# Arg-LLaDA: Argument Summarization via Large Language Diffusion Models and Sufficiency-Aware Refinement

**Anonymous ACL submission**

## Abstract

Argument summarization aims to generate concise, structured representations of complex, multi-perspective debates. While recent work has advanced the identification and clustering of argumentative components, the generation stage remains underexplored. Existing approaches typically rely on single-pass generation, offering limited support for factual correction or structural refinement. To address this gap, we introduce Arg-LLaDA, a novel large language diffusion framework that iteratively improves summaries via sufficiency-guided remasking and regeneration. Our method combines a flexible masking controller with a sufficiency-checking module to identify and revise unsupported, redundant, or incomplete spans—yielding more faithful, concise, and coherent outputs. Empirical results on two benchmark datasets demonstrate that Arg-LLaDA surpasses state-of-the-art baselines in 7 out of 10 automatic evaluation metrics. In addition, human evaluations reveal substantial improvements across core dimensions, coverage, faithfulness, and conciseness,validating the effectiveness of our iterative, sufficiency-aware generation strategy.

## 1 Introduction

Argument summarization aiming to distill salient points from multi-perspective debates (known as *Key Point Analysis (KPA)* (Bar-Haim et al., 2020a,b)) or argumentative texts into concise, faithful, and well-structured summaries (Li et al., 2024a) (Shown in Table 1). This task is particularly important for applications in civic discourse analysis (Vecchi et al., 2021), legal decision support (Habernal et al., 2024), and educational technology (Grljevic et al., 2022), where it is essential to capture the reasoning structures and factual bases behind opposing claims. Current pipeline can be decomposed into two critical stages: *(1)* identifying and collecting the core argumentative components—namely, claims and their supporting or opposing evidence and *(2)* generating high-quality summaries based on this structured input (van der Meer et al., 2024).

Recent advances in argument summarization have concentrated on what we refer to as Stage 1, which either involves: *(i)* constructing argumentative components from a fixed, pre-collected corpus; or *(ii)* retrieving such components dynamically from the open Web. Some studies target the former, automatically identifying argumentative elements—such as claims, premises, and stance—via supervised learning methods (Kapadnis et al., 2021) or prompt-based large language models (LLMs) (Tang et al., 2024a,b; Fukuma et al., 2024). Other work further groups semantically and functionally similar arguments into representative key points (Phan et al., 2021; Reimer et al., 2021), employing techniques such as contrastive learning (Alshomary et al., 2021), graph-based (Li et al., 2024b), and density-based clustering (Li et al., 2023; Khosravani et al., 2024). Meanwhile, another strand of work emphasizes dynamic retrieval of contextual evidence from external sources such as online documents or knowledge bases: models query these sources at inference time to retrieve supporting or elaborating context for argumentative components (Shan and Lu, 2025). This retrieval-augmented generation (RAG) paradigm enables LLMs to ground their outputs in external content, enhancing factuality and currency without continually training the model.

However, Stage 2 remains inadequately investigated. Firstly, most existing approaches rely on end-to-end fine-tuned models (Eden et al., 2023; Bar-Haim et al., 2021) or prompt-based LLMs that generate summaries (Tang et al., 2024b; Altemeyer et al., 2025) in an auto-regressive manner, providing no inherent mechanism for correction. Once the summary is generated, post-hoc editing becomes cumbersome and unsystematic. This greedy de-

| **Input** |
| --- |
| **(Topic)** Routine child vaccinations should be mandatory<br>**(Stance)** Oppose<br>**(Claim 1)** Vaccines or their side effects may be dangerous<br>**(Evidence 1)** Rotashield was withdrawn after being linked to bowel obstruction<br>**(Evidence 2)** CDC reports rare risks like pneumonia (chickenpox vaccine) and Guillain-Barré Syndrome (flu vaccine)<br>**(Claim 2)** Mandatory vaccination violates basic rights<br>**(Evidence 1)** The First Amendment protects religious freedom<br>**(Evidence 2)** Compulsory vaccination interferes with bodily integrity, violating international human rights conventions |
| **Summary** |
| Mandatory child vaccinations raise health and ethical concerns. Historical cases like Rotashield and CDC findings highlight medical risks. Moreover, opponents argue that such mandates may infringe on personal freedoms and human rights, suggesting vaccination should remain a personal choice guided by awareness rather than coercion. |

Table 1: Example mapping from structured arguments to generated summary. The model effectively integrates two claims and four supporting evidences into a coherent output.

coding strategy often favors highest-probability tokens at each step, resulting in factual inaccuracies or hallucinations (Andriopoulos and Pouwelse, 2023). Secondly, unlike text summarization, argument summarization demands conciseness, covers different aspects or facets without redundancy, with a clear argumentative structure (Li et al., 2025c). Existing models rarely guarantee the preservation of argument roles, such as claim-premise relations, or tightly control thematic redundancy (Li et al., 2024a; Liu et al., 2023; Cheng et al., 2022). As a result, these methods often compromise narrative coherence and structural fidelity, rendering them ill-suited for producing argument summaries that meet the required standards.

