

Spot the BlindSpots: Systematic Identification and Quantification of Fine-Grained LLM Biases in Contact Center Call Summarization

Kawin Mayilvaghanan, Siddhant Gupta*, Ayush Kumar

{kawin.m, siddhant.gupta, ayush}@observe.ai

Observe.AI
Bangalore, India

Abstract

Abstractive summarization is a core application in contact centers, where Large Language Models (LLMs) generate millions of summaries of call transcripts daily. Despite their apparent quality, it remains unclear whether LLMs systematically under- or over-attend to specific aspects of the transcript, potentially introducing biases in the generated summary. While prior work has examined social and positional biases, the specific forms of bias pertinent to contact center operations—which we term ‘Operational Bias’—have remained unexplored. To address this gap, we introduce **BlindSpot**, a framework built upon a taxonomy of 15 operational bias dimensions (e.g., disfluency, speaker, topic) for the identification and quantification of these biases. **BlindSpot** leverages an LLM as a zero-shot classifier to derive categorical distributions for each bias dimension in a pair of transcript and its summary. The bias is then quantified using two metrics: *Fidelity Gap*, measured as the Total Variation Distance (TVD) between distributions, and *Coverage*, defined as the percentage of source labels omitted. Using **BlindSpot**, we conduct an empirical study with 2500 real call transcripts and their summaries generated by 20 LLMs of varying scales and families (e.g., GPT, Llama, Claude). Our analysis reveals that biases are systemic and present across all evaluated models, regardless of size or family. We further report on bias mitigation via targeted prompting which measurably reduces bias across models.

1 Introduction and Related Works

Contact centers are central to business operations, serving as the primary interface for customer support. Their capacity to deliver superior customer service is crucial for maintaining satisfaction, cultivating loyalty, and ultimately ensuring business

*Work done during internship at Observe.AI.

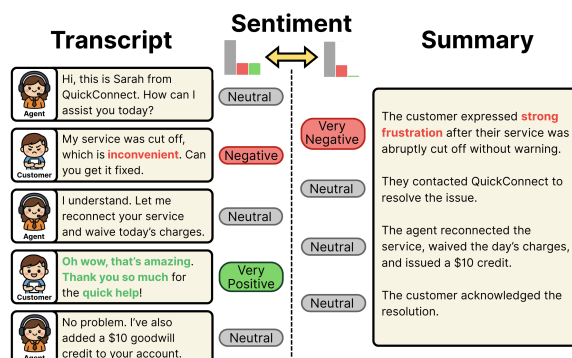


Figure 1: A call transcript (left) with mixed sentiment is contrasted with its summary (right). Although the summary is factually correct and complete, it **amplifies** the customer’s negative sentiment and **neutralizes** their positive sentiment towards resolution. This sentiment bias, invisible through contemporary summary evaluation metrics, underscores the importance of bias evaluation.

success across various industries. Within this context, abstractive call summarization (Yuan and Yu, 2019) is a critical task that enables contact center agents to effectively document interactions for regulatory compliance, contextual handoffs to other agents, and future reference. These summaries also underpin crucial downstream processes such as agent performance evaluation, business intelligence, insights discovery, and regulatory audits. For instance, supervisors use them to assess protocol adherence, while aggregated data highlights issues to inform strategic decisions.

The emergence of Large Language Models (LLMs) has facilitated the automated generation of call summaries, producing fluent and coherent summaries at scale (Sachdeva et al., 2023; Thulke et al., 2024). Evaluating the quality of LLM-generated summaries presents a multifaceted challenge. Existing metrics (Fabbri et al., 2021; Gao and Wan, 2022) effectively capture general qualities like factual correctness, relevance, and coherence, often

relying on human judgments via Likert-scale annotations. Complementing these assessments are automated reference-based metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020), which provide quantitative measures of text similarity and overlap. More recently, "LLM-as-a-Judge" (Liu et al., 2023; Kim et al., 2024) has emerged, where another LLM is utilized to evaluate the quality of a summary, offering a potentially scalable and efficient alternative to human annotation.

However, these established quality metrics overlook a crucial aspect of fidelity: a summary can be factually correct and fluent, yet still be biased in how it represents the original interaction. While the field has extensively studied bias, their work has predominantly focused on two categories. First, **social and demographic biases**, focusing on attributes such as gender, race, and nationality (Nadeem et al., 2021; Li et al., 2020; Rudinger et al., 2018; Zhu et al., 2024). Numerous methods have been proposed for detecting and mitigating these biases, including fairness-focused QA assessments (Wang et al., 2023), similarity based (Zhou and Tan, 2023) and metrics like Information Density Metric (IDM) (Wang et al., 2024), Total Variation Distance (TVD) (Steen and Markert, 2024), and Fairness Gap (Olabisi and Agrawal, 2024). Second, **structural biases**, notably position bias, the tendency to favor information based on its location in the text, have been documented (Wan et al., 2024; Olabisi and Agrawal, 2024).

