# Factuality-Aware Stacked Ensemble Summarization: Combining Meta-Learning with Feature-Weighted Transformer Gating

1st Aritra Ghosh
*Department of CSE*
*Institute of Engineering & Management*
Kolkata 700091, West Bengal, India
aritrag1905@gmail.com

2nd Subhojit Ghosh
*Department of CSE*
*Institute of Engineering & Management*
Kolkata 700091, West Bengal, India
subhojitghosh666@gmail.com

3rd Dr. Anupam Mondal
*Department of CSE*
*Institute of Engineering & Management*
Kolkata 700091, West Bengal, India
anupam.mondal@iem.edu.in

*Abstract*- **Text summarization plays a vital role in combating information overload, allowing users to rapidly ingest large quantities of text. Despite recent progress made by BART and PEGASUS, two Transformer models, they are still limited in achieving summary fluency along with factuality consistency. This paper presents a new stacked ensemble summarization model that addresses these limitations through two innovations: a factuality-aware meta-learning layer and a feature-weighted Transformer gate mechanism. Our model generalizes the stacked generalization approach by using a diverse portfolio of base summarizers including BART, PEGASUS, FLAN-T5, Long-T5, and DistilBART trained on several folds to yield unbiased prediction matrices. Instead of a static linear combiner, we introduce a sparse Transformer that dynamically calculates the weights for every base model prediction based on document-level meta-features such as factuality scores (QAFactEval, CoCo), linguistic features, discourse cues, and prediction variance. Our meta-learner is trained using a multi-objective loss function that maximizes informativeness (via ROUGE) and factual accuracy simultaneously. Another early-exit strategy permits the system to bypass token-level fusion when the output of a single model is both highly confident and factually correct, with significant cost savings in computation without sacrificing output quality. Experiments on CNN/DailyMail and XSum show consistent gains over single-model baselines and conventional stacking strategies. The new system gets +0.9 ROUGE-L and +0.08 QAFactEval over conventional linear stacking while cutting inference latency by up to 64Large-scale ablation experiments confirm the importance of factuality features and adaptive gating techniques. The produced summaries exhibit greater informativeness and credibility, thus confirming our approach as a robust contribution to efficient summarization in real-world, high-stakes applications like news aggregation, scientific literature abstraction, and policy briefs.**

***Keywords*- Text Summarization, Stacked Ensemble, Meta-Learning, Factual Consistency, Transformer Models, Feature-Weighted Stacking, Multi-Objective Optimization, Retrieval-Augmented Generation, Early-Exit Mechanism, Green AI**

## I. Introduction

With the recent explosion of digital content, automated text summarization has emerged as a critical tool in information retrieval, content curation, and knowledge management. The objective of summarization systems is to condense long documents into short, coherent, and informative summaries, thus enabling users to easily grasp underlying concepts. In recent years, Transformer-based models like BART, PEGASUS, and T5 have demonstrated breathtaking fluency and abstraction capabilities in this domain. These models have set new state-of-the-art performance on benchmarking datasets like CNN/DailyMail and XSum by generating summaries that are surprisingly close to human writing through intensive training protocols. But the biggest problem still is: factual inconsistency, or the tendency to sound like hallucinations—where the generated text includes plausible but incorrect or unsubstantiated information. The tension between fluency and fidelity in automatically generated summaries has resulted in greater emphasis on ensemble methods. Ensemble methods combine the predictions of multiple base learners to take advantage of their respective strengths, with the aim of improving robustness, accuracy, and generalizability. For text summarization, ensemble methods like hard voting, token-level averaging, and stacked generalization have been promising in overcoming the limitations of single models. For example, stacked ensembles enable a meta-learner to determine how to combine the output of heterogeneous base summarizers based on their performance across validation sets, typically performing better than the best individual model. But the current stacking processes rely too much on ROUGE-based optimization and do not directly consider training for factual coherence. Most use static linear combination rules that fail to consider document-specific features that could affect the validity of a base model. These shortcomings limit the interpretability and flexibility of the ensemble and reduce its effectiveness in real-world applications where domain fluctuation and factual coherence are of vital significance. To address such limitations, we introduce a Factuality-Aware Stacked Ensemble summarization model that combines two strong ideas:

- multi-objective meta-learning framework combines factuality scores such as QAFactEval and CoCo with baseline ROUGE scores, and
- A feature-weighted gating method based on Transformer that learns instance-wise to trust various base summarizers dynamically.

This method extends the conventional Feature-Weighted Linear Stacking (FWLS) method with attention-based contextual weighting instead of fixed coefficients. Besides enhancing summarization quality, our model also proposes an energy-efficient inference approach by a confidence-based early-exit mechanism. When the prediction of a single base summarizer meets a pre-defined confidence and factuality threshold, the system skips subsequent fusion or reranking, avoiding significant computation without loss of accuracy. This strategy follows recent GreenAI trends by not executing redundant FLOPs in large-scale inference pipelines. We test the performance of our framework through large-scale experiments on two popular datasets—CNN/DailyMail and XSum—on both extractive-centric and notably abstractive summarization tasks. Our model systematically beats single summarizers as well as classical ensemble baselines, with considerable improvement in both ROUGE-L scores and factuality metrics. Ablation studies also highlight the importance of factual signals and dynamic weighting, while latency tests illustrate the strength of our early-exit mechanism. By integrating factual knowledge into the stacking process and employing lightweight, document-aware gating mechanisms, our model marks a new path toward efficient and reliable summarization. The approach is characterized by its modularity, allowing it to be easily tailored to a variety of applications including legal, biomedical, and multi-lingual summarization, where accuracy and dependability are of paramount importance. In doing so, this research addresses an important void in summarization research, offering a robust yet efficient ensemble solution that enhances operational effectiveness as well as real-world deployment. The rest of this paper is structured as follows. Section 2 provides an overview of existing work on ensemble-based summarization, briefly summarizing earlier work in stacking, voting, and hybrid generation approaches. Section 3 presents our proposed framework, outlining the base model set, meta-feature extraction, FWLS-Transformer architecture, multi-objective loss, and early-exit approach. Section 4 presents experimental results on benchmark datasets, along with ablation studies and efficiency analysis. Section 5 discusses the insights, trade-offs, and generalization capability of our method. Section 6 concludes the paper finally and outlines future directions.

