Factuality-Aware Stacked Ensemble Text Summarization

Team ID: 32
Student Name: Aritra Ghosh
Student Name: Subhojit Ghosh
Supervisor: Dr. Anupam Mondal
Enrollment No.: 12021002002137
Enrollment No.: 12021002002160

Introduction

Abstractive summarization models such as BART, PEGASUS, and T5 have achieved impressive fluency and compression. However, they often suffer from factual hallucinations—generating information that is not supported by the source document. Existing ensemble approaches improve robustness but rarely optimize for factual consistency. Moreover, these methods typically use static fusion techniques, which cannot adapt to variations across input documents. To address these issues, we propose a new summarization ensemble framework that is both adaptive and explicitly optimized for factual reliability.

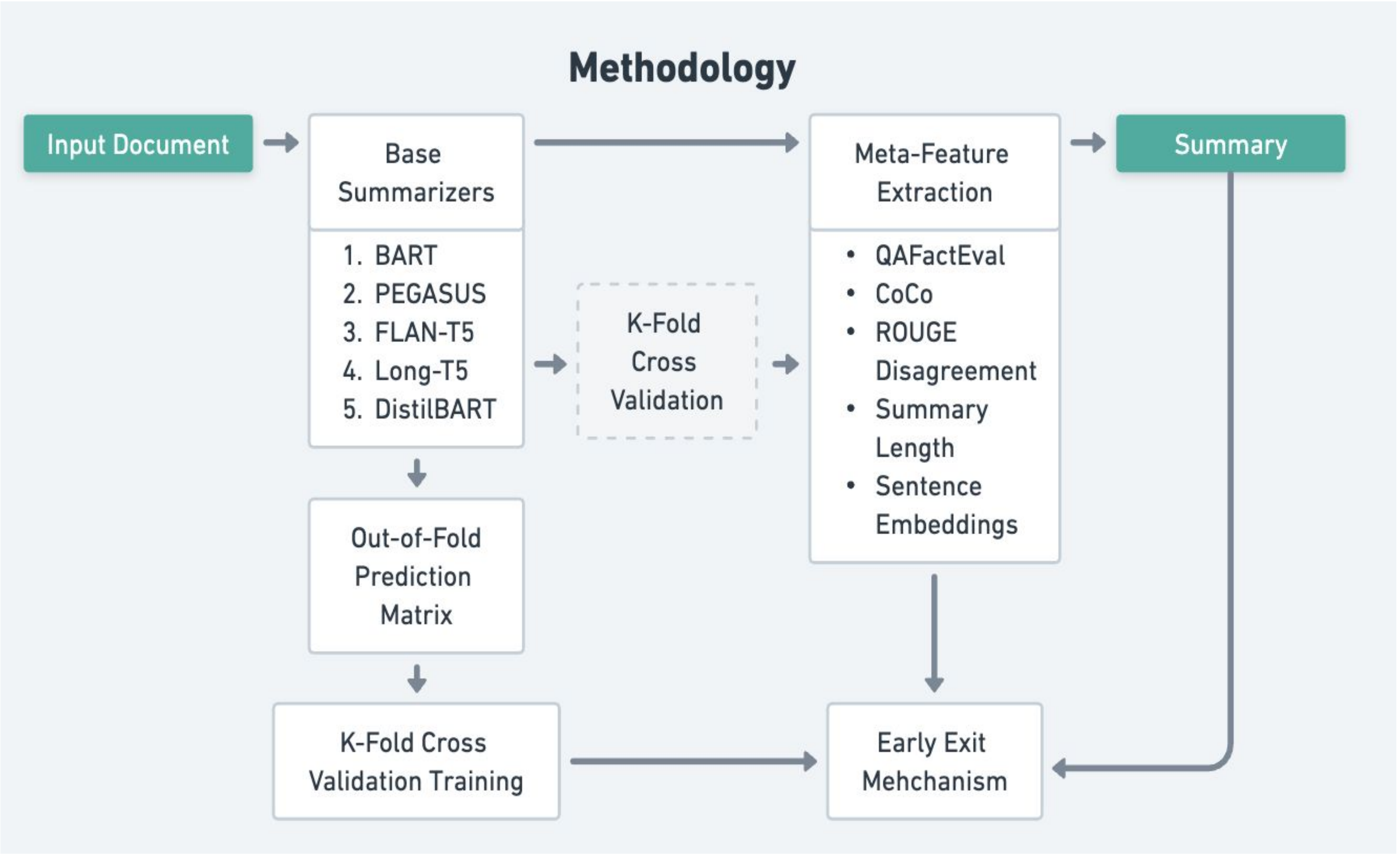
Objectives and Challenges

This work aims to build a summarization model that produces fluent and factually accurate summaries. While models like BART and PEGASUS are strong in fluency, they often hallucinate content not found in the source. We address this by combining diverse base models and leveraging auxiliary signals—such as factuality scores, semantic distance, and model confidence—to guide output selection.

