinement, and source reconstruction, with inference restricted to the first phase alone.

Summary Generation. Given a source document $\mathbf{x} = (x_1, \ldots, x_n)$ and a reference summary $\mathbf{y} = (y_1, \ldots, y_m)$, we perform a standard autoregressive forward pass through a SLM. From this forward pass, we extract the final hidden state $\mathbf{h}_T$, which encodes both the source document and generated summary.

Hidden State Refinement. A single forward pass encodes both the source and the generated summary into $\mathbf{h}_T$, but this representation may not sufficiently capture document-level content for faithful reconstruction. To enrich it, we introduce a hidden state refinement process that iteratively refines it through T sequential decoder passes, where $T = m$ is dynamically set to the summary length. At each step, the current hidden state is passed through the decoder without invoking the language model head, operating entirely in hidden space. After T steps, the refined hidden state $\mathbf{h}_T^*$ encodes richer source-grounded information, enabling more faithful reconstruction of the original document.

Rather than relying on a fixed step count that requires heuristic tuning, we tie T to the summary length, naturally scaling reasoning capacity to generation complexity without introducing an additional hyperparameter (Hao et al. 2024).

Source Reconstruction. The refined hidden state $\mathbf{h}_T^*$ is concatenated to the source token embeddings and passed through the decoder in parallel via teacher forcing, where the reconstruction loss is computed as cross-entropy over the source tokens.

The final training objective adds this reconstruction term to the summary generation loss, normalizing it by the summary length T to stabilize training across varying sequence lengths.

3 Experiments

3.1 Experimental Settings

Datasets. We evaluated our method on two summarization benchmarks: CNN/DailyMail (`CNN/DM`) (Nallapati et al., 2016) and XSum (`XSum`) (Narayan et al., 2018). Detailed dataset statistics are provided in Appendix A. We sample 8k training instances with a fixed random seed, and evaluate on a fixed set of 1,000 test instances.

Evaluation Metrics. For text quality, we evaluated outputs using ROUGE-1 (**R-1**), ROUGE-2 (**R-2**), ROUGE-L (**R-L**) (Lin, 2004), and BERTScore (**BS**) (Zhang et al., 2020). For factuality, we evaluated on **SummaC** (**SC**) (Laban et al., 2022), **AlignScore** (**AS**) (Zha et al., 2023), **FactCC** (**FC**) (Kryściński et al., 2020), **FIZZ** (**FI**) (Yang et al., 2024). Detailed evaluation metrics are provided in Appendix B.

Implementation Details. We fine-tuned Llama-3.2-3B-Instruct (Grattafiori et al., 2024) and Phi-3.5-mini-Instruct (Abdin et al., 2024) using QLoRA (Dettmers et al., 2023) with 4-bit NF4 quantization and LoRA (Hu et al., 2022) adapters on the query and value projection. All methods were trained with length priming prompt (Juseon-Do et al., 2024), with only LoRA adapter parameters updated and base model parameters kept frozen. At inference, models were merged and unloaded in BF16 precision. Detailed implementation details are provided in Appendix C.

Compared Methods. We compare against standard summarization fine-tuning (**Baseline**) and **SimCLS** (Liu and Liu, 2021), a two-stage contrastive reranking method. For our SimCLS implementation, we rerank candidates using FactCC. To avoid evaluation bias, we omit FactCC scores for +SimCLS variants. We further evaluate **ReFS (Head)**, a lightweight variant that reconstructs the source directly from $\mathbf{h}_T$ through the language

Model	Method	CNN/DailyMail								XSum							
		R-1	R-2	R-L	BS	SC	AS	FC	FI	R-1	R-2	R-L	BS	SC	AS	FC	FI
Llama 3.2-3B	Baseline	**43.82**	**19.79**	29.77	**63.65**	69.4	87.80	43.10	62.23	**38.80**	**15.36**	**30.52**	**63.09**	24.0	43.19	25.09	18.74
	ReFS (Head)	43.54	19.65	**29.92**	63.58	70.9†	89.22*	45.03*	64.27†	38.38	14.90	29.81	62.61	**24.2**	43.74	25.55	**20.41**†
	ReFS	43.18	19.55	29.47	63.25	**72.0***	**89.96***	**48.42***	**67.25***	38.08	14.98	29.60	62.17	24.1	**43.90**	**26.07**	19.70
	+SimCLS (Base)	**42.80**	19.05	28.83	**62.85**	75.2	92.55	-	66.47	**38.72**	**15.16**	**30.05**	**63.24**	23.8	40.06	-	16.95
	+SimCLS (Head)	42.75	**19.09**	**28.85**	62.85	76.5†	93.01	-	68.24†	37.38	14.17	28.71	61.94	23.7	**41.77**	-	**18.22**†
	+SimCLS (ReFS)	42.43	18.78	28.38	62.42	**77.2***	**93.29**†	-	**71.13***	37.64	14.86	29.16	62.24	**23.9**	39.94	-	17.92
Phi 3.5-mini	Baseline	43.23	19.49	**29.34**	63.58	74.6	90.03	48.34	60.99	**38.85**	**15.09**	**30.42**	**63.50**	24.3	47.02	26.65	20.93
	ReFS (Head)	42.91	19.08	29.01	63.42	74.5	90.34	47.20	63.27*	38.24	15.00	30.04	63.08	24.6	47.99	25.73	23.68*
	ReFS	**43.24**	**19.50**	29.28	**63.75**	**81.8***	**94.91***	**61.03***	**70.78***	38.07	14.68	29.89	62.93	**24.7**†	**49.00**†	**26.97**	**24.12***
	+SimCLS (Base)	**42.96**	**19.04**	28.79	**62.98**	79.3	94.16	-	68.73	**38.10**	**14.62**	**29.40**	**63.34**	**24.2**	43.57	-	18.74
	+SimCLS (Head)	42.89	18.97	**28.84**	62.97	78.8	94.13	-	69.61†	37.38	13.96	28.89	62.72	24.1	43.40	-	19.48†
	+SimCLS (ReFS)	42.58	**19.14**	28.44	62.90	**85.5***	**96.45***	-	**75.68***	37.40	13.77	28.65	62.32	24.1	**44.65**	-	**20.25***

