FOCUS: A Fine-Grained Customer-Oriented Sentiment Dialogue Summarization Dataset for Chinese Customer Service

Anonymous ACL submission

Abstract

Dialogue summarization (DS) plays a vital role in improving customer service efficiency by automatically generating concise summaries from lengthy multi-turn dialogues. However, existing studies largely overlook the fine-grained sentiment dynamics expressed by customers, and most DS datasets lack detailed sentiment annotations. These limitations hinder both accurate service quality assessment and the development of sentiment-aware summarization models. To address these challenges, we propose a three-stage approach to building an aspect-aware sentiment dataset, comprising: (1) aspect-anchored dialogue rewriting, (2) dialogue-anchored explainable label generation, and (3) label-dialogue integrated summarization. Building upon this scheme, we construct FOCUS, a **F**ine-grained customer-**O**riented **C**hinese dialog**U**e **S**ummarization dataset. FOCUS is the first Chinese dataset with 12,948 dialogues annotated for multi-level aspects, sentiment polarity, opinion content, emotions, as well as customer-oriented formatted and free-style sentiment summaries. To demonstrate the challenges and utility of FOCUS, we benchmark a range of summarization models on FOCUS and observe that current methods often exhibit misalignment between aspects and sentiments. Meanwhile, we find that a Chain-of-Thought approach can enhance faithfulness and interpretability, highlighting promising directions for future research on this dataset. FOCUS serves as a valuable resource to advance research in sentiment-aware DS and related tasks. Code: `https://anonymous.4open.science/r/FOCUS-6D6F`

1 Introduction

In recent years, AI-powered dialogue agents have been widely adopted in customer service, offering scalable and cost-effective solutions for handling large volumes of user interactions (Dias et al., 2022). However, these systems often exhibit limitations in accurately detecting users' sentiment and emotional states (Brun et al., 2025), potentially resulting in impersonal or contextually inadequate responses. This shortcoming can diminish user satisfaction and hinder merchants' ability to evaluate the performance of their deployed agents.

In this context, customer-oriented dialogue summarization (DS) serves as a critical tool—not only for distilling key information from lengthy multi-turn dialogues, but also for enabling downstream analysis of user sentiment and system performance. For instance, generating concise summaries that reflect a customer's sentiment attitude toward different aspects of the service (e.g., delivery or refunds) allows service provider to better understand user needs and monitor AI-agent behavior.

However, most existing DS research has focused on domains such as meetings (Kirstein et al., 2025) or doctor-patient dialogue (Song et al., 2020), which emphasize the objective factual information. Unlike other domains, customer service dialogue in the e-commerce domain exhibit subjective characteristics covering multi-aspect, and dynamic shifts in sentiment states. These traits pose unique challenges for the task of DS. While early efforts like JDDC (Chen et al., 2020) and CSDS (Lin et al., 2021) have contributed large-scale dialogue datasets for this domain, they lack critical sentiment annotations, and their summary formats are overly simplistic (e.g., role-specific QA pairs). Such limitations pose challenges to capturing the subtle sentiment dynamics required for assessing customer satisfaction and improving quality of service systems.

To solve the above issues, inspired by ESCoT (Zhang et al., 2024), we propose a scalable, three-stage dataset construction scheme based on LLMs to trade off data privacy and annotation costs in e-commerce scenarios. Based on this scheme, we construct FOCUS, a Fine-grained Customer-Oriented Chinese Sentiment Dialogue Summarization Dataset, one sample of which is shown in

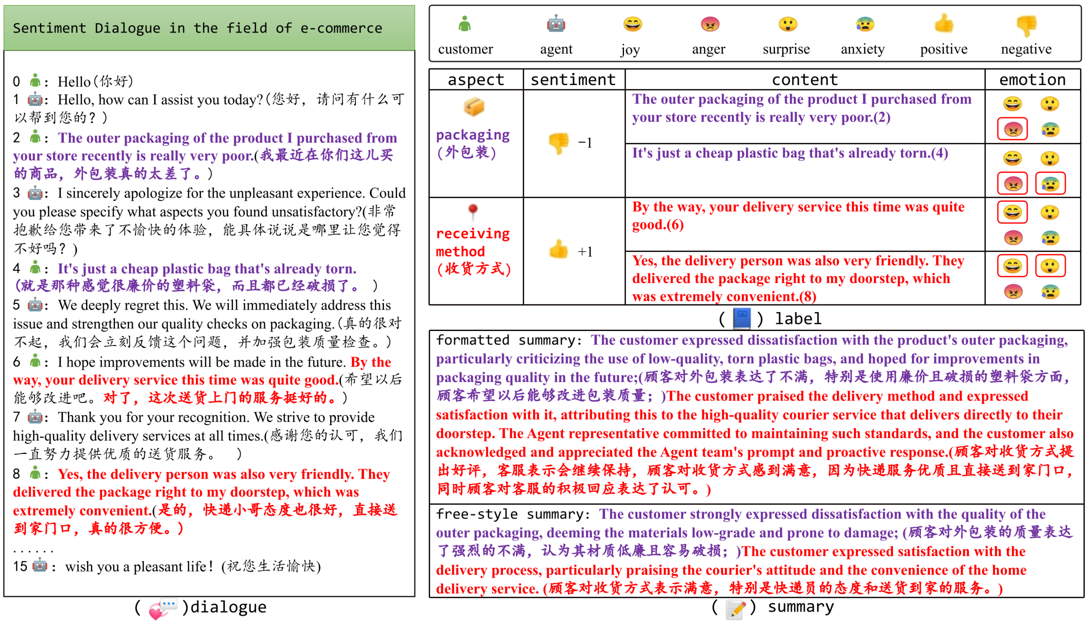

Figure 1: An example from FOCUS dataset. Sentences with the same color (purple and red) represent key informations with the same aspect (purple for *packaging* and red for *receiving method*). The numbers contained in parentheses in the content of *label* block denote indexes of key utterances.

Figure 1. Each sample in FOCUS consists of three blocks, i.e., *dialogue*, *label* and *summary*. *Dialogue* block consists of multi-turn utterances, packed with sentiment, between a customer and an agent. In *label* block, the content field is used to record the origin of a specific sentiment expressed by the customer towards corresponding aspect, which can be used as an important reference for generating explainable summary. To complement sentiment analysis, we categorize four emotion types reflecting customers' transient states, enabling dialogue systems to better detect emotional shifts. The *summary* block contains two distinct types of summaries, i.e., formatted and free-style summary. Unlike existing resources, FOCUS is designed as a multi-task dataset to support four key sentiment-centric tasks: (1) fine-grained sentiment analysis (using *dialogue* and *label* in Figure 1), (2) emotion attribution analysis (using *content* and *emotion*), (3) empathetic response generation (using *label*), and (4) sentiment dialogue summarization (using *dialogue* and *summary*).

Existing approaches consider summary generation a black-box process (Tian et al., 2024), resulting in limited consistency control and explainability. These limitations impede practical deployment in customer service, where missing key complaints and opaque reasoning are critical concerns. Wei et al. (2022) first proposed using CoT prompting to elicit reasoning in LLMs. Later, CoT is widely used to improve model performance, including explainability and faithfulness (Nachane et al., 2024; Jacovi et al., 2024). Inspired by CoT, we propose COTS[2] as a task-aligned baseline that makes aspect–sentiment reasoning explicit, serving to better diagnose faithfulness issues on FOCUS.

Our contributions are summarized as follows:

- **Dataset.** We release FOCUS, a Chinese customer-service COSDS resource with a **12,948-dialogue checked corpus** and a **5,688-dialogue human-refined benchmark subset**, annotated with fine-grained aspects, aspect-level sentiment polarity, opinion evidence, and four emotion types, along with **dual-style** (formatted and free-style) customer-oriented summaries.

- **Task and evaluation.** We formalize **Customer-Oriented Sentiment Dialogue Summarization (COSDS)** and propose task-aligned evaluation via **Aspect Coverage Rate (ACR)** and **Sentiment Accuracy per Aspect (SAA)** (both objective and human-scored), complementing standard ROUGE/BERTScore for fair comparison.

- **Benchmarks and findings.** We benchmark strong PLMs/LLMs on FOCUS and show that current models often fail at multi-aspect

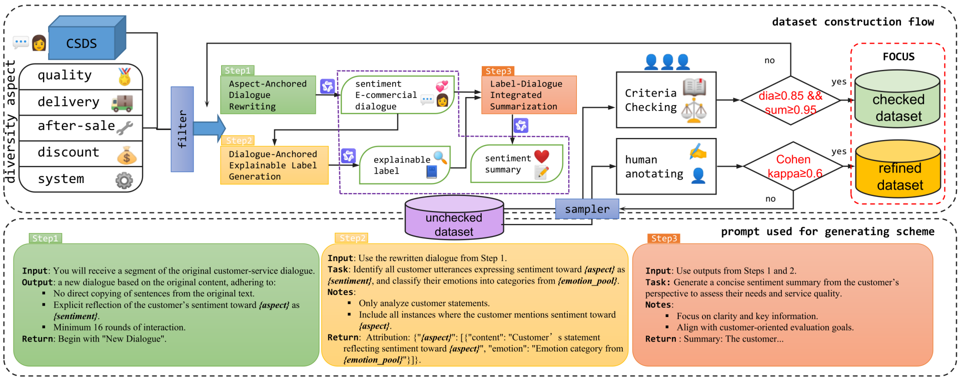

Figure 2: FOCUS construction flow (top) and prompting scheme (bottom). The LLM-generated pool (**unchecked**) is filtered into the released **checked** set (**12,948** dialogues). We then sample 6,000 checked dialogues for manual correction and retain **5,688** as the high-precision **refined** subset used for the main benchmarks.

sentiment disentanglement (e.g., aspect omissions and sentiment flips), leading to a marked misalignment between standard automatic metrics and human judgments. We further introduce **COTS**[2] as a **task-aligned baseline** that makes aspect–sentiment reasoning explicit and improves faithfulness and aspect-level correctness.

2 Data Construction

We will discuss the construction flow on how our dataset is constructed in this section, including source data collection, sample generation, and quality control, which is illustrated in Figure 2.

2.1 Data Source

Our dataset builds on the real-world CSDS (Lin et al., 2021) dataset from JD.com[1], ensuring FOCUS's authenticity. During the selection process, we follow rigorous filter criteria: prioritizing lengthy dialogues to ensure substantial informational content and greater summarization challenges; ensuring a diverse and evenly distributed range of aspects; and verifying semantic completeness across all dialogues. This guarantees the quality of our initial data.

