Reranking-based Generation for Unbiased Perspective Summarization

Narutatsu Ri and **Nicholas Deas** and **Kathleen McKeown**
Department of Computer Science, Columbia University
{wl2787, nid2107, km}@columbia.edu

Abstract

Generating unbiased summaries in real-world settings such as political perspective summarization remains a crucial application of Large Language Models (LLMs). Yet, existing evaluation frameworks rely on traditional metrics for measuring key attributes such as coverage and faithfulness without verifying their applicability, and efforts to develop improved summarizers are still nascent. We address these gaps by (1) identifying reliable metrics for measuring perspective summary quality, and (2) investigating the efficacy of LLM-based methods beyond zero-shot inference. Namely, we build a test set for benchmarking metric reliability using human annotations and show that traditional metrics underperform compared to language model–based metrics, which prove to be strong evaluators. Using these metrics, we show that reranking-based methods yield strong results, and preference tuning with synthetically generated and reranking-labeled data further boosts performance. Our findings aim to contribute to the reliable evaluation and development of perspective summarization methods.

1 Introduction

Article summarization is a key application of Large Language Models (LLMs) given their recent breakthroughs in text generation capabilities (Goyal et al., 2023; Zhang et al., 2024a). Critically, however, LLMs often exhibit undesirable behaviors and input-level biases toward spurious features (e.g., position) (Jung et al., 2019; Chhabra et al., 2024; Liu et al., 2024a), resulting in unbalanced input coverage (Zhang et al., 2024c) and hallucination (Maynez et al., 2020). These issues are especially problematic in opinionated article summarization (Amplayo et al., 2021; Iso et al., 2022), where unbiased representation of diverse viewpoints is crucial.

Recent studies in opinion summarization address these risks by developing tasks and methods that generate summaries free of framing bias (Lee et al., 2022a), fairly represent input diversity (Zhang et al., 2024c; Feng et al., 2024), or preserve the source perspectives (Lei et al., 2024; Liu et al., 2024b). Within this domain, *perspective summarization* (Deas and McKeown, 2025) serves as a representative evaluation setting, where models are tasked to generate precise, perspective-specific summaries from multi-document inputs containing diverse political views. However, two gaps remain unaddressed in this setting: (1) existing evaluation metrics are primarily derived from news summarization domains and have not been validated for measuring *perspective* summary quality, and (2) the effectiveness of LLM-based methods beyond zero-shot inference in generating unbiased, high-quality perspective summaries remains underexplored.

To address these gaps, we first identify effective metrics for measuring summary quality by constructing a test set to evaluate existing metrics. We focus on two key attributes that a desirable summary should have: *perspective coverage*—the extent to which the summary includes all key content from the intended perspective, and *perspective faithfulness*—the degree to which the summary excludes content unsupported by the source articles of the target perspective. We collect key point annotations from articles to create controlled summaries with varied key point selections and assigned ground truth scores. We find that language model-based metrics such as ALIGNSCORE (Zha et al., 2023) and prompting-based scoring (Zheng et al., 2023) serve as strong evaluators, whereas traditional metrics (ROUGE (Lin, 2004), BERTSCORE (Zhang et al., 2020)) underperform.

Following this, we evaluate methods for generating perspective summaries with improved coverage and faithfulness beyond zero-shot inference. We benchmark prompting frameworks, mechanistic methods for mitigating input biases, and reranking-based methods that select the best candidate based on proxy metrics. Using both human and automatic evaluations, we show that reranking outperforms

Findings of the Association for Computational Linguistics: ACL 2025, pages 24701–24723
July 27 - August 1, 2025

zero-shot inference and prompting-based methods, while prompting only yields marginal improvements over zero-shot inference. Notably, preference tuning with Direct Preference Optimization (DPO) (Rafailov et al., 2023) on reranked generations further boosts performance on both attributes and particularly improving faithfulness. Our results suggest that current LLMs can generate high-quality perspective summaries with strong coverage and faithfulness, and that preference-based training can further boost performance.

In summary, our contributions are as follows:

- We construct a controlled test set and identify effective metrics for measuring coverage and faithfulness for perspective summarization.
- We evaluate various generation methods and demonstrate that reranking-based approaches deliver the best performance in producing summaries with improved coverage perspective and faithfulness. Notably, preference tuning on reranked generations significantly improves both attributes, with the most pronounced gains in faithfulness.
- We conduct ablation studies and show that commonly employed prompting frameworks consistently underperform relative to reranking-based methods, even when scaled to high-resource settings.

2 Related Work

Summary Evaluation. Summary evaluation traditionally relies on reference-based metrics, including n-gram-based methods (ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), CHRF (Popović, 2015)), model-based coverage scores (BERTSCORE (Zhang et al., 2020), BLEURT (Sellam et al., 2020)), and composite measures (METEOR (Banerjee and Lavie, 2005)). In response to unreliable references, recent work proposes reference-free metrics that target aspects such as faithfulness and factual consistency. Neural approaches dominate this space, including end-to-end classifiers (FactCC (Kryscinski et al., 2020)), QA-based methods (QAGS (Wang et al., 2020), QAFactEval (Fabbri et al., 2022)), NLI models (SUMMAC (Laban et al., 2022)), and information alignment models (ALIGNSCORE (Zha et al., 2023)). Here, we focus on automatic, reference-free measures of coverage and faithfulness, but conduct a novel evaluation of their reliability in a multi-document perspective summarization task.

Beyond developing improved faithfulness metrics, prior works focus on improving the factual consistency of summarizers, with studies noting the tradeoff between abstractiveness and faithfulness (Durmus et al., 2020; Dreyer et al., 2023). Accordingly, some methods improve faithfulness without increasing extraction (Ladhak et al., 2022), while others modify training via contrastive (Nan et al., 2021), multi-task (Chen et al., 2022), or reinforcement learning (Roit et al., 2023) methods. In contrast, we show that reranking-based methods serve as a strong baseline that yields high faithfulness without sacrificing abstractiveness, and a DPO-based approach trained on reranked self-generated summaries further improves both qualities.

Perspective-Conditioned Summarization. Existing research on opinion summarization and related tasks has primarily focused on domains such as product reviews (Bražinskas et al., 2020), while recent work has broadened to a range of tasks on opinionated texts. Most single-document methods aim to preserve authorial intent (Liu et al., 2024b) or polarity (Lei et al., 2024), whereas multi-document summarization must integrate varied perspectives. For instance, Lee et al. (2022b) generates politically neutral summaries from sets of left-, right-, and center-leaning news articles. Other approaches aim to fairly represent diverse perspectives in reviews (Zhang et al., 2024c), controllably represent community perspectives (Feng et al., 2024), generate consensus summaries (Bakker et al., 2022), or produce multiple summaries reflecting distinct political perspectives (Deas and McKeown, 2025). In line with these works, we summarize the political perspective among a set of input passages while addressing the coverage and faithfulness issues observed in existing models as highlighted in these studies.

3 Measuring Summary Quality

In perspective summarization, the summarizer is given two perspectives θ_1, θ_2, each with a source article $D_{t,\theta}, \theta \in \{\theta_1, \theta_2\}$, comprising a set of documents $D_{t,\theta} = \{d_{t,\theta}^{(i)} \mid i \in \mathbb{N}\}$, that present opinions on topic t. We study the setting where the summarizer is tasked to generate a summary that encapsulates all *key points* directly supporting a specified perspective's stance. Concretely, a high-quality perspective summary should: (1) include all key points from each relevant document, and (2) avoid including any content unsupported by or in opposition to the perspective's documents. We formalize these properties as follows:

Article Topic	Ron DeSantis
Perspective Source Article (Key Points)	DeSantis has shown authoritarian tendencies throughout his time in office. DeSantis' election police proposal chills legitimate election work and threatens democracy. DeSantis' claim that Florida is the freest state contradicts restrictions on health, protest, and education.
Synthetic (High-Quality)	The article contends that DeSantis's proposal for an election police squad undermines legitimate election activities and democracy, contradicts his claim of Florida being the freest state by restricting various freedoms, and highlights his persistent authoritarian inclinations during his tenure.
Synthetic (Low-Quality)	The article highlights DeSantis's authoritarian tendencies and his contradiction in calling Florida the freest state while restricting freedoms, but praises his election police proposal for protecting elections and strengthening democracy and urges Trump to prioritize GOP leadership in Florida and retaking the House over personal pride.

Table 1: Examples of constructed summaries. For brevity, only curated key points are shown for the source article. Purple, blue, and green highlights denote relevant key points, while red and orange highlights respectively indicate adversarial and opposite key points.

- *Perspective Coverage*: The ratio of key points included in the summary relative to the total number of key points.
- *Perspective Faithfulness*: The ratio of relevant key points included in the summary relative to the total number of included key points.

Although metrics for similar properties exist in other summarization domains, it is unclear whether they effectively measure the properties as defined above for the perspective summarization task. We therefore evaluate how well these metrics capture our definitions of coverage and faithfulness.[1]

3.1 Assessing Metric Quality

Quantifying the efficacy of existing metrics requires article-summary pairs with ground truth scores for evaluation. Although perspective summarization datasets such as POLISUM (Deas and McKeown, 2025) include reference summaries, each document is paired with only one summary without assigned scores for coverage and faithfulness. Hence, we construct a test set of article-summary pairs with assigned ground truth scores for coverage and faithfulness and evaluate how well existing metrics align with these scores.

