

Quality of Rejections Matters: Preference Data Construction for Faithful Summarization

Jihwan Oh¹, Taewon Yun¹, Minkyu Che¹, Sundong Kim², Hwanjun Song^{1,3,*}

¹Korea Advanced Institute of Science and Technology (KAIST),

²Gwangju Institute of Science and Technology (GIST), ³Cluvion
{jh.oh, songhwanjun}@kaist.ac.kr

Abstract

Factual reliability is a key challenge in abstractive summarization, as LLMs continue to generate hallucinations. A widely adopted solution is preference optimization, training models to prefer faithful over unfaithful summaries, but prior work emphasized the quantity of rejections over their alignment quality. In this paper, we show that effective alignment arises when rejected summaries achieve high alignment potential, characterized by a small preference margin that keeps rejections non-trivial and a large factuality margin that enforces clear factual contrast. Through both theoretical analysis and controlled prompting, we show that three factors, hallucination level, summary length, and prompt complexity, critically shape alignment potential. Building on these insights, we propose SPICE, a simple prompting strategy that produces rejected summaries with strong alignment value. Across diverse model scales and alignment algorithms, SPICE achieves superior factuality without sacrificing coherence, relevance, or abstractiveness, outperforming existing rejection strategies.

1 Introduction

Recent advances in large language models (LLMs) have substantially advanced text summarization, producing outputs that are more coherent, fluent, and stylistically natural than ever before (Pu et al., 2023; Song et al., 2025). Nevertheless, a persistent and fundamental challenge is *hallucination*, where generated summaries contain information inconsistent with the source text (Maynez et al., 2020; Song et al., 2024; Lee et al., 2024b). Even high-performing LLMs continue to exhibit non-trivial rates of hallucination (Song et al., 2024), underscoring that fluency does not guarantee factuality. This reflects an alignment gap, a divergence between model outputs and source fidelity that undermines summary reliability.

A common remedy is preference optimization, where models learn to favor *chosen* (factually consistent) summaries over *rejected* (inconsistent) ones (Ouyang et al., 2022). Yet, collecting human feedback at scale is prohibitively costly, as humans must verify consistency against long source documents (Mishra et al., 2024; Song et al., 2025). To remedy this, recent studies instead generate preference pairs automatically. *Model-based* methods (Choi et al., 2024; Song et al., 2025) sample multiple summaries from LLMs, score them with automated evaluators such as G-Eval (Liu et al., 2023) and FineSuRE (Song et al., 2024), and construct preference pairs by selecting high-scoring candidates as chosen via Best-of- N and low-scoring ones as rejected via Worst-of- N . In contrast, *Reference-based* methods (Cao & Wang, 2021; Mishra et al., 2024) construct rejected summaries by synthetically editing a faithful, chosen reference summary by injecting or altering factual information using language models.

As rejected summaries are comparatively easier to generate, prior work has placed greater emphasis on generating rejections, often favoring quantity over quality under the assumption that more negatives yield stronger training signals (Cao & Wang, 2021; Mishra et al., 2024; Xie et al., 2025). Yet this focus on quantity leaves a crucial question underexplored: "whether these rejections truly provide the right signal?" That is, rejections should be plentiful and exhibit sufficient alignment potential, which we define as the joint condition where

*Corresponding author.

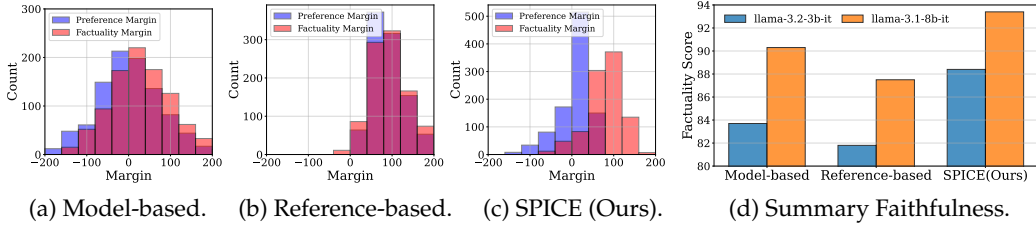


Figure 1: Comparison of (c) our SPICE with (a) the model-based SummLlama (Song et al., 2025) and (b) the reference-based SYNFAc-EDIT (Mishra et al., 2024), w.r.t preference and factuality margins between generated rejected summaries and their corresponding chosen summaries. (d) shows the average factuality of summaries from two Llama models trained via DPO (Rafailov et al., 2023) on preference pairs from each method.

the difference between the factuality margin and the preference margin is large—the rejection remains difficult for the model to distinguish from the chosen summary before training (small preference margin) while still presenting a clear factual contrast with the chosen one (large factuality margin) (refer to Section 3.1 for details).

However, our analysis in Figure 1 reveals that neither approach consistently produces rejections that satisfy both conditions. Instead, the model-based method tends to produce rejections that closely resemble the chosen summary, resulting in only subtle hallucinations and thus small preference and factuality margins. By contrast, the reference-based method seeks strong contrast with the chosen summary. Still, in doing so, it often introduces obvious factual errors, producing large preference and factuality margins that make the rejections trivial to separate. That is, existing methods fail to produce informative rejections, summaries that appear plausible to the model while remaining factually distinct.

In this paper, we theoretically and empirically establish what constitutes effective rejected summaries for faithfulness alignment through the notion of alignment potential, and in doing so, identify key factors. First, the *hallucination level* in the rejected summary determines its factual distance from the chosen one, directly influencing the factuality margin. Second, the *summary length* and *prompt complexity* significantly impact the preference margin, as longer or more complex generations tend to be easier for the model to distinguish. Crucially, by controlling the target factor while isolating the others, we demonstrate that targeted margin control not only shapes alignment potential but also translates directly into substantial gains in summary faithfulness after alignment tuning.

Building on this foundation, we introduce SPICE (Simple Prompt for Informative and Contrastive Examples), a simple yet effective prompting strategy that generates rejected summaries with high hallucination, matched length, and minimal prompt instructions, thereby maximizing “alignment potential”—achieving a small preference margin and a large factuality margin in Figure 1(c), such that the gap between the two margins is maximized. This enlarged margin gap strengthens the alignment potential by ensuring that the rejected example is not only clearly incorrect (high factuality margin) but also non-trivially distinguishable from the chosen example (small preference margin), leading to more faithful text summarization. Extensive experiments comparing SPICE with diverse rejection strategies, including rule-based methods as well as recent LLM-based approaches such as SummLlama (Song et al., 2025) and SYNFAc-EDIT (Mishra et al., 2024), demonstrate that SPICE substantially improves summary faithfulness across model sizes and alignment methods, including DPO (Rafailov et al., 2023), SimPO (Meng et al., 2024), and KTO (Ethayarajh et al., 2024). Moreover, SPICE improves faithfulness without an alignment tax: coherence, relevance, and abstractiveness remain comparable or slightly improved after alignment tuning.

Our main contributions are: (1) We present the first work to identify which rejected summaries are most effective for improving faithfulness in summarization, grounded in a theoretical and empirical analysis; (2) We introduce SPICE, a simple yet effective strategy to generate high-quality rejected summaries that maximize alignment potential; (3) We conduct extensive experiments comparing rule-based, model-based, and reference-based rejection strategies; and (4) We release a high-quality dataset for faithfulness alignment to support the development of more effective summarization models.

2 Related Work

Text Summarization. Text summarization is a long-standing task, but recent advances amplify concerns about the factuality of generated summaries (Chen et al., 2022; Ji et al., 2023). There has been a substantial effort to improve the faithfulness of summary generation. Earlier methods (Chen et al., 2022; Song et al., 2023) aimed to close the gap between generated and reference summaries by optimizing for ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020). Complementary work improves factuality through hard negative mining, using unfaithful summaries as negative samples in contrastive learning (Liu & Liu, 2021; Cao & Wang, 2021; Xu et al., 2022). For example, CLIFF (Cao & Wang, 2021) constructs hard negatives via techniques such as entity swapping to encourage consistency modeling.

With the advent of LLMs, recent efforts have focused on enhancing summary faithfulness via instruction tuning (Zhang et al., 2023; Zhou et al., 2023) and preference optimization (Stiennon et al., 2020; Lee et al., 2024a). In this context, two approaches, including model- and reference-based methods, aim at constructing large chosen–rejected summary pairs for alignment (Choi et al., 2024; Mishra et al., 2024; Song et al., 2025). While they improve faithfulness, the types of rejected summaries that best support alignment remain unclear.

Preference Optimization. This serves as an effective way to align language models with human intention (Jiang et al., 2024; Ryu et al., 2024). Two primary methods dominate this field: RM-based approaches such as PPO (Schulman et al., 2017) and RM-free methods such as DPO (Rafailov et al., 2023), SimPO (Meng et al., 2024), and KTO (Ethayarajh et al., 2024). In the summarization task, Stiennon et al. (2020) collected pairwise human preferences on Reddit posts for PPO training, while SYNFACT-EDIT (Mishra et al., 2024) and SummLlama (Song et al., 2025) construct synthetic chosen–rejected pairs and use them for DPO training.

Preference Data Quality. High-quality preference pairs are critical for alignment, as they sharpen a model’s ability to separate preferred from rejected responses (Gao et al., 2024; Duan et al., 2024). To reduce noise in preference data, recent work has explored filtering strategies (Khaki et al., 2024; Pace et al., 2024; Liu et al., 2024b), with growing attention to improving rejected responses (Liu et al., 2024d; Qiao et al., 2024; Xie et al., 2025). The quality is typically assessed by the external reward margin (score gap from external evaluators) (Khaki et al., 2024; Song et al., 2025), the internal policy margin (log-probability gap from the model itself) (Pace et al., 2024; Muldrew et al., 2024), or both (Liu et al., 2024b; Huang et al., 2025). While the optimal combination of these margins remains uncertain (Zhou et al., 2024; Cui et al., 2025), we study how varying them influences faithfulness alignment.

3 Formulation and Theoretical Analysis

3.1 Alignment Potential for Faithful Summarization

The notion of alignment potential measures how much training signal a preference pair provides (Huang et al., 2025; Deng et al., 2025). While introduced for general preference optimization, we extend it to faithful summarization. In our setting, an ideal model produces a preference margin that exactly matches the true factuality difference between chosen and rejected summaries. In practice, models deviate from this ideal, failing to clearly separate faithful from unfaithful outputs. We quantify this gap using the *preference* and *factuality margins*, which together define alignment potential.

Let x be the text document to be summarized, and let y_w and y_l be its chosen and rejected summaries. The two margins are defined as follows.

Preference Margin. Let π_θ be the summarization model prior to faithfulness alignment training, *i.e.*, the policy model. Then, the preference margin Δ_{pref} between chosen and rejected summaries is computed by:

$$\Delta_{\text{pref}} = r_\theta(y_w; x) - r_\theta(y_l; x), \text{ where } r_\theta(y; x) = \sum_{t=1}^{|y|} \log \pi_\theta(y_t | x, y_{<t}), \quad (1)$$

and y_t is the t -th token in the summary y .¹ While we do not take the absolute for analysis, a small margin reflects closeness to zero, not a negative.

Factuality Margin. Suppose that $f(\cdot)$ is an optimal external model that provides accurate assessments of the faithfulness of a summary y grounded on a source document x . Then, the factuality margin Δ_{fact} between chosen and rejected summaries is computed by:

$$\Delta_{\text{fact}} = f(y_w; x) - f(y_l; x). \quad (2)$$

This margin can be computed by either prompting an external model to assign scores to each response, as in G-Eval (Liu et al., 2023) or FineSuRE (Song et al., 2024), and then taking their difference, or by directly replacing the policy model with an external model and calculating the same implicit margin using its probabilities (see Appendix A.3), as suggested in recent work (Huang et al., 2025). Then, the discrepancy between the two margins is translated into the *alignment potential*, highlighting how much factual distinction the policy model still fails to capture and thus how much room remains for alignment as:

$$\left| \underbrace{\{f(y_w; x) - f(y_l; x)\}}_{\text{factuality margin}} - \underbrace{\{r_\theta(y_w; x) - r_\theta(y_l; x)\}}_{\text{preference margin}} \right|. \quad (3)$$

Consequently, we aim to generate rejected summaries y_l that, given a fixed chosen summary y_w , maximize the margin discrepancy $|\Delta_{\text{fact}} - \Delta_{\text{pref}}|$. Unlike prior work (Mishra et al., 2024; Song et al., 2025), our approach is the first to explicitly optimize rejection *quality* by maximizing alignment potential in faithful summarization.

3.2 Theoretical Analysis

We perform a theoretical analysis of an adversarial sampling approach, which preferentially samples rejected summaries with larger factuality-preference margin discrepancies, thus providing stronger alignment signals.

Setting. Consistent with other work (Rosset et al., 2024; Shi et al., 2025; Huang et al., 2025), we analyze the convergence behavior under the contextual bandit setup when training with DPO using stochastic gradient descent (SGD), comparing how quickly our alignment objective can be achieved when sampling rejected summaries adversarially vs. uniformly. We set the objective as minimizing the mean squared margin error (MSME) with fixed y_w :

$$\text{MSME}(\theta | y_w) = \frac{1}{|\mathcal{Y}|} \sum_{y_l \in \mathcal{Y}} (\Delta_{\text{fact}} - \Delta_{\text{pref}})^2 \quad (4)$$

Uniform vs. Adversarial Sampling. We compare two approaches for sampling losers (rejected summaries) given a fixed winner y_w : (i) *uniform sampling*, where y_l is drawn uniformly from $\mathcal{Y}$, and (ii) *adversarial sampling*, where y_l is drawn with probability proportional to its margin discrepancy $|\Delta_{\text{fact}} - \Delta_{\text{pref}}|$. Our analysis shows that adversarial sampling converges faster by prioritizing losers with large factuality margins, reducing MSME more effectively.

Analysis. Let $\Delta\text{MSME}_{\text{uni}}$, $\Delta\text{MSME}_{\text{hard-adv}}$ be the expected reduction in MSME per SGD step under uniform and hard adversarial sampling, where the latter deterministically selects the sample that maximizes the margin discrepancy. By computing the expectation of error reduction under both schemes, we can show (detailed proof in Appendix B):

$$\mathbb{E}_{\text{hard-adv}}[\Delta\text{MSME}] \geq \frac{4}{3} \mathbb{E}_{\text{uni}}[\Delta\text{MSME}]. \quad (5)$$

Therefore, hard adversarial sampling converges at least 33.3% faster than uniform sampling, as it prioritizes losers with the larger margin discrepancies, which provide the strongest corrective signals for reducing MSME. This result highlights that, when constructing preference pairs, favoring rejected summaries with high alignment potential in Eq. (3) yields high-quality preference pairs, as they effectively reduce the gap between the policy model’s preference margin and the ideal factuality margin. The result generalizes to SimPO with length normalization in Appendix A.2, though our proof is based on the DPO formulation.

¹Consistent with the DPO objective (Rafailov et al., 2023), we compute sequence-level log-probability differences. We assume a uniform reference policy and omit the normalization term in DPO, since our goal is data construction rather than training.

4 Controlling Margins via LLM Prompting

To translate our theoretical insight into practice, we examine how margins can be systematically controlled in LLM-based data construction. We identify three prompt-level factors that affect the margins: (i) hallucination level, (ii) summary length, and (iii) prompt complexity.

4.1 Base Dataset with Chosen Summary

Dataset. To study how prompting controls factuality and preference margins in summarization, we construct a dataset of source documents paired with high-quality summaries. Following prior work emphasizing source diversity for alignment (Song et al., 2025; Yu et al., 2025; Oh et al., 2025), we adopt the FEEDSUM corpus (Song et al., 2025), which comprises 13,672 documents from 7 datasets, covering diverse domains and varying lengths (both short and long), as well as both dialogue and non-dialogue formats. Each document is paired with summaries generated by 13 summarizers, including 3 non-LLMs, 7 open-source LLMs, and 3 proprietary LLMs. In addition, FEEDSUM provides faithfulness scores for each summary, enabling fine-grained analysis of factuality. The dataset includes a held-out test set of 1,400 documents used to evaluate alignment performance, while the training portion is used to construct chosen-rejected summary pairs. See Table 11 for dataset details.

Chosen Summary Selection. For further analysis, we construct a set of chosen summaries used to synthesize the corresponding rejected summaries. To obtain high-quality chosen summaries that serve as clear factual references for analyzing preference and factuality margins (Eqs. (1), (2)), we select, for each document in FEEDSUM, the highest-scoring summary among the 13 candidates with a faithfulness score above 0.8. If no such summary is available, the document is discarded. This yields 7,491 documents, each paired with a single high-quality reference summary used as the chosen one. Detailed statistics and additional post-processing steps are provided in Appendix C.

4.2 Three Factors for Effective Margin Control

We design prompting strategies to modulate each factor independently, while controlling the others to isolate its effect (see the isolation details in Appendix G.2). Given the document and its chosen summary, each prompt generates a rejected summary with controlled variation in the target aspect. By default, we use Llama3.1-70B-Instruct as the backbone, varying only the prompts, and report results with GPT-4o-mini and Gemma3-12B in Appendix G.5.

4.2.1 Factor 1: Hallucination Level

Hallucination level correlates with the factuality margin in Eq. (2), with larger margins indicating greater factual discrepancy. To control this, we vary hallucination in rejected summaries by explicitly specifying it in the prompt (e.g., edit 10% of the summary with hallucinated content). We consider four settings: None (0%, paraphrasing only), Mild (10%), Severe (over 90%), and Default (no explicit control), as shown in the prompts of Tables 29–31. We generate rejected summaries under each prompt setting and do not alter length or complexity, ensuring that the effect of hallucination is isolated. Summaries are scored for faithfulness using GPT-4o-mini with the G-EVAL prompt (Liu et al., 2023) (see Table 37).