2.2 Sentiment dialogue sample generation

Given the high cost of manual annotation and rewriting, we leverage LLM to automatically rewrite dialogues, aiming to preserve core aspects from real-world scenarios while enriching them with sentiment information. While this process introduces stylistic artifacts inherent to LLM generation, our five complementary alignment tests demonstrate that FOCUS closely matches real dialogues (see Section 2.6). LLM-based data generation strategy is widely adopted in top academic conference (Zhang et al., 2024). The dataset construction process consists of four main steps: (1) Identifying common aspects in the e-commerce domain that are likely to elicit user sentiments. (2) Starting from existing e-commerce dialogues, leveraging LLMs to integrate aspect and sentiment information into the dialogues and rewrite them accordingly, thereby generating sentiment-aware dialogues (referred to as *dialogue* in Figure 1). (3) Employing LLMs to locate specific sentences (referred to as *content* in Figure 1) within the sentiment-aware dialogues from step (2) where customers express sentiment toward a given aspect , and analyzing the emotions (referred to as *emotion* in Figure 1) conveyed in these sentences. (4) Generating a sentiment summary (referred to as *summary* in Figure 1) for customers based on dialogues obtained from (2) and the *content* obtained from (3).

Diversity Aspects We define five representative coarse-grained aspects in the e-commerce domain: *Quality*, *Delivery*, *After-sale*, *Discounts*, and *System*. Based on these categories, we select several relevant aspects from the CSDS dataset that are likely to reflect customer sentiment. The resulting two-level aspect framework comprises 25 fine-grained aspects, as presented in Table 1.

Generation of Dialogue Data After determining all the aspects, we utilize Qwen[2] to rewrite dialogues. To ensure aspect diversity and a balanced distribution, we constrain the dataset such that dialogues with one aspect comprise 40%, two

[1]https://www.jd.com

[2]https://bailian.console.aliyun.com

Quality	Delivery	After-sale	Discounts	System
packaging	delivery method	returns	membership discounts	order
missing accessories	delivery cycle	refund	threshold discounts	shopping cart
color issue	shipping damage	warranty	points discounts	payment
authenticity	incorrect delivery	after-sale installation	price protection	invoice
shelf life	receiving method	reshipment	cashback	account

Table 1: Sentiment-oriented aspects table.

aspects 50%, and three or four aspects combined account for the remaining 10%. This aspect-count distribution is **challenge-oriented**: it intentionally up-weights multi-aspect cases to stress-test models' ability to disentangle aspect-level sentiments, rather than mirroring the long-tailed frequencies of raw service logs. We designed an algorithm to ensure the diversity of numbers in each dialogue (see Appendix A) and proposed an aspect-aware sentiment DS generation scheme based on LLMs, consisting of three steps: aspect-anchored dialogue rewriting, dialogue-anchored explainable label generation, and label-dialogue integrated summarization. The prompt to each step for generating scheme is illustrated in Figure 2. (1) **Aspect-anchored Dialogue Rewriting**: First, we rewrite the extracted dialogues, integrating the aspects and sentiment into the dialogues during this process. Since we conducted a few preliminary tests before generation, we found that the model sometimes copied the original text and produced too few turns, both of which could lead to lower quality of the generated dialogues. Therefore, we added constraints to the generation process in the prompts. (2) **Dialogue-anchored Explainable Label Generation**: Second, based on the dialogues obtained in the first step, we ask the LLM to identify the sentences that express sentiment regarding the aspect and analyze what kind of emotions these sentences convey, so that our dataset has good interpretability. As the model tended to overlook emotional customer utterances or focus on agent responses, we further refined the prompts to guide its attention. (3) **Label-Dialogue Integrated Summarization**: Finally, based on the dialogues obtained in the first step and the attribution sentences generated in the second step, we generate the final customer-oriented summary. To mitigate overly verbose outputs, we apply length constraints during generation.

2.3 Summary Style

We observe that, without explicit style instructions, LLM-generated summaries exhibit diverse forms but often lack sufficient detail to meet user needs. We refer to these as *free-style* summaries and examine which summary style best aligns with the objectives of our task.

To better capture real-world diversity, we introduce additional summary styles. Based on extensive analysis, we argue that an effective dialogue summary should identify customer needs, enhance service quality, and track service progress in e-commerce scenarios. Importantly, the summary style should adapt to the customer's sentiment toward each aspect. For negative sentiment, the summary should highlight its cause and the customer's expected resolution. For positive sentiment, it should clarify whether satisfaction stems from the dynamic service process or the aspect itself. Accordingly, we design two summary templates, as shown in Table 2.

Sentiment	Template
Negative	The customer expressed {specific opinions or issues} about ***{aspect}***, particularly regarding {specific details}. The customer hopes for {desired outcome or solution}.
Positive	The customer raised {specific opinions or issues} about ***{aspect}***, and the agent {handling process}. The customer was satisfied with ***{aspect}*** because of {reasons why this aspect satisfied the customer}. Meanwhile, the customer expressed {attitude} towards {the actions of the agent}.

Table 2: Templates for summarizing dialogues with different sentiment tendencies. The negative template focuses on complaints, specific reasons, and customer expectations; the positive template highlights service experiences and customer satisfaction.

We provide both formatted and free-style summaries. A pilot study suggests formatted summaries are preferred for low-aspect dialogues, while free-style summaries better handle multi-aspect cases; details are in Appendix C.

Subset	#Dialogues	Supervision	Primary use
Checked	12,948	LLM labels + 5,000-sample audit	Released corpus; statistics; optional training
Refined	5,688	Human-corrected (from 6,000)	Main benchmarks (train/dev/test)

Table 3: FOCUS subsets used in this paper. The refined subset is sampled from the checked set and manually corrected for high-precision evaluation.

Dataset	Lang.	# Dia	Role Sum.	Aspec.	Senti.	Emoti.
AMI (Carletta et al., 2006)	EN	137	No	No	No	No
SAMSum (Gliwa et al., 2019)	EN	16,369	No	No	No	No
HET-MC (Song et al., 2020)	ZH	44,983	Yes	No	No	No
TOSDS (Zou et al., 2021)	ZH	18,860	No	No	No	No
CSDS (Lin et al., 2021)	ZH	10,701	Yes	Yes	No	No
JDDC 2.1 (Zhao et al., 2022)	ZH	246,000	Yes	No	No	No
ClidSum (Wang et al., 2022)	EN→ZH/DE	67,000	No	No	No	No
MDS (Liu et al., 2024)	EN/ZH	11,305	No	No	No	No
MISP-Meeting (HangChen et al., 2025)	ZH	163	No	No	No	No
FOCUS	ZH	12,948	Yes	Yes	Yes	Yes

Table 4: Comparison of different DS datasets. # Dia. represents total number of dialogues in each dataset. Aspec., Senti., and Emoti. indicate whether the dataset contains information related to aspect, sentiment, and emotion, respectively.

2.4 Check and Annotation

We audit 5,000 unchecked instances with 3 annotators using dialogue-level and summary-level rubrics; agreement is substantial (Fleiss' κ in [0.59, 0.71]) and averaged scores indicate high quality (details in Appendix B).

To improve the real-world applicability of our experiments, we manually refined a portion of the data during quality checking to better simulate real-world dialogues. Specifically, we randomly sampled 6,000 instances from the **checked dataset** for additional human annotation, correcting inaccurate aspect and emotion labels where necessary. After filtering, 5,688 high-quality instances were retained to form the **refined dataset**.

2.5 Dataset Statistics and Comparison

The FOCUS dataset consists of the **checked** and **refined** dataset, with the former serving as the core component and the latter tailored for real-world evaluation. Table 3 summarizes the size and role of each subset. In this section, we focus on analyzing the **checked dataset**. A comparison with existing datasets is provided in Table 4, where, to the best of our knowledge, FOCUS is the first Chinese sentiment DS dataset annotated with aspects, sentiments, and emotions.

We further present key statistics of the dataset: Figure 3 (a) shows the frequency of 25 aspects and their sentiment distributions, highlighting the aspects most frequently discussed or complained about; Figure 3 (b) presents the distribution of dialogue turns, indicating the typical length and complexity of customer service interactions; Figure 3 (c) illustrates the relationship between aspects and emotions, capturing the emotional nuances associated with different aspects; Figure 3 (d) shows the co-occurrence of emotions, demonstrating the dataset's ability to reflect complex, mixed emotional states. Additional statistics for both the checked and refined dataset are provided in the Appendix D.

2.6 Validation of Data Authenticity and Real-World Alignment

Because FOCUS includes LLM-rewritten dialogues based on real CSDS interactions, we first checked whether these rewrites introduced subtle style artifacts before running benchmarks. We compared FOCUS with 3000 held-out authentic CSDS dialogues using five tests: (1) conversation structure (Kolmogorov–Smirnov on turn counts), (2) aspect and polarity prevalence (Pearson r over 25 aspects), (3) sentiment dynamics within dialogues (polarity-shift rate), (4) stylometric n-gram divergence (character 4-gram JSD), and (5) a style-artifact detector (linear classifier AUC on stylometric features). Results show synthetic and real dialogues to be essentially the same: KS D=0.024 (p>0.05); Pearson r=0.92/0.89; shift rates 18.4% vs. 17.8%; median JSD 0.013 (95% CI $[0.012, 0.016]$); and detector AUC 0.54 (95% CI $[0.50, 0.58]$). Full details and acceptance criteria are provided in the Appendix I, supporting the authenticity and reliability of FOCUS.

3 Experiment

3.1 Task Definition

We define the **Customer-Oriented Sentiment Dialogue Summarization (COSDS)** task as follows. Given a multi-turn dialogue

$$D = \{(r_1, u_1), (r_2, u_2), \ldots, (r_n, u_n)\},$$

where each utterance u_i is associated with a speaker role $r_i \in \{\texttt{customer}, \texttt{agent}\}$, the goal is to generate a customer-oriented sentiment summary

$$S = \{(a_1, s_1, x_1), (a_2, s_2, x_2), \ldots, (a_m, s_m, x_m)\}.$$

Each segment (a_j, s_j, x_j) denotes an aspect a_j, its sentiment polarity $s_j \in \{\texttt{positive}, \texttt{negative}\}$, and an explanation x_j of the reason and expectation.

The summary focuses on customer opinions and is structured to reflect aspect-specific sentiment with explanatory context.

