Beyond Overlap Metrics: Rewarding Reasoning and Preferences for Faithful Multi-Role Dialogue Summarization

Xiaoyong Mei[1]*, Tingting Zuo[1]*, Da Chen[2]*, Guangyu Hu[3], Xiangyu Wen[4]
Chao Duan[1]†, Mingyan Zhang[1], Fudan Zheng[5]
[1]Zhejiang Normal University [2]Huawei Technologies [3]HKUST, Hong Kong
[4]CUHK, Hong Kong [5]Sun Yat-sen University
{cdmxy, mingyanzhang, duanchao, zuotingting}@zjnu.edu.cn
915102610106@njust.edu.cn ghuae@connect.ust.hk
1155186676@link.cuhk.edu.hk zhengfd5@mail.sysu.edu.cn

Abstract

Multi-role dialogue summarization requires modeling complex interactions among multiple speakers while preserving role-specific information and factual consistency. However, most existing methods optimize for automatic metrics such as ROUGE and BERTScore, which favor surface-level imitation of references rather than genuine gains in faithfulness or alignment with human preferences. We propose a novel framework that couples explicit cognitive-style reasoning with reward-based optimization for multi-role dialogue summarization. Our method first distills structured reasoning traces (e.g., step-by-step inferences and intermediate reflections) from a large teacher model and uses them as auxiliary supervision to initialize a reasoning-aware summarizer via staged supervised fine-tuning. It then applies GRPO with a dual-principle reward that blends metric-based signals with human-aligned criteria targeting key information coverage, implicit inference, factual faithfulness, and conciseness. Experiments on multilingual multi-role dialogue benchmarks show that our method matches strong baselines on ROUGE and BERTScore. Specifically, results on CSDS confirm the framework's stability in semantic consistency, while in-depth analysis on SAMSum demonstrates clear gains in factual faithfulness and model-based preference alignment. These findings underscore the value of reasoning-aware and preference-aware training for reliable dialogue summarization[1].

1 Introduction

In this paper, we focus on the multi-role dialogue summarization tasks, where the system distills key information from multi-turn conversations into concise, coherent summaries (Feng et al., 2022). Such summaries enable users to quickly grasp the main points without having to navigate complex conversational context (Chen and Yang, 2020). This capability is crucial for a range of real-world applications, including customer service interaction analysis and automatic generation of meeting minutes (Feigenblat et al., 2021; Zhao et al., 2021).

A central challenge in dialogue summarization is accurately modeling complex interaction dynamics to produce faithful summaries. Early work established baselines by fine-tuning pre-trained sequence-to-sequence models (Lewis et al., 2020; Zhang et al., 2020; Zhong et al., 2022). However, these methods largely operate as surface-level text mappers and lack deep semantic understanding. More recent research has turned to Large Language Models in an effort to overcome these limitations, considering their impressive performance across a wide range of tasks and application scenarios (Wang et al., 2023; Tian et al., 2024; Zhang et al., 2025; Wen et al., 2025b). Existing LLM-based approaches primarily seek to elicit reasoning capabilities through Chain-of-Thought prompting (Wang et al., 2023; Jin et al., 2026), or to improve alignment via instruction tuning and Reinforcement Learning from Human Feedback (RLHF) (Rafailov et al., 2023; Li et al., 2025b; Lu et al., 2025; Ye et al., 2025a).

Despite these advances, maintaining faithfulness remains a major bottleneck. In particular, we observe that models are prone to hallucinations—such as misattributing responsibility to a speaker or fabricating unsupported details—which severely undermines factual accuracy and alignment with human preferences. A key reason is that real-world conversations are characterized by multiple participants, informal and noisy language, and intricate interaction patterns (Ramprasad et al., 2024; Wen et al., 2025b), but standard instruction tuning does not enforce strict logical consistency, all of which can cause models to miss critical information or misin-

*Equal contribution.
†Corresponding author.
[1]Checkpoints and datasets are available at https://huggingface.co/collections/NebulaPixel/summorchestra-multirole-summary.

Findings of the Association for Computational Linguistics: ACL 2026, pages 23189–23203

terpret speaker intent, ultimately leading to factual inconsistencies (Ramprasad et al., 2024).

To address these challenges, we introduce a novel framework for multi-role dialogue summarization that explicitly embeds cognitive-style reasoning into the reinforcement learning process. Our method is designed around the hypothesis that, before producing a summary, the model should engage in structured reasoning procedures that mirror human cognitive behaviors—such as inferring speaker intent, tracking responsibilities, and verifying factual consistency. By orchestrating this reasoning phase prior to generation, our framework aims to substantially improve both factual faithfulness and alignment with human preferences in complex, real-world conversational settings.

Specifically, we adopt a teacher–student paradigm (Hinton et al., 2015; Wen et al., 2025a; Fang et al., 2025) to distill structured reasoning capabilities from a larger model into a base model, initializing a reasoning-aware summarizer via supervised fine-tuning. Building on this, we introduce a dual-principle reward mechanism within a group relative policy optimization (GRPO) training framework (Shao et al., 2024). A **reasoning principle** evaluates intermediate reasoning traces and summaries in terms of key information coverage, implicit understanding, and factual fidelity, while a **summary principle** assesses final outputs using lexical and semantic metrics (e.g., ROUGE, BERTScore) and length control. Together, these rewards provide fine-grained process supervision, enabling the policy model to internalize complete reasoning paths and generate summaries that are both faithful and well aligned with human preferences.

Our experiments demonstrate that our method attains performance on par with strong baselines in standard automatic metrics such as ROUGE and BERTScore, while delivering substantial gains in factual faithfulness and preference alignment evaluations. These results underscore the effectiveness of integrating explicit cognitive reasoning with reward-based optimization, and highlight it as a promising direction for faithful, preference-aligned multi-role dialogue summarization.

2 Related Work

2.1 Multi-Role Dialogue Summarization

Multi-role dialogue summarization aims to generate concise and coherent summaries from conversations involving multiple participants. Compared with single-speaker settings, multi-role dialogues pose additional challenges, including intricate turn-taking patterns, rich speaker interactions, and complex discourse structures. Early work in this area primarily relied on extractive strategies, such as graph-based ranking and keyphrase- or template-driven methods, which were widely applied to meeting and multi-party dialogue data (Murray et al., 2005; Riedhammer et al., 2008; Tixier et al., 2017).