Table 1: Main results on CNN/DM and XSum. **Bold** denotes the best score within each comparison group (Baseline vs. ReFS (Head) vs. ReFS; +SimCLS (Base) vs. +SimCLS (Head) vs. +SimCLS (ReFS)). Underline indicates improvement over the Baseline. Light gray and dark gray denote cells where ReFS (Head) and ReFS outperform the Baseline, respectively. Statistical significance is assessed via paired bootstrap resampling (Koehn, 2004) (100,000 samples): * $p<0.01$, † $p<0.05$.

model head without iterative refinement. **ReFS** denotes our full method with iterative hidden-state refinement introduced in Section 2.

3.2 Main Results

As shown in Table 1, we report 8K-scale results for both models on CNN/DM and XSum.

CNN/DM. ReFS consistently improves factuality across both models, while ReFS (Head) yields more modest and occasionally inconsistent gains, suggesting that iterative hidden state refinement is the primary driver of factuality improvement. ROUGE and BERTScore remain largely comparable to the baseline, indicating that the factuality gains do not come at a substantial cost to lexical quality. When combined with SimCLS reranking, +SimCLS (ReFS) performs comparably to or better than +SimCLS (Base) across most factuality metrics, demonstrating compatibility with two-stage reranking.

XSum. ReFS improves FI across both models, although this improvement is accompanied by a moderate decline in ROUGE and BERTScore. This trade-off is structurally plausible: XSum references are highly abstractive and often lexically distant from the source, creating a competing signal against the reconstruction objective (Dou et al., 2021). A similar trend is observed under SimCLS reranking: +SimCLS (ReFS) improves FI over +SimCLS (Base), while ROUGE and BERTScore show moderate declines.

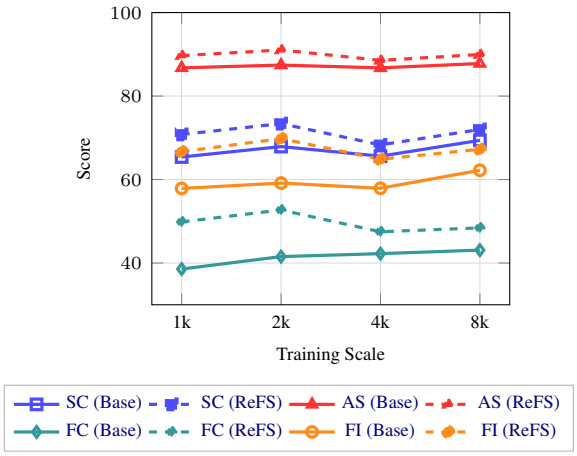

Figure 2: Factuality metrics of ReFS vs. Baseline across training scales.

4 Analysis

Effect of Training Scale on Factuality. As shown in Figure 2, ReFS consistently outperforms the baseline across all training scales on CNN/DM using Llama-3.2-3B-Instruct, with a stable performance gap. The improvement is most pronounced in FC and FI, while AS and SC exhibit smaller but consistent gains. Both configurations show a slight performance dip at the 4k scale before recovering at 8k. This suggests that source reconstruction provides a robust factuality signal across different data regimes. Full results across all scales and both models are provided in Appendix E.

Ablation Study. To isolate the contribution of hidden-state refinement, we compare two variants: **ReFS (w/o HR)**, which performs reconstruction with a single decoder pass without iterative refine-

Method	R-1	R-2	R-L	BS	SC	AS	FC	FI
Baseline	**43.82**	**19.79**	29.77	**63.65**	69.4	87.80	43.10	62.23
ReFS (w/o HR)	43.18	19.37	29.47	63.46	69.7	88.66	43.77	63.95
ReFS	43.18	19.55	29.47	63.25	**72.0***	**89.96***	**48.42***	**67.25***

Table 2: Ablation on CNN/DM (8k) using Llama-3.2-3B-Instruct. * denotes statistically significant (p<0.01) improvements, compared to the underlined scores.

