How Facility is the Small-Scale Abstractive Summarization Model: A Quantitative Study of Semantics and Syntax

Anonymous ACL submission

Abstract

Large-scale language models (LLMs) have demonstrated advancements in numerous capabilities, including factual consistency in abstractive summarization. However, the benefits of straightforward deployment and reduced invocation latency for small-scale language models (SLMs) should not be disregarded. Current evaluation metrics merely provide an abstract indication of factual score differences, leaving us uncertain about the specific areas where SLMs underperform and whether this gap is tolerable in certain contexts. This study initially illustrates the disparities between LLMs and SLMs regarding semantic knowledge and syntactic ability. Subsequently, we propose an SLM based on contrastive learning that allows tailored semantic and syntactic information and generates a parallel corpus with diverse summaries for the same document, each containing subtle semantic or syntactic flaws. By comprehensively integrating eight distinct factual evaluation metrics, we further elucidate the meaning of the gap in factual scores and identify the primary factual challenges current SLMs face in the abstractive summarization task.

1 Introduction

Previous studies suggest that the abstractive summarization model is prone to factual consistency problems (Kryściński et al., 2019). In recent years, LLMs have emerged, demonstrating superior comprehensive capabilities and performance in specific tasks compared to SLMs (fewer than 100 million parameters) (Zhao et al., 2023). They also exhibit better factual consistency in abstractive summarization (Zhang et al., 2024). However, deploying LLMs is challenging, and their invocation is costly (Yang et al., 2024; Liu et al., 2023b). Despite these challenges, SLMs retain considerable potential and are more apt for well-defined, singular tasks (Lepagnol et al., 2024). Therefore, utilizing SLMs in the right scenarios is still meaningful.

Source	**Text**
Document	It was written to author Betty Shew by the 21-year-old princess in 1947, months before her marriage. The two-page note [...]
Predicted summary	A letter written by Princess Elizabeth describing her relationship with Prince Philip has **sold for** more than **£15,000** at auction.

Table 1: Example of document and summary generated by SLMs with errors. We attribute this type of error to syntactic ability deficits because the phrase "sold for" in summary compels the model to provide the price number, a detail that cannot be discerned from the document.

Existing evaluation metrics are proficient at assessing the summaries' factual consistency, but their individual score intervals did not correlate with specific type or degrees of errors. This is largely attributed to the intricate variety of factual errors, leaving people with no idea of the usability of the model even if they know the factual scores.

This study establishes a correlation between ambiguous factual scores and specific scenarios through semantic knowledge and syntactic organization ability. Our experiments initially illustrate that these two competencies are the primary reason why LLMs exhibit superior factual consistency compared to SLMs. We propose a syntax-semantics controllable abstractive summarization model to demonstrate how these two competency deficits correspond to factual score gaps. This model generates parallels with variations in semantics or syntax, which is valuable due to the high costs of manual annotation and the unstable outputs from LLMs. By extensively integrating eight evaluation metrics for factual consistency, we further explore the correlation between specific factual scores and specific forms or degrees of errors. Concurrently, we gain a comprehensive understanding of the current factual issues encountered by SLMs in the abstractive summarization task, thereby shedding light on potential avenues for future research.

The contributions are highlighted as follows:

1. We investigate factual issues' origins from semantic and syntactic perspectives, an approach not previously proposed. These perspectives illustrate the disparity in factual consistency between SLMs and LLMs in abstractive summarization.

2. We introduce a contrastive learning-based model for controllable abstractive summarization with semantic and syntactic guidance. This method enables controllable text generation in SLMs.

3. By creating a parallel corpus comprising various summaries, each exhibiting subtle differences in semantics and syntax for the same document, we establish a correlation between factual scores and specific forms or degrees of errors. This corpus aids in identifying the primary challenges encountered by current small-scale models in the task of abstractive summarization and provides insight into potential future development directions.

2 Related work

2.1 Factual consistency problem in abstractive summarization

The factual consistency problem refers to the contradiction between the content stated in the model's summary and the document, a significant challenge encountered by abstractive summarization models. The manifestations of factual consistency issues are diverse. (Pagnoni et al., 2021) has divided factual errors into seven categories, including entity errors caused by incorrect semantic information and entity relationship errors or out-of-article errors caused by inappropriate syntactic structures, as can be seen in Table 1.

Several proposed metrics have facilitated the evaluation of factual consistency. These metrics can be grouped into two distinct categories based on their implementation: natural language inference (Kryściński et al., 2019; Laban et al., 2022), QA models (Durmus et al., 2020; Nan et al., 2021; Li et al., 2022). With the recent advent of LLMs, LLM-based evaluation metrics, such as G-eval (Liu et al., 2023a), have shown promising results. However, these metrics only provide a nebulous score for factual consistency, making it difficult to intuitively reflect the model's performance regarding semantic knowledge or syntactic structure.

2.2 Factual improvement method in abstractive summarization

The enhancement of model factual accuracy is often achieved through various modifications to the model training process. These modifications encompass adjustments to the training data (Chaudhury et al., 2022), the pre-training and fine-tuning process (Wan and Bansal, 2022), and the loss function during training (Cao and Wang, 2021; Dixit et al., 2023). LLMs have seen rapid evolution, exhibiting enhanced comprehensive capabilities. Significant improvements have also been observed in the factual consistency of the generated text (Zhang et al., 2024; Tang et al., 2024). However, these improved factual scores only abstractly reflect the trend of increasing factual consistency, failing to provide a clear picture of its practical utility. The specific errors these models make, and the frequency of such errors remain largely unknown.

2.3 Syntactic controllable text generation model

Despite their advanced prompt comprehension abilities and better overall capabilities, LLMs are associated with specific challenges in deployment and training and high invocation costs (Xu and Zhang, 2024; Wang et al., 2024). On the other hand, SLMs with adjustable parameters continue to hold substantial value for text generation tasks that require singular objectives (Lepagnol et al., 2024). Therefore, controllable text generation based on SLMs remains a significant study area.

Numerous studies have strived to regulate the text-generation process of SLMs. In the context of open-dialogue response tasks, Zhu et al. (2021) suggested a sentence-level information method in the latent space to disentangle content and style. Furthermore, Zhu et al. (2021) proposed a syntax-controlled paraphrase generator to learn the decoupling of semantics and syntax from unannotated text, thereby generating training datasets that lack parallel corpora. The conventional sequence-to-sequence framework has also been enhanced by introducing a novel two-stage decoder to impose style constraints on the generated text (Hu et al., 2022). In this study, we aim to employ a semantic-syntax controllable summarization model to generate parallel corpora with subtle semantic and syntactic differences and establish a correlation between factual scores and specific forms or degrees of errors.

3 Gap between LLM and SLM

As the scale of LLMs increases, they exhibit improved syntactic capabilities, thus enabling them to tackle more intricate problems. Empirically, LLMs demonstrate enhanced control over semantic knowledge and syntactic information. Research has substantiated that LLMs can generate abstractive summarizations with consistent factual accuracy (Zhang et al., 2024; Tang et al., 2024). We discern that it is the dual aspects of **semantic knowledge** and **syntactic structure** that endow LLMs with superior factual consistency.