With the advent of neural approaches, research has increasingly shifted toward sequence-to-sequence architectures with hierarchical encoders and attention mechanisms to better capture conversational context, discourse structure, and role-specific information (Li et al., 2019; Zhu et al., 2020; Chen and Yang, 2020; Feng et al., 2021). More recently, multi-stage and hierarchical frameworks have been proposed to address the challenge of long-range dependencies in extended dialogues and meetings (Zhang et al., 2022). Nevertheless, faithfully aggregating speaker intent across multi-turn interactions while maintaining factual consistency remains an open challenge, especially in real-world scenarios with diverse and evolving participant roles.

2.2 Factual Consistency in Summarization

In this case, factual consistency has become a central concern in abstractive summarization, where models can generate plausible yet unsupported or contradictory content (Rennard et al., 2023). To address this, a variety of metrics—such as FactCC (Kryscinski et al., 2019), QAGS (Feng et al., 2024), SummaC (Laban et al., 2021), and FIZZ (Yang et al., 2024)—have been introduced to assess the degree to which summaries are grounded in the source. In dialogue summarization, auxiliary language understanding signals have been incorporated to reduce factual errors (Akani et al., 2024), and discourse-level evaluation highlights the role of document structure, especially in long texts (Zhong and Litman, 2025).

Methodologically, prior work includes constrained decoding, post-hoc correction, and fact-checking or entailment modules. For example, graph-based knowledge grounding mitigates unfaithful responses in dialogues (Ji et al., 2023). Manakul et al. (2023) propose SelfCheckGPT to detect hallucinations post-hoc. Yu et al. (2024) propose Truth-Aware Context Selection (TACS) to fil-

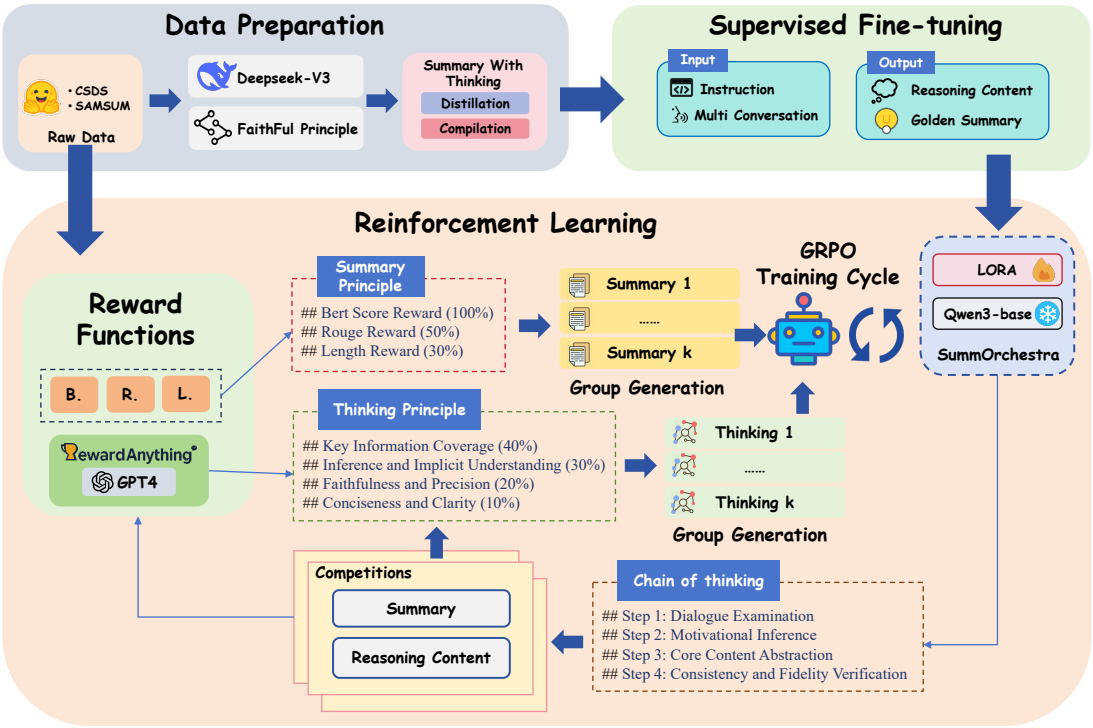


Figure 1: Overview of the proposed framework. The pipeline comprises two main parts: (1) Reasoning distillation from a teacher model to extract thinking traces for data creation; (2) a two-stage training paradigm that first initializes a reasoning-aware summarizer and then internalizes factual consistency and human preferences.

ter untruthful context before generation. However, these techniques are mostly designed for single-document or single-speaker settings. Multi-role dialogues pose additional challenges, as facts are fragmented across speakers, expressed informally, and often revised, making robust factual grounding substantially harder.

2.3 Human Preference Alignment for Text Generation

Beyond factual correctness, summarization models should align with human preferences for clarity, usefulness, and style. *Reinforcement Learning from Human Feedback (RLHF)* trains models using pairwise human preferences via a learned reward model (Christiano et al., 2017). *Direct Preference Optimization (DPO)* provides a simpler alternative without a separate reward model (Rafailov et al., 2023), with variants such as ODPO (Amini et al., 2024) and length-controlled DPO (Park et al., 2024) addressing preference strength and human bias.

Recent works further improve factuality and task-specific alignment. Aggregating multiple imperfect metrics enhances factual consistency (Ye et al., 2025b), while DPO has been applied to conversational recommendation (Tajiri and Inaba, 2025), multi-dimensional feedback (Song et al., 2025), and joint instruction-response preference learning (JPO) (Bansal et al., 2025). Despite these advances, incorporating role-sensitive salience, factuality, and narrative structure into a unified preference signal for multi-role dialogue summarization remains underexplored.

2.4 Discussion on Limitations of Existing Methods

Despite notable progress, existing approaches to dialogue summarization exhibit several limitations in multi-role settings. Extractive models struggle to capture implicit relations, discourse-level intent, and cross-speaker dependencies. Abstractive models, while more expressive, are prone to hallucinations, misattribution of roles or responsibilities, and biased emphasis on specific speakers. Moreover, most prior work treats factual consistency and human preference alignment as separate or secondary objectives, and few methods explicitly integrate both in the context of complex, multi-speaker conversations. These gaps motivate the development of frameworks that can simultaneously model role-specific dynamics, enforce factual faithfulness, and optimize for human-aligned quality—precisely the goal of the proposed framework in this work.

