From Experts to Novices: Controllable Abstractive Summarization via Audience-Aware Modeling

Aarush Verulkar
verulkar@usc.edu

Aditya Ramachandran
adityar2@usc.edu

Ashwath Subash
asubash@usc.edu

Daryl Christi
dkchrist@usc.edu

Siddhi Galada
galada@usc.edu

Venkataraman V
vv70613@usc.edu

Abstract

We propose a multi-task learning framework that addresses both factual consistency and audience adaptation through audience-aware summarization. Our two-stage pipeline uses sequence-to-sequence Transformer models. In the first stage, a T5-small model fine-tuned on WikiLarge simplifies the source text sentence-by-sentence to reduce linguistic complexity. The second stage employs a T5-base model, fine-tuned on the CNN/DailyMail dataset, to generate concise, factually grounded summaries from this pre-simplified context. To ensure the reliability of our system, we employ rigorous evaluation protocols using external factuality models, specifically FactCC (based on RoBERTa) and QAGS (based on Valhalla T5 Base) to verify the factual accuracy and audience suitability of the outputs. Experiments demonstrate that our pipeline successfully reduces the Flesch-Kincaid Grade Level from 8-10 to 6-7, significantly enhancing accessibility for non-expert readers. Crucially, this simplification does not compromise accuracy. Our models maintain FactCC scores between 0.91 and 0.99, ensuring that simplified outputs remain semantically faithful to the source.

1 Introduction

Text summarization is a fundamental natural language processing task that automatically generates concise representations of longer documents while preserving key information. However, modern text summarization systems face two critical challenges that limit their real-world deployment. The first challenge is factual inconsistency where generated summaries contain information not supported by the source document. The second challenge is a lack of audience adaptation where summaries fail to adjust their complexity and style for different target readers. These limitations prevent summarization models from functioning effectively in applications that require both accuracy and audience-specific communication.

Consider real-world scenarios where both faithfulness and audience adaptation are crucial. Medical professionals need accurate and technical summaries of research papers while patients require simplified and accessible explanations of the same information. Educational platforms also need to generate grade-appropriate summaries of complex texts without losing factual accuracy. Current approaches often treat style control and factual accuracy as separate problems. This separation creates a trade-off where improving one aspect often hurts the other.

We propose a multi-task learning framework that addresses both challenges simultaneously through audience-aware faithful summarization. Our approach uses a two-stage pipeline to ensure both accuracy and accessibility. We first apply a simplification model trained on the WikiLarge dataset to adapt the linguistic complexity of the source text. Subsequently, we utilize a summarization model fine-tuned on the CNN/DailyMail dataset to perform faithful content selection on the simplified input. We employ audience-specific prompts to guide summary generation toward target readers while leveraging advanced factuality models such as FactCC and QAGS to validate the correctness of our generated outputs.

Our contributions are threefold:

1. We present a unified framework for audience-controlled factual summarization.
2. We provide a comprehensive evaluation across multiple audiences.
3. We analyze the interaction between style control and factual consistency.

2 Related Works

2.1 Persona and Controllable Summarization

Persona- and role-aware summarization aims to adapt generated content to the needs of specific stakeholders or perspectives. Mullick et al. (2024) explore this in the context of healthcare, modeling distinct personas such as doctors, patients, and

laypeople. They fine-tune a small, domain-adapted LLM on a healthcare corpus and introduce an AI-based critiquing pipeline for summary evaluation. Their results show strong alignment between AI critiques and human judgments, demonstrating the feasibility of role-specific summarization that selectively adapts content to different personas.

In dialogue summarization, modeling role interactions is crucial for coherence. Lin et al. (2022) in Other Roles Matter! generate role-specific summaries (e.g., for merchants and consumers) while modeling inter-role dependencies using cross-attention to select relevant utterances and decoder self-attention to integrate information from other role summaries, addressing prior methods that treat roles independently. Liu and Chen (2021) present controllable neural dialogue summarization with personal named entity planning, which allows the model to focus on a specific persona or generate a comprehensive view by planning entity occurrences and using coreference information, thus maintaining coherence while adapting the perspective of the summary. Liang et al. (2022) introduce a Role-Aware Centrality (RAC) model that assigns the importance of the sentence based on both the perspective of a role and the relevance between roles, helping to prioritize content for specific stakeholders. More recently, Zhang et al. (2025) proposed Rehearse With User, a framework in which LLMs engage in role-playing and iterative feedback with a virtual user to personalize summary, ensuring that outputs align with user interests and stakeholder needs. Collectively, these studies illustrate the evolution of persona-based summarization, highlighting methods for controlling perspective, incorporating inter-role dependencies, and producing stakeholder-aligned outputs.