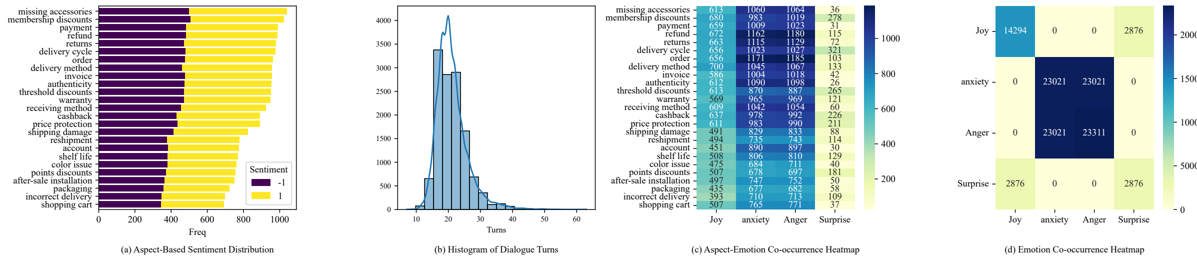

Figure 3: Analysis of FOCUS.

3.2 Evaluation of Models and Metrics

Compared to traditional DS tasks, COSDS is more challenging. In this subsection, we select baseline models as follows. **PLMs**: 1) mengzi-t5-base (T5) (Zhang et al., 2021), 2) bart-large-chinese (BART) (Yunfan SHAO, 2024); **LLMs**: 3) Llama3-Chinese-8B-Instruct (Llama) (Grattafiori et al., 2024), 4) DeepSeek-R1-Distill-Qwen-14B (DeepSeek) (DeepSeek-AI, 2025).

For evaluation metrics, we conduct comprehensive evaluations in 3 different ways(i.e., objective, human and ChatGLM (GLM-4) evaluation). Objective evaluation includes: ROUGE (Lin, 2004), BERTScore (Zhang* et al., 2020). Subjective evaluation include Faithfulness, Fluency, Informativeness, and Conciseness (Gao et al., 2023).

We report ROUGE/BERTScore mainly for comparability with prior DS work; however, they are insensitive to **aspect-level sentiment correctness**, which is central to COSDS. To evaluate the quality of COSDS, we propose two task-specific metrics: **Aspect Coverage Rate (ACR)** and **Sentiment Accuracy per Aspect (SAA)**. ACR measures whether the key aspects discussed in the dialogue are retained in the summary, while SAA evaluates whether the sentiment associated with each aspect is correctly preserved. Our evaluation combines both human judgments and automatic scores. Detailed definitions and computation procedures are provided in the Appendix G.

3.3 Experiment Setup

We conduct experiments on the **refined dataset** from FOCUS. We use the refined subset for benchmarking because our aspect-aware metrics (ACR/SAA) require high-precision ground truth. To verify robustness to scale and label noise, we additionally train on the full checked set and evaluate on the refined test set; the conclusions remain unchanged (Appendix J). The dataset is split into a training set (80%), a validation set (10%), and a test set (10%) for both *formatted* and *free-style* fine-tuning strategies.

Model Training We fine-tune PLMs with full FT and LLMs with LoRA under standard settings; full hyperparameters and prompts are in Appendix F.

3.4 Result and Analysis

3.4.1 Multi-view Evaluation

We report the overall performance of all models on both *formatted* and *free-Style* summarization settings in Table 5, evaluated by both automatic metrics and human judgments. Human evaluations were conducted by three annotators with NLP expertise. Inter-annotator agreement, measured via Fleiss Kappa, averaged 0.82, indicating high reliability. Additional evaluation results for ChatGLM are provided in the Appendix H. We summarized our findings from following aspects:

1) **In the *formatted* setting**, Our analysis reveals that while LLMs like DeepSeek achieve high ROUGE scores, they can struggle with faithfulness and sentiment accuracy. An explicit reasoning approach, as used in our COTS^2 baseline, demonstrates significant improvements in human-evaluated metrics like Faithfulness (4.90), Fluency (4.93), Informativeness (4.70), ACR (4.50), and SAA (4.80), indicating that the structured reasoning path is a key challenge presented by our dataset.

2) **In the *free-Style* setting**, COTS^2 continues to outperform other baselines in human evaluation. It surpasses all models in Faithfulness (4.80), Fluency (4.97), and SAA (4.75), indicating that COTS^2 remains robust even in less constrained generation styles. Moreover, it maintains competitive performance on ROUGE-1 (51.89), better than most other models.

3) **Existing models are prone to mismatch between aspects and sentiments** in COSDS. In contrast, our COTS^2, by leveraging an explicit reasoning path guided by label information, significantly

enhances both controllability and accuracy in complex sentiment summarization tasks.

4) **We observe a clear misalignment between automatic metrics and human evaluations** in both settings. In particular, some models (e.g., T5) obtain unexpectedly high ROUGE despite weaker human preference, likely due to issues like content repetition. We analyze such discrepancies through three representative case studies (see Appendix L).

Why standard metrics misalign with COSDS In COSDS, a summary can achieve high n-gram overlap by copying aspect keywords while **flipping sentiment polarity** or introducing unsupported claims—errors that are fatal for customer-service diagnosis but largely invisible to ROUGE/BERTScore. Table 6 decomposes common failure types. We find that models with higher ROUGE may still have higher rates of (i) missing gold aspects, (ii) sentiment flips on covered aspects, whereas COTS^2 reduces these errors and aligns better with human faithfulness.

We further test transfer to CSDS and domain-adaptive training on JDDC; results suggest COTS^2 scales and transfers beyond FOCUS (Appendix K)

3.4.2 Multi-aspect Evaluation

To investigate the robustness of models under different levels of aspect complexity, we assess the model's performance on dialogue samples that include one to four aspects in formatted summary setting. As shown in Figure 4, **all models exhibit performance fluctuations as the number of aspects increases, revealing fluctuations that highlight the inherent challenges of the task**. Specifically, traditional PLMs such as T5 and BART show a noticeable decline in all metrics with more aspects involved, indicating difficulty in capturing multi-aspect information consistently. In contrast, LLMs demonstrate stronger resilience. For example, COTS^2 and DeepSeek maintain competitive ROUGE and BERTScore across different aspect settings, even achieving performance gains when moving from single to double-aspect samples. These results highlight that while increasing aspect numbers poses significant challenges for summarization, LLMs can better adapt to complex semantic structures, showing potential in real-world multi-aspect dialogue scenarios. We provide free-style results in Appendix H.

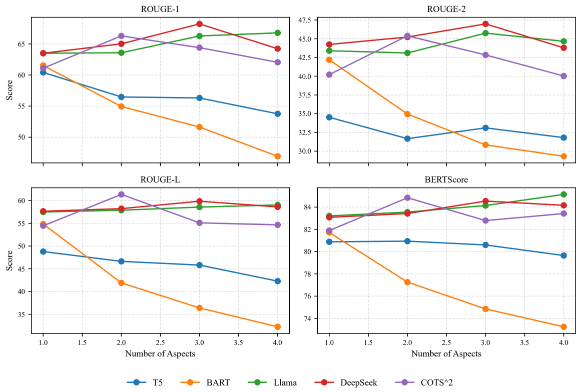

Figure 4: Model performance across dialogue samples containing one to four aspects under the formatted setting. Metrics include ROUGE-1/2/L and BERTScore.

3.4.3 Ablation and Exploration of CoT

Building on the superior performance of COTS^2 in Faithfulness, ACR, and SAA, we conducted comprehensive ablation studies to identify the optimal CoT structure. We compared our fixed three-step CoT against several variants, including step removal, order permutations, a two-step Short CoT, and CoTs generated by external LLMs (DeepSeek-R1, QwQ-32B). All experiments used DeepSeek-R1-Distill-Qwen-14B as the base model.The templates generated based on FOCUS labels and those generated by LLMs are both provided in the Appendix E.

The results in Table 7 unequivocally demonstrate the superiority of our proposed CoT. We found that the canonical order is critical, as any permutation degraded performance, especially those starting with step S. Furthermore, each step proved necessary; removing any step (A, O, or S) harmed performance, with the omission of A causing the largest drop. Simpler variant, such as the Short CoT was progressively less effective. Finally, CoTs generated by external LLMs also underperformed, likely because their over-extended and instance-specific reasoning failed to generalize well for our base model and data scale.

4 Related Work

Datasets DS is particularly challenging due to its multi-turn structure, overlapping speaker roles, and implicit sentiment shifts. Existing datasets span diverse domains such as political debates (Rennard et al., 2023), multilingual conversations (Wang et al., 2022; Feng et al., 2022), and clinical interactions (Song et al., 2020). In e-commerce, Chen et al.

Type	Models	Params	Objective metrics						Human evaluation					
			R-1	R-2	R-L	BS	ACR_{obj}	SAA_{obj}	Fai.	Flu.	Inf.	Con.	ACR_{hum}	SAA_{hum}
Formatted	T5	220M	61.64	35.86	51.78	82.49	69.80	75.40	$4.53_{(0.7)}$	$4.43_{(0.7)}$	$3.96_{(0.8)}$	$4.36_{(0.8)}$	$3.86_{(1.1)}$	$4.13_{(1.0)}$
	BART	406M	57.18	35.72	45.65	78.62	68.10	73.20	$4.60_{(0.7)}$	$4.70_{(0.7)}$	$3.53_{(0.7)}$	$4.80_{(0.3)}$	$3.56_{(1.0)}$	$3.83_{(0.9)}$
	Llama	8B	64.51	44.00	58.02	**83.72**	79.60	84.30	$4.70_{(0.6)}$	$4.90_{(0.4)}$	$4.63_{(0.7)}$	$4.53_{(0.5)}$	$4.46_{(0.8)}$	$4.60_{(0.7)}$
	DeepSeek	14B	**65.34**	**45.28**	**58.50**	83.67	78.90	83.90	$4.76_{(0.4)}$	$4.90_{(0.3)}$	$4.60_{(0.7)}$	$\mathbf{4.83}_{(0.3)}$	$4.43_{(0.8)}$	$4.60_{(0.6)}$
	COTS^2	14B	63.92	42.77	57.13	83.31	**81.80**	**86.20**	$\mathbf{4.90}_{(0.3)}$	$\mathbf{4.93}_{(0.2)}$	$\mathbf{4.70}_{(0.5)}$	$4.63_{(0.4)}$	$\mathbf{4.50}_{(0.7)}$	$\mathbf{4.80}_{(0.4)}$
Free-Style	T5	220M	51.75	**29.12**	43.19	79.53	59.20	62.30	$4.13_{(1.1)}$	$4.23_{(0.9)}$	$4.17_{(0.9)}$	$3.90_{(1.2)}$	$4.57_{(0.7)}$	$4.23_{(1.1)}$
	BART	406M	51.82	26.47	42.98	78.98	58.70	62.10	$4.53_{(0.8)}$	$4.90_{(0.3)}$	$3.97_{(1.0)}$	$\mathbf{4.67}_{(0.7)}$	$4.30_{(1.1)}$	$4.40_{(1.0)}$
	Llama	8B	51.03	26.60	**43.85**	**79.94**	66.10	70.50	$4.57_{(0.7)}$	$4.87_{(0.3)}$	$4.37_{(0.8)}$	$4.57_{(0.6)}$	$4.53_{(0.9)}$	$4.63_{(0.8)}$
	DeepSeek	14B	50.78	24.67	42.69	79.31	65.00	69.80	$4.77_{(0.5)}$	$4.90_{(0.2)}$	$4.33_{(0.9)}$	$4.56_{(0.6)}$	$4.67_{(0.8)}$	$4.71_{(0.8)}$
	COTS^2	14B	**51.89**	26.55	43.73	79.59	**66.90**	**71.20**	$\mathbf{4.80}_{(0.5)}$	$\mathbf{4.97}_{(0.2)}$	$\mathbf{4.41}_{(1.0)}$	$4.58_{(0.3)}$	$\mathbf{4.69}_{(0.9)}$	$\mathbf{4.75}_{(0.8)}$

Table 5: Overall results in terms of Objective metrics and Human evaluation. The mean and standard deviation of the evaluation scores from human evaluation are reported.