Test Set Construction. To assign meaningful ground truth scores for both attributes, we identify all key points in an article and create summaries

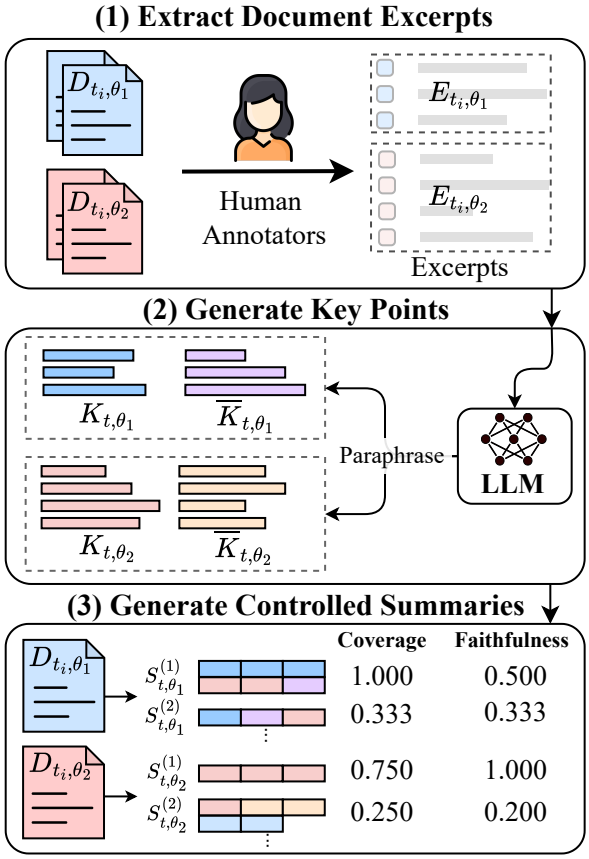


Figure 1: Pipeline for curating the synthetic testbed for metric evaluation. Annotators extract the most important excerpts $E_{t,\theta}$ from articles $D_{t,\theta}$, which are paraphrased into key points $K_{t,\theta}$ and adversarial key points $\overline{K}_{t,\theta}$. We then curate summaries with a diverse range of coverage and faithfulness scores using the key points.

using different combinations of these points. We begin with articles from POLISUM[2] and collect human annotations in which annotators highlight document excerpts supporting the perspective's stance. See §D.2 for the annotation interface.

Formally, given an article $D_{t,\theta}$, we collect a set of excerpts $E_{t,\theta}$ defined as:

$$E_{t,\theta} = \{e^{(i)}_{t,\theta} \mid e^{(i)}_{t,\theta} \subseteq d^{(i)}_{t,\theta}, d^{(i)}_{t,\theta} \text{ contains a key point}\},$$

where $|E_{t,\theta}| \leq |D_{t,\theta}|$ (i.e., not all documents contain key points, and each document has at most one key point). As an excerpt $e^{(i)}_{t,\theta}$ may not clearly convey the main argument, we use an LLM $f : E_{t,\theta} \to K_{t,\theta}$ to rewrite excerpts into key points to form the set $K_{t,\theta}$:[3]

$$K_{t,\theta} = \{k^{(i)}_{t,\theta} \mid k^{(i)}_{t,\theta} = f(e^{(i)}_{t,\theta}),\ e^{(i)}_{t,\theta} \in E_{t,\theta}\}.$$

Given $K_{t,\theta}$, we construct summaries $S^{(i)}_{t,\theta}$ by selecting k_g key points from $K_{t,\theta}$ and k_b from a set

[1] We note that our notions of coverage and faithfulness differ from prior work (Zhang and Bansal, 2021; Tang et al., 2024; Song et al., 2024), as we assess both attributes with respect to the correct inclusion of key points. For brevity, we use perspective coverage and faithfulness interchangeably with coverage and faithfulness.

[2] In the POLISUM dataset, $|D_{t,\theta}|$ has a mean of 5.31 with a standard deviation of 1.45.

[3] See §B.4 for further details.

```
You are an evaluator. Your task is to determine how well a
generated summary captures all of the main arguments from a
source article. This is a measure of "coverage," which does
not necessarily address factual accuracy (faithfulness) but
focuses on completeness of content. The scale for coverage is:
1. No Coverage: The summary does not include any of the main
arguments from the article.
2. Low Coverage: The summary includes only a few of the main
arguments from the article, omitting most.
3. Medium Coverage: The summary contains around half of the
article's main arguments.
4. High Coverage: The summary contains most of the main
arguments from the article, missing only a few.
5. Perfect Coverage: The summary includes all major points
mentioned in the article, leaving out nothing important.

Follow these steps carefully:
(Omitted for Brevity)

# Source Article:
(article)
# Summary:
(summary)
# Coverage Score (1~5 only):
```

Figure 2: Example prompt for LLM-Coverage. We follow the prompt instruction format in Wu et al. (2024). Portions of the prompt are omitted for brevity. See §B.2 for complete prompt instructions.

of unfaithful key points. We generate unfaithful key points by sampling key points from the opposing perspective (e.g., using key points from the left-leaning document for right-perspective summaries), and by reversing the content of key points in $K_{t,\theta}$ to form adversarial key points $\overline{K}_{t,\theta}$ (Laban et al., 2022). We then define:

$$\text{Coverage}(S^{(i)}_{t,\theta}) = \frac{k_g}{|K_{t,\theta}|}, \quad (1)$$

$$\text{Faithfulness}(S^{(i)}_{t,\theta}) = \frac{k_g}{k_g + k_b}. \quad (2)$$

We provide examples of summaries with varying scores in Table 1. With this procedure, we produce summaries with error levels ranging from few minor omissions to many faithfulness errors. We collect annotations for 50 documents from 5 annotators and generate a varying number of summaries for each document, ultimately curating 370 article-summary pairs in total. We illustrate the process in Figure 1. See §D.1 for further annotation details.

Benchmarked Metrics. As baselines, we respectively use the recall and precision variants of **ROUGE** (Lin, 2004) and **BERTSCORE** (Zhang et al., 2020) for measuring coverage and faithfulness. We also report **BLEURT** (Sellam et al., 2020) as an additional coverage metric. For faithfulness, we test **SUMMAC** (Laban et al., 2022) (NLI-based inconsistency detection metric), **ALIGNSCORE** (Zha et al., 2023) (factual consistency metric), the consistency dimension of **UniEval** (Zhong et al., 2022) (T5-based multi-task evaluator), **MiniCheck** (Tang et al., 2024) (FLAN-T5 model for fact-checking via entailment), and the faithfulness dimension of **FineSurE** (Song et al., 2024) (span-level fact verification). See §B.1 for details on metric configurations and model checkpoints.

Metric	Coverage		Faithfulness	
	Corr. (ρ_s)	Winrate	Corr. (ρ_s)	Winrate
ROUGE$_L$ (R)	0.473***	0.780 ± 0.048	−0.038	0.393 ± 0.063
BERTSCORE (R)	0.527***	0.815 ± 0.018	−0.032**	0.415 ± 0.015
BLEURT	0.086	0.530 ± 0.067	−0.014	0.527 ± 0.063
LLM-Coverage	0.707***	0.739 ± 0.047	0.393***	0.431 ± 0.115
ROUGE$_L$ (P)	−0.169**	0.443 ± 0.056	0.333***	0.714 ± 0.076
BERTSCORE (P)	0.073**	0.510 ± 0.030	0.366***	0.655 ± 0.020
SUMMAC	0.028	0.491 ± 0.084	−0.016	0.315 ± 0.066
ALIGNSCORE	0.261***	0.503 ± 0.074	0.650***	0.773 ± 0.061
UniEval (C)	0.267***	0.545 ± 0.055	0.629***	0.768 ± 0.054
MiniCheck	0.099	0.435 ± 0.066	0.578***	0.747 ± 0.074
FineSurE (F)	0.271***	0.288 ± 0.076	0.084	0.216 ± 0.072
LLM-Faithfulness	0.462***	0.398 ± 0.055	0.706***	0.537 ± 0.091

Table 2: Comparison of Spearman correlation (ρ_s) and Winrate with 95% Confidence Interval (CI) across all metrics. Darker shading indicates better performance. Asterisks indicate significance levels (*,** ,*** for $p < 0.05, 0.01, 0.001$, respectively). P and R denote the precision and recall variants of each metric. Note the random baseline for Winrate is 0.500.

Furthermore, recent studies suggest that LLMs serve as effective evaluators (Chiang and Lee, 2023; Dubois et al., 2023; Chen et al., 2023), including for some dimensions of summary qualities (Jain et al., 2023; Wu et al., 2024). Hence, we examine two LLM-as-a-Judge settings where the source article and generated summary are passed as input alongside tailored prompts (Liu et al., 2023). We respectively term these **LLM-Coverage** and **LLM-Faithfulness** for convenience. As an example, see Figure 2 for the LLM-Coverage prompt instruction. We use `Mistral-7B-Instruct-v0.3` as the default backbone based on evaluation performance. See §B.3 for results using alternative models.

Note that, by our formulation, coverage corresponds to recall and faithfulness to precision in key point inclusion. Hence, we report results on both attributes for all metrics and show that recall-based metrics do not capture faithfulness and vice versa to verify the reliability of our curated test set.

Evaluation Criteria. We examine two measures of evaluating metrics: (1) **Correlation**, assessed via Spearman correlation between metric-assigned and ground truth scores, and (2) **Winrate**, the accuracy for which the metric correctly selects the summary with the higher ground truth score. For each source article, we form summary pairs and compute the average ratio of correctly ranked pairs. A desirable metric should achieve high scores for both measures, as correlation gauges true model

performance whereas winrate measures the metric's accuracy in selecting the better summarizer.

3.2 Results

We present our results in Table 2. Overall, LLM-Coverage and ALIGNSCORE serve as reliable metrics for coverage and faithfulness respectively, which we use as automatic evaluators in §5. Notably, metrics for coverage do not effectively measure faithfulness and vice versa, indicating that our testbed assesses these dimensions separately.

We see that although both variants of ROUGE and BERTSCORE do not achieve the highest correlation, they display moderate correlation ($0.376 \sim 0.527$) and winrate alignment ($0.722 \sim 0.815$ on average). In contrast, we see that BLEURT and SUMMAC exhibit poor results for both attributes.

In particular, LLM-Coverage exhibits strong coverage performance with a Spearman correlation of 0.707 and a winrate of 0.739. For faithfulness, LLM-Faithfulness performs the best on correlation, but ALIGNSCORE, UniEval, and MiniCheck exhibit better winrates, also corroborating prior work that suggest LLMs are not yet reliable as standalone measures of faithfulness (Parcalabescu and Frank, 2024; Siegel et al., 2024).

4 Method Evaluation

With reliable metrics established in §3, we now investigate methods for generating improved perspective summaries beyond zero-shot prompting. Notably, due to the absence of large-scale training data, we examine several well-established methods and variants that do not rely on training data. We use `Llama-3.1-8B-Instruct` as the default backbone for all methods.