To address these challenges, we introduce **Arg-LLaDA**, a novel large language diffusion framework specifically tailored for argument summarization. Rather than generating summaries in a one-shot, greedy fashion, **Arg-LLaDA** employs an iterative remasking and regeneration strategy designed to incrementally refine outputs through multiple editing cycles. Central to our method is a novel enhanced span-infilling paradigm (Nie et al., 2025) with a flexible mask-remask controller, which allows for multiple masking strategies—such as sentence-level masking, thereby enabling precise and context-aware rewriting of deficient summary segments while preserving global coherence. Additionally, we incorporate a sufficiency-checking module, which detects unsupported, redundant, or logically incomplete spans based on the principle of conclusion reconstructability. We validate the effectiveness of our design on the ASE and ArgKP benchmarks. Arg-LLaDA outperforms state-of-the-art LLM baselines such as Gemini 2.5 and DeepSeek-R1, achieving relative improvements with 7 of 10 metrics best. Human evaluation further validates these gains, with Arg-LLaDA improving coverage, faithfulness, and conciseness by up to 5.0% compared to... Ablation studies demonstrate that the iterative refinement and sufficiency diagnosis modules contribute orthogonally to model improvements, showing their independent effectiveness.

In summary, our key contributions are as follows: *(1)* **Arg-LLaDA**, a novel self-refinement model that improves argument summaries via sufficiency-guided remasking and iterative generation. *(2)* We empirically validate our approach on benchmark datasets, demonstrating significant improvements in content fidelity, conciseness, and argumentative

structure over prior baselines.

## 2 Related work

**Argument Summarization.** Traditional approaches often relied on extractive methods, such as clustering-based selection (Misra et al., 2017; Reimers et al., 2019; Ajjour et al., 2019), which group similar arguments without producing fluent summaries. Later work reformulated the task as one of claim generation (Wang and Ling, 2016) or aspect-controlled generation (Schiller et al., 2021), enabling abstractive summaries that better capture the underlying stance. A prominent research line is *Key Point Analysis (KPA)* (Bar-Haim et al., 2020a,b), which identifies representative key points across a collection of arguments. Follow-up work enhances this pipeline through unsupervised clustering (Alshomary et al., 2021), generation-based key point synthesis (Li et al., 2023), and extractive key point selection with structure-aware scoring (Khosravani et al., 2024). Recent studies explore end-to-end generation models that condition on structured argumentative inputs (Chen et al., 2022) or integrate content selection via reinforcement learning (Roit et al., 2023).

**Diffusion Language Models.** Recent work has increasingly focused on diffusion-based modeling, encompassing both continuous and discrete approaches (Li et al., 2025b). Notably, masked diffusion models, a type of discrete diffusion technique, have emerged as the leading variant in terms of performance. Theoretical studies have provided rigorous foundations for these models and verified their competitiveness with autoregressive architectures at the GPT-2 scale (Shi et al., 2024; Ou et al., 2025; Shao et al., 2024; Li et al., 2025a). Building on this progress, LLaDA (Nie et al., 2025) scaled masked diffusion modeling to 8 billion parameters—becoming the first diffusion-based LLM to rival models like LLaMA 3 (Dubey et al., 2024) in a variety of downstream tasks. Its multimodal variant, LLaDA-V (You et al., 2025), outperforms the LLMs baseline across several multimodal benchmarks—indicating potential inherent advantages of the masked-diffusion framework in broader, multimodal contexts.

## 3 Methodology

In this section, we introduce **Arg-LLaDA**, a diffusion-based framework for argument summarization. We first formalize the task (Section 3.1), then describe an iterative masked denoising process with sufficiency-guided masking (Section 3.2). To support targeted refinement, we develop a sufficiency diagnosis module using both prompting and supervised classification (Section 3.4). Finally, we outline training objectives and implementation details (Section 3.5).

### 3.1 Problem Definition

We study the task of *argument summarization*, where the input is a collection of $N$ claim–evidence pairs:

$$\{(c_i, E_i)\}_{i=1}^{N} \tag{1}$$

with each claim $c_i$ supported by its corresponding set of evidence $E_i$. Our goal is to generate a summary $S$ satisfying the following criteria: *(i)* **Comprehensiveness**: The summary should include all relevant claims from the input. *(ii)* **Conciseness**: The output must avoid redundancy and unnecessary verbosity. *(iii)* **Faithfulness**: All summarized claims and conclusions must be grounded in the provided evidence. We formalize the problem as finding a summary

$$S^* = \arg\max_{S} \text{ Quality}(S \mid \{(c_i, E_i)\}) \tag{2}$$

under the above desiderata.

### 3.2 Diffusion Generation Framework

Inspired by recent advances in masked diffusion language models (Nie et al., 2025), we adopt a non-autoregressive, iterative generation paradigm for argument summarization. Unlike autoregressive decoders, which generate outputs token-by-token in a fixed order and struggle with revising earlier errors, our diffusion-based approach refines summaries through repeated masked denoising steps. This iterative design allows for flexible editing, span-level abstraction, and improved controllability—characteristics well-aligned with the requirements of factual and concise argument summarization.