Table 1 shows the distribution of rejected summaries across faithfulness intervals and their factuality margins, illustrating the effect of hallucination control. The results indicate that hallucination level effectively controls the factuality margin: each level produces a distinct distribution, yielding measurable margin differences. Thus, controlling hallucination is key to constructing chosen-rejected pairs with higher alignment potential. Notably, the Default setting performs comparably to Severe, suggesting that large factuality margins can arise even without explicit control, reducing the need for complex prompts.

Score	None	Mild	Severe	Default
7-10	96.8%	31.4%	0.0%	1.4%
4-6	2.3%	51.5%	1.1%	7.3%
1-3	0.9%	17.1%	98.9%	91.3%
Δ_{fact}	0.88	4.06	7.81	6.89

Table 1: Faithfulness score distribution across hallucination levels along with factuality margin.

4.2.2 Factor 2: Summary Length

Token length is a potential factor influencing the preference margin in Eq. (1), as longer (or shorter) summaries amplify log-probability differences with the chosen summary. To assess this, we control rejected summary length via prompts specifying Shorter, Similar, or Longer relative to the chosen summary. Following Tables 32–34, we extend the Default hallucination prompt with a single length instruction, allowing us to isolate the effect of length with minimal confounding from hallucination level and prompt complexity.

Figure 2 compares the token length gap and preference margin of rejected summaries across three length settings: Shorter, Similar, and Longer. Controlling rejected summary length effectively modulates the preference margin. Fewer tokens shift the margin left, while more tokens shift it right. This stems from the preference margin being computed over token-level log-likelihoods, as in Eq. (1); thus, longer summaries yield greater differences, amplifying the model’s preference, while shorter ones reduce it.

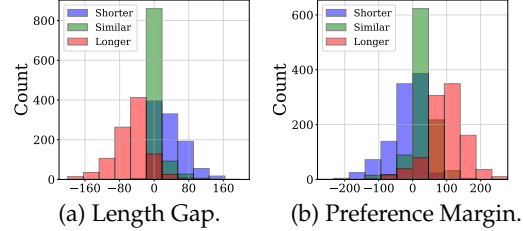


Figure 2: Token length gaps and preference margins over rejected summaries across three summary length settings.

4.2.3 Factor 3: Prompt Complexity

The final factor is the complexity of the prompt (*i.e.*, number of instructions) for rejected summary generation. While it does not directly affect hallucination level or summary length, higher complexity introduces structural and stylistic constraints (e.g., entity switching or numerical alterations), leading to lower-probability token generation and thus increasing the preference margin, a pattern similarly observed in (Kim et al., 2025). This encourages models to rely on trivial cues rather than factual differences. To examine its effect on alignment potential, we compare two prompts for generating high-level hallucinations (Section 4.2.1), both with hallucination and length controlled to avoid confounding from length gaps. The prompts differ only in instruction complexity: Simple, a length-controlled variant of Default (Table 33), and Complex, a length-controlled variant of Severe (Table 35).

Figure 3 compares factuality and preference margins under two prompts with different complexity. Both induce high hallucination and thus large factuality margins, but their preference margins differ sharply: Complex produces many trivially dispreferred rejections (far right of zero), while Simple yields margins closer to zero. This indicates that simpler prompts generate more effective contrastive pairs with higher alignment potential. We provide an in-depth analysis of how the instruction complexity and hallucination types of the Complex prompt affect the quality of rejected summaries in Appendix G.8.

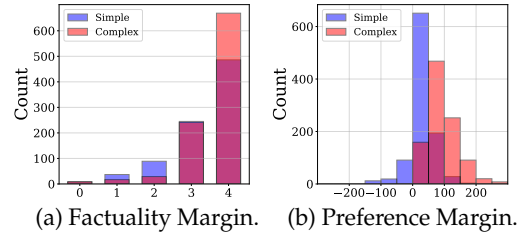


Figure 3: Factuality and preference margins over rejected summaries using Simple and Complex prompts.

4.3 Connection to Enhancing Summary Faithfulness

While our analysis shows how factors affect the two margins and guide prompt design, it remains unclear if they enhance faithfulness in alignment tuning. We therefore test whether their improved alignment potential yields measurable gains. Table 2 reports summary faithfulness, measured by FineSurE, G-Eval, and AlignScore, after DPO training using Llama3.2-3B-Inst on rejected summaries generated under eight prompt designs. Rows 3–9 correspond to the three investigated factors. The results show that alignment potential increases

Factor	FineSurE	G-Eval	AlignScore
Default	0.857	8.743	0.823
Severe	0.856	8.688	0.812
Mild	0.828	8.265	0.786
None	0.795	7.537	0.748
Longer	0.854	8.645	0.819
Shorter	0.859	8.712	0.821
Complex	0.850	8.682	0.810
Simple	0.885	9.037	0.854

Table 2: Comparison of rejected summaries w.r.t summary faithfulness.

You are given a document and a reference summary.
 Your task is to generate a factually inconsistent summary based on the provided document and reference summary.
Ensure that the generated summary has the same length as the reference summary.
Document: {Document}
Reference Summary: {Chosen Summary}
 Your answer MUST be in JSON format.
 The dictionary key should be "hallucinated_summary" as a string.

Table 3: Final prompt used by SPICE to generate rejected summaries for alignment training.

with larger factuality margins and smaller preference margins, which in turn lead directly to improved summary faithfulness.

First, the Severe hallucination setting yields higher faithfulness scores than Mild and None, as its rejections exhibit larger factuality margins. Second, the Simple(=Similar) prompt design, which combines matched-length and concise instructions, consistently improves faithfulness by reducing preference margins compared to Longer, Shorter, and Complex. In particular, we observe that this trend remains consistent across different LLMs used to generate rejected summaries, including GPT-4o-mini and Gemma3-12B-Instruct, in addition to Llama3.1-70B-Instruct, as detailed in Appendix G.5.

5 Proposed Prompt: SPICE

We introduce SPICE, a prompting strategy grounded in three insights for generating more effective rejected summaries that improve faithfulness alignment, as detailed:

Insight 1. Rejected summaries should *contain a high level of hallucination*, ensuring a large factuality margin and stronger contrast with chosen summaries.

Insight 2. Rejected summaries should be *similar in length* to the chosen ones, keeping the preference margin small so it reflects factuality differences rather than length.

Insight 3. Prompts should *be kept sufficiently simple* to avoid trivially dispreferred summaries, preserving a small preference margin and informative contrast.

By these principles, our prompt in Table 3 remains simple, using the base prompt with a single added length control instruction (highlighted in purple). We omit explicit hallucination control for conciseness, as LLMs naturally produce high-level hallucinations (Section 4.2.1). This matches the Simple prompt in Section 4.2.3 and, as shown in Figure 3 (blue), achieves large factuality margins with small preference margins. Accordingly, SPICE adopts this prompt to generate rejected summaries for the 7,491 training examples from Section 4.1.

6 Evaluation

To evaluate SPICE, we compare summarization quality after alignment tuning using rejected summaries generated by our method versus existing rejection methods.

Baseline Method. We compare SPICE with two reference-based and one model-based methods. (i) CLIFF (Cao & Wang, 2021) injects factual errors using two variants: Entity, which replaces named entities with same-type alternatives using a rule-based approach, and Mask, which uses a masked language model to generate plausible but incorrect content. (ii) SYNFACT-EDIT (Mishra et al., 2024) uses an LLM to generate factually inconsistent summaries. Designed for the clinical domain, we adapt its prompt for general summarization while preserving the core structure (Table 42). (iii) SummLlama (Song et al., 2025) selects rejected summaries from 13 summarizers based on faithfulness scores computed by FineSurE (Song et al., 2024). Details appear in Appendix D.

We further include two baselines: a zero-shot model (Zero-Shot) without alignment tuning; and the model trained with supervised fine-tuning (SFT) using only the chosen summaries. For all reference-based methods, including ours, rejected summaries are generated using Llama3.1-70B-Inst for consistency. Results with GPT-4o-mini and Gemma3-12B are discussed in Section G.5.

Method	Backbone: Llama3.2-3B-Instruct				Backbone: Llama3.1-8B-Instruct			
	Human	FineSurE	G-Eval	AlignScore	Human	FineSurE	G-Eval	AlignScore
Zero-Shot	0.736 _(0.00)	0.798 _(0.00)	7.519 _(0.00)	0.751 _(0.00)	0.824 _(0.00)	0.864 _(0.00)	8.317 _(0.00)	0.796 _(0.00)
SFT	0.693 _(0.00)	0.823 _(0.00)	7.947 _(0.00)	0.792 _(0.00)	0.835 _(0.00)	0.867 _(0.00)	8.789 _(0.00)	0.818 _(0.00)
CLIFF-Entity	0.743 _(0.00)	0.816 _(0.00)	8.011 _(0.00)	0.785 _(0.00)	0.787 _(0.00)	0.854 _(0.00)	8.742 _(0.00)	0.820 _(0.00)
CLIFF-Mask	0.730 _(0.00)	0.828 _(0.00)	7.981 _(0.00)	0.794 _(0.00)	0.789 _(0.00)	0.861 _(0.00)	8.773 _(0.00)	0.826 _(0.00)
SYNFAC-EDIT	0.765 _(0.00)	0.812 _(0.00)	8.264 _(0.00)	0.789 _(0.00)	0.816 _(0.00)	0.855 _(0.00)	8.797 _(0.00)	0.834 _(0.00)
SummLlama	0.792 _(0.01)	0.843 _(0.01)	8.473 _(0.01)	0.813 _(0.01)	0.879 _(0.02)	0.908 _(0.02)	9.038 _(0.01)	0.848 _(0.01)
SPICE (Ours)	0.833	0.885	9.037	0.854	0.909	0.934	9.438	0.895

Table 4: Faithfulness of summaries from two LLM backbones using six training methods, along with zero-shot inference (parentheses indicate statistical significance, $p < 0.05$).

Configuration. We train three instruct-tuned LLM backbones with varying sizes, including Llama3.2-1B/3B and Llama3.1-8B, with DPO using 7,491 chosen–rejected pairs across all methods. We fine-tune the models with QLoRA for 6,000 steps using AdamW, with an initial learning rate of $1e-6$ (cosine decayed). See Appendix E for additional training details.

Metric. We evaluate rejected summaries by measuring faithfulness improvement on 1,400 test documents from FeedSum (seven domain datasets) (Song et al., 2025). Faithfulness is measured using three reliable automatic factuality metrics: FineSurE (Song et al., 2024), G-Eval (Liu et al., 2023), and AlignScore (Zha et al., 2023), along with human evaluation (Human). G-Eval uses a 1–10 Likert scale, whereas others use a 0–1 scale. Human evaluation was conducted on Amazon MTurk following the protocol of Lee et al. (2024b). We evaluated 1,400 document–summary pairs, comprising 100 samples from each of the 14 summarizers in Table 4, with three annotators per pair providing sentence-level factual consistency judgments. Inter-annotator agreement was Krippendorff’s $\alpha = 0.53$, comparable to prior sentence-level factuality annotation studies. Metric details are in Appendix F, including evaluation protocol (Section F.3) and human evaluation scale (Section F.4.1).

6.1 Detailed Analysis of SPICE’s Practical Properties

Summary Faithfulness Improvement. Table 4 presents the faithfulness scores of summaries generated by 7 methods across two backbone models. Our proposed method, SPICE, consistently achieves the highest factuality scores across all model sizes and metrics, including both human and automated evaluations. Specifically in human evaluation, SPICE achieves the highest factuality scores of 0.833 (3B) and 0.909 (8B), clearly surpassing other methods by significant margins. Notably, the smaller 3B model trained with SPICE achieves faithfulness scores exceeding those of the larger 8B zero-shot baseline; it achieves higher scores in Human (0.833 vs. 0.824), FineSurE (0.885 vs. 0.864), G-Eval (9.037 vs. 8.317) and AlignScore (0.854 vs. 0.796), highlighting its exceptional efficiency and alignment effectiveness. All main-table results are statistically significant ($p < 0.05$); see Appendix F.1 for details. These results show that our simple prompting, by effectively controlling factuality and preference margins, enables strong faithfulness alignment in text summarization.

Factuality and Preference Margin. Table 5 shows the standardized factuality margin ($\tilde{\Delta}_{\text{fact}}$) and preference margin ($\tilde{\Delta}_{\text{pref}}$), along with alignment potential ($|\tilde{\Delta}_{\text{fact}} - \tilde{\Delta}_{\text{pref}}|$), computed from training datasets used by each method. Both margins are first computed in log-probability space: Δ_{fact} is obtained implicitly from the external factuality evaluator, while Δ_{pref} is obtained from the target policy model. Following Appendix B.3, we standardize the two margins separately using the global mean and standard deviation pooled over all instance-level comparisons:

Method	$\tilde{\Delta}_{\text{fact}}$	$\tilde{\Delta}_{\text{pref}}$	$ \tilde{\Delta}_{\text{fact}} - \tilde{\Delta}_{\text{pref}} $
CLIFF-Entity	-0.16	0.08	0.24
CLIFF-Mask	0.78	0.51	0.27
SYNFAC-EDIT	-0.12	-0.44	0.32
SummLlama	0.49	-0.36	0.85
SPICE (Ours)	1.56	0.12	1.44

Table 5: Alignment potential of rejected summaries over five methods. Llama3.1-70B-Inst (external evaluator) and Llama3.2-3B-Inst (policy model) are used to compute $\tilde{\Delta}_{\text{fact}}$ and $\tilde{\Delta}_{\text{pref}}$.

$$\tilde{\Delta}_{\text{fact}} = (\Delta_{\text{fact}} - \mu_{\text{fact}}^{\text{pool}}) / \sigma_{\text{fact}}^{\text{pool}}, \quad \tilde{\Delta}_{\text{pref}} = (\Delta_{\text{pref}} - \mu_{\text{pref}}^{\text{pool}}) / \sigma_{\text{pref}}^{\text{pool}}. \quad (6)$$

Equation 6 removes the scale mismatch between the two margins while preserving comparability across methods. Negative standardized values indicate margins below the pooled mean and do not correspond to negative raw log-probability margins.

SPICE achieves the highest alignment potential, reaching 1.44 on generated rejected summaries, significantly exceeding other approaches and indicating more informative, non-trivial rejections for faithfulness alignment. Although it does not minimize the preference margin, it combines a large factuality margin with a moderate preference margin to produce the largest gap. This highlights that maximizing their interaction, rather than optimizing each margin independently, is key to high-quality rejections (Section 3.2).

In contrast, CLIFF-Entity and CLIFF-Mask yield small alignment potentials of 0.24 and 0.27, respectively, due to limited separation between their factuality and preference margins. SYNFACT-EDIT also shows weak factual contrast, resulting in 0.32. SummLlama achieves the strongest baseline result of 0.85, but its factuality margin remains substantially lower than that of SPICE. Detailed analyses are provided in Appendix D.

6.2 Detailed Analysis of SPICE’s Practical Performance

Data Efficiency. To show the data efficiency of SPICE, we compare chosen–rejected pairs generated by SPICE with those from the second-best method, SummLlama (Table 4). Table 6 reports faithfulness scores as the number of chosen–rejected training examples is reduced. Notably, SPICE achieves performance comparable to SummLlama using only 1/16 of the data, and improves as more samples are added, demonstrating robust faithfulness alignment in low-resource settings. Appendix G.7 shows SPICE achieves a $2.1\times$ speed-up in data generation for the same number of preference pairs.

Alignment Tax. Faithfulness alone does not define summary quality, as alignment may incur an alignment tax (Song et al., 2025) that diminishes performance. Table 7 presents performance differences across four quality dimensions before and after alignment tuning with SPICE. The three dimensions, faithfulness, coherence, and relevance, are measured using G-Eval (Liu et al., 2023) (see prompts in Tables 37 to 39), while abstractiveness is assessed using the novel n -gram method. Detailed descriptions and formulations are in Appendix F. The results show that while SPICE substantially improves faithfulness, it incurs no alignment tax, maintaining comparable or slightly better relevance (8.579 vs. 8.563) and abstractiveness (0.689 vs. 0.686), with negligible change in coherence (8.059 vs. 8.102). This suggests that SPICE provides strong factual alignment signals while preserving overall summary quality.

Comparison with Adversarial Sampling. We further analyze whether adversarial sampling (*i.e.*, top- k selection of rejected summaries with the largest alignment potential from a large candidate pool) can replace our simple SPICE prompt that directly generates high alignment-potential rejections. For comparison, we adversarially sample rejected summaries from the full FeedSum seed set. For each of the 7,491 unique documents (each with 13 rejected-summary candidates), we select the top-

Method	FineSurE	G-Eval	AlignScore
SummLlama	0.843	8.473	0.813
SPICE	0.885	9.037	0.854
SPICE ^{1/2}	0.874	8.814	0.846
SPICE ^{1/4}	0.864	8.736	0.833
SPICE ^{1/8}	0.858	8.689	0.825
SPICE ^{1/16}	0.846	8.503	0.812

Table 6: Effectiveness of SPICE’s rejected summaries for DPO on Llama3.2-3B-Inst. We sample 7,491 training examples at four scales: 3,745 (1/2), 1,872 (1/4), 936 (1/8), and 468 (1/16).

Method	Faith (↑)	Coh. (↑)	Rel. (↑)	Abs. (↑)
Zero-Shot	7.519	8.102	8.563	0.686
DPO-SPICE	9.037	8.059	8.579	0.689

Table 7: Multi-dimensional summary quality on Llama3.2-3B-Inst. before/after training with SPICE across four dimensions: faithfulness (Faith.), coherence (Coh.), relevance (Rel.), and abstractiveness (Abs.).