Prompting-Based Approaches. Much work on LLMs proposes inference-time methods that elicit reasoning and planning (Wang et al., 2023a; Press et al., 2023; Huang et al., 2023; Weng et al., 2023; Zhang et al., 2024b). Such methods have proven effective across various tasks (Wang et al., 2023b; Jacob et al., 2024; Saha et al., 2024; Dhuliawala et al., 2024) and improve factual consistency (Xu et al., 2024). As such, we consider two methods: (1) **Multi-Agent Debate** (Du et al., 2024), where multiple LLMs iteratively update their responses based on one another, and (2) **Self-Refine** (Madaan et al., 2023), where an LLM iteratively critiques and revises its own output. We use the default settings of three agents over three rounds for Debate and three iterations for Self-Refine.

Mechanistic Approach. A natural alternative to zero-shot inference is to direct the model's attention to salient input segments that support the overall perspective. Similar methods have been proposed to mitigate position biases in LLMs using calibration-based (Hsieh et al., 2024) and mechanistic approaches (Ratner et al., 2023; Hu et al., 2024; Liu et al., 2024a). In particular, **PINE** (Wang et al., 2024) modifies causal attention bidirectionally and increases the weight on specified segments. We examine whether controlling the model's attention to segments corresponding to the desired perspective can improve coverage and faithfulness. See §A.1 for additional details.

Reranking Generations. We examine a **Reranking (RR)** approach in which an untrained backbone generates multiple summaries and we select the highest-scoring summary based on LLM-Coverage and LLM-Faithfulness. Prior work has explored similar methods (Vijayakumar et al., 2018; Suzgun et al., 2022) with notable success (Wei et al., 2022; Xu et al., 2024). Benchmarking reranking examines whether the backbone is inherently capable of generating high-quality summaries. In particular, comparing reranking with prompting-based methods, which are more commonly used to improve inference-time performance, assesses the optimal approach for perspective summarization. For reranking-based methods, we use `Qwen2.5-14B-Instruct` as the scorer backbone to avoid incorporating signals from the automatic coverage evaluation, and we generate nine summaries per input for reranking by default.

Preference Tuning with Reranking. Many studies employ reinforcement learning-based training for instruction following (Ouyang et al., 2022; Bai et al., 2022; Nakano et al., 2022), with applications in summary generation (Stiennon et al., 2020; Gooding and Mansoor, 2023; Huang et al., 2024; Lee et al., 2024). However, these approaches typically rely on human feedback (e.g., RLHF) and labeled preference pairs (e.g., DPO (Rafailov et al., 2023)). Here, we examine whether preference-based training on synthetic, reranking-generated data can improve perspective summarization performance. Namely, we consider a **DPO with Reranking (DPO+RR)** approach that iteratively repeats the procedure of generating summaries with the backbone model, scoring them with LLM-Coverage and LLM-Faithfulness, and creating preference pairs by designating higher-scoring sum-

Article Highlights	Two years after a Ruth Bader Ginsburg's death made the elimination of that right more likely than not, President Joe Biden entered the chat. . . The fumbling about for a spell before he awakens to the fire is a recurring pattern. Biden has been notably tucked away. . . With Roe, the situation is particularly galling. . . damning indictment of the administration's messaging. . . Biden appears to be trapped in a vicious cycle. . . he and his advisers appear to be gripped with anxiety that doing anything will only make things worse. Democrats need to give voters reasons to believe. . . the only way to reverse Biden's sliding popularity is for him to step forward and start providing these reasons. Our immediate goal within the Democratic Party is to "dump Biden"
Zero-Shot	The Left believes that President Joe Biden's slow response to the Supreme Court's elimination of the constitutional right to an abortion and his overall lack of effective communication and decisive action are major reasons for his declining popularity and the Democrats' electoral struggles.
DPO+RR	The Left views President Joe Biden's delayed response to the SCOTUS abortion ruling, characterized by "fumbling" actions, as further exacerbating his declining public image and hindering his ability to provide meaningful solutions amidst various national crises, underscoring concerns that his leadership style may undermine democratic values and ultimately harm Democrats' chances at re-election.

Table 3: Example summary generated by Zero-Shot and DPO+RR. Highlights indicate excerpts marked by an annotator. Zero-Shot captures only one of the three excerpts, whereas DPO+RR captures all three key points.

maries as preferred over lower-scoring ones, which are then used to train the backbone. We split the POLISUM dataset (1816 article pairs) into train (1716) and test (100) splits to ensure that synthetic training data is generated exclusively from the train split, and repeat over 10 epochs.

4.1 Evaluation Setup

Automatic Evaluation. We automatically evaluate summary quality using two criteria. First, we assign numerical scores to summaries using LLM-Coverage and ALIGNSCORE (cf. §3). Second, we compute instance-level rankings across all test articles to assess relative method performance. However, as automatic metrics do not rank methods perfectly (cf. Table 2), we address this by fitting a Bradley-Terry model to the pairwise comparisons derived from raw scores and performing bootstrap resampling over the test documents to obtain 95% confidence intervals. This avoids naive "rank-then-average" methods that can yield cyclic or inconsistent preferences when pairwise comparisons do not form a strict total ordering. We use the split test set of 200 input documents for automatic evaluation. Refer to §A.2 for details on the ranking procedure.

Human Evaluation. We collect human judgments on summary quality by having annotators review input documents and their corresponding model-generated summaries. Analogous to the procedure in §3.1, annotators first extract key points from both documents and summaries, then identify which document key points each summary includes or omits, and which summary key points appear in the document. This process yields coverage and faithfulness scores computed as in Eqs. (1) and (2). See §D.1 for further annotation details.

5 Results

We present coverage and faithfulness results in Figures 3a (automatic evaluation) and 3b (human evaluation), and provide generated example summaries in Table 3. We include additional examples in §C.3.

Ratio	Document	Summary
$R(\cdot \mid \cdot)$	0.672 ± 0.262	0.918 ± 0.173
Random	0.235 ± 0.322	0.650 ± 0.380

Table 4: Inter-Annotator Agreement (IAA) results. Values lie between the interval $[0, 1]$. We observe substantial agreement for both document- and summary-level key point extraction.

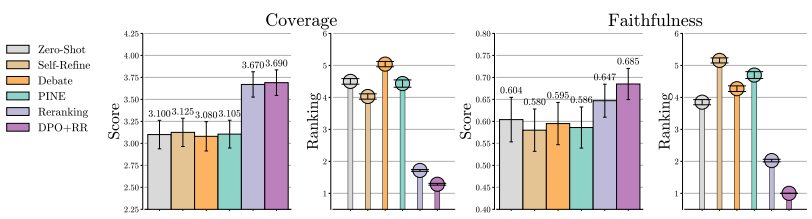
5.1 Automatic Evaluation

We observe that DPO+RR achieves the highest performance on both metrics, improving coverage and faithfulness scores by 0.590 and 0.081, corresponding to approximately 12% and 8% gains, respectively. Reranking is a strong baseline, outperforming all other methods by considerable margins, corroborating prior work on the benefits of re-ranking (e.g., (Horvitz et al., 2024)). In contrast, zero-shot inference, prompting methods, and PINE show minimal score differences. Although Self-Refine marginally improves coverage over zero-shot inference, all methods except reranking yield lower faithfulness scores.

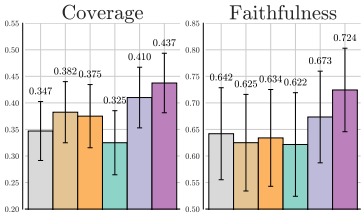
5.2 Human Evaluation

DPO+RR achieves the highest human evaluation scores (0.437 for coverage and 0.724 for faithfulness), with Reranking close behind (0.410 and 0.673, respectively). Prompting-based methods improve coverage over zero-shot inference (0.347 for coverage and 0.642 for faithfulness) but yield similar faithfulness scores. PINE shows no performance gains for either attribute.

Inter-Annotator Agreement (IAA). We measure IAA by counting the number of excerpts with non-trivial overlap between annotators. Formally, given excerpts from two annotators A and B (from a document or a summary), denoted as $E_A = \{e_1^A, e_2^A, \dots\}$ and $E_B = \{e_1^B, e_2^B, \dots\}$, we define a matching function $M(E_A, E_B)$ that



(a) Automatic evaluation results. Higher values indicate better performance for Score (Bars), while lower values are better for Ranking (Lolipops). Coverage scores range from 1 to 5, while faithfulness scores lie in the interval $[0, 1]$. DPO+RR achieves the highest scores and best average rank, followed by Reranking. Other methods show similar performance in both coverage and faithfulness.



(b) Human evaluation results. Higher is better for both attributes. DPO+RR achieves the best performance for both attributes, followed by Reranking. Scores lie in $[0, 1]$ for both attributes.

Figure 3: Automatic (left) and human (right) evaluation results. For clarity, note that y-axes do not begin at 0 in score plots. Reranking-based methods perform best across both evaluation regimes, with DPO+RR achieving the highest overall performance in both coverage and faithfulness. Error bars represent 95% confidence intervals (CI).

Method	$\lvert K_D \cap K_S \rvert$	$\lvert K_D \setminus K_S \rvert$	$\lvert K_S \setminus K_D \rvert$
Zero-Shot	1.338 ± 0.894	3.059 ± 1.254	0.765 ± 0.855
Self-Refine	1.412 ± 1.097	2.912 ± 1.288	0.794 ± 0.729
Debate	1.368 ± 0.847	2.971 ± 1.291	0.735 ± 0.790
PINE	1.206 ± 0.854	3.235 ± 1.350	0.706 ± 0.799
Reranking	1.544 ± 0.916	2.882 ± 1.320	0.735 ± 0.828
DPO+RR	1.721 ± 0.889	2.500 ± 1.080	0.618 ± 0.739

Table 5: Statistics for key point inclusion for each method with standard deviation. $|K_D \cap K_S|$, $|K_D \setminus K_S|$, and $|K_S \setminus K_D|$ denote the average number of key points included, omitted, and hallucinated, respectively.

counts the number of strings in E_A matched to at most one string in E_B. We then compute $R(A \mid B) = |M(E_A, E_B)|/|E_A|$ and $R(B \mid A) = |M(E_A, E_B)|/|E_B|$, and take their average to obtain the overall annotator overlap $R(\cdot \mid \cdot)$. To assess overlap, we provide overlapping annotations to pairs of annotators across five documents and evaluate agreement for both document-level and summary-level key point extraction. Additionally, we establish a random baseline for annotator overlap by sampling highlight counts and lengths for documents and summaries that match the observed mean and variance in the real annotations. Further details are provided in §C.1.