Dataset	Method	R-1	R-2	R-L
CNN/DM	Baseline	34.76	10.16	16.05
	ReFS	**40.39***	**11.86***	**17.58***

Table 3: ROUGE scores for source-document reconstruction using Llama-3.2-3B-Instruct. The notation follows Table 2.

ment, and **ReFS**, which applies T sequential decoder passes before reconstruction. As shown in Table 2, ReFS consistently outperforms ReFS (w/o HR) across factuality metrics while maintaining comparable ROUGE scores, confirming that iterative refinement is important for improving factuality.

Does ReFS Improve Source Reconstruction? To assess source reconstruction quality, we prompt the trained model to reconstruct the source document from a given summary and evaluate the overlap between the generated and reference articles using ROUGE. As shown in Table 3, ReFS consistently achieves higher reconstruction scores on CNN/DM, suggesting that its reconstruction-based training encourages the model to retain richer document-level information. Prompt templates are provided in Appendix F.

Case Study. Table 4 presents a representative example from CNN/DM. The Baseline summary includes information that is inconsistent with the source article, whereas ReFS better preserves the source content. This example suggests that the reconstruction objective helps the model use source-document information more effectively, reducing unsupported content in the generated summary.

Human Evaluation. We conduct a human evaluation on 100 randomly sampled CNN/DM test instances. Three annotators, all holding bachelor's degrees and having expertise in summarization, rated summaries from the Gold reference, Baseline, and ReFS on Factuality (Fact.) and Informativeness (Info.) using a five-point Likert scale.

Table 5 shows that ReFS scores higher on both criteria, suggesting that its reconstruction-based objective improves perceived factual quality and informativeness. Gold references receive lower scores than ReFS, consistent with prior findings on the limited quality of CNN/DailyMail reference summaries (Zhang et al., 2023).

Case Study: CNN/DailyMail

Article

Just weeks after Marvel's "Daredevil" premiered its first season on Netflix, the company confirmed a second season will be coming in 2016. The show focuses on attorney **Matt Murdock** (played by **Charlie Cox**), who was blinded as a child, as he fights injustice **by day using the law**. **By night**, he continues the fight, **becoming the superhero Daredevil**. **It's gotten rave reviews.**

Gold Summary

The critically acclaimed "Daredevil" will be back for season 2. Charlie Cox plays a blind attorney by day who is a superhero by night.

Baseline:

Marvel's "Daredevil" premiered its first season on Netflix in February. The show focuses on attorney Matt Murdock as he fights injustice by day and night.

ReFS:

The show focuses on attorney Matt Murdock as he fights injustice by day and becomes the superhero Daredevil at night. It's gotten rave reviews.

Table 4: Case study using Llama-3.2-3B-Instruct. Bold tokens in the *Article* are factuality-critical spans.

System	Fact.	Info.	Overall
Gold	4.61	3.38	4.00
Baseline	4.02	3.61	3.82
ReFS	**4.78***	**4.49***	**4.64***

Table 5: Human evaluation results. Annotators rated each summary on a 1–5 scale across three aspects (5 is the best). The overall score is the average across all aspects. The notation follows Table 2.

5 Conclusion

We proposed ReFS, a single-stage training framework that revisits reconstruction-based autoencoding for decoder-only SLMs through source reconstruction and iterative hidden-state refinement. ReFS improves factuality on CNN/DailyMail and XSum without introducing additional inference-time overhead. These results show that reconstruction-based regularization offers a practical way to improve source faithfulness in decoder-only summarization models. ReFS is also compatible with two-stage reranking, yielding further factuality gains when combined with SimCLS.

6 Limitations

ReFS demonstrates consistent improvements across multiple metrics and training scales. However, several limitations remain.

Model Scale. Our experiments are limited to small language models of approximately 3 billion parameters. Whether the reconstruction regularizer yields consistent gains for larger models remains an open question, as stronger pretrained representations may reduce the marginal benefit of the auxiliary objective.

Training Overhead. The hidden state refinement and source reconstruction phases together introduce $T + 1$ additional decoder forward passes per training step relative to standard fine-tuning, where T is dynamically set to the summary length (capped at $T_{\max} = 30$). This overhead scales linearly with T and can be reduced by lowering $T_{\max}$, at the expense of fewer refinement steps. While this may be non-trivial at larger scales, inference cost remains entirely unchanged.

Refinement Schedule. The ablation study on the effect of hidden state refinement demonstrates that iterative refinement substantially outperforms a single decoder pass, but does not establish whether $T = m$ is optimal relative to fixed or alternative schedules. Different choices of T may yield further factuality gains, and a more comprehensive sensitivity analysis is left for future work.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Singh Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio C'esar Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allison Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Young Jin Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Hong Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Corby Rosset, Sambudha Roy, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xianmin Song, Olatunji Ruwase, Praneetha Vaddamanu, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Andre Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonal Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Cheng yuan Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yunan Zhang, Xiren Zhou, and Yifan Yang. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *ArXiv*, abs/2404.14219.