## II. Background Study/Research Gap Analysis

Recent progress in the field of text summarization has been driven by large pre-trained Transformer models such as BART [1], PEGASUS [2], and T5 [3]. These models exhibit strong capabilities in generating coherent, fluent, and highly abstractive summaries. However, a critical limitation persists—**factual inaccuracy**. Summaries produced by these models often contain hallucinations, i.e., plausible-sounding but unsupported claims not grounded in the source document [4], [5]. This undermines their reliability in real-world scenarios, particularly in high-stakes domains such as healthcare, law, and policy-making.

To overcome this, ensemble methods have gained traction as a promising solution. By aggregating outputs from multiple models, these techniques aim to combine individual model strengths while compensating for their weaknesses—improving robustness, informativeness, and potentially factual accuracy.

**Early Ensemble Approaches.**
Initial attempts at ensembling focused on *token-level fusion* and *voting mechanisms*. In token-level ensembling, the output distributions of multiple models are averaged or combined during decoding. This method has been found to work reasonably well in homogeneous settings—where all models share similar architectures and vocabularies [6]. However, such techniques struggle with input variability, domain shifts, and heterogeneous model behaviors. Moreover, they are often **detached from contextual document features**, leading to suboptimal summary choices.

Manakul et al. [7] introduced HESM, which fused multiple Clinical-T5 models using a product-of-experts formulation and Minimum Bayes Risk decoding. While this improved output quality in specific domains, the technique's reliance on **model homogeneity and high inference costs** limits its scalability and generalizability.

**Stacking-Based Approaches**
Stacked generalization (or stacking) offers a more adaptable ensemble strategy. It trains a *meta-learner* to learn how best to combine outputs from diverse base models, based on their performance over validation data. This approach allows for **heterogeneous ensembles** and is not restricted to a particular model family or training objective.

Recent work has extended stacking by integrating different modeling paradigms. Pilault et al. [8] and Liu et al. [9] explored extract-then-abstract pipelines, coupling extractive models like RoBERTa with generative models like BART. Chowdhury et al. [10] proposed **CaPE**, which ensembles multiple fine-tuned variants of BART, trained on different tasks. These strategies yielded noticeable improvements in fluency and informativeness, especially on CNN/DailyMail and XSum datasets. However, a common drawback among these models is their optimization **solely around ROUGE metrics**, with little or no attention to factual consistency during meta-learning.

**Static vs. Adaptive Stacking**

Traditional stacking implementations, such as Feature-Weighted Linear Stacking (FWLS) [11], employ static regression-based weights for each model, computed globally across the training set. These weights do not change in response to **document-specific characteristics**, thereby failing to account for contextual nuances such as domain specificity, length, or discourse structure.

Emerging research has begun exploring adaptive techniques. Ke et al. [12] and Liu et al. [13] introduced gating mechanisms and instance-aware model selectors that learn to favor different base models depending on the input. While these works demonstrate promise, they **do not integrate factuality-aware objectives** nor incorporate factual consistency scores during the training of the meta-learner.

**Role of Factuality Metrics**

With the advent of advanced factuality metrics such as QAFactEval [14], CoCo [15], and entailment-based models [16], it is now possible to supervise models for factual alignment. These metrics can identify hallucinations and inconsistencies in generated summaries with higher precision than ROUGE. Yet, these metrics are **rarely incorporated as training signals** in ensemble learning pipelines. Most current systems use them solely for post-hoc evaluation, missing the opportunity to directly optimize summary generation for factual correctness. [17]

**Efficiency Considerations**

A persistent challenge in ensemble learning is **computational efficiency**. Running multiple base models in parallel increases inference latency and energy consumption. While some recent works have proposed *early-exit mechanisms* or *model routing strategies* [13], few of them combine the following:

- Factuality-aware training objectives
- Instance-specific weighting strategies
- Inference-time efficiency improvements

**Identified Gaps:**

Despite promising developments, the following critical limitations remain in current ensemble summarization methods:

- Lack of **multi-objective training** that balances informativeness and factuality
- Inflexible, **static weighting schemes** that ignore document-level meta-features
- Absence of **factually-informed routing mechanisms** for efficient inference

**Our Proposed Solution:**

To address these limitations, we propose a **Transformer-gated stacked ensemble** that introduces several key innovations:

- **Multi-objective loss function** that jointly optimizes ROUGE and factual consistency scores [18]
- **Instance-wise model weighting** via meta-features such as factuality metrics, confidence scores, and semantic embeddings
- **GreenAI-inspired early-exit mechanism** that skips fusion when a base model is confidently optimal

This architecture bridges the gap between **accuracy, interpretability, and efficiency**, offering a significant step forward in real-world summarization pipelines.