2.2 Factual Consistency for Summarization

Dixit et al. (2023) introduce EFACTSUM, a two-stage framework for improving factual consistency in abstractive summarization. It first generates multiple candidate summaries, then applies a contrastive ranking objective combining factuality and similarity metrics to favor accurate yet fluent summaries. Similarly, Wang et al. (2023) present an element-aware framework that integrates expert-defined content elements like entities, events, and results into both evaluation and generation to reduce noise, hallucination, and redundancy. Finally, Feng et al. (2024) propose Contrastive Preference Optimization (CPO), a fine-tuning method using contrastive pairs of faithful and fabricated summaries with probing-based supervision to detect inconsistencies. Compared to reinforcement learning approaches, CPO offers more stable training, lower annotation costs, and improved factual consistency without sacrificing fluency.

3 Methodology

We propose a comprehensive two-stage pipeline designed to align abstractive summarization with audience readability levels. Unlike standard summarization systems that optimize solely for information retention, our system introduces a control mechanism to generate two distinct outputs per input article: a *Normal Summary* for expert readers and a *Simplified Summary* tailored for novices or non-native speakers.

3.1 Pipeline Architecture

The core of our approach is a bifurcated workflow that integrates text simplification with abstractive summarization. The pipeline processes raw input text through a preprocessing stage before branching based on the desired output complexity.

Our system leverages the T5 (Text-to-Text Transfer Transformer) encoder–decoder architecture for both tasks due to its state-of-the-art performance in sequence-to-sequence generation.

- **Summarizer Model:** We utilize a T5-Base model fine-tuned on the **CNN/DailyMail 3.0.0** dataset. This component is responsible for condensing the content while retaining high semantic density. The CNN/DailyMail dataset, consisting of news articles, provides a robust foundation for abstractive summarization tasks.

- **Simplifier Model:** For the simplification task, we employ a T5-Small model fine-tuned on the **WikiLarge** dataset. This dataset consists of aligned pairs of complex and simplified Wikipedia sentences, allowing the model to learn transformations that reduce syntactic complexity without altering the core meaning.

3.2 Execution Workflow

The input text undergoes distinct processing stages depending on the target complexity level:

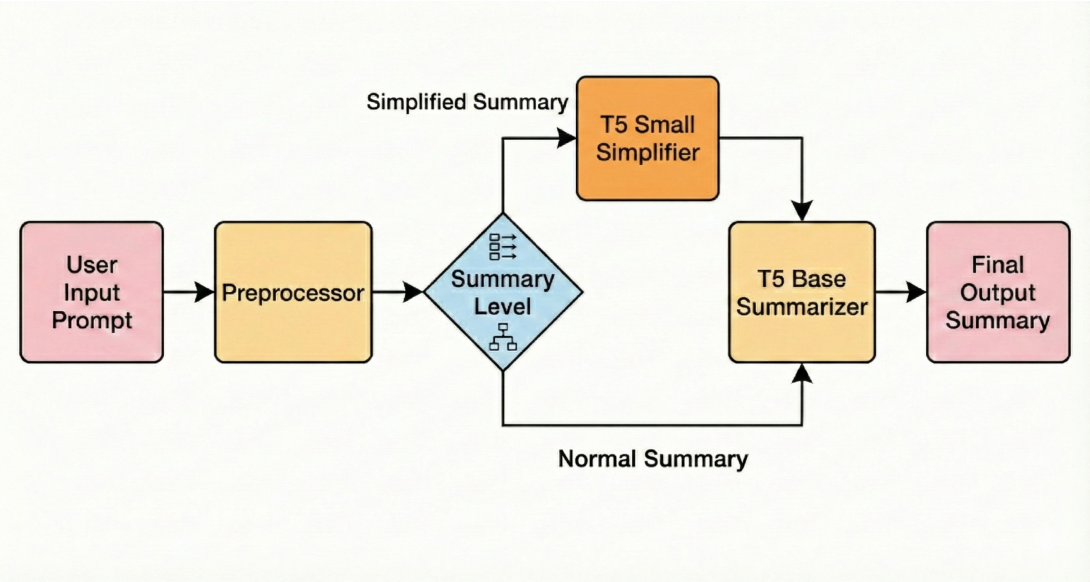

Figure 1: Two stage summarizer model

3.2.1 Preprocessing

Raw input is first cleaned to remove extraneous artifacts. For the simplification track, we utilize the Natural Language Toolkit (NLTK) to perform sentence segmentation, ensuring that the simplifier operates on granular linguistic units rather than full paragraphs.