Christos Baziotis, Ion Androutsopoulos, Ioannis Konstas, and Alexandros Potamianos. 2019. SEQ^3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 673–681, Minneapolis, Minnesota. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.

Thibault Févry and Jason Phang. 2018. Unsupervised sentence compression using denoising auto-encoders. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 413–422, Brussels, Belgium. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yanzhu Guo, Chloé Clavel, Moussa Kamal Eddine, and Michalis Vazirgiannis. 2022. Questioning the validity of summarization datasets and improving their factual consistency. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5716–5727, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*.

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

Juseon-Do, Jingun Kwon, Hidetaka Kamigaito, and Manabu Okumura. 2024. Instructcmp: Length control in sentence compression through instruction-based large language models.

Hidetaka Kamigaito, Katsuhiko Hayashi, Tsutomu Hirao, and Masaaki Nagata. 2018. Higher-order syntactic attention network for longer sentence compression. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1716–1726, New Orleans, Louisiana. Association for Computational Linguistics.

Hidetaka Kamigaito and Manabu Okumura. 2020. Syntactically look-ahead attention network for sentence compression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8050–8057.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9332–9346.

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

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

Yixin Liu and Pengfei Liu. 2021. Simcls: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072.

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.

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.

Yishu Miao and Phil Blunsom. 2016. Language as a latent variable: Discrete generative models for sentence compression. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 319–328, Austin, Texas. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, 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.

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.

Joonho Yang, Seunghyun Yoon, ByeongJeong Kim, and Hwanhee Lee. 2024. FIZZ: Factual inconsistency detection by zoom-in summary and zoom-out document. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Miami, Florida, USA. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

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

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

A Dataset Statistics

We use two standard summarization benchmarks: CNN/DailyMail (CNN/DM) and XSum. Table 6 summarizes the dataset statistics used in our experiments.

B Evaluation Metrics

In this work, we use the following seven metrics.

ROUGE: ROUGE (Lin, 2004) is a reference-based evaluation metrics that measures summary quality

Dataset	# Num			Avg of Len.	
	Train	Valid	Test	Doc.	Sum.
CNN/DM	287K	13K	11K	768.6	55.7
XSum	203K	11K	11K	429.2	23.3

Table 6: Dataset statistics for CNN/DailyMail and XSum. Avg. Len. reports the average token length of source documents and reference summaries.

based on lexical overlap between generated summaries and reference summaries. It captures similarity through unigram overlap, bigram overlap, and longest common subsequence via ROUGE-1, ROUGE-2, and ROUGE-L, respectively.

BERTScore: BERTScore (Zhang et al., 2020) evaluates semantic similarity between a generated summary and a reference summary using contextualized token embeddings. By computing token-level similarity based on contextual representations, it offers greater flexibility in handling paraphrases and diverse surface-level variations compared to lexical overlap-based metrics.

SummaC: SummaC (Laban et al., 2022) is a metric that evaluates the factual consistency of summaries using natural language inference (NLI). It assesses whether the summary factually matches the input document by verifying whether each sentence in the generated summary entails from the source document.

AlignScore: AlignScore (Zha et al., 2023) is a unified alignment-based factuality metric that measures the degree of information alignment between a generated summary and a source document. AlignScore evaluates factual consistency between two texts through a scoring function trained on various text-pair alignment settings, without relying on a specific task formulation.

FactCC: FactCC (Kryściński et al., 2020) is a classifier-based metric proposed to evaluate the factual consistency of abstractive summarization. It treats the factual consistency between source documents and generated summaries as a classification problem, determining whether the generated summary conflicts with the content of the input documents or contains information not found in the source.

FIZZ: FIZZ (Yang et al., 2024) is a fine-grained factual inconsistency detection metric designed to evaluate summarization factuality more precisely. By breaking down generated summaries into smaller factual units and evaluating whether each unit of information is supported by a source document, it enables a more detailed analysis of factual errors than summary-level or sentence-level evaluations.

Human Evaluation: Zhang et al. (2023) conduct a large-scale human evaluation demonstrating that LLM-generated summaries are frequently preferred over CNN/DailyMail gold references on coherence, faithfulness, and informativeness, attributing this to the origin of references as journalistic highlight bullets rather than self-contained summaries. Consistent with these findings, our human evaluation results (Table 5) show that ReFS scores higher than the Gold reference on Factuality and Informativeness.

Hyperparameter	Llama-3.2-3B	Phi-3.5-mini
LoRA rank (r)	32	32
LoRA α	64	64
Target modules	q_proj, v_proj	qkv_proj
LoRA dropout	0.05	0.05
Batch size (train)	1	
Gradient accumulation	32	
Learning rate	2×10^{-4}	
Max epochs	10	
Checkpoint selection	Best validation loss	
Optimizer	Paged AdamW 8-bit	
Quantization	4-bit NF4 + double quant	
Summary length cap ($T_{\max}$)	30	
Batch size (inference)	16	
Decoding	Greedy	
Max new tokens	512	
Max source length	512 tokens	
Inference precision	BF16	
Seed	42	

Table 7: Hyperparameter configurations. The upper block shows model-specific LoRA settings; middle blocks show shared training and inference settings.