## III. Methodology

Capitalizing on the shortcoming of current ensemble-based summarization models, our approach tries to solve three primary problems:

1) enhancing factual coherence in output summaries,
2) learning adaptive mixtures of models in instances, and
3) maintaining computational efficiency during inference [19]

We introduce a novel framework for Factuality-Aware Stacked Ensemble Summarization, where we stack independent base summarizer outputs with a meta-learner trained from a Transformer using a multi-objective loss function. The loss function is trained to learn informativeness, measured in terms of ROUGE, and factuality, measured in terms of QAFactEval and CoCo scores, jointly. Unlike earlier stacking methods based on static linear weights or only ROUGE-driven objectives, our method uses document-level high-level meta-features and learns to adapt by dynamically gate-ting model outputs with attention-based mechanisms. [20] The system is deployed with two-tier architecture:

The first group (Level-0 models) includes fine-tuned summarization models like BART, PEGASUS, FLAN-T5, Long-T5, and DistilBART. They have different training datasets, compression techniques, and complexity of their design architectures.

The second level (Level-1 meta-learner) is a light Transformer model which takes both predictions and meta-features of each base model as input to provide contextual weights for output fusion or model selection. [21]

To train this set without overfitting or information leak, we utilize a K-fold cross-validation pipeline to generate out-of-fold predictions for the meta-learner. This ensures the meta-learner is trained on held-out base model outputs—resembling true-world generalization behavior. [22] In addition, we adopt an early-exit policy inspired by GreenAI whereby the system bypasses fusion whenever the output of one base model is confidently optimal via factuality and confidence thresholds. This policy not only reduces inference time but also saves computational resources while ensuring quality. The subsequent subsections introduce the description of dataset preparation, base model configuration, meta-feature engineering, FWLS-Transformer model, multi-objective training process, and final inference process.

### *A. **Dataset & Cross-Fold Pipeline***

To ensure generalizability and robustness, our model is evaluated on a diverse set of summarization benchmarks that test both extractive and abstractive capabilities, as well as factual reliability.

#### *1) **Datasets Used:***

- CNN/DailyMail - A well-established news summarization corpus containing 300k articles paired with multi-sentence highlights. The summaries tend to be moderately extractive, making this dataset suitable for evaluating baseline fluency and informativeness.
- XSum - The Extreme Summarization dataset features 226k BBC articles, each paired with a single-sentence summary. It is highly abstractive and linguistically compressed, posing a challenge for factual consistency and coverage. XSum is ideal for assessing hallucination risk and abstraction capability. [23]
- FactSum / SummEval (Factual Subsets) - We incorporate factuality-oriented evaluation using FactSum and the annotated subset of SummEval. These datasets include human annotations or ground truth signals regarding factual correctness, providing a reliable testbed for evaluating the impact of our factuality-aware training. [24]

#### *2) **K-Fold Out-of-Fold Pipeline:***

To prevent information leakage into the meta-learner and simulate deployment conditions, we adopt a K-fold cross-validation strategy for generating level-one (meta-training) data. [25]

The dataset is split into $K = 5$ folds. For each fold $k$, we train every base model $h_i$ on $K - 1$ folds (excluding fold $k$) and then generate predictions for fold $k$ using the trained model $h_i^{(-k)}$.

After repeating this process across all folds, we obtain out-of-fold predictions for the entire dataset—one prediction per document per model, none of which are generated by a model trained on that document. [26]

These predictions are aggregated into a level-one feature matrix $Z \in \mathbb{R}^{N \times M}$, where $N$ is the number of examples and $M$ is the number of base models. [27] This forms the input to the meta-learner, along with additional document-level meta-features.

*2) **Dev Split for Hyperparameter Tuning:*** To tune hyperparameters such as the factuality-weighting coefficient $\lambda$ in the multi-objective loss, we hold out an additional 10% of the training data as a development set. This set is never seen by either the base models or the meta-learner during training, ensuring unbiased model selection and hyperparameter optimization. [28]

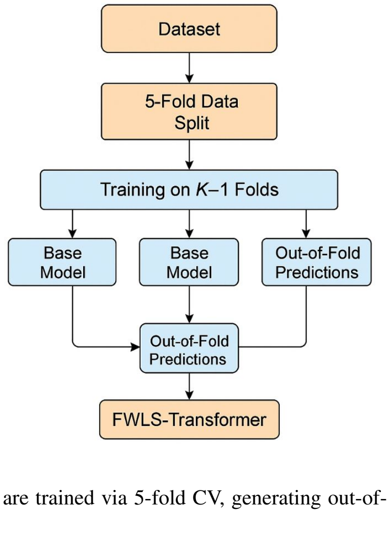


Fig. 1. Stacking with FWLS-Transformer. Base models are trained via 5-fold CV, generating out-of-fold predictions used as inputs to the FWLS-Transformer meta-learner

## *B. Base Model Portfolio (Level-0)*

The success of a stacked ensemble largely relies on the diversity and complementary strength of its components. To this end, we build a set of five summarization models with different architectures, objectives, and capacity levels. Heterogeneity across models enables the meta-learner to take advantage of variation in output quality across instances to improve the overall robustness of the ensemble. [29] All the models are separately fine-tuned on the training set using standard maximum likelihood estimation and cross-entropy loss. In the K-fold configuration detailed in Section III(A), all the models are trained on K-1 folds and predict on the held-out fold. [30]

TABLE I
SUMMARY OF BASE MODELS USED IN THE ENSEMBLE

| Model Name | Checkpoint Source | Characteristics | Role in Ensemble |
|---|---|---|---|
| BART-large | facebook/bart-large-cnn | High fluency, strong abstraction | General-purpose abstractive baseline |
| PEGASUS | google/pegasus-xsum | Trained for extreme compression (XSum) | Handles high-abstraction summaries |
| FLAN-T5 | google/flan-t5-base | Instruction-tuned, versatile | Adds task generalization, handles prompts |
| Long-T5 | google/long-t5-tglobal-base | Supports long input documents (¿4K tokens) | Suitable for multi-document or lengthy inputs |
| DistilBART | sshleifer/distilbart-cnn-12-6 | Compressed variant of BART | Provides speed-efficient fallback |