Given an input set $\mathcal{X} = \{(c_i, E_i)\}_{i=1}^{N}$ consisting of claims and their supporting evidence, the goal is to generate a coherent and faithful summary $\hat{S}$. To achieve this, we initialize with a fully masked summary $\tilde{S}_T$ and iteratively denoise it using a learned reverse process. At each timestep $t$, we sample a cleaner intermediate summary $\tilde{S}_{t-1}$ as follows:

$$\tilde{S}_{t-1} \sim q_\theta(\tilde{S}_{t-1} \mid \tilde{S}_t, \mathcal{X}), \quad t = T, ..., 1 \tag{3}$$

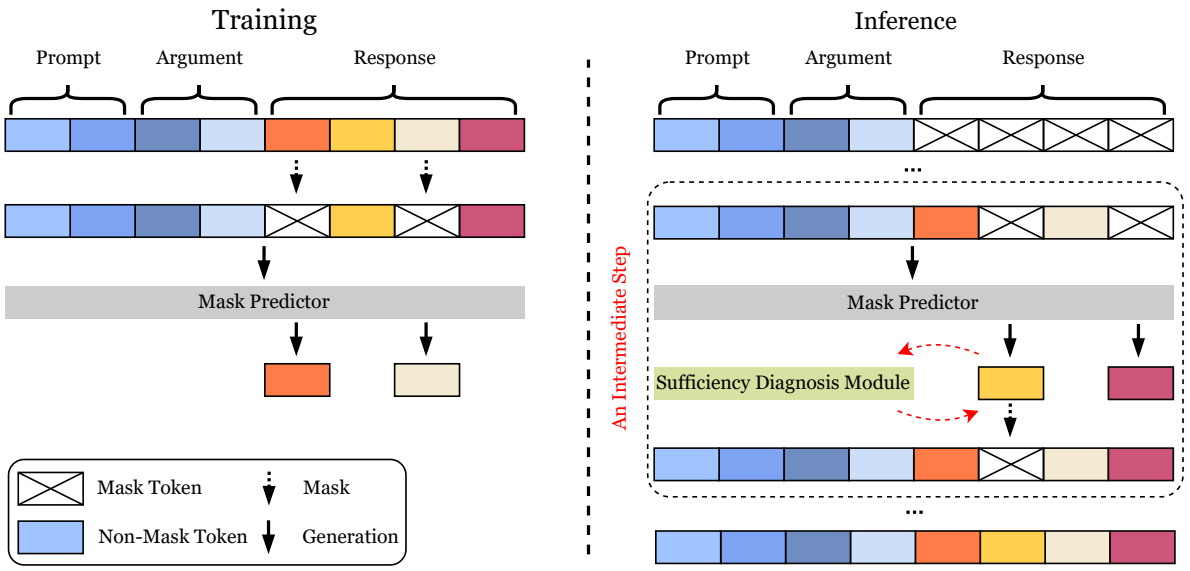


Figure 1: **Training and inference process of Arg-LLaDA.** The model iteratively generate the output through a masked denoising diffusion process. At each timestep, the sufficiency diagnosis module assigns token-level sufficiency scores based on the claim-evidence context, guiding a selective masking controller to focus regeneration on unsupported or redundant spans. This allows Arg-LLaDA to perform semantically grounded and targeted refinement toward factually faithful and concise argument summaries.

This process continues until convergence, yielding the final summary $\hat{S} = \tilde{S}_0$.

**Training Objective.** Following Nie et al. (2025), the model is trained via a denoising score matching objective. Specifically, given a reference summary $\hat{S}$, we randomly corrupt a subset of tokens $\mathcal{M} \subset \hat{S}$ and optimize the model to reconstruct the original tokens conditioned on the unmasked context and input arguments:

$$\mathcal{L}_{\text{diff}} = -\mathbb{E}_{\mathcal{M}} \sum_{t \in \mathcal{M}} \log p_\theta(\hat{S}_t \mid \hat{S}_{\setminus \mathcal{M}}, \mathcal{X}) \quad (4)$$

This objective encourages the model to generate contextually appropriate and semantically grounded content. Importantly, masking during training remains random to ensure broad denoising capability, while targeted refinement is deferred to inference.

### 3.3 Inference-Time Sufficiency Guided Masking

In line with recent advances in diffusion-based text generation, Arg-LLaDA adopts an inference-time selective masking strategy to refine summaries across iterative denoising steps. Instead of relying solely on random corruption, our framework integrates two complementary signals—token-level confidence and span-level sufficiency. This combined design allows the model to address both locally uncertain expressions and confidently generated but unsupported statements, thereby achieving more faithful and concise summaries.

**Token-level low-confidence remasking.** At each inference step, the model produces probability distributions over candidate tokens. We compute token confidence as the maximum probability assigned to any candidate token. Tokens with confidence below a threshold $\tau$ are deemed unreliable, and the lowest-$r$ proportion of such tokens are selected for resampling. This mechanism ensures that local corrections are directed toward genuinely uncertain regions while avoiding unnecessary revisions of fluent and stable tokens. Following prior work on diffusion-based generation (Nie et al., 2025), we apply confidence-based remasking only at inference, not training. This separation ensures that the diffusion backbone learns a broad, task-agnostic denoising capability, while inference-time confidence signals specialize revision to context-specific uncertainties.

**Sufficiency-guided span masking.** Low-confidence detection alone cannot address confidently incorrect content or omissions. To mitigate this, we integrate an external sufficiency module $\mathcal{C}_\phi$, which evaluates whether candidate summary spans

are adequately supported by the input claims and evidence. Spans are segmented primarily at the sentence level, aligned with punctuation and discourse boundaries. For compound sentences, we further split into clause-level units to obtain semantically coherent argumentative statements. Each span is assigned a sufficiency score or label, and tokens within spans flagged as insufficient are projected into the corruption distribution.