Model	ROUGE-L↑	ACR_{obj} ↑	SAA_{obj} ↑	Aspect omission↓	Sentiment error↓
T5	51.78	69.80	75.40	30.20	24.60
BART	45.65	68.10	73.20	31.90	26.80
Llama	58.02	79.60	84.30	20.40	15.70
DeepSeek	58.50	78.90	83.90	21.10	16.10
COTS^2	57.13	81.80	86.20	18.20	13.80

Table 6: Error-type decomposition in the formatted setting. These error types are central to COSDS but are weakly reflected by ROUGE.

CoT Variant	R-1	R-2	R-L	BS	ACR_{obj}	SAA_{obj}
Ablations / Variants of our 3-step CoT						
Ours	**46.58**	**19.27**	**32.90**	**74.23**	**85.10**	**92.50**
w/o Aspect recognition	40.82	12.89	28.81	72.75	75.21	88.13
w/o Opinion localization	43.15	16.03	30.14	73.49	80.54	90.32
w/o Sentiment extraction	42.51	15.72	29.56	73.11	81.23	87.55
Order swap: O→A→S	45.76	18.55	32.20	73.90	84.55	91.84
Order swap: A→S→O	45.10	18.00	31.95	73.68	84.12	91.53
Order swap: O→S→A	45.15	18.05	32.00	73.70	84.19	91.60
Order swap: S→A→O	44.65	17.40	31.25	73.35	83.51	90.98
Order swap: S→O→A	44.38	17.20	31.05	73.28	83.30	90.71
Short CoT: A→S only	44.78	17.65	31.40	73.55	83.82	91.22
Short CoT: O→S only	44.36	17.28	31.05	73.32	83.41	90.83
External CoTs						
QwQ	42.34	11.90	28.57	73.01	78.90	89.54
DS-R1	40.15	12.51	28.56	72.58	77.34	88.91

Table 7: Ablation and variant results for our fixed 3-step CoT (aspect recognition → opinion content localization → sentiment extraction). Rows remove one step, swap order, or shorten the CoT; External CoTs are LLM-generated chains for comparison.

(2020) introduced JDDC, a large-scale Chinese dataset based on real-world customer service dialogues. Lin (2021) proposed CSDS, a fine-grained dataset emphasizing dialogue summarization with role information. JDDC 2.1 (Zhao et al., 2022) further expanded the scope to multimodal data with over 246K sessions and 507K images. In English, Feigenblat et al. (2021) released TWEETSUMM, a high-quality dataset of annotated summaries for customer service. However, these datasets lack sentiment-level annotations and often ignore customer emotion, limiting their utility for assessing service satisfaction or building sentiment-aware summarization models. In contrast, **FOCUS** introduces fine-grained aspect-sentiment labels, four emotion categories, and dual-style summaries, enabling research into explainable COSDS.

Methods Early DS approaches adapted document summarization techniques directly to dialogues (Gliwa et al., 2019). Later work proposed dialogue-specific enhancements, including the use of dialogue acts (Goo and Chen, 2018), key phrases and entity tracking (Narayan et al., 2021), and role-aware modeling (Lin et al., 2022). To improve performance, recent studies explored MoE architectures (Tian et al., 2024). However, most prior methods treat summarization as a black-box generation task (Zhong and Litman, 2025), lacking faithfulness, controllability, and interpretability. In contrast, **COTS^2** integrates explainable aspect-sentiment reasoning into the generation process. This improves alignment between user opinions, emotional tone, and the resulting summaries, setting new benchmarks for interpretable DS in real-world applications.

5 Conclusion

In this paper, we construct a novel Chinese customer-oriented sentiment dialogue summary dataset named FOCUS ,which is the first Chinese dataset with 12,948 dialogues annotated for multi-level aspects, sentiment polarity, opinion content, emotions, as well as customer-oriented summaries. We also do elaborate experiments on FOCUS and draw some instructive conclusions on method performance and dataset difficulties. Furthermore, we find that a CoT approach significantly enhances the explainability of the summary generation process, along with improvements in faithfulness and aspect matching accuracy. FOCUS serves as a multi-task dataset designed to support sentiment-centric summarization and related tasks. Future work will explore diverse downstream applications extending beyond DS, such as fine-grained sentiment analysis, emotion attribution analysis, and the generation of empathetic responses.

Limitations

Although FOCUS is packed with sentiment labels and summary styles, due to the high cost of manual annotation, over 10,000 raw data entries were generated by LLMs in FOCUS. Despite human checking is performed, only 5,688 samples were meticulously annotated by humans, reaching a certain quality standard. However, the overall quality of the dataset still requires further optimization.

Furthermore, the current scope of FOCUS is monolingual, exclusively featuring Chinese dialogues. This presents a clear opportunity for future work to extend the data construction methodology to other languages, thereby creating a valuable multilingual resource for sentiment-aware dialogue summarization. The dataset is also domain-specific, centered on e-commerce customer service. Consequently, the established aspect taxonomy and observed conversational patterns may not directly transfer to other domains like technical support, healthcare, or financial services, where customer needs and sentiment expression can differ significantly. Finally, our emotion analysis is confined to four categories (joy, anger, surprise, anxiety), which simplifies the rich spectrum of human emotional states. Future work could explore more granular emotion taxonomies to capture finer nuances in customer sentiment.

Although we introduce automated metrics for scalable evaluation, our analysis reveals critical limitations that motivate the need for future work in this area:

- **Lack of Precision in ACR:** The definition of ACR as $ACR = \frac{|A_{\text{gold}} \cap A_{\text{pred}}|}{|A_{\text{gold}}|}$ exclusively measures recall. It does not penalize the model for including irrelevant or factually incorrect aspects, thus failing to capture the precision of the generated summary.
- **Conditional Dependence of SAA:** The SAA metric evaluates sentiment accuracy only for the intersection of correctly predicted and gold-standard aspects ($|A_{\text{gold}} \cap A_{\text{pred}}|$). This makes it a conditional probability that can be misleading. For instance, a model with a very low ACR could still achieve a perfect SAA, masking its inability to identify most relevant aspects.
- **Absence of a Unified Score:** The two metrics operate independently, requiring a separate interpretation of aspect coverage and sentiment accuracy. This complicates model comparison and lacks a single, holistic score that captures the overall quality of the aspect-based summary.

These shortcomings underscore the need for a more comprehensive objective metric that is both efficient and more closely aligned with a nuanced human assessment of summary quality. Therefore, there is an urgent need to propose a more reasonable, efficient, and comprehensive objective evaluation metric for assessing summary quality.

Ethics Statement

Data and privacy. We build on an anonymized CSDS corpus; content is further paraphrased and filtered to reduce re-identification. We never attempt to link dialogues to real individuals, and examples are illustrative.

LLM rewriting and authenticity. To mitigate synthetic artifacts, we compare rewritten dialogues with held-out authentic data using complementary tests (structure statistics, aspect/polarity prevalence, shift rates, n-gram/stylometric divergence, and a small discriminator near chance).

Human evaluation. Annotators were trained, consented, compensated, and worked independently; agreement (Fleiss' κ) and evaluation rubrics are reported to support reproducibility.

Risks, intended use, and licensing. Labels may reflect domain biases and rewriting may shift distributions; users should report disaggregated results and avoid high-stakes use without validation. The release is for research only, prohibits re-identification or harmful use, and includes a takedown channel.

Compute and limitations. We favor parameter-efficient tuning and report configurations to aid transparency. Residual stylistic artifacts and subjective ratings may remain, and cultural/domain shifts can affect generalization.

References

Antonin Brun, Ruying Liu, Aryan Shukla, Frances Watson, and Jonathan Gratch. 2025. Exploring emotion-sensitive llm-based conversational ai. *Preprint*, arXiv:2502.08920.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav

Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI Meeting Corpus: A Pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg. Springer Berlin Heidelberg.

Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. The JDDC corpus: A large-scale multi-turn Chinese dialogue dataset for E-commerce customer service. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 459–466, Marseille, France. European Language Resources Association.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Isabel Dias, Ricardo Rei, Patrícia Pereira, and Luisa Coheur. 2022. Towards a sentiment-aware conversational agent. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, IVA '22, New York, NY, USA. Association for Computing Machinery.

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.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. MSAMSum: Towards benchmarking multi-lingual dialogue summarization. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 1–12, Dublin, Ireland. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382. Place: US Publisher: American Psychological Association.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *Preprint*, arXiv:2304.02554.

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.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, , and et. al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

HangChen HangChen, Chao-Han Huck Yang, Jia-Chen Gu, Sabato Marco Siniscalchi, and Jun Du. 2025. MISP-meeting: A real-world dataset with multimodal cues for long-form meeting transcription and summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15479–15492, Vienna, Austria. Association for Computational Linguistics.

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva. 2024. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4615–4634, Bangkok, Thailand. Association for Computational Linguistics.

Frederic Thomas Kirstein, Terry Lima Ruas, and Bela Gipp. 2025. Is my meeting summary good? estimating quality with a multi-LLM evaluator. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 561–574, Abu Dhabi, UAE. Association for Computational Linguistics.