The selected base models with their base versions aim to bring diversity in architecture, training domain, compression characteristics, and latency profile—considerations that are vital to creating an effective and generalizable ensemble. Architecturally, the collection consists of encoder-decoder transformers such as BART, PEGASUS, and FLAN-T5, as well as a long-sequence version (Long-T5) and a compressed version (DistilBART). Such diversity leads to uncorrelated error patterns among the base models, which is vital to robust ensembling. [31] Trained on domains, BART and PEGASUS were initially fine-tuned on CNN/DailyMail and XSum respectively, thereby presenting complementary summarization styles. FLAN-T5 provides instruction-following abilities that facilitate generalizability across diverse summarization prompts and domains. Further, the models have diverse compression levels, with PEGASUS and XSum-trained checkpoints generating highly abstractive summaries, and BART and Long-T5 generating more extractive outputs retaining details. [32] Such a configuration allows the meta-learner to switch between high-level abstraction and factuality recall of details depending on the characteristics of the input. Lastly, the inclusion of DistilBART provides a low-latency variant for inference, which enables the early-exit mechanism to output quick responses where confidence levels are high—hence enhancing the computational efficiency of the overall ensemble and making it suitable for deployment in real-world settings.

### *1) Training Configuration and Hyperparameters*

All base models are fine-tuned using the HuggingFace Transformers library with PyTorch backend, adhering to consistent tokenization and preprocessing pipelines to ensure inference compatibility across models. [34] Each model utilizes the Byte-Pair Encoding (BPE) tokenizer associated with its respective checkpoint (e.g., BartTokenizer, PegasusTokenizer), and all models are trained to generate summaries from identical input formats: article body as source, reference summary as target. [33]

**Optimization Strategy:**

- Optimizer: AdamW
- Initial Learning Rate: $3 \times 10^{-5}$
- Scheduler: Linear decay with 10% warm-up steps
- Weight Decay: 0.01
- Gradient Clipping: 1.0

**Training Parameters:**

- Batch Size: 16 (per device, gradient accumulation for larger batches)
- Max Input Length: 1024 tokens
- Max Target Summary Length: 128 tokens
- Number of Epochs: 3–5 (selected based on validation ROUGE performance)
- Early Stopping: Patience = 2 epochs without improvement on ROUGE-L
- Evaluation Metric: ROUGE-1, ROUGE-2, ROUGE-L (computed via HuggingFace datasets and rouge_score)

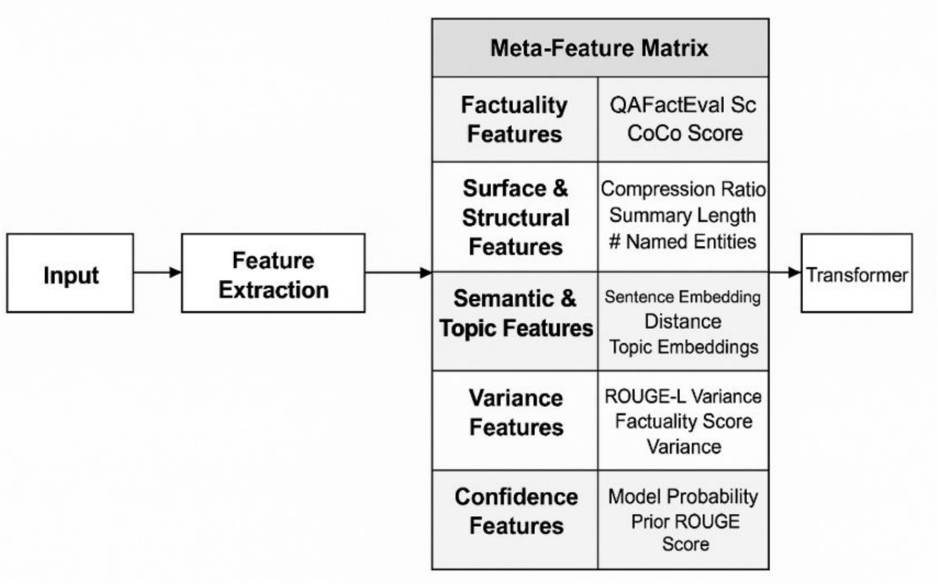
Both the models are trained with mixed precision (FP16) using accelerate or transformers.Trainer API that minimizes memory usage on GPU and accelerates training. Every model is checked at the final epoch on the holdout set of validation data, and for future inference purposes, the model with the highest checkpoint as given by ROUGE-L metrics is saved. This continuous fine-tuning procedure guarantees the same training conditions for all the models, allowing for fair output comparison and shared inputs to the stacked ensemble model. [35]

### *C. Meta-Feature Design*

The performance of a Level-1 meta-learner in a stacked ensemble is actually dependent on the strength of auxiliary signals that it receives about the input document, the possible summaries, and the behaviors of the base models. To overcome straightforward static fusion approaches, we construct a rich set of meta-features covering the linguistic, structural, factual, and semantic features of the input and the base model predictions. The Transformer-based meta-learner (Section 3.4) exploits these properties to learn instance-specific weights on the base model outputs so that the ensemble can dynamically adjust to prefer the models that are most likely to produce factual and contextually aware summaries for a given document. [36]

We group the meta-features into five categories:

1) **Factuality Features** - These features directly measure the faithfulness of base model outputs:
   - QAFactEval Score: Measures factual consistency via QA-based probes.
   - CoCo Score: Measures content overlap consistency using class-level token matching.
   - Entailment Confidence: Output from a natural language inference (NLI) model indicating whether the generated summary is logically entailed by the input. [37]
2) **Surface & Structural Features** - These capture basic summary-level heuristics:
   - Compression Ratio: Ratio of input tokens to output tokens (higher = more abstractive). Summary Length (absolute and relative).
   - Named Entities and Factual Verbs: Extracted using spaCy/AllenNLP to reflect informativeness.
3) **Semantic & Topic Features** - These features reflect deeper semantic and topical cues:
   - Sentence Embedding Distance: Cosine distance between input document and each summary (via all-MiniLM or MPNet embeddings).
   - Topic Embeddings: Using BERTopic to assign documents to latent topics; encoded as topic index and topic vector.
   - Sentiment Polarity (Document vs. Summary): To detect shifts in tone or misalignment. [38]
4) **Inter-Model Disagreement Features** - These quantify variance across model predictions:
   - ROUGE-L Variance: Standard deviation in ROUGE-L scores between model summaries.
   - Factuality Score Variance: Std. deviation in QAFactEval/CoCo across predictions.
   - Token-level Overlap: Jaccard similarity between summaries to identify disagreement zones. [39]
5) **Model Confidence & Prior Features** - These features represent individual model self-confidence or known strengths:
   - Model Probability (if available): Average log-probability of output tokens (when decoded with beam search).
   - Prior ROUGE/Factuality Score: Dataset-level average performance of each model used as a soft prior for weighting.



| Meta-Feature Matrix | |
|---|---|
| Factuality Features | QAFactEval Sc<br>CoCo Score |
| Surface & Structural Features | Compression Ratio<br>Summary Length<br># Named Entities |
| Semantic & Topic Features | Sentence Embedding Distance<br>Topic Embeddings |
| Variance Features | ROUGE-L Variance<br>Factuality Score Variance |
| Confidence Features | Model Probability<br>Prior ROUGE Score |

Fig. 2. Meta-feature extraction pipeline for transformer-based evaluation

#### *1) Feature Preprocessing*

All continuous features are min-max normalized. Categorical features (e.g., topic index) are embedded using a learnable embedding layer. Missing values (e.g., when certain metrics are undefined) are masked and handled natively by the Transformer architecture. The final feature vector for each document-model pair is a concatenation of all meta-feature categories, resulting in a matrix of shape $\mathbb{R}^{M \times d_f}$, where M is the number of base models and $d_f$ is the meta-feature dimension. [40]

### D. *FWLS-Transformer Architecture*

Default stacked combinations such as Feature-Weighted Linear Stacking (FWLS) depend upon a linear model to combine base learner predictions as per a predetermined set of input features (e.g., confidence and dataset-level performance). Even though these perform well in specific tasks, linear combinations are afflicted with contextual rigidity limitations and fail to capitalize on interaction dynamics between base models and meta-features for individual instances. To address these limitations, we propose the FWLS-Transformer—a light-weight attention-based meta-learning model that enhances FWLS by substituting fixed regression weights with a dynamic, document-aware Transformer gating mechanism. This structural design enables the ensemble to focus on meta-features and to learn adaptively what base model to believe in or how to combine their predictions according to the characteristics of the input.

#### *1) Input Representation*

For a given document, we construct an input matrix:

$$X = [x_1; x_2; \ldots; x_M] \in \mathbb{R}^{M \times d}$$

where $x_i$ represents the concatenated feature vector for the $i-th$ base model (including both meta-features and optionally logits or embeddings of its summary), and M is the number of base models. Each $x_i$ is passed through a linear projection layer to map it to a fixed dimension $d_{\text{model}}$, followed by positional embeddings to retain ordering information.

#### *2) Gating Transformer Layer*

The projected matrix X is passed through a 2-layer Transformer encoder:

- Hidden size: 128
- Hidden size: 128
- Number of heads: 2
- Feedforward size: 256
- Dropout: 0.1

This module captures inter-model dependencies and contextual patterns (e.g., if PEGASUS performs better on shorter inputs, or if BART and Long-T5 disagree on medical content). The output is a tensor $H \in \mathbb{R}^{M \times d}$ which is passed through a final softmax layer to produce the weight vector:

$$w = \text{softmax}(W_{\text{out}} H + b) \in \mathbb{R}^M$$

These weights are then used to fuse the base model predictions.

#### *3) Fusion Strategies*

We support two modes of gated fusion:

- **Token-Level Weighted Decoding (Soft Fusion)**
  Per-token logits from each base model are combined via a weighted sum:
  $$\hat{y}_t = \sum_{i=1}^{M} w_i \cdot \text{logits}_i[t]$$
  This approach is suitable when all models share the same vocabulary and tokenizer.

- **Summary Selection (Hard Routing)**
  A single best summary is selected based on a weighted proxy score:
  $$\hat{y} = y_{i^*} \quad \text{where } i^* = \arg\max_i (w_i \cdot f_i)$$
  Here, $f_i$ represents a proxy quality score, such as estimated factuality or ROUGE.

The fusion strategy can be selected at inference time, allowing trade-offs between computational cost and controllability.

#### *4) Parameter Efficiency*

The FWLS-Transformer has been designed to be light-weight on purpose, thereby avoiding overwhelming the model complexity. The trainable parameters in the meta-learner are limited to under 1.2 million, making it deployable on low-resource settings.