C Implementation Details

All models were fine-tuned for up to 10 epochs with a batch size of 1 and gradient accumulation over 32 steps, using the Paged AdamW 8-bit optimizer with a learning rate of 2×10^{-4}. The best checkpoint was selected based on validation loss; in practice, the best checkpoint typically occurred within the first 2 epochs. Training employed 4-bit NF4 quantization with double quantization and BF16 compute dtype. Gradient checkpointing was enabled to reduce memory consumption. The summary length cap was set to $T_{\max} = 30$. The random seed was fixed at 42 across all experiments.

At inference, LoRA adapters were merged and the model was unloaded in BF16 precision. Source articles were truncated to 512 tokens, and sum-

maries were generated using greedy decoding with a maximum of 512 new tokens. Length targets for the length priming prompt were derived from the gold summary length of each test instance. LoRA configurations differ between models as shown in Table 7.

For SimCLS, 4 candidate summaries are generated via beam search and reranked using FactCC as the scoring criterion; accordingly, FactCC scores are not reported for +SimCLS variants to avoid evaluation bias.

D Background

SimCLS: Liu and Liu (2021) propose a two-stage summarization framework in which a pool of candidates is first generated via beam search and then reranked by a scoring model trained with a contrastive objective. We combine ReFS with SimCLS, using FactCC (Kryściński et al., 2020) as the reranking criterion; FactCC scores are accordingly excluded from evaluation for all `+SimCLS` variants to avoid evaluation bias.

Latent Reasoning: Hao et al. (2024) propose Coconut (Chain of Continuous Thought), training LLMs to reason in continuous latent space by feeding the hidden state at the final sequence position back as the input embedding for the subsequent step, rather than sampling a discrete token, for a pre-specified number of steps T. ReFS adopts the same iterative hidden state passing mechanism to enrich $\mathbf{h}_T$ with source-grounded representations prior to reconstruction, but dynamically ties the number of refinement steps to the summary length ($T = m$) rather than treating T as a fixed hyperparameter, avoiding performance variance from heuristic step selection.

Length Priming: Juseon-Do et al. (2024) introduce length priming, an instruction approach for better length control that supplies the target word count directly in the instruction prompt, enabling instruction-tuned models to generate outputs of a specified length without architectural modification. Following this approach, all methods in our experiments employ the same length priming template with target lengths derived from the gold summary of each test instance, reflecting the realistic deployment setting in which a desired output length is specified by the user.

E Performance at Various Training Scales

Tables 8 and 9 report full results for Llama-3.2-3B and Phi-3.5-mini across all training scales (1k, 2k, 4k, 8k) on CNN/DailyMail and XSum. Figure 3 provides a visual summary of factuality trends for both Llama-3.2-3B and Phi-3.5-mini across all training scales.

CNN/DailyMail. Our ReFS consistently outperforms the baseline across all scales under both models, with factuality gains observed across SC, AS, FactCC, and FIZZ. On CNN/DailyMail, the improvement margin remains stable as training data increases, suggesting that the reconstruction objective provides a reliable factual grounding signal regardless of data size. Notably, FactCC scores show the largest absolute gains at smaller scales (e.g., +11.17 at 2K for Llama), indicating that the reconstruction regularizer is particularly effective when training data is limited. When SimCLS reranking is applied, +SimCLS (ReFS) consistently achieves the highest factuality scores across all scales, demonstrating that our method produces a stronger candidate pool for two-stage approaches even at lower training scales.

XSum. On XSum, factuality improvements are observed across all scales, with the gains being most pronounced at smaller training sizes. As training scale increases from 1K to 8K, the improvement margin over the baseline tends to narrow, particularly for AlignScore and FIZZ. This trend suggests that the reconstruction objective provides a stronger signal on XSum when training data is scarce, with its relative advantage narrowing as the baseline benefits from more data. SummaC scores remain largely stable across all scales, consistent with the limited metric sensitivity observed in the main results. Fluency and relevance metrics show a moderate decline across all scales, reflecting an inherent tension between factuality-oriented training and surface-level generation quality.

Output Length. We report **ΔCR**, the difference between the model-generated and gold compression ratios (Kamigaito et al., 2018; Kamigaito and Okumura, 2020); lower absolute values indicate closer adherence to the reference length. Across all configurations, ΔCR values under ReFS remain close to those of the baseline, suggesting that the reconstruction objective does not systematically alter output length.