**Combined corruption distribution.** Both signals are unified into a single masking distribution:

$$p_{\text{mask}}(i) \propto (1-\text{Conf}(s_i))+\gamma\cdot(1-\text{Suff}(s_i)), \quad (5)$$

where $\text{Conf}(s_i)$ is the token-level confidence, $\text{Suff}(s_i)$ the span-level sufficiency assigned to the token's enclosing span, and $\gamma$ a balancing coefficient. At each denoising step, we sample tokens for resampling according to this distribution.

### 3.4 Sufficiency Diagnosis Module

To implement the global component of inference-time masking, we introduce a sufficiency diagnosis module that evaluates whether generated spans are adequately grounded in the input arguments. Here, a summary span is defined at the sentence level (or clause-level segmentation for long sentences) to ensure coherent semantic units. A span is considered *sufficient* if it is consistent with the provided claims and evidence while avoiding redundancy with other spans. This module provides span-level sufficiency scores that complement token-level confidence, enabling the system to revise both uncertain tokens and confidently produced but unsupported statements. We explore two complementary approaches.

**Chain-of-Thought Prompting.** We construct few-shot prompts for large language models to explicitly reason about sufficiency. Given a candidate span $s$ and its supporting arguments $\{(c_i, E_i)\}$, the model is prompted to determine whether $s$ is fully supported, insufficient, or redundant. This prompting strategy provides interpretable rationales and flexible adaptation across domains, though it may introduce variability depending on model scale and prompting quality.

**Supervised Sufficiency Classifier.** In parallel, we develop a RoBERTa-based classifier to provide stable sufficiency signals. For each span $s$, claim $c_i$, and evidence $E_i$, the classifier predicts a binary sufficiency label:

$$\mathcal{C}_\theta(s, c_i, E_i) = \sigma\left(f_\theta\left([\![s; c_i; E_i]\!]\right)\right), \quad (6)$$

where $f_\theta$ is the encoder and $[\![\cdot]\!]$ denotes concatenation with segment indicators. Training data are constructed via controlled perturbations (hallucinated, contradictory, or unsupported variants) to ensure coverage of common insufficiency types. During inference, the classifier provides span-level sufficiency scores $\text{Suff}(s_i)$ that are integrated into the corruption distribution, complementing low-confidence token masking.

### 3.5 Training and Implementation Details

We adopt a two-stage training strategy for **Arg-LLaDA**, comprising diffusion-based denoising and sufficiency classification. Importantly, sufficiency guidance is introduced only at inference time: training uses diverse random masking to teach general denoising skills, while inference selectively leverages sufficiency signals for targeted refinement. This separation avoids overfitting to noisy sufficiency pseudo-labels and maintains generalization.

**Diffusion Backbone Fine-tuning.** We initialize the diffusion model with the LLaDA checkpoint (Nie et al., 2025), pretrained on large-scale web corpora with masked diffusion objectives. Its architecture and noise scheduling are retained. For fine-tuning, we supervise the model on domain-specific argument summarization data. Given input tuples of claims and evidence with reference summaries, we apply stochastic masking (30% ratio). The model is then optimized to denoise the masked summary using the diffusion objective (Section 3.2).

**Sufficiency Classifier.** We train a RoBERTa-base encoder (Liu et al., 2019) to predict span-level sufficiency labels. Each training instance consists of a sentence-level span $s$, its associated claim $c_i$, and supporting evidence $E_i$. Labels are automatically generated through controlled perturbations, including hallucinations, contradictions, and omissions. The classifier is trained with binary cross-entropy loss and applied only during inference to guide selective span masking.

**Optimization and Setup.** We use the AdamW optimizer with linear learning rate decay. Batch size is set to 32, with early stopping based on validation BLEURT score. All experiments are conducted on two NVIDIA A100 80GB GPUs with FP16 precision.

## 4 Experimental Setup

Broadly speaking, we aim to investigate the effectiveness of our proposed **Arg-LLaDA** framework in addressing the limitations of current argument summarization methods. Specifically, we pose the following research questions: *(i)* Does **Arg-LLaDA** improve the overall quality of argument summaries compared to standard one-pass or prompt-based methods? *(ii)* To what extent does the sufficiency-based span diagnosis contribute to reducing redundancy and enhancing faithfulness? *(iii)* What is the impact of different sufficiency checking modules?

**Baselines.** We compare **Arg-LLaDA** against ten state-of-the-art baselines, categorized into three groups: *(i)* Zero-shot proprietary LLMs, including GPT-4 (OpenAI, 2023), GPT-o3 [1], Gemini 2.5 (Anil et al., 2023), Claude 3.7 [2], and DeepSeek-R1 (DeepSeek-AI et al., 2025), which are closed-source commercial models accessed via API without any task-specific adaptation; *(ii)* Zero-shot open-source LLMs, such as LLaMA 3 (Dubey et al., 2024), LLaDA (Nie et al., 2025), Mistral (Jiang et al., 2023), Gemma (Rivière et al., 2024) and FlanT5 (Chung et al., 2024), which are publicly released and evaluated in their instruction-tuned form without additional training; *(iii)* Full-shot open-source LLMs, which are further fine-tuned on benchmark datasets to adapt them explicitly to the argument summarization task.