3.1 More robust semantic knowlwdge

Hallucinations significantly contribute to factual inconsistency in the task of abstractive summarization. It has been established that hallucinations transpire when models generate summaries without referencing the document, relying solely on their internal information storage (Chae et al., 2024). Despite being inevitable (Xu et al., 2024), if the knowledge stored within the model aligns with the actual scenario, the model can generate also factually accurate text. Empirical evidence suggests that in the context of news text summarization, LLMs possess a more accurate reserve of semantic knowledge. This capability allows LLMs to retain critical information in the summary, even without referencing the document, thereby generating summaries with higher factual consistency.

Numerical information in the summary is challenging to infer solely from context. The ability of the model to reduce these words partly reflects the model's semantic knowledge. We use Spacy[1] to select numerals from the summary and mask them, requiring the model to restore them without referring to the document. Empirical evidence in table 2 shows that the restoration ability of LLMs far exceeds that of SLMs. They can restore the critical information in the summary without referring to the document. This means that even when LLMs rely directly on their knowledge storage, the information output is factually consistent.

LLMs utilize their semantic knowledge flexibly, as opposed to a blind application. We manipulated crucial information such as time, place, and person in the document using LLM, thereby generating a fabricated news document that contradicts objective facts and the knowledge stored in the model. Empirical evidence indicates that the summaries generated by the model under these circumstances maintain factual consistency, with an average G-eval score of 4.83, even significantly surpassing the scores of golden summaries in the dataset.

[1] https://spacy.io/

Model	Total (sampled from 1000)	Restored Number	Proportion
Bart_large	399	66	16.5%
GPT4	399	195	48.9%

Table 2: The proportion of numerical words recovered by different models. GPT4 stores more semantic information and has a more robust recovery capability. Not all abstracts contain numerical information; we selected 399 summaries from 1000 that contain numeral words.

3.2 More flexible syntactic structures

The syntactic structure is also a crucial factor influencing the factual consistency of abstractive summarization. When the syntactic structure of the summary does not match the information that the document can provide, the model may be forced to choose incorrect words to fill in to meet the basic requirements of grammatical accuracy. As shown in Table 1, the prepositional phrase in the model forces the model to fill in numerals, but this information cannot be obtained from the document. LLMs have superior abilities to coordinate syntactic structures, avoiding such problems affecting the factual consistency of the summary.

In many specific grammatical constructions, we select numeral prepositional phrases as our study's focus, as an illustrative example akin to those in Table 1. To assess their respective capabilities quantitatively, we emulate a scenario where the document lacks numerals, thereby examining whether the model necessitates using numerals in the summary. Utilizing Spacy, we substitute numerals in the document with an unk-token and record the changes in numeral prepositional phrases in the summary, pre and post masking. The experiment substantiates that SLMs lack the flexibility to manipulate the summary's syntactic structure. Despite the document's inability to provide pertinent information, most of its summaries persist in employing numeral prepositional phrase structures, thereby introducing erroneous information. In contrast, LLMs exhibit superior control over syntactic structure.

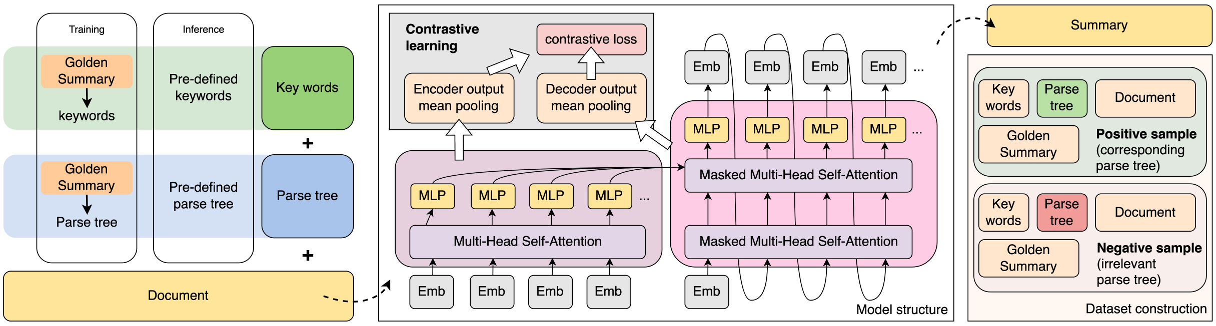


Figure 1: An overview of the semantic-syntax controllable summarization model. Its input includes semantic knowledge, syntactic information, and the document. We use contrastive learning strategy in the training process.

Model	Original proportion	Proportion after numeral mask	Percentage of decline
Bart_large	0.127	0.103	18.6%
GPT4	0.145	0.0514	64.6%

Table 3: Proportional changes in the numeral prepositional phrases number. The larger the drop, the more sensible syntactic structure it uses.

4 Semantic-syntax controllable text summarization model

4.1 Motivation

In this section, we will propose a text generation model with controllable semantics information and syntactic structure. We will present the design details of this model and demonstrate its validation.

4.2 Model design

We achieve controllable semantics and syntax by modifying the format of model input and introducing contrastive learning methods. We categorize number words, nouns, and proper nouns as semantic information and the parse tree as syntactic information. We use the bracket notation method[2] to transform the parse tree into a string. As shown in the figure1, we concatenate the semantic information, syntactic information, and the document as the model input. The model output should be a summary that adheres to semantic and syntactic criteria. We utilize the contrastive loss similar to FactPegasus (Wan and Bansal, 2022):

$$l_{I_i,S_i} = -\log \frac{\exp(sim(z_{I_i}, z_{S_i})/\tau)}{\sum_{I_j \in \mathcal{N} \cup \{I_i\}} \exp(sim(z_{I_j}, z_{S_i})/\tau)} \tag{1}$$

The parse tree of the negative example is unrelated to that of the generated summary. We denote the input and generated summary as I_i and S_i, respectively, where z_{I_i} and z_{S_i} represent their representations. These representations, z_I and z_S, are generated by applying mean pooling to the final hidden layer of the encoder and decoder outputs, respectively. The function $sim(\cdot,\cdot)$ signifies the cosine similarity between the representations, while τ represents the temperature parameter. The final loss is computed as the sum of the cross-entropy loss L_{CE} and the contrastive loss L_{CL}, with λ being a scalar. The equation is as follows:

$$L = L_{CE} + \lambda L_{CL} \tag{2}$$

In this manner, the guidance signals influence the generated summary's semantic information and syntactic structure. We demonstrate the effectiveness of semantic and syntactic guidance signals through two sets of experiments. Meanwhile, the model does not simply combine words but generates summary text by referring to the document. In subsequent factual evaluations, the generated summary maintains a high factual consistency when both semantics and syntax are appropriate.

4.3 Semantic controllability verification

Our model strongly correlates semantic information in abstractive summarization and semantic guidance signals. For our experimental design, we selected a sample size of 100 instances. For each instance of semantic information, we opted for parse trees of diverse depths to serve as syntactic guidance signals. To mitigate the influence of individual cases on the experimental results, we employed multiple sampling strategies, with each group containing ten samples drawn from different golden summaries. The experimental results are demonstrated in Figure 2.

Semantic control signals are effective. The trend of the two solid lines representing the maximum

[2]Examples can be found in the table 8

values tells us that the model will utilize semantic guidance information as much as possible. It is also seen that syntactic structures limit the use of semantic information. When the parse tree in the guidance signal is too shallow, only a tiny part of the semantic information can be utilized. As the depth of the parse tree gets deeper, the increased recall indicates that the model is trying its best to use all semantic information. At the same time, when the parse tree is too deep, the model has to add extra information for syntactic structure completeness, and precision decreases as well.