Despite the complexity of these metrics, a critical gap remains: they fail to address a category of distortions that, while not necessarily factual errors, can severely undermine a summary's utility in a business context. This raises crucial questions of fidelity: do summaries accurately preserve customer sentiment? Do they equitably represent all parts of the conversation, or do they overstate the efficacy of an agent's proposed solution? We term these systematic deviations as operational biases: distortions in a summary that misrepresent the context of the original interaction. Such biases carry significant downstream consequences for agent evaluation, business intelligence, and customer satisfaction. To systematically identify and quantify these biases, our work makes the following contributions:

1. **Taxonomy of Operational Bias:** We define a taxonomy of 15 bias dimensions specific to

the operational requirements of contact center summarization, grouped into five classes.

2. **The *BlindSpot* Framework:** We introduce a fully-automated framework that quantifies bias by comparing the distributional properties of source transcripts and their summaries.
3. **An Empirical Audit:** We conduct the first comprehensive benchmark of operational bias, evaluating 20 LLMs on a corpus of 2500 contact center transcripts.

Our analysis extends beyond aggregate bias scores, using the *BlindSpot* framework to provide a fine-grained view of representation. This allows us to identify specific labels that are systematically over- or under-represented by each model and reveal common failure modes. Crucially, this analysis is actionable: a targeted system prompt engineered from our findings reduced bias across nine different models, increasing average *Coverage* by up to +5.87% and measurably reducing the *Fidelity Gap*.

Ultimately, this work provides a crucial toolset for moving beyond quality metrics toward a rigorous evaluation of summary biases. By systematically identifying and quantifying these biases, we lay the groundwork for building more accountable, reliable summarization systems for practical environments.

2 Methodology

In this section, we detail our methodology for identifying and quantifying biases in summaries.

2.1 Taxonomy of Operational Bias

To evaluate operational bias, we propose a taxonomy of 15 dimensions (Table 1). The framework moves beyond simple bias identification to link specific bias dimension to tangible operational outcomes, grouping dimensions into five classes based on core functional requirements of a contact center summary.

The first three classes address the foundational integrity of the summary: its narrative structure, and participant representation. ***Content & Information Fidelity*** ensures the summary is a reliable and actionable record; for instance, *Entity Type Bias* can render a summary useless by omitting key identifiers, while *Solution Bias* corrupts business metrics like First Call Resolution. ***Conversational***

Bias Dimension	Business & KPI Impact
<i>1. Content & Information Fidelity Dimensions</i>	
Entity Type	Omitting identifiers like a case number, product ID, or callback number renders the summary less useful for follow-up actions, breaking continuity and raising Average Handling Time (AHT).
Topic	Over-representing a brief billing mention in a call mainly about technical support leads to misclassification, which corrupts Business Intelligence dashboards, and distorts Call Driver categorizations.
Solution	Misreporting a partial fix as a full resolution artificially inflates FCR and conceals product/service failures, and Agent Performance KPIs.
Information Repetition	Collapsing repetitions removes critical interaction cues: customer self-repetition signals frustration. Losing these harms Churn Prediction, Soft-Skill QA.
<i>2. Conversational Structure & Flow Dimensions</i>	
Position	Favoring the start of calls risks omitting mid-conversation resolution steps or escalations. This distorts Resolution Rate, misleads Compliance Audits, and creates incomplete Customer Journey records.
Turn Length	Short turns (e.g., “Yes, cancel my service” or “I don’t consent”) often capture legally binding intent. If summaries omit these while overweighting long monologues, they risk missing consent/refusal and creating compliance audit failures.
Temporal Sequence	Misordering events, e.g., placing a customer’s frustration after the proposed solution, breaks cause-and-effect. This undermines Root Cause Analysis, and creates unfair Agent Evaluations.
<i>3. Speaker & Role Representation Dimensions</i>	
Speaker	Over-representing agent speech while underrepresenting the customer creates a distorted picture of the interaction, masking dissatisfaction and skewing CSAT/NPS, and Voice of Customer analytics.
Agent Action	Omitting key actions such as acknowledging, escalating, or following compliance scripts provides an incomplete QA record. This undermines Agent Coaching metrics, and distorts QA scorecards.
<i>4. Linguistic & Stylistic Dimensions</i>	
Language Complexity	Over-simplifying technical explanations hides agent expertise, while failing to capture simplicity of agent’s explanation misses evidence of good communication. Thus, agent’s performance score gets impacted.
Disfluency	Removing hesitations and false starts erases signals of customer uncertainty, while cluttering summaries with all disfluencies reduces clarity. This distorts CSAT, masks Training opportunities, and misrepresents Escalation Prevention KPIs.
Politeness	Sanitizing a rude interaction or making a professional agent curt conceals actual behavior. This skews Soft-Skill QA scores, misguides Coaching Notes.
<i>5. Affective & Pragmatic Interpretation Dimensions</i>	
Sentiment	Systematic bias toward a particular sentiment (e.g., overemphasizing negativity or positivity across summaries) corrupts trend analysis in VoC and CX dashboards, inflates or deflates CSAT/NPS benchmarks, and misguides churn prediction models.
Emotion Shift	Downplaying strong frustration as mild “dissatisfaction” hides churn risk. This undermines Churn Prediction, corrupts VoC analytics and CX KPIs.
Urgency	Failing to preserve urgent requests like “I need to cancel this fraudulent transaction right now!” masks time-sensitivity, resulting in SLA breaches, financial/reputational damage, and higher churn.