Principle	Description	Weight
Key Information Coverage	Does the summary capture the core facts of the dialogue? Must include: the request/proposal, the refusal, the insistence, and any implied motivation if present. Missing major elements is a critical error.	40%
Inference and Implicit Understanding	Does the summary correctly reflect implied attitudes, motives, or emotional tone (e.g., sarcasm, concern, frustration)? Reasonable inference is rewarded. Fabrication is penalized.	30%
Faithfulness and Precision	No hallucinations or incorrect claims beyond what can be safely inferred. The summary must not change the meaning of the original dialogue.	20%
Conciseness and Clarity	The summary should be brief, readable, and well-structured. Overly verbose summaries lose points even if factually correct.	10%

Table 1: Evaluation criteria and weighting for dialogue summarization.

3 Methodology

Our framework for multi-role dialogue summarization is explicitly designed to improve both factual consistency and alignment with human preferences. As illustrated in Figure 1, our approach follows a structured pipeline comprising tailored data construction and a two-stage training process, which together equip the model with reasoning-aware and preference-aligned summarization capabilities.

3.1 Data Construction and Reasoning Distillation

As shown in the first module of Figure 1, we construct high-quality training data for multi-role dialogue summarization by distilling explicit reasoning signals from raw dialogues. Given the complexity of reasoning over multiple speakers and interaction patterns, we employ DeepSeek-v3 (Liu et al., 2024) as a teacher model to generate structured reasoning traces—such as step-by-step reasoning paths and intermediate reflections—for each dialogue. These distilled traces are incorporated as auxiliary supervision alongside the original summaries, enabling the student model to internalize both factual content and the underlying reasoning logic that supports faithful summarization and subsequent preference-based training. The distillation process follows the prompt template provided in Appendix C.

3.2 Two-Stage Training for Internalizing Faithfulness and Human Preference

To more effectively internalize reasoning logic and produce summaries that are both faithful and aligned with human preferences, we adopt a two-stage training paradigm. In the first stage, we perform supervised fine-tuning (SFT) to initialize the base model as a reasoning-aware summarizer. The model is initially trained on the original dialogue–summary pairs without reasoning annotations, and is then further fine-tuned on the augmented dataset enriched with distilled reasoning traces from the teacher model. This staged SFT procedure encourages the model to learn the desired reasoning format and leverage intermediate reasoning before generating the final summary, providing a stable and structured initialization for subsequent reinforcement learning.

In the second stage, we apply group relative policy optimization (GRPO) to further align the model's outputs with our semantic quality and factual principles. For each input dialogue, the policy samples a group of G candidate summaries, and optimization is carried out based on their relative rewards under the dual-principle reward mechanism. This reinforcement learning phase refines the model beyond supervised imitation, driving it toward summaries that are more factually faithful and better reflect human preferences in complex multi-role conversational settings. The group-level reward is defined in Eq. 1:

$$\begin{aligned} R_{\text{group}} &= \frac{1}{G}\sum_{i=1}^{G} R_i^{\text{base}}, \\ R_i^{\text{base}} &= \lambda_b\, R_{\text{bertscore},i} + \lambda_r\, R_{\text{rouge},i} \\ &\quad + \lambda_l\, R_{\text{length},i} + \lambda_p\, R_{\text{principle},i} \end{aligned}, \quad (1)$$

where $\lambda_b : \lambda_r : \lambda_l : \lambda_p = 1 : 0.5 : 0.3 : 1$.

We design four complementary reward components to guide GRPO training. R_{rouge} is defined as the average of ROUGE-1, ROUGE-2, and ROUGE-L, encouraging reasonable lexical overlap with the reference without overfitting to surface forms. $R_{\text{bertscore}}$ measures semantic similarity via contextual embeddings, promoting preservation of sentence-level meaning. R_{length} rewards summaries whose length is close to that of the reference,

discouraging both under- and over-generated outputs. Finally, $R_{\text{principle}}$ leverages the RewardAnything framework (Yu et al., 2025) with GPT-4 as a teacher model to score summaries against four human-aligned criteria—key information coverage, implicit inference, factual faithfulness, and conciseness (Table 1).

We set the reward weights, as shown in Table 1, to emphasize semantic consistency and factual correctness while preserving concise expression. In particular, the ROUGE-based component is assigned a relatively lower weight to avoid mere imitation of reference phrasing, whereas the principle-based reward plays a dominant role in steering the model toward factually grounded, preference-aligned summaries in multi-role dialogue settings.

4 Experimental Results

4.1 Experimental Setup

Benchmark Details. We conduct experiments on two widely used dialogue summarization benchmarks: CSDS (Lin et al., 2021) and SAMSum (Gliwa et al., 2019). CSDS is a large-scale Chinese dialogue summarization dataset constructed from multi-turn, multi-party conversations in diverse, real-world scenarios (e.g., customer service, daily chat, and task-oriented discussions). It provides human-written abstractive summaries for each dialogue, with conversations typically longer and more complex than in standard English benchmarks. This makes CSDS particularly suitable for evaluating models' ability to handle informal language, topic drift, and information sparsity in Chinese conversational settings.

SAMSum is an English dialogue summarization dataset composed of short, messenger-style conversations that resemble everyday chat on platforms such as WhatsApp or Facebook Messenger. Each dialogue is paired with a concise abstractive summary written by linguists, focusing on capturing key events, decisions, and user intentions. Compared to CSDS, SAMSum dialogues are generally shorter and more informal, with a strong emphasis on ellipsis, pragmatics, and conversational implicature, thus providing a complementary testbed for evaluating cross-lingual robustness and performance on casual English conversations.

Evaluation Metrics. We assess models with complementary metrics covering lexical overlap, semantic similarity, factual consistency, and human preference, using language-aware configurations for multilingual data. Together, these metrics provide a compact yet comprehensive evaluation of multi-role dialogue summaries, balancing surface overlap, meaning preservation, factual grounding, and alignment with human preferences.

We report ROUGE-1, ROUGE-2, and ROUGE-L with language-specific toolchains (Lin, 2004). Language is detected via a reference-based classifier (langid [2]). For Chinese, we use rouge_chinese [3] with THULAC (Li and Sun, 2009) word-level tokenization; for English, we use py-rouge [4]. This avoids artifacts from applying a single, non-adapted ROUGE variant across languages. Details are present in Table 2.

ROUGE Setting	Description
Language detection	langid (reference-based)
Chinese ROUGE	`rouge_chinese`
Chinese tokenization	THULAC (word-level)
Token granularity	Word-level (not character-level)
Character substitution	Not used
English ROUGE	`py-rouge`
Metrics	ROUGE-1 / ROUGE-2 / ROUGE-L

Table 2: Evaluation protocol for ROUGE.

Semantic similarity is measured with BERTScore (Zhang et al., 2019) using language-specific backbones: BERT-base Chinese [5] for Chinese texts and BERT-base uncased [6] for English. Scores are computed with cosine similarity over the top 12 layers, and we report mean precision, recall, and F1, without IDF weighting. Details are present in Table 3.