		CNN/DailyMail									XSum								
		Fluency & Relevance				Factuality				Len.	Fluency & Relevance				Factuality				Len.
Size	**Method**	**R-1**	**R-2**	**R-L**	**BS**	**SC**	**AS**	**FC**	**FI**	**ΔCR**	**R-1**	**R-2**	**R-L**	**BS**	**SC**	**AS**	**FC**	**FI**	**ΔCR**
1K	Baseline	**43.61**	**19.53**	**29.03**	**63.59**	65.4	86.76	38.57	57.87	+1.70	**37.07**	**13.91**	**28.78**	**61.80**	24.1	47.22	23.03	22.65	+0.46
	ReFS (Head)	43.25	19.19	28.90	63.38	66.1†	86.87	36.70	58.54†	+0.83	36.30	13.23	27.86	61.30	24.3	49.85*	25.26†	23.78†	**+0.34**
	ReFS	42.73	19.18	28.86	62.72	**70.8***	**89.69***	**49.85***	**66.67***	**+0.63**	34.28	11.90	26.39	60.01	**25.2†**	**56.76***	**28.77***	**29.38***	+0.46
	+SimCLS (Base)	**42.82**	**18.83**	28.04	**63.02**	71.2	89.41	-	62.87	+2.27	**36.71**	**13.83**	**28.25**	**61.56**	23.8	44.83	-	21.10	**+0.40**
	+SimCLS (Head)	42.60	18.80	**28.22**	62.83	72.2†	90.84†	-	63.97†	+1.32	35.21	12.37	26.86	60.51	23.9	48.88*	-	22.42†	+0.55
	+SimCLS (ReFS)	41.40	17.78	27.24	61.67	**76.8***	**92.56***	-	**73.16***	**+0.89**	32.71	10.66	24.73	58.42	**25.9***	**55.55***	-	**29.97***	+1.39
2K	Baseline	**43.67**	**19.83**	**29.38**	**63.67**	67.9	87.43	41.54	59.16	+1.18	**37.59**	**14.40**	**29.30**	**62.30**	24.2	45.04	22.91	21.67	+0.35
	ReFS (Head)	43.05	19.42	29.21	63.13	70.6*	88.70†	44.79*	63.47*	+0.75	36.98	13.79	28.64	61.62	24.2	46.05	25.28*	21.29	+0.30
	ReFS	42.52	18.80	28.73	62.68	**73.4***	**91.05***	**52.71***	**69.81***	**+0.51**	35.80	13.10	27.78	60.79	**24.6**	**51.15***	**27.59***	**25.73***	**−0.13**
	+SimCLS (Base)	**42.91**	**18.97**	**28.20**	**62.88**	73.5	92.05	-	64.33	+1.53	**37.19**	**14.42**	**28.85**	**62.03**	24.0	42.63	-	19.38	+0.44
	+SimCLS (Head)	41.78	18.29	27.68	62.17	76.1*	93.14*	-	68.02*	+0.99	36.20	13.39	27.76	61.05	23.9	42.66	-	20.19	+0.50
	+SimCLS (ReFS)	41.67	18.06	27.49	61.85	**78.8***	**93.23***	-	**73.72***	**+0.84**	34.68	12.21	26.64	60.08	**24.9†**	**49.04***	-	**24.38***	**+0.28**
4K	Baseline	**43.60**	19.44	29.25	**63.66**	65.6	86.75	42.25	57.91	+0.86	**38.27**	**15.01**	**29.80**	**62.76**	24.1	43.76	24.42	20.41	+0.24
	ReFS (Head)	43.32	**19.48**	**29.44**	63.46	68.2*	**88.60***	43.14†	61.54*	+0.51	37.48	14.12	28.78	61.87	24.2	47.21*	27.05†	22.07†	+0.36
	ReFS	42.56	18.92	28.84	62.69	**68.3***	88.55*	**47.49***	**64.87***	**+0.31**	37.21	14.03	28.91	61.67	**24.3**	**50.33***	**27.43†**	**23.20***	**+0.12**
	+SimCLS (Base)	**43.15**	**19.26**	**28.69**	**63.22**	71.6	91.83	-	62.31	+1.26	**37.95**	**14.37**	**29.12**	**62.45**	24.1	41.51	-	18.44	+0.39
	+SimCLS (Head)	42.47	18.86	28.24	62.61	**75.9***	**93.33***	-	68.13*	+0.80	36.90	13.91	28.23	61.52	23.8	43.28*	-	20.10†	+0.59
	+SimCLS (ReFS)	42.14	18.45	27.85	62.04	74.5*	92.28†	-	**71.42***	**+0.53**	36.34	13.48	28.01	61.08	**24.5**	**45.77***	-	**21.62***	**+0.36**
8K	Baseline	**43.82**	**19.79**	29.77	**63.65**	69.4	87.80	43.10	62.23	+0.41	**38.80**	**15.36**	**30.52**	**63.09**	24.0	43.19	25.09	18.74	+0.20
	ReFS (Head)	43.54	19.65	**29.92**	63.58	70.9†	89.22*	45.03*	64.27†	+0.29	38.38	14.90	29.81	62.61	**24.2**	43.74	25.55	**20.41†**	+0.50
	ReFS	43.18	19.55	29.47	63.25	**72.0***	**89.96***	**48.42***	**67.25***	**+0.08**	38.08	14.98	29.60	62.17	24.1	**43.90**	**26.07**	19.70	**+0.01**
	+SimCLS (Base)	**42.80**	19.05	28.83	**62.85**	75.2	92.55	-	66.47	+0.65	**38.72**	**15.16**	**30.05**	**63.24**	23.8	40.06	-	16.95	**+0.20**
	+SimCLS (Head)	42.75	**19.09**	**28.85**	62.85	76.5†	93.01	-	68.24†	+0.45	37.38	14.17	28.71	61.94	23.7	**41.77**	-	**18.22†**	+0.60
	+SimCLS (ReFS)	42.43	18.78	28.38	62.42	**77.2***	**93.29†**	-	**71.13***	+0.31	37.64	14.86	29.16	62.24	**23.9**	39.94	-	17.92	+0.44

Table 8: Llama-3.2-3B across training scales on CNN/DailyMail and XSum.