Results are presented in Table 4. Observe that annotators exhibit substantial overlap in both document- and summary-level annotations that considerably exceed the random baseline. We also see higher agreement for summaries than for documents, which we attribute to summaries being more concise and explicitly including key points.

Overall, our results show that while prompting-based and attention modification methods offer little improvement over zero-shot prompting, reranking-based methods significantly improves coverage and faithfulness. In particular, employing DPO-based training further boosts faithfulness, even when using self-generated synthetic data.

Method	**Novel 4-gram (↑)**	**EF Density (↓)**
Zero-Shot	0.930 ± 0.104	1.815 ± 1.614
Self-Refine	0.946 ± 0.094	1.470 ± 1.307
Debate	0.954 ± 0.088	1.571 ± 1.162
PINE	0.848 ± 0.074	3.340 ± 4.801
Reranking	0.949 ± 0.217	1.445 ± 0.914
DPO+RR	0.953 ± 0.079	1.415 ± 1.039

Table 6: Abstractiveness statistics for each method, measured by novel n-gram ratios and extractive fragment density. Arrows indicate higher abstractiveness.

6 Analysis

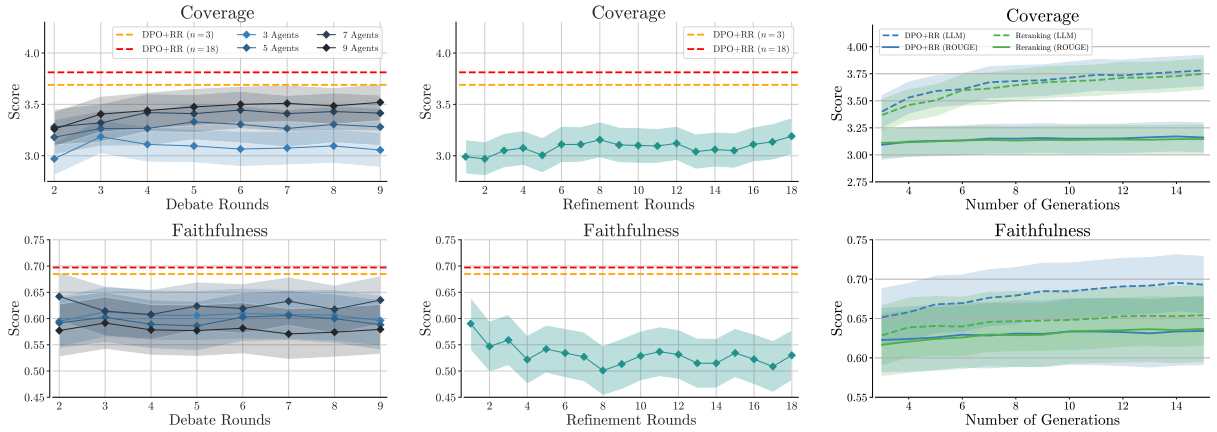
6.1 Summary Characteristics

Here, we examine the summaries generated by each method and assess their key point inclusion patterns, abstractiveness, and length.

Key Point Inclusion. Beyond coverage and faithfulness, we evaluate how each method incorporates key points. For an article D with key points K_D and a summary S with key points K_S, we compute the average number of key points included ($|K_D \cap K_S|$), omitted ($|K_D \setminus K_S|$), and hallucinated ($|K_S \setminus K_D|$).

Results are shown in Table 5. We observe that DPO+RR includes more relevant key points while minimizing hallucinations and omissions compared to other methods. In contrast, PINE is more conservative, reducing hallucinations but omitting more key points. Self-Refine retains additional key points yet introduces more hallucinations, while Debate shows only slight improvements over the zero-shot baseline.

Summary Abstractiveness. We assess abstractiveness using two metrics: (1) *Novel n-gram ratios* (See et al., 2017), which measure the proportion of n-grams in the summary absent from the source (with $n = 4$), and (2) *Extractive fragment density*



(a) Debate performance across varying agent counts and debate rounds.

(b) Self-Refine performance across varying refinement rounds.

(c) Comparison of LLM- and ROUGE-based proxies for reranking methods.

Figure 4: Ablation study results. Figures 4a and 4b show that both prompting-based methods consistently underperform compared to reranking-based methods across all resource settings. Figure 4c shows that using a ROUGE-based proxy metric yields worse performance than LLM-based proxy metrics.

(Grusky et al., 2018), which quantifies the continuity of extracted spans. Higher novel n-gram ratios and lower extractive fragment density indicate greater abstractiveness. We include additional results for analysis in §C.2.

Table 6 shows our results. Notably, PINE exhibits lower novel n-gram ratios and higher extractive fragment density than other methods, indicating that PINE favors more extractive summaries. In contrast, DPO+RR yields higher abstractiveness than zero-shot inference while also improving faithfulness (cf. §5). This suggests that DPO+RR not only encourages extraction of source content but also generates summaries with more novel tokens.

6.2 Ablation Studies

We conduct ablation studies to determine whether prompting-based methods outperform Reranking and DPO+RR under more resourceful generation settings, as measured by automated metrics. See Figure 4 for all results.

Debate: Agents and Rounds. We vary the number of rounds $n \in \{2, 3, \ldots, 9\}$ and agents $m \in \{3, 5, 7, 9\}$ in Multi-Agent Debate. For reference, we report results for DPO+RR in two settings: generating 3 (base setting) and generating 18 summaries (approximate upper bound).

From Figure 4a, we observe that increasing the number of agents improves coverage but not faithfulness. With $m = 9$ agents, Debate slightly outperforms DPO+RR with 3 reranked generations for $n \geq 4$ in coverage but falls short of DPO+RR with 18 generations. For faithfulness, Debate remains consistently below DPO+RR in all settings.

Self-Refine: Refinement Rounds. We evaluate Self-Refine over various numbers of refinement rounds ($n \in \{2, 3, \ldots, 18\}$). Results are shown in Figure 4b. We observe that coverage improves with more rounds, whereas faithfulness does not. Nevertheless, Self-Refine underperforms DPO+RR in both settings across all rounds.

Reranking: ROUGE as Proxy Metric. To assess the effectiveness of LLM-based proxy metrics, we compare Reranking and DPO+RR with variants that use ROUGE as the proxy. Following the training procedure in §4, we replace the original proxy with the average ROUGE_n score (for $n \in \{1, 2, L\}$) computed across both precision and recall. As shown in Figure 4c, the ROUGE-based variant underperforms across all settings.

7 Conclusion

In this paper, we identify reliable evaluation metrics for measuring perspective summary quality and investigate LLM-based methods for generating improved summaries beyond zero-shot inference. We construct a test dataset using human annotations to benchmark existing summarization metrics for coverage and faithfulness. We find that traditional metrics such as ROUGE and BERTSCORE underperform, while language model–based metrics such as ALIGNSCORE and prompting-based scoring serve as strong evaluators. Using these metrics, we show that reranking-based methods outperform prompting frameworks and significantly improve performance over zero-shot inference. Moreover, preference tuning with self-generated, reranking-labeled data further boosts performance, particu-

larly in terms of faithfulness. We recommend that future work examine the transferability of our findings to domains beyond political perspectives and whether similar improvements can be achieved in other multi-document summarization tasks.

Limitations

We acknowledge two limitations in our work. First, we focus on evaluating existing summarization metrics commonly used in the literature and benchmark those applied to perspective summarization. As we show that existing metrics achieve satisfactory accuracy for evaluating perspective summaries, we do not investigate the development of a novel metric tailored specifically for measuring coverage and faithfulness in this setting. We leave this as a promising direction for future work. Second, we primarily investigated methods for perspective summary generation that do not rely on human-labeled training data, given the infeasibility of collecting such data. Although our experiments with preference tuning using synthetically generated data show performance improvements, future studies should examine the benefits of human-curated training data.

Ethical Considerations

In this paper, we focus on metrics to accurately measure the unbiasedness of perspective summaries through the attributes of coverage and faithfulness, and we show that certain methods yield higher performance on these attributes. Our work aims to ensure fair representation and reduce hallucinations in opinion-based summarization. While it is unclear whether these findings could be misused to generate more biased summaries, we acknowledge that such risks are not negligible.

Acknowledgements

This work was supported in part by the Knight First Amendment Institute at Columbia University, National Science Foundation Graduate Research Fellowship DGE-2036197, the Columbia University Provost Diversity Fellowship, and the Columbia School of Engineering and Applied Sciences Presidential Fellowship. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Knight First Amendment Institute or National Science Foundation. We thank the anonymous reviewers for providing feedback on an earlier draft of the work.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, and Christopher Summerfield. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. In *Advances in Neural Information Processing Systems*, volume 35, pages 38176–38189. Curran Associates, Inc.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Xiuying Chen, Mingzhe Li, Xin Gao, and Xiangliang Zhang. 2022. Towards improving faithfulness in abstractive summarization. In *Advances in Neural Information Processing Systems*, volume 35, pages 24516–24528. Curran Associates, Inc.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. Association for Computational Linguistics.

Anshuman Chhabra, Hadi Askari, and Prasant Mohapatra. 2024. Revisiting zero-shot abstractive summarization in the era of large language models from the perspective of position bias. *Preprint*, arXiv:2401.01989.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *Preprint*, arXiv:2307.08691.

Nicholas Deas and Kathleen McKeown. 2025. Summarization of opinionated political documents with varied perspectives. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8088–8108, Abu Dhabi, UAE. Association for Computational Linguistics.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Markus Dreyer, Mengwen Liu, Feng Nan, Sandeep Atluri, and Sujith Ravi. 2023. Evaluating the tradeoff between abstractiveness and factuality in abstractive summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2089–2105, Dubrovnik, Croatia. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171, Miami, Florida, USA. Association for Computational Linguistics.

Sian Gooding and Hassan Mansoor. 2023. The impact of preference agreement in reinforcement learning from human feedback: A case study in summarization. *Preprint*, arXiv:2311.04919.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3. *Preprint*, arXiv:2209.12356.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Zachary Horvitz, Ajay Patel, Kanishk Singh, Chris Callison-Burch, Kathleen McKeown, and Zhou Yu. 2024. TinyStyler: Efficient few-shot text style transfer with authorship embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13376–13390, Miami, Florida, USA. Association for Computational Linguistics.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14982–14995, Bangkok, Thailand. Association for Computational Linguistics.

Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Tianfu Wang, Zhengyu Chen, Nicholas Jing Yuan, Jianxun Lian, Kaize Ding, and Hui Xiong. 2024. Explaining length bias in llm-based preference evaluations. *Preprint*, arXiv:2407.01085.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Shengyi Huang, Michael Noukhovitch, Arian Hosseini, Kashif Rasul, Weixun Wang, and Lewis Tunstall. 2024. The n+ implementation details of RLHF with PPO: A case study on TL;DR summarization. In *First Conference on Language Modeling*.

Hayate Iso, Xiaolan Wang, Stefanos Angelidis, and Yoshihiko Suhara. 2022. Comparative opinion summarization via collaborative decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3307–3324, Dublin, Ireland. Association for Computational Linguistics.

Athul Paul Jacob, Yikang Shen, Gabriele Farina, and Jacob Andreas. 2024. The consensus game: Language model generation via equilibrium search. In *The Twelfth International Conference on Learning Representations*.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8487–8495, Toronto, Canada. Association for Computational Linguistics.

Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. 2019. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. *Preprint*, arXiv:1908.11723.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1410–1421, Dublin, Ireland. Association for Computational Linguistics.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2024. RLAIF: Scaling reinforcement learning from human feedback with AI feedback.

Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022a. NeuS: Neutral multi-news summarization for mitigating framing bias. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3131–3148, Seattle, United States. Association for Computational Linguistics.

Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022b. NeuS: Neutral multi-news summarization for mitigating framing bias. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3131–3148, Seattle, United States. Association for Computational Linguistics.

Yuanyuan Lei, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Ruihong Huang, and Dong Yu. 2024. Polarity calibration for opinion summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5211–5224, Mexico City, Mexico. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yuhan Liu, Shangbin Feng, Xiaochuang Han, Vidhisha Balachandran, Chan Young Park, Sachin Kumar, and Yulia Tsvetkov. 2024b. P^{3}Sum: Preserving author's perspective in news summarization with diffusion language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2154–2173, Mexico City, Mexico. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder,

Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback. *Preprint*, arXiv:2112.09332.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089, Bangkok, Thailand. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. Parallel context windows for large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6383–6402, Toronto, Canada. Association for Computational Linguistics.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard 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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6252–6272, Toronto, Canada. Association for Computational Linguistics.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2024. Branch-solve-merge improves large language model evaluation and generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8352–8370, Mexico City, Mexico. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Noah Siegel, Oana-Maria Camburu, Nicolas Heess, and Maria Perez-Ortiz. 2024. The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models. In *Pro-*

ceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 530–546, Bangkok, Thailand. Association for Computational Linguistics.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained summarization evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2222, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. 2024. Eliminating position bias of language models: A mechanistic approach. *Preprint*, arXiv:2407.01100.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575, Singapore. Association for Computational Linguistics.

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2024. Less is more for long document summary evaluation by LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 330–343, St. Julian's, Malta. Association for Computational Linguistics.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024. LLMRefine: Pinpointing and refining large language models via fine-grained actionable feedback. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1429–1445, Mexico City, Mexico. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Shiyue Zhang and Mohit Bansal. 2021. Finding a balanced degree of automation for summary evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6617–6632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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

Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming

```
Given texts from both Left-leaning and Right-leaning
perspectives, summarize only the Left-leaning perspective in
one sentence, starting with 'The Left '. ONLY RETURN THE
SUMMARY AND NOTHING ELSE.

Left:
(left-perspective article)

Right:
(right-perspective article)
```

Figure 5: Prompt instruction for zero-shot inference when generating summaries from the left-leaning perspective.

Lu. 2024b. Self-contrast: Better reflection through inconsistent solving perspectives. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3602–3622, Bangkok, Thailand. Association for Computational Linguistics.

Yusen Zhang, Nan Zhang, Yixin Liu, Alexander Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, Kathleen McKeown, and Rui Zhang. 2024c. Fair abstractive summarization of diverse perspectives. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3404–3426, Mexico City, Mexico. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Supplementary Details

A.1 Experimental Setup

Unless otherwise specified, all inference is run using the `transformers` library with 16-bit floating point precision using FLASH ATTENTION 2 (Dao, 2023). We train the DPO-based models on four NVIDIA A100-SXM4-80GB GPUs, with each model requiring approximately 2∼3 days of training. For other experiments, including inference and evaluation using small-scale language models, we use a variable number of NVIDIA A100-SXM4-80GB GPUs depending on the model size. For completeness, we also provide the prompt instruction for the zero-shot inference setting in Figure 5.

DPO Training. We train our DPO-based models using 4 batches and the default hyperparameter settings from the `DPOConfig` class in the `transformers` library. This corresponds to using an (adaptive) learning rate of 5.0×10^{-5}, a β value of 0.1, and reverse KL divergence for f-divergence regularization.

PINE. We use the codebase available at `github.com/wzq016/PINE.git` for the PINE implementation. In our setup, the input is formatted as

$$[\texttt{INS} \mid d^{(1)}_{t,\theta_1} \mid d^{(2)}_{t,\theta_1} \mid \ldots \mid d^{(1)}_{t,\theta_2} \mid d^{(2)}_{t,\theta_2} \mid \ldots \mid \texttt{EOS}],$$

where INS is the prompt instruction, $D_{t,\theta_1} = d^{(1)}_{t,\theta_1} \mid d^{(2)}_{t,\theta_1} \mid \ldots$ represents the left-leaning source documents, $D_{t,\theta_2} = d^{(1)}_{t,\theta_2} \mid d^{(2)}_{t,\theta_2} \mid \ldots$ the right-leaning source documents, and EOS is the end-of-sequence token. PINE reformats the input by designating a target segment (e.g., D_{t,θ_1} when the target perspective is the left-leaning view) to ensure that all segments are attended to uniformly, regardless of their original positions.

Dataset. We use the POLISUM dataset (Deas and McKeown, 2025) as our primary testbed for perspective summarization. We remove duplicates from the dataset and obtain 1816 article pairs (left- and right-leaning). We split the data into 1716 article pairs for training DPO+RR and 100 article pairs for testing. Although most methods we investigate do not rely on training, we maintain a strict separation between train and test sets to avoid inflating DPO+RR performance.

A.2 Ranking Methods

To obtain accurate ranking results using automated metrics, we fit a Bradley-Terry model (Bradley and Terry, 1952) to the per-method scores for each bootstrap resample of the test set and derive confidence intervals for each method's ability estimate.

Specifically, let there be M methods with latent abilities $\{\theta_1, \theta_2, \ldots, \theta_M\}$. For each pair of methods (i, j), the model posits that the probability of i "winning" over j in a pairwise comparison is given by a logistic function:

$$\Pr[i \text{ beats } j] = \frac{1}{1 + \exp\big(-(\theta_i - \theta_j)/\sigma\big)}, \quad (3)$$

where $\sigma > 0$ is a noise or scale parameter. We treat method i as having beaten method j if i's

aggregated raw score exceeds j's, resolving exact ties randomly.

We estimate the abilities by maximizing the log-likelihood of all observed pairwise outcomes:

$$\ell(\theta_1, \ldots, \theta_M) = \sum_{(i,j)\in\mathcal{D}} \Big[\mathbb{1}[i \text{ beats } j] \log \Pr[i \text{ beats } j] + \mathbb{1}[j \text{ beats } i] \log\big(1 - \Pr[i \text{ beats } j]\big) \Big],$$

where $\mathcal{D}$ denotes all pairwise comparisons from the current (re)sample, and $\mathbb{1}(\cdot)$ is an indicator function. We perform this fitting procedure via numerical optimization (L-BFGS). To account for variability, we employ bootstrap resampling over the test set: each resample draws the test documents (with replacement), averages each method's raw scores within that resample, and re-fits the Bradley-Terry model to generate a new set of abilities $\{\theta_m\}$. We repeat for $B = 500$ iterations and obtain an empirical distribution of ability estimates for each method. We then rank methods by their mean estimated ability across all bootstrap replicates and derive 95% confidence intervals from the resulting bootstrap distributions.

B Metric Evaluation

Here, we provide supplementary content for benchmarking evaluation metrics for measuring coverage and faithfulness.

B.1 Metric Configurations

For ROUGE, we use the `rouge-score` Python library. For BERTSCORE and BLEURT, we use the `deberta-large-xnli` and `BLEURT-20-D6` checkpoints respectively, due to their higher correlations with human judgments. For ALIGNSCORE, we employ the `AlignScore-large` checkpoint from Zha et al. (2023). For SUMMAC, we use the `tals/albert-xlarge-vitaminc-mnli` model, which is the default setting for the SUMMAC evaluation metric.

B.2 Prompt Instructions

Prompt-based Scoring: LLM-Coverage and LLM-Faithfulness. We provide the full prompt instructions for both LLM-Coverage and LLM-Faithfulness in Figures 6b and 6a, respectively. While we experiment with prompt variations such as using binary and ternary scoring and removing step-by-step procedures, these modifications result in lower performance. We omit these alternate prompts for brevity.

B.3 Backbone Scoring Evaluation

Table 7 presents additional results for various LLM backbones. Notably, prompt-based scoring generally performs better on coverage than on faithfulness. In particular, `Mistral-7B-Instruct-v0.3` and `Qwen2.5-14B-Instruct` exhibit the best performance across both metrics. Based on these results, we use `Mistral-7B-Instruct-v0.3` as the evaluator and `Qwen2.5-14B-Instruct` as the proxy metric. For coverage, larger model sizes weakly correlate with higher performance, though the gains are marginal. To keep inference time reasonable, we therefore use smaller-scale models that still exhibit good performance. For faithfulness, the `Llama` models consistently underperform compared to other backbones on both correlation and ranking. However, as all backbones perform close to the random baseline on winrate, we avoid using prompt-based scoring for faithfulness.

