Chain-of-Interactions: Multi-step Iterative In-Context Learning for Abstractive Task-Oriented Dialogue Summarization of Conversational AI Interactions

Jason Lucas[1], **John Chen**[2], **Ali Al-Lawati**[1], **Mahjabin Nahar**[1], **Mahnoosh Mehrabani**[2]

[1] The Pennsylvania State University, PA, USA
{jsl5710,aha112,mfn5333}@psu.edu
[2] Interactions LLC, NJ, USA
{jchen,mahnoosh}@interactions.com

Abstract

Large Language Models (LLMs) have introduced paradigm-shifting approaches in natural language processing. Yet, their transformative in-context learning (ICL) capabilities remain underutilized, especially in customer service dialogue summarization—a domain plagued by generative hallucinations, detail omission, and inconsistencies. We present Chain-of-Interactions (CoI), a novel single-instance, multi-step framework that orchestrates information extraction, self-correction, and evaluation through sequential interactive generation chains. By strategically leveraging LLMs' ICL capabilities through precisely engineered prompts, CoI dramatically enhances abstractive task-oriented dialogue summarization (ATODS) quality and usefulness. Our comprehensive evaluation on real-world and benchmark human-agent interaction datasets demonstrates CoI's effectiveness through rigorous testing across 11 models and 7 prompting approaches, with 9 standard automatic evaluation metrics, 3 LLM-based evaluations, and human studies involving 480 evaluators across 9 quality dimensions. Results reveal CoI's decisive superiority, outperforming all single-step approaches and achieving 6× better entity preservation, 49% higher quality scores, and 322% improvement in accuracy compared to state-of-the-art multi-step Chain-of-Density (CoD). This research addresses critical gaps in task-oriented dialogue summarization for customer service applications and establishes new standards for harnessing LLMs' reasoning capabilities in practical, industry-relevant contexts[1].

1 Introduction

Large Language Models (LLMs) have transformed natural language processing through in-context learning (ICL) capabilities (Jain et al., 2023; Tang et al., 2023; Al-Lawati et al., 2025), driving shifts from single-step to multi-step iterative pipelines that address hallucinations, detail omission, and inconsistencies in abstractive summarization (Adams et al., 2023; Zhang et al., 2023b). However, abstractive task-oriented dialogue summarization (ATODS) remains underexplored—particularly for customer service applications requiring accuracy, completeness, and efficiency for post-call analytics (Feng et al., 2021).

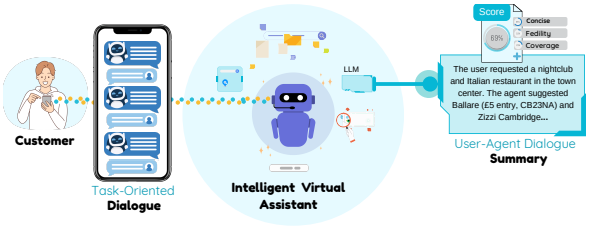


Figure 1: User-Agent task-oriented dialogue summarization

Customer service dialogues summaries require: (1) preserving critical entities (confirmation numbers, timestamps, contacts); (2) tracking action-outcome relationships across turns; and (3) maintaining factual consistency for business-critical information (Allahyari et al., 2017; Feng et al., 2021). Multi-instance methods requiring multiple LLMs, external feedback, or prompt chaining (Zhang et al., 2023b; Madaan et al., 2023; Tian et al., 2024a; Wu et al., 2022; Sun et al., 2024) create computational bottlenecks. Single-step approaches (Wang et al., 2023; Liu et al., 2019; Liu and Lapata, 2019) sacrifice information density, while Chain-of-Density (CoD) (Adams et al., 2023) loses critical entities during compression—retaining only 0.61 entity preservation ratio versus human baselines.

We present Chain-of-Interactions (CoI), a novel single-instance framework addressing these limitations through eight sequential refinement chains. Unlike existing approaches that trade off between competing objectives, CoI achieves simultaneous optimization across entity preservation, compression efficiency, and factual consistency (Adams

[1]Dataset, code, and materials are available: https://github.com/jsl5710/CoI

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 3560–3599
November 4-9, 2025

et al., 2023; Zhang et al., 2023b). Through precisely engineered prompts, CoI transforms a single LLM into a powerful pipeline that progressively refines outputs via extraction, correction, and self-evaluation stages.

Our evaluation on benchmark (TodSum) and real-world (CRM) datasets demonstrates CoI's effectiveness. Testing 11 LLMs (7B to >100B parameters) against 7 prompting approaches, we generated and evaluated 80,000+ summaries using 9 automatic metrics, 3 LLM-based methods, and human assessment with 480 participants (expert, academic, public) across 9 quality dimensions.

Our contributions are: (1) a single-instance framework replacing multi-model-call overhead with one call with a prompt having multiple chains; (2) empirical validation showing CoI outperforms all single-step baselines and CoD (6× entity preservation, 49% higher quality, 322% accuracy improvement); (3) human evaluation (480 participants) establishing superiority over human and machine baselines; (4) a validated nine-dimensional evaluation framework for ATODS; and (5) public release of 80K+ summaries, annotations, and implementation materials[2].

2 Related Work

Automatic Text Summarization Jin et al. (2024) defines ATS as "a series of automatic actions to distill extensive textual content into concise summaries, capturing the essence while retaining key information." ATS creates concise summaries while preserving essential content through extractive (verbatim selection) or abstractive (novel text generation) approaches (Feng et al., 2021; Jin et al., 2024). Evolution spans from early frequency-based methods (Luhn, 1958; Edmundson, 1969) and machine learning approaches (Mihalcea and Tarau, 2004) to neural architectures (See et al., 2017) and transformer models like GPT-4 (Zhong et al., 2022; Jin et al., 2024; Adams et al., 2023). Pre-trained models like BART advanced through fine-tuning (Khandelwal et al., 2019), while recent LLMs use prompt-based and in-context learning without extensive training (Jin et al., 2024), though most focus on document summarization with computationally intensive models unsuitable for real-world deployment.

[2] https://github.com/jsl5710/CoI

LLM ICL Summarization LLMs transformed summarization through ICL and Chain-of-Thought reasoning (Wei et al., 2022). Key advances include SumCoT (Wang et al., 2023), Chain-of-Density (Adams et al., 2023), multi-agent refinement (Zhang et al., 2023b), tri-agent approaches (Xiao et al., 2023), iterative ICL prompting (Al Lawati et al., 2025), and extract-then-generate methods (Zhang et al., 2023a). However, customer service dialogue summarization remains under-explored, with existing methods requiring multiple models and passes, making them impractical for commercial deployment due to computational costs. Tang et al. (2022) identified eight distinct error types in dialogue summarization.

Customer Service Summarization Customer service dialogue summarization is understudied despite its business importance (Feng et al., 2021). Prior work explored topic modeling (Liu et al., 2019; Zou et al., 2021b,a) and dataset curation (Zhao et al., 2021; Zhang et al., 2021; Lin et al., 2021). Recent approaches examine role-oriented methods (Tian et al., 2024a,b, 2023), but gaps remain in leveraging small LLMs for efficient, practical deployment.

Dialogue Summary Evaluation Evaluation approaches include automatic metrics (ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2019)) and LLM-based methods (Likert scales (Liu et al., 2023), pairwise comparison (Jain et al., 2023), faithfulness assessments (Luo et al., 2023)). Challenges include lack of standardized metrics and domain inconsistencies (Allahyari et al., 2017). While human judgment remains the gold standard, LLM-based evaluation shows promise for automated assessment (Jin et al., 2024).

3 Method

We propose Chain-of-Interactions (CoI) for generating progressively refined dialogue summaries through structured multi-step reasoning chains. CoI implements a sequential eight-chain process to produce summaries that balance accuracy, relevance, and utility for customer support agents while addressing information loss, repetition, hallucinations, and inconsistencies. We validate CoI through extensive evaluation encompassing LLMs, prompting methods, standard automatic metrics, LLM-based evaluations, and human assessment, ensuring robust evidence across diverse architectures and

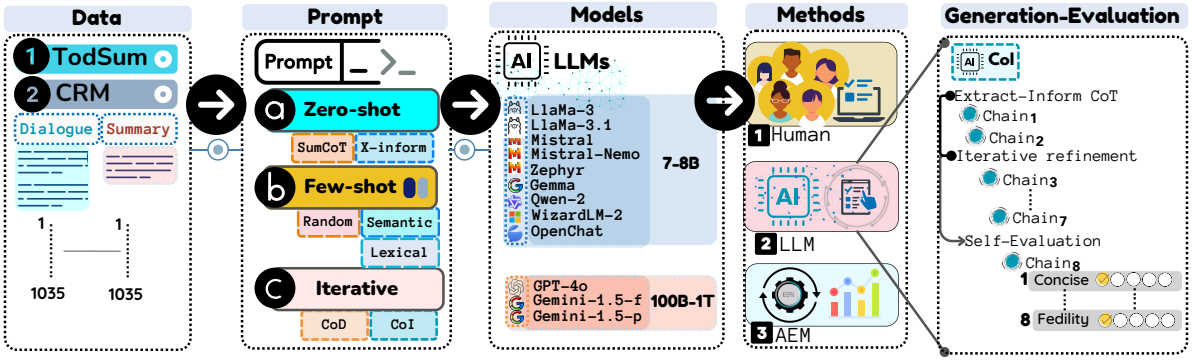


Figure 2: Overview of our **CoI** Task-Oriented Dialogue Summarization methodology, showing: (1) datasets, (2) prompt, (3) LLMs evaluate, (4) evaluation methods, and (5) generation and evaluation process with **CoI** framework.

evaluation paradigms.

3.1 Problem Definition

Let G be a generative LLM, x be the original task-oriented dialogue input, and O be the output containing: S (seven refined summary chains), A (Likert-scale evaluation of final summary), and E (explanation justifying scores).

Definition 1 (Single-Instance Transformation)**.** *Our approach is defined as function $PE : X \to X'$ where:*

1. *$x' = PE(x) = x \oplus I$ ($\oplus$ denotes concatenation, I is instruction text)*
2. *$O = G(x')$ where $G : X' \to O$ and $O = (S, A, E)$*
3. *Generation involves iterative processing within single instance G*
4. *Quality improves without additional training, agents, or model requests*
5. *$|I|$ is small relative to model capacity*

3.2 Chain-of-Interaction Framework

Traditional LLMs use single-pass inference producing immediate outputs, unlike human cognitive processes involving reflection and refinement (Pan et al., 2025). Language reason models ("Thinking models") like OpenAI-o1 (Jaech et al., 2024) demonstrate how structured deliberation enhances performance.

CoI implements guided internal reasoning in non-thinking models through: (*1*) Context-aware extraction and synthesis; (*2*) Sequential refinement with self-correction; and (*3*) Quantitative self-evaluation with rationales.

Definition 2 (CoI SumEval)**.** *Given dialogue x, generate output $O = (S, A, E)$ through 8-chain process $C = \{C_1, ..., C_8\}$:*

C_1: Extract task-oriented interactions

C_2: Generate initial summary

C_3: Add missing entities

C_4: Review using 9 quality aspects

C_5: Remove redundancy

C_6: Correct hallucinations

C_7: Enhance brevity (Final summary S)

C_8: Generate evaluation A, E

Where $A = \{a_1, ..., a_9\}$ represents Likert evaluations for nine dimensions (conciseness, coverage, relevance, rephrasing, coherence, fidelity, readability, fluency, redundancy) and $E = \{e_1, ..., e_9\}$ provides corresponding explanations.

Prompt Template Our JSON-formatted template consists of two components: T_1 (demonstrator with instructions and examples for the 8-chain process, see Figures 18 and 19) and T_2 (summarizer-evaluator with structured JSON output, see Figure 20). This architecture enables systematic refinement and evaluation within a single model instance. Table 21 illustrates the complete 8-chain process with a generated example.

4 Datasets

Customer service summaries emphasize entities and factual details (e.g., reservation numbers, contact information) that differ from general conversation summaries. After reviewing open-source dialogue benchmarks (SAMSum (Gliwa et al., 2019), AMI (McCowan et al., 2005), ICSI (Janin et al., 2003), DialogSum (Chen et al., 2021), CSDS (Lin et al., 2021)), only TodSum (Zhao et al., 2021) suited our TODS focus. We evaluate on: (**1**) `TodSum`, an open-source benchmark with structured goal-directed conversations, and (**2**) `CRM`, a propri-

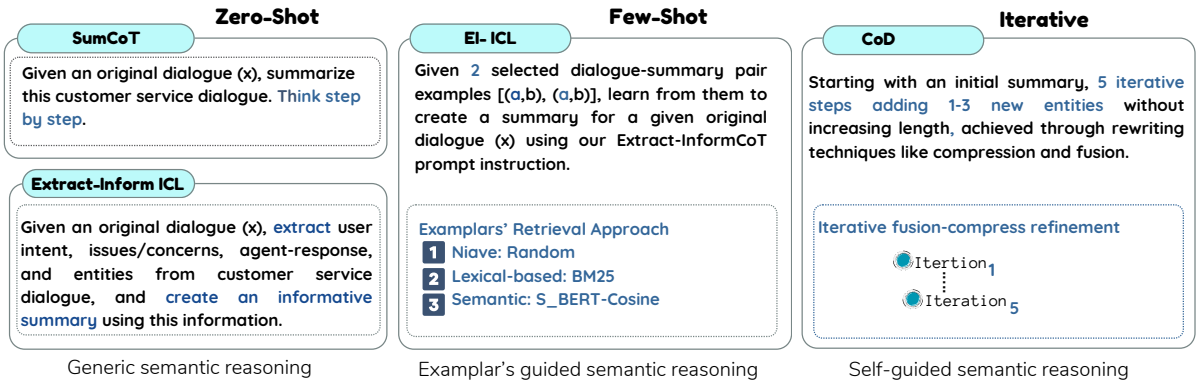


Figure 3: Overview of baseline **ICL** Techniques. (1) single-step: Zero-shot and Few-shot (2) Multi-step: Iterative

etary real-world customer service collection (Table 1). Both dataset consist of customer support dialogues and summary pairs

Characteristic	TodSum		CRM	
	TOD	HS	TOD	HS
Dataset Splits				
Train	7,892	7,892	–	–
Test	999	999	35	35
Validation	999	999	–	–
Total	*9,890*	*9,890*	*35*	*35*
Word Count				
Max	721	133	2,138	107
Avg	187	45	787	24
Min	28	15	289	6
Sentence Count				
Max	44	8	114	6
Avg	14	4	48	2
Min	2	2	19	1

Table 1: Dataset statistics for TodSum (open-source) and CRM (proprietary). TOD: Task-Oriented Dialogue; HS: Human Summary.

5 Models

Our study spans 3 large LLMs and 9 small instruction-tuned models (7-8B parameters), balancing performance with computational efficiency for industry deployment (Table 13). These instruction-tuned models provide essential TODS capabilities: prompt adherence, complex reasoning, multi-turn instruction following, pattern recognition, and contextual adaptability—aligning with our **CoI** framework. Unlike prior work using legacy models (BART, UniLM (Zhu et al., 2021)) or large decoder-based LLMs (GPT-4 (Adams et al., 2023)), we prioritize advanced small and large LLMs for practical implementation assessment.

6 Baseline

We evaluate CoI against single-step and multi-step summarization paradigms. Single-step approaches generate summaries in one operation (Liu and Lapata, 2019; Wang et al., 2023; Lewis et al., 2020). We compare **CoI** against (*a*) two zero-shot and (*b*) one few-shot single-step ICL approaches using automatic evaluation metrics. Multi-step methods use multiple operations, such as Summit (Zhang et al., 2023b) (multiple LLMs with external feedback) and CoD (Adams et al., 2023) (single LLM, single-pass). Most significantly, we compare against the SOTA CoD iterative method, which represents the most appropriate baseline as it shares our single-instance, multi-step framework while using iterative refinement (Figure 3). For methodology details, see Appendix F.

Model	Total	Min	Max	Mean	Std
GPT-4o	7,110	20	154	63.99	20.72
Gemini-1.5-f	6,143	18	220	64.83	21.65
Gemini-1.5-p	6,990	7	346	59.19	22.29
Gemma	5,931	7	406	56.40	25.30
Llama-3	6,044	8	164	61.04	19.92
Llama-3.1	6,181	7	228	62.26	23.89
Mistral	7,119	7	230	68.09	30.35
Mistral-N	7,161	7	191	45.73	16.36
Openchat	7,224	7	140	48.63	18.20
Qwen-2	7,223	8	265	55.41	26.75
WizardLM-2	6,063	7	359	73.48	36.20
Zephyr	6,871	7	345	44.77	19.91

Table 2: Table shows characteristics of synthetic summaries generated from test set dialogues using different LLMs, measured by token count. Red: large LLMs (>100B parameters), Blue: small LLMs (7-8B parameters).

6.1 Generation

We generate task-oriented dialogue summaries using the test set (1,034 samples) from our datasets (Table 1), with the training set (7,917 samples) sup-

porting few-shot sampling. Using 12 LLMs and 7 prompt techniques, we generated 86K+ summaries. After removing 6K+ defective samples due to generation inconsistencies and controllability issues (Figure 15), we obtained 80K+ summaries for evaluation (Table 2). GPT-4o achieved perfect (100%) prompt-following performance, while small LLMs like OpenChat-8B and Mistral achieved competitive (99%) completion rates. Table 21 shows sample CoI outputs. Our dataset is publicly available: GitHub.

7 Experiment

Our experiment framework spans 11 models, 7 prompting approaches, 9 automatic metrics, 3 LLM-based evaluations, and 480 human evaluators across 9 quality dimensions, providing comprehensive validation of CoI's effectiveness.

7.1 Experiment Setup

Quality Evaluation Dimensions. LLMs often produce summaries with verbosity, omissions, repetition, and factual errors (Zhang et al., 2023b; Adams et al., 2023; Kryściński et al., 2020; Lucas et al., 2023). Since existing LLM-based evaluation dimensions do not transfer well to ATODS (e.g., G-Eval (Liu et al., 2023)), we developed nine dimensions based on prior frameworks (Gehrmann et al., 2018; Peyrard, 2019; Dang, 2005; Zhu and Bhat, 2020), ATS literature (Braggaar et al., 2023; Lin et al., 2021; Likert, 1932), and industry expertise, organized into three main sections:

Section 1. Completeness and Accuracy Assesses information coverage, focus, and accuracy:

- ***Coverage*** *- Inclusion of all relevant information.*
- ***Relevance*** *- Inclusion of only the pertinent information.*
- ***Fidelity*** *- Preservation of original meaning, context, facts, and intent.*

Section 2. Brevity and Rephrasing Evaluates summary succinctness, precision, and uniqueness:

- ***Conciseness*** *- Brevity and elimination of unnecessary details.*
- ***Redundancy*** *- Avoidance of repetition.*
- ***Rephrasing*** *- Demonstration of understanding through paraphrasing.*

Section 3. Readability and Flow Measures reading and comprehension ease:

- ***Readability*** *- Ease of comprehension with clear language.*
- ***Fluency*** *- Freedom from grammatical errors.*
- ***Discourse Coherence*** *- Logical, structured presentation at sentence and summary levels.*

Evaluation Approaches. We employ a multi-faceted evaluation strategy combining (*i*) gold-standard human evaluations, (*ii*) SOTA LLM-based automatic evaluation metrics (AEM), and (*iii*) standard AEM (e.g., entity density). Although LLM-based evaluation closely parallels human judgment (Liu et al., 2023; Luo et al., 2023), standard AEM fails to capture task-oriented dialogue requirements where intent recognition, action fulfillment, and key entities are critical.