Method	$\tilde{\Delta}_{\text{fact}}$	$\tilde{\Delta}_{\text{pref}}$	$ \tilde{\Delta}_{\text{fact}} - \tilde{\Delta}_{\text{pref}} $
Adv. Sampling	0.51	-0.45	0.96
SPICE (Ours)	1.56	0.12	1.44

Table 8: Factuality and preference margins on two training datasets. Llama3.1-70B-Inst (evaluator) and Llama3.2-3B-Inst (policy) are used to compute $\tilde{\Delta}_{\text{fact}}$ and $\tilde{\Delta}_{\text{pref}}$.

Method	Non-Dialogue				Dialogue			Average
	CNNDM	WikiHow	GovReport	PubMed	DialogSum	MediaSum	MeetingBank	
Zero-Shot	0.850 _(0.00)	0.723 _(0.00)	0.901 _(0.00)	0.891 _(0.00)	0.685 _(0.00)	0.785 _(0.00)	0.751 _(0.00)	0.798 _(0.00)
SFT	0.864 _(0.00)	0.796 _(0.00)	0.913 _(0.00)	0.878 _(0.00)	0.726 _(0.01)	0.819 _(0.00)	0.765 _(0.00)	0.823 _(0.00)
CLIFF-Entity	0.861 _(0.00)	0.712 _(0.00)	0.952 _(0.02)	0.916 _(0.00)	0.666 _(0.00)	0.839 _(0.00)	0.766 _(0.00)	0.816 _(0.00)
CLIFF-Mask	0.865 _(0.00)	0.751 _(0.00)	0.953 _(0.02)	0.930 _(0.01)	0.710 _(0.00)	0.809 _(0.00)	0.778 _(0.00)	0.828 _(0.00)
SYNFAC-EDIT	0.854 _(0.00)	0.736 _(0.00)	0.938 _(0.01)	0.906 _(0.00)	0.658 _(0.00)	0.821 _(0.00)	0.771 _(0.00)	0.812 _(0.00)
SummLlama	0.856 _(0.00)	0.805 _(0.00)	0.952 _(0.02)	0.929 _(0.01)	0.734 _(0.02)	0.846 _(0.01)	0.779 _(0.00)	0.843 _(0.01)
SPICE (Ours)	0.924	0.854	0.978	0.964	0.768	0.882	0.825	0.885

Table 9: FineSurE score (faithfulness) performance comparison across datasets in Llama3.2-3B-Instruct with DPO (parentheses indicate statistical significance, $p < 0.05$).

1 candidate with the largest potential, ensuring the resulting dataset matches the size of SPICE for fair comparison. Table 8 summarizes factuality and preference margins with alignment potential. Even after selecting the best among 13 candidates, alignment potential remains substantially lower than SPICE, indicating that large pools from unoptimized prompts (e.g., SummLlama) fail to produce sufficiently high-quality rejections. Moreover, adversarial sampling is highly inefficient, requiring over $200\times$ more candidates to match SPICE’s performance (Appendix G.10).

Domain-level Performance Analysis Table 9 shows the faithfulness performance across our test datasets. Our proposed method (SPICE) consistently outperforms baselines across both Non-Dialogue and Dialogue domains. Specifically, it achieves notable improvements in WikiHow (0.723 vs. 0.854) in the non-dialogue domain, and MediaSum (0.785 vs. 0.882) in the dialogue domain. Overall, SPICE attains the highest average FineSurE score (0.885), highlighting its robustness and effectiveness in factuality evaluation across diverse summarization domains. These findings highlight the effectiveness of SPICE in achieving robust factuality improvements across diverse summarization domains.

Optimization with KTO and SimPO. We further fine-tune Llama3.1-8B-Instruct using KTO and SimPO to examine whether the benefits of SPICE generalize beyond DPO. As shown in Table 10, SPICE improves all three faithfulness metrics across DPO, KTO, and SimPO relative to configurations without explicit control (Default), demonstrating the effectiveness of equal-length filtering through margin alignment. Moreover, all SPICE-trained models outperform both the zero-shot 8B baseline and the nearly $10\times$ larger zero-shot 70B model. These results demonstrate that SPICE remains effective across different optimization objectives.

Method	FineSurE	G-Eval	AlignScore
Zero-Shot	0.864 _(0.00)	8.317 _(0.00)	0.796 _(0.00)
Zero-Shot-70B	0.910 _(0.02)	8.938 _(0.00)	0.838 _(0.00)
DPO-Default	0.902 _(0.00)	8.916 _(0.00)	0.843 _(0.00)
KTO-Default	0.886 _(0.00)	8.857 _(0.00)	0.838 _(0.00)
SimPO-Default	0.908 _(0.00)	9.081 _(0.00)	0.852 _(0.00)
DPO-SPICE	0.934 _(0.09)	9.438 _(0.13)	0.895 _(0.07)
KTO-SPICE	0.912 _(0.02)	9.178 _(0.01)	0.857 _(0.01)
SimPO-SPICE	0.945	9.512	0.908

Table 10: Comparison of DPO with KTO and SimPO using SPICE on Llama3.1-8B-Instruct.

SimPO-SPICE achieves the highest scores overall, although its improvements over DPO-SPICE are not statistically significant ($p > 0.05$). KTO yields lower scores than DPO and SimPO in our setup, because its batch-level estimation was less stable with our batch size. However, KTO-SPICE still outperforms every Default setting across all metrics.

7 Conclusion

We provided theoretical and empirical analyses of what makes rejected summaries effective for faithfulness alignment. We showed that high hallucination, length matching, and simple prompts are key to balancing preference and factuality margins, shaping alignment potential. Based on these, we proposed SPICE, which generates higher-value rejections and substantially improves factuality while maintaining coherence, relevance, and abstractiveness. We further showed that SPICE achieves the highest alignment potential among all methods and is highly data-efficient, matching competitive baselines with a fraction of the data.

Ethics Statement

Our work investigates the use of LLM-generated feedback for alignment in text summarization and does not introduce significant ethical concerns, as all data is generated and evaluated using publicly available models or APIs. For human evaluation, we adhere to established protocols from prior work to ensure fairness and annotation quality. Annotators were compensated above the U.S. minimum wage, with additional bonuses provided based on consistency and performance.

Reproducibility Statement

We provide a comprehensive description of our data and experimental setup to support reproducibility. Section 4.1 and Appendices C and D detail the data selection and processing procedures. Section 6 describes the baseline methods, model configurations, and evaluation metrics. Training details, including hyperparameters, are provided in Appendix E, while evaluation metrics are further described in Appendix F. We also specify the publicly available checkpoints for open-source models (e.g., Hugging Face) and note that proprietary models are accessed via paid APIs such as OpenAI. Additional model and checkpoint details are summarized in Table 27, and all prompts used in this work are included in the Appendix.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00334343), Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00445087, Enhancing AI Model Reliability Through Domain-Specific Automated Value Alignment Assessment), and the National Research Foundation of Korea (NRF) funded by Ministry of Science and ICT. (RS-2022-NR068758)

References

- Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *AISTATS*, 2024.
- Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4), 324–345, 1952.
- Shuyang Cao and Lu Wang. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *EMNLP*, 2021.
- Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *NeurIPS*, 2021.
- Xiuying Chen, Mingzhe Li, Xin Gao, and Xiangliang Zhang. Towards improving faithfulness in abstractive summarization. In *NeurIPS*, 2022.
- Jaepill Choi, Kyubyung Chae, Jiwoo Song, Yohan Jo, and Taesup Kim. Model-based preference optimization in abstractive summarization without human feedback. In *EMNLP*, 2024.
- Guofeng Cui, Pichao Wang, Yang Liu, Zemian Ke, Zhu Liu, and Vimal Bhat. CRPO: Confidence-reward driven preference optimization for machine translation. In *ACL Findings*, 2025.
- Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. Less is more: Improving llm alignment via preference data selection. In *NeurIPS*, 2025.

- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *NeurIPS*, 2023.
- Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. In *EMNLP Findings*, 2023.
- Shitong Duan, Xiaoyuan Yi, Peng Zhang, Yan Liu, Zheng Liu, Tun Lu, Xing Xie, and Ning Gu. Negating negatives: Alignment with human negative samples via distributional dispreference optimization. In *EMNLP Findings*, 2024.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model alignment as prospect theoretic optimization. In *ICML*, 2024.
- Yang Gao, Dana Alon, and Donald Metzler. Impact of preference noise on the alignment performance of generative language models. In *COLM*, 2024.
- Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Zexu Sun, Bowen Sun, Huimin Chen, Ruobing Xie, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Controllable preference optimization: Toward controllable multi-objective alignment. In *EMNLP*, 2024.
- Kexin Huang, Junkang Wu, Ziqian Chen, Xue Wang, Jinyang Gao, Bolin Ding, Jiancan Wu, Xiangnan He, and Xiang Wang. Larger or smaller reward margins to select preferences for llm alignment? In *ICML*, 2025.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.
- Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. A survey on human preference learning for large language models. *arXiv preprint arXiv:2406.11191*, 2024.
- Saeed Khaki, JinJin Li, Lan Ma, Liu Yang, and Prathap Ramachandra. RS-DPO: A hybrid rejection sampling and direct preference optimization method for alignment of large language models. In *NAACL Findings*, 2024.
- Joongwon Kim, Anirudh Goyal, Aston Zhang, Bo Xiong, Rui Hou, Melanie Kambadur, Dhruv Mahajan, Hannaneh Hajishirzi, and Liang Tan. A systematic examination of preference learning through the lens of instruction-following. In *NAACL*, 2025.
- Klaus Krippendorff. Computing krippendorff’s alpha-reliability, 2011.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. In *ICML*, 2024a.
- Yuho Lee, Taewon Yun, Jason Cai, Hang Su, and Hwanjun Song. UniSumEval: Towards unified, fine-grained, multi-dimensional summarization evaluation for llms. In *EMNLP Findings*, 2024b.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020.
- Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.
- Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. Chain of hindsight aligns language models with feedback. In *ICLR*, 2024a.

- Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. In *ICLR*, 2024b.
- Wei Liu, Yang Bai, Chengcheng Han, Rongxiang Weng, Jun Xu, Xuezhi Cao, Jingang Wang, and Xunliang Cai. Length desensitization in direct preference optimization. *arXiv preprint arXiv: 2409.06411*, 2024c.
- Xiao Liu, Xixuan Song, Yuxiao Dong, and Jie Tang. Extensive self-contrast enables feedback-free language model alignment. *arXiv preprint arXiv:2404.00604*, 2024d.
- Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In *EMNLP-IJCNLP*, 2019.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: Nlg evaluation using gpt-4 with better human alignment. In *EMNLP*, 2023.
- Yixin Liu and Pengfei Liu. Simcls: A simple framework for contrastive learning of abstractive summarization. In *ACL*, 2021.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *ACL*, 2020.
- Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. In *NeurIPS*, 2024.
- Hyangsuk Min, Yuho Lee, Minjeong Ban, Jiaqi Deng, Nicole Hee-Yeon Kim, Taewon Yun, Hang Su, Jason Cai, and Hwanjun Song. Towards multi-dimensional evaluation of llm summarization across domains and languages. In *ACL*, 2025.
- Prakamya Mishra, Zonghai Yao, Parth Vashisht, Feiyun Ouyang, Beining Wang, Vidhi Dhaval Mody, and Hong Yu. SYNFACT-EDIT: Synthetic imitation edit feedback for factual alignment in clinical summarization. In *EMNLP*, 2024.
- William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. In *ICML*, 2024.
- Jihwan Oh, Jeonghwan Choi, Nicole Hee-Yoen Kim, Taewon Yun, and Hwanjun Song. Learning to verify summary facts with fine-grained LLM feedback. In *COLING*, 2025.
- 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. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.
- Alizée Pace, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. West-of-n: Synthetic preferences for self-improving reward models. *arXiv preprint arXiv:2401.12086*, 2024.
- Xiao Pu, Mingqi Gao, and Xiaojun Wan. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*, 2023.
- Shiqi Qiao, Ning Xu, Biao Liu, and Xin Geng. Negative-prompt-driven alignment for generative language model. *arXiv preprint arXiv:2410.12194*, 2024.
- Haoyi Qiu, Kung-Hsiang Huang, Jingnong Qu, and Nanyun Peng. Amrfact: Enhancing summarization factuality evaluation with amr-driven negative samples generation. In *NAACL*, 2024.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

- Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacrose, Ahmed Awadallah, and Tengyang Xie. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv: 2404.03715*, 2024.
- Sangwon Ryu, Heejin Do, Yunsu Kim, Gary Lee, and Jungseul Ok. Multi-dimensional optimization for text summarization via reinforcement learning. In *ACL*, 2024.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.
- Ruizhe Shi, Runlong Zhou, and Simon Shaolei Du. The crucial role of samplers in online direct preference optimization. In *ICLR*, 2025.
- Hwanjun Song, Igor Shalyminov, Hang Su, Siffr Singh, Kaisheng Yao, and Saab Mansour. Enhancing abstractiveness of summarization models through calibrated distillation. In *EMNLP Findings*, 2023.
- Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. FineSurE: Fine-grained summarization evaluation using llms. In *ACL*, 2024.
- Hwanjun Song, Taewon Yun, Yuho Lee, Jihwan Oh, Gihun Lee, Jason Cai, and Hang Su. Learning to summarize from llm-generated feedback. In *NAACL*, 2025.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *NeurIPS*, 2020.
- Liyan Tang, Igor Shalyminov, Amy Wing-mei Wong, Jon Burnsky, Jake W Vincent, Yu'an Yang, Siffr Singh, Song Feng, Hwanjun Song, Hang Su, et al. TofuEval: Evaluating hallucinations of llms on topic-focused dialogue summarization. In *NAACL*, 2024.
- Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. Confit: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In *NAACL*, 2022.
- Shuo Xie, Fangzhi Zhu, Jiahui Wang, Lulu Wen, Wei Dai, Xiaowei Chen, Junxiong Zhu, Kai Zhou, and Bo Zheng. Mppo: Multi pair-wise preference optimization for llms with arbitrary negative samples. In *COLING*, 2025.
- Shusheng Xu, Xingxing Zhang, Yi Wu, and Furu Wei. Sequence level contrastive learning for text summarization. In *AAAI*, 2022.
- Tianshu Yu, Ting-En Lin, Yuchuan Wu, Min Yang, Fei Huang, and Yongbin Li. Diverse ai feedback for large language model alignment. *Transactions of the Association for Computational Linguistics*, 13:392–407, 2025.
- Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function. In *ACL*, 2023.
- Haopeng Zhang, Xiao Liu, and Jiawei Zhang. Summit: Iterative text summarization via chatgpt. In *EMNLP Findings*, 2023.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with bert. In *ICLR*, 2020.
- Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In *EMNLP Findings*, 2023.
- Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. WPO: Enhancing rlhf with weighted preference optimization. In *EMNLP*, 2024.
- Rongxin Zhu, Jianzhong Qi, and Jey Han Lau. Annotating and detecting fine-grained factual errors for dialogue summarization. In *ACL*, 2023.

A Discussion

A.1 Limitations and Future Work

In our study, we focus primarily on the task of faithfulness alignment within the text summarization domain. Rather than broadening the scope across multiple tasks, we deliberately focus on summarization, where we conduct an in-depth and controlled investigation using SOTA baselines and evaluation metrics, and extensive experimental analysis. However, SPICE has the potential to be applied to a wide range of alignment objectives with simple modification, such as helpfulness and safety, owing to its simplicity and ease of implementation. It can also be adapted to other generation tasks beyond summarization, including machine translation, question answering, and code generation. We leave the validation of its generality and robustness in these broader settings to future work.

A.2 Preference Definition and Length Normalization

A natural question arises regarding the role of sequence length in defining model preference. The distinction stems from the optimization objectives. Since we can assume a uniform reference policy $\pi_{ref}(y|x) = \frac{1}{|Y|}$ according to [Shi et al. \(2025\)](#), DPO ([Rafailov et al., 2023](#)) defines the loss using the *unnormalized cumulative log-probability* of the chosen and rejected summaries:

$$\mathcal{L}_{DPO} \propto \log \pi_{\theta}(y_w|x) - \log \pi_{\theta}(y_l|x),$$

which directly compares raw sequence likelihoods. Because cumulative log-probabilities naturally decrease with sequence length, this formulation is inherently length-sensitive [Liu et al. \(2024c\)](#). Note that if we set $\pi_{ref} = \pi_0$ in the DPO implementation, then $\frac{\pi_{\theta}(y|x)}{\pi_{ref}(y|x)} = 1$ prior to training, collapsing all preferences into a trivial ratio. *Hence, the reference term must be omitted in the data metric level.*

In contrast, SimPO ([Meng et al., 2024](#)) employs *length-normalized* log-probabilities:

$$\mathcal{L}_{SimPO} \propto \frac{1}{|y_w|} \log \pi_{\theta}(y_w|x) - \frac{1}{|y_l|} \log \pi_{\theta}(y_l|x).$$

Given that the primary distinction between these two objectives lies in the treatment of sequence length, our theoretical derivation of alignment potential for the DPO objective in Section 3.2 naturally generalizes to SimPO-style formulations.

It is worth noting that our analysis primarily focuses on the standard DPO objective, where the loss is defined on the cumulative (unnormalized) log probabilities as in Eq. 1. Accordingly, we did not apply length normalization, and instead explicitly incorporated length as an independent factor (Factor 2) in our framework. However, normalization-based variants such as SimPO apply token-level average log probabilities to remove length bias. In such cases, the effect of length is already normalized away, and the length factor can be ignored.

Interestingly, our results in Section 4.2.2 show that alignment is most effective when the chosen and rejected summaries are of similar length. This implies that a form of length normalization naturally arises at the data-construction level, which makes the DPO-style formulation effectively consistent with length-normalized variants such as SimPO. We view this as a positive signal: despite the differences in objective formulation, both approaches converge to the same practical guideline regarding length, highlighting the robustness of this factor in preference optimization.