BERTScore Setting	Description
Language detection	langid (reference-based)
Chinese backbone	BERT-base Chinese
English backbone	BERT-base uncased
Layer selection	Top 12 transformer layers
Score aggregation	Mean Precision / Recall / F1
IDF weighting	Not used
Token alignment	Cosine similarity

Table 3: Evaluation protocol for BERTScore.

Factual consistency is quantified as the proportion of a summary's factual claims that are supported by the source dialogue. We first decom-

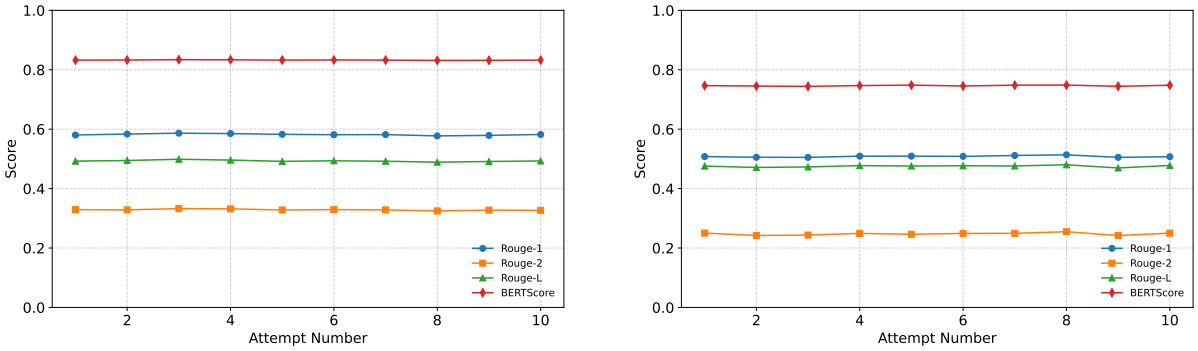
[2] https://github.com/saffsd/langid.py

[3] https://github.com/Isaac-JL-Chen/rouge_chinese

[4] https://github.com/diegoantognini/py-rouge

[5] https://huggingface.co/google-bert/bert-base-chinese

[6] https://huggingface.co/google-bert/bert-base-uncased



(a) CSDS dataset

(b) SAMSum dataset

Figure 2: Performance of our model over ten resampling trials on the test datasets of CSDS and SAMSum

pose each summary into atomic statements, then verify each against the dialogue using HHEM-2.1-Open (Li et al., 2024), a T5-based hallucination detector. The faithfulness score is: Faithfulness $= \frac{\#supported\ claims}{\#total\ claims}$, yielding a value in $[0, 1]$.

We adopt WorldPM-72B-RLHFLOW (Wang et al., 2025) as an automated evaluator to approximate human preference judgments over model outputs. WorldPM has been extensively validated and shown to match or surpass the reliability of human annotators across a wide range of evaluation tasks (Wang et al., 2025), providing a strong empirical basis for its use as a proxy for human evaluation. Leveraging WorldPM enables us to systematically and scalably measure preference signals without relying exclusively on expensive, time-consuming human annotation. This design choice is consistent with emerging best practices in alignment research, where high-capacity reward models and LLM-as-judge frameworks are increasingly used to approximate human feedback in both RLHF pipelines and automatic preference evaluation (Li et al., 2025a; Wu et al., 2025).

Training Detail. The SFT stage employs a LoRA-based parameter-efficient training strategy. Models are trained for 8 epochs with a per-device batch size of 16, learning rate 1×10^{-5}, and gradient accumulation of 1 step. Adam optimizer is applied to all linear modules. Maximum sequence length is 2048, with mixed precision and gradient checkpointing enabled. Validation and checkpointing occur every 50 steps, with early stopping after 3 intervals without improvement. LoRA parameters are set to rank 16 and alpha 32, with 8 dataloader workers and a warmup ratio of 0.05.

Following SFT, the GRPO stage further refines the model using a training reward composed of four components: ROUGE, BERTScore, length, and a **principle reward** implemented via RewardAnything. GRPO training is conducted for 100 epochs with a per-device batch size of 12, learning rate 5×10^{-6}, maximum completion length of 2048, gradient clipping of 0.5, and mixed precision (bfloat16). Dataset processing uses 8 processes with 4 generations per example. Reward ablation experiments are conducted on subsets of the four reward components: B (BERTScore), R (ROUGE), L (length), and P (principle).

Evaluation Setting. For evaluation, we measure **faithfulness**, ensuring factual consistency with source dialogues and distilled reasoning, and **human preference**, assessing alignment with human judgments of clarity, informativeness, and relevance. This distinction clarifies that the training rewards and evaluation metrics are not identical.

All experiments are conducted on 8 Ascend NPU 910B cards, each with 64GB of memory. We evaluate Qwen2.5 on CSDS and Qwen3 on SAMSum, selecting the base models that best align with each dataset's language characteristics and reasoning requirements (see Appendix A). Training follows a two-stage paradigm: supervised fine-tuning (SFT) followed by guided reinforcement learning with preferences (GRPO). More training details can be found in Appendix A. For reproducibility, all model and datasets are publicly available.[7]

4.2 Experimental Results Analysis

To ensure the reliability of our evaluation, we first assessed the stability of the trained models on the CSDS and SAMSum test sets using ten resampling trials. As shown in Figure 2, the results remain

[7] https://huggingface.co/collections/NebulaPixel/summorchestra-multirole-summary

Model	Training Method	B	R	L	P	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
gpt-4.1	-	-	-	-	-	42.91	14.08	32.04	77.37
gpt-4.1-mini	-	-	-	-	-	42.78	14.11	32.31	77.40
gpt-4o	-	-	-	-	-	48.52	19.59	36.55	79.51
gpt-4o-mini	-	-	-	-	-	45.87	16.93	33.82	78.28
gpt-5	-	-	-	-	-	39.72	11.47	28.98	75.63
gpt-5-mini	-	-	-	-	-	41.54	12.95	30.15	76.68
qwen2.5-3B	SFT	-	-	-	-	50.87	27.35	43.40	76.58
qwen2.5-3B	GRPO(B)	✓	-	-	-	47.21	25.78	41.85	83.10
qwen2.5-3B	GRPO(B+R)	✓	✓	-	-	58.00	32.95	48.90	83.22
qwen2.5-3B	GRPO(B+R+L)	✓	✓	✓	-	58.50	33.25	49.20	83.28
qwen2.5-7B	SFT	-	-	-	-	51.11	28.19	44.05	77.29
qwen2.5-7B	GRPO(B)	✓	-	-	-	48.40	23.45	42.35	83.42
qwen2.5-7B	GRPO(B+R)	✓	✓	-	-	58.80	34.00	49.75	83.50
qwen2.5-7B	GRPO(B+R+L)	✓	✓	✓	-	59.05	34.25	50.05	83.58