Table 1: Bias dimensions and their Business & KPI impacts in contact center summarization. See Appendix A and Table 4 for detailed examples.

Structure & Flow maintains narrative integrity, as *Temporal Sequence Bias* can alter cause-and-effect interpretations, and *Position Bias* can omit

crucial mid-conversation resolution steps. Finally, **Speaker & Role Representation** ensures fair attribution, with *Speaker Bias* being critical for balanced performance evaluations.

The remaining two classes evaluate more nuanced aspects of the interaction that are vital for risk management and quality assurance. The **Linguistic & Stylistic** class addresses distortions in conversational tone; *Politeness Bias*, for example, can conceal agent behavior vital for performance evaluation, while *Disfluency Bias* can mask customer confusion. Similarly, **Affective & Pragmatic Interpretation** focuses on subtext and intent. *Sentiment Bias* captures systematic skew toward a particular sentiment (positive or negative), while *Emotion Shift Bias* captures changes in intensity or polarity within an individual interaction, which can obscure significant customer dissatisfaction and churn risks. *Urgency Bias* addresses failure to capture time-sensitive requests.

The proposed taxonomy therefore provides a structured framework that connects summarization bias to specific operational requirements. Although not exhaustive, this approach offers a crucial tool for holistically assessing a summary’s true operational value and guiding its improvement, moving beyond generic metrics. A detailed description of each dimension is provided in Appendix A.

2.2 Problem Formulation

Let $\mathcal{T} = \{T_1, \dots, T_N\}$ be a corpus of N contact-center transcripts. Each transcript T_i consists of n_i turns, where a turn is a continuous utterance from a single speaker. An LLM summarizer \mathcal{M} produces a summary S_i composed of m_i propositions—atomic units of information, typically a single claim or clause: $S_i = \mathcal{M}(T_i) = (s_{i,1}, s_{i,2}, \dots, s_{i,m_i})$.

We define 15 bias dimensions d , each associated with a discrete set of labels $\mathcal{C}_d = \{c_{d,1}, \dots, c_{d,k}\}$. For any unit u , a turn or proposition, a multi-label classifier LLM \mathcal{L}_d assigns a subset of these labels:

$$\mathcal{L}_d(u) \subseteq \mathcal{C}_d \quad \forall u \in T_i \cup S_i.$$

For each transcript T_i and dimension d , we compute the label distribution

$$P_{i,d}(c) = \frac{1}{n_i} \left| \{t_{i,j} \in T_i : c \in \mathcal{L}_d(t_{i,j})\} \right|, \quad c \in \mathcal{C}_d.$$

Likewise for the summary S_i :

$$Q_{i,d}(c) = \frac{1}{m_i} \left| \{s_{i,j} \in S_i : c \in \mathcal{L}_d(s_{i,j})\} \right|.$$