Chihkai Lin. 2021. A corpus-based analysis of prosodic pauses in bă, gěi and ràng constructions in Taiwan Mandarin. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 640–645, Shanghai, China. Association for Computational Lingustics.

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.

Haitao Lin, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2022. Other roles matter! enhancing role-oriented dialogue summarization via role interactions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2558, Dublin, Ireland. Association for Computational Linguistics.

Zhipeng Liu, Xiaoming Zhang, Litian Zhang, and Zelong Yu. 2024. MDS: A fine-grained dataset for multi-modal dialogue summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11123–11137, Torino, Italia. ELRA and ICCL.

Saeel Sandeep Nachane, Ojas Gramopadhye, Prateek Chanda, Ganesh Ramakrishnan, Kshitij Sharad Jadhav, Yatin Nandwani, Dinesh Raghu, and Sachindra Joshi. 2024. Few shot chain-of-thought driven reasoning to prompt LLMs for open-ended medical question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 542–573, Miami, Florida, USA. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Virgile Rennard, Guokan Shang, Damien Grari, Julie Hunter, and Michalis Vazirgiannis. 2023. FREDSum: A dialogue summarization corpus for French political debates. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4241–4253, Singapore. Association for Computational Linguistics.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729, Barcelona, Spain (Online). International Committee on 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.

Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. ClidSum: A benchmark dataset for cross-lingual dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Yitao LIU Junqi DAI Hang YAN Fei YANG Zhe LI Hujun BAO Xipeng QIU Yunfan SHAO, Zhichao GENG. 2024. Cpt: a pre-trained unbalanced transformer for both chinese language understanding and generation. *SCIENCE CHINA Information Sciences*, 67(5):152102–.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. ESCoT: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412, Bangkok, Thailand. Association for Computational Linguistics.

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

Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. 2021. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese. *Preprint*, arXiv:2110.06696.

Nan Zhao, Haoran Li, Youzheng Wu, and Xiaodong He. 2022. JDDC 2.1: A multimodal Chinese dialogue dataset with joint tasks of query rewriting, response generation, discourse parsing, and summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12037–12051, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yang Zhong and Diane Litman. 2025. From information to insight: Leveraging LLMs for open aspect-based educational summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1914–1947, Vienna, Austria. Association for Computational Linguistics.

Yicheng Zou, Lujun Zhao, Yangyang Kang, Jun Lin, Minlong Peng, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2021. Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14665–14673.

A How to control the diversity of aspects

To ensure the diversity of aspects embedded in the dialogues and the controllability of dataset quality, we design the following rules. For single-aspect, we sequentially identify all dialogue samples corresponding to the aspect from the original CSDS and rewrite the dialogues by incorporating either positive or negative sentiment towards the aspect. For multi-aspect, taking double aspects as an example, we have a total of 25 fine-grained aspects, which allows for 300 possible combinations. Due to cost constraints, we plan to select 100 combinations out of 300. To achieve this, we devise Algorithm 1 to perform selection, in which we define a dynamic weight for each aspect. The weight is related to the number of times it has been selected. This ensures uniformity during the selection process. Each time, we select from a weight distribution, a process that guarantees randomness in aspect selection. Finally, when adding aspects to the set, we remove duplicate combinations, guarantees the uniqueness of aspect pair selection. For samples with triple aspects, we select 20 from 2300 combination. For samples with quadruple aspects, we select 5 combinations.

Algorithm 1 Uniform Double-Aspect Combinations Selection

Input: A set of aspect identifiers $T = \{1, 2, \ldots, 25\}$
Output: A list R containing 100 selected pairs

$C \leftarrow \{(i, j) \mid i, j \in T, i < j\}$ {Generate all valid pairs}
for all $t \in T$ **do**
  count$[t] \leftarrow 0$; weight$[t] \leftarrow 1.0$
end for
$R \leftarrow \emptyset$
for iteration $= 1$ **to** 100 **do**
  for all $(i, j) \in C$ **do**
    pair_weight$(i, j) \leftarrow$ weight$[i]$ + weight$[j]$
  end for
  total_weight $\leftarrow \sum_{(i,j)\in C}$ pair_weight(i, j)
  for all $(i, j) \in C$ **do**
    probability$(i, j) \leftarrow \frac{\text{pair_weight}(i,j)}{\text{total_weight}}$
  end for
  Sample (i, j) from C according to probability
  Append (i, j) to R
  for $t \in \{i, j\}$ **do**
    count$[t] \leftarrow$ count$[t] + 1$
    weight$[t] \leftarrow \exp(-$count$[t])$
  end for
end for
return R

B Manual evaluation

We randomly sampled 5,000 instances from the unchecked dataset for manual evaluation across two levels and eight dimensions: dialogue-level (Aspect Relevance, Polarity Relevance, Aspect Coherence, Emotion Relevance) and summary-level (Faithfulness, Fluency, Informativeness, Conciseness). Each instance was independently annotated by three evaluators. Inter-annotator agreement, measured by Fleiss' kappa (Fleiss, 1971), ranged from 0.59 to 0.71 across the dimensions (0.67, 0.63, 0.60, 0.64, 0.61, 0.59, 0.61, 0.71). Given the high agreement, we aggregated the scores to assess data quality. The averaged dimension scores were: 0.86 for dialogue-level (Aspect Relevance: 0.83, Polarity Relevance: 0.94, Aspect Coherence: 0.92, Emotion Relevance: 0.78), and 0.96 for summary-level (Faithfulness: 0.92, Fluency: 1.00, Informativeness: 0.93, Conciseness: 1.00). As shown in Figure 2, the **checked dataset** meets our quality standards across both evaluation levels.

	maximum	**average**	**minimum**
turn	7	21.13	63
dialogue	124	461.94	1094
summary	17	50.13	169

Table 8: Statistics of FOCUS.

C Pilot study

To compare the practical effectiveness of different summary styles, we conducted a pilot study. We randomly sampled 40 dialogues, each paired with both a free-style and a formatted summary, and asked three annotators to evaluate them. As shown in Figure 5, the vote counts for the two styles are comparable overall. Formatted summaries performed better when dialogues involved fewer aspects, benefiting from their concise and structured presentation. However, as the number of aspects increased, their rigidity led to redundancy, whereas free-style summaries proved more effective due to their flexibility.

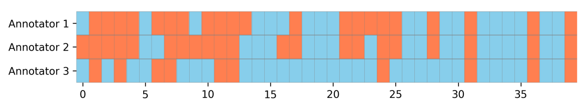

Figure 5: A pilot comparison of free-style versus formatted summaries. Skyblue and coral bars indicate annotator votes for each style. The vertical axis shows vote counts. Samples 0–9 contain one aspect, 10–19 two, and so forth, up to four aspects in 30–39.

D Dataset statistics

The proportion distributions of our FOCUS dataset over single, double, triple and quadruple aspects are 38.36%(4978), 54.74%(7085), 5.79%(741) and 1.11%(144), respectively, which is much close to the scheduled proportion specified in Diversity Aspects of Section generation. The maximum, minimum, and average number of dialogue turns, as well as the character counts in both the dialogues and summaries, are presented in Table 8.

The numbers of occurrence for each coarse-grain aspect appeared in FOCUS are 4218, 4378, 4439, 4511 and 4375 for *Quality*, *Delivery*, *After-sale*, *Discounts* and *System*, respectively, which show the balance of aspect distribution. The word cloud diagrams for dialogue and summary are illustrated in Figure 6, from which, we can see that the comparable word cloud patterns in both dialogue and summary suggest that the summary adequately captures the dialogue's essential content.

E CoT construction

We use the explainable label of samples in FOCUS to construct CoT for $COTS^2$. To exemplify this, an example is shown in Figure 7, in which CoT mainly contains three steps: aspect recognition, opinion content location, sentiment extraction. Three label elements involving three steps are highlighted in CoT prompt template for $COTS^2$ with bold and underline in braces.

F Model training

We conducted three types of training settings. For PLM training, we applied full fine-tuning on a single A100 GPU with batch size 32, learning rate 2×10^{-5}, warmup ratio 0.1, and 10 epochs. For LLM training, we employed LoRA for parameter-efficient fine-tuning using the same GPU and learning rate, but with batch size 16 and 5 epochs. For $COTS^2$ training, we performed instruction tuning initialized with DeepSeek-R1-Distill-Qwen-14B, also using LoRA for efficiency, with batch size 16, learning rate 2×10^{-5}, warmup ratio 0.1, 5 epochs, and a maximum sequence length of 2048. The instruction prompt guided the model to generate a sentiment summary from customer service dialogues, incorporating a CoT process comprising aspect recognition, opinion content localization, and sentiment extraction, with detailed CoT construction provided in the Appendix E.

G Evaluation criteria for ACR and SAA

To provide a comprehensive assessment, we evaluate these metrics through two complementary approaches:

G.1 Automated Evaluation

To create objective, automated versions of ACR and SAA, we define them mathematically based on the ground-truth labels and the generated summary.

Let $G = \{(a_i, s_i)_{\text{gold}}\}$ be the set of ground-truth aspect-sentiment pairs from the dialogue's label. Let $P = \{(a_j, s_j)_{\text{pred}}\}$ be the set of aspect-sentiment pairs extracted from the model-generated summary. Let $A_{\text{gold}} = \{a_i | (a_i, s_i)_{\text{gold}} \in G\}$ be the set of unique aspects in the ground truth. Let $A_{\text{pred}} = \{a_j | (a_j, s_j)_{\text{pred}} \in P\}$ be the set of unique aspects in the prediction.

Rule-based extraction of aspect–sentiment pairs We extract P deterministically with a rule-based procedure (no LLM judging / LLM-as-a-judge / model-based scoring). Aspects are detected by dictionary matching over the fixed inventory of 25 fine-grained aspects (Table 1), and each detected aspect is assigned a polarity using nearby sentiment trigger words.

Normalization and segmentation. Given a model-generated summary, we first normalize it by (i) Unicode NFKC (full-width/half-width conversion), (ii) lowercasing Latin letters, and (iii) collapsing repeated whitespace and punctuation into single separators. We then split the normalized text into segments using line breaks, bullet/number markers, and sentence-final punctuation; the same purely rule-based segmentation is applied to both formatted (line/bullet-like) and free-style (sentence-like) summaries.