Table 4: Results on the CSDS dataset. B, R, L represent the rewards enabled in GRPO training: B for BERTScore, R for ROUGE, L for length reward. Checkmarks indicate which rewards were used in each training method.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Aligned Pref.	Faithfulness
Original Dataset	–	–	–	–	0.4973	0.7548
gpt-4.1	42.91	14.08	32.04	77.37	0.5012	0.6124
gpt-4.1-mini	42.78	14.11	32.31	77.40	0.4983	0.6021
gpt-4o	48.52	19.59	36.55	79.51	0.5223	0.6354
gpt-4o-mini	45.87	16.93	33.82	78.28	0.5434	0.6132
gpt-5	39.72	11.47	28.98	75.63	0.5635	0.6144
gpt-5-mini	41.54	12.95	30.15	76.68	0.4832	0.5344
qwen3-fast[a](BRL)	48.85	22.89	45.86	72.64	0.4435	0.6659
qwen3-fast[a](BRLP)	47.40	21.79	44.73	70.34	**0.5563**	**0.7959**
qwen3-slow[b](BRL)	50.78	25.03	47.58	74.67	0.4883	0.6852
qwen3-slow[b](BRLP)	49.78	24.12	46.43	73.32	**0.6683**	**0.8352**

[a] "fast" denotes the no-thinking (fast inference) mode.
[b] "slow" denotes the thinking (slow inference) mode.

Table 5: Results on the SAMSum dataset. For qwen3, both modes are evaluated after supervised fine-tuning and GRPO training. **The principle reward (P) enforces factual consistency and structural validity, and does not directly optimize ROUGE or BERTScore.**

consistently stable across these trials, indicating that the final models generalize well and do not exhibit signs of overfitting.

Building upon this, Table 4 summarizes the performance of various models on the CSDS dataset. It can be seen that supervised fine-tuning (SFT) with Qwen2.5 models leads to substantial improvements over baseline LLMs. Furthermore, the GRPO stage with semantic and lexical rewards (B+R+L) consistently improves over SFT, demonstrating the framework's effectiveness in optimizing standard metrics on large-scale data.

During the GRPO stage, the choice of reward signals significantly influences model behavior. Optimizing solely for semantic similarity tends to reduce surface-level overlap, while incorporating additional rewards, such as ROUGE and length, progressively enhances both semantic and structural quality. These observations suggest that different reward components contribute complementary benefits, and a careful combination is crucial for balanced performance.

4.3 Ablation Study

To further dissect the effects of reward design, we analyze results on the SAMSum dataset, comparing fast and slow inference regimes for Qwen3 models. Incorporating the principle reward leads to modest reductions in conventional metrics but substantially improves factual faithfulness and preference alignment. As illustrated in Figure 3, the principle reward markedly enhances the preference score and factual reliability, demonstrating that structured rewards can guide models toward outputs that better

align with human preference criteria. Notably, our model (0.6683) surpasses the original human references (0.4973) in preference scores, suggesting it generates more coherent and useful summaries than the noisy ground truth.

4.4 Summary of Findings

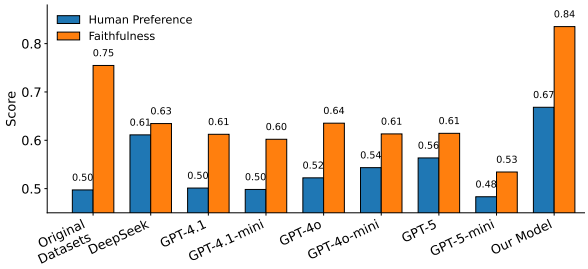


Figure 3: Human Preference and Faithfulness on the SAMSUM Dataset

The above experiments reveal a consistent pattern: traditional automatic metrics alone do not fully capture summary quality, particularly regarding factual consistency and user-perceived value. Multi-reward GRPO training improves both semantic and structural aspects of summaries, while the addition of the principle reward shifts optimization toward structural validity and factual reliability. The improvements in human preference and faithfulness, corroborated by Figure 3, underscore the importance of integrating both surface-level and principle-oriented objectives when optimizing models for multi-role dialogue summarization, providing a holistic measure of model performance.

5 Conclusion and Future Work

We presented a two-stage framework for multi-role dialogue summarization that explicitly targets factual consistency and alignment with human preferences rather than merely optimizing surface-level metrics. First, a teacher–student SFT stage distills structured cognitive-style reasoning into a reasoning-aware summarizer. Then, a GRPO-based reinforcement learning stage applies a dual-principle reward that jointly supervises intermediate reasoning traces and final summaries along dimensions of key information coverage, implicit inference, factual faithfulness, and conciseness.

Experiments on CSDS and SAMSum show that our approach matches strong baselines on ROUGE and BERTScore while achieving clear gains in factual faithfulness and model-based preference alignment, supported by qualitative evidence of more accurate and coherent summaries. These results underscore the limitations of traditional metrics for multi-role dialogue summarization and highlight the value of explicitly integrating reasoning and preference-aware objectives into training.

Future work includes constructing datasets that more directly capture factual consistency and human preferences, scaling our framework to larger and more diverse multilingual dialogue corpora, and exploring broader applications of reasoning-aware, preference-aligned optimization beyond summarization.

Limitations

While our two-stage framework demonstrates strong performance in multi-role dialogue summarization, several limitations remain.

First, evaluation on the Chinese CSDS dataset is constrained by the lack of suitable factual-consistency and preference evaluators. Because HHEM-2.1-Open cannot be applied to Chinese dialogues, the GRPO stage for CSDS omits the principle reward, limiting our ability to directly optimize and assess faithfulness on this dataset. In addition, SAMSum is relatively small, predominantly English, and exhibits limited topical diversity, which may restrict the generalization of our approach to more complex, multilingual, or domain-specific multi-role dialogues.

Second, our GRPO reward function combines automatic metrics (ROUGE, BERTScore) with model-based proxies for factual consistency and human preference. Although this design captures several aspects of summary quality, the available human preference annotations are limited in scale and potentially subjective, and existing factual-consistency metrics may miss subtle hallucinations or nuanced errors.