Semantic information utilization is relatively low on average value. It tells us that a fixed syntactic structure often struggles to accommodate all semantic information, and in most cases, semantic information cannot be fully utilized.

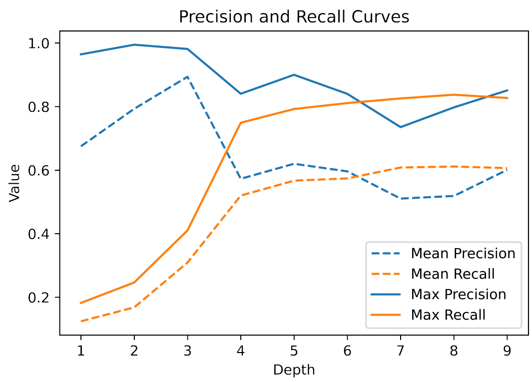


Figure 2: Trends of precision and recall in generated summary semantic information as the depth of the parse tree changes. We have multiple templates for each group of syntactic templates related to different parse tree depths, and for each case, we take the maximum and average values, respectively.

4.4 Syntactic controllability verification

We establish the correlation between syntactic structures by computing the Rouge score (Lin, 2004) of the bracket notation string derived from the input syntactic signals and the parse tree of the output summary. This is an unprecedented method. However, intuitively, the computational principle indicates that the similarity of the string is strongly associated with the similarity of the parse tree.

In the absence of contrastive learning strategies, it remains a challenge for the model to discern the correlation between the output summary and the syntactic signals in the input. The document component in the input provides sufficient information, leading the model to generate summaries

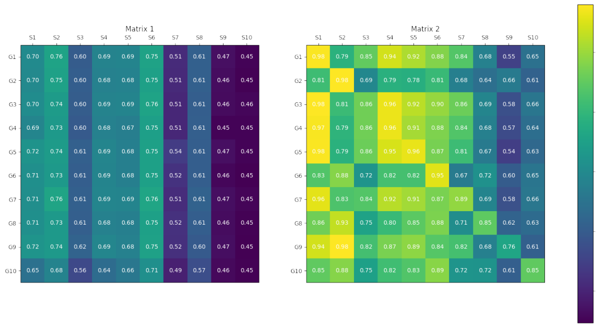


Figure 3: Correlation of summary syntactic structure and guidance Signals. $Matrix1$ represents direct training and $Matrix2$ represents training with contrastive learning strategies. S_i represents the summaries generated under the conditions of the guidance signal G_i.

based predominantly on this. In Figure 3 $Matrix1$, the summaries produced under varying guidance signals exhibit considerable similarity. However, the introduction of contrastive learning reveals a significant correlation between the input syntactic structure signals and the output summaries. This is illustrated $Matrix2$, which shows the trend of the highest diagonal values in the matrix.[3]

5 Quantitative Analysis for semantic and syntactic

5.1 Implementation details

We conduct experiments based on the XSUM dataset (Narayan et al., 2018). Given that factual errors also exist in the golden summaries (Maynez et al., 2020), we selectively sample summaries exhibiting superior factual consistency for our experiments, specifically those with a G-eval score 5.

We chose the Bart-base model as base model, which has approximately 140 million parameters. Our models were trained on an RTX 4090 GPU using PyTorch, with a training epoch of 10 and a batch size of 4. We utilize Spacy for semantics and syntax analysis. The semantic guidance is annotated from the nouns, numerals, and proper nouns extracted from the golden summaries. The syntactic information is represented by transforming the corresponding parse tree of the sentence into a bracket notation string. The guidance signals and the document are concatenated using a separator token in the input.

We widely adopt various factual evaluation metrics introduced in the related works. FactCC (Kryś-

[3]The precision in Rouge-L is depicted in the figure, and comprehensive results can be found in the Appendix D

ciński et al., 2019), DAE (Goyal and Durrett, 2021), SummaC (Laban et al., 2022), ANLI (Nie et al., 2019) are based on NLI model, ClozE (Li et al., 2022), FEQA (Durmus et al., 2020),Q2 (Honovich et al., 2021) are based on cloze or QA models, G-eval(Liu et al., 2023a) are based on LLM. All of them have open-source codes. The complete experimental results can be found in the appendix. In terms of summary model selection, in addition to Bart (Lewis et al., 2019) and Pegasus (Zhang et al., 2020), we also chose some models optimized for factual consistency. CLIFF(Cao and Wang, 2021), FactPegasus (Wan and Bansal, 2022) and EFACTSUM (Dixit et al., 2023) are all implemented based on publicly available checkpoint files. For X-factor (Chaudhury et al., 2022), we executed our implementation based on the statement in the paper.

5.2 Semantic influence

5.2.1 Data Enhancement Method

We quantify the impact of incorrect semantic information by adjusting the guidance provided to the model. The semantic information and syntactic information utilized in the experiments are derived from the golden summary. The data augmentation method encompasses the following two types:

Replace: We replace the keywords in the semantic guidance with irrelevant words, simulating the summary generated when erroneous semantic information is introduced. The irrelevant words are randomly extracted from the golden summaries of other cases, ensuring consistency in word type. Here, R-N represents the number of words replaced with incorrect words.

Mask: As inferred from the above experiments, the model will select appropriate words to supplement when the semantic guidance is insufficient. We artificially reduce the number of words in the semantic guidance extracted from the golden summary, simulating a scenario where most semantic information has been accurately chosen. However, the model needs to select a few additional words. Here, M-N represents the number of words reduced in the semantic information.

5.2.2 Results analysis

The results of the experiment can be seen in Fig 4. Detailed scores can be found in the Appendix. The experiment leads us to the following conclusions:

(1) Erroneous semantic information substantially impacts factual consistency, whereas the quantity

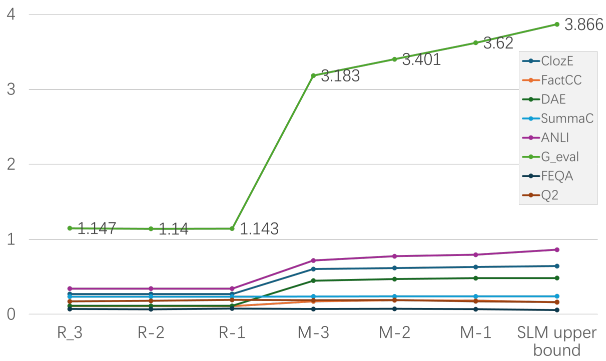


Figure 4: The relationship between factual scores and different types of semantic information inside. Upper bound refers to summaries generated under fully correct syntactic and semantic signals.

Method	Variance
Replace	0.162
Mask	2.232

Table 4: Fact score (G-eval) variance of summaries generated by different data enhancement methods.

of erroneous information has a less effect. The factual consistency score (R-N) for summaries generated from erroneous words is markedly lower than those from keyword masking (M-N). Notably, the factual consistency score does not diminish further with an increase in erroneous words.

(2) The likelihood of a model making errors increases with the amount of semantic information it needs to process. As can be seen in summaries generated by keyword masking (M-N), The more masked words, the worse the factual scores are. The occurrence of factual errors in this case is probabilistic because the variance of the factual score in the mask method is higher, as seen in Table 4.