B.4 Paraphrasing Excerpts to Key Points

As mentioned in §3.1, we use an LLM to paraphrase highlighted excerpts into key points. We also employ an LLM to generate adversarial key points $\overline{K}_{t,\theta}$ from the curated key points $K_{t,\theta}$, using the prompts provided in Figures 7a and 7b. We use `Qwen2.5-32B-Instruct` for both paraphrasing and key point generation.

C Benchmarking Methods

Here, we provide supplementary details on our evaluation procedure along with additional analysis on the generated summaries by each method.

C.1 Inter-Annotator Agreement

We first provide additional information on the matching function $M(\cdot,\cdot)$ in §5.2. For each element $s_i^A \in S_A$, the function finds the first unmatched element $s_j^B \in S_B$ that meets a matching condition. The first criterion is exact containment: if s_i^A is a substring of s_j^B or vice versa, they are considered a match. If no exact containment is found, we compute the longest common subsequence (LCS) between s_i^A and s_j^B. If the LCS length divided by the length of the shorter string exceeds a predefined threshold τ, we consider them a match. Each element in S_A is matched to at most one element in S_B, and vice versa, and is removed from further matching once paired. By default, we set $\tau = 0.5$.

Random Baseline for IAA. We simulate random highlight selection as follows. First, we compute the mean and variance of the number of highlights

Metric	Model	Coverage		Faithfulness	
		Corr. (ρ_s)	Winrate	Corr. (ρ_s)	Winrate
LLM-Coverage	`Mistral-7B-Instruct-v0.3`	0.707^{***}	0.739 ± 0.047	0.393^{***}	0.431 ± 0.115
	`Mixtral-8x7B-Instruct-v0.1`	0.720^{***}	0.771 ± 0.050	0.335^{***}	0.475 ± 0.087
	`Llama-3.1-8B-Instruct`	0.606^{***}	0.648 ± 0.051	0.188^{***}	0.313 ± 0.093
	`Llama-3.3-70B-Instruct`	0.724^{***}	0.753 ± 0.058	0.280^{***}	0.415 ± 0.100
	`Qwen2.5-7B-Instruct`	0.650^{***}	0.624 ± 0.081	0.343^{***}	0.349 ± 0.106
	`Qwen2.5-14B-Instruct`	0.732^{***}	0.749 ± 0.049	0.334^{***}	0.380 ± 0.081
	`Qwen2.5-32B-Instruct`	0.721^{***}	0.709 ± 0.060	0.302^{***}	0.343 ± 0.097
LLM-Faithfulness	`Mistral-7B-Instruct-v0.3`	0.504^{***}	0.494 ± 0.061	0.646^{***}	0.498 ± 0.113
	`Mistral-Large-Instruct-2411`	0.722^{***}	0.688 ± 0.076	0.579^{***}	0.479 ± 0.108
	`Llama-3.1-8B-Instruct`	0.577^{***}	0.303 ± 0.074	0.439^{***}	0.188 ± 0.079
	`Llama-3.3-70B-Instruct`	0.558^{***}	0.283 ± 0.080	0.735^{***}	0.343 ± 0.112
	`Qwen2.5-7B-Instruct`	0.589^{***}	0.536 ± 0.064	0.644^{***}	0.503 ± 0.087
	`Qwen2.5-14B-Instruct`	0.702^{***}	0.671 ± 0.099	0.616^{***}	0.519 ± 0.086
	`Qwen2.5-32B-Instruct`	0.712^{***}	0.675 ± 0.063	0.670^{***}	0.590 ± 0.096

Table 7: Comparison of Spearman rank correlation (**Corr.** (ρ_s)) and Winrate (**Winr.**) across different backbone models. LLM-Coverage exhibits moderate to high correlation and winrate across all backbones, while `Mistral-7B-Instruct-v0.3` and `Qwen2.5-14B-Instruct` achieve the best performance for faithfulness.

Method	Summary Length	EF Coverage	Comp. Ratio
Zero-Shot	40.77 ± 6.212	0.719 ± 0.113	14.958 ± 4.165
Self-Refine	43.94 ± 10.73	0.692 ± 0.107	15.097 ± 5.774
Debate	41.50 ± 11.609	0.692 ± 0.103	16.192 ± 4.903
PINE	38.17 ± 7.401	0.776 ± 0.125	19.589 ± 20.23
Reranking	37.13 ± 13.856	0.651 ± 0.157	14.166 ± 4.181
DPO+RR	42.94 ± 8.245	0.661 ± 0.149	14.391 ± 4.421

Table 8: Supplementary statistics for each method, measured by summary length, extractive fragment coverage, and compression ratio.

and their lengths separately for documents and summaries. Using these statistics, we sample the number of highlights and the length of each highlight from a normal distribution $\mathcal{N}(\cdot, \cdot)$ for each document or summary. We repeat this process independently twice and compute the overlap between the two instances as described in §5.2. This procedure simulates a non-trivial, semi-realistic random highlighting of excerpts in documents and summaries.

C.2 Supplementary Analysis

In addition to coverage-density plots, we report additional results for summary lengths, *extractive fragment coverage* (quantifying the extent of copying from the source), and *compression ratios* (Grusky et al., 2018) (assessing summary length relative to the source document).

Table 8 shows our results. Consistent with findings in §6.1, PINE exhibits lower abstractiveness compared to other methods. Moreover, both compression ratios and summary lengths indicate that PINE tends to generate shorter summaries relative to other methods.

Extractive Fragment Plots. We also include Coverage-Density plots for both the source and opposing perspective documents in Figure 8. Overall, we observe similar coverage-density structures for all by PINE, which exhibits a wider spectrum of coverage and density. This indicates that the abstractiveness of PINE exhibits high variance, whereas for other methods the abstractiveness is relatively stable. Furthermore, we also see that the coverage-density plots for the opposing side is slightly lower than for the target source articles.

C.3 Additional Example Summaries

In Table 9, we provide additional sampled examples for the summaries generated by each method.

D Annotation Information

D.1 Annotation Details

For both annotation procedures, annotators consented to having their annotated excerpts used for research purposes (cf. Figures 9 and 11). All human evaluations in this work were conducted under an approved IRB protocol.

Test Set for Metric Evaluation. We recruited 5 graduate annotators, each assigned 10 documents for excerpt highlighting. Each annotator received $15 as compensation.

Summary Evaluation. Annotators were recruited from undergraduate Political Science students with self-reported knowledge of conservative and liberal beliefs to ensure the required expertise to judge summary perspectives. Four annotators

```
You are an evaluator. Your task is to determine how
well a generated summary captures all of the main
arguments from a source article. This is a measure of "
coverage," which does not necessarily address factual
accuracy (faithfulness) but focuses on completeness of
content.

The scale for coverage is:
1. No Coverage: The summary does not include any of the
 main arguments from the article.
2. Low Coverage: The summary includes only a few of the
 main arguments from the article, omitting most.
3. Medium Coverage: The summary contains around half of
 the article's main arguments.
4. High Coverage: The summary contains most of the main
 arguments from the article, missing only a few.
5. Perfect Coverage: The summary includes all major
points and arguments mentioned in the article, leaving
out nothing important.

Follow these steps carefully:

1. **Read the Source Article**: Examine the text
provided in the article. Identify all major points,
arguments, or facts it contains.
2. **Read the Summary**: Look at the text in the
summary. List each argument or point the summary
includes.
3. **Compare for Completeness**:
- Check if each major point from the source article is
present in the summary.
- Count how many major points are covered versus how
many are omitted.
4. **Determine the Score**:
- Assign a score from 1 (no coverage) to 5 (perfect
coverage), based on how many main arguments are
included in the summary relative to the source.
5. **Output Instructions**:
- Output only the final numeric score (1, 2, 3, 4, or 5)
 without any explanation or additional text.

---

# Source Article:

(article)

# Generated Summary:

(summary)

# Final Coverage Score (1~5 only):
```

(a) Full prompt instructions for LLM-Coverage.

```
You are an evaluator. Your task is to analyze how faithfully a
generated summary represents the information found in the source
article. Faithfulness here means the absence of factual errors---
i.e., any claims in the summary must be either directly stated,
heavily implied, or logically entailed by the source article.

The scale for faithfulness is:
1. Unfaithful: The summary is almost entirely incorrect or
unrelated to the source.
2. Mostly Unfaithful: The summary includes only a few relevant
arguments or correct details but is largely incorrect or missing.
3. Neutral: The summary has about half of the important points
correct, but also includes notable inaccuracies or omissions.
4. Mostly Faithful: The summary reflects most of the article's
arguments accurately, with only minor errors or omissions.
5. Perfectly Faithful: The summary includes all of the article's
main arguments accurately and does not introduce any
contradictory or unsupported claims.

Follow these steps carefully:

1. **Read the Source Article**: Examine the text provided in the
article. Identify the main points, arguments, or facts it
contains.
2. **Read the Summary**: Look at the text in the summary. Itemize
 or note each claim or statement made in the summary.
3. **Compare for Accuracy**:
- Check if each claim in the summary is explicitly or logically
supported by the source.
- Mark any claim that appears to be contradicting the source or
not found in the source.
- Check if the summary omits major arguments that are central to
the source.
4. **Determine the Score**:
- Assign a score from 1 (completely unfaithful) to 5 (perfectly
faithful), based on how many claims match (and do not contradict)
 the source article and whether key points are included.
5. **Output Instructions**:
- Output only the final numeric score (1, 2, 3, 4, or 5) without
any additional explanation or text.

---

# Source Article:

(article)

# Generated Summary:

(summary)

# Final Faithfulness Score (1~5 only):
```

(b) Full prompt instructions for LLM-Faithfulness.

Figure 6: Complete prompt instructions for both attributes in prompting-based scoring. The model is provided with descriptions of each score value and a step-by-step procedure for evaluating the summary based on the article.

participated—three annotated 20 documents each and one annotated 15. To measure inter-annotator agreement, overlapping annotations were collected for 10 documents, with each document annotated by two annotators. This process yielded a total of 75 document-summary annotations per method. Annotators were compensated at $22.50 per hour and spent approximately 15 ± 2.5 minutes per page.

D.2 Annotation Interfaces

We provide the annotation interfaces for the human studies described in §3.1 and §4.1 in Figures 9 and 10 (for metric evaluation) and Figures 11 and 12 (for summary evaluation). Both interfaces were built using the `streamlit` Python library and hosted on the Streamlit Community Cloud platform. Annotator results were stored using Amazon Web Services (AWS) Simple Storage Service (S3).