**Dataset.** The ArgKP dataset consists of approximately 27k labeled argument–key point pairs covering 31 diverse debate topics, with each pair annotated for stance polarity (Bar-Haim et al., 2020a,b). Following prior work, we partition the dataset into training and test sets in a 28:3 ratio while preserving topic distribution. In addition, we employ and extand the ASE dataset, which provides a comprehensive end-to-end benchmark for argument (Li et al., 2024a) summarization. It spans the full pipeline—from claim and evidence detection to the generation and evaluation of debate-ready summaries—enabling holistic assessment of argument generation systems.

**Metrics.** We report **ROUGE** (Lin, 2004), which common used in text summarization. While ROUGE captures surface-level similarity, it may overlook semantic equivalence in cases of lexical variation. To do so, we consider **BERTScore** (Zhang et al., 2020) and **BLEURT** (Sellam et al., 2020) evaluates summary quality based on contextualized token embeddings obtained from a pre-trained BERT model. They better account for semantic similarity than ROUGE and are less sensitive to exact word matching.

**Human Evaluation.** While automatic metrics provide efficient ways to evaluate summarization quality, they are known to exhibit several limitations—particularly in the context of argument summarization. These metrics primarily rely on surface-level lexical or semantic overlap with reference summaries, and thus often fail to capture deeper aspects such as logical structure, factual grounding, and the nuanced coverage of multiple argumentative facets. Moreover, they may assign high scores to outputs that are fluent but contain hallucinations or overlook important viewpoints.

To address these shortcomings, we complement our automatic evaluation with a human assessment focused on three essential dimensions: **Coverage**, **Faithfulness**, and **Conciseness** (Altemeyer et al., 2025). These criteria reflect core desiderata of argumentative summaries—namely, the ability to accurately capture all salient perspectives, avoid unsupported or fabricated content, and maintain clarity without redundancy.

Each summary is evaluated by three trained annotators with backgrounds in argument mining or discourse analysis. Given the input claims and evidence, annotators rate the summary using a 5-point Likert scale (1 = very poor, 5 = excellent) according to:

- **Coverage**: Does the summary reflect all relevant and diverse viewpoints from the input?
- **Faithfulness**: Are the summarized statements well-grounded in the original claims and evidence, without introducing hallucinated or contradictory content?
- **Conciseness**: Is the summary free from redundancy and excessive verbosity, while still conveying the complete argumentative structure?

All outputs are anonymized and randomly ordered to mitigate evaluator bias. We report the average scores across all annotators, and assess inter-annotator agreement via Krippendorff's alpha. This

[1] https://openai.com/index/introducing-o3-and-o4-mini/
[2] https://www.anthropic.com/news/claude-3-7-sonnet

| | Methods | *ASE* | | | | | *ArgKP* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R-1 | R-2 | R-L | BLE | BER | R-1 | R-2 | R-L | BLE | BER |
| **Zero-shot** | *GPT-4* | 0.314 | 0.091 | 0.302 | 0.512 | 0.703 | 0.402 | 0.192 | 0.374 | 0.656 | 0.809 |
| | *GPT-o3* | 0.333 | 0.105 | 0.311 | 0.566 | 0.722 | 0.453 | 0.233 | 0.419 | 0.619 | 0.811 |
| | *Gemini 2.5* | 0.341 | 0.128 | 0.337 | 0.577 | 0.744 | 0.468 | 0.245 | 0.436 | 0.637 | 0.823 |
| | *Claude* | 0.334 | 0.119 | 0.326 | 0.568 | 0.740 | 0.454 | 0.229 | 0.418 | 0.626 | 0.820 |
| | *DeepSeek-R1* | 0.361 | 0.137 | 0.342 | **0.587** | **0.746** | 0.475 | 0.257 | 0.441 | 0.601 | 0.769 |
| | *LLaMA 3.1* | 0.328 | 0.097 | 0.304 | 0.524 | 0.721 | 0.443 | 0.215 | 0.406 | 0.614 | 0.818 |
| | *LLaDA* | 0.332 | 0.105 | 0.312 | 0.545 | 0.721 | 0.455 | 0.231 | 0.410 | 0.618 | 0.921 |
| | *Mistral* | 0.291 | 0.083 | 0.277 | 0.503 | 0.683 | 0.344 | 0.228 | 0.305 | 0.512 | 0.781 |
| | *Gemma* | 0.282 | 0.081 | 0.276 | 0.544 | 0.714 | 0.457 | 0.233 | 0.422 | 0.629 | 0.822 |
| | *FlanT5* | 0.266 | 0.076 | 0.235 | 0.492 | 0.667 | 0.342 | 0.197 | 0.313 | 0.402 | 0.754 |
| **Fine-tuning** | *LLaMA 3.1* | 0.353 | 0.122 | 0.338 | 0.571 | 0.728 | 0.493 | 0.266 | 0.421 | 0.635 | 0.822 |
| | *LLaDA* | 0.355 | 0.131 | 0.347 | 0.575 | 0.738 | 0.501 | 0.267 | 0.428 | 0.629 | 0.813 |
| | *Mistral* | 0.325 | 0.099 | 0.307 | 0.533 | 0.726 | 0.382 | 0.183 | 0.347 | 0.537 | 0.784 |
| | *Gemma* | 0.300 | 0.099 | 0.291 | 0.532 | 0.710 | 0.511 | 0.269 | **0.455** | 0.643 | 0.830 |
| | *FlanT5* | 0.290 | 0.085 | 0.281 | 0.503 | 0.684 | 0.404 | 0.203 | 0.382 | 0.551 | 0.797 |
| | **Arg-LLaDA** | **0.363** | **0.138** | **0.349** | 0.580 | 0.741 | **0.525** | **0.271** | 0.453 | **0.658** | **0.834** |