3.2.2 Normal Summary Generation

For the standard output, the cleaned original article is passed directly into the *Summarizer* model. This path prioritizes information density and structural fidelity to the source text.

$$S_{normal} = \text{Summarizer}(T_{original}) \quad (1)$$

3.2.3 Simplified Summary Generation

To generate the audience-aware output, we implement a "Simplify-then-Summarize" strategy. This intermediate step ensures that the dense vocabulary and complex syntax often found in technical or news sources are mitigated before the summarization logic is applied.

1. **Text Simplification:** The segmented source text is processed sentence-by-sentence by the *Simplifier* model. This transforms the input into a lower-readability version while preserving factual content.

2. **Abstractive Summarization:** The fully simplified article is then fed into the *Summarizer* model. Because the input context is already simplified, the resulting summary naturally inherits simpler linguistic features while maintaining the abstractive nature of the T5 output.

$$S_{simple} = \text{Summarizer}(\text{Simplifier}(T_{original})) \quad (2)$$

This modular design allows for independent optimization of the simplification and summarization objectives, ensuring that readability improvements do not come at the cost of factual consistency.

4 Evaluation Metrics

We evaluate summaries on three axes: (1) Content Quality, (2) Factual Consistency, and (3) Audience Alignment.

4.1 Content Quality

We report ROUGE-N (Eq. 3) to measure n-gram (unigram, bigram, and LCS) overlap between the generated (S_{gen}) and reference (S_{ref}) summaries.

$$\text{ROUGE-N} = \frac{\sum_{\text{gram}n \in S\text{ref}} \text{Countmatch}(\text{gram}n)}{\sum \text{gram}n \in S\text{ref} \text{Count}(\text{gram}n)}. \quad (3)$$

To capture semantic similarity, we also report BERTScore (F1). It computes recall (Eq.4) and precision (Eq.5) based on the cosine similarity of token embeddings, combined into an F1 score (Eq.6).

$$\text{R}_{\text{BERT}} = \frac{1}{|S\text{gen}|} \sum_{x_i \in S_{\text{gen}}} \max_{y_j \in S_{\text{ref}}} (x_i^\top y_j) \quad (4)$$

$$\mathrm{P_{BERT}} = \frac{1}{|S_{\text{ref}}|} \sum_{y_j \in S_{\text{ref}}} \max_{x_i \in S_{\text{gen}}} (x_i^\top y_j) \tag{5}$$

$$\mathrm{F1_{BERT}} = 2 \cdot \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \tag{6}$$

BLEU In addition to ROUGE, we report BLEU (Bilingual Evaluation Understudy) to measure n-gram precision between the generated summary (S_{gen}) and the reference summary (S_{ref}). BLEU computes a geometric mean of modified n-gram precisions, combined with a brevity penalty (BP) to discourage overly short outputs.

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{7}$$

where p_n denotes the modified precision for n-grams of size n, w_n are uniform weights (typically $w_n = \frac{1}{N}$), and BP is defined as:

$$\mathrm{BP} = \begin{cases} 1, & \text{if } |S_{\text{gen}}| > |S_{\text{ref}}| \\ \exp\left(1 - \frac{|S_{\text{ref}}|}{|S_{\text{gen}}|}\right), & \text{otherwise} \end{cases} \tag{8}$$

BLEU primarily emphasizes surface-level overlap and fluency, making it complementary to recall-focused metrics such as ROUGE.

4.2 Factual Consistency (Faithfulness)

Factual consistency evaluates whether a generated summary preserves facts stated in the source document without introducing hallucinations or unsupported claims. We employ two complementary automatic metrics: **FactCC** and **QAGS**.

FactCC. FactCC is a supervised factual consistency metric formulated as a natural language inference (NLI) task. Given a source document D_{src} and a generated summary S_{gen}, FactCC predicts whether the summary is *entailed* by the source, or whether it contains factual inconsistencies such as contradictions or unsupported statements.