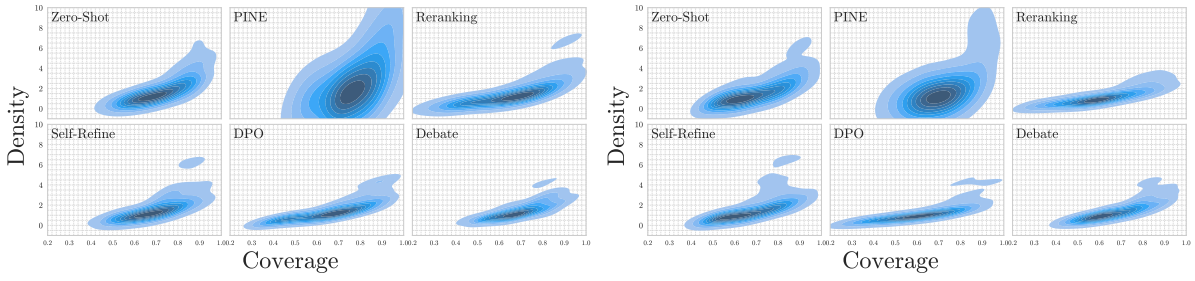
```
[TASK]
You are given an article that makes an
argument related to the provided topic. An
excerpt from the document highlights the main
key argument that the author of the article is
 trying to assert. Please write a concise,
short, one-sentence paraphrase (as short as
possible) that reflects the argument implied
or present in the provided excerpt. **Your
paraphrase should begin with "The article
argues"**.

---

Topic: (topic)

Article: (article)

Excerpt: (excerpt)

---

One-Line Argument Summary starting with "The
article argues":
```

(a) Full prompt instructions for paraphrasing highlighted excerpts to key points.

```
[TASK]
You are given one main argument from a political news article (either left-
leaning or right-leaning). **Rewrite the argument so that the argument is
completely reversed or semantically opposite.** If the original argument
supports or praises a policy/idea/group, the reversed version should
criticize or oppose it, and vice versa. Only return the reversed argument
itself, with no extra commentary or analysis.

[EXAMPLES]
1.
ORIGINAL: The article argues that stricter immigration laws help protect
domestic jobs and strengthen national identity.
REVERSED: The article argues that relaxed immigration laws create more job
opportunities and enhance cultural diversity.

2.
ORIGINAL: The article insists that climate change is primarily caused by
human activity and demands immediate government intervention.
REVERSED: The article insists that human activity has minimal impact on
climate change and calls for minimal government involvement.

[INFERENCE]
ORIGINAL: (original key point)
REVERSED:
```

(b) Prompt for generating adversarial key points $\overline{K}_{t,\theta}$ from the curated key points $K_{t,\theta}$.

Figure 7: Prompts used for portions of the procedure for curating the benchmarking test set for metrics.



(a) Coverage-Density plots for the source document.

(b) Coverage-Density plots for the opposing document.

Figure 8: Coverage-density plots for source and opposing documents. PINE exhibits higher variance in coverage, while other methods follow a similar structure.

Topic	Student Loans
Zero-Shot	The Right argues that Biden's student loan relief plan disproportionately benefits those making less than $75,000 per year and that it is a Band-Aid solution that fails to address the root cause of the issue, which is the increasing cost of higher education due to colleges and universities jacking up their prices.
Self-Refine	The Right argues that President Biden's student loan relief plan benefits low-income individuals, excludes the truly rich, and is not a giveaway to the elite, while the root cause of the student debt crisis lies in government subsidies to colleges, which have led to inflated tuition prices.
Debate	The Right argues that President Biden's student loan relief plan is a reasonable measure to help lower-income individuals, who make up 90% of those benefiting, due to the government's subsidization of colleges and universities leading to inflated tuition costs.
PINE	The Right argues that the student debt relief plan is a Band-Aid that will collapse and that colleges should be expected to finance the education they provide.
Reranking	The Right suggests that Biden's student debt relief plan is a Band-Aid solution that primarily benefits those making less than $75,000 per year and does not address the root cause of rising college costs, which can be mitigated by requiring institutions to cut costs and improve their financial efficiency.
DPO+RR	The Right views Biden's student-loan debt forgiveness plan as a Band-Aid solution that unfairly excludes the truly rich and argues that colleges and universities should be held accountable for their role in inflating prices and should provide debt-free education to students.

Topic	Mitt Romney
Zero-Shot	The Left sees the failure of Republicans to convict Trump as a decision made "functionally at gunpoint" due to the perceived threat of violence from the extremist right, rather than through conviction or political calculation.
Self-Refine	The Left views the failure of Republicans to convict Trump as a result of the perceived violent threat from the extremist right, which swayed their decision-making due to fear for their families' safety, representing one of the most consequential decisions in American history.
Debate	The Left views Romney's decision to stay in the Republican Party as a failure to challenge the extremist right and uphold democratic values, having been swayed by perceived threats of violence rather than conviction or principle.
PINE	The Left believes that Mitt Romney's decision to leave the Republican Party is a testament to his commitment to principle and his willingness to stand up to the party's extremist elements.
Reranking	The Left portrays Romney as revealing a stark contrast between his authentic disdain for the extreme ideologies within his own party in private conversations, versus his careful and cautious language used in public statements, highlighting the profound impact of fear generated by far-right extremism on congressional decision-making processes, specifically in relation to Trump's impeachment trial and acquittal.
DPO+RR	The Left views Senator Romney's decision not to convict Donald Trump due to fears of potential retribution from extremists as indicative of the broader breakdown within the Republican party, illustrating its susceptibility to authoritarianism and inability to stand up against violence-based intimidation despite holding supposedly "right" convictions.

Topic	Democrats Sign Letter Opposing Pelosi
Zero-Shot	The Left emphasizes that Nancy Pelosi's experience, legislative skill, and ability to pass significant bills, such as the Affordable Care Act, make her the ideal choice to lead the House Democrats.
Self-Refine	The Left believes Nancy Pelosi's experience and effectiveness make her the best choice for Speaker, despite potential risks to LGBTQ rights and ordinary women's concerns.
Debate	The Left emphasizes the importance of retaining Nancy Pelosi as Speaker of the House due to her extensive experience, legislative skill, and ability to pass key progressive policies, with her being a crucial figure in landmark legislative successes and there not being a clear alternative.
PINE	The Left argues that Nancy Pelosi is a highly effective legislator who has consistently been in the top third of most liberal Democrats in the House, and that ousting her would be a blow to ordinary women and LGBTQ people whose concerns will lose priority status if the anti-Pelosi wing prevails.
Reranking	The Left believes that keeping Nancy Pelosi as Speaker due to her vast experience and proven ability to effectively navigate Congress towards Democratic achievements and maintain party unity is vital amidst any challenges faced or proposed replacements who do not offer comparable qualifications or stronger alliances, especially regarding marginalized groups interests.
DPO+RR	The Left emphasizes that retaining Nancy Pelosi as Speaker of The House is crucial due to her extensive experience, successful legislation passing track record, ability to effectively oversee the Democratic Party, and importance in maintaining progress on key issues like LGBTQ rights and access to affordable healthcare.

Topic	China and Coronavirus
Zero-Shot	The Right suggests that the United States should not completely decouple from China on public-health matters, as past collaboration has been effective, but rather demands institutional changes in China's domestic food safety policies and calls for greater economic independence from China.
Self-Refine	The Right believes that China's disastrous cover-up and propaganda efforts have been matched by the US's own bumbling response to the crisis.
Debate	The Right views China's actions as deceitful and propaganda-driven, but also acknowledges that both the US and China have mishandled the pandemic, with the US's decoupling from China on public health matters and the US's bumbling response to the pandemic exacerbating the crisis.
PINE	The Right argues that the United States should distance itself from China and prioritize economic independence as a permanent national security goal.
Reranking	The Right perspective emphasizes that the US and the world should be cautious of China's intentions and actions, particularly its propaganda efforts, and that the US should prioritize economic independence and not rely on China for aid or information, while also acknowledging that the US has made mistakes in its response to the pandemic.
DPO+RR	The Right emphasizes that while China failed catastrophically during the COVID-9 crisis through its aggressive disinformation campaign and inadequate initial response, the sudden shift towards dependency on China-led relief efforts actually highlights broader risks inherent in the interconnection between nations, necessitating reconsideration of international relationships and prioritization of economic sovereignty.

Table 9: Additional example summaries generated by each method.