Aspect extraction (dictionary matching). Let $\mathcal{A}$ be the fixed set of 25 canonical aspect names (Table 1). For each $a \in \mathcal{A}$, we define a small fixed alias set $\text{alias}(a)$ containing the canonical name and common surface forms (2–4 examples: *packaging*$\mapsto$`packaging`, `package`, `outer package`, `pkg`; *receiving method*$\mapsto$`receiving method`, `delivery option`, `shipping method`; *refund*$\mapsto$`refund`, `money back`, `returned payment`; *price protection*$\mapsto$`price protection`, `price-match`, `price guarantee`). We scan each segment for any alias: ASCII aliases are matched with case-insensitive word-boundary patterns (to avoid partial matches), while non-ASCII aliases (e.g., Chinese surface forms used in the implementation) are matched

(a) dialogue word cloud (b) summary word cloud

Figure 6: Word cloud diagram for dialogue and summary.

Models	Fai.	Flu.	Inf.	Con.	ACR	SAA	Fai.	Flu.	Inf.	Con.	ACR	SAA
T5	3.80	4.38	3.51	3.83	3.43	3.67	3.90	3.92	3.66	3.11	3.87	3.79
BART	3.87	4.41	3.56	3.88	3.62	3.79	3.89	4.00	3.69	3.21	3.93	3.85
Llama	3.97	4.43	3.63	3.90	3.70	4.00	4.00	4.23	3.97	3.27	4.53	4.02
DeepSeek	4.05	4.57	3.77	4.13	3.97	4.27	4.03	4.30	3.90	3.37	4.33	3.90
COTS[2]	4.10	4.49	3.81	3.97	4.01	4.13	4.09	4.20	3.98	3.23	4.48	4.03

Table 9: ChatGLM evaluation results for free-style and formatted strategies

by exact substring. Repeated mentions are deduplicated so each aspect contributes at most once to P.

Sentiment polarity (trigger words, no LLM). For each detected aspect mention in a segment, we search for sentiment triggers within a fixed window in the same segment (within ±30 characters around the matched alias span). Positive triggers include, e.g., good, great, `satisfied`, happy, `efficient`, `helpful`, `resolved`; negative triggers include, e.g., `bad`, `poor`, `unsatisfied`, `angry`, `complain`, `delay`, `damaged`, `wrong`, `missing`. If both polarities are triggered in-window, we select the polarity whose nearest trigger occurrence (by character distance to the aspect span) is closer; ties deterministically default to *negative*. If no trigger is found in the window, we drop the aspect (i.e., we do not add any $(a, s)^{pred}$ to P for that aspect).

Output used by ACR/SAA. The resulting P is the set of *unique* predicted (a, s) pairs used in the ACR/SAA equations in Appendix G.1 This rule-based extraction is used only for objective automated scores (ACRobj/SAAobj) as a fast large-scale metric, complementary to human and ChatGLM-based evaluations.

Aspect Coverage Rate (ACR): The automated ACR can be defined as the recall of the ground-truth aspects in the generated summary. This directly measures the model's ability to identify and retain all the key topics discussed by the customer.

$$ACR = \frac{|A_{\text{gold}} \cap A_{\text{pred}}|}{|A_{\text{gold}}|}$$

Where $|A_{\text{gold}} \cap A_{\text{pred}}|$ is the number of unique aspects correctly identified by the model, and $|A_{\text{gold}}|$ is the total number of unique aspects in the ground truth.

Sentiment Accuracy per Aspect (SAA): The automated SAA evaluates the sentiment polarity accuracy exclusively for the aspects that were correctly identified by the model. This isolates the model's ability to correctly interpret sentiment from its ability to identify aspects.

$$SAA = \frac{\sum_{(a,s)_{\text{pred}} \in P} \mathbb{I}[(a, s)_{\text{pred}} \in G]}{|A_{\text{gold}} \cap A_{\text{pred}}|}$$

Where the numerator sums the number of correctly predicted aspect-sentiment pairs. $\mathbb{I}[\cdot]$ is the indicator function, which is 1 if the condition is true and 0 otherwise. The denominator is the number of correctly identified aspects. If no aspects are correctly identified (i.e., the denominator is 0), SAA is defined as 0.

These automated metrics allow for rapid, large-scale model comparison and address the need for more efficient evaluation protocols.

Models	R-1	R-2	R-L	BS
T5	41.80/53.62/58.86/52.54	17.69/32.11/36.51/29.49	33.24/45.03/51.27/40.45	74.16/80.09/83.57/81.36
BART	49.82/51.24/54.11/53.05	24.31/26.64/28.25/26.82	42.54/43.89/42.88/40.87	77.84/79.31/79.96/77.96
Llama	46.02/52.27/54.80/ 49.32	21.41/27.88/30.20/26.14	38.87/45.26/47.80/40.56	76.31/80.95/82.36/79.50
DeepSeek	43.76/52.37/55.02/53.26	16.56/28.3/27.66 /26.85	36.09/44.11/46.38/46.61	75.13/80.54/81.55/80.59
COTS^2	44.56/56.14/52.43/57.96	18.41/29.57/28.74/34.96	36.88/45.94/46.39/49.04	76.42/81.10/80.54/80.97

Table 10: The objective metric results for various aspect number in free-style strategy. Each block has 4 values, representing single, double, triple and quatra aspect appeared in each dialogue from left to right.

Score	**Aspect Coverage Rate (ACR)**	**Sentiment Accuracy per Aspect (SAA)**
5	All key aspects in the dialogue are accurately covered in the summary	All aspect-level sentiments are correctly preserved and naturally expressed
4	Most aspects are covered, with minor omissions	Most sentiments are correctly preserved with minor mismatches
3	Around half of the key aspects are covered	Sentiments are partially preserved, with some errors or ambiguities
2	Only a few aspects are mentioned	Most sentiments are missing or incorrect
1	No key aspect is preserved	Sentiment expression is entirely wrong or missing

Table 11: Annotation criteria for ACR and SAA scores on a 5-point Likert scale.

G.2 Human Evaluation

We engaged ten volunteers for human evaluation: five were undergraduate students majoring in computer science, and the other five were first-year graduate students also majoring in computer science. Ultimately, we conducted human evaluations on 60 summaries outcomes produced by predictions from different models. On average, each dialogue-summary pair received assessments from 3 annotators.

To ensure consistency in the evaluation process, we trained these 10 students on six criteria and required them to conduct assessments using a back-to-back approach, meaning they scored independently without knowledge of each other's ratings. We used the kappa value to adjust and estimate the consistency of their evaluation results; If the kappa score did not exceed 0.55, we reconvened the evaluators for discussions aimed at reaching a consensus on the criteria. ACR and SAA are metrics first proposed in this paper. Therefore, we specifically drafted a scoring standard table(as elaborated in Table 11) and held pre-assessment discussions with the evaluators based on this table to reach a consensus on scoring standards.

This manual process captures the nuances of language that automated metrics might miss and serves as the ultimate benchmark for model performance.

H Experimental result

In this subsection, We provide the evaluation results of ChatGLM in Table 9, which show that, regardless of formatted-based or free-style strategy, our COTS^2 demonstrate advantages in FAI., ACR, and SAA. Meanwhile, all models exhibit insufficient performance in terms of informativeness, further indicating that this task is highly challenging. Additionally, compared to free-style strategy, all models experience a noticeable decline in Con. when dealing with formatted summaries, as they are longer and contain more information.

We present results of different models under free-style summaries with varying numbers of aspects in Table 10. Results show that most models perform best on triple-aspect data. We argue that when the number of aspects is small, the model may identify additional aspects, leading to a decrease in performance. Prompt templates for CoT generation and ChatGLM evaluation are shown in Table 19.

I Evaluation of LLM-generated Dialogues and Real-World Interaction Patterns

To further address concerns regarding the use of LLM-generated data, we conducted a comprehensive evaluation of real-world customer service dialogues and compared their interaction patterns to those in our synthetic FOCUS dataset. This analysis demonstrates that our generated data closely aligns with real-world dialogues in terms of structural and sentiment characteristics, thereby mitigating potential biases introduced by automatic generation.

I.1 Data Collection

We randomly sampled 3000 dialogues from the original CSDS corpus that were not used in FOCUS construction. These real-world dialogues encompass diverse customer inquiries across the five coarse-grained aspects (Quality, Delivery, After-sale, Discounts, System) and exhibit natural variations in turn length and sentiment expression.

I.2 Methodology

Our evaluation focused on five key dimensions:

1. **Turn Distribution**: We compared the dialogue turns in real-world versus LLM-generated dialogues, using Kolmogorov–Smirnov tests to assess distributional similarity.

2. **Aspect and Sentiment Coverage**: We calculated the frequency of each fine-grained aspect and associated sentiment polarity in both sets, measuring correlation via Pearson's r.

3. **Sentiment Dynamics**: We analyzed the co-occurrence of sentiment shifts (e.g., negative-to-positive within a dialogue) to verify that the generated data preserved realistic emotional transitions.

4. **Stylometric n-gram Divergence (char 4-gram JSD)**: We measured Jensen–Shannon divergence (JSD) between per-dialogue character 4-gram distributions and a bootstrapped real baseline, reporting median JSD and 95% CI, complemented by a permutation test. The protocol segmented text at the character level, computed TF distributions per dialogue, and averaged via bootstrap over real samples. Acceptance required median JSD < 0.02 and 95% CI < 0.03, with qualitative salience analysis as a complement.

5. **Style-Artifact Detector (linear classifier AUC)**: We trained a lightweight classifier to test whether systematic style artifacts could separate synthetic from real data. Features included TF–IDF of char n-grams (3–5), function-word and punctuation rates, and turn-length statistics. Models used logistic regression with 5×2 cross-validation; AUC was evaluated with DeLong 95% CI. An ablation kept only non-lexical stylometry to attribute separation. Acceptance required AUC CI including 0.5 with the upper bound < 0.60. Protocol standardized features, stratified folds by dialogue length, and reported macro-AUC.

I.3 Results

Turn Distribution The Kolmogorov–Smirnov statistic ($D = 0.024$, $p > 0.05$) indicates no significant difference in their distributions, suggesting comparable conversation lengths. Detailed turn-count frequencies are provided in Table 12.

Turn Count	Real-World	LLM-Generated
12	90	78
13	192	174
14	288	300
15	330	324
16	432	384
17	390	420
18	324	336
19	270	252
20	228	240
21	162	180
22	108	120
23	72	60
24	48	60
25	30	36
26–30	36	36

Table 12: Dialogue turn distribution for real-world and LLM-generated samples (3000 dialogues each).

Aspect and Sentiment Coverage Across 25 fine-grained aspects, the Pearson correlation between real and synthetic frequency counts is $r = 0.92$ ($p < 0.001$), and for sentiment polarity distributions $r = 0.89$ ($p < 0.001$), confirming high alignment in topic and sentiment prevalence.