(3) The sensitivity of all metrics to semantic alterations is not uniform, as illustrated in the Appendix E. The two QA-based metrics, FEQA and Q2, appear insensitive to semantic errors. This insensitivity is likely attributable to the question-asking method, which can only sample and verify a limited amount of semantic information.

5.3 Syntactic influence

5.3.1 Data Enhancement Method

We have devised two ways to select syntactic structure signals. In this experiment, the semantic information comes directly from the golden summary.

Fixed Syntactic Structure: We have manually constructed a collection of guidance signals with fixed syntactic structures by sampling gold summaries fulfill specific attributes. These sets include dozens of syntactic structures that share standard features in terms of the depth of the parse tree, the type of the top-level syntactic structure, and the number of modifiers[4]. To avoid bias in the experimental results outcomes induced by a single syntactic structure, each group of syntactic structures includes ten specific cases that meet the feature.

Syntactic Structure from Document In pursuit of a syntactic structure more aligned with semantic information, we also attempt to select the syntactic structures present in the document as guidance signals. We count the number of times each sentence in the document is hit by the words in the semantic information and prioritize the sentences with a higher number of hits as guidance signals.

5.3.2 Results analysis

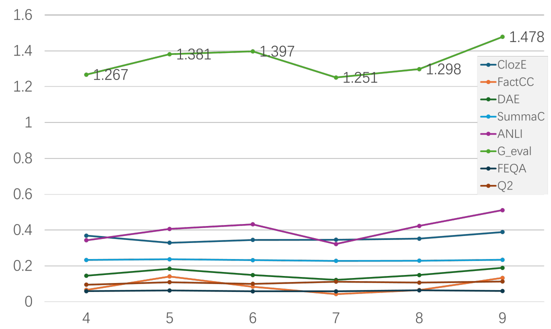
We tabulated the average fact score for each scenario. Moreover, for this case, in the fixed syntactic structure, we statistically characterize the effect of template features on the factual consistency. Our experiment yields the following conclusions:

(1) In the realm of factual consistency, improper syntactic structures can have detrimental effects. As depicted in Table 5, the factual consistency of summaries generated by fixed syntactic structures is generally low, with values closely mirroring the average scores of those utilizing incorrect semantic information. This indicates that even if the model can select appropriate information from the document under accurate semantic guidance, inappropriate syntactic structures may compel the model to convey incorrect semantics. It is also evident that aligning syntactic structures with semantic information can pose a significant challenge.

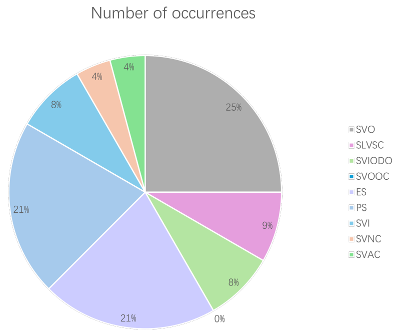
(2) Based on fixed syntactic structures, distinct syntactic configurations can result in varying probabilities of factual consistency issues, thereby altering the error margin. This insight could aid in comprehending how the text generation models manage semantic information and syntactic structure. We conduct a comparison of summaries generated under the influence of diverse syntactic structure groups and arrive at the subsequent conclusion:

1. It is observed that an increase in the depth of the parse tree often leads to the introduction of additional words by the model. However, this does not necessarily imply a decrease in

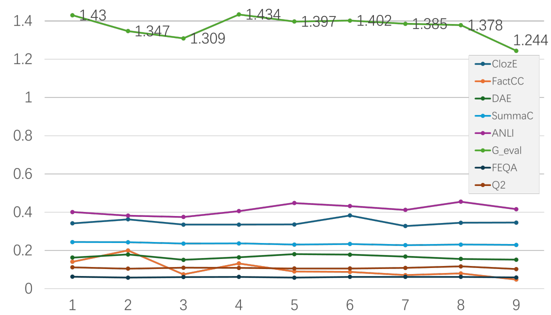
[4] we define modifiers as words involved in various syntactic dependency relations, such as adjectival modifiers (amod), adverbial modifiers (advmod), quantity modifiers (quantmod)



(a) Relationship between factual scores and parse tree depths



(b) Relationship between factual scores and top-level syntactic structures is counted by the top-level syntactic structures with the top three scores in each metric. The greater the proportion of top-level syntactic structures in the pie chart, the stronger the factual advantage it proves to be. The meaning of the abbreviations can be found in the appendix A



(c) Relationship between factual scores and the modifier number

Figure 5: The factual scores of summaries generated under different fixed syntactic guidance signal.

Situation	Scores
Fixed syntactic structure	1.388
Syntactic structure from document	1.539
Incorrect semantic information	1.143
Semantic information deternined by model	3.401
Upper bound	3.866

Table 5: The average factual score (G-eval) of generated summaries under different conditions. As can be seen from the table, inappropriate syntactic information can have a significant impact on the factual consistency of the model, which is nearly as influential as the introduction of incorrect semantic information.

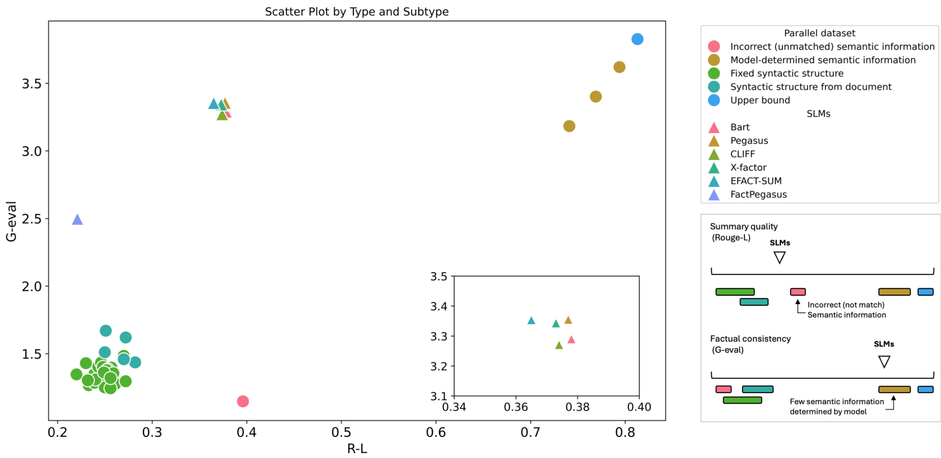


Figure 6: The distribution of summaries in terms of summary quality and factual consistency. The △ represent summaries generated by SLM-based summary models, while the ○ represent summaries in the parallel dataset.

factual consistency. More complex syntactic structures may enhance the model's ability to position existing words accurately. As shown in Fig 5a, the model achieves optimal factual consistency when the parse tree depth is 9.

2. As shown in Fig 5b, the model often exhibits higher factual consistency when its syntactic structure incorporates subject-predicate-object constructs and existential or passive sentences.
3. As shown in Fig 5c, the more modifiers there are in the syntactic structure, the more semantic information the model aims to convey and the higher the probability of error.
4. As shown in Table 5, utilizing syntactic structures with shared keywords in document sentences enhances factual consistency, largely due to these structures' capacity to incorporate corresponding semantic content effectively.