Implications. These formulations confirm that alignment potential serves as a unifying framework: the margin computation simply aligns with the objective, whether unnormalized for DPO or normalized for SimPO. Therefore, length normalization is a consistent design choice within this generalized perspective, matching the specific learning objective.

A.3 How Log-probability Can Reflect Factuality?

A natural question arises regarding how log-probabilities of model outputs can serve as a factuality-aligned signal.

By definition, the log-probability represents the model’s likelihood of producing a given output (e.g., a summary) conditioned on a specific input prompt (e.g., “Factually summarize the input document.”). This formulation aligns with prior work on token-based preference control [Chen et al. \(2021\)](#); [Dong et al. \(2023\)](#); [Liu et al. \(2024a\)](#); [Guo et al. \(2024\)](#), where explicit tokens in the input prompt are used to condition model outputs along specific alignment dimensions. Our setup explicitly prioritizes factuality as the primary preference dimension by including the “factuality” token in the input prompt.

Implications. We anchor the alignment dimension specifically to factuality (i.e., mitigating hallucinations), rather than other summarization attributes such as coherence, relevance, or abtractiveness, using token-based preference control.

A.4 Alignment Potential vs. Hard Negative Mining

Alignment potential explicitly quantifies the learning value of a rejected summary by jointly considering two complementary factors: *negativeness* and *hardness*. Specifically, the negativeness is measured by the factuality margin Δ_{fact} using an external evaluator, capturing how clearly the rejection violates factual consistency (factual divergence) relative to the chosen summary, while the hardness is measured by the preference margin Δ_{pref} , reflecting how difficult it is for the policy model to distinguish the rejection from the chosen output. Alignment potential is then computed by combining these two signals, enabling a quantitative assessment of whether a rejection provides an informative and non-trivial learning signal.

In contrast, hard negative mining focuses solely on hardness, using Δ_{pref} of the policy model, and assumes that all rejected samples are valid negatives. As a result, hard negative mining evaluates the learning signal using a single criterion, namely the preference margin of the model being trained, without explicitly accounting for how negative a rejection is in terms of factual correctness.

Implications. While both alignment potential and hard negative mining aim to assess the usefulness of training signals, the key difference lies in whether the evaluation relies on a single internal model signal or jointly considers complementary signals from both an external and the internal policy model.

B Proof of Section 3.2

B.1 Convergence of Fixed–Winner DPO with Sampled Loser

Contextual bandit setting. We frame the problem in a contextual bandit setting like previous works in RLHF (Rafailov et al., 2023; Rosset et al., 2024; Huang et al., 2025) with a context space $\mathcal{X}$ and a full action (response) space $\mathcal{Y}$. For each $x \in \mathcal{X}$, a policy $\pi_\theta(\cdot | x)$ is a probability distribution on $\mathcal{Y}$, and all evaluation metrics are defined on this space.

Setting and population loss. To optimize the policy π_θ using preference data, we define the DPO population loss in its soft-label form. The objective is to minimize the expected negative log-likelihood:

$$\mathbb{E}_{(x, y_w, y_l) \sim D} [\mathcal{L}(x, y_w, y_l; \theta)] = \mathbb{E} \left[-p(y_w \succ y_l) \log \sigma(r_\theta(y_w; x) - r_\theta(y_l; x)) \right] \quad (7)$$

$$= \mathbb{E} \left[-\sigma(f(y_w; x) - f(y_l; x)) \log \sigma(r_\theta(y_w; x) - r_\theta(y_l; x)) \right], \quad (8)$$

where $p(y_w \succ y_l)$ represents the ground-truth preference probability from the data distribution D . In Eq. (8), we specifically assume that these preferences follow the Bradley-Terry model (Bradley & Terry, 1952) governed by an underlying oracle reward function f . Here, $r_\theta(y; x) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}$ denotes the implicit reward defined by the policy π_θ relative to the reference policy π_{ref} .

Under stochastic gradient descent, at iteration t we draw a triple (x^t, y_w^t, y_l^t) and update θ^t using the empirical loss. Because the sampled context and pair allow us to evaluate both $\mathcal{L}(x^t, y_w^t, y_l^t; \theta^t)$ and its swapped counterpart $\mathcal{L}(x^t, y_l^t, y_w^t; \theta^t)$, we follow prior work (Azar et al., 2024; Rosset et al., 2024; Huang et al., 2025) and adopt the symmetric update

$$\theta^{t+1} = \theta^t - \frac{1}{2} \eta^t \nabla_\theta \left(\mathcal{L}(x^t, y_w^t, y_l^t; \theta^t) + \mathcal{L}(x^t, y_l^t, y_w^t; \theta^t) \right),$$

where η^t is the step size.

Following (Shi et al., 2025), we take the reference policy to be uniform, $\pi_{ref}(y | x) = 1/|\mathcal{Y}|$. Under this choice, the model preference satisfies $\hat{r}_\theta(x, y) = \beta \theta(x, y) + \log Z(x)$, and the ground truth model θ^* characterized by $f(x, y) = \hat{r}_{\theta^*}(x, y) + \log Z^*(x)$ implies $\beta \theta^*(x, y) = f(x, y) + c(x)$ where $c(x)$ is constant.

Therefore, we can define the potential (misalignment) of summary and its variance:

$$\gamma_y^t := \beta \theta^t(x, y) - f(x, y), \quad \bar{\gamma}^t := \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \gamma_y^t, \quad V^t := \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} (\gamma_y^t - \bar{\gamma}^t)^2.$$

Then MSME can be written *explicitly* in terms of the γ 's as

$$\begin{aligned} \text{MSME}(\theta^t | y_w) &= \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} (\gamma_{y_w}^t - \gamma_y^t)^2 \quad \text{where } \text{SME} = (\gamma_{y_w}^t - \gamma_y^t)^2 = (\Delta_{\text{fact}} - \Delta_{\text{pref}})^2. \\ &= \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \left((\gamma_{y_w}^t - \bar{\gamma}^t) - (\gamma_y^t - \bar{\gamma}^t) \right)^2 \\ &= \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \left[(\gamma_{y_w}^t - \bar{\gamma}^t)^2 + (\gamma_y^t - \bar{\gamma}^t)^2 - 2(\gamma_{y_w}^t - \bar{\gamma}^t)(\gamma_y^t - \bar{\gamma}^t) \right] \\ &= (\gamma_{y_w}^t - \bar{\gamma}^t)^2 + \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} (\gamma_y^t - \bar{\gamma}^t)^2 - \underbrace{\frac{2}{|\mathcal{Y}|} (\gamma_{y_w}^t - \bar{\gamma}^t) \sum_{y \in \mathcal{Y}} (\gamma_y^t - \bar{\gamma}^t)}_{=0} \\ &= (\gamma_{y_w}^t - \bar{\gamma}^t)^2 + V^t. \end{aligned} \quad (9)$$

Thus controlling MSME is equivalent to controlling V^t up to the (winner–mean) offset $(\gamma_{y_w}^t - \bar{\gamma}^t)^2$.

Per-step decrease. With the optimal learning rate by step t , the variance decreases in expectation as

$$\mathbb{E}_{y \sim D}[V^{t+1} - V^t] = -\frac{1}{2|\mathcal{Y}|} \mathbb{E}_{y \sim D}[(\gamma_{y_w}^t - \gamma_y^t)^2], \quad (10)$$

as shown in previous work (Huang et al., 2025).

(1) Uniform loser sampling. If D is uniform over $\mathcal{Y}$, then

$$\mathbb{E}_{y \sim D}[(\gamma_{y_w}^t - \gamma_y^t)^2] = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} (\gamma_{y_w}^t - \gamma_y^t)^2 = (\gamma_{y_w}^t - \bar{\gamma}^t)^2 + V^t = \text{MSME}(\theta^t | y_w).$$

Plugging (11) into (10) yields, for the uniform sampler:

$$\mathbb{E}[V^{t+1} - V^t] = -\frac{1}{2|\mathcal{Y}|} \mathbb{E}_{y \sim D}[(\gamma_{y_w}^t - \gamma_y^t)^2] = -\frac{1}{2|\mathcal{Y}|} \text{MSME}(\theta^t | y_w) \quad (11)$$

$$\mathbb{E}[V^{t+1}] = V^t - \frac{1}{2|\mathcal{Y}|} \text{MSME}(\theta^t | y_w), \quad (12)$$

Under the stepsize choice used to derive (10), $\delta^t := \gamma_{y_w}^{t+1} - \gamma_{y_w}^t = -\frac{1}{2}(\gamma_{y_w}^t - \bar{\gamma}^t)$. Thus, for the uniform sampler:

$$\begin{aligned} \mathbb{E}[(\gamma_{y_w}^{t+1} - \bar{\gamma}^t)^2] &= \mathbb{E}[(\gamma_{y_w}^t - \bar{\gamma}^t + \delta^t)^2] \\ &= (\gamma_{y_w}^t - \bar{\gamma}^t)^2 + 2(\gamma_{y_w}^t - \bar{\gamma}^t) \mathbb{E}[\delta^t] + \mathbb{E}[(\delta^t)^2] \\ &= (\gamma_{y_w}^t - \bar{\gamma}^t)^2 - (\gamma_{y_w}^t - \bar{\gamma}^t)^2 + \frac{1}{4} \mathbb{E}[(\gamma_{y_w}^t - \gamma_y^t)^2] \\ &= \frac{1}{4} ((\gamma_{y_w}^t - \bar{\gamma}^t)^2 + V^t) = \frac{1}{4} \text{MSME}(\theta^t | y_w), \end{aligned} \quad (13)$$

where $\mathbb{E}[\delta^t] = -\frac{1}{2}(\gamma_{y_w}^t - \bar{\gamma}^t)$ and $\mathbb{E}[(\delta^t)^2] = \frac{1}{4} \mathbb{E}[(\gamma_{y_w}^t - \gamma_y^t)^2] = \frac{1}{4} ((\gamma_{y_w}^t - \bar{\gamma}^t)^2 + V^t)$.

Combining (12) and (13), the next-step MSME obeys the exact recursion

$$\begin{aligned} \mathbb{E}[\text{MSME}(\theta^{t+1} | y_w)] &= \mathbb{E}[(\gamma_{y_w}^{t+1} - \bar{\gamma}^t)^2] + \mathbb{E}[V^{t+1}] \\ &= \frac{1}{4} \text{MSME}(\theta^t | y_w) + V^t - \frac{1}{2|\mathcal{Y}|} \text{MSME}(\theta^t | y_w) \\ &= \left(\frac{1}{4} - \frac{1}{2|\mathcal{Y}|}\right) \text{MSME}(\theta^t | y_w) + V^t. \end{aligned} \quad (14)$$

(2) Adversarial loser sampling. Let adversarial sampler $y \sim q_\tau$ with

$$q_\tau(y) = \frac{|\gamma_{y_w}^t - \gamma_y^t|^\tau}{Z_\tau}, \quad Z_\tau = \sum_{y \in \mathcal{Y}} |\gamma_{y_w}^t - \gamma_y^t|^\tau, \quad \tau \geq 0.$$

Here, the parameter τ controls the sharpness of the sampling distribution. When $\tau = 0$, the sampler reduces to uniform sampling, while larger τ values place higher probability mass on losers whose margin discrepancy $|\gamma_{y_w}^t - \gamma_y^t| = |\Delta_{\text{fact}} - \Delta_{\text{pref}}|$ are larger, corresponding to more adversarial selections. In this part, we show that selecting rejected responses y_l in an adversarial manner that is, sampling those with larger absolute margins $|\gamma_{y_w}^t - \gamma_y^t|$ leads to a greater per-step decrease in MSME compared to uniform sampling, and thus accelerates convergence. In other words, given a fixed winner y_w , we analyze why choosing the loser y_l adversarially provides a clear advantage over uniform selection.

Define

$$\mu_\tau^t := \mathbb{E}_{y \sim q_\tau}[\gamma_y^t], \quad \text{Var}_\tau^t := \mathbb{E}_{y \sim q_\tau}[(\gamma_y^t - \mu_\tau^t)^2].$$

Then,

$$\begin{aligned} G_\tau^t &= \mathbb{E}_{y \sim q_\tau} \left[\left((\gamma_{y_w}^t - \mu_\tau^t) - (\gamma_y^t - \mu_\tau^t) \right)^2 \right] \\ &= \mathbb{E}_{y \sim q_\tau} \left[(\gamma_{y_w}^t - \mu_\tau^t)^2 \right] - 2(\gamma_{y_w}^t - \mu_\tau^t) \mathbb{E}_{y \sim q_\tau} [\gamma_y^t - \mu_\tau^t] + \mathbb{E}_{y \sim q_\tau} \left[(\gamma_y^t - \mu_\tau^t)^2 \right] \\ &= (\gamma_{y_w}^t - \mu_\tau^t)^2 + \text{Var}_\tau^t. \end{aligned}$$

Then the winner term expands as

$$\begin{aligned} \mathbb{E}_{q_\tau} [(\gamma_{y_w}^{t+1} - \bar{\gamma}^t)^2] &= \mathbb{E}_{q_\tau} [(\gamma_{y_w}^t - \bar{\gamma}^t + \delta^t)^2] \\ &= (\gamma_{y_w}^t - \bar{\gamma}^t)^2 + 2(\gamma_{y_w}^t - \bar{\gamma}^t) \mathbb{E}_{q_\tau} [\delta^t] + \mathbb{E}_{q_\tau} [(\delta^t)^2] \\ &= (\gamma_{y_w}^t - \bar{\gamma}^t)^2 - (\gamma_{y_w}^t - \bar{\gamma}^t)(\gamma_{y_w}^t - \mu_\tau^t) + \frac{1}{4} G_\tau^t, \end{aligned}$$

and the variance potential updates as

$$\mathbb{E}_{q_\tau} [V^{t+1}] = V^t - \frac{1}{2|\mathcal{Y}|} G_\tau^t.$$

Therefore we obtain the *exact* recursion:

$$\mathbb{E}_{q_\tau} [\text{MSME}(\theta^{t+1} | y_w)] = \text{MSME}(\theta^t | y_w) - (\gamma_{y_w}^t - \bar{\gamma}^t)(\gamma_{y_w}^t - \mu_\tau^t) + \left(\frac{1}{4} - \frac{1}{2|\mathcal{Y}|} \right) G_\tau^t. \quad (15)$$

(2.1) Adversarial is always faster than uniform. From the exact adversarial recursion we have equation 15 and for the uniform sampler ($\tau = 0$),

$$\mathbb{E}_{\text{uni}} [\text{MSME}(\theta^{t+1} | y_w)] = \text{MSME}(\theta^t | y_w) - (\gamma_{y_w}^t - \bar{\gamma}^t)^2 + \left(\frac{1}{4} - \frac{1}{2|\mathcal{Y}|} \right) G_0^t, \quad (16)$$

$$G_0^t = (\gamma_{y_w}^t - \bar{\gamma}^t)^2 + V^t. \quad (17)$$

Subtracting,

$$\begin{aligned} \Delta_\tau^t &:= \mathbb{E}_{\text{uni}} [\text{MSME}(\theta^{t+1} | y_w)] - \mathbb{E}_{q_\tau} [\text{MSME}(\theta^{t+1} | y_w)] \\ &= -(\gamma_{y_w}^t - \bar{\gamma}^t)(\mu_\tau^t - \bar{\gamma}^t) + \alpha (G_0^t - G_\tau^t), \quad \alpha := \frac{1}{4} - \frac{1}{2|\mathcal{Y}|} \in (0, \frac{1}{4}]. \end{aligned} \quad (18)$$

We will show $\Delta_\tau^t \geq 0$ for all $\tau \geq 0$.

(2.1.1) Base case. ($\tau = 0$). Since $\mu_0^t = \bar{\gamma}^t$ and $G_0^t = G_0^t$, clearly $\Delta_0^t = 0$.

(2.1.2) Recursive relation. ($\tau \rightarrow \tau + 1$).

By definition,

$$q_{\tau+1}(y) = \frac{\Delta \gamma^t q_\tau(y)}{\mathbb{E}_{q_\tau} [\Delta \gamma^t]},$$

so for any function f ,

$$\mathbb{E}_{q_{\tau+1}} [f(y)] = \frac{\mathbb{E}_{q_\tau} [f(y) \Delta \gamma^t]}{\mathbb{E}_{q_\tau} [\Delta \gamma^t]}.$$

Therefore,

$$\mu_{\tau+1}^t - \mu_\tau^t = \frac{\text{Cov}_{q_\tau}(\gamma_{y_w}^t, \Delta \gamma^t)}{\mathbb{E}_{q_\tau} [\Delta \gamma^t]}, \quad G_{\tau+1}^t - G_\tau^t = \frac{\text{Cov}_{q_\tau}((\Delta \gamma^t)^2, \Delta \gamma^t)}{\mathbb{E}_{q_\tau} [\Delta \gamma^t]}.$$

Thus from (18),

$$\begin{aligned} \Delta_{\tau+1}^t - \Delta_\tau^t &= -(\gamma_{y_w}^t - \bar{\gamma}^t)(\mu_{\tau+1}^t - \mu_\tau^t) + \alpha (G_\tau^t - G_{\tau+1}^t) \\ &= \frac{-1}{\mathbb{E}_{q_\tau} [\Delta \gamma^t]} \left[(\gamma_{y_w}^t - \bar{\gamma}^t) \text{Cov}_{q_\tau}(\gamma_{y_w}^t, \Delta \gamma^t) + \alpha \text{Cov}_{q_\tau}((\Delta \gamma^t)^2, \Delta \gamma^t) \right] \end{aligned} \quad (19)$$

We split the loser sampling subspace $\mathcal{Y}$ into two disjoint subsets:

$$\mathcal{Y}_1 = \{y \in \mathcal{Y} : \gamma_{y_w}^t < \gamma_y^t\}, \quad \mathcal{Y}_2 = \{y \in \mathcal{Y} : \gamma_{y_w}^t \geq \gamma_y^t\}.$$

Hence, the entire subspace can be written as

$$\mathcal{Y} = \mathcal{Y}_1 \cup \mathcal{Y}_2, \quad \mathcal{Y}_1 \cap \mathcal{Y}_2 = \emptyset.$$

Here, y_w is the minimizer of $\{\gamma_y^t : y \in \mathcal{Y}_1\}$, and y_w is the maximizer of $\{\gamma_y^t : y \in \mathcal{Y}_2\}$.