Sentiment Dynamics We counted the number of dialogues exhibiting at least one within-dialogue sentiment polarity shift. Out of 3000 samples per set, 552 real-world dialogues and 534 synthetic dialogues contained shifts, corresponding to shift rates of 18.4% and 17.8%, respectively. The full contingency is given in Table 13.

Character Distribution Character 4-gram distributions show very small divergence between *FOCUS* and real dialogues. The median JSD is 0.013 (95% CI [0.012, 0.016]); a permutation test finds no practically meaningful shift ($p = 0.21$). These

Outcome	Human-Authored	LLM-Generated
With Shift	552	534
Without Shift	2448	2466

Table 13: Counts of dialogues containing at least one sentiment polarity shift (3000 dialogues each).

values are comfortably within our acceptance region (median < 0.02, CI < 0.03). The full result is given in Table 14.

Set	Median JSD	Mean JSD	95% CI (Median)	Permutation p
Real vs. Real (bootstrap baseline)	0.012	0.013	[0.011, 0.015]	–
FOCUS vs. Real	**0.013**	0.014	**[0.012, 0.016]**	0.21

Table 14: Stylometric divergence : JSD over character 4-grams. We report median, mean, and a bootstrap 95% CI for the median across dialogues.

Stylometric Discrimination A lightweight discriminator trained on stylometric features performs at *near-chance* levels. With full features (char 3–5-gram TF–IDF + function/punctuation rates + turn-length stats), mean AUC is 0.54 (DeLong 95% CI [0.50, 0.58]). Using only non-lexical stylometry (function/punctuation/turn stats) yields 0.51 [0.48, 0.55]. Char n-grams alone reach 0.56 [0.52, 0.59]. All CIs include 0.5 and remain < 0.60 at the upper bound, satisfying our acceptance criterion and indicating no learnable, systematic style fingerprint beyond weak lexical cues. The full result is given in Table 15.

I.4 Discussion

These findings corroborate that our LLM-based generation scheme reliably imitates real-world dialogue structures and sentiment patterns. By validating against real-world data, we alleviate concerns about synthetic artifacts and reinforce the credibility of the FOCUS dataset for downstream research and practical deployment.

J Training on the Full Checked Set

To test whether conclusions drawn from the refined benchmark generalize to a larger, noisier training pool, we additionally train models on the full **checked** set and evaluate on the same refined test set.

K Domain Adaptation and Cross-Dataset Transfer

K.1 Domain-adaptive training on JDDC

We further test whether domain-adaptive training on a larger in-domain corpus (JDDC) improves performance before applying the COSDS prompting/tuning.

Feature set	AUC (mean)	95% CI
Full stylometric + char n-grams	**0.54**	**[0.50, 0.58]**
Non-lexical stylometry only	0.51	[0.48, 0.55]
Char n-grams only	0.56	[0.52, 0.59]

Table 15: Style-artifact detector (M5): Linear classifier AUC with DeLong 95% CI and feature ablations.

Strategy	Model	ROUGE-L↑	BERTScore↑	ACR_{obj} ↑	SAA_{obj} ↑	Faithfulness↑
Formatted	DeepSeek	57.92	83.32	78.52	82.32	4.72
	$COTS^2$	57.02	83.19	**81.26**	**85.94**	**4.83**
Free-style	DeepSeek	43.65	80.21	66.12	70.14	4.81
	$COTS^2$	**44.72**	**80.59**	**68.35**	**73.34**	**4.84**

Table 16: Generalization when trained on the full checked set. Models are trained on checked and evaluated on the refined test set. $COTS^2$ remains robust and consistently improves task-aligned metrics (ACR/SAA) and faithfulness.

K.2 Cross-dataset evaluation on CSDS

Although CSDS does not contain the same fine-grained aspect/emotion supervision as FOCUS, we evaluate transferability by directly testing a model trained on FOCUS on CSDS.

L Detail of case study

To demonstrate the effectiveness of $COTS^2$, we present a sample case generated by different models in Figure 9, where three aspects are covered, i.e., *returns*, *reshipment* and *invoice*. First, T5 successfully identifies all three aspects, but its summary contained factual inconsistencies regarding *returns*. Specifically, the summary mentions that the customer expressed gratitude for the explanation provided by agent, whereas in the original dialogue, the customer only indicated satisfaction with the service instead of gratitude. A similar issue exists in DeepSeek, where two aspects are identified, and the summary mentions that the customer expressed frustration about *invoice* issues, while the original dialogue convey that the customer was puzzled about the invoice. For BART, the summary only covers two aspects, resulting in aspects missing. Although Llama successfully identifies all aspects, the summary related to *invoice* issue lacks detailed information. In Conclusion, it is evident that existing models are prone to factual inconsistencies, aspect missing, and mismatches between aspects and sentiments when generating summaries, while our $COTS^2$ can effectively mitigate all the issues mentioned above. Besides, $COTS^2$ outlined the reasoning process, with the three steps in the chain

Setting	Model	Training Data	ROUGE-L↑	BERTScore↑	ACR_{obj} ↑	SAA_{obj} ↑
Formatted	DeepSeek	FOCUS only	58.50	83.67	78.90	83.90
	COTS^2	FOCUS only	57.13	83.31	81.80	86.20
	COTS^2 + JDDC	JDDC + FOCUS	**59.62**	**84.32**	**82.14**	**87.32**
Free-style	DeepSeek	FOCUS only	42.69	79.31	65.00	69.80
	COTS^2	FOCUS only	43.73	79.59	66.90	71.20
	COTS^2 + JDDC	JDDC + FOCUS	**45.13**	**81.29**	**68.48**	**72.94**

Table 17: Domain adaptation on JDDC before COSDS training.

Training → Test	Model	R-1↑	R-2↑	R-L↑	BERTScore↑	Faithfulness↑
FOCUS → CSDS	COTS^2	**65.12**	**45.21**	**61.32**	**88.73**	**4.93**
FOCUS → FOCUS	COTS^2	63.92	42.77	57.13	83.31	4.90

Table 18: Cross-dataset evaluation of COTS^2 on CSDS without additional fine-tuning.

of thought (aspect identification, sentence localization, and sentiment recognition) ensuring the controllability of summaries generation process.

From Figure 8, sentences with the same color represent duplication for the same content. We see that in these two cases, there indeed exists a significant amount of information redundancy in the outputs generated by T5. Through manual random sampling of 20 samples, it is found that 11 of the predicted outcomes contain instances of information repetition. Additionally, 4 of the samples exhibited severe redundancy, meaning that more than two core pieces of information were repeated. To demonstrate the effectiveness of COTS^2, we present a sample case generated by different models in Figure 9, where three aspects are covered, i.e., *returns*, *reshipment* and *invoice*. The red-shaded content indicates issues with the generated summaries. From the case shown in this figure, it can be seen that T5 has inconsistencies with the facts; BART predictions lack an *invoice* aspect; The summary produced by Llama suffers from insufficient information; DeepSeek's prediction results exhibit mismatches between sentiment and aspect, and missing key aspects.

Task	Prompt Template
CoT generation	Based on the following dialogue and summary in the field of e-commerce, generate a chain of thought that demonstrates how the given dialogue is distilled into the given summary. The chain of thought should be presented in paragraph form and explain how the key information in the dialogue is extracted and integrated into the summary. Dialogue: ***{dialogue}*** Summary: ***{summary}*** Requirements: Analyze the key information points in the dialogue. Explain how these information points were selected and integrated into the summary. Present the reasoning process from the dialogue to the summary.
ChatGLM evaluation	You will be provided with a dialogue and its paired summary. Your task is to evaluate the quality of the summary based on the dialogue. Please rate each summary on six dimensions: Faithfulness: whether the summary is correct based on the dialogue; Fluency: whether the summary is fluent and grammatically correct; Informativeness: whether the summary includes all key information; Conciseness: whether the summary is very concise (not redundant); ACR: whether the summary covers the aspects present in the dialogue; SAA: whether the aspects and sentiments expressed in the dialogue are correctly reflected in the summary. Return the output in the following JSON format: {Faithfulness: value, "Fluency": value, "Informativeness": value, "Conciseness": value, "ACR": value, "SAA": value}. You should rate on a scale from 1 (worst) to 5 (best). Do not provide detailed explanations. Dialogue: ***{dialogue}*** Summary: ***{summary}***

Table 19: Prompting templates for CoT generation and ChatGLM evaluation

COT prompt template for COTS[2]

Well, now I need to generate a summary of a customer service dialogue from the customer's perspective. First, I need to carefully read through the entire dialogue content, especially focusing on what the customer said. To understand the key points of communication between the customer and the agent. Since many aspects elements are involved in the dialogue, I need to identify the aspects that appear in the current dialogue. Because this summary will be subjective, I also need to recognize the customer's emotions regarding these aspects. The aspect appearing in this dialogue include: ***{aspects}*** . Regarding the aspect of ***{aspect}*** , the customer mentioned: '***{content}***'. At the same time, for the aspect related to ***{aspect}***, the customer's emotion was ***{sentiment}***; From the customer's point of view, the summary should highlight their main problems and feelings. What the customer says should be the focus of the summary information. Meanwhile, I should also pay attention to what the agent has said, using it as supplementary information for the summary. Finally, it's important to ensure the summary is concise and clear, allowing readers to quickly understand the customer's position and decisions.

0 : Hello
1 : Hello, how can I assist you today?
2 : **The outer packaging of the product I purchased from your store recently is really very poor.**
3 : I sincerely apologize for the unpleasant experience. Could you please specify what aspects you found unsatisfactory?
4 : **It's just a cheap plastic bag that's already torn.**
5 : We deeply regret this. We will immediately address this issue and strengthen our quality checks on packaging.
6 : I hope improvements will be made in the future. **By the way, your delivery service this time was quite good.**
7 : Thank you for your recognition. We strive to provide high-quality delivery services at all times.
8 : **Yes, the delivery person was also very friendly. They delivered the package right to my doorstep, which was extremely convenient.**
9 : I'm glad to hear that feedback. We will definitely continue to improve.
10 : However, the packaging issue still needs to be resolved.
11 : Absolutely, we will work on it and make improvements as soon as possible. Is there anything else we can assist you with?
12 : That's all for now. Thank you for your response.
13 : You're very welcome! Please feel free to contact us anytime if you have any further questions.
14 : Okay, goodbye.
15 : wish you a pleasant life!