6 Quantitative analysis of SLMs in sbstractive summarization task

Overall quality: We used the Rouge to assess the overall quality of the summaries. As shown in Figure 6, in the parallel data, summary quality is at its lowest when the syntactic strcuture is inappropriate. Quality is also affected when there are inappropriate words. Compared with these specific situations, SLMs can generate summaries structurally similar to the golden summary but cannot fully restore the words in the golden summaries. Given that a summary is not unique, this does not necessarily indicate an error in the summary generated. [5]

[5] We replicated the FactPegasus model utilizing the publicly available checkpoint, but its performance is relatively ordinary.

Factual consistency: The influence of particular semantic or syntactic anomalies on factual consistency has been previously deliberated. As depicted in Figure 6, it is evident that the issues presently faced by SLMs are semantic rather than syntactic. The factual scores of the model significantly exceed those obtained when employing fixed syntactic structures. SLMs also accurately discern most semantic information. Factual inconsistencies may arise when the model makes decisions on the final one or two semantic messages, which is the primary challenge encountered by current SLMs and constitutes the principal discrepancy with LLMs.

7 Conclusion

Summaries generated by SLMs often lack the factual consistency of those LLMs produce. However, current evaluation metrics only quantify this discrepancy through numerical scores, making the difference unclear. This paper elucidates the disparity between SLMs and LLMs regarding semantic knowledge and syntactic information. Then, we introduce a semantic-syntax controllable summarization model. By utilizing parallel data generated by this model, we highlight the semantic and syntactic shortcomings of the generated summaries that may correspond to varying factual scores. This approach allows us to understand better the factual performance of SLMs in the abstractive summarization task. In this manner, we gain a more precise understanding of the factual performance of SLMs in the abstractive summarization task. We can leverage the deployment simplicity and reduced latency of SLMs by applying them in suitable scenarios.

8 Limitations

(1) we use Spacy to extract semantic and syntactic information in this study. While Spacy is a widely used parsing tool with notable performance, its analytical results are still inconsistent.
(2) A variety of factual consistency evaluation metrics are employed in our assessment. While the output scores from most of these metrics align with our findings, there are exceptions. A detailed analysis of the discrepancies among these metrics was not conducted due to the number of implementation details and space constraints of the article

References

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

Kyubyung Chae, Jaepill Choi, Yohan Jo, and Taesup Kim. 2024. Mitigating hallucination in abstractive summarization with domain-conditional mutual information. *arXiv preprint arXiv:2404.09480*.

Subhajit Chaudhury, Sarathkrishna Swaminathan, Chulaka Gunasekara, Maxwell Crouse, Srinivas Ravishankar, Daiki Kimura, Keerthiram Murugesan, Ramón Fernandez Astudillo, Tahira Naseem, Pavan Kapanipathi, et al. 2022. X-factor: A cross-metric evaluation of factual correctness in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7100–7110.

Tanay Dixit, Fei Wang, and Muhao Chen. 2023. Improving factuality of abstractive summarization without sacrificing summary quality. *arXiv preprint arXiv:2305.14981*.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*.

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.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870.

Zhe Hu, Zhiwei Cao, Hou Pong Chan, Jiachen Liu, Xinyan Xiao, Jinsong Su, and Hua Wu. 2022. Controllable dialogue generation with disentangled multi-grained style specification and attribute consistency reward. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:188–199.

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.

Pierre Lepagnol, Thomas Gerald, Sahar Ghannay, Christophe Servan, and Sophie Rosset. 2024. Small language models are good too: An empirical study of zero-shot classification. In *LREC-COLING 2024*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yiyang Li, Lei Li, Qing Yang, Marina Litvak, Natalia Vanetik, Dingxin Hu, Yuze Li, Yanquan Zhou, Dongliang Xu, and Xuanyu Zhang. 2022. Just cloze! a fast and simple method for evaluating the factual consistency in abstractive summarization. *arXiv preprint arXiv:2210.02804*.

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. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Re, et al. 2023b. Deja vu: Contextual sparsity for efficient llms at inference time. In *International Conference on Machine Learning*, pages 22137–22176. PMLR.

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.

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021. Improving factual consistency of abstractive summarization via question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6881–6894, Online. 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. *arXiv preprint arXiv:1808.08745*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

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*.

Liyan Tang, Igor Shalyminov, Amy Wing-mei Wong, Jon Burnsky, Jake W Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, et al. 2024. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. *arXiv preprint arXiv:2402.13249*.

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.

Yuxin Wang, Yuhan Chen, Zeyu Li, Zhenheng Tang, Rui Guo, Xin Wang, Qiang Wang, Amelie Chi Zhou, and Xiaowen Chu. 2024. Towards efficient and reliable llm serving: A real-world workload study. *arXiv preprint arXiv:2401.17644*.

Jing Xu and Jingzhao Zhang. 2024. Random masking finds winning tickets for parameter efficient fine-tuning. *arXiv preprint arXiv:2405.02596*.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Qingfu Zhu, Weinan Zhang, Ting Liu, and William Yang Wang. 2021. Neural stylistic response generation with disentangled latent variables. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4391–4401.

A Top Grammar structure

1. Subject + Verb + Object (SVO)
   Explanation: The subject performs the action, and the object is the receiver of the action.
   Example sentence: She reads books.
   Grammar structure: Subject (She) + Verb (reads) + Object (books)

2. Subject + Linking Verb + Subject Complement (SLVSC)
   Explanation: The linking verb connects the subject and the subject complement, and the subject complement describes the subject.
   Example sentence: The sky is blue.
   Grammar structure: Subject (The sky) + Linking Verb (is) + Subject Complement (blue)

3. Subject + Verb + Indirect Object + Direct Object (SVIODO)
   Explanation: The indirect object is the beneficiary of the action, and the direct object is the object of the action.
   Example sentence: He gave her a gift.
   Grammar structure: Subject (He) + Verb (gave) + Indirect Object (her) + Direct Object (a gift)

4. Subject + Verb + Object + Object Complement (SVOOC)
   Explanation: The object complement further explains the object.
   Example sentence: They named the baby John.
   Grammar structure: Subject (They) + Verb (named) + Object (the baby) + Object Complement (John)

5. Existential Sentence: There is/are + Subject (ES)
   Explanation: Used to describe the existence of something or someone.
   Example sentence: There are many stars in the sky.
   Grammar structure: Existential structure (There are) + Subject (many stars) + Adverbial (in the sky)

6. Passive Sentence: Subject + Auxiliary Verb + Past Participle + (by Agent) (PS)
   Explanation: Passive sentences are used to emphasize the receiver of the action, not the doer of the action. In passive sentences, the doer of the action is usually introduced by the preposition "by".
   Example sentence: The cake was eaten by the children.
   Grammar structure: Subject (The cake) + Auxiliary Verb (was) + Past Participle (eaten) + Prepositional Phrase (by the children)

7. Subject + Verb + Infinitive (SVI)
   Explanation: The infinitive serves as the object or complement.
   Example sentence: She wants to travel.
   Grammar structure: Subject (She) + Verb (wants) + Infinitive (to travel)

8. Subject + Verb + Noun Clause (SVNC)
   Explanation: The noun clause serves as the object, subject, or complement.
   Example sentence: Tom believes that Mary is a good student.
   Grammar structure: Subject (Tom) + Verb (believes) + Noun Clause (that Mary is a good student)

9. Subject + Verb + Adjective Clause (SVAC)
   Explanation: The adjective clause modifies a noun or pronoun.
   Example sentence: The book that you gave me is fascinating.
   Grammar structure: Subject (The book) + Adjective Clause (that you gave me) + Verb (is) + Complement (fascinating)

B Construction of syntactic guidance signals

We explore three main aspects of syntactic structural features that affect the factual accuracy of summaries: the depth of the parse tree, the type of the top-level syntactic structure, and the number of modifiers. We sample from gold summaries and select the parse trees of the summaries that satisfy specific features as guide signals. Table 6 and Table 7 represent the distribution of the depth of the parse tree and the number of modifiers in the gold summary.