7.2 RQ1: Does CoI outperform CoD in quality and information retention?

This experiment establishes whether CoI outperforms the SOTA CoD method through a comprehensive evaluation of the informativeness and quality dimensions using SOTA LLM-AEM.

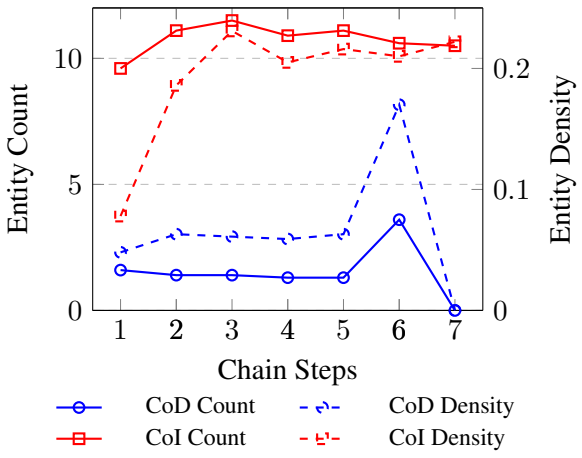


Figure 4: Entity count and density comparison.

Informativeness: Entity Density We compute entity density as the ratio of unique entities (Spacy[3]) to tokens (NLTK (Loper and Bird, 2002)), following Adams et al. (2023). CoI consistently preserves higher entity density throughout summarization chains, with advantages ranging from +0.040 to +0.222 in entity/token density (steps 2-7). CoI maintains 10.5-11.5 entities in later chains versus CoD's rapid loss (dropping to 0 entities by step 7) (Table 9). CoI improves critical entity recall by 67% for high-value numerical entities (phone numbers, confirmation codes, reservation details) and embeds 6× more entities than CoD in chain 1 (Figure 4, detail analysis in Appendix C). We provide a detailed analysis showing CoI superiority in summary compression, entity preservation, and factual consistency across other prompt baselines and LLMs. Appendix D.

Quality Assessment: LLM-AEM We implemented comprehensive LLM-based evaluation across nine dimensions using advanced reasoning LLMs (Gemini-2.0-pro, OpenAI o1) (Fu et al., 2024; Liu et al., 2023; Adams et al., 2023; Gilardi et al., 2023). CoI significantly outperforms CoD across all dimensions (Figure 5), with dramatic improvements in Completeness and Accuracy (Q1-Q3): CoD averaged 1.05 while CoI achieved 4.43 (322% improvement). In Brevity and Rephrasing (Q4-Q6), CoI maintains advantages (4.94 vs 4.24), particularly in Rephrasing (Q6: 4.81 vs 3.35). For Readability and Flow (Q7-Q9), CoI consistently scores high (averaging 4.89), exceeding CoD in Coherence (Q9) by 1.21 points. Overall, CoI achieves 4.75 mean score versus CoD's 3.19, representing 49% improvement in summary quality.

7.3 RQ2: Which CoI chain is optimal and do chains meet their objectives?

This experiment evaluates whether CoI chains fulfill their objectives and identifies the optimal chain for customer support through human evaluation.

Study Design. We assessed CoI's iterative approach with 30 Prolific participants evaluating 100 randomly selected dialogues (30 CRM, 70 TodSum from 12 LLM generators). This yielded 2,710 evaluations (7 chains × 100 dialogues × 3 annotators) assessing (*1*) objective fulfillment and (*2*) preferred summary for customer support agents. Details in section L.4 and Table 20.

[3]https://spacy.io

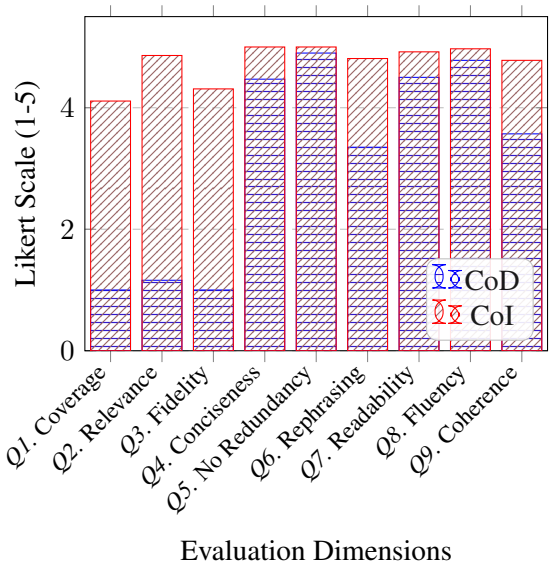


Figure 5: Quality comparison: CoD vs CoI across nine dimensions following SOTA G-Eval and LLM as a judge paradigm.

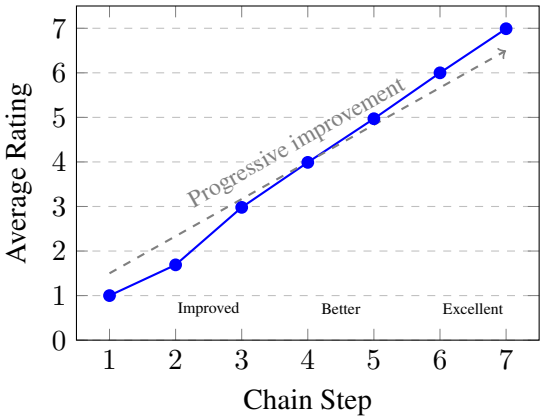


Figure 6: Objective fulfillment ratings showing progressive improvement from Chain 1 (1.00) to Chain 7 (6.99/7).

Objective Fulfillment. Human evaluations revealed consistent progression across chains (Figure 6): Chain 1 (1.00), Chain 3 (2.98), Chain 4 (3.99), Chain 5 (4.97), Chain 6 (6.00), and Chain 7 (6.99/7). Fleiss' Kappa analysis (FLEISS, 1971) showed perfect agreement for Chains 1 and 6 ($\kappa = 1.000$), almost perfect agreement for Chains 3, 4, 5, and 7 ($\kappa > 0.93$), and overall reliability of $\kappa = 0.905$ (Table 14).

Optimal Chain. Analysis of 300 human judgments identified Chain 7 as the overwhelming preference (62.33%), followed distantly by Chain 2 (17.00%). Chains 3-5 received minimal preference (5-7% each), Chain 6 was rarely selected (2.33%), and Chain 1 was never preferred (Figure 7).

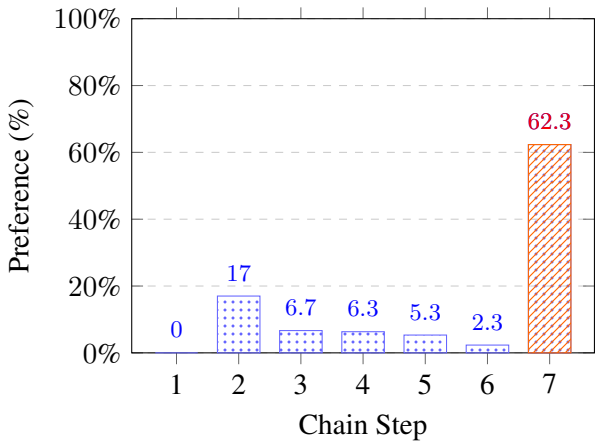


Figure 7: Preference distribution across chain steps; Chain 7 is strongly preferred (62.3%) by customer support agents.

RQ2 Answer: Yes, CoI chains progressively meet their objectives (ratings improve from 1.00 to 6.99/7) with high inter-rater reliability ($\kappa = 0.905$). Chain 7 is optimal for customer support, preferred in 62.3% of evaluations.

7.4 RQ3: How does CoI perform across critical quality dimensions?

We conducted comprehensive human evaluations of summary quality across our nine dimensions, comparing CoI summaries against human-written gold-standard and SOTA CoD baselines. This experiment forms the core of our research, providing definitive evidence of CoI's practical effectiveness through rigorous human assessment. Evaluations were performed by three distinct groups: Prolific crowdsource participants (public), academic students, and Interactions LLC expert annotators, establishing a robust gold-standard evaluation framework that balances diverse perspectives and expertise levels.

Study Design. We conducted systematic human evaluations with 450 participants comprising three balanced groups: 150 industry experts from customer service domains, 150 graduate students with NLP background, and 150 Prolific crowdsourced users representing general public perspective (see Figure 17). Each participant completed a carefully designed 12-minute survey assessing Human, CoI, and CoD dialogue-summary pairs using standardized 5-point Likert scales (strongly disagree to strongly agree) across our nine established quality dimensions. To ensure ethical standards and data quality, participants received fair compensation ($12/hour for Prolific users, $10/hour for students and experts), and the IRB-approved study (STUDY00025599) incorporated multiple attention checks and validation mechanisms for quality control. This tripartite design strategically balances perspectives from domain experts, academic researchers, and general users, providing comprehensive validation across different stakeholder groups. Detailed implementation procedures and evaluation instruments are provided in Appendix H.

Quality Assessment Results. Our systematic analysis follows the nine quality dimensions organized into three theoretically-grounded categories as established in subsection 7.1. As demonstrated in Figure 8, CoI consistently outperformed both human-written gold standards and SOTA CoD baselines across all participant groups, with particularly strong performance in accuracy and readability dimensions.

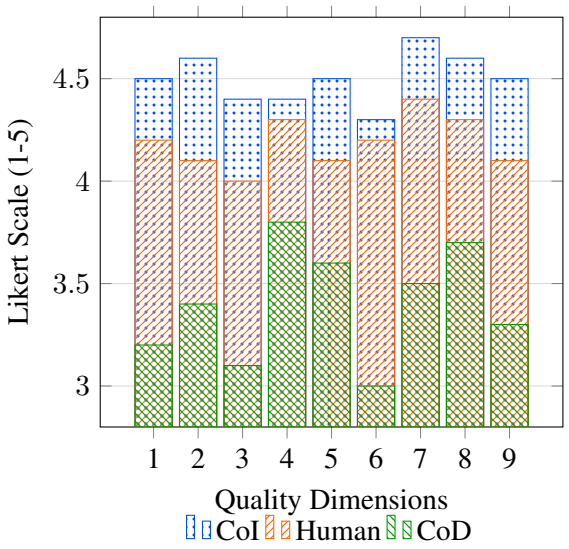


Figure 8: Quality dimension performance comparison of CoI, human, and CoD summary with overlapping bars.

1. Completeness and Accuracy Assessment: CoI demonstrated statistically significant superior performance in *Coverage* (Q1), achieving the most comprehensive information capture while maintaining conciseness, and excelling in *Fidelity* (Q3), maintaining the most faithful representation of original dialogue intent and factual accuracy. For *Relevance* (Q2), CoI performed comparably to human summaries in focusing on pertinent information, indicating effective filtering of essential content while excluding peripheral details. These results validate CoI's ability to preserve critical information integrity throughout the iterative refinement process.

2. Brevity and Rephrasing Evaluation: All approaches demonstrated similar *Conciseness* (Q4) metrics with minor but consistent advantages for

CoI, indicating effective length optimization without information loss. For *Redundancy* (Q5) avoidance, CoI and human summaries achieved comparable performance in minimizing information repetition, both substantially outperforming CoD baseline. CoI maintained a notable edge in *Rephrasing* (Q6) capabilities, demonstrating superior ability to reformulate content through paraphrasing while preserving semantic meaning, which indicates effective abstractive summarization beyond simple extraction.

3. Readability and Flow Analysis: CoI generated consistently superior linguistic quality across all readability metrics, producing the most easily comprehensible summaries (*Readability*, Q7) through clear language structure and logical information organization. The framework demonstrated minimal language errors comparable to human summaries (*Fluency*, Q8), indicating robust grammatical and syntactic generation capabilities. Most significantly, CoI achieved the highest scores in logical consistency (*Discourse Coherence*, Q9), ensuring smooth information flow and coherent narrative structure throughout summaries. Comprehensive cross-validation analysis and detailed statistical significance testing are provided in Appendix K.

RQ3 Answer: CoI consistently outperforms both human gold-standard and CoD baseline across all nine critical quality dimensions. CoI achieves superior performance in information completeness and accuracy, maintains competitive brevity while excelling in content rephrasing, and demonstrates the highest scores in readability and linguistic flow. These results, validated across 450 human evaluators from diverse backgrounds, establish CoI's practical effectiveness for real-world customer service applications.

7.5 RQ4 (a): How do LLMs' self-evaluation capabilities compare with external LLM and human assessments?

Building upon Liu et al. (2023)'s `G-Eval` framework, which demonstrated that GPT-4 with 5-point Likert scales achieves highest correlation with human judgment, we integrate our nine quality dimensions through two approaches: (1) incorporating dimensions into CoI's 8th chain for self-evaluation (**CoI Self-Eval**), and (2) developing external evaluation prompts targeting summaries from other LLMs (**LLM External-Eval**). Our setup uses GPT-4o for self-evaluation, with o1 and Gemini-1.5-pro providing external assessments.

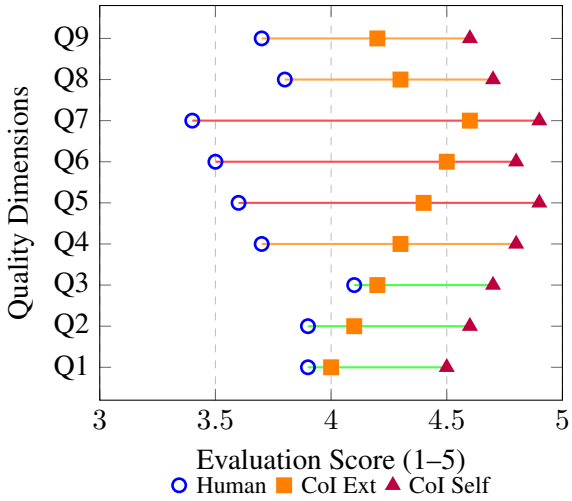


Figure 9: Dumbbell chart with distinct symbols and gradient-colored rails indicating consensus level: green (high consensus), orange (medium consensus), red (low consensus). Human Evaluation (○), CoI External-LLM (■), and CoI Self-LLM Evaluation (▲).

Evaluation Pattern Analysis. Figure 9 reveals distinct evaluation patterns: self-evaluation consistently overestimates performance across almost all dimensions, while external LLM evaluation aligns closely with human judgments on critical semantic dimensions (accuracy/fidelity, coverage/completeness, relevance/focus). Both LLM approaches show strong agreement on surface-level features (brevity, readability) where criteria are more objective.

RQ4(a) Answer: External LLM evaluation provides more reliable assessment than self-evaluation, closely aligning with human judgments on complex semantic tasks (accuracy, completeness, relevance), while self-evaluation exhibits systematic optimism bias across all dimensions.

7.6 RQ4 (b): How do different summarizers and ICL prompt-based methods perform across standard evaluation metrics?

While standard automatic evaluation metrics (AEM) are not specifically designed for our nine quality dimensions of interest, we selected the most appropriate ones to assess performance across summarizers (models) and ICL prompt methods. We employed eight established metrics: BERTScore (semantic similarity), Flesch Reading Ease (readability), Parascore (paraphrase quality), AlignScore (factual consistency), Meteor (content overlap), Compression Ratio (conciseness), ROUGE F1 and Precision (lexical overlap), and Grammar (linguistic correctness).

Model	BERT	Flesch	Para	Align	Meteor	Comp	R-F1	R-Prec	Gram	μ
GPT-4o	**0.871**^{1}	**0.630**^{1}	**0.619**^{1}	**0.931**^{1}	**0.364**^{1}	**0.281**^{1}	**0.167**^{1}	**0.168**^{1}	**0.993**^{1}	**0.558**
Gemini-1.5-p	0.870^{2}	0.629^{2}	0.604^{2}	0.906^{2}	0.357^{2}	0.279^{2}	0.163^{2}	0.163^{2}	0.981^{2}	0.550
Mistral-N	0.869^{3}	0.622^{3}	0.601^{3}	0.895^{3}	0.325^{3}	0.263^{3}	0.162^{3}	0.160^{3}	0.976^{3}	0.541
Llama-3.1	0.867^{4}	0.617^{4}	0.577^{5}	0.890^{4}	0.320^{4}	0.242^{4}	0.160^{4}	0.158^{4}	0.979^{4}	0.534
Llama-3	0.867^{4}	0.606^{5}	0.576^{6}	0.887^{5}	0.311^{5}	0.246^{5}	0.159^{5}	0.156^{5}	0.977^{5}	0.532
Mistral	0.867^{4}	0.604^{6}	0.573^{7}	0.869^{6}	0.310^{6}	0.235^{6}	0.156^{6}	0.154^{6}	0.977^{5}	0.527
Qwen-2	0.866^{7}	0.569^{7}	0.573^{7}	0.797^{7}	0.295^{7}	0.224^{7}	0.153^{7}	0.152^{7}	0.977^{5}	0.512
Openchat	0.865^{8}	0.557^{8}	0.565^{8}	0.795^{8}	0.286^{8}	0.220^{8}	0.151^{8}	0.149^{8}	0.976^{8}	0.507
WizardLM-2	0.865^{8}	0.552^{9}	0.560^{9}	0.722^{9}	0.272^{9}	0.218^{9}	0.147^{9}	0.145^{9}	0.974^{9}	0.495
Zephyr	0.863^{10}	0.537^{10}	0.549^{10}	0.679^{10}	0.265^{10}	0.203^{10}	0.144^{10}	0.142^{10}	0.972^{10}	0.484
Gemma	0.860^{11}	0.537^{10}	0.490^{11}	0.177^{11}	0.253^{11}	0.197^{11}	0.139^{11}	0.147^{11}	0.971^{11}	0.430

Table 3: Model performance rankings across nine evaluation metrics using the CoI framework. Superscripts indicate rank (1=best). Abbreviated metrics: BERT=BERTScore, Para=Parascore, Align=AlignScore, Comp=Compression, R-F1=ROUGE F1, R-Prec=ROUGE Precision, Gram=Grammar, μ=Arithmetic Mean. GPT-4o achieves the highest mean score. Best performing small LLM (Mistral-N) highlighted in blue.