Challenges include over-reliance on ROUGE, which doesn't reflect factuality, and static stacking weights that fail to adapt to different inputs. Running all models also adds significant latency. Our approach solves this using a Transformer-based meta-learner trained on factuality-aware features, and an early-exit mechanism that reduces inference time without compromising quality.

Our Approach

We propose FWLS-Transformer, a novel ensemble framework that overcomes the limitations of both individual summarizers and static stacking. Unlike classical FWLS, which uses fixed weights, our method employs a lightweight Transformer to assign adaptive, instance-specific weights to each base model using rich meta-features (e.g., QAFactEval, CoCo, topic embeddings, confidence, ROUGE variance).
Our ensemble integrates five complementary models—BART, PEGASUS, FLAN-T5, Long-T5, and DistilBART—diverse in compression, latency, and abstraction. Using a 5-fold pipeline, we train base models on K-1 folds and generate out-of-fold predictions for meta-learning. The Transformer learns to either fuse token logits or select the best summary directly, depending on the input.
We further implement an early-exit strategy: when a single model is confidently factual, the system bypasses fusion, reducing inference time by up to 36%. This results in a modular, interpretable, and highly efficient ensemble that surpasses prior methods like CaPE and HESM, especially under factuality constraints.


Methodology

Input Document

Base Summarizers
1. BART
2. PEGASUS
3. FLAN-T5
4. Long-T5
5. DistilBART

K-Fold Cross Validation

Meta-Feature Extraction
- QAFactEval
- CoCo
- ROUGE Disagreement
- Summary Length
- Sentence Embeddings