In case of $\mathcal{Y}_1$, $\Delta\gamma^t = -\gamma_{y_w}^t + \gamma_y^t$ and hence $\gamma_y^t = \gamma_{y_w}^t + \Delta\gamma^t$.

Thus,

$$\text{Cov}_{q_\tau}(\gamma_y^t, \Delta\gamma^t) = \text{Var}_{q_\tau}(\Delta\gamma^t) \geq 0,$$

while monotonicity ensures

$$\text{Cov}_{q_\tau}((\Delta\gamma^t)^2, \Delta\gamma^t) \geq 0.$$

Thus,

$$\Delta_{\tau+1}^t - \Delta_\tau^t = \frac{1}{\mathbb{E}_{q_\tau}[\Delta\gamma^t]} \left(-(\gamma_{y_w}^t - \bar{\gamma}^t) \text{Var}_{q_\tau}(\Delta\gamma^t) - \alpha \text{Cov}_{q_\tau}((\Delta\gamma^t)^2, \Delta\gamma^t) \right). \quad (20)$$

Bounding the covariance by:

$$\text{Cov}_{q_\tau}((\Delta\gamma^t)^2, \Delta\gamma^t) \leq \left(\max_y \Delta\gamma^t + \mathbb{E}_{q_\tau}[\Delta\gamma^t] \right) \text{Var}_{q_\tau}(\Delta\gamma^t).$$

$$\begin{aligned} \text{Cov}_{q_\tau}((\Delta\gamma^t)^2, \Delta\gamma^t) &= \mathbb{E}_{q_\tau}[(\Delta\gamma^t)^3] - \mathbb{E}_{q_\tau}[(\Delta\gamma^t)^2] \mathbb{E}_{q_\tau}[\Delta\gamma^t] \\ &= \mathbb{E}_{q_\tau}[(\Delta\gamma^t - \mathbb{E}_{q_\tau}[\Delta\gamma^t])^2 \cdot \Delta\gamma^t] + \mathbb{E}_{q_\tau}[\Delta\gamma^t] \text{Var}_{q_\tau}(\Delta\gamma^t). \end{aligned}$$

And using $\max_y \Delta\gamma^t \geq \Delta\gamma^t$:

$$\text{Cov}_{q_\tau}((\Delta\gamma^t)^2, \Delta\gamma^t) \leq \left(\max_y \Delta\gamma^t + \mathbb{E}_{q_\tau}[\Delta\gamma^t] \right) \text{Var}_{q_\tau}(\Delta\gamma^t).$$

Since $|\mathcal{Y}|$ is sufficiently large, γ^t follows a normal distribution (see the detailed theoretical and empirical analysis in Appendix B.3), and we connect their consequences based on two principles of asymptotic statistics. By the law of large numbers, the sample mean $\bar{\gamma}^t$ converges to the population mean. Separately, by extreme value theory in order statistics, the sample mid-range also converges to the population mean for a symmetric distribution. Hence, we obtain $\max_y \Delta\gamma^t = 2(\bar{\gamma}^t - \gamma_{y_w}^t) = -2(\gamma_{y_w}^t - \bar{\gamma}^t)$.

Therefore,

$$\begin{aligned} & -(\gamma_{y_w}^t - \bar{\gamma}^t) - \alpha \left(\max_y \Delta\gamma^t + \mathbb{E}_{q_\tau}[\Delta\gamma^t] \right) \\ & \geq -(\gamma_{y_w}^t - \bar{\gamma}^t) - 2\alpha \max_y \Delta\gamma^t \\ & = \frac{1}{2} \max_y \Delta\gamma^t - 2\alpha \max_y \Delta\gamma^t \\ & = \left(\frac{1}{2} - 2\alpha \right) \max_y \Delta\gamma^t \end{aligned}$$

Since $\alpha \leq \frac{1}{4}$, the bracket is nonnegative, and by equation 20, $\Delta_{\tau+1}^t - \Delta_\tau^t \geq 0$.

We have $\Delta_0^t = 0$ and $\Delta_{\tau+1}^t - \Delta_\tau^t \geq 0$. Therefore, $\Delta_\tau^t > 0$ for all $\tau > 0$,

$$\mathbb{E}_{\text{uni}, \mathcal{Y}_1}[\text{MSME}(\theta^{t+1}|y_w)] > \mathbb{E}_{q_\tau, \mathcal{Y}_1}[\text{MSME}(\theta^{t+1}|y_w)], \quad \forall \tau > 0.$$

For the set $\mathcal{Y}_2$, the condition $|\gamma_{y_w}^t - \gamma_{y_l}^t| = \gamma_{y_w}^t - \gamma_{y_l}^t$ allows us to use equation 19 to derive the same bounds as those found for $\mathcal{Y}_1$:

$$\mathbb{E}_{\text{uni}, \mathcal{Y}_2}[\text{MSME}(\theta^{t+1}|y_w)] > \mathbb{E}_{q_\tau, \mathcal{Y}_2}[\text{MSME}(\theta^{t+1}|y_w)], \quad \forall \tau > 0.$$

By law of total expectation and basic properties of inequalities over the space $\mathcal{Y}$,

$$\boxed{\mathbb{E}_{\text{uni}}[\text{MSME}(\theta^{t+1}|y_w)] > \mathbb{E}_{q_\tau}[\text{MSME}(\theta^{t+1}|y_w)], \quad \forall \tau > 0,}$$

showing that adversarial loser sampling is always faster than uniform in expectation.

(2.2) Hard-adversarial case. For the one step change for hard-adv,

$$\Delta_{\text{hard-adv}}^t = \text{MSME}(\theta^t|y_w) - \mathbb{E}_{q_\infty}[\text{MSME}(\theta^{t+1}|y_w)] = \frac{1}{4}(\gamma_{y_w}^t - \gamma_{y_l}^t)^2,$$

while the uniform step (with $\alpha = \frac{1}{4}$) gives

$$\Delta_{\text{uni}}^t = \text{MSME}(\theta^t|y_w) - \mathbb{E}_{\text{uni}}[\text{MSME}(\theta^{t+1}|y_w)] = \frac{3}{4}(\gamma_{y_w}^t - \bar{\gamma}^t)^2 - \frac{1}{4}V^t.$$

Using $(\gamma_{y_w}^t - \bar{\gamma}^t)^2 = \frac{1}{4}(\gamma_{y_w}^t - \gamma_{y_l}^t)^2$, we obtain

$$\frac{\Delta_{\text{hard-adv}}^t}{\Delta_{\text{uni}}^t} = \frac{\frac{1}{4}(\gamma_{y_w}^t - \gamma_{y_l}^t)^2}{\frac{3}{16}(\gamma_{y_w}^t - \gamma_{y_l}^t)^2 - \frac{1}{4}V^t} = \frac{4(\gamma_{y_w}^t - \gamma_{y_l}^t)^2}{3(\gamma_{y_w}^t - \gamma_{y_l}^t)^2 - 4V^t}.$$

Since $V^t \geq 0$,

$$\boxed{\mathbb{E}_{\text{hard-adv}}[\Delta \text{MSME}] \geq \frac{4}{3} \mathbb{E}_{\text{uni}}[\Delta \text{MSME}].}$$

This result demonstrates that **MSME reduction from the hard adversarial sampling converges at least 33.3% faster than uniform sampling.**

Conclusion. As we noted in Section 3.2, given a fixed chosen summary y_w , constructing a preference pair by selecting a rejected summary y_l that maximizes the factuality–preference margin $|\Delta_{\text{fact}} - \Delta_{\text{pref}}|$ allows us to reduce the MSME (i.e., the average margin discrepancy with respect to the target model) more rapidly, thereby accelerating convergence.

B.2 Training-Dynamics Evidence for Alignment Potential



Figure 4: DPO training dynamics across datasets with different alignment potentials: (a) evaluation loss, (b) reward margin, and (c) gradient norm.

To directly examine the relationship between alignment potential and DPO optimization, we track evaluation loss, reward margin, and gradient norm throughout training. Compared with lower-alignment-potential baselines, SPICE exhibits a faster loss reduction, a larger increase in reward margin, and a generally decreasing gradient norm.

B.3 Justification of Alignment Potential Normality

In this section, we empirically show that the alignment potential follows a normal distribution. While the external evaluation model and the policy model operate within inherently different logit spaces and exhibit distinct conditional densities, the margin discrepancy $\Delta_{\text{fact}} - \Delta_{\text{pref}}$ follows a normal distribution based on empirical observations:

1. Addressing score-scale mismatch via normalization.

Although the two margins share the same functional form, their absolute scales may differ significantly because they are produced by different models. To remove this scale mismatch of log-probability, like (Huang et al., 2025), we standardize each margin using its empirical mean and standard deviation:

$$\tilde{\Delta}_{\text{fact}} = \frac{\Delta_{\text{fact}} - \mu_{\text{fact}}}{\sigma_{\text{fact}}}, \quad \tilde{\Delta}_{\text{pref}} = \frac{\Delta_{\text{pref}} - \mu_{\text{pref}}}{\sigma_{\text{pref}}}.$$

2. Empirical Validation.

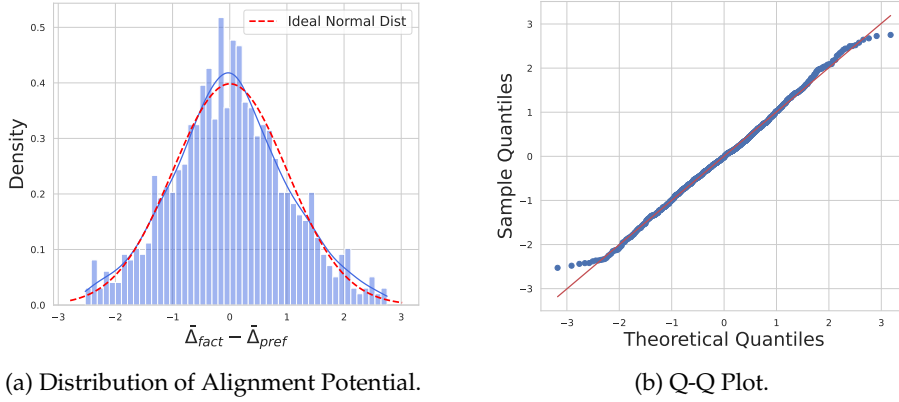


Figure 5: Results of the empirical validation of alignment potential normality. The dataset is constructed using SPICE, with values normalized from those shown in Figure 1c. The distribution closely follows a normal distribution (left), and the Q-Q plot further confirms that the alignment potential approximately satisfies normality (right).

We empirically validate that alignment potential normality holds in practice. As visualized in Figure 5, we plot the actual distribution of the standardized margin discrepancy (i.e., the alignment potential), computed from 1,000 SPICE samples. The empirical density (Figure 5a) exhibits a clear, unimodal, and symmetric bell-shaped curve that tightly aligns with the ideal Gaussian distribution. Furthermore, the corresponding Q-Q plot (Figure 5b) demonstrates a strong linear relationship, confirming the validity of the normality.

Dataset	Type	Document Length	Domain	# of Documents	Document Word Count (Min–Max)	Summary Word Count (Min–Max)
CNNM	Non-Dialogue	Short	News	1,603	664.8 (81–1919)	78.2 (8–221)
WikiHow			Lifestyle	551	124.8 (20–690)	41.9 (8–108)
GovReport		Long	Report	157	2388.8 (349–3368)	126.9 (10–348)
PubMed			Medical	713	1715.2 (125–3514)	102.2 (8–328)
DialogSum	Dialogue	Short	Daily Life	1,659	144.4 (54–765)	47.0 (6–136)
MediaSum		Long	Interview	1,481	1241.5 (80–3226)	80.1 (8–327)
MeetingBank			Meeting	1,327	998.3 (101–3384)	74.7 (9–292)

Table 11: Statistics of the SPICE training set, including the average word counts of input documents and reference summaries, with min–max ranges in parentheses. Documents with over 1K words are considered "long".

C Chosen Summary Detail

In the FeedSum (Song et al., 2025) training dataset, we select the highest-scoring summary with a faithfulness score above 0.8 (out of 1.0) as the chosen summary. We include documents containing between 100 and 4,000 tokens, as measured by the Llama3 tokenizer, and exclude cases where the summary is identical to the document to avoid copy bias. Detailed statistics are provided in Table 11. We use these chosen summaries across all experimental settings, including SFT and others.

D Rejected Summary Detail

We construct one-to-one rejected summaries for each baseline, aligning them with the document-chosen summary pairs constructed in Section 4.1. In section 6, we use Llama3.1-70B-Instruct, GPT-4o-mini, and Gemma3-12B-Instruct for both SYNFACT-EDIT and SPICE. The rejected summary examples for each baseline and SPICE, along with their factuality and preference scores, are provided in Table 44 and Table 45.

CLIFF (Cao & Wang, 2021). The Entity Swap strategy generates negative summaries by replacing named entities in the reference summary with randomly selected entities of the same type from the source document. Named entity recognition (NER) is applied to both the source and summary using spaCy to extract entities and their types. For each entity in the summary, one negative sample is created by substituting it with a same-type entity from the source. We implement the SwapEnt strategy by replacing all eligible named entities in the reference summary with randomly selected entities of the same type from the source, rather than swapping only one entity as in the original CLIFF implementation. To ensure that at least one substitution is made even when no matching entities are found, we incorporate a fallback mechanism that selects replacements from a predefined entity pool.

The Mask&Fill strategy generates negative summaries by replacing named entities in the reference summary with [MASK] tokens and using BART (Lewis et al., 2020) to fill in the masked spans. This simulates extrinsic factual errors by leveraging the generative ability of masked language models to produce fluent but potentially unfaithful content. Named entities are identified using spaCy, and for each entity, the reference summary is masked and decoded by a pre-trained BART model. Among the generated candidates, those containing entities not present in the source article or original reference are used as negative samples. We mask all named entities in the reference summary simultaneously rather than one at a time, and generate five candidate summaries using BART_large with top-p sampling. We set the top-p value to 0.9, and generate five return sequences. We select the first candidate that introduces a novel entity, and apply a fallback masking strategy based on randomly selected tokens when no named entities are detected.

Dataset	Rouge1	RougeL	Δ_{fact} (G-Eval)	Δ_{fact} (Log-prob.)
Original	0.883	0.856	0.36	49.3
Ours	0.908	0.882	0.15	45.2

Table 12: Comparison of original and reproduced SYNFACT-EDIT datasets in terms of factuality margins and ROUGE scores.

SYNFAC-EDIT (Mishra et al., 2024). We use the SYNFACT-EDIT prompt to generate a rejected summary based on the document and the chosen summary. Since the original prompt was designed for the clinical domain, we adapt it for all summarization domains while preserving the key points and overall structure, as shown in Table 42. However, when applying this prompt to generate rejected summaries, we observed that it often failed to induce sufficient hallucination, resulting in outputs that closely resembled the original chosen summary (see Table 44 and 45). To clarify why SYNFACT-EDIT exhibits this behavior and why it nevertheless remains a meaningful baseline for comparison, we provide a detailed analysis below:

–A. Faithful Reproduction and Analysis of SYNFACT-EDIT.

(1) *Inherently Small Faithfulness Margin.* To ensure our findings were not implementation artifacts, we analyzed the original SYNFACT-EDIT dataset¹ and found that their chosen–rejected pairs already exhibit very small faithfulness margins, indicating the method inherently produces minimal factual divergence. In our reproduction, we sampled 1,000 examples from each dataset and computed the Rouge score and factuality margin using two approaches: (1) prompting an external model for factuality scores (G-Eval with GPT-4o-mini) and (2) computing the log-probability margin using an external model (Llama 3.1-70B). As shown in the Table 12, our reproduction showed the same pattern with the original, and this behavior remained consistent under our experimental setup. This convergence across both the original and reproduced data strongly indicates that the small faithfulness margin is an inherent characteristic of SYNFACT-EDIT, rather than a flaw in our implementation.

(2) *Low Factuality Margin from Edit-Based Mechanism.* The small factuality margin is a direct consequence of SYNFACT-EDIT’s edit-based design. The method primarily performs localized rejection edits, such as adding, deleting, or swapping words, which inherently limit the degree of semantic deviation it can introduce. As also supported by prior works (Tang et al., 2022; Qiu et al., 2024), such operations indeed rarely produce the substantial meaning shifts required to generate informative hallucinations. Therefore, the low factuality margin reflects a fundamental limitation of the mechanism itself.

(3) *Prompt Revisions Are Not Sufficient.* Although we explored prompt variations within the intended edit-based paradigm (e.g., varying the number and balance of Add/Omit operations, relaxing edit constraints, adjusting the allowed word budget), the factuality margin remained low. Meaningful improvement would require abandoning the edit-based mechanism or introducing overly complex prompt changes that go beyond a revision and amount to creating a new method. Within reasonable revisions, we observed no improvement.

–B. Experimental Results Reinforce SYNFACT-EDIT’s Limitations.

First, SYNFACT-EDIT employs an entity-swap style similar to CLIFF-Entity. As shown in Table 5, both methods consistently produce small factuality margins, which directly explains their weaker performance in Table 4. Second, despite its extensive instructions, SYNFACT-EDIT does not reliably induce semantic-level hallucinations. Thus, while it remains an important baseline for contrast, its limitations stem from its mechanism, not our implementation.