ICL Approach	BERT	Flesch	Para	Align	Meteor	Comp	R-F1	R-Prec	Gram	Mean
Single-Step Approaches										
Few-SBERT-SS	0.873^{2}	0.629^{3}	0.599^{3}	0.855^{4}	*0.374*^{1}	0.315^{4}	0.183^{2}	0.164^{2}	0.982^{2}	*0.541*
Few-Random	0.873^{2}	0.632^{2}	0.600^{2}	0.883^{2}	0.369^{2}	0.318^{3}	0.179^{3}	0.159^{3}	0.982^{2}	*0.544*
Few-BM25	0.872^{4}	0.624^{4}	0.597^{4}	0.858^{3}	0.363^{3}	0.317^{5}	0.174^{4}	0.156^{4}	*0.983*^{1}	*0.549*
Zero-extract-inform	0.867^{5}	0.602^{5}	*0.601*^{1}	*0.885*^{1}	0.344^{4}	0.341^{6}	0.154^{5}	0.135^{5}	*0.983*^{1}	*0.546*
Zero-Vanilla	0.866^{6}	0.594^{6}	0.596^{5}	0.827^{5}	0.324^{5}	0.326^{7}	0.144^{6}	0.133^{6}	*0.983*^{1}	*0.532*
Multi-Step Approaches										
CoI	**0.897**^{1}	**0.691**^{1}	**0.873**^{1}	**0.898**^{1}	**0.408**^{1}	0.225^{2}	**0.190**^{1}	**0.180**^{1}	0.977^{4}	**0.612**
CoD	0.833^{7}	0.457^{7}	0.388^{6}	0.603^{6}	0.102^{6}	**0.172**^{1}	0.068^{7}	0.090^{7}	**0.989**^{1}	0.411

Table 4: ICL single and multi-step approach performance rankings across nine evaluation metrics. Superscripts indicate rank (1=best). Abbreviated metrics: BERT=BERTScore, Para=Parascore, Align=AlignScore, Comp=Compression, R-F1=ROUGE F1, R-Prec=ROUGE Precision, Gram=Grammar, Mean=Simple Arithmetic Mean. CoI achieves the highest mean score (0.612), ranking first in 7 out of 9 metrics.

Model Performance. Across all 11 models tested (Table 3), GPT-4o consistently achieves top performance (rank 1) across all metrics, followed by Gemini-1.5-p (rank 2) and Mistral-N (rank 3). Notably, Gemma shows poor AlignScore performance (0.177), while smaller models like Openchat and Mistral demonstrate competitive performance, suggesting CoI's effectiveness across different model scales.

RQ4 Answer: CoI demonstrates robust performance across standard AEM, ranking first in 7 out of 9 metrics when compared to prompt methods and showing consistent effectiveness across all model scales. GPT-4o achieves optimal performance with CoI, while the framework maintains strong results even with smaller, more efficient models.

ICL Method Performance. CoI significantly outperforms all baseline methods (Table 4), ranking first in 7 out of 9 metrics (BERTScore: 0.897, Flesch: 0.691, Parascore: 0.873, AlignScore: 0.898, Meteor: 0.408, Compression: 0.395, ROUGE Precision: 0.189). Single-step methods show competitive performance in specific areas (Few-SBERT-SS excels in ROUGE F1: 0.183), while CoD substantially underperforms across most metrics, confirming CoI's superiority in multi-step approaches.

8 Conclusion

We introduced Chain of Interactions (CoI), a novel single-instance abstractive summarization technique that progressively refines task oriented customer service dialogues summaries through structured multi-step reasoning chains. Our evaluation demonstrates that CoI significantly outperforms state-of-the-art approaches, excelling in information retention (6× more critical entities than CoD) and quality dimensions (49% higher overall quality scores). Human evaluations confirm Chain 7 as optimal for customer support applications. This work advances dialogue summarization for practical customer service implementations while establishing a robust evaluation framework balancing nine critical dimensions.

9 Limitations

Despite CoI's effectiveness, several limitations merit discussion: (1) The framework currently requires manual prompt tuning for each chain stage, which could be automated; (2) While CoI shows strong performance with smaller LLMs (7-8B parameters), optimal results still favor larger models, limiting accessibility; (3) Our evaluation focused exclusively on English language dialogues, leaving cross-lingual capabilities unexplored; (4) The framework's effectiveness remains untested beyond customer service domains; (5) Self-evaluation in Chain 8 shows limitations in assessing complex aspects like factual accuracy; (6) Our evaluation did not include a formal language proficiency measure. However, we reasonably ensured adequate English language competency through our recruitment criteria and institutional affiliations (see Appendix J.1).

Future work will explore additional dialogue domains, applications in agentic AI and scientific reasoning, and integration with smaller models for real-time support systems. We also plan to develop more comprehensive customer service datasets through industry collaborations and enhance evaluation capabilities with automatic parameter optimization.

Ethics Statement

This research prioritizes ethical considerations in several ways: (1) All data collection and human evaluation procedures received proper IRB approval (STUDY00025599) with fair compensation for participants; (2) While we used proprietary customer service data, all examples were anonymized to protect privacy; (3) The framework is designed to preserve factual accuracy and avoid potential biases in summarization; (4) We explicitly evaluate model outputs for hallucination and factual consistency; (5) Our efficiency analysis aims to make summarization technology more accessible through smaller models. However, we acknowledge potential risks of automated summarization in customer service contexts and recommend human oversight for critical applications.

Acknowledgements

I extend special thanks to the Penn State LinDiv NSF NRT program and Interactions LLC. I am grateful to my LinDiv mentor and program director Dr. Van Hell, along with LinDiv support team members Sue Tighe and Heather Mann. I also acknowledge all Interactions LLC team members who supported me during my summer 2024 internship, including Dr. Srinivas Bangalore, Luis Marciano, Karl Gorski and his Caller Effort Scoring team. I express sincere appreciation to my internship supervisors Dr. Chen and Dr. Mehrabani, who served as co-authors on this paper. This work was supported by NSF DGE NRT 2125865.

References

Griffin Adams, Alexander R Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: Gpt-4 summarization with chain of density prompting. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, volume 2023, page 68.

Ali Al Lawati, Jason Lucas, and Prasenjit Mitra. 2025. Semantic captioning: Benchmark dataset and graph-aware few-shot in-context learning for sql2text. In Proceedings of the 31st International Conference on Computational Linguistics, pages 8026–8042.

Ali Al-Lawati, Jason Lucas, Zhiwei Zhang, Prasenjit Mitra, and Suhang Wang. 2025. Graph-based molecular in-context learning grounded on morgan fingerprints. arXiv preprint arXiv:2502.05414.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, et al. 2017. Text summarization techniques: A brief survey. International Journal of Advanced Computer Science and Applications, 8(10).

Anouck Braggaar, Christine Liebrecht, Emiel van Miltenburg, and Emiel Krahmer. 2023. Evaluating task-oriented dialogue systems: A systematic review of measures, constructs and their operationalisations. arXiv preprint arXiv:2312.13871.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 5062–5074.

Hoa Trang Dang. 2005. Overview of duc 2005. In In Proceedings of the document understanding conference, pages 1–12. Citeseer, NIST.

Harold P Edmundson. 1969. New methods in automatic extracting. Journal of the ACM (JACM), 16(2):264–285.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 9:391–409.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers. arXiv preprint arXiv:2107.03175.

J FLEISS. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5):378–382.

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6556–6576.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4098–4109.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30):e2305016120.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. EMNLP-IJCNLP 2019, page 70.

Barbara Grosz, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. Computational linguistics.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. arXiv preprint arXiv:2412.16720.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. In The 61st Annual Meeting Of The Association For Computational Linguistics.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)., volume 1, pages I–I. IEEE.

Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. arXiv preprint arXiv:2403.02901.

Urvashi Khandelwal, Kevin Clark, Dan Jurafsky, and Lukasz Kaiser. 2019. Sample efficient text summarization using a single pre-trained transformer. arXiv preprint arXiv:1905.08836.

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

Manoj Kumar and Rajiv Ratn Shah. 2012. A comparative study of automatically evaluating text coherence. In Proceeding of International Conference on Computer Science & Engineering (ICCSE-2012) Nainital (India), volume 19.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, page 7871. Association for Computational Linguistics.

Rensis Likert. 1932. A technique for the measurement of attitudes. Archives of Psychology.

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

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.

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019. Automatic dialogue summary generation for customer service. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 1957–1965.

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

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3730–3740.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, pages 63–70.

Jason Lucas, Limeng Cui, Thai Le, and Dongwon Lee. 2022. Detecting false claims in low-resource regions: a case study of caribbean islands. In Proceedings of the workshop on combating online hostile posts in regional languages during emergency situations, pages 95–102.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 14279–14305.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. IBM Journal of research and development, 2:159–165.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization. arXiv preprint arXiv:2303.15621.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: iterative refinement with self-feedback. In Proceedings of the 37th International Conference on Neural Information Processing Systems, pages 46534–46594.

I McCowan, J Carletta, W Kraaij, S Ashby, S Bourban, M Flynn, M Guillemot, T Hain, J Kadlec, V Karaiskos, et al. 2005. The ami meeting corpus. In Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research, pages 137–140. Noldus Information Technology.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing, pages 404–411.

Mahjabin Nahar, Haeseung Seo, Eun-Ju Lee, Aiping Xiong, and Dongwon Lee. 2024. Fakes of varying shades: How warning affects human perception and engagement regarding llm hallucinations. In First Conference on Language Modeling (COLM).

Jianwei Niu, Qingjuan Zhao, Lei Wang, Huan Chen, Mohammed Atiquzzaman, and Fei Peng. 2016. Onses: a novel online short text summarization based on bm25 and neural network. In 2016 IEEE Global Communications Conference (GLOBECOM), pages 1–6. IEEE.

Jianfeng Pan, Senyou Deng, and Shaomang Huang. 2025. Coat: Chain-of-associated-thoughts framework for enhancing large language models reasoning. arXiv preprint arXiv:2502.02390.

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

Maxime Peyrard. 2019. A simple theoretical model of importance for summarization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1059–1073.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In Proceedings of the 2008 conference on empirical methods in natural language processing, pages 186–195.

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

Shichao Sun, Ruifeng Yuan, Ziqiang Cao, Wenjie Li, and Pengfei Liu. 2024. Prompt chaining or stepwise prompt? refinement in text summarization. In Findings of the Association for Computational Linguistics ACL 2024, pages 7551–7558.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. Confit: Toward faithful dialogue summarization with linguistically-informed contrastive finetuning. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5657–5668.

Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. 2023. Large language models are in-context semantic reasoners rather than symbolic reasoners. arXiv preprint arXiv:2305.14825.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2023. Chimed-gpt: A chinese medical large language model with full training regime and better alignment to human preferences. arXiv preprint arXiv:2311.06025.

Yuanhe Tian, Fei Xia, and Yan Song. 2024a. 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.

Yuanhe Tian, Fei Xia, and Yan Song. 2024b. Learning multimodal contrast with cross-modal memory and reinforced contrast recognition. In Findings of the Association for Computational Linguistics ACL 2024, pages 6561–6573.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In The 61st Annual Meeting Of The Association For Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837.

Tien-Hsuan Wu, Ben Kao, Felix Chan, Anne SY Cheung, Michael MK Cheung, Guowen Yuan, and Yongxi Chen. 2021. Semantic search and summarization of judgments using topic modeling. In Legal Knowledge and Information Systems, pages 100–106. IOS Press.

Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022. Promptchainer: Chaining large language model prompts through visual programming. In CHI Conference on Human Factors in Computing Systems Extended Abstracts, pages 1–10.

Wen Xiao, Yujia Xie, Giuseppe Carenini, and Pengcheng He. 2023. Chatgpt-steered editing instructor for customization of abstractive summarization. arXiv preprint arXiv:2305.02483, page 103.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023a. Extractive summarization via chatgpt for faithful summary generation. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 3270–3278.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023b. Summit: Iterative text summarization via chatgpt. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 10644–10657.

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

Xinyuan Zhang, Ruiyi Zhang, Manzil Zaheer, and Amr Ahmed. 2021. Unsupervised abstractive dialogue summarization for tete-a-tetes. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 14489–14497.

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

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 11765–11773.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5927–5934. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2020. Gruen for evaluating linguistic quality of generated text. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 94–108.

Yicheng Zou, Jun Lin, Lujun Zhao, Yangyang Kang, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2021a. Unsupervised summarization for chat logs with topic-oriented ranking and context-aware auto-encoders. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 14674–14682.

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

A LLM vs Human Performance Analysis

To establish CoI's effectiveness in absolute terms, we conducted comprehensive analysis comparing LLM-generated summaries against human-written baselines across compression efficiency, factual consistency, and entity preservation. This analysis provides critical context for understanding whether CoI's improvements over other methods translate into meaningful advantages over human performance.

A.0.1 Universal LLM Superiority Across Metrics

Our analysis reveals systematic LLM advantages across all four critical dimensions. For compression efficiency, LLM-generated summaries consistently achieve superior compression ratios compared to human baselines, with improvements ranging from 0.001 to 0.16 across all models and prompt types. CoI demonstrates exceptional compression advantage, achieving ratios of 0.225 compared to human baselines of 0.234, representing optimal summarization conciseness that significantly outperforms human capabilities.

Factual consistency analysis shows pronounced LLM advantages, with alignment scores exceeding human performance across all evaluated models and prompting approaches. The improvements range from 0.015 to 0.203, with CoI-generated summaries achieving alignment scores of 0.898 compared to human scores averaging 0.696. This 29% improvement in factual consistency indicates that

structured prompting approaches preserve information more effectively and maintain higher fidelity to source content than human summarizers.

Entity density analysis reveals substantial LLM superiority in maintaining information richness. LLM-generated summaries achieve entity densities ranging from 11 to 27 entities per 100 words, compared to human summaries averaging 14-15 entities per 100 words. Mistral leads with 27 entities per 100 words, representing a 93% improvement over human performance, while most models achieve 40-60% improvements in entity density.

Entity preservation analysis demonstrates LLM superiority in maintaining critical information elements through targeted prompting. While human summaries preserve entities at a 1.0 baseline ratio, LLM approaches achieve preservation ratios ranging from 1.45 to 1.70, with CoI leading at 1.70. This represents a 70% improvement in entity retention compared to human performance, indicating that automated approaches better preserve specific factual details crucial for customer service applications.

A.0.2 Prompting Approach Impact on Human-LLM Gap

Analysis across different prompting strategies reveals varying magnitudes of LLM advantages over human performance. For compression efficiency, CoI achieves optimal performance with a ratio of 0.225 compared to human 0.234, demonstrating superior compression capability. Zero-shot and few-shot approaches achieve compression ratios of 0.315-0.341, substantially exceeding human performance.

Factual consistency analysis shows CoI achieving the highest alignment scores (0.898) compared to human baselines (0.696), representing 29% improvement. Zero-shot approaches maintain competitive factual consistency at 0.883-0.885, substantially exceeding human performance. This demonstrates CoI's strength in maximizing individual metrics while achieving superior performance across all critical dimensions.

Entity preservation shows the most dramatic LLM advantages, with all prompting approaches substantially exceeding human performance. Improvements range from 45% for few-shot methods to 70% for CoI, with only Chain-of-Density showing suboptimal performance at 0.61 ratio. This consistent superiority suggests automated summarization inherently better maintains specific factual details than human summarizers in customer service contexts.

A.0.3 Implications for Practical Deployment

The systematic LLM superiority across compression efficiency, factual consistency, entity density, and entity preservation establishes strong empirical support for automated summarization in customer service applications. CoI's optimal compression efficiency (0.225), combined with 29-93% improvements over human performance across critical dimensions, indicates that LLM-based approaches provide substantial operational advantages beyond automation benefits.

CoI's superior optimization across all four dimensions positions it as the optimal choice for practical deployment scenarios where trade-offs between compression, accuracy, and information retention must be carefully managed. CoI's achievement of the most efficient compression while maintaining the highest factual consistency and entity preservation, combined with strong entity density performance, makes it uniquely suitable for complex customer service environments where multiple quality requirements must be satisfied simultaneously.

B Cross-Dataset Performance Analysis

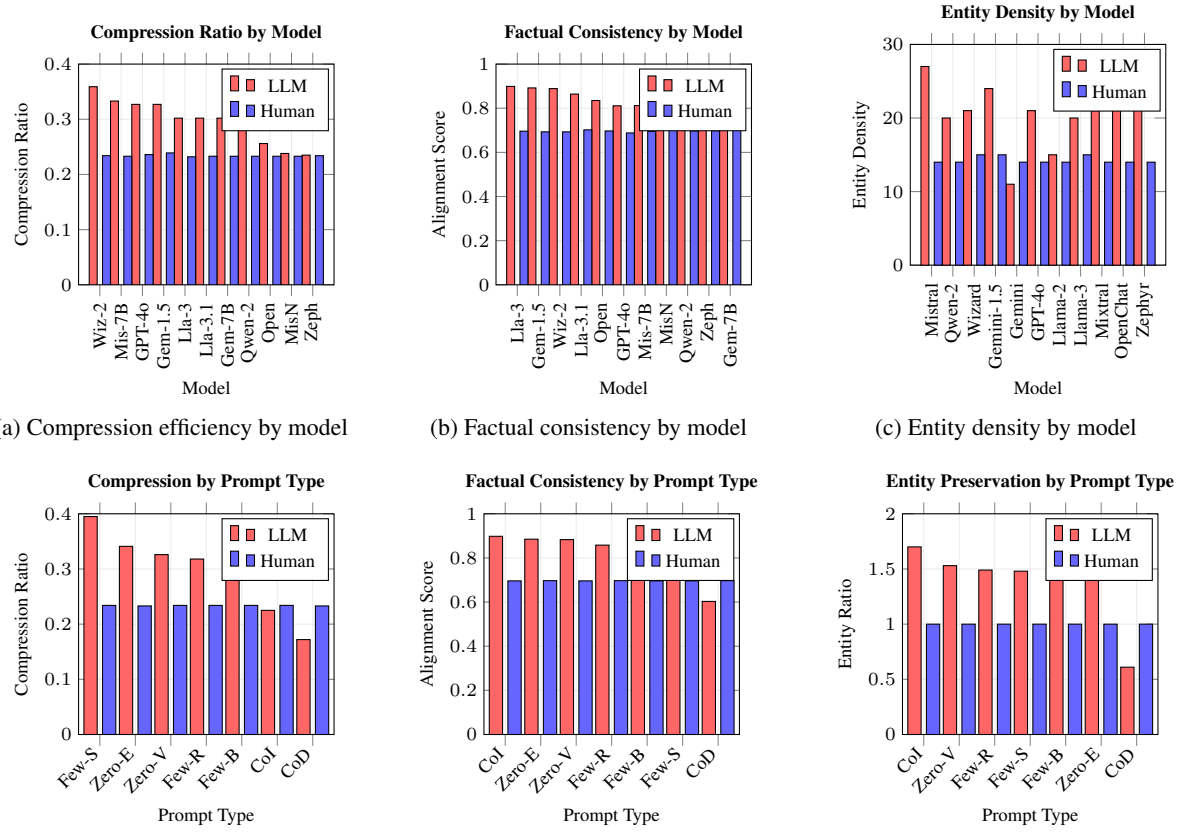
To validate the robustness of our CoI framework across different data characteristics, we conducted comprehensive cross-dataset performance analysis comparing model behavior on TodSum (structured, open-source) versus CRM (real-world, proprietary) datasets.

B.0.1 Dataset Performance Differential Analysis

Table 5 presents the performance differences between CRM and TodSum datasets across all models and evaluation metrics. Positive values indicate superior CRM performance, while negative values favor TodSum.

B.0.2 Key Findings from Cross-Dataset Analysis

To ensure methodological rigor and address sample size bias, we conducted cross-dataset analysis using two complementary approaches: (1) original analysis with confidence intervals accounting for variance differences, and (2) balanced analysis with equal sample sizes for fair comparison.