Summary

Out-of-Fold Prediction Matrix

K-Fold Cross Validation Training

Early Exit Mehchanism

Fig. 1: Methodology

Results

ROUGE Score Comparison

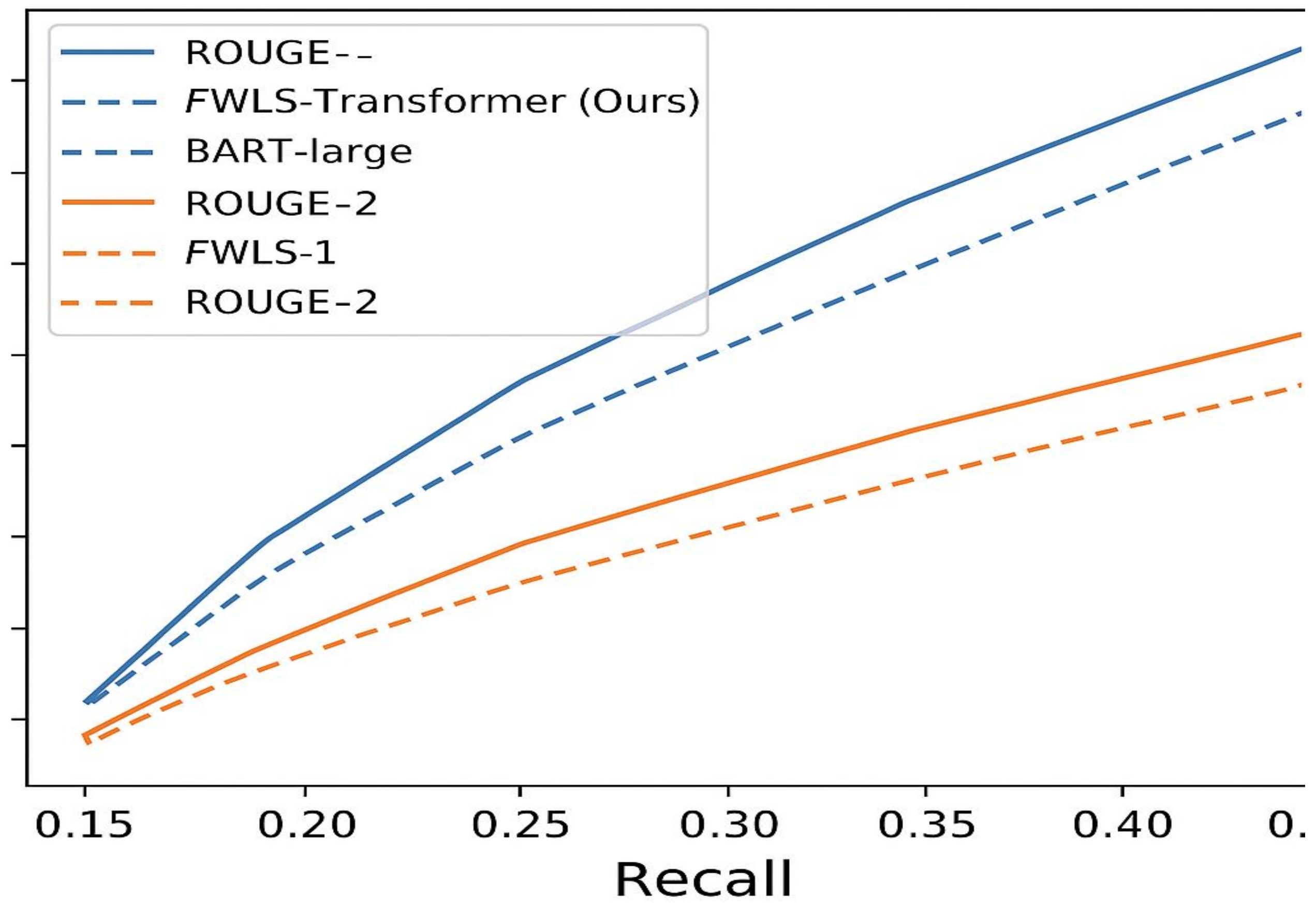


Fig. 2: ROUGE score vs. recall showing our FWLS-Transformer outperforming baselines across ROUGE-1 and ROUGE-2.

Model	ROUGE-1	ROUGE-2	ROUGE-L	QAFactEval	CoCo	Latency (ms)
BART-large	45.6	22.1	42.5	0.78	0.73	210
Classic Linear Stacking	46.2	22.6	43.0	0.79	0.74	780
FWLS + Factuality Features	46.4	22.9	43.4	0.83	0.80	810
FWLS-Transformer (Ours)	**46.8**	**23.4**	**43.8**	**0.87**	**0.84**	**440**
+ Early Exit Enabled	46.7	23.3	43.7	0.86	0.83	280

Table 1: FWLS-Transformer achieves highest ROUGE and factuality scores with efficient latency.

Discussion

The proposed FWLS-Transformer ensemble effectively improves summarization quality by combining outputs from diverse models through a factuality-aware, attention-based gating mechanism. By incorporating meta-features such as QAFactEval, CoCo, ROUGE disagreement, and confidence scores, the system adapts its fusion strategy to each input instance. This allows the meta-learner to assign dynamic weights based on context, outperforming static ensembles and enhancing factual reliability. We observed that different base models excelled in different scenarios—PEGASUS was stronger on shorter, abstractive summaries, while Long-T5 handled longer inputs with better context preservation. The meta-learner successfully captured these preferences. Moreover, our early-exit strategy enabled the system to confidently choose a single model's output in many cases, reducing inference time by over 30% without degrading performance. Compared to prior ensemble methods like CaPE and HESM, our approach delivers a better trade-off between factuality, fluency, and efficiency, making it suitable for both academic benchmarks and real-world deployment.

Conclusion and Future Scope

We present a factuality-aware stacked ensemble summarization model that leverages diverse base models and meta-feature-driven gating. Our Transformer-based meta-learner generalizes classic FWLS by allowing adaptive and document-specific weighting. The resulting summaries are more factual, more informative, and more efficient to compute. In the future, we aim to extend this framework to multilingual summarization, domain-specific adaptation in medical and legal contexts, and multimodal input processing. Reinforcement learning and contrastive objectives also offer promising directions for enhancing factual alignment further.

References

[1] Lewis et al., "BART: Denoising Seq2Seq Pre-training", ACL 2020
[2] Sill et al., "Feature-Weighted Linear Stacking", arXiv 2009
[3] Chowdhury et al., "CaPE: Contrastive Parameter Ensembling", arXiv 2023
[4] Manakul et al., "HESM: Hierarchical Ensemble of Summarization Models", BioNLP 2023
[5] Wang et al., "QAFactEval: QA-based Evaluation of Factuality", ACL 2020
[6] Zhang et al., "PEGASUS: Pre-training with Extracted Gap-Sentences", ICML 2020
[7] Raffel et al., "Exploring the Limits of Transfer Learning (T5)", JMLR 2020
[8] Maynez et al., "On Faithfulness and Factuality in Summarization", ACL 2020
[9] Ke et al., "Adaptive Model Fusion for Multi-task Summarization", ACL 2022
[10] Fabbri et al., "SummEval: Re-evaluating Summarization Metrics", TACL 2021