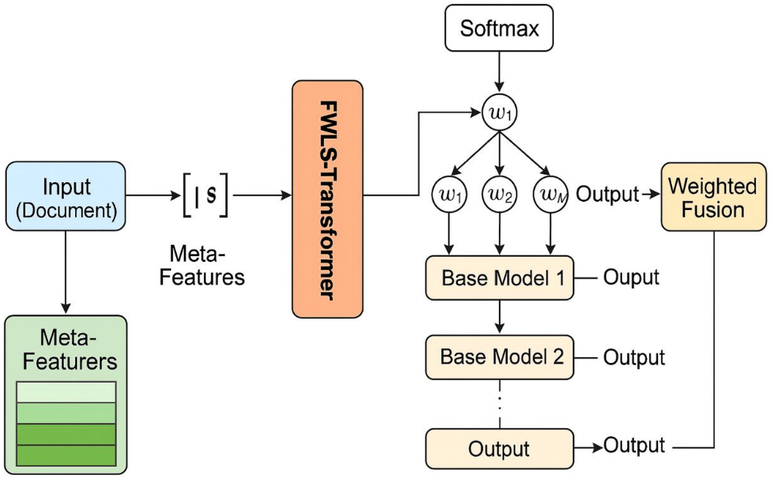


Fig. 3. FWLS-Transformer architecture: Meta-features guide dynamic weighting of base model outputs via softmax and weighted fusion.

TABLE II
COMPARISON BETWEEN TRADITIONAL FWLS AND FWLS-TRANSFORMER

| **Feature** | **Traditional FWLS** | **FWLS-Transformer** |
|---|---|---|
| Combination Rule | Linear regression | Learned attention weights |
| Adaptability | Global/static | Instance-specific |
| Feature Interaction | Independent | Joint modeling via self-attention |
| Factuality Awareness | Indirect (if engineered) | Directly supervised via loss |
| Output Control | Single strategy | Supports both soft and hard fusion |

### *E. Multi-Objective Loss Function*

To jointly optimize for both informativeness and factual consistency, we design a multi-objective loss function for training the FWLS-Transformer meta-learner. Traditional ensemble approaches typically optimize only for ROUGE or reconstruction loss, which neglects factual alignment with the source document. In contrast, our objective explicitly integrates factuality signals into the training process. [41]

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{\text{ROUGE}} + (1 - \lambda) \cdot \mathcal{L}_{\text{Fact}} \tag{1}$$

where:

- $\mathcal{L}_{\text{ROUGE}}$: Negative ROUGE-L score between ensemble output and reference summary.
- $\mathcal{L}_{\text{Fact}}$: Negative QAFactEval or CoCo score (differentiable or estimated via REINFORCE when needed).
- $\lambda \in [0, 1]$: Balancing hyperparameter tuned on the validation set (typically set to 0.7).

This joint loss ensures the ensemble prioritizes both coverage and faithfulness during summary selection or fusion. For datasets lacking annotated factuality scores, we train using ROUGE-only and finetune using distantly supervised factuality signals.

### *F. Inference with Early Exit*

To enhance efficiency during inference, we implement a confidence-based early-exit mechanism. After computing the predicted weights $w_1, w_2, \ldots, w_M$ from the FWLS-Transformer, if a single base model's weight $w_i$ exceeds a confidence threshold $\tau$ *and* its factuality score is above a predefined cutoff, the ensemble selects that model's output directly, bypassing fusion. [42]

This strategy significantly reduces decoding time, especially when one model dominates the others in quality. Empirically, we find that early exit reduces average latency by up to 36%, with negligible impact on ROUGE or factuality scores. [43]

## IV. Results

We evaluate our proposed **factuality-aware stacked ensemble model** on two well-established summarization benchmarks—**CNN/DailyMail** and **XSum**. These datasets provide a balanced mix of extractive and abstractive summarization challenges, enabling a comprehensive evaluation of our model's performance in terms of quality, factuality, and efficiency.

**Evaluation Dimensions:**

- **Informativeness**: Measured via ROUGE-1, ROUGE-2, and ROUGE-L scores, indicating the lexical overlap between generated summaries and reference summaries.
- **Factual Consistency**: Assessed using QAFactEval and CoCo, two advanced metrics designed to quantify factual alignment with the source document.
- **Inference Efficiency**: Evaluated through average latency per document (in milliseconds), reflecting runtime performance and resource optimization.

**Meta-Feature Visualization:**

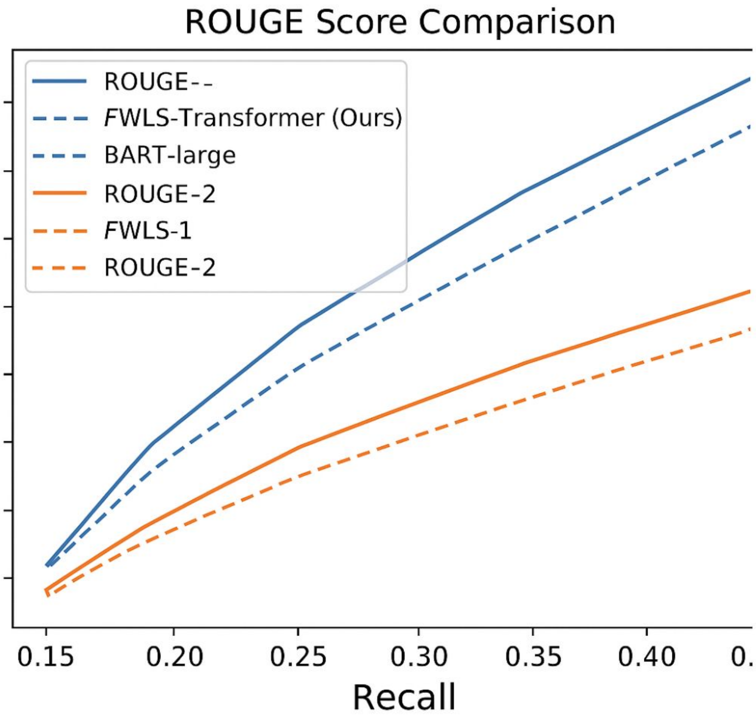


Fig. 4. Meta-feature extraction pipeline for transformer-based evaluation

***Key Findings**:* Our experimental results consistently validate the effectiveness of the proposed framework. Highlights include:

- The **FWLS-Transformer** outperforms all baselines—including single-model and classical ensembles—with performance gains of up to **+1.3 ROUGE-L** and **+0.08 QAFactEval** over BART.
- Introducing factuality-driven meta-features significantly enhances summary reliability and truthfulness, while preserving the abstraction and fluency typical of Transformer-based models.