Parse Tree Depth	Number of Samples
1	1
2	7
3	16
4	110
5	211
6	210
7	196
8	105
9	76
10	35
11	22
12+	11

Table 6: The distribution of parse tree depth in 1000 golden summaries sampled from the XSUM datasets.

Number of Modifiers	Number of Samples
0	6
1	17
2	50
3	107
4	153
5	164
6	160
7	139
8	175
9	62
10	35
11	13
12+	19

Table 7: The distribution of modifier number in 1000 golden summaries sampled from the XSUM datasets.

C Examples of generated summaries

Table 8 summarizes the same semantic information output under different syntactic structures guidance signals. In order to demonstrate the consistency of the output summaries and the input syntactic guidance signals, we have chosen to use shorter guidance signals. It also demonstrates that the model is not simply sentences but is doing its best to generate summaries based on the document's content. hyperref

D Validation of contrastive learning

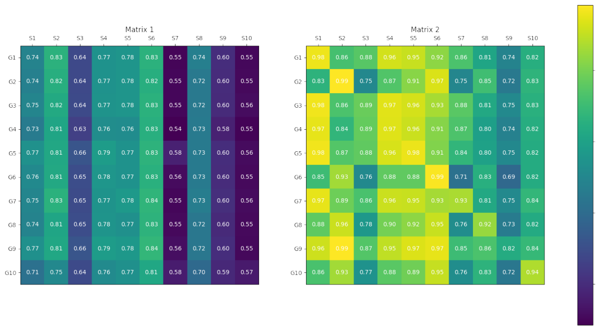
Figure 7 illustrates direct training versus training using contrast learning.

E Complete experimental results

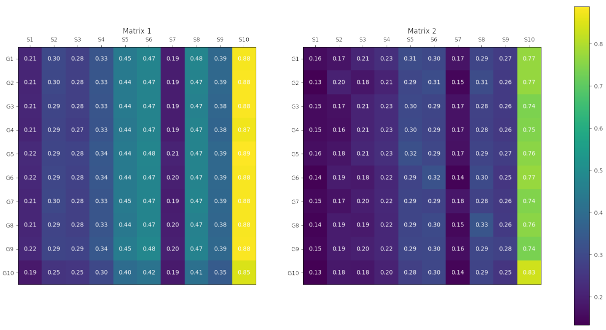
The complete experimental results can be seen in Figure 8, Figure 9 and Figure 10.

Parse Tree	Example	Generated Summary
VERB (PRON NOUN PUNCT)	She reads books.	Apple denies misleading customers.
AUX(NOUN (DET) ADJ PUNCT)	The sky is blue.	The firm is misleading customers.
VERB (PRON PRON NOUN (DET) PUNCT)	He gave her a gift.	Apple denies it broke the law.
VERB (PRON NOUN (DET) PROPN PUNCT)	They named the baby John.	What does the firm claim Apple?
VERB (PRON NOUN (ADJ ADP (NOUN (DET))) PUNCT)	There are many stars in the sky.	Apple faces legal action from the US.
VERB (NOUN (DET) AUX ADP (NOUN (DET)) PUNCT)	The cake was eaten by the children.	The firm has apologised to some customers.
VERB (PRON VERB (PART) PUNCT)	She wants to travel.	Apple plans to apologise.
VERB (PROPN AUX (SCONJ PROPN NOUN (DET ADJ)) PUNCT)	Tom believes that Mary is a good student.	Apple says that US is the only firm.
AUX (NOUN (DET VERB (PRON PRON PRON)) ADJ PUNCT)	The book that you gave me is fascinating.	The firm that sells it is misleading.
VERB (PROPN (NOUN (NOUN (PROPN))) AUX VERB (PART NOUN (ADJ VERB (PRON AUX ADP (NOUN (DET NOUN (NUM) ADP (PROPN (DET ADJ))))))) PUNCT)	Us technology firm Aple has offered to refund Australian customers who felt misled about the 4G capabilities of the new ipad.	Us technology firm Apple has promised to explain misleading customers it is selling with the 4G capabilities of the new iPad.

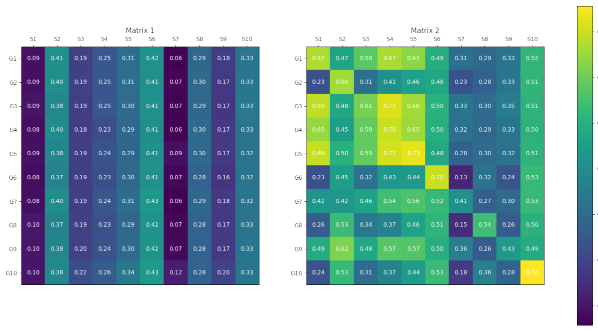
Table 8: Summaries of the same semantic information are generated under varying grammatical guidelines. The semantic knowledge is provided by these words in random order: ’US,’ ’technology,’ ’firm,’ ’Apple,’ ’customers,’ ’4G’, ’capability,’ and ’iPad.’ The source can be accessed via this URL.



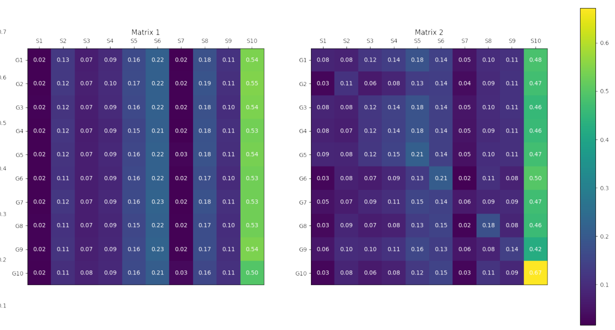
(a) Rouge-1 precision



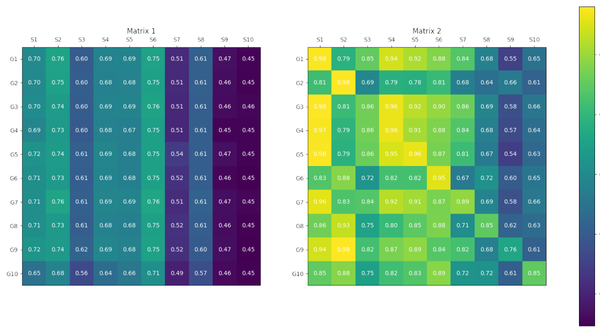
(b) Rouge-1 recall



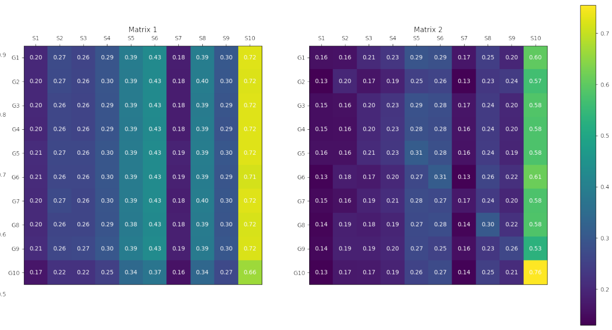
(c) Rouge-2 precision



(d) Rouge-2 recall



(e) Rouge-L precision



(f) Rouge-L recall

Figure 7: Evaluation of the validity of syntactic guidance signals using the Rouge metrics.In each subgraph, $Matrix1$ represents direct training and $Matrix2$ represents training with contrastive learning strategies. S_i represents the summaries generated under the conditions of the guidance signal G_i