Table 2: Comparison of **Arg-LLaDA** against zero-shot and fine-tuned LLMs on two argument summarization datasets. We report ROUGE (R-1, R-2, R-L), BLEURT, and BERTScore. The best and second-best results are respectively **bolded** and underlined.

evaluation offers a more comprehensive view of summary quality, complementing automatic scores with expert human judgment.

## 5 Results and Analysis

To evaluate the effectiveness of our proposed **Arg-LLaDA** framework, we revisit the three core research questions posed earlier.

**Arg-LLaDA significantly outperforms both zero-shot and fine-tuned large language models (LLMs) in generating high-quality argument summaries.** As shown in Table 2), Arg-LLaDA achieves the highest ROUGE-1 (0.525), ROUGE-2 (0.271), and ROUGE-L (0.453) on the *ArgKP* dataset, outperforming both fine-tuned (*Gemma*, *LLaDA*) and zero-shot (*Gemini 2.5*, *DeepSeek-R1*) baselines. These results are consistent across the *ASE* dataset, where Arg-LLaDA also leads in ROUGE-L (0.349), BLEURT (0.580), and BERTScore (0.741), indicating strong lexical and semantic overlap with human-written references. Compared to its closest competitors—*Gemini 2.5* and *DeepSeek-R1*—Arg-LLaDA shows modest but consistent improvements across all dimensions, particularly in BLEURT and BERTScore, which are more sensitive to semantic adequacy and contextual fidelity.

Human evaluation (Figure 2) further corroborates these findings. Arg-LLaDA receives the highest average scores across all three dimensions: Coverage (4.7), Faithfulness (4.6), and Conciseness (4.4). These results suggest that Arg-LLaDA not only captures a broader spectrum of argumentative content but also avoids common issues such as hallucination or redundancy. In contrast, while models like *Gemini 2.5* (4.5/4.4/4.2) and *DeepSeek-R1* (4.4/4.3/4.1) exhibit strong performance, they tend to generate longer outputs that introduce occasional repetition or irrelevant detail, resulting in slightly lower conciseness scores.

**The integration of sufficiency-based span diagnosis plays a pivotal role in improving both the conciseness and faithfulness of generated summaries.** As shown in Table **??**, Arg-LLaDA outperforms all baselines in the Faithfulness (4.6) and Conciseness (4.4) dimensions. These improvements can be attributed to the model's explicit mechanism for identifying and filtering spans that lack sufficient argumentative support before generation. Compared to its fine-tuned counterparts (e.g., *Gemma*, 4.3/4.2; *LLaDA*, 4.2/4.1), Arg-LLaDA demonstrates superior control over hallucinated

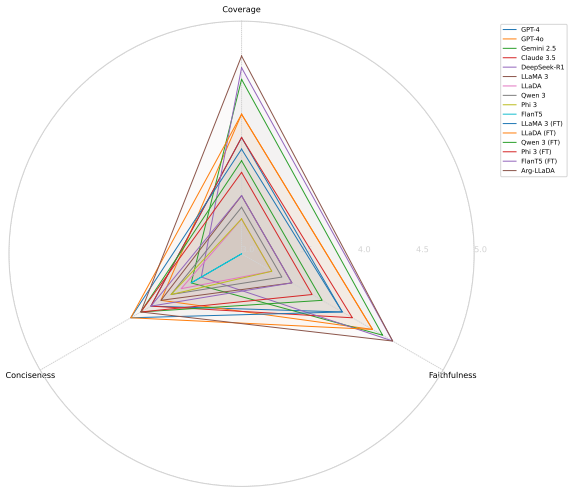


Figure 2: Human evaluation results across three dimensions, using a 5-point Likert scale.

or redundant content. This is particularly evident when contrasting with zero-shot models like *Claude* or *Qwen 3*, whose summaries frequently exhibit ungrounded generalizations or verbose phrasing, as reflected in lower average ratings for faithfulness and conciseness (both $< 4.0$).

| Iterations | R-L | BLEURT | BERTScore |
|---|---|---|---|
| 0 | 0.429 | 0.629 | 0.814 |
| 1 | 0.448 | 0.650 | 0.829 |
| 2 | 0.451 | 0.657 | 0.831 |
| 3 | 0.453 | 0.658 | 0.834 |

Table 3: Ablation study: varying the number of iterative refinement steps.

**The design and integration of the sufficiency checking module critically affect final summarization performance, both in lexical overlap and semantic adequacy.** As shown in Table 4, removing the sufficiency diagnosis module ("No Diagnosis") results in a significant drop across all evaluation metrics—ROUGE-L drops from 0.453 to 0.428, BLEURT from 0.658 to 0.629, and BERTScore from 0.834 to 0.813. This indicates that without explicit sufficiency filtering, the generation becomes more verbose and includes unsupported content, thereby reducing both faithfulness and conciseness.