¹<https://huggingface.co/datasets/bio-nlp-umass/SYNFACT-EDIT>

–C. SummLlama as the Most Recent and Strongest Baseline.

SummLlama is the most recent method and has already demonstrated superior performance over SYNFACT-EDIT within the paper. Therefore, it serves as a strong baseline for comparison, and outperforming SummLlama is a meaningful indicator of effectiveness. In all our experiments, we compare directly against SummLlama.

SummLlama (Song et al., 2025). We sample the rejected summaries from the FeedSum dataset by selecting the lowest-faithfulness summary that shares the same document as the chosen and satisfies $f(y_w; x) - f(y_l; x) \geq 0.2$.

E Training Detail

E.1 Training Configuration

Supervised Fine-tuning (SFT). We fine-tune the models using QLoRA (Dettmers et al., 2023) on a single NVIDIA H100 GPU. The model is trained for 3,000 steps with AdamW as the optimizer, using a batch size of 8, an initial learning rate of $1e-4$, and a weight decay of 0.05. We apply the best chosen summaries for training, which are used in all experiment configurations.

Direct Preference Optimization (DPO). Since the instruction model has already completed pretraining and instruction tuning, we directly perform DPO (Rafailov et al., 2023) optimization on it. Like SFT, we utilize QLoRA to fine-tune the model on a single NVIDIA H100 GPU. The model is optimized for 6,000 steps using AdamW with $\beta = 0.1$, a batch size of 8, an initial learning rate of $1e-6$, and a weight decay of 0.05.

Simple Preference Optimization (SimPO). We additionally train the models using SimPO (Meng et al., 2024). Similarly to DPO, we utilize QLoRA to fine-tune the model on a single NVIDIA H100 GPU. The model is optimized for 6,000 steps using AdamW with $\beta = 2.5$, $\gamma/\beta = 0.55$, a batch size of 8, an initial learning rate of $1e-6$, and a weight decay of 0.05.

Kahneman-Tversky Optimization (KTO). Same as DPO, we utilize QLoRA to fine-tune the model on a single NVIDIA H100 GPU using KTO (Ethayarajh et al., 2024). The model is optimized for 6,000 steps using AdamW with $\beta = 0.1$, a batch size of 8, an initial learning rate of $1e-6$, and a weight decay of 0.05.

E.2 Input and Output Format

Table 43 shows an example input with its corresponding chosen and rejected summaries used for DPO training. The prompt includes a summarization instruction and the input document. In this example, the chosen summary is sampled from the SummLlama dataset, and the rejected summary is generated using the SPICE framework. SimPO and KTO follow a similar setup, referring to the dataset format guide provided by Hugging Face.²

For SFT, the input–response format is nearly identical to that of DPO, except that each document is paired with a single reference summary, without distinguishing between chosen and rejected outputs. Instead, reference summaries are directly used to train the model through teacher forcing.

²https://huggingface.co/docs/trl/main/en/dataset_formats

F Summary Quality Evaluation

F.1 Significance Testing

We conducted an approximate randomization test with 10,000 iterations on 1,400 test samples to assess the statistical significance of the results reported in Tables such as 4, 9, 17, 16, 18 and 10. For each table, we compared the best-performing method (bolded) against all other methods. Values below 0.05 are assumed to be statistically significant.

F.2 Metric for Summary Quality

Faithfulness Score. The faithfulness score measures how accurately the content of a summary reflects factual information from the source document. Let $S = \{s_1, \dots, s_N\}$ be a summary composed of N sentences, where s_i indicates the i -th sentence, and $S_{\text{fact}} \subseteq S$ denotes the subset of sentences identified as factually correct. The score is defined as the proportion of factually accurate sentences within the summary, $\text{Faithfulness}(D, S) = |S_{\text{fact}}|/|S|$, which reflects the density of factual correctness at the sentence level.

Abtractiveness Score. The abtractiveness score evaluates the degree to which a summary introduces novel content beyond the exact phrasing of the source. Following standard approaches Liu & Lapata (2019); Song et al. (2023), let $n\text{-gram}_{D\text{-shared}}$ be the set of overlapping n -grams between the summary and the document, and $n\text{-gram}_{\text{summary}}$ represent the complete set of n -grams in the summary. The novelty ratio for n -grams is computed as $N_n = 1 - |n\text{-gram}_{D\text{-shared}}|/|n\text{-gram}_{\text{summary}}|$, and the final abtractiveness score is obtained by averaging the novelty ratios across 1-gram, 3-gram, and 5-gram levels: $\text{Abtractiveness}(D, S) = (N_1 + N_3 + N_5)/3$.

F.3 Automated Evaluator Details

We use FINESURE evaluated with the Llama3.3-70B-Instruct, and G-EVAL evaluated with GPT-4o-mini³. To encourage deterministic decoding, we set the temperature to zero. Tables 36 and 37 present the prompts used for FINESURE and G-EVAL, respectively. For ALIGNSCORE, we employ the RoBERTa-large in the official checkpoint with a batch size of 64, using the nli_sp evaluation setting. We found that all three metrics produced nearly identical values, with differences of less than 0.001 percentage points when rerun, suggesting stable and consistent outputs. Additionally, unlike G-EVAL, both FINESURE and ALIGNSCORE evaluate summary quality at the sentence level by aggregating binary labels assigned to individual sentences. These automated evaluators enable a more fine-grained and localized analysis of summary quality. Despite these more granular evaluations, the overall trends remain consistent across different methods, as demonstrated in Table 4.

These three automatic factuality evaluation metrics (FineSurE, G-Eval, and AlignScore) are all recent, verified, SOTA in summarization domain. They show strong agreement with human judgments in factuality evaluation, achieving a balanced accuracy up to 0.92 and a Pearson correlations coefficient over 0.84 (Song et al., 2024; Liu et al., 2023; Zha et al., 2023). Across the three evaluators, prior work reports that FineSurE improves correlation with human judgments over traditional metrics such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020) by average margins of 0.52 and 0.74, respectively. Similarly, G-Eval shows gains of 0.44 and 0.49, while AlignScore achieves improvements of 0.24 and 0.31 over the same baselines. Therefore, the metrics align with human judgments far more strongly than traditional metrics and are now widely used as practical substitutes for human evaluation in dataset construction, model assessment, and benchmarking for text summarization.

³Even with other backbone LLMs for FINESURE and G-EVAL, the experimental results are consistent. Please see Appendix G.9 for detailed results.

F.4 Human Evaluation Details

We conduct a human evaluation for the factuality verification task, following Lee et al. (2024b). In this task, annotators identify each sentence in a model summary as either factually consistent or inconsistent with the source document. Based on human annotation results, we compute faithfulness by calculating the proportion of factually accurate sentences, as described in Appendix F.

Due to the high cost of fine-grained human evaluation, we randomly sample 1400 document-summary pairs, specifically 100 (documents) $\times$ 14 (summarizers) in Table 4. We recruited Amazon Mechanical Turk (MTurk) workers residing in Australia (AU), Canada (CA), New Zealand (NZ), Great Britain (GB), and the United States (US). We allowed only annotators with an approval rating above 95% and at least 1,000 approved HITs to ensure the quality of annotations. We compensate annotators at a rate exceeding the U.S. minimum wage, with a total cost of about \$1,000 for all annotations. For these annotations, we achieved a sufficiently high Inter-Annotator Agreement (IAA) score of 0.53, as measured by Krippendorff’s α Krippendorff (2011).

F.4.1 Regarding Human Evaluation Scale

Compared to our baselines (Cao & Wang, 2021; Mishra et al., 2024; Song et al., 2025), our human evaluation was conducted at both a substantial scale and with a hierarchical structure. As shown in Table 11, we evaluated summaries from seven dataset domains spanning dialogue, non-dialogue, short-form, and long-form documents, with each summary containing roughly six sentences on average. Each sentence was annotated independently by three annotators, and the final summary-level factuality score was computed through majority voting over sentence-level judgments. This constitutes a hierarchical and fine-grained evaluation protocol: factuality is assessed at the sentence level and then aggregated to the summary level, mirroring the human annotation procedure used in FineSurE. We believe that this setup satisfies both large-scale coverage and hierarchical evaluation, ensuring that our results are both reliable and comprehensive.

Method	Human Eval Size	# of Datasets	# of Annotators	Granularity
SYNFAC-EDIT	100	2	2	summary-level
SYNFAC-EDIT	10	1	3	sentence-level
SummLlama	70	7	3	sentence-level
SPICE (Ours)	100	7	3	sentence-level

Table 13: Comparison of human-evaluation settings across methods.

In particular, Table 13 demonstrates that our human evaluation covers a broader and more diverse range of data compared to our three baselines. Here, Human Eval Size indicates the number of summaries evaluated by humans within a single test set. SPICE evaluates on the largest number of datasets (7), with 100 human-evaluated summaries per dataset, which is higher than any other method. Each summary is assessed by three independent annotators and evaluated at the sentence level, offering a more fine-grained analysis. Taken together, these aspects demonstrate that SPICE provides the most quantitative and fine-grained human evaluation among the baselines.

F.4.2 Regarding IAA.

In the context of sentence-level factuality and faithfulness evaluation, $\alpha = 0.53$ is considered high. Prior work using comparable sentence-level factuality annotation settings reports the inter-annotator agreement for factuality on average as 0.40 in CLIFF (Cao & Wang, 2021), 0.49 in BERTMULTI (Zhu et al., 2023), 0.49 in SYNFAC-EDIT (Mishra et al., 2024), 0.38 in TofuEval (Tang et al., 2024), and 0.58 in MSumBench (Min et al., 2025). Therefore, our score of 0.53 lies at the upper end of this range and indicates strong inter-annotator consistency for this evaluation setting.

Method	Backbone: Llama3.2-1B-Instruct				Backbone: Gemma3-1B-Instruct			
	FineSurE	G-Eval	AlignScore	Abs.	FineSurE	G-Eval	AlignScore	Abs.
Zero-Shot	0.551	7.078	0.679	0.685	0.586	6.645	0.624	0.225
SFT	0.695	7.627	0.815	0.521	0.357	4.212	0.373	0.303
CLIFF-Entity	0.527	7.001	0.685	0.701	0.552	6.596	0.626	0.257
CLIFF-Mask	0.534	7.284	0.671	0.708	0.567	6.552	0.633	0.323
SYNFAC-EDIT	0.520	6.954	0.657	0.741	0.591	6.646	0.630	0.257
SummLlama	0.594	7.787	0.719	0.695	0.589	6.702	0.629	0.241
SPICE (Ours)	0.638	8.098	0.787	0.680	0.630	6.877	0.639	0.243

Table 14: Faithfulness and abtractiveness comparison on Llama3.2-1B-Instruct and Gemma3-1B-Instruct with DPO. Bold values indicate the highest score in each column.

G Additional Experiments

G.1 Performance on Smaller Models

Table 14 compares performance on smaller models, Llama3.2-1B-Instruct and Gemma3-1B-Instruct. First, on Gemma, our proposed method (SPICE) achieves the highest overall performance across metrics. However, the abtractiveness of Gemma is considerably lower than Llama’s (0.685 vs. 0.225), indicating limited capacity for generating genuinely abtractive summaries. Similar observations are reported in Song et al. (2025), highlighting that the small size of Gemma is not appropriate for abtractive summarization with long documents.

Second, on Llama, the SFT model shows high factuality, but with significantly reduced abtractiveness (0.685 vs. 0.521). Specifically, 214 out of 1,400 SFT summaries are **exact copies of the input**. This reflects a trade-off in summarization, where extractive summaries often exhibit higher factuality but lower abtractiveness (Choi et al., 2024). Since this issue, we did not report significance test results for the best scores in this section. In contrast, SPICE leverages DPO to significantly improve factuality (0.551 vs. 0.638) while preserving abtractiveness (0.685 vs. 0.680), generating only 11 summaries exactly the same as a document out of 1,400, thus achieving a better balance between the two objectives. These results highlight **SPICE enhances faithfulness without compromising abtractiveness, even on smaller models**.

G.2 Controlling Three Factors.

To isolate the effect of each factor (hallucination level, summary length, and prompt complexity), we carefully controlled the remaining conditions during data construction. For hallucination level, we varied the rejected summaries across Severe, Mild, and None conditions while maintaining similar summary lengths and prompt structures. As shown in Table 15, the token length differences between chosen and rejected summaries remain minimal (-4.65 to +3.38 tokens on average), and no additional instructions were introduced beyond those needed to adjust hallucination. For summary length, we varied between Shorter, Similar, and Longer setups while using the same prompts, all of which tended to induce severe hallucinations. The resulting factuality margins were nearly identical (6.89–7.06), confirming that only length was varied. Finally, for prompt complexity, we compared Simple vs. Complex prompts, both designed to induce severe hallucinations, and explicitly controlled summary length by requiring equal output lengths. Token length differences and factuality margins were comparable across the two conditions, indicating that only prompt complexity was varied. Overall, **these controls ensure that the observed differences can be attributed to the intended factors**.

Factor	Fact. Margin	Token Diff	Prompt
Severe	7.81 _(0.37)	-4.65 _(4.51)	Simple
Mild	4.06 _(1.84)	-4.12 _(4.89)	Simple
None	0.88 _(0.98)	+3.38 _(4.94)	Simple
Longer	7.06 _(0.77)	-48.16 _(26.85)	Simple
Shorter	6.92 _(0.53)	+45.95 _(24.62)	Simple
Complex	7.12 _(0.87)	+5.31 _(4.30)	Complex
Simple	6.89 _(0.48)	+3.79 _(4.41)	Simple

Table 15: Controlled analyses of hallucination level, summary length, and prompt complexity. Values are reported as mean with standard deviation in parentheses.

Setting	Generator: Llama3.1-70B-Inst.			Generator: GPT-4o-mini			Generator: Gemma3-12B-Inst.		
	FineSurE	G-Eval	AlignScore	FineSurE	G-Eval	AlignScore	FineSurE	G-Eval	AlignScore
Default	0.857 _(0.02)	8.743 _(0.02)	0.823 _(0.01)	0.855 _(0.02)	8.719 _(0.02)	0.822 _(0.02)	0.858 _(0.02)	8.793 _(0.02)	0.818 _(0.01)
Severe	0.856 _(0.02)	8.688 _(0.01)	0.812 _(0.01)	0.840 _(0.01)	8.596 _(0.00)	0.810 _(0.01)	0.851 _(0.01)	8.610 _(0.00)	0.812 _(0.00)
Mild	0.828 _(0.00)	8.265 _(0.00)	0.786 _(0.00)	0.811 _(0.00)	8.142 _(0.00)	0.782 _(0.00)	0.818 _(0.00)	8.204 _(0.00)	0.790 _(0.00)
None	0.795 _(0.00)	7.537 _(0.00)	0.748 _(0.00)	0.784 _(0.00)	7.488 _(0.00)	0.736 _(0.00)	0.788 _(0.00)	7.505 _(0.00)	0.752 _(0.00)
Longer	0.854 _(0.02)	8.645 _(0.01)	0.819 _(0.01)	0.847 _(0.01)	8.534 _(0.00)	0.806 _(0.00)	0.846 _(0.01)	8.601 _(0.00)	0.810 _(0.00)
Shorter	0.859 _(0.02)	8.712 _(0.02)	0.821 _(0.01)	0.854 _(0.02)	8.695 _(0.01)	0.815 _(0.01)	0.854 _(0.01)	8.706 _(0.01)	0.818 _(0.01)
Complex	0.850 _(0.01)	8.682 _(0.00)	0.810 _(0.01)	0.843 _(0.01)	8.588 _(0.00)	0.806 _(0.01)	0.841 _(0.01)	8.572 _(0.00)	0.804 _(0.00)
Simple	0.885	9.037	0.854	0.881	9.012	0.855	0.883	9.098	0.851

Table 16: Summary faithfulness, measured by FineSurE, G-Eval, and AlignScore, after DPO training using Llama3.2-3B on rejected summaries generated under eight prompt designs.

G.3 Impact of Input Source

Table 17 shows the impact of input settings on model alignment. Although our method is originally designed to generate rejected summaries based on both the document and the reference summary, we conduct an ablation study to verify whether using both inputs is indeed necessary. To this end, we test (1) document-only and (2) summary-only variants, with prompt examples shown in Table 40 and Table 41, respectively. Providing both the document and summary as input yields the best performance across all metrics. Using only the document or only the summary leads to performance degradation, indicating that **providing both the source document and reference summary is essential for improving factuality**.

Method	FineSurE	GEval	AlignScore
Zero-Shot	0.798 _(0.00)	7.519 _(0.00)	0.751 _(0.00)
Doc-Only	0.855 _(0.02)	8.756 _(0.02)	0.828 _(0.02)
Summary-Only	0.864 _(0.03)	8.845 _(0.02)	0.833 _(0.02)
Doc+Summary	0.885	9.037	0.854

Table 17: Faithfulness comparison of three input settings on Llama3.2-3B-Instruct with DPO: document with summary, document-only, and summary-only.

G.4 Impact of Quality Control

Table 16 reports the effects of hallucination level and length control on rejected summary generation across three models. This demonstrates consistent performance patterns across diverse backbone architectures and model sizes: (1) We observe that higher hallucination levels correspond to improved alignment performance, likely due to increased factuality margins. (2) Likewise, rejected summaries with lengths similar to the chosen summaries consistently yield better alignment than those that are longer or shorter, as they result in smaller preference margins. These results highlight that **high hallucination and length similarity as critical factors for improving factuality, consistently observed across LLMs of varying sizes and types**.

G.5 Impact of Rejection Models

We examine the effect of varying the model used to generate rejected summaries with SPICE. Hence, we replace Llama3.1-70B-Instruct with two other alternatives, including GPT-4o-mini and Gemma3-12B-Instruct. Table 16 compares three rejection models that differ in architecture, with their summary quality evaluated using three metrics.