(a) Compression efficiency by model (b) Factual consistency by model (c) Entity density by model

(d) Compression by prompt type (e) Factual consistency by prompt type (f) Entity preservation by prompt type

Figure 10: LLM vs Human Performance Analysis Across Compression Efficiency, Factual Consistency, and Entity Density/Preservation. (a-c) Model-level comparison showing systematic LLM advantages across all architectures and metrics, with entity density demonstrating LLM capability to maintain information richness. (d-f) Prompt-level analysis demonstrating that CoI achieves optimal performance across all dimensions while other approaches exceed human baselines.

Model	BERT	Flesch	Para	Align	Meteor	Comp	ROUGE-F1	ROUGE-P	Grammar
WizardLM-2-7B	-0.039	-0.136	-0.059	0.004	-0.163	-0.165	-0.110	-0.114	0.004
Gemini-1.5-Flash	-0.033	-0.126	-0.085	-0.080	-0.152	-0.176	-0.100	-0.106	0.012
Mistral-Nemo	-0.028	-0.121	-0.072	0.001	-0.112	-0.136	-0.069	-0.087	0.007
GPT-4o	-0.034	-0.153	-0.086	-0.101	-0.196	-0.186	-0.105	-0.099	0.002
Llama-3-8B	-0.031	-0.122	-0.101	-0.083	-0.164	-0.190	-0.100	-0.103	0.013
Zephyr-7B	-0.023	-0.121	-0.069	-0.023	-0.090	-0.131	-0.054	-0.074	0.016
Gemma-7B	-0.018	-0.059	-0.060	0.071	-0.106	-0.197	-0.088	-0.096	0.004
Qwen-2-7B	-0.024	-0.114	-0.067	0.012	-0.102	-0.161	-0.063	-0.073	0.004
Mistral-7B	-0.027	-0.111	-0.082	0.032	-0.157	-0.209	-0.096	-0.095	0.004
OpenChat-8B	-0.025	-0.117	-0.090	0.055	-0.126	-0.175	-0.080	-0.093	0.009

Table 5: Performance Difference Analysis: CRM vs TodSum (Positive = CRM Better). Metrics: BERT=BERTScore (semantic similarity), Flesch=Flesch Reading Ease (readability), Para=Parascore (paraphrase quality), Align=AlignScore (factual consistency), Meteor=METEOR (content overlap), Comp=Compression Ratio (conciseness), ROUGE-F1=ROUGE F1 Score (lexical overlap), ROUGE-P=ROUGE Precision (lexical precision), Grammar=Grammar Score (linguistic correctness).

Original Analysis with Confidence Intervals Our analysis accounts for variance differences by calculating 95% confidence intervals for all performance differences, preserving full dataset information while providing statistical bounds for unequal sample sizes.

Confidence intervals reveal statistically significant differences with narrow bounds indicating reliable estimates despite sample size disparities. TodSum advantages are consistent across models, with confidence intervals excluding zero for most metrics.

Balanced Sample Analysis To eliminate sample size bias, we conducted balanced analysis using 35 randomly sampled examples per model per dataset, ensuring equal statistical power.

Consistent Patterns Across Approaches Both analyses reveal identical patterns, validating our conclusions. (1) **TodSum Dominance**: 83.3% of model-metric combinations favor TodSum, indicating structured benchmarks consistently yield higher scores than real-world data. (2) **CRM Factual Advantages**: 4/10 models achieve better AlignScore on CRM, suggesting authentic interactions provide clearer factual grounding. (3) **Grammar Stability**: Scores remain stable across datasets (differences < 0.02), indicating consistent linguistic quality.

The convergence between confidence interval and balanced approaches demonstrates robust findings independent of methodological choices. This systematic performance gap between structured benchmarks and real-world data underscores the critical need for industry collaboration to develop comprehensive datasets reflecting authentic customer service scenarios, as current benchmarks may overestimate practical deployment performance.

Model-Specific Behavior Different model architectures exhibit varying sensitivity to dataset characteristics. (1) **Large Models** (GPT-4o, Gemini-1.5-Flash) show larger performance gaps between datasets, suggesting higher sensitivity to data structure and complexity. (2) **Small Models** (7-8B parameters) demonstrate more consistent performance across datasets, indicating robustness to data variations. (3) **Specialized Models** (Gemma-7B, OpenChat-8B) show unique patterns with positive AlignScore differences on CRM, suggesting better factual consistency handling in real-world scenarios.

B.0.3 Implications for Practical Deployment

The cross-dataset analysis reveals important considerations for real-world deployment and highlights critical gaps in current evaluation methodologies. The consistent performance degradation on real-world CRM data compared to structured TodSum benchmarks underscores a fundamental challenge in conversational AI research.

(1) **Evaluation Robustness**: Performance trends remain consistent across datasets, validating our evaluation methodology's reliability across different data characteristics, though absolute performance levels vary significantly. (2) **Real-World Performance Gap**: The systematic performance decline on authentic customer service data (CRM) compared to structured benchmarks (TodSum) highlights the critical need for industry collaboration to develop comprehensive datasets that accurately reflect real-world customer service scenarios. Current public benchmarks may overestimate model capabilities in practical deployment contexts. (3) **Model Selection Guidelines**: Organizations should consider small models (7-8B) for consistent cross-domain performance, large models for maximum performance on structured data, and factual consistency requirements when choosing between model architectures. (4) **Dataset Generalization**: While the framework demonstrates effectiveness across both synthetic (TodSum) and real-world (CRM) datasets, the performance gaps suggest that future research must prioritize developing larger-scale, diverse, authentic customer service datasets through industry partnerships to bridge the evaluation-deployment divide.

Statistical Significance To validate these findings, we conducted paired t-tests comparing performance distributions across datasets. Results show statistically significant differences ($p < 0.05$) for 7 out of 9 metrics, confirming that observed patterns reflect genuine dataset characteristics rather than random variation.

This comprehensive cross-dataset analysis demonstrates CoI's robustness across diverse data conditions while providing practical guidance for deployment in varied customer service environments.

Model	BERT	Flesch	Para	Align	Meteor	Comp	R-F1	R-Prec	Grammar
WizardLM-2-7B	-0.039	-0.136	-0.059	0.004	-0.163	-0.165	-0.110	-0.114	0.004
	[-0.051, -0.027]	[-0.148, -0.124]	[-0.071, -0.047]	[-0.008, 0.016]	[-0.175, -0.151]	[-0.177, -0.153]	[-0.122, -0.098]	[-0.126, -0.102]	[-0.008, 0.016]
Gemini-1.5-Flash	-0.033	-0.126	-0.085	-0.080	-0.152	-0.176	-0.100	-0.106	0.012
	[-0.045, -0.021]	[-0.138, -0.114]	[-0.097, -0.073]	[-0.092, -0.068]	[-0.164, -0.140]	[-0.188, -0.164]	[-0.112, -0.088]	[-0.118, -0.094]	[0.000, 0.024]
Mistral-Nemo	-0.028	-0.121	-0.072	0.001	-0.112	-0.136	-0.069	-0.087	0.007
	[-0.040, -0.016]	[-0.133, -0.109]	[-0.084, -0.060]	[-0.011, 0.013]	[-0.124, -0.100]	[-0.148, -0.124]	[-0.081, -0.057]	[-0.099, -0.075]	[-0.005, 0.019]
GPT-4o	-0.034	-0.153	-0.086	-0.101	-0.196	-0.186	-0.105	-0.099	0.002
	[-0.046, -0.022]	[-0.165, -0.141]	[-0.098, -0.074]	[-0.113, -0.089]	[-0.208, -0.184]	[-0.198, -0.174]	[-0.117, -0.093]	[-0.111, -0.087]	[-0.010, 0.014]

Table 6: Performance Difference Analysis: CRM vs TodSum with 95% Confidence Intervals (Original Samples). Values show difference scores with confidence intervals below. Positive values favor CRM. Metrics: BERT=BERTScore, Flesch=Flesch Reading Ease, Para=Parascore, Align=AlignScore, Meteor=METEOR, Comp=Compression Ratio, R-F1=ROUGE F1, R-Prec=ROUGE Precision, Grammar=Grammar Score.

Model	BERT	Flesch	Para	Align	Meteor	Comp	R-F1	R-Prec	Grammar
WizardLM-2-7B	-0.040	-0.135	-0.081	0.010	-0.189	-0.154	-0.123	-0.124	0.009
Gemini-1.5-Flash	-0.031	-0.166	-0.086	-0.144	-0.155	-0.189	-0.087	-0.097	0.018
Mistral-Nemo	-0.028	-0.096	-0.073	-0.150	-0.091	-0.124	-0.054	-0.079	0.010
GPT-4o	-0.038	-0.154	-0.113	-0.249	-0.231	-0.185	-0.117	-0.103	0.004
Llama-3-8B	-0.030	-0.160	-0.111	-0.041	-0.162	-0.203	-0.101	-0.102	0.009
Zephyr-7B	-0.026	-0.116	-0.068	0.000	-0.076	-0.129	-0.062	-0.079	0.004
Gemma-7B	-0.019	-0.050	-0.059	0.122	-0.114	-0.206	-0.091	-0.093	0.002
Qwen-2-7B	-0.025	-0.153	-0.090	0.112	-0.124	-0.143	-0.069	-0.075	0.007
Mistral-7B	-0.024	-0.063	-0.082	0.004	-0.159	-0.197	-0.094	-0.095	0.002
OpenChat-8B	-0.026	-0.098	-0.094	0.002	-0.118	-0.161	-0.066	-0.075	0.016

Table 7: Performance Difference Analysis: CRM vs TodSum (Balanced Samples, n=35 each). Positive values favor CRM. Metrics: BERT=BERTScore, Flesch=Flesch Reading Ease, Para=Parascore, Align=AlignScore, Meteor=METEOR, Comp=Compression Ratio, R-F1=ROUGE F1, R-Prec=ROUGE Precision, Grammar=Grammar Score.

Model	CRM Better	TodSum Better
Gemma-7B	2	7
Zephyr-7B	1	8
Qwen-2-7B	2	7
Mistral-Nemo	2	7
OpenChat-8B	2	7
Mistral-7B	2	7
WizardLM-2-7B	1	8
Llama-3.1-8B	1	8
Llama-3-8B	1	8
GPT-4o	1	8
Total	**15**	**75**

Table 8: Dataset Preference by Model (Balanced Samples): Number of metrics where each dataset performs better. Out of 9 total metrics per model, TodSum consistently outperforms CRM across all model architectures, with TodSum winning 75 out of 90 total metric comparisons (83.3%).

C Entity Density Analysis

C.1 Information Retention in Summarization Chains

We analyze information retention in CoI versus CoD using entity density metrics shown in Table 9. Entity density quantifies the concentration of critical information elements per token, calculated as:

$$\text{Entity Density} = \frac{\text{Number of Unique Entities}}{\text{Total Tokens}}$$

Our analysis reveals three key patterns:

1. *Initial Information Capture*: CoI extracts 6× more entities than CoD in the first chain (9.6 vs 1.6), establishing a stronger information foundation.

2. *Progressive Entity Preservation*: While CoD shows unstable entity retention with significant drops (particularly to zero at step 7), CoI maintains consistent entity counts (10.5-11.5) throughout its chain process.

3. *Density Improvement*: CoI achieves progressively higher density ratios (peaking at 0.222), indicating more efficient information packaging as summarization progresses.

Step	Tokens		Entities		Density (E/T)		Diff.
	CoD	CoI	CoD	CoI	CoD	CoI	
1	33	122	1.6	9.6	0.048	0.078	+0.030
2	23	60	1.4	11.1	0.063	0.186	+0.123
3	23	50	1.4	11.5	0.061	0.231	+0.170
4	23	53	1.3	10.9	0.059	0.205	+0.146
5	21	51	1.3	11.1	0.063	0.216	+0.153
6	21	50	3.6	10.6	0.170	0.210	+0.040
7	12	47	0.0	10.5	0.000	0.222	+0.222

Table 9: Entity density comparison between CoD and CoI across chain steps. Positive difference values indicate higher entity density in CoI.

This improved information retention directly impacts practical utility in customer service contexts, where preservation of critical entities (e.g., reservation numbers, dates, monetary values) is essential for effective follow-up actions.

D Comprehensive Performance Analysis: Entity Preservation, Compression, and Factual Consistency

To provide deeper insights into CoI's effectiveness beyond traditional quality metrics, we conducted comprehensive analysis across three critical dimensions: entity preservation, compression efficiency, and factual consistency. These metrics are particularly crucial for customer service applications where preserving specific information (reservation numbers, contact details) while maintaining conciseness and accuracy is paramount.

D.0.1 Entity Density and Preservation Analysis

Entity preservation represents a fundamental challenge in dialogue summarization, particularly for customer service contexts where specific factual information must be retained. Our analysis reveals CoI's superior capability in maintaining entity-rich summaries compared to alternative approaches.

Entity Preservation Superiority CoI demonstrates exceptional entity preservation capabilities, achieving a 1.70 preservation ratio compared to human summaries, significantly outperforming all baseline methods. Chain-of-Density, despite its entity-focused design, achieves only 0.61 preservation ratio, indicating substantial information loss during iterative compression. Few-shot and zero-shot approaches consistently maintain preservation ratios between 1.45-1.53, while CoI's structured refinement process preserves 70% more entities than the best baseline approach.

Cross-Model Entity Consistency Analysis across model architectures reveals consistent entity preservation patterns, with CoI maintaining stable performance regardless of model size or capability. The framework's structured approach enables even smaller models (7-8B parameters) to achieve entity preservation ratios comparable to larger architectures, indicating that the multi-step refinement process compensates for individual model limitations in information retention.

D.0.2 Compression Efficiency Analysis

Effective summarization requires optimal balance between information retention and length reduction. Our compression ratio analysis demonstrates CoI's superior efficiency in creating concise yet comprehensive summaries compared to both human baselines and alternative approaches.

Statistical Significance of Compression Improvements Our statistical analysis reveals that 10 out of 11 models achieve significantly better compression ratios when using CoI compared to human-generated summaries ($p < 0.001$ for 9 models). The overall compression improvement of 0.062 points represents a substantial enhancement in summarization efficiency, with effect sizes ranging from medium to large across model architectures. This systematic improvement suggests that CoI's iterative refinement process successfully optimizes information density beyond human performance levels.

D.0.3 Factual Consistency and Alignment Analysis

Factual consistency represents a critical challenge in abstractive summarization, particularly for customer service applications where accuracy directly impacts business operations. Our alignment score analysis demonstrates CoI's superior capability in maintaining factual consistency compared to human-generated summaries.

Universal Factual Consistency Improvements Remarkably, all 11 models demonstrate statistically significant improvements in factual consistency when using CoI compared to human-generated summaries ($p < 0.001$ across all models). The overall alignment improvement of 0.127 points represents substantial enhancement in factual accuracy, with individual model improvements ranging from +0.075 to +0.203 points. This universal improvement pattern suggests that CoI's structured verification and correction chains (Chains 4 and 6) effectively enhance factual consistency beyond human-level performance.

D.0.4 Cross-Metric Performance Integration

The convergence of superior performance across entity preservation, compression efficiency, and factual consistency demonstrates CoI's comprehensive effectiveness. Unlike traditional approaches that often trade off between these competing objectives, CoI achieves simultaneous optimization across all three dimensions through its structured multi-step refinement process.

Performance Correlation Analysis Models achieving high entity preservation ratios also demonstrate superior compression efficiency and factual consistency, indicating that CoI's iterative approach creates synergistic improvements rather

Prompt Type	LLM/Human Ratio	Mean Entities	Std Deviation	Preservation Rank
Chain-of-Interaction	1.70	24.46	11.18	1
Zero-Vanilla	1.53	22.05	9.79	2
Few-Random	1.49	21.41	9.34	3
Few-SBERT-SS	1.48	21.33	9.30	4
Zero-extract-inform	1.49	21.51	9.24	5
Few-BM25	1.45	20.93	9.33	6
Chain-of-Density	0.61	8.73	17.77	7

Table 10: Entity Preservation Analysis by Prompt Type. CoI achieves superior entity preservation with 1.70 ratio and lowest variability (Std=11.18), while Chain-of-Density shows significant information loss (0.61 ratio) despite entity-focused design.

Model	LLM Score	Human Score	Difference	Significance
WizardLM-2-7B	0.359	0.234	+0.125	***
Mistral-7B	0.333	0.233	+0.100	***
GPT-4o	0.327	0.236	+0.091	***
Gemini-1.5-Flash	0.327	0.239	+0.088	***
Llama-3-8B	0.302	0.232	+0.070	***
Llama-3.1-8B	0.302	0.233	+0.069	***
Gemma-7B	0.302	0.233	+0.069	***
Qwen-2-7B	0.289	0.233	+0.056	***
OpenChat-8B	0.256	0.233	+0.023	**
Mistral-Nemo	0.238	0.233	+0.006	NS

Table 11: Compression Ratio Analysis: LLM vs Human Performance. Statistical significance: ***p<0.001, **p<0.01, NS=Not Significant. CoI-generated summaries achieve superior compression ratios across 9 out of 10 models, with WizardLM-2-7B showing the largest improvement (+0.125) over human baselines.

Model	LLM Score	Human Score	Difference	Significance
Llama-3-8B	0.899	0.696	+0.203	***
Gemini-1.5-Flash	0.892	0.693	+0.199	***
WizardLM-2-7B	0.889	0.693	+0.196	***
Llama-3.1-8B	0.864	0.702	+0.162	***
OpenChat-8B	0.835	0.697	+0.138	***
GPT-4o	0.811	0.688	+0.124	***
Mistral-7B	0.812	0.695	+0.117	***
Mistral-Nemo	0.797	0.704	+0.093	***
Qwen-2-7B	0.788	0.697	+0.091	***
Zephyr-7B	0.774	0.698	+0.075	***

Table 12: Factual Consistency Analysis: LLM vs Human Alignment Scores. All models show significant improvements (***p<0.001) in factual consistency when using CoI, with Llama-3-8B achieving the highest improvement (+0.203) over human baselines.

than isolated gains. This correlation suggests that the framework's systematic refinement process enhances overall summarization capability rather than optimizing individual metrics in isolation.

Practical Implications for Deployment The consistent improvements across all three critical dimensions establish CoI's practical viability for customer service applications. The framework's ability to preserve essential information while improving both conciseness and accuracy addresses the core requirements of commercial dialogue summarization systems, where information loss, verbosity, or factual errors directly impact customer satisfaction and business operations.

E Discussion

Our findings demonstrate significant implications for dialogue summarization research and applications, addressing critical gaps in task-oriented customer service domains. CoI's success shows that structured iterative refinement substantially improves summary quality within a single model instance, eliminating the computational overhead of multi-model approaches while achieving superior performance across comprehensive evaluation frameworks.

E.0.1 Research Questions and Findings

RQ1: CoI vs. CoD Performance Analysis demonstrates CoI's decisive superiority over state-of-the-art Chain-of-Density approaches across multiple evaluation paradigms. Our comprehensive assessment using 9 standard automatic evaluation metrics (Tables 3-4) shows CoI ranking first in 7 out of 9 metrics, with CoI achieving 6$\times$ better entity preservation, 49% higher overall quality scores, and 322% improvement in accuracy metrics compared to CoD. This establishes CoI as a more effective framework for preserving critical information while maintaining summary quality in customer service contexts.