When comparing alternative sufficiency strategies, we observe that simple classifier-based diagnosis yields moderate improvements over the no-diagnosis baseline (+0.019 in ROUGE-L), while CoT-style prompting achieves slightly better results. However, our proposed combined approach, which integrates classifier predictions with chain-of-thought (CoT) rationale prompting, consistently achieves the best results across all metrics—demonstrating a synergistic effect when both implicit (reasoning-driven) and explicit (discriminative) sufficiency signals are fused.

| Sufficiency Module | R-L | BLEURT | BERTScore |
|---|---|---|---|
| No Diagnosis | 0.428 | 0.629 | 0.813 |
| CoT Prompting Only | 0.451 | 0.649 | 0.827 |
| Classifier Only | 0.447 | 0.645 | 0.823 |
| Combine | 0.453 | 0.658 | 0.834 |

Table 4: Ablation study: comparing sufficiency diagnosis strategies.

## 6 Conclusion

In this work, we introduced **Arg-LLaDA**, a novel large language diffusion framework for argument summarization that integrates sufficiency-guided remasking and iterative regeneration. Our method addresses key limitations of existing approaches by enabling controllable refinement of summaries through a flexible masking controller and sufficiency diagnosis module. Empirical results across two benchmark datasets demonstrate that Arg-LLaDA consistently outperforms state-of-the-art baselines. These findings highlight the potential of diffusion-based self-refinement in enhancing the factuality and structure of argumentative summaries.

## Limitations

Despite the promising results of **Arg-LLaDA**, several limitations warrant discussion. First, although we trained annotators with backgrounds in argument mining and provided clear evaluation guidelines, human judgments remain inherently subjective. Variations in interpretation or attention to detail may affect the consistency of annotation, especially for nuanced aspects like sufficiency and redundancy. Second, while our datasets cover a wide range of debate topics, some of these topics may contain controversial or sensitive content. But such content is not the focus of this work.

## References


Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. In *EMNLP/IJCNLP (1)*, pages 2922–2932. Association for Computational Linguistics.

Milad Alshomary, Timon Gurcke, Shahbaz Syed, Philipp Heinisch, Maximilian Spliethöver, Philipp Cimiano, Martin Potthast, and Henning Wachsmuth. 2021. Key point analysis via contrastive learning and extractive argument summarization. In *ArgMining@EMNLP*, pages 184–189. Association for Computational Linguistics.

Moritz Altemeyer, Steffen Eger, Johannes Daxenberger, Tim Altendorf, Philipp Cimiano, and Benjamin Schiller. 2025. Argument summarization and its evaluation in the era of large language models. *CoRR*, abs/2503.00847.

Konstantinos Andriopoulos and Johan Pouwelse. 2023. Augmenting llms with knowledge: A survey on hallucination prevention. *CoRR*, abs/2309.16459.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020a. From arguments to key points: Towards automatic argument summarization. In *ACL*, pages 4029–4039. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Eden, Yoav Kantor, Roni Friedman, and Noam Slonim. 2021. Every bite is an experience: Key point analysis of business reviews. In *ACL/IJCNLP (1)*, pages 3376–3386. Association for Computational Linguistics.

Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020b. Quantitative argument summarization and beyond: Cross-domain key point analysis. In *EMNLP (1)*, pages 39–49. Association for Computational Linguistics.

Xiuying Chen, Mingzhe Li, Xin Gao, and Xiangliang Zhang. 2022. Towards improving faithfulness in abstractive summarization. In *NeurIPS*.

Liying Cheng, Lidong Bing, Ruidan He, Qian Yu, Yan Zhang, and Luo Si. 2022. IAM: A comprehensive and large-scale dataset for integrated argument mining tasks. In *ACL (1)*, pages 2277–2287. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25:70:1–70:53.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes,

Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Lilach Eden, Yoav Kantor, Matan Orbach, Yoav Katz, Noam Slonim, and Roy Bar-Haim. 2023. Welcome to the real world: efficient, incremental and scalable key point analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 483–491.

Tomoki Fukuma, Koki Noda, Toshihide Ubukata, Kousuke Hosoi, Yoshiharu Ichikawa, Kyosuke Kambe, Yu Masubuchi, and Fujio Toriumi. 2024. QANA: llm-based question generation and network analysis for zero-shot key point analysis and beyond. *CoRR*, abs/2404.18371.

Olivera Grljevic, Zita Bosnjak, and Aleksandar Kovacevic. 2022. Opinion mining in higher education: a corpus-based approach. *Enterp. Inf. Syst.*, 16(5).

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2024. Mining legal arguments in court decisions. *Artif. Intell. Law*, 32(3):1–38.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Manav Nitin Kapadnis, Sohan Patnaik, Siba Smarak Panigrahi, Varun Madhavan, and Abhilash Nandy. 2021. Team enigma at argmining-emnlp 2021: Leveraging pre-trained language models for key point matching. In *ArgMining@EMNLP*, pages 200–205. Association for Computational Linguistics.