F Prompt Templates

Table 10 shows the prompt templates used in our experiments. `{src}` denotes the input article and `{src_len}` denotes the word count of the article.

		CNN/DailyMail									XSum								
		Fluency & Relevance				Factuality				Len.	Fluency & Relevance				Factuality				Len.
Size	**Method**	**R-1**	**R-2**	**R-L**	**BS**	**SC**	**AS**	**FC**	**FI**	**ΔCR**	**R-1**	**R-2**	**R-L**	**BS**	**SC**	**AS**	**FC**	**FI**	**ΔCR**
1K	Baseline	41.94	18.18	28.42	63.11	72.6	89.06	33.89	58.13	+0.32	**36.79**	**13.81**	**28.42**	**61.75**	24.8	49.31	25.00	23.77	+0.36
	ReFS (Head)	**42.10**	**18.47**	**28.62**	**63.23**	**73.2**†	90.08	37.78*	59.00	+0.46	35.26	12.81	27.34	60.93	25.4	51.17†	24.69	23.80	**+0.04**
	ReFS	40.35	16.89	27.05	62.89	64.5	**91.30***	**42.10***	**61.58***	**+0.22**	33.70	11.73	26.24	59.81	**25.8**†	**56.43***	**30.03***	**29.92***	−0.97
	+SimCLS (Base)	**41.92**	**18.47**	**28.35**	62.81	76.8	**92.13**	-	65.19	+1.46	**35.52**	**12.52**	**26.77**	**61.07**	24.3	46.10	-	20.52	+0.91
	+SimCLS (Head)	41.59	18.32	28.17	62.64	**77.0**	91.43	-	**65.48**	+1.56	34.01	11.76	25.72	60.05	25.4†	47.86†	-	21.96	+0.87
	+SimCLS (ReFS)	40.23	16.86	26.62	**62.83**	65.4	91.17	-	61.10	**+1.14**	33.59	11.45	25.91	59.88	25.6†	**50.74***	-	**25.46***	−0.30
2K	Baseline	**42.85**	**19.19**	**29.31**	63.26	75.5	90.94	43.63	62.29	+0.23	**37.03**	**13.98**	**29.06**	**62.08**	24.8	48.51	26.28	23.70	−0.71
	ReFS (Head)	42.67	18.92	29.14	63.15	74.6	90.35	44.77	63.14	**+0.14**	36.26	13.37	28.15	61.24	**25.6**†	50.96†	26.12	24.68	−0.04
	ReFS	41.96	18.59	27.94	**63.27**	**79.2***	**93.85***	**51.85***	**66.30***	+4.04	34.19	11.89	26.80	60.71	25.1	**61.04***	**29.63**†	**28.86***	−0.93
	+SimCLS (Base)	42.29	18.68	28.60	**62.83**	79.3	93.38	-	70.13	**+0.64**	**36.22**	**13.24**	**27.99**	**61.63**	24.4	46.19	-	21.13	−0.51
	+SimCLS (Head)	**42.36**	**18.95**	**28.78**	62.76	79.7	93.34	-	70.16	**+0.64**	35.29	12.55	26.73	60.72	**25.7**†	47.47†	-	22.07	+0.30
	+SimCLS (ReFS)	41.10	18.17	27.17	62.65	**83.1***	**95.03***	-	**71.40**†	+5.30	33.35	10.92	25.23	60.13	**24.9**	**57.65***	-	**28.05***	−0.52
4K	Baseline	**43.18**	19.49	29.37	**63.58**	76.0	91.32	48.96	63.50	+0.89	**37.64**	**14.33**	**29.37**	**62.65**	24.3	47.49	25.89	22.60	−0.35
	ReFS (Head)	43.02	**19.51**	**29.45**	63.49	77.0†	91.73	49.08	65.39†	**+0.45**	37.28	14.13	28.97	62.14	**25.3**†	50.11*	25.63	**25.06***	−0.23
	ReFS	42.32	18.88	28.48	63.07	**80.6***	**94.12***	**51.96**†	**69.33***	+2.19	36.30	13.31	27.86	61.77	24.8	**54.66***	**27.01**†	23.60	**−0.23**
	+SimCLS (Base)	**42.39**	18.78	28.52	**62.85**	80.0	93.76	-	69.29	+1.57	**36.53**	**13.37**	**28.13**	**62.26**	24.1	43.49	-	20.25	**−0.09**
	+SimCLS (Head)	42.31	**18.92**	**28.69**	62.77	80.4	94.08	-	71.90*	+0.87	35.77	12.85	27.35	60.96	**25.3**†	47.85*	-	**23.53***	+0.10
	+SimCLS (ReFS)	41.42	18.30	27.61	62.26	**82.9***	**94.99***	-	**73.63***	+2.96	34.91	11.96	26.26	60.55	25.0†	**53.11***	-	22.44*	+1.22
8K	Baseline	43.23	19.49	**29.34**	63.58	74.6	90.03	48.34	60.99	+1.09	**38.85**	**15.09**	**30.42**	**63.50**	24.3	47.02	26.65	20.93	**−0.43**
	ReFS (Head)	42.91	19.08	29.01	63.42	74.5	90.34	47.20	63.27*	**+1.06**	38.24	15.00	30.04	63.08	24.6	47.99	25.73	23.68*	−0.66
	ReFS	**43.24**	**19.50**	29.28	**63.75**	**81.8***	**94.91***	**61.03***	**70.78***	+1.73	38.07	14.68	29.89	62.93	**24.7**†	**49.00**†	**26.97**	**24.12***	−0.48
	+SimCLS (Base)	**42.96**	19.04	28.79	**62.98**	79.3	94.16	-	68.73	+1.54	**38.10**	**14.62**	**29.40**	**63.34**	**24.2**	43.57	-	18.74	−0.27
	+SimCLS (Head)	42.89	18.97	**28.84**	62.97	78.8	94.13	-	69.61†	**+1.52**	37.38	13.96	28.89	62.72	24.1	43.40	-	19.48†	+0.24
	+SimCLS (ReFS)	42.58	**19.14**	28.44	62.90	**85.5***	**96.45***	-	**75.68***	+1.57	37.40	13.77	28.65	62.32	24.1	**44.65**	-	**20.25***	**−0.13**