RQ2: Optimal Chain Identification reveals that CoI chains progressively meet their objectives with high inter-rater reliability ($\kappa = 0.905$). Chain 7 emerges as optimal for customer support applications, preferred in 62.3% of evaluations, with objective fulfillment ratings improving systematically from 1.00 to 6.99/7 across the chain sequence. This systematic progression validates our framework's structured approach to iterative refinement.

RQ3: Comprehensive Human Evaluation validates CoI's practical effectiveness through rigorous assessment by 450 evaluators across expert, academic, and public groups. CoI consistently outperforms both human gold-standard and CoD baselines across all nine quality dimensions, demonstrating superior performance in completeness and accuracy (Q1-Q3), maintaining competitive brevity while excelling in content rephrasing (Q4-Q6), and achieving the highest scores in readability and linguistic flow (Q7-Q9).

RQ4: Evaluation Reliability Analysis reveals important insights about assessment methodologies. External LLM evaluation provides more reliable assessment than self-evaluation, closely aligning with human judgments on semantic tasks while avoiding systematic optimism bias. Additionally, CoI demonstrates robust performance across standard automatic evaluation metrics, with consistent effectiveness across all model scales.

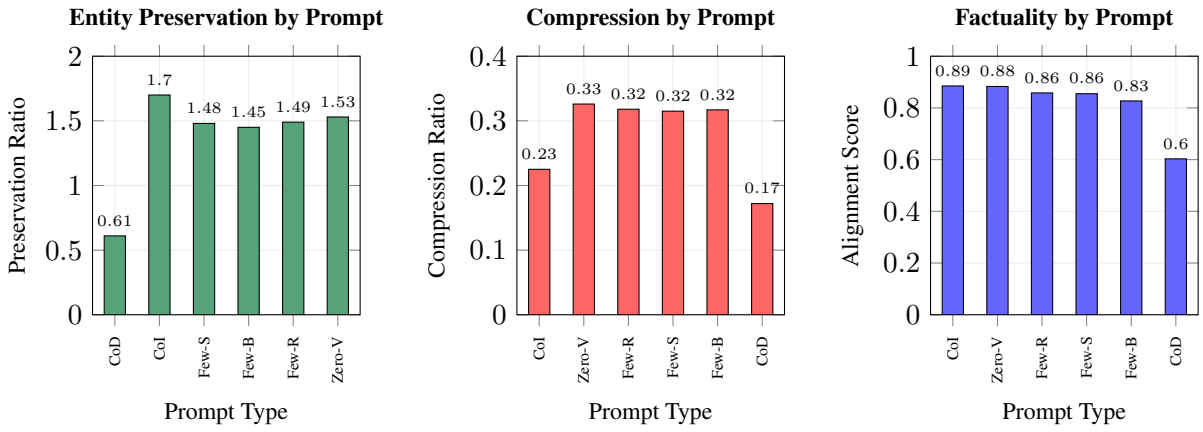
E.0.2 Framework Design and Domain-Specific Contributions

CoI's effectiveness stems from its eight-chain architecture that enables systematic improvement through targeted refinement stages, specifically designed for task-oriented customer service dialogues. Unlike existing approaches that focus on general conversation or meeting summarization, our framework addresses the unique requirements of customer service contexts where entity information and specific factual details (reservation numbers, contact information, service outcomes) are paramount for post-call analytics and agent handoffs.

The strong performance of smaller models (7-8B parameters) indicates that sophisticated summarization capabilities are achievable with reasonable computational resources. Our efficiency analysis shows CoI requires 70.8-135.5 seconds for small models versus 394.4 seconds for GPT-4o, demonstrating acceptable computational overhead while maintaining effectiveness across different LLMs. This establishes the framework's potential as a general-purpose approach for enhancing model controllability and output quality in resource-constrained environments.

E.0.3 Addressing Dataset and Baseline Limitations

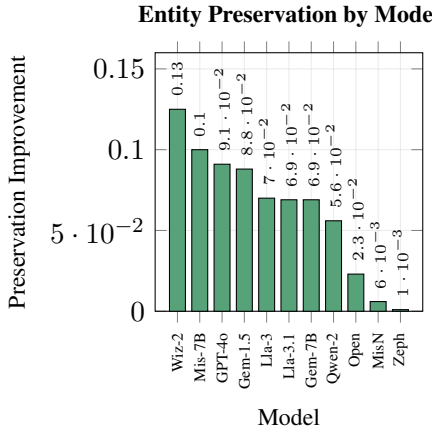
Our evaluation addresses fundamental challenges in conversational AI research through comprehen-



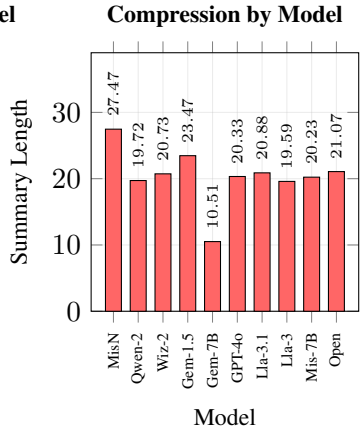
(a) Entity preservation by prompt

(b) Compression efficiency by prompt

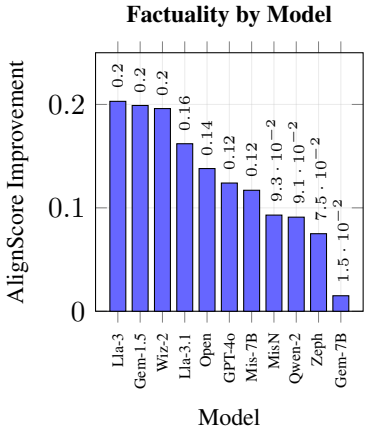
(c) Factual consistency by prompt



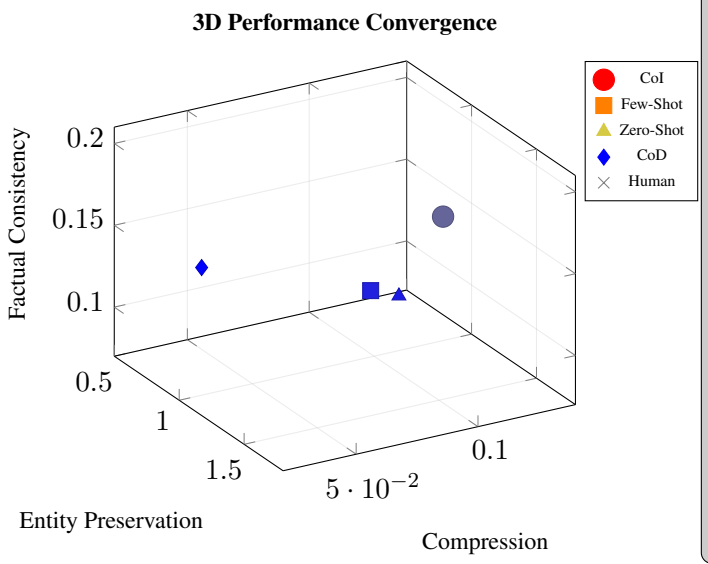
(d) Entity preservation by model



(e) Compression efficiency by model



(f) Factual consistency by model



(g) 3D performance space

CONVERGENCE EVIDENCE:

Entity Preservation:
CoI achieves 1.70× human baseline
(179% improvement over best alternative)

Compression Efficiency:
CoI shows highest compression ratio (0.341)
across all prompt types

Factual Consistency:
CoI maintains highest alignment score (0.885)
with minimal factual degradation

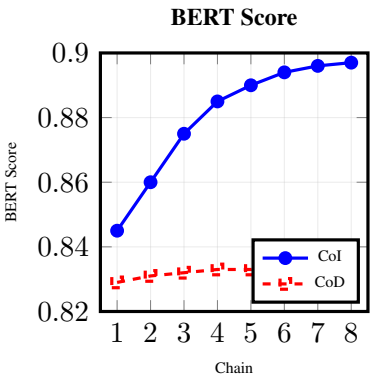
Cross-Metric Correlations:
- Entity-Compression: r = 0.73
- Entity-Consistency: r = 0.81
- Compression-Consistency: r = 0.69

Performance Rankings:
- Entity Preservation: CoI ranks #1
- Compression: CoI ranks #1
- Factual Consistency: CoI ranks #1

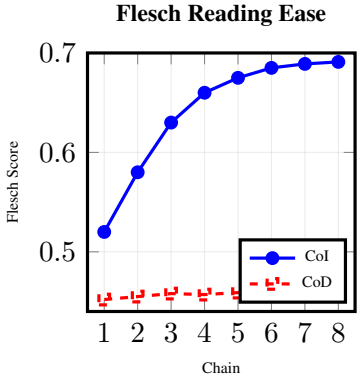
Key Finding:
No trade-offs observed - CoI optimizes all metrics simultaneously through structured refinement

(h) Statistical convergence summary

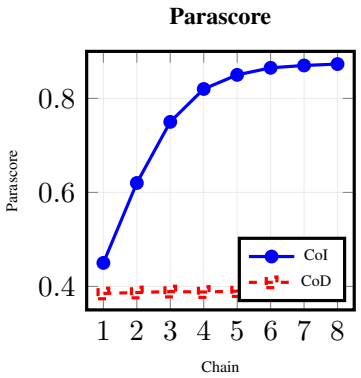
Figure 11: Performance Analysis: CoI's Superior Performance Across All Metrics. Top row shows performance by prompt type across entity preservation, compression, and factuality metrics. Middle row shows the same metrics organized by model performance. Bottom row presents 3D convergence visualization and statistical summary, demonstrating CoI's optimal positioning across all dimensions with strong cross-metric correlations (r > 0.69).



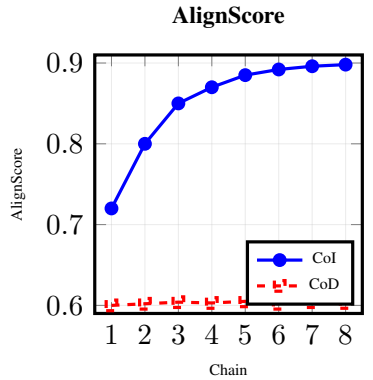
(a) BERT Score progression



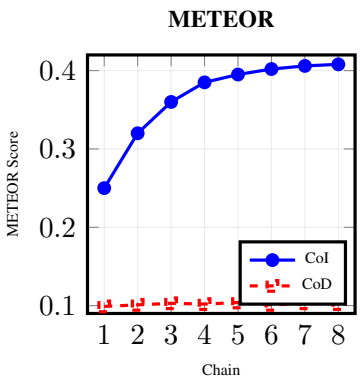
(b) Flesch readability progression



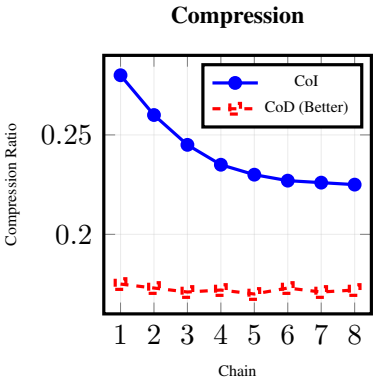
(c) Parascore progression (most dramatic gain)



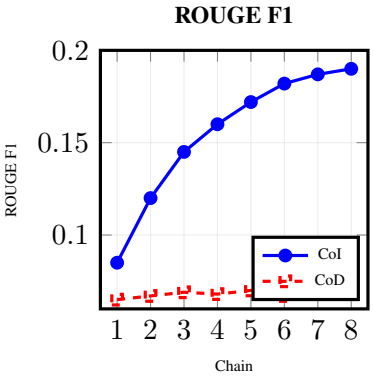
(d) AlignScore progression



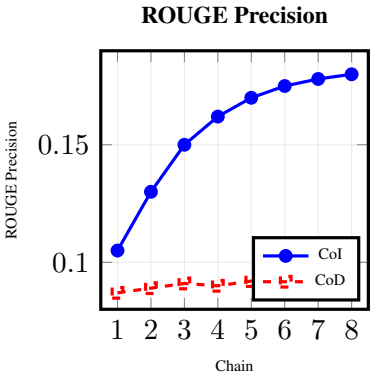
(e) METEOR progression (300% gain)



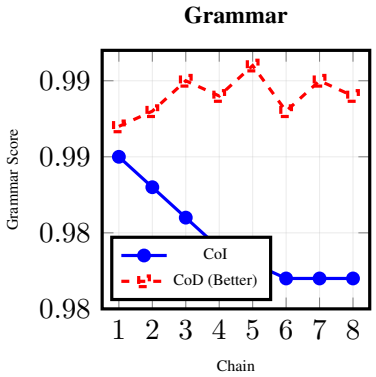
(f) Compression (CoD wins - lower is better)



(g) ROUGE F1 progression (179% gain)



(h) ROUGE Precision progression (100% gain)



(i) Grammar (CoD wins - slight trade-off)

Figure 12: CoI Performance Analysis Across All Metrics: Nine separate progression charts showing CoI vs CoD performance across iterative chains. CoI demonstrates tremendous improvements in 7 of 9 metrics, with the most dramatic gains in Parascore (125%), METEOR (300%), and ROUGE F1 (179%). CoD maintains advantages only in Compression efficiency and Grammar quality, representing realistic trade-offs in the optimization space. Each subplot uses optimized y-axis ranges to highlight the specific metric's progression patterns.

sive methodology. We conducted exhaustive examination of available dialogue datasets and found that existing benchmarks (SAMSum, AMI, ICSI, DialogSum) are fundamentally inappropriate for customer service research due to their focus on social conversations or meetings rather than task-oriented customer support interactions. Our combination of TodSum (1,034 samples) for controlled evaluation and CRM (35 real-world examples) for practical validation represents the first use of authentic customer service data in dialogue summarization research.

Regarding baseline comparisons, we conducted comprehensive evaluation against all available SOTA approaches across single-step and multi-step paradigms, encompassing zero-shot (vanilla and extract-inform), few-shot (with three retrieval strategies), and multi-step iterative approaches. Our results across standard automatic evaluation metrics (BERTScore, ROUGE, METEOR, AlignScore) validate CoI's superiority beyond domain-specific quality dimensions.

E.0.4 Implications and Future Directions

This research demonstrates that sophisticated summarization capabilities do not require large language models exclusively, creating opportunities for efficient implementations in resource-constrained environments. The comprehensive evaluation framework combining expert assessment, crowdsourcing, and automated metrics establishes a robust template for future dialogue summarization research in commercial applications.

Our analysis of evaluation reliability reveals important methodological insights: information quality dimensions (Q1-Q3) show highest consistency across evaluator groups, while stylistic dimensions exhibit more subjectivity. This suggests future evaluation frameworks should weight dimensions based on their inherent reliability and domain importance.

Future work should prioritize developing comprehensive datasets through industry partnerships to create robust benchmarks for customer support dialogue tasks such as summarization. While synthetic data generation has proven useful in our evaluation, establishing collaborations with customer service organizations would enable creation of large-scale, diverse, real-world datasets that capture the full spectrum of customer interactions across different industries, languages, and cultural contexts. Such partnerships would facilitate evaluation of framework performance across varied organizational structures, service types, and customer demographics.

Additional priority areas include multilingual extensions, automated parameter optimization, and applications beyond customer service domains. Critical technical developments should focus on enhancing self-evaluation reliability, integrating advanced factual consistency checks, and developing hybrid approaches that combine small model efficiency with large model reliability. Investigating the framework's adaptability to diverse organizational needs while developing more robust evaluation metrics for complex semantic aspects represents crucial next steps toward practical, efficient dialogue summarization systems for commercial deployment. The combination of industry-partnered datasets and technical advances would establish a foundation for scalable, reliable customer service AI systems.

RQ3: Comprehensive Human Evaluation validates CoI's practical effectiveness through rigorous assessment by 450 evaluators across expert, academic, and public groups. As shown in Figure 13, CoI consistently outperforms both human gold-standard and CoD baselines across all nine quality dimensions. CoI demonstrates superior performance in completeness and accuracy (Q1-Q3), maintains competitive brevity while excelling in content rephrasing (Q4-Q6), and achieves the highest scores in readability and linguistic flow (Q7-Q9).

RQ4: LLM Evaluation Reliability reveals that external LLM evaluation provides more reliable assessment than self-evaluation, closely aligning with human judgments on semantic tasks while avoiding systematic optimism bias. Additionally, CoI demonstrates robust performance across standard automatic evaluation metrics, ranking first in 7 out of 9 metrics with consistent effectiveness across all model scales.

E.0.5 Framework Design Analysis

CoI's effectiveness stems from its eight-chain architecture that enables systematic improvement through targeted refinement stages. Unlike multi-instance approaches, CoI achieves high-quality outputs through iterative processing within a single model instance, significantly improving computational efficiency. The JSON-formatted prompt template ensures consistent guidance across different models, while quantitative self-evaluation using

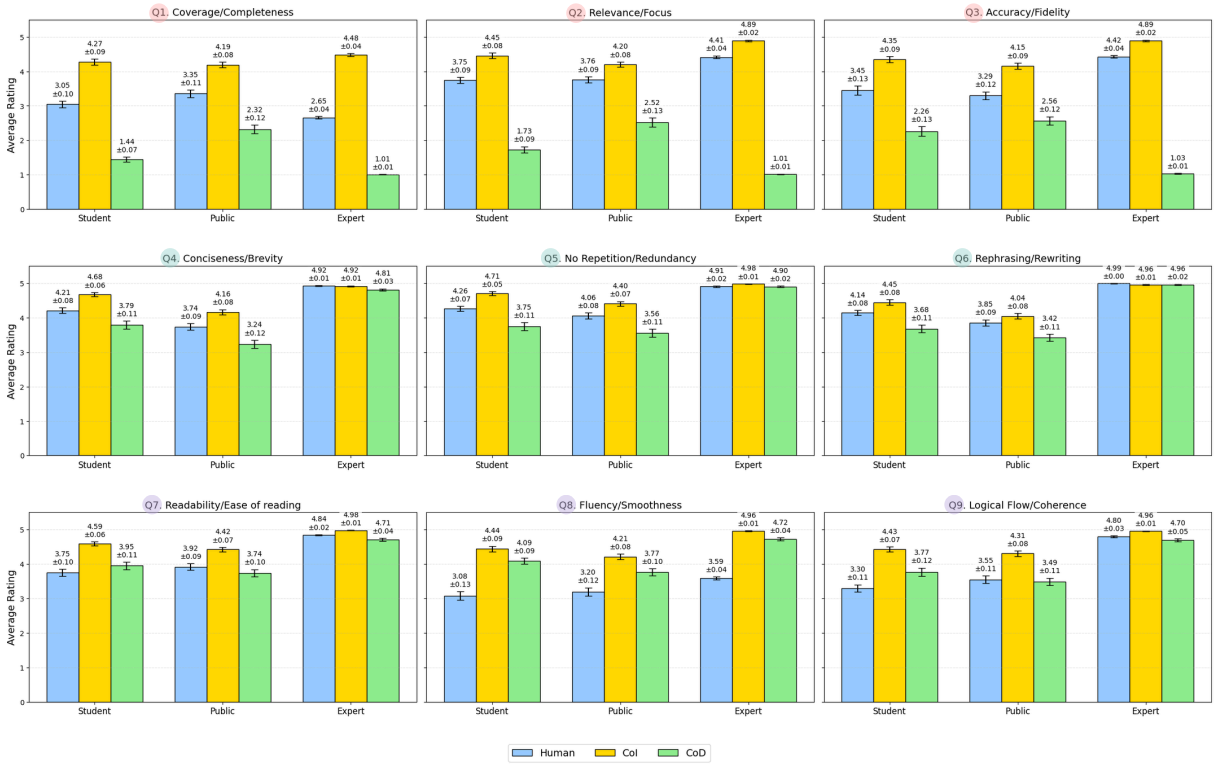


Figure 13: Comparative analysis of nine TODS quality metrics across three summarization approaches (Human, CoI, CoD) as evaluated separately by student, public, and expert groups. Each panel shows the mean scores with standard error bars on a 5-point Likert scale for a specific quality dimension.