Formally, FactCC is a binary classifier:

$$\mathrm{FactCC}(D_{\text{src}}, S_{\text{gen}}) = \begin{cases} 1, & \text{if } D_{\text{src}} \models S_{\text{gen}} \\ 0, & \text{otherwise} \end{cases}$$

where $\models$ denotes textual entailment. A summary is considered *faithful* only if all of its claims are entailed by the source document. The final FactCC score is computed as the proportion of summaries classified as faithful.

QAGS. QAGS (Question Answering for Generating Summaries) evaluates factual consistency by measuring answer agreement to questions derived from the generated summary. Unlike FactCC, QAGS does not require labeled entailment data and instead relies on question answering models.

Given a generated summary S_{gen}, a set of questions $\mathcal{Q} = \{q_1, q_2, \ldots, q_n\}$ is automatically generated from its content. For each question q_i, answers are extracted independently from the summary and the source document:

$$a_i^{(S)} = \mathrm{QA}(q_i, S_{\text{gen}}), \quad a_i^{(D)} = \mathrm{QA}(q_i, D_{\text{src}})$$

The consistency score for each question is computed using a token-level F1 overlap between $a_i^{(S)}$ and $a_i^{(D)}$. The overall QAGS score is then defined as:

$$\mathrm{QAGS}(D_{\text{src}}, S_{\text{gen}}) = \frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} \mathrm{F1}(a_i^{(S)}, a_i^{(D)})$$

Higher QAGS scores indicate stronger factual alignment between the summary and the source document. By focusing on answer consistency, QAGS captures fine-grained factual errors that may not be detected by entailment-based classifiers.

4.3 Audience Alignment

To measure style adaptation, we report the Flesch-Kincaid Grade Level (FKGL) (Eq.9) to estimate the text's readability.

$$\mathrm{FKGL} = (0.39 \times \mathrm{ASL}) + (11.8 \times \mathrm{ASW}) - 15.59 \tag{9}$$

where ASL = Average Sentence Length and ASW = Average Word length

4.4 Simplification

SARI To evaluate text simplification quality, we report SARI (System output Against References and the Input). Unlike traditional overlap-based metrics, SARI explicitly measures how well the model edits the source document (D_{src}) to produce a simplified summary (S_{gen}), relative to reference simplifications (S_{ref}).

SARI evaluates three operations: *Add*, *Delete*, and *Keep*. The final score is computed as the average of F1 scores for these operations across n-grams:

$$\mathrm{SARI} = \frac{1}{3}\left(F1_{\text{Add}} + F1_{\text{Keep}} + F1_{\text{Delete}}\right) \tag{10}$$

where each $F1$ score balances precision and recall for the corresponding operation. By explicitly rewarding appropriate deletions and additions, SARI aligns well with human judgments of simplicity and readability.

5 Results and Discussion

The model evaluation began with an assessment of the individual component models to establish a baseline for performance. The T5-base summarization model, fine-tuned on the CNN/DailyMail dataset, demonstrated strong capabilities in **information retention** and **semantic similarity**. As detailed in Table 1, the model achieved a **ROUGE-1** score of **40.46** and a **BERTScore** of **88.36**, indicating a high degree of overlap with reference summaries and preservation of meaning.

Metric	Score
ROUGE-1	40.46
ROUGE-2	18.63
ROUGE-L	28.42
ROUGE-Lsum	37.70
BERTScore F1	88.36

Table 1: Summarization Model Metric Scores

The T5-small simplification model, fine-tuned on the WikiLarge dataset, was evaluated for its ability to adapt text complexity. The model demonstrated effective performance as detailed in Table 2 with a **SARI** score of **51.38**, a standard metric to measure semantic similarity, and a **BLEU** score of **37.23**, which measures exact word overlap. Both of these results confirm that the foundational blocks of the pipeline were sufficiently robust to support subsequent audience-aware generation tasks.

Metric	Score
SARI	51.38
ROUGE-1	62.65
ROUGE-2	45.87
ROUGE-L	58.69
ROUGE-Lsum	58.65
BLEU	37.23

Table 2: Simplification Model Metric Scores

To measure the effectiveness of the pipeline in controlling syntactic complexity, we analyzed readability through the **Flesch-Kincaid Grade Level (FKGL)** metric using eq. 9 on a sample of 10 articles. The findings revealed a distinct separation between the two output types. While the normal summaries typically produced scores that indicated a reading grade level of **8 to 10**, the simplified summaries consistently achieved lower scores in the **6 to 7 range**. These results indicate that the two-stage pipeline successfully simplifies sentence structures and reduces reading levels, thereby enhancing accessibility for non-expert audiences.