Table 9: Phi-3.5-mini across training scales on CNN/DailyMail and XSum.

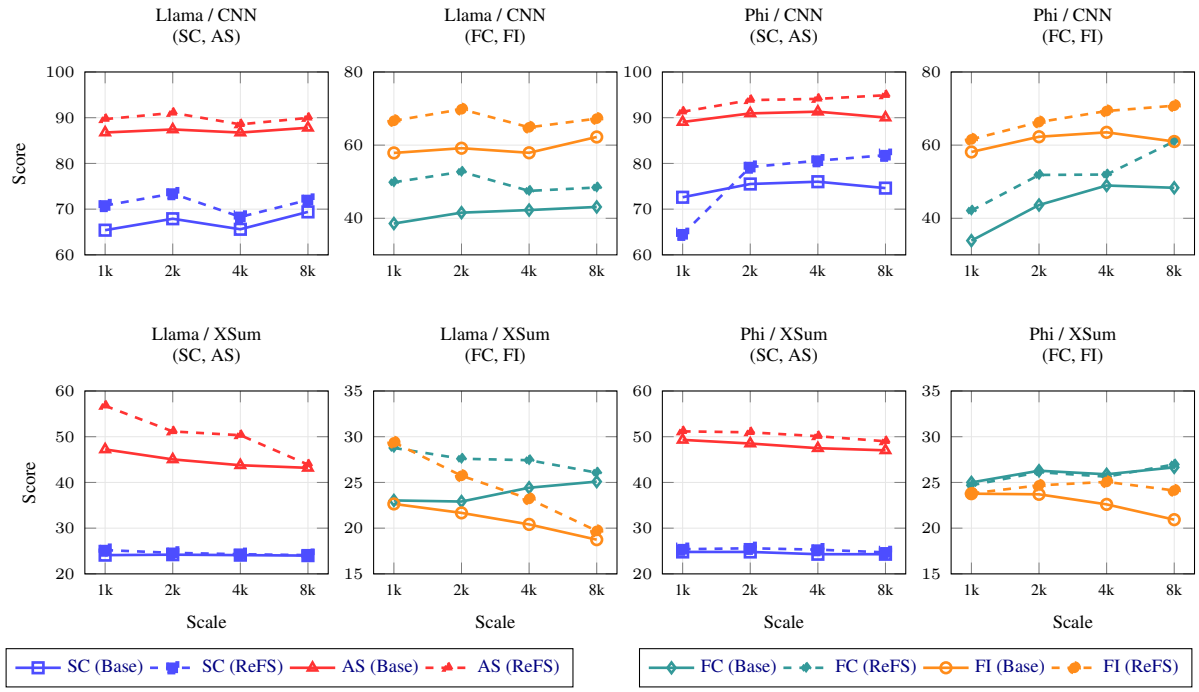

Figure 3: Factuality metrics across training scales for Llama-3.2-3B-Instruct and Phi-3.5-mini-Instruct on CNN/DailyMail (top row) and XSum (bottom row). Solid = Baseline; dashed = ReFS.

#	Constraint	Instruction
1	Priming	Article with {src len} words:\n {src} \n Summarize in {keep} words without the less important {del} words would be: \n
2	Reconstruction	You are given a news summary. \n Reconstruct a detailed original news article that this summary was derived from.\n \n Summary:\n {summary} \n \n Original Article:

Table 10: Prompt templates.