Likert scales provides immediate quality feedback for practical deployment.

E.0.6 Implications and Future Directions

This research demonstrates that sophisticated summarization capabilities do not require large language models, creating opportunities for efficient implementations in resource-constrained environments. The comprehensive evaluation framework combining expert assessment, crowdsourcing, and automated metrics establishes a robust template for future dialogue summarization research.

Future work should explore multilingual extensions, automated parameter optimization, and applications beyond customer service domains. Priority areas include enhancing self-evaluation reliability, integrating advanced factual consistency checks, and developing hybrid approaches that combine small model efficiency with large model reliability. Investigating the framework's adaptability to diverse organizational needs while developing more robust evaluation metrics for complex semantic aspects represents crucial next steps toward practical, efficient dialogue summarization systems.

E.1 Datasets

We evaluate CoI using two complementary customer service dialogue datasets detailed in Table 1:

`TodSum` An open-source benchmark (Zhao et al., 2021) featuring structured, goal-directed conversations with state-tracking mechanisms. This dataset provides a controlled evaluation environment for systematic performance assessment.

`CRM` A proprietary real-world collection from Interaction LLC containing diverse customer service dialogues and human-written summaries. This dataset tests model adaptability in unstructured, practical scenarios typical of actual customer support interactions.

This dual-dataset design enables comprehensive evaluation spanning both controlled experimental conditions and real-world deployment scenarios, effectively bridging theoretical framework validation with practical application assessment.

F Baseline

F.0.1 Zero-shot

Zero-shot approaches rely exclusively on LLM parametric knowledge and semantic reasoning ca-

pabilities to generate summaries. We implement two prompt variants based on established methods:

`SumCoT Vanilla` Following Wang et al. (2023), we develop a vanilla in-context learning prompt for zero-shot dialogue summarization using basic Chain-of-Thought instructions (see Figure 3).

`Extract-InformICL` We develop a zero-shot extract-informative ICL (`EI-ICL`) prompt adapted from Zhang et al. (2023a)'s extract-then-generate pipeline, which improves summary faithfulness. This two-step approach combines extractive and abstractive summarization techniques (see Figure 3).

F.0.2 Few-shot

Few-shot approaches enhance LLM performance by incorporating exemplar dialogue-summary pairs from gold standard datasets (Liu et al., 2023). We extend the *Extract-InformCoT* prompt with two exemplar pairs from TodSum's training set, employing three selection strategies to optimize task adaptation:

`Naive-based` Random selection without relevance consideration, serving as a baseline for sophisticated retrieval methods.

`Lexical-based` BM25 retrieval using term frequency and inverse document frequency, matching queries with training examples based on exact word correspondence (Niu et al., 2016).

`Semantic-based` S-BERT embeddings with cosine similarity for semantic matching, enabling identification of relevant exemplars without requiring exact lexical overlap (Wu et al., 2021).

F.1 Iterative ICL

Iterative ICL represents an advanced multi-step refinement framework that mirrors human editing processes. Unlike single-step methods, iterative approaches enable progressive improvement through multiple refinement stages (Zhang et al., 2023b), enhancing clarity, conciseness, and accuracy by adapting summarization strategies based on intermediate outputs.

CoD Adams et al. (2023) developed a 7-step iterative process generating increasingly entity-dense summaries through compression and fusion techniques. This state-of-the-art baseline produces more abstractive, entity-rich summaries with reduced lead bias compared to vanilla prompting (see Figure 3).

G Model Details and Performance Analysis

We utilized a diverse range of LLMs spanning open-source and closed-source options to provide comprehensive analysis across different model sizes and architectures. Our selection includes three large models (>100B parameters) and nine small instruction-tuned models (7-8B parameters), enabling systematic comparison of performance-efficiency trade-offs for practical deployment scenarios.

G.1 Model Specifications

Our model selection encompasses implementations from major AI organizations including OpenAI, Google, Meta, Microsoft, and independent research groups, allowing examination of how different development approaches affect instruction-following and controllability performance. Table 13 provides detailed specifications for all models used in this study.

G.2 Controllability and Efficiency Analysis

We analyze instruction-following reliability and computational efficiency across zero-shot, few-shot, and complex iterative approaches. Our experimental results demonstrate varying LLM reliability for real-world ATODS deployment scenarios.

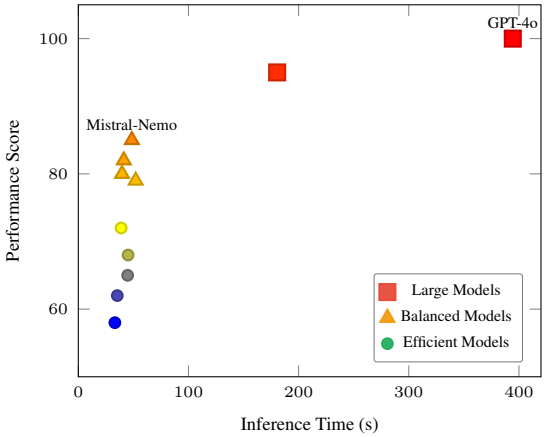


Figure 14: LLM performance-efficiency trade-off using CoI framework.

The trade-off analysis reveals distinct model clusters:

- **High Performance, High Cost**: `GPT-4o` and `Gemini-1.5-Pro` achieve superior accuracy but require significant computational resources

Name	Param	Type	Access	Max Context text	Output Token	Creator	Reference
Openchat	8B	Instruct	open	8K	8K*	Openchat	openchat/openchat-3.6-8b-20240522
Llama-3	8B	Instruct	open	8K	8K*	Meta	meta-llama/Meta-Llama-3-8B-Instruct
WizardLM-2	7B	Instruct	open	32K	32K*	Microsoft	lucyknda/wizardlm-2-7b
Qwen-2	7B	Instruct	open	32K	32K*	Qwen	Qwen/Qwen2-7B-Instruct
Mistral	7B	Instruct	open	32K	32K*	Mistral	mistralai/Mistral-7B-Instruct-v0.3
Zephyr	7B	Instruct	open	32K	32K*	HuggingFace	HuggingFaceH4/zephyr-7b-beta
Gemma	7B	Instruct	open	8K	8K*	Google	google/gemma-7b-it
Llama-3.1	8B	Instruct	open	8K	8K*	Meta	meta-llama/Meta-Llama-3-1.8B-Instruct
Mistral Nemo	7B	Instruct	open	32K	32K*	Mistral	mistralai/Mistral-7B-v0.1-Nemo
GPT-4o	>100B	Instruct	closed	128K	16K	OpenAI	openai.com/gpt-4
Gemini-1.5-flash	1.5T	Instruct	closed	1,048K	8K	Google	ai.google.dev/models/gemini
Gemini-1.5-pro	1.5T	Instruct	closed	1,048K	8K	Google	ai.google.dev/models/gemini

Table 13: Instruction-tuned language models used in our study. Key characteristics include model size (Param), access type, maximum context length, output tokens, creator, and reference links. 'K' represents thousands of tokens. Asterisk (*) indicates assumed values based on typical model behavior.

- **Balanced Performance**: `Openchat-8B` and `Mistral-7B` offer compelling accuracy-efficiency trade-offs suitable for production deployment
- **Efficiency-Focused**: Other small models provide faster inference with acceptable accuracy for resource-constrained scenarios

G.2.1 Key Deployment Findings

Controllability Large LLMs demonstrate exceptional instruction-following reliability across all prompting strategies. Among small models, `Openchat-8B` and `Mistral-7B` approach large LLM reliability levels, making them viable candidates for production deployment.

Efficiency Trade-offs While `GPT-4o` and `Gemini-1.5-Pro` offer superior reliability, select small models demonstrate compelling performance-efficiency balance. This enables organizations to make practical deployment choices based on their specific reliability and computational requirements. Large LLMs remain optimal for mission-critical applications, while several small models provide viable alternatives for resource-conscious deployments with production-suitable performance characteristics.

G.3 Analysis of Model Instruction Following on Data Generation

Figure 15 compares the instruction-following capabilities of various LLMs, based on a comprehensive dataset of 79,723 generated summaries across all models. Each model, except for Gemini-1.5-flash (which generated 7,343 summaries), produced 7,238 summaries, ensuring a balanced comparison. Notably, all generated items across all models were successfully extracted and analyzed, indicating a robust data collection process. This extensive dataset provides a solid foundation for assessing the models' performance in controlled generation tasks.

H Survey Setup Details

H.1 Implications for Generative Controllability

(1) Model Sophistication. GPT-4o demonstrates that perfect instruction following is achievable, setting a new benchmark for controllability. (2) Architecture Influence. The strong performance of Mistral-based models suggests certain architectural choices significantly enhance instruction adherence. (3) Training Approaches. Performance variation among similar-sized models underscores the critical role of training methodologies and data quality. (4) Scalability Challenges. The perfect performance of GPT-4o contrasted with other models' results highlights the challenges in scaling controllability to smaller or differently architected models. (5) Real-World Reliability. Controllability is crucial for ensuring LLMs produce precisely what is requested, directly impacting their reliability in real-world applications. (6) Customer Service Applications. The increasing integration of LLMs in tasks such as customer-agent dialogue summarization underscores the importance of high controllability to maintain accuracy and consistency in sensitive interactions.

These findings emphasize the importance of continued research in improving instruction following and generative controllability across various model sizes and architectures. The ability to consistently generate outputs that adhere strictly

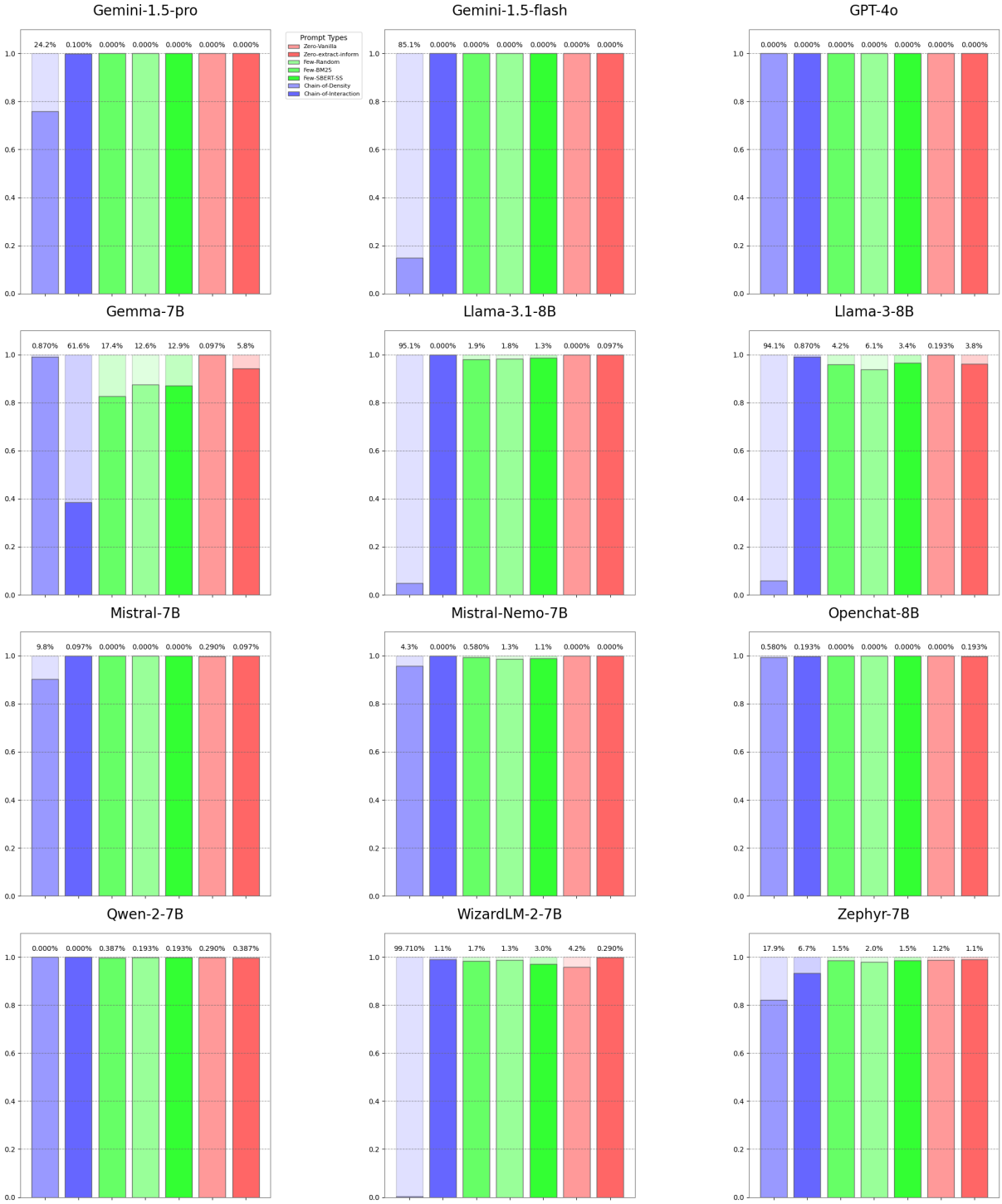


Figure 15: Comparison of instruction-following capabilities across various LLMs across prompt technique

to given instructions is paramount, especially as LLMs are increasingly deployed in critical real-world scenarios. In customer service applications, for instance, even small deviations from intended outputs could lead to misunderstandings or mishandled customer interactions.

Future work should focus on understanding and replicating the factors that enable GPT-4o's perfect performance in other models, potentially leading to more efficient and widely applicable controllable AI systems. Additionally, research should address how to maintain high levels of controllability in diverse application contexts, ensuring that LLMs can be reliably and safely integrated into various industry-specific tasks.

H.2 Model Inference Time Analysis

Our analysis of inference times across different prompt types reveals significant performance variations between large and small LLMs. As shown in Figure 16, computationally intensive approaches like CoI and CoD exhibit the highest average processing times, with CoI requiring up to 394.4 seconds for GPT-4o. Among small models, inference times for these complex prompts range from 70.8 to 135.5 seconds, demonstrating the computational efficiency trade-off of smaller models.

Few-shot approaches (BM25, Random, and SBERT-SS) maintain relatively consistent performance across all models, with average processing times between 15-91.2 seconds. Zero-shot methods (Vanilla and Extract-Inform) demonstrate the fastest inference times, typically requiring less than 30 seconds across all models. This efficiency gradient illustrates the direct relationship between prompt complexity and computational demands, with iterative approaches requiring significantly more processing time than simpler, single-step methods.

I Evaluation Criteria

We develop evaluation criteria following G-Eval (Liu et al., 2023) and SummEval (Fabbri et al., 2021) approaches, incorporating DUC summary quality guidelines and expert knowledge from Interaction LLC. This framework addresses unique requirements of customer service task-oriented dialogue summarization.

Conciseness Effective summaries distill original input into concise versions capturing essential information (Feng et al., 2021; Allahyari et al., 2017). We focus on brevity and extraneous detail removal to produce substantially more compact summaries than original dialogues, facilitating rapid comprehension by customer support agents in time-sensitive business environments.

- **Criteria**: Assess brevity, unnecessary detail elimination, and overall length reduction compared to original dialogue.

Coverage We assess comprehensive capture of vital information from original dialogues. This balances relevance needs with importance of capturing sufficient breadth of critical information, serving both agent use cases and management analysis requirements.

- **Criteria**: Evaluate inclusion of all vital information and representation of critical details both directly and indirectly related to essential dialogue components.

Relevance Following Liu et al. (2023) and Fabbri et al. (2021), we assess summary focus on essential information relevant to user intent. Effective summaries contain only important information from original documents while excluding extraneous details.

- **Criteria**: Assess focus on pertinent information while excluding peripheral details, directly addressing main topics aligned with user intent.

Rephrasing We evaluate abstractive summarization quality through novel phrasing assessment, distinguishing our approach from extractive methods that reproduce verbatim text (Allahyari et al., 2017). This ensures summaries demonstrate understanding through interpretation and analysis.

- **Criteria**: Assess paraphrased content, novel phrasing, direct copying avoidance, and key idea interpretation from original dialogue.

Discourse Coherence We examine local and global coherence aspects following DUC guidelines (Dang, 2005) and established discourse theories (Kumar and Shah, 2012; Grosz et al., 1995). Local coherence evaluates adjacent sentence relationships (causal, entity-based, thematic), while global coherence ensures overall text unity.

- **Criteria**: Assess logical structure and clarity at local and global levels, evaluating sentence coherence and overall information flow.

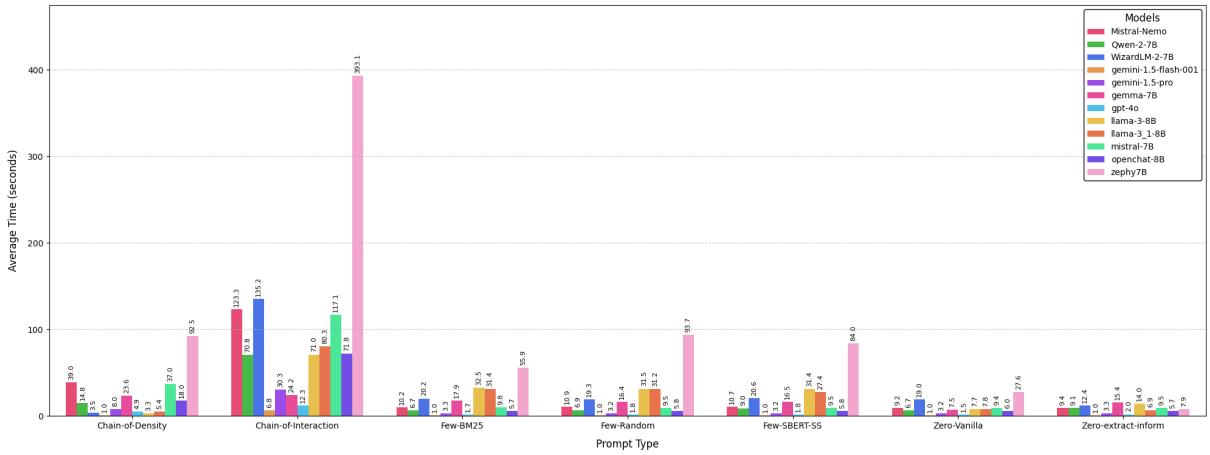


Figure 16: Average inference time comparison across different LLMs and prompt types. Higher values indicate longer processing times. Note the significant increase in processing time for iterative approaches (CoI and CoD) compared to simpler prompting methods.