Finally, we examined whether the simplification process compromised the accuracy of the content by evaluating factual consistency using FactCC and semantic faithfulness using QAGS. The results demonstrated that the pipeline maintains strong factual grounding, with both normal and simplified summaries achieving high FactCC scores ranging from **0.91 to 0.99**. Regarding semantic faithfulness, the simplified summaries actually outperformed the normal versions slightly, achieving QAGS values of approximately **0.62 to 0.64** compared to the normal summaries **0.60 to 0.62**. This suggests that simplification does not degrade factual correctness or introduce hallucinations.

Metric	Level	Score
FKGL	Normal	8–10
FKGL	Simplified	6–7
FactCC	Normal	0.91–0.99
FactCC	Simplified	0.91–0.99
QAGS	Normal	0.60–0.62
QAGS	Simplified	0.62–0.64

Table 3: Readability and Factual Consistency Evaluation of the Two-Stage Pipeline

6 Conclusion

This work presents a two-stage pipeline for controllable abstractive summarization that addresses both factual consistency and audience adaptation. By combining text simplification with summarization, we demonstrate that summaries can be tailored to different reading levels without sacrificing factual accuracy. This capability is essential for real-world applications in healthcare, education, and news dissemination where both precision and accessibility matter.

Our "Simplify-then-Summarize" approach validates an important architectural insight: separating simplification and summarization into distinct stages avoids the trade-offs present in end-to-end systems. The simplifier focuses on linguistic accessibility while the summarizer prioritizes infor-

mation extraction. This modular design proves effective, as simplified summaries achieve substantially lower reading grade levels while maintaining strong factual grounding.

Interestingly, we found that simplification does not inherently reduce factual quality. In several cases, simplified summaries actually showed improved consistency, suggesting that removing linguistic complexity may help the model focus on core factual content. This challenges the assumption that accessibility and accuracy must compete, and instead points toward their potential compatibility when approached thoughtfully.

As summarization systems become more prevalent in real-world applications, adapting content for diverse audiences while maintaining truthfulness becomes increasingly important. Our two-stage pipeline establishes that audience-aware summarization and factual faithfulness can be achieved simultaneously through careful design, paving the way for more accessible and trustworthy summarization.

7 Future Scope

Our system currently produces binary outputs (normal vs. simplified), but real-world audiences exist on a spectrum of expertise. Techniques such as adapter modules, prefix tuning, or controllable generation could support continuous audience scales or multi-dimensional style profiles that adjust vocabulary, sentence complexity, and technical depth independently.

Also, human evaluation remains essential for validating practical utility. While automated metrics provide valuable signals, they cannot fully capture whether simplified summaries genuinely serve novice readers or non-native speakers. User studies with actual target audiences would provide crucial insights and reveal gaps that computational metrics may miss.

References

Tanay Dixit, Fei Wang, and Muhao Chen. 2023. Improving factuality of abstractive summarization without sacrificing summary quality. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 902–913, Toronto, Canada. Association for Computational Linguistics.

Huawen Feng, Yan Fan, Xiong Liu, Ting-En Lin, Zekun Yao, Yuchuan Wu, Fei Huang, Yongbin Li, and Qianli Ma. 2024. Improving factual consistency of news summarization by contrastive preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11084–11100, Miami, Florida, USA. Association for Computational Linguistics.

Xinnian Liang, Chao Bian, Shuangzhi Wu, and Zhoujun Li. 2022. Towards modeling role-aware centrality for dialogue summarization. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 43–50, Online only. Association for Computational Linguistics.

Haitao Lin, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2022. Other roles matter! enhancing role-oriented dialogue summarization via role interactions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2558, Dublin, Ireland. Association for Computational Linguistics.

Zhengyuan Liu and Nancy Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 92–106, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ankan Mullick, Sombit Bose, Rounak Saha, Ayan Bhowmick, Pawan Goyal, Niloy Ganguly, Prasenjit Dey, and Ravi Kokku. 2024. On the persona-based summarization of domain-specific documents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14291–14307, Bangkok, Thailand. Association for Computational Linguistics.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.

Yanyue Zhang, Yulan He, and Deyu Zhou. 2025. Rehearse with user: Personalized opinion summarization via role-playing based on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15194–15211, Vienna, Austria. Association for Computational Linguistics.