Average value										
Factual consistency metrics	Datasets		ClozE	FactCC	DAE	SummaC	ANLI	G_eval	FEQA	Q2
	Replace	1	0.27	0.11	0.113	0.236	0.342	1.143	0.077	0.192
		2	0.27	0.11	0.113	0.236	0.342	1.14	0.067	0.18
		3	0.27	0.11	0.113	0.236	0.342	1.147	0.071	0.172
	Mask	1	0.631	0.183	0.482	0.24	0.795	3.62	0.068	0.174
		2	0.617	0.188	0.47	0.239	0.775	3.401	0.073	0.191
		3	0.604	0.173	0.449	0.238	0.718	3.183	0.071	0.188
	Reference	SLM upper bound	0.644	0.16	0.483	0.241	0.861	3.866	0.057	0.164
		Golden summary	0.668	0.2	0.541	0.242	0.928	4.929	0.062	0.183
Rouge	Datasets		Rouge-1			Rouge-2			Rouge-L	
			f	p	r	f	p	r	p	r
	Replace	1	0.354	0.306	0.428	0.194	0.163	0.244	0.284	0.396
		2	0.354	0.306	0.428	0.194	0.163	0.244	0.284	0.396
		3	0.354	0.306	0.428	0.194	0.163	0.244	0.284	0.396
	Mask	1	0.824	0.83	0.818	0.674	0.674	0.674	0.806	0.794
		2	0.797	0.803	0.791	0.638	0.638	0.637	0.779	0.769
		3	0.768	0.774	0.764	0.598	0.598	0.598	0.75	0.741
	Reference	SLM upper bound	0.844	0.85	0.839	0.7	0.701	0.7	0.826	0.815
		Golden summary	1	1	1	1	1	1	1	1
Variance										
Factual consistency metrics	Datasets		ClozE	FactCC	DAE	SummaC	anli	G_eval	FEQA	q2
	Replace	1	0.03	0.082	0.03	0.005	0.151	0.15	0.01	0.054
		2	0.03	0.082	0.03	0.005	0.151	0.172	0.009	0.046
		3	0.03	0.082	0.03	0.005	0.151	0.163	0.01	0.042
	Mask	1	0.044	0.141	0.082	0.004	0.109	2.129	0.01	0.05
		2	0.038	0.151	0.082	0.004	0.114	2.25	0.013	0.051
		3	0.046	0.14	0.085	0.004	0.141	2.316	0.009	0.054
	Reference	SLM upper bound	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		Golden summary	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Rouge	Datasets		Rouge-1			Rouge-2			Rouge-L	
			f	p	r	f	p	r	p	r
	Replace	1	0.031	0.025	0.046	0.026	0.019	0.039	0.024	0.046
		2	0.031	0.025	0.046	0.026	0.019	0.039	0.024	0.046
		3	0.031	0.025	0.046	0.026	0.019	0.039	0.024	0.046
	Mask	1	0.016	0.015	0.017	0.045	0.044	0.045	0.022	0.024
		2	0.017	0.016	0.018	0.045	0.045	0.046	0.023	0.025
		3	0.02	0.02	0.021	0.049	0.049	0.049	0.028	0.029
	Reference	SLM Upper bound	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		Golden summary	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Figure 8: Evaluation results of summaries guided by different semantic information.

Average Value										
	Datasets		ClozE	FactCC	DAE	SummaC	ANLI	G_eval	FEQA	Q2
Factual consistency metrics	Prase tree depth	4	0.369	0.066	0.145	0.233	0.343	1.267	0.059	0.095
		5	0.329	0.141	0.184	0.237	0.406	1.381	0.063	0.109
		6	0.345	0.084	0.149	0.232	0.432	1.397	0.058	0.1
		7	0.346	0.043	0.122	0.228	0.322	1.251	0.058	0.112
		8	0.352	0.065	0.149	0.229	0.423	1.298	0.064	0.107
		9	0.389	0.133	0.189	0.234	0.511	1.478	0.06	0.113
	Top-level suntactic structure	SVO	0.373	0.109	0.165	0.232	0.502	1.404	0.063	0.111
		SLVSC	0.37	0.129	0.14	0.233	0.409	1.349	0.063	0.104
		SVIODO	0.381	0.102	0.141	0.234	0.43	1.318	0.062	0.108
		SVOOC	0.349	0.079	0.13	0.23	0.386	1.285	0.059	0.107
		ES	0.355	0.114	0.156	0.235	0.506	1.404	0.063	0.106
		PS	0.364	0.1	0.173	0.235	0.474	1.484	0.06	0.11
		SVI	0.355	0.11	0.132	0.238	0.467	1.392	0.061	0.092
		SVNC	0.378	0.092	0.178	0.228	0.351	1.276	0.059	0.105
		SVAC	0.402	0.082	0.153	0.229	0.405	1.25	0.057	0.097
	Modifier number	1	0.342	0.141	0.163	0.244	0.401	1.43	0.063	0.112
		2	0.363	0.2	0.179	0.243	0.382	1.347	0.058	0.105
		3	0.335	0.075	0.151	0.236	0.375	1.309	0.061	0.11
		4	0.335	0.132	0.164	0.237	0.406	1.434	0.062	0.109
		5	0.336	0.09	0.181	0.231	0.448	1.397	0.058	0.106
		6	0.383	0.088	0.178	0.234	0.432	1.402	0.062	0.106
		7	0.328	0.071	0.168	0.228	0.412	1.385	0.062	0.109
		8	0.345	0.08	0.156	0.231	0.455	1.378	0.062	0.117
		9	0.346	0.048	0.152	0.229	0.416	1.244	0.059	0.103
	Top similar sentences from document	0	0.401	0.08	0.271	0.227	0.333	1.436	0.061	0.11
		1	0.466	0.11	0.314	0.231	0.321	1.62	0.055	0.094
		2	0.409	0.08	0.254	0.229	0.377	1.459	0.06	0.106
		3	0.434	0.07	0.255	0.232	0.359	1.511	0.049	0.098
		4	0.453	0.1	0.26	0.234	0.435	1.67	0.051	0.084
	Reference	SLM upper bound	0.642	0.17	0.478	0.242	0.859	3.827	0.055	0.092
		Golden summary	0.668	0.2	0.541	0.242	0.928	4.934	0.061	0.133
Rouge	Datasets		R1			R2			RL	
			f	p	r	f	p	r	p	r
	Prase tree depth	4	0.344	0.375	0.327	0.098	0.108	0.093	0.267	0.233
		5	0.36	0.384	0.352	0.097	0.102	0.096	0.267	0.247
		6	0.357	0.369	0.358	0.092	0.094	0.095	0.255	0.251
		7	0.342	0.33	0.367	0.091	0.087	0.1	0.226	0.253
		8	0.352	0.342	0.376	0.092	0.088	0.102	0.235	0.26
		9	0.368	0.361	0.391	0.096	0.093	0.105	0.247	0.269
	Top-level suntactic structure	SVO	0.331	0.314	0.363	0.081	0.075	0.093	0.215	0.25
		SLVSC	0.332	0.336	0.342	0.081	0.081	0.085	0.233	0.239
		SVIODO	0.338	0.333	0.358	0.084	0.08	0.094	0.224	0.244
		SVOOC	0.334	0.319	0.366	0.082	0.078	0.093	0.205	0.239
		ES	0.337	0.329	0.36	0.083	0.08	0.091	0.22	0.243
		PS	0.387	0.4	0.389	0.109	0.111	0.112	0.276	0.27
		SVI	0.362	0.364	0.376	0.096	0.095	0.101	0.244	0.255
		SVNC	0.347	0.33	0.382	0.085	0.079	0.096	0.226	0.261
		SVAC	0.335	0.321	0.366	0.076	0.072	0.086	0.217	0.25
	Modifier number	1	0.354	0.432	0.317	0.1	0.122	0.091	0.315	0.23
		2	0.334	0.403	0.3	0.097	0.118	0.088	0.298	0.22
		3	0.35	0.378	0.34	0.099	0.107	0.096	0.268	0.24
		4	0.362	0.393	0.346	0.102	0.11	0.098	0.279	0.246
		5	0.359	0.364	0.368	0.094	0.094	0.1	0.252	0.257
		6	0.356	0.368	0.357	0.089	0.091	0.092	0.254	0.248
		7	0.345	0.338	0.365	0.095	0.092	0.103	0.232	0.253
		8	0.338	0.331	0.36	0.087	0.083	0.096	0.23	0.252
		9	0.336	0.309	0.38	0.082	0.074	0.096	0.206	0.256
	Top similar sentences from document	0	0.335	0.302	0.394	0.093	0.081	0.116	0.213	0.282
		1	0.352	0.341	0.387	0.099	0.095	0.113	0.24	0.272
		2	0.346	0.327	0.381	0.098	0.092	0.11	0.234	0.27
		3	0.333	0.322	0.362	0.081	0.077	0.091	0.221	0.25
		4	0.332	0.329	0.355	0.086	0.084	0.093	0.232	0.251
	Reference	SLM upper bound	0.848	0.855	0.842	0.701	0.701	0.701	0.824	0.813
		Golden summary	1	1	1	1	1	1	1	1