While intermediate reflections improve summarization performance on the evaluated benchmarks, their effectiveness across other domains, languages, and more complex dialogue settings remains to be validated. Additionally, reflection introduces extra training and inference overhead, posing challenges for large-scale deployment.

Future work will focus on expanding and diversifying datasets, developing more reliable evaluators (especially for non-English settings), and improving training and inference efficiency to enhance robustness, generalization, and real-world applicability.

Acknowledgments

This work was supported in part by the Research Project on Ideological and Political Work under the Philosophy and Social Science Planning Program of Zhejiang Province (No. 25GXSZ009YB), and in part by the National Key Research and Development Program of China (No. 2022YFC3303600).

References

Eunice Akani, Benoît Favre, Frédéric Béchet, and Romain Gemignani. 2024. Increasing faithfulness in human-human dialog summarization with spoken language understanding tasks. *CoRR*, abs/2409.10070.

Afra Amini, Tim Vieira, and Ryan Cotterell. 2024. Direct preference optimization with an offset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9954–9972, Bangkok, Thailand. Association for Computational Linguistics.

Hritik Bansal, Ashima Suvarna, Gantavya Bhatt, Nanyun Peng, Kai-Wei Chang, and Aditya Grover. 2025. Comparing bad apples to good oranges aligning large language models via joint preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 701–723, Vienna, Austria. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. Association for Computational Linguistics.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Luyang Fang, Xiao-Xing Yu, Jiazhang Cai, Yongkai Chen, Shushan Wu, Zheng Liu, Zhenyuan Yang, Haoran Lu, Xilin Gong, Yufang Liu, Terry Ma, Wei Ruan, Ali Abbasi, Jing Zhang, Tao Wang, Ehsan Latif, Wei Liu, Wei Zhang, Soheil Kolouri, and 5 others. 2025. Knowledge distillation and dataset distillation of large language models: emerging trends, challenges, and future directions. *Artificial Intelligence Review*, 59.

Guy Feigenblat, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. 2021. TWEETSUMM - a dialog summarization dataset for customer service. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 245–260, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Huawen Feng, Yan Fan, Xiong Liu, Ting-En Lin, Zekun Yao, Yuchuan Wu, Fei Huang, Yongbin Li, and Qianli Ma. 2024. Improving factual consistency of news summarization by contrastive preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11084–11100, Miami, Florida, USA. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. A survey on dialogue summarization: Recent advances and new frontiers. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5453–5460. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2021. Dialogue discourse-aware graph model and data augmentation for meeting summarization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3808–3814. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng, and Pascale Fung. 2023. Rho: Reducing hallucination in open-domain dialogues with knowledge grounding. In *Findings of the association for computational linguistics: ACL 2023*, pages 4504–4522.

Keyan Jin, Yapeng Wang, Leonel Santos, Tao Fang, Xu Yang, Sio Kei Im, and Hugo Gonçalo Oliveira. 2026. Reasoning or not? A comprehensive evaluation of reasoning llms for dialogue summarization. *Expert Syst. Appl.*, 299:129831.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. In *Conference on Empirical Methods in Natural Language Processing*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training

for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025a. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

Miaoran Li, Rogger Luo, and Ofer Mendelevitch. 2024. HHEM-2.1-Open.

Raymond Li, Chuyuan Li, Gabriel Murray, and Giuseppe Carenini. 2025b. Hierarchical attention adapter for abstractive dialogue summarization. In *Proceedings of The 5th New Frontiers in Summarization Workshop*, pages 17–30, Hybrid. Association for Computational Linguistics.

Zhongguo Li and Maosong Sun. 2009. Punctuation as implicit annotations for chinese word segmentation. *Computational Linguistics*, 35(4):505–512.

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

Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. CSDS: A fine-grained Chinese dataset for customer service dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4436–4451, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yen-Ju Lu, Ting-Yao Hu, Hema Swetha Koppula, Hadi Pouransari, Jen-Hao Rick Chang, Yin Xia, Xiang Kong, Qi Zhu, Xiaoming Simon Wang, Oncel Tuzel, and Raviteja Vemulapalli. 2025. Mutual reinforcement of LLM dialogue synthesis and summarization capabilities for few-shot dialogue summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7237–7256, Albuquerque, New Mexico. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Gabriel Murray, Steve Renals, and Jean Carletta. 2005. Extractive summarization of meeting recordings. In *9th European Conference on Speech Communication and Technology, INTERSPEECH-Eurospeech 2005, Lisbon, Portugal, September 4-8, 2005*, pages 593–596. ISCA.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4998–5017, Bangkok, Thailand. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Sanjana Ramprasad, Elisa Ferracane, and Zachary Lipton. 2024. Analyzing LLM behavior in dialogue summarization: Unveiling circumstantial hallucination trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12549–12561, Bangkok, Thailand. Association for Computational Linguistics.

Virgile Rennard, Guokan Shang, Julie Hunter, and Michalis Vazirgiannis. 2023. Abstractive meeting summarization: A survey. *Transactions of the Association for Computational Linguistics*, 11:861–884.

Korbinian Riedhammer, Benoît Favre, and Dilek Hakkani-Tür. 2008. A keyphrase based approach to interactive meeting summarization. In *2008 IEEE Spoken Language Technology Workshop, SLT 2008, Goa, India, December 15-19, 2008*, pages 153–156. IEEE.

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

Hwanjun Song, Taewon Yun, Yuho Lee, Jihwan Oh, Gihun Lee, Jason Cai, and Hang Su. 2025. Learning to summarize from LLM-generated feedback. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 835–857, Albuquerque, New Mexico. Association for Computational Linguistics.

Manato Tajiri and Michimasa Inaba. 2025. Refining text generation for realistic conversational recommendation via direct preference optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28628–28649, Suzhou, China. Association for Computational Linguistics.

Yuanhe Tian, Fei Xia, and Yan Song. 2024. Dialogue summarization with mixture of experts based on large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7143–7155, Bangkok, Thailand. Association for Computational Linguistics.

Antoine Tixier, Polykarpos Meladianos, and Michalis Vazirgiannis. 2017. Combining graph degeneracy and submodularity for unsupervised extractive summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 48–58, Copenhagen, Denmark. Association for Computational Linguistics.

Binghai Wang, Runji Lin, Keming Lu, Le Yu, Zhenru Zhang, Fei Huang, Chujie Zheng, Kai Dang, Yang Fan, Xingzhang Ren, and 1 others. 2025. Worldpm: Scaling human preference modeling. *arXiv preprint arXiv:2505.10527*.