Despite these differences, all variants produce high-quality rejected summaries with comparable factuality scores. For instance, the stronger model (Llama3.1-70B) yields a FineSurE score of 0.885, G-Eval score of 9.037 and AlignScore of 0.854, while the weaker model (Gemma3-12B) achieves a similarly FineSurE score of 0.883, G-Eval score of 9.098 and AlignScore of 0.851. We attribute this to the simplicity of our prompt design, which enables even smaller and weaker models to reliably generate effective rejected summaries. This finding suggests that **SPICE generalizes well across LLM types and scales, offering lower-cost rejection generation without compromising alignment performance**.

G.6 Impact of Chosen Quality

To examine the impact of chosen summary quality on alignment performance, we utilize low-quality (low factuality) reference summaries from the FeedSum (Song et al., 2025) dataset. As shown in Table 18, compared to the results in Table 4, all methods show performance drops due to the low-quality chosen summaries. Nonetheless, our proposed method still achieves the best overall results among all rejection strategies, **highlighting its effectiveness even with the low-quality chosen summaries.**

Method	FineSurE	G-Eval	AlignScore
Zero-Shot	0.798 _(0.00)	7.519 _(0.00)	0.751 _(0.00)
SFT ^H	0.698 _(0.00)	6.810 _(0.00)	0.656 _(0.00)
CLIFF-Entity ^H	0.757 _(0.00)	7.313 _(0.00)	0.740 _(0.00)
CLIFF-Mask ^H	0.784 _(0.00)	7.478 _(0.00)	0.752 _(0.00)
SYNFAC-EDIT ^H	0.778 _(0.00)	7.542 _(0.00)	0.754 _(0.00)
SummLlama ^H	0.790 _(0.00)	7.664 _(0.00)	0.756 _(0.00)
SPICE ^H (Ours)	0.841	8.882	0.833

Table 18: Faithfulness performance of Llama3.2-3B-Instruct trained with DPO using low-quality human-written chosen summaries (denoted as ^H) and corresponding rejected summaries.

G.7 Data Generation Latency

Table 19 reports the latency and cost of generating rejections for 1,000 randomly sampled summaries from the Feedsum training set using GPT-4o-mini, along with the resulting average faithfulness scores in Table 4. SPICE (Ours) is approximately 2.1× faster than SummLlama and 3.9× faster than SYNFAC-EDIT. It is also 1.75× and 1.49× more cost-efficient than SummLlama and SYNFAC-EDIT, respectively. In terms of faithfulness, SPICE achieves the highest score, with gains of +4.1%p over SummLlama and +7.3%p over SYNFAC-EDIT. Overall, these results demonstrate that **SPICE delivers superior faithfulness while being substantially faster and more cost-efficient than existing rejection construction methods.**

Method	Time (sec)	Cost (\$)	Faithfulness
SYNFAC-EDIT	21,137	1.33	0.822
SummLlama	11,490	1.56	0.854
SPICE (Ours)	5,425	0.89	0.895

Table 19: Comparison of efficiency and faithfulness across methods.

G.8 Fine-grained Analysis of Prompt Complexity

In Section 4.2.3, our initial comparison between Simple and Complex prompts is a bit coarse-grained, as this binary contrast does not reveal which aspects of prompt complexity actually drive the observed effects. To obtain a more precise understanding, (1) we conduct a finer-grained analysis that isolates the role of each edit instruction, (2) evaluate how alignment potential evolves as instructions are added incrementally, (3) examine whether partially relaxing the complex prompt preserves its benefits, and (4) analyze the resulting distribution of hallucination types. These investigations allow us to identify the specific mechanisms through which prompt complexity shapes the quality of rejected summaries. In Table 35, the Complex prompt consists of eight edit instructions. Since length preservation (Instructions 3–4) is a core component of SPICE, we keep this constraint fixed by default. We then group the remaining instructions into five categories:

- (1) Heavy hallucination induction (Instructions 1–2)
- (2) Numerical alterations (Instruction 5)
- (3) Named-entity substitutions (Instruction 6)
- (4) Causal relation edits (Instruction 7)
- (5) Preserving grammar and fluency (Instruction 8)

G.8.1 Effect of Each Instruction

We perform ablations over these five categories while keeping length preservation fixed. The results are shown in Table 20. For the heavy hallucination category, the factuality margin increases, however, the preference margin rises by a similar degree, making the rejected summaries easier for the model to distinguish. As a consequence, the overall alignment potential (55.0) becomes lower than that of SPICE (69.9). For the remaining instruction categories(2-5), some variants reduce the preference margin (making the rejection harder to distinguish), but this benefit is offset by a simultaneous drop in factuality margin. As a result, **all ablated variants yield a lower alignment potential compared to the original SPICE.**

Method	Δ_{fact}	Δ_{pref}	$ \Delta_{\text{fact}} - \Delta_{\text{pref}} $
SPICE	102.0	32.1	69.9
SPICE + (1)	120.3	65.3	55.0
SPICE + (2)	80.3	31.6	48.7
SPICE + (3)	86.1	33.2	52.9
SPICE + (4)	79.5	25.4	54.1
SPICE + (5)	76.7	38.3	38.4

Table 20: Average factuality and preference margins computed on datasets generated from the Complex prompt with each instruction. Llama3.1-70B-Inst (evaluator) and Llama3.1-8B-Inst (policy) are used to compute Δ_{fact} and Δ_{pref} , respectively.

G.8.2 Effect of Increasing Prompt Complexity

We extend the default SPICE setting by incrementally adding each instruction category to provide intermediate comparisons, ranging from simple to more complex prompts. The final configuration is exactly the same as the Complex. As shown in the Table 21, while the factuality margin remains roughly constant or increases slightly as more instructions are added, the preference margin grows substantially. This makes the rejected summaries progressively easier for the model to distinguish, ultimately reducing alignment potential. These results indicate that increasing the number of instructions consistently enlarges the preference margin, which is undesirable. This supports our design choice in SPICE: **allowing the model to introduce errors freely, without enforcing specific edit types, produces harder and more informative negative examples.**

Method	Δ_{fact}	Δ_{pref}	$ \Delta_{\text{fact}} - \Delta_{\text{pref}} $
SPICE	102.0	32.1	69.9
SPICE + (1)	120.3	65.3	55.0
SPICE + (1-2)	118.6	68.4	50.2
SPICE + (1-3)	119.6	72.6	47.0
SPICE + (1-4)	122.7	78.9	43.8
SPICE + (1-5)	126.5	85.3	41.2

Table 21: Average factuality and preference margins computed on datasets generated from the Complex prompt while incrementally adding each instruction. Llama3.1-70B-Inst (evaluator) and Llama3.1-8B-Inst (policy) are used to compute Δ_{fact} and Δ_{pref} , respectively.

G.8.3 Effect of Relaxing Complex Instructions

We extend the Complex prompt by adding the following instruction to increase the freedom with which hallucinations can be introduced: *"The following instructions are just examples and do not need to be followed entirely; you may apply them selectively and flexibly as long as the overall summary remains factually inconsistent."* and denote this variant as Complex-Flex. As shown in the Table 22, introducing this flexibility reduces the preference margin more than the factuality margin, resulting in an improvement in alignment potential compared to the original Complex setting. However, the overall alignment potential remains lower than that achieved by SPICE, which allows **hallucinations to emerge more naturally without enforcing specific edit types**.

We further modify the SPICE+(1) setting by removing the instruction *"nearly all words should be edited"*, and denote this variant as SPICE+(1A). As expected, Table 23 shows that reducing the required extent of edits results in less aggressive hallucinations, which in turn leads to a smaller preference margin and therefore a higher alignment potential. This indicates that the **extreme instruction contributed to unnatural phrasing rather than generating informative hallucinations**. However, the alignment potential still fell short of that achieved by SPICE.

G.8.4 Distribution of Hallucination Categories

We defined four error types from the Complex prompt and modified the FineSurE evaluation prompt 36 to detect these categories by replacing the original error labels. This allows us to compare the distributions of error categories between SPICE-generated data and Complex-generated data. Table 24 shows a clear contrast between the two settings. Since neither SPICE nor Complex produces grammatical or fluency errors, we omit this category from the results. Complex generates hallucinations that closely follow the instructed categories, with numerical, entity, and causal errors together accounting for about 70% of all detected cases. In contrast, SPICE produces a broader and more diverse distribution of hallucination types, with "other" errors alone comprising 56.5% of instances. This indicates that **SPICE induces a wider and more natural spectrum of factual inconsistencies, rather than concentrating on predefined error types, which likely contributes to its stronger alignment potential**. The error categories follow:

- Numerical errors: Incorrectly modifying numerical values (e.g., dates, quantities, percentages) in contradiction to the source document.
- Entity errors: Replacing people, places, or organizations with incorrect but plausible alternatives.
- Causal-relation errors: Altering cause–effect relations to assert statements that contradict the document.
- Grammatical and fluency errors: Reducing grammatical correctness or fluency, leading to unnatural or awkward phrasing.
- Other errors: Factuality errors not included in the categories above.

Method	Δ_{fact}	Δ_{pref}	$ \Delta_{\text{fact}} - \Delta_{\text{pref}} $
SPICE	102.0	32.1	69.9
Complex	126.5	85.3	41.2
Complex-Flex	114.2	57.6	56.6

Table 22: Average factuality and preference margins computed on datasets generated from the Complex prompt and its flexible variant. Llama3.1-70B-Inst (evaluator) and Llama3.1-8B-Inst (policy) are used to compute Δ_{fact} and Δ_{pref} , respectively.

Method	Δ_{fact}	Δ_{pref}	$ \Delta_{\text{fact}} - \Delta_{\text{pref}} $
SPICE	102.0	32.1	69.9
SPICE + (1)	120.3	65.3	55.0
SPICE + (1A)	108.7	49.4	59.3

Table 23: Average factuality and preference margins computed on datasets generated from the SPICE + (1) and SPICE + (1A). Llama3.1-70B-Inst (evaluator) and Llama3.1-8B-Inst (policy) are used to compute Δ_{fact} and Δ_{pref} , respectively.

Method	Num.E	Ent.E	Rel.E	Other.E
SPICE	9.8%	14.0%	19.7%	56.5%
Complex	12.4%	35.7%	21.5%	30.4%

Table 24: Distribution of hallucination error categories for rejected summaries generated by SPICE and Complex. Numerical (Num.E), entity (Ent.E), relation (Rel.E), and other factual errors (Other.E) are detected using the modified FineSurE.

Method	Backbone: Llama3.2-3B-Instruct						Backbone: Llama3.1-8B-Instruct					
	FineSurE			G-Eval			FineSurE			G-Eval		
	GPT	Llama	Qwen	GPT	Llama	Qwen	GPT	Llama	Qwen	GPT	Llama	Qwen
SummLlama	0.788 _(0.00)	0.843 _(0.01)	0.766 _(0.00)	8.473 _(0.00)	8.102 _(0.00)	8.422 _(0.00)	0.865 _(0.00)	0.908 _(0.02)	0.854 _(0.00)	9.038 _(0.00)	8.796 _(0.00)	9.182 _(0.00)
SPICE	0.833	0.885	0.811	9.037	8.629	8.996	0.906	0.934	0.898	9.438	9.182	9.568

Table 25: Faithfulness comparison of SummLlama and SPICE using two evaluation metrics (FineSurE and G-Eval) and three evaluator model families (GPT, Llama, Qwen).

G.9 Impact of Automatic Evaluation Models

In our automatic evaluations, each of our metrics relies on a single evaluator: G-Eval uses GPT-4o-mini, FineSurE uses Llama3.3-70B-Instruct, and AlignScore is based on a dedicated pretrained RoBERTa-Large model. To examine whether our conclusions are sensitive to the choice of evaluator model family for each metric, we conducted additional evaluation for Table 4 using two evaluation metrics (FineSurE, G-EVAL) with three evaluator model families (GPT-4o-mini, Llama3.3-70B-Instruct, Qwen3-Next-80B-A3B-Instruct) for each metric. As AlignScore is tightly coupled to its original pretrained architecture, it is not designed to be replaced with alternative model families. In this analysis, we focus on SummLlama, since it consistently outperformed other baselines. The results in the Table 25 show that **SPICE consistently outperforms SummLlama across two evaluation metrics and three evaluator families, showing the robustness of the automatic evaluation results.**

G.10 Inefficiency of Adversarial Sampling

The table 26 presents the DPO training results using the adversarially sampled dataset. We observe that its performance is comparable to only 1/16 of what SPICE achieves, as reported in Table 6. This implies that adversarial sampling would require generating over $200\times$ more rejected candidates ($13 \text{ candidates} \times 16 \times \text{gap}$) to reach performance comparable to SPICE. This is extremely costly in practice. These findings suggest that **achieving strong performance through adversarial sampling demands an extremely large candidate pool, which is not feasible in realistic settings.** In contrast, SPICE provides an efficient and effective alternative: it produces high-quality rejected summaries without requiring massive cost, making it far more practical for real-world preference optimization. For these reasons, we did not adopt adversarial sampling as part of our main approach.

Method	FineSurE	G-Eval	AlignScore
Adv. Sampling	0.848	8.551	0.818
SPICE ^{1/16}	0.846	8.503	0.812
SPICE	0.885	9.037	0.854

Table 26: Faithfulness comparison of DPO-trained models between adversarial sampling(Adv. Sampling) and SPICE using Llama3.2-3B-Inst.

Model Name	Checkpoints
Rejected Summary Generation	
Bart-Large	facebook/bart-large-cnn
Llama3.1-70B-Inst.	meta-llama/Llama3.1-70B-Instruct
GPT-4o-mini	gpt-4o-mini-2024-07-18 (OpenAI)
Gemma3-12B-Inst.	google/gemma-3-12b-it
Model Training	
Llama3.2-1B-Inst.	meta-llama/Llama3.2-1B-Instruct
Llama3.2-3B-Inst.	meta-llama/Llama3.2-3B-Instruct
Llama3.1-8B-Inst.	meta-llama/Llama3.1-8B-Instruct
Gemma3-1B-Inst.	google/gemma-3-1b-it
Summary Evaluation	
Llama3.3-70B-Inst.	meta-llama/Llama3.3-70B-Instruct
GPT-4o-mini	gpt-4o-mini-2024-07-18 (OpenAI)
Roberta-Large	yzha/AlignScore-large

Table 27: Checkpoints of 4 rejected summary generators, 4 training models, and 3 automated summary evaluators used in our experiments. For open-source models, we use publicly available checkpoints from HuggingFace. For proprietary models, we utilize paid API services provided by OpenAI.

User Prompt
You are given a document and a reference summary. Your task is to generate a factually inconsistent summary based on the provided document and a reference summary. Document: {Document} Reference Summary: {Reference Summary} Your answer MUST be in JSON format. The dictionary key should be hallucinated_summary as a string.

Table 28: **Default prompt** for generating a negative summary.

User Prompt
You are given a text. Your task is to paraphrase the text while preserving its original meaning. Ensure that the generated summary has the same length as the reference summary. Text: {Text} Your answer MUST be in JSON format. The dictionary key should be paraphrased_text as a string.

Table 29: Prompt for generating a **no hallucination** summary with similar length.

User Prompt
You are given a document and a reference summary. Your task is to generate a factually inconsistent summary based on the provided document and a reference summary. Ensure that the hallucinated summary includes factually incorrect information compared to the document and reference summary by following the instruction below. <Instruction> 1. About 10% of the summary should be edited with hallucinated content. 2. The summary is almost entirely accurate, containing very few hallucinated sentences and almost no errors. A few words should be edited. 3. Do not add additional sentences to generate hallucinations. 4. Ensure that the generated summary has the same length as the reference summary. Document: {Document} Reference Summary: {Reference Summary} Your answer MUST be in JSON format. The dictionary key should be hallucinated_summary as a string.

Table 30: Prompt for generating a **mildly hallucinated** summary with similar length.

User Prompt
You are given a document and a reference summary. Your task is to generate a factually inconsistent summary based on the provided document and a reference summary. Ensure that the hallucinated summary includes factually incorrect information compared to the document and reference summary by following the instruction below. <Instruction> 1. About 90-100% of the summary should be edited with hallucinated content. 2. The summary is almost entirely hallucinated, containing very little accurate information and a large number of errors. Nearly all words should be edited. 3. Do not add additional sentences to generate hallucinations. 4. Ensure that the generated summary has the same length as the reference summary. Document: {Document} Reference Summary: {Reference Summary} Your answer MUST be in JSON format. The dictionary key should be hallucinated_summary as a string.

Table 31: Prompt for generating a **severely hallucinated** summary with similar length.

User Prompt
You are given a document and a reference summary. Your task is to generate a factually inconsistent summary based on the provided document and a reference summary. Ensure that the generated summary is shorter than the reference summary. Document: {Document} Reference Summary: {Reference Summary} Your answer MUST be in JSON format. The dictionary key should be hallucinated_summary as a string.

Table 32: Prompt for generating a highly hallucinated summary with a **shorter length**.

User Prompt
You are given a document and a reference summary. Your task is to generate a factually inconsistent summary based on the provided document and a reference summary. Ensure that the generated summary has the same length as the reference summary. Document: {Document} Reference Summary: {Reference Summary} Your answer MUST be in JSON format. The dictionary key should be hallucinated_summary as a string.

Table 33: Prompt for generating a highly hallucinated summary with **similar length**.

User Prompt
You are given a document and a reference summary. Your task is to generate a factually inconsistent summary based on the provided document and a reference summary. Ensure that the generated summary is longer than the reference summary. Document: {Document} Reference Summary: {Reference Summary} Your answer MUST be in JSON format. The dictionary key should be hallucinated_summary as a string.

Table 34: Prompt for generating a highly hallucinated summary with a **longer length**.