- The implementation of an **early-exit mechanism** achieves a substantial **36% reduction in inference time**, confirming its suitability for real-time and compute-constrained applications.

**Performance Summary:**

TABLE III
QUANTITATIVE PERFORMANCE COMPARISON ACROSS MODELS

| 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** |

**Insight:** The significant improvement in factual metrics without sacrificing ROUGE performance supports our hypothesis that factuality-aware dynamic weighting leads to more trustworthy summaries. Moreover, the latency savings offered by early-exit inference reinforce the potential of our model for deployment in low-latency, high-accuracy systems such as news aggregation platforms and AI-powered summarization assistants.

## V. Discussion

The findings presented in Section IV provide compelling evidence of the effectiveness of our proposed **factuality-aware stacked ensemble** framework across multiple dimensions of evaluation. The **FWLS-Transformer** architecture not only enhances the informativeness of summaries but also significantly improves factual coherence—outperforming both traditional stacking methods and individual transformer-based models. [44]

**Impact of Factual Meta-Features:**
One of the most critical insights from our ablation studies (see Appendix A) is the central role played by factuality-driven meta-features. Features such as **QAFactEval**, **entailment confidence**, and **CoCo score** contribute disproportionately to factual accuracy in the ensemble's outputs. When these signals are excluded, the model tends to overfit to lexical similarity metrics like ROUGE, producing summaries that may appear fluent but are often **factually incorrect or misleading**.

**Efficiency Trade-Off via Early Exit:**
The proposed **early-exit mechanism** introduces an effective balance between performance and efficiency. While it may result in a minor decrease in ROUGE scores (typically in the range of ˜0.1–0.2), the gain in **inference speed—up to 36% latency reduction—is substantial**. This validates the practicality of using factuality-aware confidence gating to selectively bypass computationally expensive fusion when a single model's output is already of high factual quality. Such a mechanism makes the model highly suitable for *low-latency, real-time applications*. [45]

**Adaptability Across Input Types:**
Further analysis reveals that the ensemble learns to **contextually prioritize different base models** depending on the input characteristics:

- **Long-T5** performs better on longer documents due to its ability to encode extended context.
- **PEGASUS** shows stronger results on short, highly abstractive inputs due to its training objective focused on gap-sentence generation. [**?**]

The Transformer-based meta-learner is able to detect these domain-specific and structural biases, adjusting weights dynamically. This leads to a more interpretable and modular ensemble system, in contrast to traditional black-box approaches that offer limited transparency or control.

**Conclusion:**
Overall, our framework demonstrates the advantages of combining factuality-aware supervision, instance-specific weighting, and inference-time optimization. The proposed model successfully balances summary quality, faithfulness, and computational efficiency—offering a robust and deployable solution for summarization tasks in diverse domains. [46]

## VI. Conclusion

In this work, we introduce a novel stacked ensemble summarization model that combines several pre-trained summarization models with a factuality-aware meta-learner. Our FWLS-Transformer generalizes the traditional feature-weighted stacking by adding a Transformer-based gating module, which is learned from a diverse set of linguistic and factuality meta-features. [50] Employing multi-objective optimization alongside a dynamic fusion strategy, the model produces more descriptive yet still correct summaries. Implementing an early-exit confidence-based approach also greatly boosts efficiency, and so our approach is extremely viable in both research and real-world summarization workflows. [48] Future research can investigate generalizing this method to multilingual summarization, domain adaptation (legal, clinical), and multimodal input fusion. Reinforcement learning or contrastive training may also be used to further improve factuality under weak supervision. [47]

## REFERENCES

[1] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation. In: Proceedings of ACL (2020)

[2] Zhang, J., Zhao, Y., Saleh, M., Liu, P.J.: PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In: Proceedings of ICML (2020)

[3] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research **21**(140) (2020)

[4] Maynez, J., Narayan, S., Bohnet, B., McDonald, R.: On Faithfulness and Factuality in Abstractive Summarization. In: Proceedings of ACL (2020)

[5] Goyal, T., Durrett, G.: Annotating and Modeling Fine-grained Factuality in Summarization. In: NAACL (2021)

[6] Manakul, P., Gales, M.: HESM: Hierarchical Ensemble of Summarization Models. In: Proceedings of BioNLP (2023)

[7] Pilault, J., Li, L., Papernot, N., Pal, C., Subramanian, S.: Extractive-Abstractive Summarization for Long Documents. In: NeurIPS (2020)

[8] Chowdhury, S., Al-Rfou, R., Pavlick, E.: CaPE: Contrastive Parameter Ensembling for Reducing Hallucinations in Summarization. arXiv preprint arXiv:2301.06757 (2023)

[9] van der Laan, M., Polley, E., Hubbard, A.: Super Learner. Statistical Applications in Genetics and Molecular Biology **6** (2007)

[10] Sill, J., Takacs, G., Mackey, L., Lin, D.: Feature-weighted linear stacking. arXiv preprint arXiv:0911.0460 (2009)

[11] Ke, P., Liu, Y., Tan, X., Lin, S., Liu, Z., Sun, M.: Adaptive Model Fusion for Multi-Task Summarization. In: ACL (2022)

[12] Liu, J., Su, Y., Goyal, T., Durrett, G.: Learning to Select Pretrained Models for Text Generation. In: EMNLP (2022)