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

Xiangyu Wen, Junhua Huang, Zeju Li, Min Li, Jianyuan Zhong, Zhijian Xu, Mingxuan Yuan, Yongxiang Huang, and Qiang Xu. 2025a. Reasoning scaffolding: Distilling the flow of thought from llms. *ArXiv*, abs/2509.23619.

Xiangyu Wen, Jianyuan Zhong, Zhijian Xu, and Qiang Xu. 2025b. Guideline compliance in task-oriented dialogue: The chained prior approach. In *North American Chapter of the Association for Computational Linguistics*, pages 6750–6776.

Jie Wu, Yu Gao, Zi-Nuo Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, Yangyang Zeng, and Weilin Huang. 2025. Rewarddance: Reward scaling in visual generation. *ArXiv*, abs/2509.08826.

Joonho Yang, Seunghyun Yoon, Byeongjeong Kim, and Hwanhee Lee. 2024. FIZZ: factual inconsistency detection by zoom-in summary and zoom-out document. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 30–45. Association for Computational Linguistics.

Jing Ye, Rui Wang, Yuchuan Wu, Victor Ma, Feiteng Fang, Fei Huang, and Yongbin Li. 2025a. CPO: Addressing reward ambiguity in role-playing dialogue via comparative policy optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 297–323, Suzhou, China. Association for Computational Linguistics.

Yuxuan Ye, Raul Santos-Rodriguez, and Edwin Simpson. 2025b. Optimising factual consistency in summarisation via preference learning from multiple imperfect metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17342–17355, Suzhou, China. Association for Computational Linguistics.

Tian Yu, Shaolei Zhang, and Yang Feng. 2024. Truth-aware context selection: Mitigating hallucinations of large language models being misled by untruthful contexts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10862–10884, Bangkok, Thailand. Association for Computational Linguistics.

Zhuohao Yu, Jiali Zeng, Weizheng Gu, Yidong Wang, Jindong Wang, Fandong Meng, Jie Zhou, Yue Zhang, Shikun Zhang, and Wei Ye. 2025. Rewardanything: Generalizable principle-following reward models. *CoRR*, abs/2506.03637.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, and Fei Wu. 2025. Instruction tuning for large language models: A survey. *ACM Comput. Surv.* Just Accepted.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summ^n: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics.

Lulu Zhao, Fujia Zheng, Keqing He, Weihao Zeng, Yuejie Lei, Huixing Jiang, Wei Wu, Weiran Xu, Jun Guo, and Fanyu Meng. 2021. Todsum: Task-oriented dialogue summarization with state tracking. *CoRR*, abs/2110.12680.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained

model for long dialogue understanding and summarization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11765–11773. AAAI Press.

Yang Zhong and Diane Litman. 2025. Discourse-driven evaluation: Unveiling factual inconsistency in long document summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2050–2073, Albuquerque, New Mexico. Association for Computational Linguistics.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online. Association for Computational Linguistics.

A Training Details

This appendix provides detailed descriptions of the dataset-specific training strategies and the two-stage optimization procedure adopted in our framework.

A.1 Dataset-Specific Training Strategies

We adopt tailored training strategies for different datasets to accommodate their language characteristics and annotation availability. The dataset statistics and splitting strategies are also summarized for completeness.

- **CSDS (Chinese)**: The CSDS dataset contains 9,101 training samples, 800 validation samples, and 800 test samples. For the Chinese multi-role dialogue dataset CSDS, we employ the fast-thinking model Qwen2.5. Since CSDS does not provide reliable intermediate reasoning annotations and automatic faithfulness evaluators are not readily available for Chinese, we directly train the model using supervised fine-tuning (SFT) on dialogue–summary pairs.

  Specifically, SFT is performed on the union of the training and validation sets, with the validation split also used for model selection. For the reinforcement learning stage, we further utilize the full dataset, and reserve the first 200 samples from the validation set as the evaluation set during GRPO training. Final performance is reported on the held-out test set.

- **SAMSum (English)**: The SAMSum dataset consists of 14,700 training samples, 818 validation samples, and 819 test samples. For the English SAMSum dataset, we use the slow-thinking model Qwen3, which supports explicit reasoning representations.

  We first perform knowledge distillation to augment the original dataset with distilled intermediate reasoning signals. The augmented data are then used for supervised fine-tuning, where both the training and validation sets are jointly used for training and model selection. Subsequently, guided reinforcement learning with preferences (GRPO) is applied on the full dataset, with the first 200 samples from the validation set reserved as the evaluation set during training.

Final evaluation is conducted on the test set to ensure fair comparison.

A.2 Two-Stage Training for Internalizing Faithfulness and Human Preference

We employ a two-stage optimization pipeline consisting of supervised fine-tuning and preference-guided reinforcement learning.

Supervised Fine-Tuning. We first initialize the base models via supervised fine-tuning. For Qwen2.5, we directly fine-tune the model on the CSDS dataset using standard dialogue–summary supervision. For Qwen3, we adopt a two-stage SFT strategy on the SAMSum dataset: the model is first fine-tuned on the original data without reasoning annotations, and subsequently fine-tuned on an augmented version containing distilled reasoning signals. This procedure provides a stable initialization for reinforcement learning.

Guided Reinforcement Learning with Preferences. After supervised fine-tuning, we further optimize the model using Guided Reinforcement Learning with Preferences (GRPO). For each input dialogue, the model samples a group of G candidate summaries, and optimization is performed based on their relative rewards. This stage encourages the model to generate summaries that are not only semantically informative but also factually consistent and aligned with human judgment.

B Evaluation Metrics Details and Modification

We evaluate model performance using automatic metrics with language-aware configurations to ensure fair and accurate assessment across multilingual dialogue datasets. Unlike prior work that applies a unified evaluation toolkit regardless of language, we explicitly distinguish between Chinese and English inputs and adopt language-specific implementations for both ROUGE and BERTScore.

ROUGE. As summarized in Table 2, we first identify the dialogue language using reference-based language detection. For Chinese samples, we apply the `rouge_chinese` toolkit with THULAC-based word-level tokenization, rather than character-level segmentation. For English samples, we use the standard `py-rouge` implementation. In both cases, we report ROUGE-1, ROUGE-2, and ROUGE-L scores. This design avoids inaccuracies caused by applying non-adapted ROUGE variants to languages with different tokenization characteristics.

BERTScore. BERTScore is computed following the protocol in Table 3, with language-specific backbone models. We use BERT-base Chinese for Chinese texts and BERT-base uncased for English texts, and aggregate similarity scores across the top 12 transformer layers. Scores are computed using cosine similarity without IDF weighting, and we report the mean precision, recall, and F1 scores. Language detection is performed consistently with the ROUGE evaluation to ensure aligned metric computation.