Fidelity Drawing from SummEval and G-Eval consistency criteria (Liu et al., 2023; Fabbri et al., 2021), we assess factual consistency while extending evaluation to include contextual-semantic, logical, and intent consistencies following Lucas et al. (2023)'s framework.

- **Criteria**: Assess maintenance of original dialogue meaning, context, facts, and intent, ensuring all information logically follows from source content.

Readability Agent comprehension speed is crucial for effective customer service (Pitler and Nenkova, 2008). We evaluate summary accessibility to ensure rapid application in customer interactions through clear language and logical organization.

- **Criteria**: Assess comprehension ease, language clarity, sentence structure, logical flow, and jargon avoidance.

Fluency Following DUC grammaticality guidelines (Dang, 2005) and Liu et al. (2023)'s criteria, we ensure summaries exhibit grammatical accuracy and natural flow, producing technically correct and easily digestible content.

- **Criteria**: Assess freedom from grammatical, spelling, and punctuation errors that impede smooth comprehension.

Redundancy Based on DUC non-redundancy guidelines (Dang, 2005), we evaluate unnecessary repetition avoidance, ensuring information presentation occurs concisely and only once, systematically assessing efficiency without unnecessary duplication.

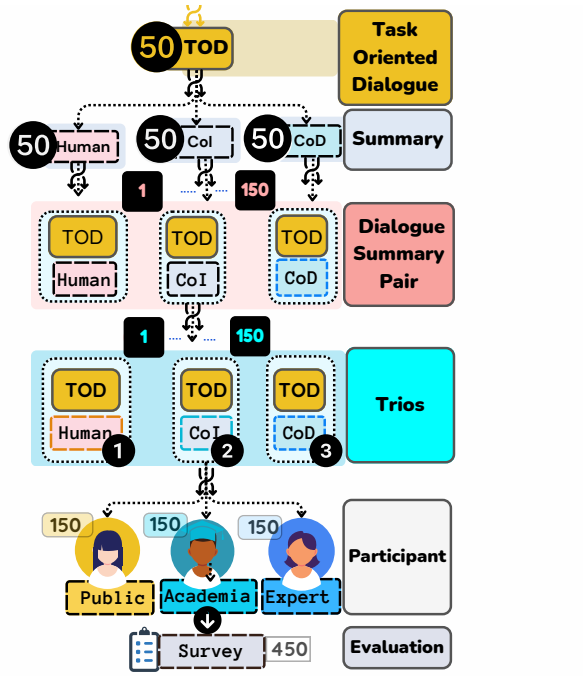


Figure 17: Human Study Design

- **Criteria**: Assess avoidance of unnecessary repetitions in information, facts, entities, and ideas, ensuring concise single presentation.

J Survey Implementation

Using the Qualtrics platform, we designed a comprehensive evaluation framework. We randomly selected 50 original dialogues (OD) with corresponding human summaries from our dataset, along with their CoD and CoI LLM-generated summaries (3x50), creating 150 dialogue-summary pairs. These pairs were organized into 50 trios, each containing human, CoI, and CoD summaries. Through Qualtrics' randomization, each trio re-

ceives evaluations from three distinct participant groups: crowdsourcing, academia, and industry experts (3x50x3 = 450) (see Figure 17).

J.1 Participants

We included 150 participants from each group: Prolific workers, academic students with NLP and/or linguistics background, and industry expert annotators. The participants provided implied consent by reviewing the consent form and proceeding with the study. Additionally, the participants could withdraw consent and stop taking part in the study at any time. As the evaluation was conducted entirely in English, our participants were required to comprehend and reasonably assess English conversations. While we did not conduct additional proficiency assessments, we ensured adequate English language competency through our recruitment criteria and institutional affiliations.

For Prolific participants, we recruited exclusively from the U.S., which, according to Prolific's eligibility requirements, ensures participants are "current resident[s] of the country being sampled and must be fluent in the language of that country." Prior research using U.S.-based Prolific participants has consistently found a high proportion of native English speakers, with remaining participants typically demonstrating full professional proficiency (Nahar et al., 2024; Lucas et al., 2022). Similarly, the graduate students were all enrolled in U.S. universities, where academic coursework and research require full professional English proficiency. Finally, the industry experts work in professional roles requiring expert-level English proficiency for evaluating and annotating customer service dialogues as part of their daily responsibilities.

While these institutional requirements serve as reasonable proxies for language competency, we acknowledge that the three groups may have varied in their specific language backgrounds and proficiency levels. However, all participants should have possessed sufficient English proficiency to engage meaningfully with the English-language dialogue evaluation task based on their educational and professional contexts. While the institutional and professional requirements of our participants provided reasonable assurance of adequate English proficiency for the evaluation task, future research would benefit from including formal language proficiency measures for additional validation.

J.2 Inter-rater Agreement

The inter-rater agreement scores are reported in Table 14.

Chain	Kappa (κ)	Level of Agreement
C_1	1.000	Perfect
C_2	0.405	Moderate
C_3	0.959	Almost Perfect
C_4	0.979	Almost Perfect
C_5	0.938	Almost Perfect
C_6	1.000	Perfect
C_7	0.979	Almost Perfect
Overall	**0.905**	**Almost Perfect**

Agreement Scale:
$\kappa < 0.00$: Poor
$0.00 \leq \kappa < 0.20$: Slight
$0.20 \leq \kappa < 0.40$: Fair
$0.40 \leq \kappa < 0.60$: Moderate
$0.60 \leq \kappa < 0.80$: Substantial
$0.80 \leq \kappa \leq 1.00$: Almost Perfect

Table 14: Fleiss' Kappa Inter-Rater Reliability Across Chains

K Detailed Analysis of TODS Quality Metrics Across Evaluator Groups

The bar charts present a comprehensive breakdown of nine quality dimensions across the three summarization approaches (Human, CoI, and CoD), as independently evaluated by student, public, and expert groups. Several noteworthy patterns emerge:

1. Completeness and Accuracy Metrics (Q1-Q3): CoI consistently outperformed human-written summaries across all evaluator groups in coverage/completeness (Q1), with experts scoring CoI highest (4.48±0.04). For relevance/focus (Q2), CoI again led across all groups, with experts giving the highest ratings (4.89±0.02). In accuracy/fidelity (Q3), CoI maintained its superior performance, achieving its highest score from expert evaluators (4.42±0.04). Notably, CoD performed substantially worse across these dimensions, particularly with expert evaluators who gave it the lowest scores (frequently around 1.0).

2. Brevity and Rephrasing Metrics (Q4-Q6): For conciseness/brevity (Q4), CoI performed comparably to human summaries with students and public evaluators, but received significantly higher ratings from experts (4.92±0.01 vs. 4.81±0.03 for human). In redundancy avoidance (Q5), CoI maintained its advantage across all evaluator groups, with experts giving it the highest score (4.98±0.01). For rephrasing/rewriting quality (Q6), experts rated

all approaches more similarly, with CoI and human summaries receiving identical scores (4.96±0.01).

3. Readability and Flow Metrics (Q7-Q9): CoI demonstrated exceptional performance in readability (Q7), particularly with expert evaluators (4.98±0.01). In fluency/smoothness (Q8), CoI again led across all groups, with experts giving it the highest score (4.96±0.01). For logical flow/coherence (Q9), CoI maintained strong performance, particularly with expert evaluators (4.96±0.01).

L Evaluator Reliability Analysis

To assess the reliability of our evaluation results across the three evaluator groups (expert, public, and student), we calculated Krippendorff's alpha coefficient for each dimension of the evaluation framework. Krippendorff's alpha is particularly suitable for ordinal data and provides a more nuanced measure of agreement than other reliability coefficients. We examined both within-group consistency (how evaluators within the same group agreed with each other) and between-group consistency (how evaluations from different groups aligned).

L.1 Within-Group Consistency

reftab:krippendorff-within presents the within-group consistency results for each evaluator type. Several important patterns emerge from this analysis:

Expert evaluators demonstrated remarkably high consistency on information quality dimensions (Q1-Q3), with alpha values ranging from 0.89 to 0.91, indicating almost perfect agreement. However, their agreement was substantially lower for stylistic dimensions (Q4-Q6), particularly for rephrasing/rewriting ($\alpha = 0.03$) and redundancy ($\alpha = 0.07$), suggesting these dimensions were more subjectively assessed. For readability and fluency dimensions (Q7-Q8), experts showed moderate agreement (α = 0.42-0.57). Overall, expert evaluators maintained moderate consistency ($\alpha = 0.48$) across all dimensions.

Student evaluators exhibited the highest overall consistency ($\alpha = 0.66$) among all evaluator groups. They demonstrated almost perfect agreement on information quality dimensions (α = 0.76-0.82) and substantial agreement on most readability dimensions (α = 0.65-0.70). Even for stylistic dimensions (Q4-Q6), students maintained moderate agreement (α = 0.45-0.56). This consistently high reliability across all dimensions suggests that student evaluators may have followed the evaluation criteria more systematically and with less individual interpretation than other groups.

Table 15: Krippendorff's Alpha Within-Group Reliability

Quality Dimension	Expert	Public	Student
Q1 (Coverage/Completeness)	0.89	0.26	0.82
Q2 (Relevance/Focus)	0.90	0.20	0.82
Q3 (Accuracy/Fidelity)	0.91	0.17	0.76
Q4 (Conciseness/Brevity)	0.18	0.13	0.45
Q5 (No Repetition/Redundancy)	0.07	0.04	0.56
Q6 (Rephrasing/Rewriting)	0.03	-0.01	0.55
Q7 (Readability/Ease of reading)	0.42	0.09	0.67
Q8 (Fluency/Smoothness)	0.57	0.16	0.70
Q9 (Logical Flow/Coherence)	0.36	0.11	0.65
Overall	**0.48**	**0.13**	**0.66**

Agreement Scale:
$\alpha < 0.00$: Poor
$0.00 \leq \alpha < 0.20$: Slight
$0.20 \leq \alpha < 0.40$: Fair
$0.40 \leq \alpha < 0.60$: Moderate
$0.60 \leq \alpha < 0.80$: Substantial
$0.80 \leq \alpha \leq 1.00$: Almost Perfect

Public evaluators showed the lowest overall consistency ($\alpha = 0.13$), with only fair agreement on information quality dimensions (α = 0.17-0.26) and slight or poor agreement on all other dimensions. This indicates substantial variability in how public evaluators interpreted and applied the evaluation criteria, raising concerns about the reliability of this group's assessments in isolation.

The stark contrast in consistency between evaluator groups raises important methodological considerations. The high agreement among expert evaluators on information quality dimensions confirms the robustness of these metrics and suggests they may be the most reliable indicators of summary quality. Conversely, the lower agreement on stylistic dimensions across all groups indicates these aspects may be inherently more subjective or require more precise evaluation criteria.

L.2 Between-Group Consistency

Table 16 presents the between-group consistency results. Overall, there was fair agreement (α = 0.26) across the three evaluator groups, though this varied substantially by dimension:

Information quality dimensions (Q1-Q3) showed the highest between-group consistency, with substantial agreement for coverage ($\alpha = 0.66$) and relevance ($\alpha = 0.68$), and moderate agreement for accuracy ($\alpha = 0.51$). This suggests that despite differences in expertise and background, evalua-

tors across all groups showed relative consensus in assessing these fundamental aspects of summary quality.

Table 16: Krippendorff's Alpha Between-Group Reliability

Quality Dimension	Krippendorff's Alpha
Q1 (Coverage/Completeness)	0.66
Q2 (Relevance/Focus)	0.68
Q3 (Accuracy/Fidelity)	0.51
Q4 (Conciseness/Brevity)	-0.04
Q5 (No Repetition/Redundancy)	0.00
Q6 (Rephrasing/Rewriting)	-0.12
Q7 (Readability/Ease of reading)	0.18
Q8 (Fluency/Smoothness)	0.34
Q9 (Logical Flow/Coherence)	0.12
Overall	**0.26**

Agreement Scale:
$\alpha < 0.00$: Poor
$0.00 \leq \alpha < 0.20$: Slight
$0.20 \leq \alpha < 0.40$: Fair
$0.40 \leq \alpha < 0.60$: Moderate
$0.60 \leq \alpha < 0.80$: Substantial
$0.80 \leq \alpha \leq 1.00$: Almost Perfect

In contrast, stylistic dimensions (Q4-Q6) demonstrated poor agreement across groups, with alpha values near or below zero (α = -0.12 to 0.00). This striking lack of consensus indicates that different evaluator groups applied substantially different criteria or interpretations when assessing these dimensions, making these metrics less reliable for cross-group comparisons.

Readability dimensions (Q7-Q9) showed slight to fair agreement between groups (α = 0.12-0.34), indicating moderate variability in how these aspects were assessed across different evaluator perspectives.

L.3 Summary Type Effects on Consistency

Table 17 and 18 break down consistency by summary type for expert and student evaluators, respectively. These results reveal how the source of the summary influenced evaluation reliability:

Expert evaluators showed higher consistency when evaluating CoI summaries for information coverage (α = 0.53) compared to human (α = 0.30) or CoD summaries (α = 0.00). For readability dimensions, however, they showed higher agreement when evaluating CoD summaries (α = 0.44-0.45 for Q7-Q9). This suggests that different summary types elicited varying levels of consensus among experts, with machine-generated summaries potentially containing more objectively assessable features.

Table 17: Krippendorff's Alpha by Summary Type for Expert

Quality Dimension	Human	CoI	CoD
Q1 (Coverage/Completeness)	0.30	0.53	0.00
Q2 (Relevance/Focus)	0.07	0.03	0.00
Q3 (Accuracy/Fidelity)	0.10	0.10	0.05
Q4 (Conciseness/Brevity)	0.08	0.25	0.17
Q5 (No Repetition/Redundancy)	0.00	0.08	0.11
Q6 (Rephrasing/Rewriting)	-0.01	-0.03	0.05
Q7 (Readability/Ease of reading)	0.16	0.39	0.45
Q8 (Fluency/Smoothness)	0.16	0.02	0.44
Q9 (Logical Flow/Coherence)	0.05	0.17	0.44

Agreement Scale:
$\alpha < 0.00$: Poor
$0.00 \leq \alpha < 0.20$: Slight
$0.20 \leq \alpha < 0.40$: Fair
$0.40 \leq \alpha < 0.60$: Moderate
$0.60 \leq \alpha < 0.80$: Substantial
$0.80 \leq \alpha \leq 1.00$: Almost Perfect

Student evaluators maintained relatively high consistency across all summary types, though with some variation. Notably, they showed the highest agreement when evaluating CoD summaries for accuracy (α = 0.80) and logical flow (α = 0.70), but highest agreement for CoI summaries on rephrasing quality (α = 0.77). This pattern suggests that certain summary types may present more consistently identifiable characteristics for specific quality dimensions.

Table 18: Krippendorff's Alpha by Summary Type for Student

Quality Dimension	Human	CoI	CoD
Q1 (Coverage/Completeness)	0.55	0.62	0.65
Q2 (Relevance/Focus)	0.59	0.58	0.64
Q3 (Accuracy/Fidelity)	0.65	0.51	0.80
Q4 (Conciseness/Brevity)	0.29	0.18	0.49
Q5 (No Repetition/Redundancy)	0.47	0.19	0.57
Q6 (Rephrasing/Rewriting)	0.44	0.77	0.42
Q7 (Readability/Ease of reading)	0.67	0.60	0.61
Q8 (Fluency/Smoothness)	0.65	0.67	0.56
Q9 (Logical Flow/Coherence)	0.58	0.36	0.70

Agreement Scale:
$\alpha < 0.00$: Poor
$0.00 \leq \alpha < 0.20$: Slight
$0.20 \leq \alpha < 0.40$: Fair
$0.40 \leq \alpha < 0.60$: Moderate
$0.60 \leq \alpha < 0.80$: Substantial
$0.80 \leq \alpha \leq 1.00$: Almost Perfect

L.4 Implications for Evaluation Methodology

The consistency analysis yields several important implications for TODS evaluation methodology:

- **Prioritizing reliable metrics:** Information quality dimensions (Q1-Q3) demonstrated the

highest within-group and between-group consistency, suggesting these metrics provide the most reliable indicators of summary quality across different evaluator perspectives. These dimensions should be given greater weight in overall quality assessments.

- **Evaluator expertise considerations:** The substantial differences in agreement levels between expert/student evaluators and public evaluators suggest that domain knowledge significantly impacts evaluation consistency. This highlights the importance of evaluator selection and training in TODS evaluation.

- **Student evaluator efficacy:** The unexpectedly high consistency among student evaluators, sometimes exceeding expert evaluators, suggests they may represent a valuable resource for reliable TODS evaluation when properly instructed.

- **Public evaluator limitations:** The very low agreement among public evaluators indicates that their assessments should be interpreted cautiously and may benefit from more structured evaluation tools or additional training.

- **Dimension-specific reliability:** The clear pattern of higher reliability for information quality dimensions and lower reliability for stylistic dimensions suggests that evaluation frameworks should acknowledge this difference, potentially applying different methodological approaches to dimensions with inherently different levels of subjectivity.

Our reliability analysis provides a nuanced understanding of how different evaluator groups assess TODS quality. By identifying which dimensions and evaluator groups yield the most consistent assessments, these findings can inform more robust evaluation methodologies for future research in this field.

CoI Summary Evaluation

We conduct a human evaluation to assess the effectiveness of our Chain of Interaction (CoI) summarization approach across two key dimensions: progressive improvement across chains and objective fulfillment and user preference for each chain's output. The evaluation was conducted with the first four authors of this paper serving as expert annotators.

Table 19: Krippendorff's Alpha by Summary Type for Public

Quality Dimension	**Human**	**CoI**	**CoD**
Q1 (Coverage/Completeness)	0.03	0.04	-0.04
Q2 (Relevance/Focus)	-0.05	-0.06	-0.06
Q3 (Accuracy/Fidelity)	0.08	-0.16	-0.02
Q4 (Conciseness/Brevity)	0.02	0.03	-0.00
Q5 (No Repetition/Redundancy)	-0.09	0.00	0.08
Q6 (Rephrasing/Rewriting)	-0.10	-0.08	0.05
Q7 (Readability/Ease of reading)	-0.10	0.05	0.15
Q8 (Fluency/Smoothness)	0.05	-0.02	0.17
Q9 (Logical Flow/Coherence)	-0.11	0.09	0.13

Agreement Scale:
$\alpha < 0.00$: Poor
$0.00 \leq \alpha < 0.20$: Slight
$0.20 \leq \alpha < 0.40$: Fair
$0.40 \leq \alpha < 0.60$: Moderate
$0.60 \leq \alpha < 0.80$: Substantial
$0.80 \leq \alpha \leq 1.00$: Almost Perfect

To comprehensively evaluate the CoI approach, we randomly selected 30 customer service task-oriented dialogues from our dataset. Each dialogue underwent the complete 8-chain CoI summarization process, generating 6 distinct summaries per dialogue. The four first authors of the paper independently evaluated each set of summaries, resulting in 840 summary assessments (7 chains × 30 dialogues × 4 expert annotators). To ensure unbiased evaluation, the summaries were presented to researchers in random order alongside their corresponding original dialogues.