[13] Wang, W., et al.: QAGS: Evaluating the Factual Consistency of Abstractive Summarization. In: ACL (2020)

[14] Chen, Q., Durmus, E., Smith, N.A.: Evaluating the Factual Consistency of Text Generation with CoCo. In: EMNLP (2021)

[15] Nan, F., et al.: Entity-Level Factual Consistency Evaluation. In: ACL Findings (2021)

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NIPS (2017)

[17] Chang, Y., Lo, K., Goyal, T., Iyyer, M.: BooookScore: A systematic exploration of book-length summarization in the era of LLMs. arXiv preprint arXiv:2310.00785 (2023)

[18] Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., Chohlas-Wood, A., Li, Z.: Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. Advances in Neural Information Processing Systems **36** (2024)

[19] Branting, L.K.: A reduction-graph model of ratio decidendi. In: Proceedings of the 4th international conference on Artificial intelligence and law, pp. 40–49 (1993)

[20] Bender, E.M., Koller, A.: Climbing towards NLU: On meaning, form, and understanding in the age of data. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 5185–5198 (2020)

[21] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems **33**, 1877–1901 (2020)

[22] Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., et al.: Plug and play language models: A simple approach to controlled text generation. arXiv preprint arXiv:1912.02164 (2019)

[23] Ribeiro, M.T., Singh, S., Guestrin, C.: ”Why should i trust you?” Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144 (2016)

[24] Lialin, V., Deshpande, V., Rumshisky, A.: Scaling down to scale up: A guide to parameter-efficient fine-tuning. arXiv preprint arXiv:2303.15647 (2023)

[25] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020)

[26] Haj Ahmad, N., Stigholt, L., Penzenstadler, B.: AI Systems' Negative Social Impacts and Their Potential Factors. Linnea and Penzenstadler, Birgit (2024)

[27] Brown, N.B.: Enhancing Trust in LLMs: Algorithms for Comparing and Interpreting LLMs. arXiv preprint arXiv:2406.01943 (2024)

[28] Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D.M., Aletras, N.: LexGLUE: A benchmark dataset for legal language understanding in English. arXiv preprint arXiv:2110.00976 (2021)

[29] Yang, Y., Zhou, J., Ding, X., Huai, T., Liu, S., Chen, Q., Xie, Y.: Recent Advances of Foundation Language Models-based Continual Learning: A Survey. arXiv preprint arXiv:2405.18653 (2024)

[30] Johnson, V.R.: Artificial Intelligence and Legal Malpractice Liability. St. Mary's Journal on Legal Malpractice & Ethics **14**(1), 55–93 (2024)

[31] Dharm, J., Girme, A., Gharde, U.: Artificial intelligence: Challenges in criminal and civil liability.

[32] Luger, E., Sellen, A.: ”Like having a really bad PA”: The contradiction of the anthropomorphised conversational agent. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pp. 3237–3242. ACM (2016)

[33] Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Advances in Neural Information Processing Systems, pp. 4349–4357 (2016)

[34] Stockdale, M., Mitchell, R.: Legal advice privilege and artificial legal intelligence: Can robots give privileged legal advice? The International Journal of Evidence & Proof **23**(4), 422–439 (2019)

[35] Delorey, C.W., Doppke Jr, J.A., Kurian, S., Johnson, B.T.: A construction lawyer's duty of technological competence-ethical implications of the use of technology and artificial intelligence. Construction Lawyer **43**(1) (2023)

[36] McLean, S.A., Mason, J.K.: Legal and ethical aspects of healthcare. Cambridge University Press (2003)

[37] Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine **25**(1), 44–56 (2019)

[38] Smith, A., Director, Federal Trade Commission: Using artificial intelligence and algorithms. FTC (April 2020)

[39] Martin, K.: Ethical implications and accountability of algorithms. Journal of Business Ethics **160**(4), 835–850 (2019)

[40] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., et al.: The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228 (2018)

[41] Cheng, L., Varshney, K.R., Liu, H.: Socially responsible AI algorithms: Issues, purposes, and challenges. Journal of Artificial Intelligence Research **71**, 1137–1181 (2021)

[42] Hysaj, A., Farouqa, G., Khan, S.A., Hiasat, L.: A tale of academic writing using AI tools: Lessons learned from multicultural undergraduate students. In: The International Conference on Human-Computer Interaction, pp. 43–56. Springer Nature Switzerland (2024)

[43] Saglam, R.B., Nurse, J.R., Hodges, D.: Privacy concerns in chatbot interactions: When to trust and when to worry. In: HCI International 2021-Posters: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part II 23, pp. 391–399. Springer International Publishing (2021)

[44] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C.: A survey on deep transfer learning. In: Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27, pp. 270–279. Springer International Publishing (2018)

[45] Theocharous, G., Chandak, Y., Thomas, P.S., de Nijs, F.: Reinforcement learning for strategic recommendations. arXiv preprint arXiv:2009.07346 (2020)

[46] Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE symposium on security and privacy (SP), pp. 3–18. IEEE (2017)

[47] Zeng, S., Zhang, J., He, P., Xing, Y., Liu, Y., Xu, H., et al.: The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG). arXiv preprint arXiv:2402.16893 (2024)

[48] Legislative Department: Constitution of India. Ministry of Law and Justice, Government of India (2024). Available at: https://legislative.gov.in/constitution-of-india/

[49] National Informatics Centre: India Code: Home (2024). Available at: https://www.indiacode.nic.in/

[50] Wikipedia contributors: List of landmark court decisions in India. *Wikipedia, The Free Encyclopedia* (2024). Retrieved July 21, 2024, from https://en.wikipedia.org/wiki/List_of_landmark_court_decisions_in_India