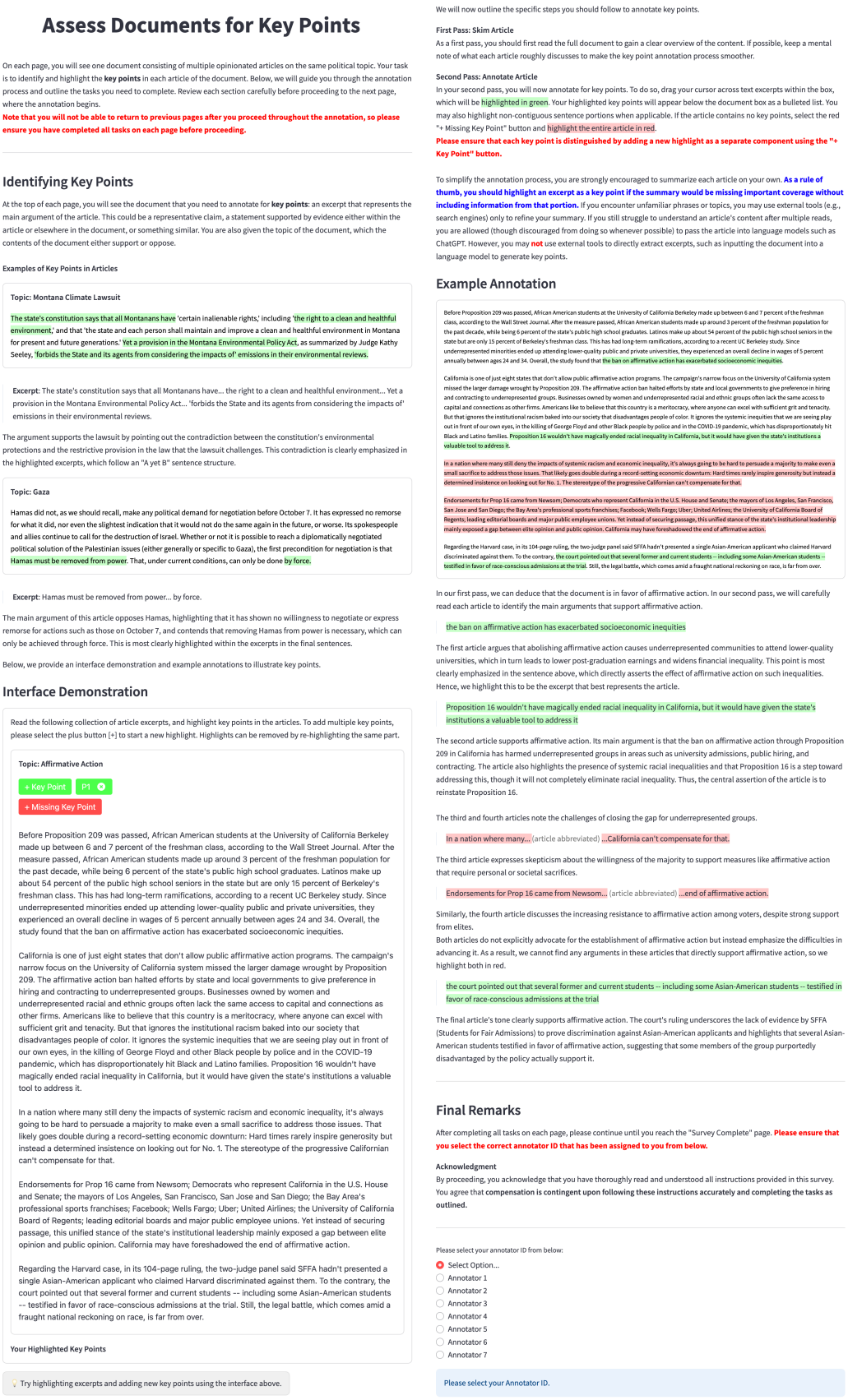


Figure 9: Introduction page for Annotation interface for annotating for article excerpts to evaluate metrics. Annotators are provided with definitions and an example annotated document.

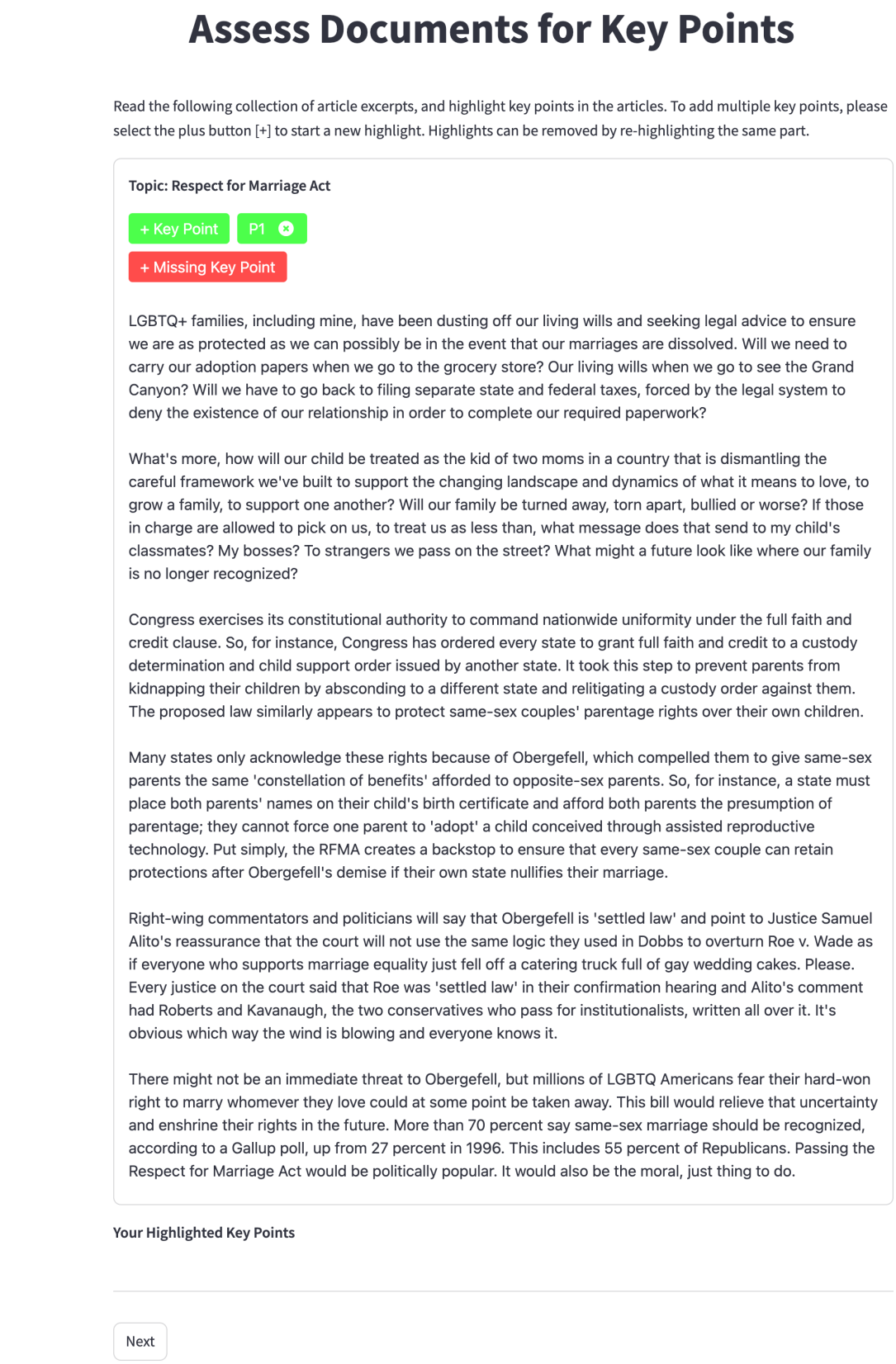


Figure 10: Example of annotation page for Annotation interface for annotating for article excerpts to evaluate metrics. Annotators are provided with an interface for highlighting sentences in the article.

Figure 11: Introduction page of annotation interface for annotating for document and summary excerpts for evaluating method-generated summaries.

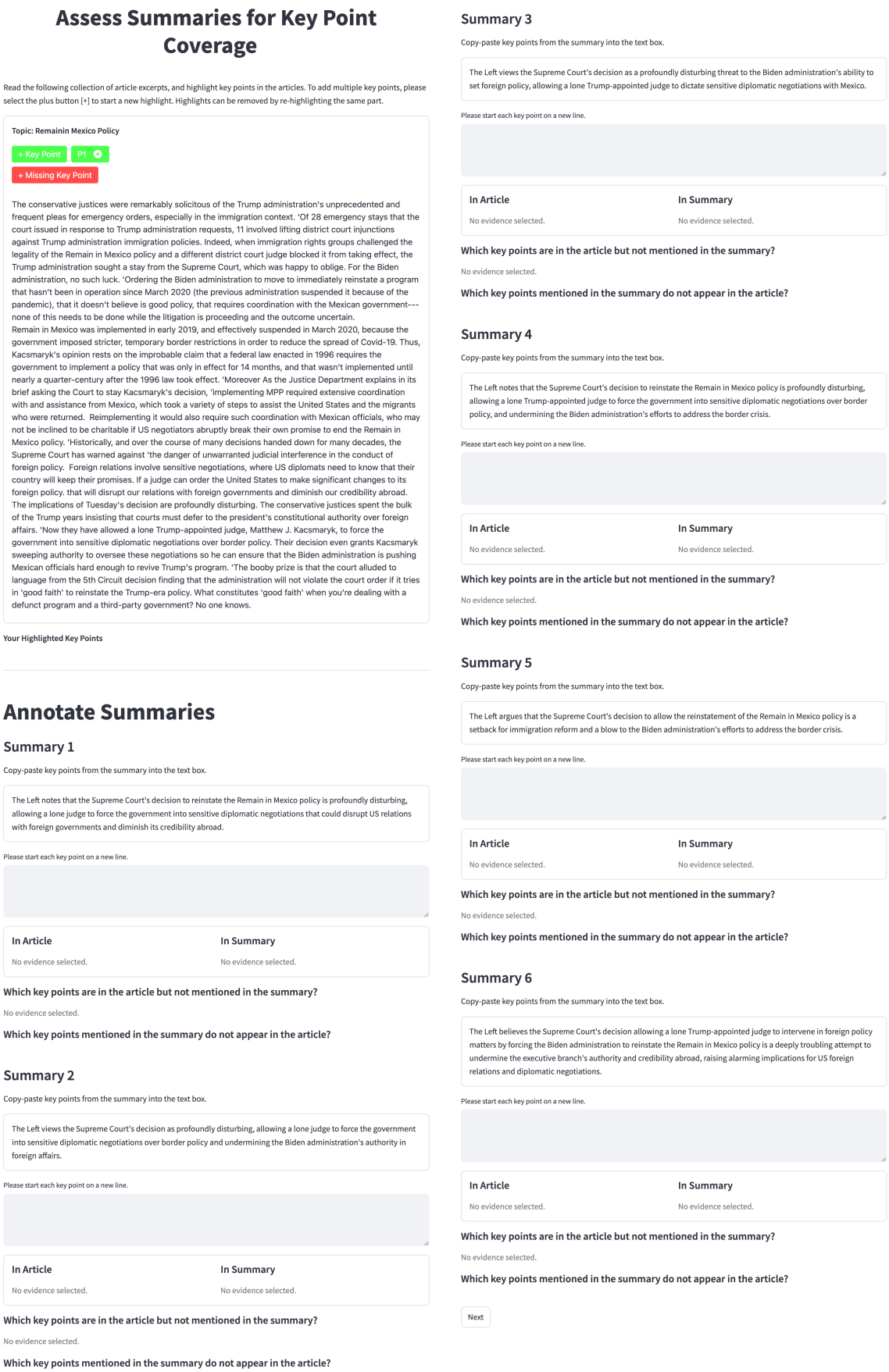


Figure 12: Example of annotation page for Annotation interface for document and summary excerpts for evaluating method-generated summaries.