Mohammad Khosravani, Chenyang Huang, and Amine Trabelsi. 2024. Enhancing argument summarization: Prioritizing exhaustiveness in key point generation and introducing an automatic coverage evaluation metric. In *NAACL-HLT*, pages 8212–8224. Association for Computational Linguistics.

Hao Li, Bowen Deng, Chang Xu, Zhiyuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, and Jiang Bian. 2025a. MIRA: medical time series foundation model for real-world health data. *CoRR*, abs/2506.07584.

Hao Li, Yu-Hao Huang, Chang Xu, Viktor Schlegel, Ren-He Jiang, Riza Batista-Navarro, Goran Nenadic, and Jiang Bian. 2025b. BRIDGE: bootstrapping text to control time-series generation via multi-agent iterative optimization and diffusion modelling. *CoRR*, abs/2503.02445.

Hao Li, Viktor Schlegel, Riza Batista-Navarro, and Goran Nenadic. 2023. Do you hear the people sing? key point analysis via iterative clustering and abstractive summarisation. In *ACL (1)*, pages 14064–14080. Association for Computational Linguistics.

Hao Li, Viktor Schlegel, Yizheng Sun, Riza Batista-Navarro, and Goran Nenadic. 2025c. Large language models in argument mining: A survey. *arXiv preprint arXiv:2506.16383*.

Hao Li, Yuping Wu, Viktor Schlegel, Riza Batista-Navarro, Tharindu Madusanka, Iqra Zahid, Jiayan Zeng, Xiaochi Wang, Xinran He, Yizhi Li, and Goran Nenadic. 2024a. Which side are you on? A multi-task dataset for end-to-end argument summarisation and evaluation. In *ACL (Findings)*, pages 133–150. Association for Computational Linguistics.

Xiao Li, Yong Jiang, Shen Huang, Pengjun Xie, Gong Cheng, and Fei Huang. 2024b. Exploring key point analysis with pairwise generation and graph partitioning. In *NAACL-HLT*, pages 5657–5667. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. 2023. Argugpt: evaluating, understanding and identifying argumentative essays generated by GPT models. *CoRR*, abs/2304.07666.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Amita Misra, Shereen Oraby, Shubhangi Tandon, Sharath T. S., Pranav Anand, and Marilyn A. Walker. 2017. Summarizing dialogic arguments from social media. *CoRR*, abs/1711.00092.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. *CoRR*, abs/2502.09992.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. 2025. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *ICLR*. OpenReview.net.

Viet Hoang Phan, Tien Long Nguyen, Duc Long Nguyen, and Ngoc Khanh Doan. 2021. Matching the statements: A simple and accurate model for key point analysis. In *ArgMining@EMNLP*, pages 165–174. Association for Computational Linguistics.

Jan Heinrich Reimer, Thi Kim Hanh Luu, Max Henze, and Yamen Ajjour. 2021. Modern talking in key point analysis: Key point matching using pretrained encoders. In *ArgMining@EMNLP*, pages 175–183. Association for Computational Linguistics.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *ACL (1)*, pages 567–578. Association for Computational Linguistics.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos Garea, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. In *ACL (1)*, pages 6252–6272. Association for Computational Linguistics.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *NAACL-HLT*, pages 380–396. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *ACL*, pages 7881–7892. Association for Computational Linguistics.

Huawei Shan and Dongyuan Lu. 2025. Debate divides: Argument relation-based contrastive opinion summarization via multi-task learning for online discussions. *Neurocomputing*, 638:130124.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. 2024. Simplified and generalized masked diffusion for discrete data. In *NeurIPS*.

An Quang Tang, Xiuzhen Zhang, and Minh Ngoc Dinh. 2024a. Aspect-based key point analysis for quantitative summarization of reviews. In *EACL (Findings)*, pages 1419–1433. Association for Computational Linguistics.

An Quang Tang, Xiuzhen Zhang, Minh Ngoc Dinh, and Erik Cambria. 2024b. Prompted aspect key point analysis for quantitative review summarization. In *ACL (1)*, pages 10691–10708. Association for Computational Linguistics.

Michiel van der Meer, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2024. An empirical analysis of diversity in argument summarization. In *EACL (1)*, pages 2028–2045. Association for Computational Linguistics.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *ACL/IJCNLP (1)*, pages 1338–1352. Association for Computational Linguistics.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *HLT-NAACL*, pages 47–57. The Association for Computational Linguistics.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. 2025. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*. OpenReview.net.

## A Instruction for Zero-shot Inference.

We provide the instruction we used to prompt baseline model to complete the summary task on both dataset:

*You are a debate writing assistant. Based on the structured information provided below, generate a* ***concise and persuasive debate speech*** *in support of the motion:* ***"[Insert Debate Topic]"****.*

*The speech should follow this structure:*

- ***Introduction****: Clearly state the affirmative stance and why the topic matters.*
- ***Body****: Present each supporting claim, explain its reasoning, and integrate the evidence in a natural, narrative style. You do not need to use all the evidence—just enough to make each point persuasive.*
- ***Conclusion****: Summarize the main points and restate why the position is valid and important.*

*Use the following input fields to guide the speech:*

*Guidelines:*

- *Select evidence selectively—strong, relevant examples are enough.*
- *Write in fluent, professional, and persuasive English.*
- *Output should be in paragraph format, suitable for public speaking.*
- *Avoid exaggeration; keep the reasoning grounded in the given claims and evidence.*