Explainable labels

aspect	sentiment	content
packaging	-1	**The outer packaging of the product I purchased from your store recently is really very poor.(2)**
		It's just a cheap plastic bag that's already torn.(4)
receiving method	+1	**By the way, your delivery service this time was quite good.(6)**
		Yes, the delivery person was also very friendly. They delivered the package right to my doorstep, which was extremely convenient.(8)

COT for current dialogue sample

Well, now I need to generate a summary of a customer service dialogue from the customer's perspective. First, I need to carefully read through the entire dialogue content, especially focusing on what the customer said. To understand the key points of communication between the customer and the agent. Since many aspects elements are involved in the dialogue, I need to identify the aspects that appear in the current dialogue. Because this summary will be subjective, I also need to recognize the customer's emotions regarding these aspects. The aspect appearing in this dialogue include: **packaging** and **receiving method**. Regarding the aspect of ***packaging*** , the customer mentioned: '**The outer packaging of the product I purchased from your store recently is really very poor.**'; '**It's just a cheap plastic bag that's already torn.**' For the aspect of **receiving method**, the customer said: '**By the way, your delivery service this time was quite good.**'; '**Yes, the delivery person was also very friendly. They delivered the package right to my doorstep, which was extremely convenient.**'. At the same time, for the aspect related to **packaging**, the customer's emotion was unsatisfaction; for the aspect of **receiving method**, the emotion was satisfaction. From the customer's point of view, the summary should highlight their main problems and feelings. What the customer says should be the focus of the summary information. Meanwhile, I should also pay attention to what the agent has said, using it as supplementary information for the summary. Finally, it's important to ensure the summary is concise and clear, allowing readers to quickly understand the customer's position and decisions.

Figure 7: CoT prompt for COTS[2]

Dialogue 1

Customer: Are you there?
Agent: Hello, how may I assist you?
Customer: How long is the shelf life of your products? Are the production dates fresh?
Agent: Please hold on a moment while I check that for you.
Customer: Alright.
Agent: Our warehouse updates its inventory every day, so we cannot confirm the exact production date of the product you will receive. But rest assured, we strictly control the shelf life of all our products and offer comprehensive after-sales service.
Customer: Can you at least tell me how long the shelf life is?
Agent: I'm very sorry, but for this product, we haven't clearly marked the shelf life; specific information needs to refer to the expiry date shown on the physical item.
Customer: What does that mean? Is there not even a basic indication of the shelf life?
...
Agent: ...., Wishing you a pleasant life.

summary generated by T5 and reference for Dialogue 1

T5: **The customer expressed concerns about the product's shelf life**, especially regarding the labeling of the expiration date. The customer hopes to receive **clear information about the shelf life**, **clear information about the shelf life**, and **clear information about the shelf life**. **The customer expressed concerns about the product's shelf life**, especially the labeling of the shelf life. The customer hopes to receive **clear information about the shelf life** and solution.

reference: The customer expressed concerns about the product's shelf life, especially regarding the lack of clear expiration date labeling. They hope to receive more explicit shelf life information and assurance that the received product will not be close to expiration.

Dialogue 2

Customer: Hello, I would like to inquire about receiving goods.
Agent: Hello, this is Appliance Assistant [number], happy to serve you! How may I assist you?
Customer: I plan to buy a refrigerator, but I might not be convenient to receive the goods for now. Do you support delayed receipt of goods?
Agent: Dear customer, could you please provide your order number?
Customer: I haven't placed the order yet.
Agent: Dear customer, it can indeed be postponed. We support pre-ordering for delayed delivery even before placing an order.
Customer: Then how long can it be postponed?
Agent: It can be postponed for [number] days, dear customer.
Customer: I see. If I choose cash on delivery, will that be convenient? I have some concerns about the method of direct door-to-door delivery by the courier and then making payment.
...
Customer: Not for now, thank you.

summary generated by T5 and reference for Dialogue 2

T5: **The customer expressed dissatisfaction with the delivery method**, especially regarding cash-on-delivery home delivery by the courier. **The customer hopes to choose a more reliable way to receive the goods**; **The customer expressed dissatisfaction with the delivery method**, especially regarding the option of scheduling a delivery delay before placing an order. The customer hopes to choose a more reliable delivery method; **The customer expressed dissatisfaction with the delivery method**, especially regarding the option of reserving a delayed delivery before placing an order. **The customer hopes for a more secure method of receiving the goods.**

reference: The customer has voiced concerns about the delivery method, particularly in terms of the delayed receipt period and the safety of the cash-on-delivery payment option. They are hoping for a more secure and trustworthy method of receiving their order..

Figure 8: Two cases: summaries generated by T5. Sentences with the same color represent duplication for the same content.

Dialogue

Customer: Hello, I recently purchased a product that has some issues, and I'd like to return it. Could you tell me the process?
Agent: Hello! We completely understand your situation. You can apply for a return directly on the order page, and we will process it as soon as possible.
Customer: Okay, do I need to prepare any materials?
Agent: You just need to provide photos of the product and explain the reason for the return.
Customer:That's convenient. **Your return policy is really efficient, and I'm quite satisfied.**
Agent: We're glad to hear that you're satisfied with our service. Is there anything else I can assist you with?
Customer: Actually, there was another item that was out of stock, and you mentioned you would send a replacement. Do you have any updates on that?
Agent: Let me check… Ah, I see the record. The reshipment has already been arranged and is expected to arrive within three days.
Customer: Great! **I'm also very satisfied with the speed of this reshipment. Thank you.**
Agent: You're welcome—it's our pleasure. Also, do you have any questions regarding invoices?
Customer: Actually, yes. I want to ask about issuing a special VAT invoice. I'm not sure how to proceed with that.
Agent: You can select the option to issue a special VAT invoice when placing your order and upload the required documents. Once the finance team approves them, the invoice will be issued.
Customer: **I didn't receive an electronic invoice for my previous order. What happened**?
Agent: Please wait a moment while I check… Sorry, the system shows that the invoice hasn't been issued yet.
Customer: **This is really stressing me out. I've waited for a long time. I hope you can resolve this issue as soon as possible**.
Agent: We sincerely apologize for the inconvenience caused. We will handle this immediately and issue the invoice as soon as possible.
Customer: Okay, I hope you can resolve it quickly.
...
Agent: Thank you for your understanding and cooperation. Wishing you a pleasant day ahead!

summary generated by various models

T5: The customer inquired about the **return** process, and the agent provided a detailed explanation of the steps involved. The customer expressed satisfaction with the return process and thanked the agent for their clear guidance. Additionally, the customer asked about the status of a **reshipment** for an out-of-stock item. After checking, the agent informed the customer that the reshipment had been arranged and would arrive within three days. The customer was pleased with the quick response and expressed gratitude for the timely update. Finally, the customer raised an issue regarding an unreceived electronic **invoice**, particularly noting that the system showed it had not been issued. The customer hoped the issue would be resolved as soon as possible.(Inconsistency with the facts)

BART: The customer reported an issue with a product and inquired about the **return** process. The agent guided the customer to apply for a return on the order page and informed them that they needed to provide photos of the product along with an explanation of the return reason. The customer found the return policy convenient and efficient and expressed satisfaction with the agent's guidance. Additionally, the customer inquired about the progress of a **reshipment** for an out-of-stock item. Upon learning that the reshipment had been arranged and was expected to arrive within three days, the customer expressed satisfaction, noting the quick handling of the reshipment, and thanked the agent. (lack of invoice aspect)

Llama: The customer raised a question about the **return** process, and the agent instructed them to apply for a return on the order page, providing photos of the product and stating the reason for the return. The customer was satisfied with the simple and efficient return process and appreciated the quick response from the agent. The customer also inquired about the status of a reshipment for an out-of-stock item. After confirming that the **reshipment** had been arranged and would arrive within three days, the customer expressed satisfaction with the swift handling of the reshipment and thanked the agent for their prompt assistance. Lastly, the customer expressed confusion regarding **invoices**, specifically about not receiving an electronic invoice, and hoped the issue with issuing the invoice could be resolved as soon as possible.(info deficiency)

COT prompt for COTS²

Well, now I need to generate a summary of a customer service dialogue from the customer's perspective. First, I need to carefully read through the entire dialogue content, especially focusing on what the customer said. To understand the key points of communication between the customer and the agent. Since many aspects elements are involved in the dialogue, I need to identify the aspects that appear in the current dialogue. Because this summary will be subjective, I also need to recognize the customer's emotions regarding these aspects. The aspect appearing in this dialogue include: ***returns***, ***reshipment*** , and ***invoice***. Regarding the aspect of *returns*, the customer mentioned: '**Your return policy is really efficient, I'm quite satisfied.**' For the aspect of reshipmen out-of-stock items, the customer said: 'Great! **I'm also very satisfied with the speed of this reshipment, thank you.**' Concerning the invoice issue, the customer expressed: '**I didn't received an electronic invoice for my previous order. What happened?**' and '**This is quite frustrating, I've waited for a long time, I hope you can resolve this issue as soon as possible.**' At the same time, for the aspect of ***returns***, the customer's emotion was satisfaction; for the aspect of ***reshipment***, the emotion was also satisfaction; but for ***invoice*** issue, the emotion was dissatisfaction. From the customer's point of view, the summary should highlight their main problems and feelings. What the customer says should be the focus of the summary information. Meanwhile, I should also pay attention to what the agent said, using it as supplementary information for the summary. Finally, it's important to ensure the summary is concise and clear, allowing readers to quickly understand the customer's position and decisions.

summary generated by DeepSeek & COTS²

DeepSeek: The customer raised a question about applying for a **return**, and the agent guided them to initiate the return process on the order page. The customer was satisfied with how convenient and efficient the return process was and expressed gratitude for the timely assistance from the representative. Additionally, the customer expressed frustrated about not receiving an electronic invoice, particularly regarding the process for issuing a special VAT **invoice**. The customer hoped the issue with generating the invoice could be resolved as soon as possible. (emotion mismatch aspect) &(lack of reshipment)

COTS2: The customer asked about how to initiate a **return**, and the agent guided them to apply for a return on the order page. **The customer was satisfied with the convenient and efficient return process and expressed appreciation for the prompt handling by the agent**. Additionally, the customer inquired about a reshipment for an out-of-stock item. The agent confirmed that the **reshipment** had been arranged and would arrive within three days. **The customer was pleased with the quick response and expressed satisfaction with the handling of the reshipment**. Finally, the customer raised the issue of not receiving an electronic invoice. The agent assured the customer that it would be processed immediately and the **invoice** issued as soon as possible. While **the customer expressed some concern about the progress of resolving the invoice issue**, they also conveyed hope that the matter would be resolved quickly based on theagent's commitment.

Figure 9: Case analysis and CoT prompt for COTS^2