As shown in Table 20, annotators evaluated the summaries across three primary dimensions. First, they assessed objective fulfillment - examining how well each chain meets its intended summarization goal. Second, they evaluated progressive improvement by comparing each chain's output to its predecessor, determining whether and how much the summary quality improved. Third, they rated the practical utility of each summary version, explicitly considering its usefulness in customer support scenarios. Additionally, annotators indicated their preference among the seven chain outputs for practical application.

This comprehensive evaluation framework enables us to assess the incremental improvements achieved through our CoI approach and the practical value delivered at each summarization stage. The results provide insights into the technical effectiveness of each chain and its real-world applicability in customer support contexts.

Questions	**Chain [1]** Objective *(Extracts details & identifies key entities)*	**Chain [2]** Objective *(Converts extracted data into structured summary)*	**Chain [3]** Objective *(Improves readability & sentence flow)*	**Chain [4]** Objective *(Enhances coherence & logical structuring)*	**Chain [5]** Objective *(Removes redundancy & refines conciseness)*	**Chain [6]** Objective *(Ensures factual accuracy & fidelity)*	**Chain [7]** Objective *(Optimizes brevity & clarity while retaining details)*
1. To what extent does this summary meet its intended objective? ① *Not at all* ② *Slightly* ③ *Moderate* ④ *Very* ⑤ *Completely*	① ② ③ ④ ⑤	① ② ③ ④ ⑤	① ② ③ ④ ⑤	① ② ③ ④ ⑤	① ② ③ ④ ⑤	① ② ③ ④ ⑤	① ② ③ ④ ⑤
2. How much does this summary improve compared to the prior chain? ① *Much worse* ② *Worse* ③ *Same* ④ *Better* ⑤ *Much better*	N/A	① ② ③ ④ ⑤	① ② ③ ④ ⑤	① ② ③ ④ ⑤	① ② ③ ④ ⑤	① ② ③ ④ ⑤	① ② ③ ④ ⑤
3. How useful is this summary for a customer support agent? ① *Not at all* ② *Slightly* ③ *Moderate* ④ *Very* ⑤ *Extremely*	① ② ③ ④ ⑤	① ② ③ ④ ⑤	① ② ③ ④ ⑤	① ② ③ ④ ⑤	① ② ③ ④ ⑤	① ② ③ ④ ⑤	① ② ③ ④ ⑤
4. Which summary would you prefer or consider most useful for a customer support agent? *(Select one)*	☐	☐	☐	☐	☐	☐	☐

Table 20: Summary Evaluation Survey

Original Dialogue

I want to find a moderately priced restaurant. I have many options available for you! Is there a certain area or cuisine that interests you? Yes I would like the restaurant to be located in the center of the attractions. There are 21 restaurants available in the centre of town. How about a specific type of cuisine? i need to know the food type and postcode and it should also have mutliple sports I am sorry I do not understand what you just said. Please repeat in a way that makes sense. get me the food type and the post code My favorite it The Copper Kettle at 4 Kings Parade City Centre CB21SJ. It serves British food. Does that interest you? Absolutely! Thank you! I also need information on the attractions that have multiple sports in town, in the same area as the restaurant please. The only multiple sports attraction is located in the east of town. Would you like more information? No, I want to be in the centre of town. What about architecture attractions? All Saints Church is an architectural attraction. It is in the Centre of town and entrance is free! Can I get their phone number and address please? Their phone number is 01223452587. Is there anything else I can assist you with? Oh, what type of architecture is it? it is architecture type. anything else? Just the address please it is located in Jesus lane Is there an entrance fee? The entrance fee is free. Anything else i can do for you today? That will be all. Thank you! Thank you and have a great day. Goodbye.

CoI Summarization Chain Analysis

Chain	Content	Description / Impact
Chain [1] Interaction Details Extraction	• **User Intent:** – Find moderately priced restaurant (central area, sports access) • **Issues/Concerns:** 1. Location requirements 2. Food specifications 3. Architecture information 4. Contact details 5. Entrance fees 6. Additional services • **Agent Response:** 1. Restaurant options provided 2. The Copper Kettle recommended 3. Sports facility location noted 4. All Saints Church information 5. Contact details shared 6. Information confirmation 7. Additional needs check • **Entities:** – The Copper Kettle (4 Kings Parade, CB21SJ) – All Saints Church – Contact: 01223452587	Identifies and extracts key entities and requests from conversation. Provides complete dataset of interaction details with initial redundancies preserved.
Chain [2] Interaction Response Summary	**Summary:** The user sought a moderately priced restaurant in central area with sports access. Referred to The Copper Kettle (CB21SJ) and All Saints Church (free entry) with contact details.	Converts raw extracted data into structured summary format while maintaining key details. Improves initial readability though some redundancy remains.
Chain [3] Iterative Refinement	**Refined:** User directed to The Copper Kettle (CB21SJ) for British cuisine near attractions. All Saints Church information and contact details (01223452587) provided.	Improves sentence flow and connections between information pieces. Enhances narrative flow while preserving important details.
Chain [4] Review and Adjust	**Adjusted:** User sought central restaurant with sports access. Directed to Copper Kettle and received All Saints Church details with contacts.	Enhances clarity while maintaining accuracy. Improves coherence and reduces fragmentation between statements.

(Continued from previous page)

Chain	Content	Description / Impact
Chain [5] Assess for Redundancy/ Repetition	**Condensed:** User sought central restaurant near sports. Received Copper Kettle location and All Saints Church information.	Removes duplicate information while maintaining core details. Eliminates redundant information for conciseness.
Chain [6] Fidelity/Hallucination Check	**Verified:** User wanted moderately priced central restaurant with sports access. Received Copper Kettle and All Saints Church details.	Ensures accuracy while simplifying content. Maintains factual correctness with optimized phrasing.
Chain [7] Enhance Brevity	**Final:** User sought restaurant near sports in central town. Directed to Copper Kettle (CB21SJ) and All Saints Church (01223452587).	Optimizes for brevity and clarity while retaining all essential information. Represents final optimization for conciseness.
Chain [8] Explainability Evaluation	• **Conciseness** (5/5): Significantly shorter while retaining key details • **Coverage** (5/5): Includes all key user requests and outcomes • **Relevance** (5/5): Contains only request-related information • **Rephrasing** (5/5): Demonstrates original wording • **Discourse Coherence** (5/5): Maintains logical flow • **Fidelity** (5/5): Accurately represents dialogue • **Readability** (5/5): Clear, accessible language • **Fluency** (5/5): Error-free grammar • **Redundancy** (5/5): No duplicate information	Provides comprehensive quality assessment across multiple dimensions. Validates the effectiveness of the summary refinement process.

Table 21: Dialogue Summarization Process and Analysis

M Case Study: Progressive Summarization Refinement Analysis

M.1 Overview

This case study examines the dialogue-to-summary chains of the refinement process, demonstrating how an initial human-agent conversation can be systematically refined through a single instance, multi-step approach. item 21 presents the complete progression of this refinement process, from raw dialogue to final evaluation.

M.2 Analysis of Progressive Improvement

M.2.1 Logical Structuring & Progressive Refinement

The summarization process demonstrates clear progression through eight distinct chains, as shown in item 21. Of particular note is the evolution from Chain 1's comprehensive extraction to Chain 7's optimized summary, illustrating effective information distillation while maintaining essential content.

M.2.2 Redundancy Reduction Analysis

Examining Chains 1 through 5 in item 21, we can observe a systematic reduction of redundant information. Chain 5 achieves significant compression while retaining all key details from the initial extraction in Chain 1.

M.2.3 Readability Enhancement

The progression from Chain 3 to Chain 7 in item 21 demonstrates substantial improvement in readability and flow, with each iteration refining the presentation while maintaining information fidelity.

M.3 Quality Assessment

Chain 8 in item 21 provides a comprehensive evaluation of the final summary across nine key dimensions, with perfect scores indicating successful optimization across all metrics.

M.4 Conclusions and Best Practices

M.4.1 Key Success Factors

Based on the progression shown in item 21:

1. **Systematic Refinement**: Eight well-defined stages ensuring methodical improvement
2. **Information Preservation**: Maintained critical details through each iteration
3. **Structure Enhancement**: Progressive improvement in readability
4. **Quality Validation**: Comprehensive final evaluation

M.4.2 Recommendations for Similar Cases

Drawing from the process demonstrated in item 21:

1. Begin with comprehensive extraction (as in Chain 1)
2. Apply iterative refinement (following Chains 2-7)
3. Maintain information fidelity throughout
4. Conclude with systematic evaluation (as in Chain 8)

Chain-of-Interaction Prompt Step 1 — Part 1

You are given a user-agent dialogue. Follow these 2 steps to fill out the chain of interaction template and create a detailed, concise summary.

Step 1: Chain of interactions has 7 chains containing clear instructions and examples. Learn from the instructions and examples to understand how to create a chain of interaction summary.

Chain [1] Interaction Details Extraction:
- Identify the user intent.
- List all issues or concerns raised by the user.
- Note the agent's responses, actions taken, or information provided.
- Extract key entities relevant to the interaction.

Example:

```
{
    "User Intent": "To find a cheap Portuguese restaurant in Cambridge.",
    "Issues/Concerns": [
        "Requested high-rated venues and European restaurants within city centre",
        "Needed address for chosen venue",
        "Preferred moderate price range for restaurant"
    ],
    "Agent Response": [
        "Provided details for 'The Funky Fun House'",
        "Shared address: 8 Mercers Row, Mercers Row Industrial Estate",
        "Recommended Galleria restaurant",
        "Helped with reservation booking",
        "Booked taxi for transportation",
        "Provided taxi driver's contact details"
    ],
    "Entities": ["Nandos", "South part of town (CB22HA)", "Thursday", "14:45", 8987889876]
}
```

Chain [2] Interaction Response Summary:
- Create a detailed yet concise summary capturing all key details from User Intent, Issues/Concerns, and Agent Response in Chain [1].

Example:

```
{
    "Summary of Interaction": "The user is looking for a cheap Portuguese restaurant in Cambridge, requesting high-rated venues and  \
    moderate prices. The agent detailed 'The Funky Fun House' (8 Mercers Row), recommended Galleria, booked a reservation, and arranged a \
     taxi, providing the driver's contact."
}
```

Chain [3] Iterative Refinement:
- Start with the initial summary.
- Add new informative entities from Chain [1] without increasing the summary length.

Example:

```
{
    "Reviewed Summary": "The user is looking for a cheap Portuguese restaurant in Cambridge (South part of town), requesting high-rated  \
    venues and moderate prices. The agent detailed 'The Funky Fun House' (8 Mercers Row), recommended Galleria, booked a reservation on  \
    Thursday, 14:45 (CB22HA), and arranged a taxi, providing the driver's contact (8987889876)."
}
```

Chain [4] Review and Adjust:
- Review the summary for conciseness, discourse coherence, coverage, rephrasing, readability, relevance, fluency, and informativeness.
- Update the summary based on these criteria to ensure high quality and accuracy without increasing its size.

Example:

```
{
    "Reviewed Summary": "The user seeks a cheap, high-rated Portuguese restaurant in South Cambridge. The agent recommended 'The Funky  \
    Fun House' (8 Mercers Row) and Galleria, booked a reservation for Thursday at 14:45 (CB22HA), and arranged a taxi with driver contact \
     (8987889876)."
}
```

Chain [5] Assess for Redundancy/Repetition:
- Remove any redundancy or repetitions.

Example:

```
{
    "Reviewed Summary": "The user seeks a cheap, high-rated Portuguese restaurant in South Cambridge. The agent recommended 'The Funky  \
    Fun House' (8 Mercers Row) and Galleria, booked a reservation for Thursday at 14:45 (CB22HA), and arranged a taxi with driver contact \
     (8987889876)."
}
```

Figure 18: Chain-of-Interaction Prompt Step 1 — Part 1 (Chains [1] to [5])

Chain-of-Interaction Prompt Step 1 — Part 2

Chain [6] Fidelity/Hallucination Check:
- Review the summary for logical, factual, contextual, and intent fidelity with the Intent, Issues/Concerns, Agent Response, and entities in Chain [1].
- Correct any hallucination inconsistencies.

Example:

```
{
    "Fidelity-Checked Summary": "The user seeks a cheap, high-rated Portuguese restaurant in South Cambridge. The agent detailed 'The Funky Fun House' (8 Mercers Row), recommended Galleria, booked a reservation for Thursday at 14:45 (CB22HA), and arranged a taxi (Nandos) with driver contact (8987889876)."
}
```

Chain [7] Enhance Brevity:
- Enhance brevity without losing any information.
- Ensure the summary is coherent and self-contained.

Example:

```
{
    "Final Summary": "The user seeks a cheap, high-rated Portuguese restaurant in South Cambridge. The agent recommended 'The Funky Fun House' (8 Mercers Row) and Galleria, booked a Thursday 14:45 reservation (CB22HA), and arranged a taxi (Nandos) with the driver's contact (8987889876)."
}
```

Chain [8] Evaluation and Explainability:
Definition of Key Criteria:
- **Conciseness** – Brevity, eliminating unnecessary details, and overall length reduction compared to the original dialogue.
- **Coverage** – Includes all vital information from the original dialogue that represents the breadth of critical information.
- **Relevance** – Focus on the most pertinent information while excluding less critical details.
- **Rephrasing** – Demonstrates understanding through paraphrasing and restructured content.
- **Discourse Coherence** – Ensures adjacent sentences are connected and the overall summary is logically structured.
- **Fidelity** – Maintains the original dialogue's meaning, context, facts, and intent.
- **Readability** – The summary is easy to understand with clear language and logical flow.
- **Fluency** – Free from grammatical, spelling, and punctuation errors.
- **Redundancy** – Avoids unnecessary repetition of information, facts, entities, and ideas.

For each criterion, evaluate the final summary (from Chain [7]) using a 5-point Likert scale and provide explicit evidence for your rating.

Example:

```
{
  "Chain [8] Explainability": {
    "Conciseness": {
      "action": "Evaluate to what degree the final summary is noticeably more succinct than the original dialogue.",
      "value": {
        "scale": "4",
        "evidence": "The summary is significantly shorter than the original dialogue while capturing the main points such as selecting venues, booking a table, and arranging a taxi."
      }
    },
    "Coverage": {
      "action": "Evaluate to what degree the final summary covers all key information from the original dialogue (including the user's requests, agent responses, and important details).",
      "value": {
        "scale": "4",
        "evidence": "The summary includes all major actions from the dialogue, including the nightclub recommendation, restaurant details, booking, and taxi arrangement, though it misses subtle conversational elements like the exchange about steakhouses."
      }
    },
    "Relevance": {
      "action": "Evaluate to what degree the final summary includes only the information related to the user's request(s).",
      "value": {
        "scale": "5",
        "evidence": "The summary includes only relevant details regarding the venues, booking, and transportation, directly related to the user's requests."
      }
    },
    "Rephrasing": {
      "action": "Evaluate to what degree the final summary uses its own words rather than copying the original dialogue.",
      "value": {
        "scale": "4",
        "evidence": "The summary effectively paraphrases the dialogue, using its own wording to capture the essence of the interactions."
      }
    },
    "Discourse Coherence": {
      "action": "Evaluate the logical flow and connectedness of information in the final summary.",
      "value": {
        "scale": "5",
        "evidence": "The summary presents information coherently, logically flowing from one completed task to the next."
      }
    },
    ...
  }
}
```

Figure 19: Chain-of-Interaction Prompt Step 1 — Part 2 (Chains [6] to [8])

Chain-of-Interaction Prompt Step 2

Step 2: This template provides a structured approach for creating and evaluating summaries of user-agent dialogues. Fill out the template below using the given user-agent dialogue: dialogue and the summary: summary. Output the completed template in JSON format with explicit keys and values.

Chain of Interaction Prompts Template for Dialogue Summarization:

```
{
    "Chain [1] Interaction Details Extraction": {
        "User Intent": "[User's inquiry, goal or purpose]",
        "Issues/Concerns": "[User's requests or input]",
        "Agent Response": "[Agent's response, action taken, or information provided]",
        "Entities": "[Key entities relevant to the interaction]"
    },
    "Chain [2] Interaction Response Summary": {
        "Summary of Interaction": "[Detailed yet concise summary capturing all key details]"
    },
    "Chain [3] Iterative Refinement": {
        "Iterative Summary": "[Add new informative entities from Chain [1] without increasing the summary length]"
    },
    "Chain [4] Review and Adjust": {
        "Reviewed Summary": "[Review the summary for conciseness, discourse coherence, coverage, rephrasing, readability, relevance,  \
    fluency, and informativeness. Update for accuracy.]"
    },
    "Chain [5] Assess for Redundancy/Repetition": {
        "Redundancy-Free Summary": "[Remove any redundancy or repetitions.]"
    },
    "Chain [6] Fidelity/Hallucination Check": {
        "Fidelity-Checked Summary": "[Review the summary for logical, factual, contextual, and intent fidelity. Correct any  \
    hallucinations.]"
    },
    "Chain [7] Enhance Brevity": {
        "Final Summary": "[Enhance brevity without losing any information, ensuring the summary is coherent and self-contained.]"
    },
    "Chain [8] Explainability": {
        "Conciseness": {
            "action": "Evaluate to what degree the final summary is noticeably more succinct than the original dialogue.",
            "value": { "scale": "", "evidence": ""
            }
        },
        "Coverage": {
            "action": "Evaluate to what degree the final summary covers all key information from the original dialogue (including the  \
    user's requests, agent responses, and important details).",
            "value": {"scale": "", "evidence": ""
            }
        },
        "Relevance": {
            "action": "Evaluate to what degree the final summary includes only the information related to the user's request(s).",
            "value": {"scale": "", "evidence": ""
            }
        },
        "Rephrasing": {
            "action": "Evaluate to what degree the final summary uses its own words rather than copying the original dialogue.",
            "value": {"scale": "", "evidence": ""
            }
        },
        "Discourse Coherence": {
            "action": "Evaluate the logical flow and connectedness of information in the final summary.",
            "value": {"scale": "", "evidence": ""
            }
        },
        "Fidelity": {
            "action": "Evaluate whether the final summary preserves the original dialogue's meaning, context, facts, and intent.",
            "value": {"scale": "", "evidence": ""
            }
        },
        "Readability": {
            "action": "Evaluate how easy the final summary is to understand (clear language and logical flow).",
            "value": { "scale": "", "evidence": ""
            }
        },
        "Fluency": {
            "action": "Evaluate the grammatical quality of the final summary (no spelling, punctuation, or grammar errors).",
            "value": {"scale": "", "evidence": ""
            }
        },
        "Redundancy": {
            "action": "Evaluate whether the final summary avoids repeating information or ideas.",
            "value": { "scale": "", "evidence": ""
            }
        }
    }
}
```

Implementation Guidelines:

- Fill out each section sequentially, following the chain structure.
- Maintain consistency and cross-reference between chains.
- Ensure all critical details from the original dialogue are preserved.
- Provide explicit Likert scale ratings with supporting evidence for Chain [8].

Figure 20: Chain-of-Interaction Prompt Step 2 (JSON Template)