Faithfulness. In addition to lexical and semantic similarity metrics, we evaluate factual faithfulness to assess whether the generated summaries are consistent with the source dialogue content. Faithfulness measures the degree to which the factual statements in a model-generated summary are supported by the corresponding dialogue context, with scores ranging from 0 to 1, where higher values indicate better factual consistency.

Formally, a summary is considered faithful if all of its factual claims can be inferred from the source dialogue. The faithfulness score is computed as the proportion of supported claims among all identified claims in the summary:

$$\text{Faithfulness} = \frac{\text{\# supported claims}}{\text{\# total claims in the summary}}. \tag{2}$$

To operationalize this metric in the multi-role, multi-turn dialogue summarization setting, we first decompose each generated summary into a set of atomic factual statements. Each statement is then individually verified against the original dialogue context. For this verification step, we employ **HHEM-2.1-Open** (Li et al., 2024), a T5-based classifier model released by Vectara, which is specifically trained to detect hallucinations in LLM-generated text. Given a candidate statement and the dialogue context, the model predicts whether the statement can be inferred from the context. The final faithfulness score is obtained by averaging the binary verification results over all statements in the summary.

Human Preference. To further assess summary quality beyond automatic similarity and factual consistency metrics, we incorporate a human preference evaluation. This metric aims to capture holistic quality aspects that are difficult to quantify au-

tomatically, including informativeness, coherence, readability, and overall usefulness of the summary.

Human preference scores are computed using the **WorldPM-72B-RLHFLOW** (Wang et al., 2025) model, a large-scale preference model trained on extensive human feedback data. The training corpus of WorldPM-72B-RLHFLOW includes a wide range of instruction-following and summarization tasks, enabling it to reliably approximate human judgments in dialogue summarization scenarios. Given a pair consisting of the source dialogue and a candidate summary, the model produces a scalar preference score reflecting the likelihood that a human annotator would favor the summary.

We report the average human preference score across all evaluation samples. By combining human preference evaluation with ROUGE, BERTScore, and faithfulness metrics, we provide a more comprehensive assessment of summary quality, especially in multi-role dialogue settings where factual accuracy and overall coherence are both critical.

Overall, this language-aware evaluation protocol provides a more reliable comparison across multilingual dialogue summaries and reduces potential bias introduced by mismatched evaluation tools.

C Distillation Prompt for Multi-role Dialogue Summarization

Multi-role Dialogue Summarization Distillation Prompt

You are a large language model participating in a *knowledge distillation* process for multi-role, multi-turn dialogue summarization.

Task Definition. Given a dialogue and its corresponding *reference summary* (generated by a stronger teacher model or human annotators), your goal is to reproduce a high-quality summary by explicitly reasoning over the dialogue content. The reference summary is provided *only as supervision* and should not be copied verbatim.

Reasoning Requirement. Before producing the final summary, you must explicitly perform structured reasoning by outputting a `<think></think>` block that follows the four steps below:

1. **Dialogue Examination**:
   Identify speakers, dialogue turns, and key utterances across the multi-role interaction.
2. **Motivational Inference**:
   Infer speakers' intentions, preferences, or implicit goals behind their statements.
3. **Core Content Abstraction**:
   Extract and abstract the essential information, including requests, responses, agreements, or conflicts.
4. **Consistency and Fidelity Verification**:
   Verify that the abstracted content is faithful to the original dialogue and aligned with the reference summary, avoiding omissions, distortions, or hallucinated information.

Example Input:

```
Conversation:
Amanda: I baked cookies. Do you
    want some?
Jerry: Sure! Amanda: I'll bring
    you tomorrow :-)

Summary:
Amanda baked cookies and will
    bring Jerry some tomorrow.
```

Example Output:

```
<think>I observed Amanda stating
    she baked cookies and offering
    some to Jerry, who accepted,
    followed by Amanda confirming
    she will bring them tomorrow;
    this sequence shows a clear
    offer, acceptance, and planned
    delivery, leading me to
    conclude Amanda baked cookies
    and will share them with Jerry
    the next day.</think>
Amanda baked cookies and will
    bring Jerry some tomorrow.
```

Note. This prompt is used exclusively during the *distillation stage* to enable the student model to learn structured reasoning behaviors when trained on the SAMSum dataset, whose original annotations contain summaries without explicit thinking or reasoning traces.

D Summary Reward Principle

Summary Reward Principle

Evaluate summaries using these criteria:

1. **Key Information Coverage (40%)** - Does the summary capture the core facts of the dialogue? - Must include: the request/proposal, the refusal, the insistence, and any implied motivation if present. - Missing major elements is a critical error.

2. **Inference and Implicit Understanding (30%)** - Does the summary correctly reflect implied attitudes, motives, or emotional tone (e.g., sarcasm, concern, frustration)? - Reasonable inference is rewarded. Fabrication is penalized.

3. **Faithfulness and Precision (20%)** - No hallucinations or incorrect claims beyond what can be safely inferred. - The summary must not change the meaning of the original dialogue.

4. **Conciseness and Clarity (10%)** - The summary should be brief, readable, and well-structured. - Overly verbose summaries lose points even if factually correct.

Priority: Key information coverage $>$ faithfulness $>$ inference quality $>$ clarity.

E Principle-based Scorer Prompt Template

This appendix provides the prompt template used for principle-based evaluation. The evaluator is instructed to assess model outputs strictly according to a predefined evaluation principle.

System Prompt for Principle-based Scorer

SYSTEM PROMPT
You are an evaluator judging model responses based on a given **evaluation principle**. Your primary goal is to assess how well each response adheres to the principle, prioritizing this over general preferences, while avoiding endorsement of harmful content.

Evaluation Procedure:

1. Carefully read the principle, the input prompt, and all candidate responses. Briefly consider how each response aligns with the principle in a concise reasoning process.

2. Assign a score to each response on a 1–5 scale:
   - 5: Perfect adherence with excellent overall quality
   - 4: Strong adherence with minor limitations
   - 3: Basic adherence
   - 2: Partial adherence with important omissions
   - 1: Poor adherence or contradiction to the principle

3. Rank all responses from best to worst. Responses with identical scores should still be ordered based on relative quality.

Use the full scoring range when necessary. Avoid score compression if there are clear quality differences among responses.

Output Format (JSON only):

```
{
  "scores": {"model-1": 2, "model-2": 4, ...},
  "best-to-worst": ["model-2", "model-1", ...]
}
```