Figure 9: Evaluation results (Average value) of summaries guided by different syntactic information.

Variance										
	Datasets		ClozE	FactCC	DAE	SummaC	ANLI	G_eval	FEQA	Q2
Factual consistency metrics	Prase tree depth	4	0.05	0.055	0.031	0.004	0.18	0.301	0.005	0.019
		5	0.042	0.106	0.033	0.004	0.193	0.442	0.008	0.024
		6	0.039	0.07	0.025	0.003	0.19	0.459	0.007	0.027
		7	0.034	0.038	0.017	0.003	0.166	0.246	0.007	0.031
		8	0.032	0.067	0.034	0.003	0.181	0.3	0.007	0.026
		9	0.038	0.119	0.037	0.003	0.199	0.536	0.006	0.027
	Top-level suntactic structure	SVO	0.031	0.112	0.026	0.003	0.201	0.416	0.007	0.03
		SLVSC	0.045	0.117	0.028	0.004	0.181	0.347	0.007	0.026
		SVIODO	0.028	0.089	0.027	0.004	0.19	0.284	0.007	0.023
		SVOOC	0.029	0.078	0.019	0.003	0.192	0.259	0.007	0.029
		ES	0.029	0.093	0.028	0.003	0.198	0.428	0.007	0.02
		PS	0.035	0.089	0.033	0.004	0.199	0.596	0.006	0.028
		SVI	0.035	0.115	0.025	0.005	0.193	0.345	0.007	0.027
		SVNC	0.031	0.091	0.029	0.003	0.176	0.272	0.007	0.022
		SVAC	0.031	0.083	0.025	0.002	0.191	0.225	0.006	0.022
	Modifier number	1	0.058	0.131	0.04	0.004	0.194	0.681	0.006	0.023
		2	0.052	0.153	0.045	0.005	0.186	0.429	0.007	0.029
		3	0.044	0.074	0.038	0.004	0.19	0.352	0.006	0.028
		4	0.042	0.115	0.029	0.003	0.187	0.531	0.006	0.027
		5	0.039	0.078	0.04	0.004	0.194	0.427	0.006	0.028
		6	0.04	0.091	0.039	0.004	0.194	0.463	0.007	0.023
		7	0.036	0.059	0.03	0.002	0.192	0.425	0.007	0.025
		8	0.035	0.076	0.024	0.003	0.191	0.405	0.007	0.034
		9	0.032	0.043	0.023	0.002	0.198	0.229	0.005	0.026
	Top similar sentences from document	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		4	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Reference	SLM upper bound	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		Golden summary	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Rouge	Datasets		R1			R2			RL	
			f	p	r	f	p	r	p	r
	Prase tree depth	4	0.016	0.021	0.017	0.007	0.009	0.007	0.015	0.011
		5	0.013	0.018	0.015	0.007	0.008	0.008	0.012	0.012
		6	0.015	0.018	0.019	0.007	0.007	0.008	0.011	0.013
		7	0.014	0.015	0.017	0.005	0.005	0.007	0.009	0.011
		8	0.012	0.015	0.016	0.007	0.007	0.009	0.01	0.013
		9	0.012	0.015	0.016	0.007	0.006	0.008	0.011	0.013
	Top-level suntactic structure	SVO	0.013	0.015	0.015	0.006	0.005	0.007	0.008	0.011
		SLVSC	0.013	0.017	0.016	0.005	0.005	0.006	0.009	0.009
		SVIODO	0.014	0.018	0.015	0.007	0.006	0.008	0.01	0.011
		SVOOC	0.013	0.016	0.014	0.005	0.005	0.007	0.008	0.008
		ES	0.013	0.016	0.014	0.006	0.006	0.007	0.01	0.01
		PS	0.013	0.018	0.016	0.008	0.008	0.01	0.012	0.012
		SVI	0.014	0.019	0.016	0.007	0.007	0.008	0.011	0.012
		SVNC	0.013	0.015	0.016	0.005	0.005	0.007	0.009	0.011
		SVAC	0.012	0.016	0.014	0.005	0.005	0.006	0.009	0.009
	Modifier number	1	0.017	0.025	0.02	0.01	0.015	0.009	0.021	0.014
		2	0.02	0.025	0.022	0.009	0.012	0.008	0.018	0.013
		3	0.016	0.02	0.019	0.007	0.009	0.007	0.015	0.012
		4	0.017	0.021	0.018	0.009	0.009	0.009	0.013	0.013
		5	0.014	0.017	0.018	0.007	0.007	0.008	0.011	0.013
		6	0.015	0.018	0.017	0.007	0.008	0.008	0.012	0.012
		7	0.013	0.014	0.017	0.007	0.006	0.009	0.009	0.011
		8	0.013	0.014	0.018	0.006	0.006	0.008	0.009	0.012
		9	0.012	0.012	0.017	0.005	0.005	0.008	0.007	0.011
	Top similar sentences from document	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		4	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Reference	SLM upper bound	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		Golden summary	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Figure 10: Evaluation results (Variance) of summaries guided by different syntactic information.