User Prompt
You are given a document and a reference summary. Your task is to generate a factually inconsistent summary based on the provided document and a reference summary. Ensure that the hallucinated summary includes factually incorrect information compared to the document and reference summary by following the instructions below. <Instruction> 1. About 90-100% of the summary should be edited with hallucinated content. 2. The summary is almost entirely hallucinated, containing very little accurate information and a large number of errors. Nearly all words should be edited. 3. Do not add additional sentences to generate hallucinations. 4. Ensure that the generated summary has the same length as the reference summary. 5. Include at least one numerical value that contradicts the document (e.g., wrong year, quantity, or percentage). 6. Replace named entities (people, places, organizations) with incorrect but realistic alternatives. 7. Add contradictory causal relations (e.g., claim X caused Y, even if the document says otherwise). 8. Maintain grammatical fluency and readability so that the hallucinations are not trivially detectable by poor language quality. Document: {Document} Reference Summary: {Reference Summary} Your answer MUST be in JSON format. The dictionary key should be hallucinated_summary as a string.

Table 35: **Complex prompt** with multiple constraints for generating a highly hallucinated summary of similar length.

User Prompt

You will receive a document followed by a corresponding summary.
Your task is to assess the factuality of each summary sentence across nine categories:

- * no error: the statement aligns explicitly with the content of the document and is factually consistent with it.
- * out-of-context error: the statement contains information not present in the document.
- * entity error: the primary arguments (or their attributes) of the predicate are wrong.
- * predicate error: the predicate in the summary statement is inconsistent with the document.
- * circumstantial error: the additional information (like location or time) specifying the circumstance around a predicate is wrong.
- * grammatical error: the grammar of the sentence is so wrong that it becomes meaningless.
- * coreference error: a pronoun or reference with wrong or non-existing antecedent.
- * linking error: error in how multiple statements are linked together in the discourse (for example temporal ordering or causal link).
- * other error: the statement contains any factuality error which is not defined here.

Instruction:

First, compare each summary sentence with the document.

Second, provide a single sentence explaining which factuality error the sentence has.

Third, answer the classified error category for each sentence in the summary.

Provide your answer in JSON format. The answer should be a list of dictionaries whose keys are "sentence", "reason", and "category":

```
[{"sentence": "first sentence", "reason": "your reason", "category": "no error", "sentence": "second sentence", "reason": "your reason", "category": "out-of-context error", "sentence": "third sentence", "reason": "your reason", "category": "entity error"},]
```

Document: {Document}

Summary with {# of sentences} sentences:
{sentences}

JSON Output:

Table 36: Prompt of the FineSurE (Song et al., 2024) for factuality.

User Prompt
You will then be given one summary written for the source text. Your task is to rate the summary on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. Evaluation Criteria: Factual Consistency (1-10): The factual alignment between the summary and the source text. A factually consistent summary contains only statements that are entailed by the source text. Evaluation Steps: 1. Read the article carefully and identify the main facts and details it presents. 2. Read the summary and compare it to the source text. Check if the summary contains any factual errors that are not supported by the source text. 3. Assign a score for factual consistency on a scale of 1 to 10, where 1 is the lowest and 10 is the highest based on the Evaluation Criteria. Source Text: {Source Text} Summary: {Summary} Factual Consistency (scores ONLY):

Table 37: Prompt of the modified G-Eval (Liu et al., 2023) for factuality. We adjusted the score range to 1–5, 1–9, or 1–10 depending on the task. Specifically, a 1–5 scale is used in Figure 1 and Figure 3 for visualization clarity, while Table 1 adopted a 1–9 scale to better separate three levels of quality. In the experimental results (Section 6), we used a 1–10 scale for consistency with other evaluation scores on a tens-place scale.

User Prompt
You will be given one summary written for an article. Your task is to rate the summary on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. Evaluation Criteria: Coherence (1–10) – the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby “the summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to a coherent body of information about a topic.” Evaluation Steps: 1. Read the article carefully and identify the main topic and key points. 2. Read the summary and compare it to the article. Check if the summary covers the main topic and key points of the article, and if it presents them in a clear and logical order. 3. Assign a score for coherence on a scale of 1 to 10, where 1 is the lowest and 10 is the highest based on the Evaluation Criteria. Example: Source Text:{Document} Summary:{Summary} Coherency (scores ONLY):

Table 38: Prompt of the modified G-Eval (Liu et al., 2023) for coherency. We adjusted the score range to 1–10.

User Prompt
You will be given one summary written for an article. Your task is to rate the summary on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. Evaluation Criteria: Relevance (1–10) – selection of important content from the source. The summary should include only important information from the source document. Annotators were instructed to penalize summaries which contained redundancies and excess information. Evaluation Steps: 1. Read the summary and the source document carefully. 2. Compare the summary to the source document and identify the main points of the article. 3. Assess how well the summary covers the main points of the article, and how much irrelevant or redundant information it contains. 4. Assign a relevance score from 1 to 10. Example: Source Text:[Document] Summary:[Summary] Relevance (scores ONLY):

Table 39: Prompt of the modified G-Eval (Liu et al., 2023) for relevance. We adjusted the score range to 1–10.

User Prompt
You are given a document. Your task is to generate a factually inconsistent summary based on the provided document. Document: {Document} Your answer MUST be in JSON format. The dictionary key should be hallucinated_summary as a string.

Table 40: Prompt for generating a negative summary based on the document only.

User Prompt
You are given a reference summary. Your task is to generate a factually inconsistent summary based on the provided reference summary. Ensure that the generated summary has the same length as the reference summary. Reference Summary: {Reference Summary} Your answer MUST be in JSON format. The dictionary key should be hallucinated_summary as a string.

Table 41: Prompt for generating a negative summary based on the reference summary only.

User Prompt
»»»» Instruction »»»» You are a writing assistant who is in edit mode. You are tasked with generating a hallucinated summary based on the provided article and a reference summary for the article. The goal is to edit the reference summary to generate a hallucinated summary that sounds plausible but includes edits introduced through an edit operation, which can be one of the following: Add Operation: Intentionally add contextually important words from the source that are not necessary for a basic understanding but carry implications for accurate interpretation or decision-making. Omit Operation: Intentionally omit contextually important words in the reference summary that are necessary for accurate interpretation or decision-making. For these operations, focus on words that, if missing or incorrect in the hallucinated summary, could lead to significant misunderstandings or incorrect conclusions. Maintain coherence while deliberately excluding essential terms. The hallucinated summary should be concise, contain no more than five extra words compared to the reference summary, and have an equal number of Add and Omit operations. Steps for generating the hallucinated summary: Step 1: List the proposed edit operations to introduce hallucination on the reference summary. Step 2: Use the proposed edit operations to edit the reference summary. »»»» Output Format »»»» The output format is: Numbered List hallucination edits made: Edit 1, Edit 2, Edit 3 ... Hallucinated Summary: »»»» Follow the above Instructions, Hallucination Method and Output Format »»»» Now, let's start. Generate the hallucinated summary: Article: {Article} Reference Summary: {Reference Summary}

Table 42: Modified SYNFACT-EDIT (Mishra et al., 2024) prompt for generating rejected summaries, adapted from its original use in clinical summarization to be applicable across diverse summarization domains. The core logic and structure are preserved with minimal modifications.

Input	< begin_of_text >< start_header_id >user< end_header_id > Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction: Please factually summarize the input document. ### Input: (CNN) – On Tuesday, new Apple CEO Tim Cook is expected to help unveil Apple’s latest iPhone at its first big event since co-founder Steve Jobs stepped aside in August. Cook will no doubt be scrutinized for how he fulfills his new role Tuesday as chief pitchman for Apple’s products. And the new iPhone, expected to go on sale sometime in mid-October, will almost certainly draw lines of shoppers outside Apple’s stores. Then the question becomes: Can Apple make enough iPhones fast enough? As demand for its gadgets continues to skyrocket, keeping Apple growing at its amazing pace will be a key hurdle for Cook’s regime. The consumer shift toward cheaper, portable computers – and Apple’s huge success there with the iPhone and iPad – means creating and selling an ever-increasing number of devices each year. Right now, for example, Apple will have to ramp up production of millions of new iPhones, and possibly new iPods, for the holidays. After that there will be more iPads, Macs, and other products. It’s important that Apple’s new devices be stylish and technically impressive. But it’s equally important that they be in stock when people want to buy them, something that Apple hasn’t always been able to deliver. Only recently, for example, has the iPad 2 supply been able to meet demand. And as time goes on, thanks in part to new markets such as China, demand for each gadget will only grow. It’s not enough for Cook and his team to keep coming up with exciting new products. They also must continue Apple’s logistics revolution, so that consumers don’t have to wait forever to buy new Apple devices – or turn instead to the competition. Consider the total number of devices Apple must now have built and shipped per year. In 2011, including all iPhones, iPods, iPads, and Macs, Apple could end up shipping more than 175 million gadgets – some 50 million more than it did last year, representing roughly 40% growth. Next year, Apple is expected to ship an all-time high of around 215 million devices, or another 40 million more than this year. And so on. It wouldn’t be insane for Apple to have to ship 500 million units per year by the end of the decade. So Cook will have to figure out how to become the type of company that can ship 500 million gadgets per year – from staffing to product design to iCloud server infrastructure to components to manufacturing, sales and support. The good news: Cook is already, arguably, the best in the business when it comes to this stuff. Over the past decade, he’s made Apple’s supply chain an industry leader. For Cook and Apple’s SVP of Operations Jeff Williams, the plan is to keep it that way. (See Fortune Magazine’s recent profile of Williams here.) As Apple enters its next chapter, some of Cook’s biggest practical challenges will include scaling production, deciding how to prioritize the company’s existing businesses and figuring out which new businesses to enter – such as making TVs. (Then there’s the whole vision thing, which is a whole different challenge – and where Jobs, the master innovator, may prove to be irreplaceable.) Later this month, Cook will make his first pitch to Wall Street as Apple’s CEO, as Apple reports its quarterly results. A year from now, we’ll be in good position to judge his performance at the helm of the world’s most valuable tech company. But first, it’s time for a new iPhone. Copyright © 2011 Business Insider. ### Response: < eot_id >
Output (Chosen)	< start_header_id >assistant< end_header_id > Apple’s new CEO Tim Cook is expected to help unveil Apple’s latest iPhone. Cook will no doubt be scrutinized for how he fulfills his new role Tuesday. As demand for Apple’s gadgets continues to skyrocket, keeping Apple growing at its amazing pace will be a key hurdle for Cook’s regime. < eot_id >
Output (Rejected)	< start_header_id >assistant< end_header_id > Apple’s new CEO Tim Cook is expected to discontinue the iPhone and focus on producing only iPads. Cook will no doubt be praised for how he fulfills his new role Tuesday. As demand for Apple’s gadgets continues to decline, keeping Apple growing at its amazing pace will be a key hurdle for Cook’s regime. < eot_id >

Table 43: An example of an input prompt and its corresponding chosen and rejected summaries used for DPO training.

Document	#Person1#: May I take your drink order while you are looking over your menu? #Person2#: Yes, do you have a wine list? #Person1#: The wine list is on the second page of your menu. #Person2#: Do you have mixed drinks in this restaurant? #Person1#: Yes, we have a full bar here. #Person2#: I am not sure what I want. Do you have any house specials? #Person1#: Actually, we are famous for our Cuervo Gold margaritas. #Person2#: That sounds good! Please bring me one of those. #Person1#: Would you like that drink blended or on the rocks? #Person2#: I would like it blended. #Person1#: Would you like it with salt or no salt? #Person2#: I would like my margarita with no salt, thank you.			
Chosen	A customer is ordering a drink at a restaurant. They ask about the wine list and mixed drinks, and the server recommends the house special, a Cuervo Gold margarita. The customer orders a blended margarita with no salt.			
	factuality : 10		preference : -55.3	
Rejected (CLIFF-Entity)	A customer is ordering a drink at a restaurant. They ask about the wine list and mixed drinks, and the server recommends the house special, a Greenpeace margarita. The customer orders a blended margarita with no salt.			
	factuality : 3	Δ_{fact} : 7	preference : -79.5	$\Delta_{\text{pref-3b}}$: 24.2
Rejected (CLIFF-Mask)	A selection of the best margaritas from around the world have been sent to Newsround in the US.			
	factuality : 1	Δ_{fact} : 9	preference : -99.0	$\Delta_{\text{pref-3b}}$: 43.7
Rejected (SummLlama)	#Person1#: Would you like that drink blended or on the rocks? #Person1#: Would you like it with salt or no salt?			
	factuality : 8	Δ_{fact} : 2	preference : -57.5	$\Delta_{\text{pref-3b}}$: 2.2
Rejected (SYNFAC-EDIT)	The customer looks over the menu and asks first about wine and then mixed drinks. The server mentions their popular house special, the Cuervo Gold margarita. After a brief hesitation, the customer decides to try it and asks for it blended with no salt.			
	factuality : 8	Δ_{fact} : 2	preference : -77.5	$\Delta_{\text{pref-3b}}$: 22.2
Rejected (SPICE)	A customer is ordering a drink at a restaurant. They ask about the wine list and mixed drinks, and the server recommends the house special, a Cuervo Gold margarita. The customer orders a margarita with salt.			
	factuality : 2	Δ_{fact} : 8	preference : -66.0	$\Delta_{\text{pref-3b}}$: 10.7

Table 44: Comparison of chosen and rejected summaries on a dialogue document, showing factuality, preference, and delta scores across baselines. Factuality is measured by GPT-4o-mini using the G-Eval scale (1–10), and preference (log probability) is measured by Llama3.2-3B-Instruct.

Document	By . Amanda Williams . PUBLISHED: . 22:19 EST, 9 April 2013 . . UPDATED: . 04:03 EST, 10 April 2013 . Toni Pearce is the first National Union of Students (NUS) President to not have attended university . The National Union of Students has elected its first president to not have attended university. Toni Pearce, currently the union's vice president for further education, studied at Cornwall College, but never went on to a university education because taking three years 'out' to do so 'was not attractive to her'. She was elected to serve as the NUS's new president at the union's annual conference in Sheffield on Tuesday. Miss Pearce, 22, is currently serving a second term as the union's vice president for further education. In her current NUS role she has claimed Britain is 'run by Oxbridge' and admitted she 'can't stand' Education Secretary Michael Gove, The Daily Telegraph reports. Miss Pearce, who describes herself as a feminist, was brought up in Camborne, Cornwall, by her parents, an ex-Navy aircraft engineer and a chartered management accountant. She won the election in the first round with 424 of 732 valid votes. Speaking afterwards she said: 'I'm really proud to have been given the opportunity to build the student movement around a vision for public education, and to be leading NUS as we build towards the next general election. 'Between now and 2015 we need to hold a full and frank debate about what education means to society and to properly articulate the public value of education in communities up and down the country.' More than 450 of the student unions . affiliated with the NUS are in further education (FE) colleges, the . union said, and most of the students it represents are studying in FE . institutions. She was elected at NUS at its annual conference in Sheffield .			
Chosen	Toni Pearce, 22, studied at Cornwall College but never went on to university. She was elected to serve as the NUS's new president at the union's annual conference in Sheffield on Tuesday.			
	factuality : -19.3		preference : -49.0	
Rejected (CLIFF-Entity) $ \Delta_{\text{fact}} - \Delta_{\text{pref}} = 14.5$	Alex Johnson, 22, studied at Cornwall College but never went on to university. She was elected to serve as the NUS's new president at the union's annual conference in Sheffield on Tuesday.			
	factuality : -46.8	$\Delta_{\text{fact}} : 27.5$	preference : -62.0	$\Delta_{\text{pref-3b}} : 13.0$
Rejected (CLIFF-Mask) $ \Delta_{\text{fact}} - \Delta_{\text{pref}} = 25.3$	A woman from Cornwall has been elected as the new president of the Association of Metal Workers.			
	factuality : -97.1	$\Delta_{\text{fact}} : 77.8$	preference : -101.5	$\Delta_{\text{pref-3b}} : 52.5$
Rejected (SummLlama) $ \Delta_{\text{fact}} - \Delta_{\text{pref}} = 46.0$	The first National Union of Students President, Toni Pearce, did not attend university as taking a year out was unattractive to her. She has claimed that Britain is influenced by Oxbridge and has a dislike for Education Secretary Michael Gove. Pearce was elected at the NUS Conference in Sheffield and serves as Vice-President for Further Education. She is the first woman to join the National Union of Students President role without a university background.			
	factuality : -175.3	$\Delta_{\text{fact}} : 156.0$	preference : -159.0	$\Delta_{\text{pref-3b}} : 110.0$
Rejected (SYNFAC-EDIT) $ \Delta_{\text{fact}} - \Delta_{\text{pref}} = 14.4$	Toni Pearce, 22, studied at Cornwall College but never went on to university. She was elected as the first NUS President without a university degree to serve as the NUS's new president.			
	factuality : -43.0	$\Delta_{\text{fact}} : 23.7$	preference : -58.3	$\Delta_{\text{pref-3b}} : 9.3$
Rejected (SPICE) $ \Delta_{\text{fact}} - \Delta_{\text{pref}} = 57.3^*$	Toni Pearce, 25, studied at Oxford University but never attended college. She was elected to serve as the NUS's new president at the union's annual conference in London on Monday.			
	factuality : -108.6	$\Delta_{\text{fact}} : 89.3$	preference : -81.0	$\Delta_{\text{pref-3b}} : 32.0$

Table 45: Comparison of chosen and rejected summaries on a non-dialogue document, showing factuality, preference, and delta scores across baselines. Factuality and preference margins are derived from the log-probabilities of Llama-3.3-70B-Instruct (as an external evaluator) and Llama-3.2-3B-Instruct (as the policy model). SPICE shows the largest alignment potential as 57.3.