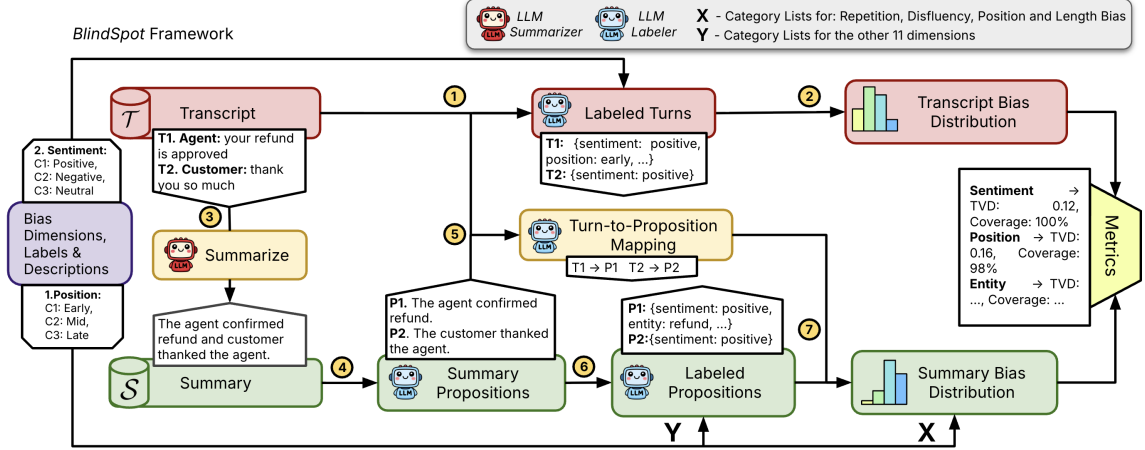


Figure 2: The *BlindSpot* framework evaluates bias in call summaries. The transcript pipeline (red) creates a reference bias distribution by labeling turns for each dimension with an *LLM Labeler*. The summary pipeline (green) creates a summary distribution by labeling propositions. Bias is quantified by comparing these distributions using *Fidelity Gap* (TVD) and *Coverage %*. White boxes provide examples. (Best viewed in color.)

We measure *fidelity gap* for pair (T_i, S_i) in dimension d via Total Variation Distance (TVD):

$$\begin{aligned} \text{FidelityGap}_{i,d} &= \text{TVD}(P_{i,d}, Q_{i,d}) \\ &= \frac{1}{2} \sum_x |P_{i,d}(x) - Q_{i,d}(x)|. \end{aligned}$$

The overall *fidelity gap* in d is $\text{FidelityGap}_d = \frac{1}{N} \sum_{i=1}^N \text{FidelityGap}_{i,d}$. To detect outright omissions, we also define *coverage* for dimension d :

$$\begin{aligned} \text{Coverage}_{i,d} &= \frac{\#\{c : P_{i,d}(c) > 0, Q_{i,d}(c) > 0\}}{\#\{c : P_{i,d}(c) > 0\}}, \\ \text{Coverage}_d\% &= \frac{1}{N} \sum_{i=1}^N \text{Coverage}_{i,d} \times 100. \end{aligned}$$

Thus, for each bias dimension d , two complementary metrics— FidelityGap_d and $\text{Coverage}_d\%$ —jointly quantify how summaries distort or omit labels relative to the original transcripts.

2.3 Framework Design and Workflow

The *BlindSpot* framework quantifies operational bias in three stages: generating the reference distribution from transcript, deriving the summary distribution, and computing bias scores from their comparison.

Transcript Pipeline: To establish a ground-truth representation, we first generate a categorical distribution P_d for each bias dimension from the source transcript. Turn-level labels are produced using

a hybrid approach. For dimensions requiring semantic interpretation (e.g., *Sentiment*, *Topic*, *Politeness*, *Entity Type*), we leverage an *LLM Labeler* \mathcal{L} to identify labels. For structural dimensions, we use direct computation: *Speaker* is extracted from metadata, while *Turn Length* and *Position* are calculated from turn and its index. Finally, derived dimensions like *Emotion Shift* and *Temporal Sequence* are inferred from the labels of *Sentiment* and *Position*. The *LLM Labeler*, GPT-4o, was validated against a human-annotated set, achieving 93.7% accuracy (see Appendix B.2 for validation details).

Summary Pipeline: Next, we generate a distribution Q_d from the summary produced by the *LLM* under evaluation. First, the model generates a complete summary from the full transcript, in single forward pass, mirroring real-world application. To enable fine-grained analysis, this summary is then decomposed into minimal semantic units, or "propositions," using an *LLM*. This step ensures a uniform and granular basis for labeling. Each proposition is then annotated using the same hybrid methodology as the transcript turns. To handle turn-dependent dimensions (e.g., *Position*, *Disfluency*), we perform a mapping step, linking each proposition back to the one or more source turns it summarizes.

Bias Quantification. Finally, we quantify bias for each dimension by calculating the *Fidelity Gap* and *Coverage* between the transcript distribution P_d and the summary distribution Q_d . For derived

Metric / Bias	llama-3.1-1b	llama-3.1-3b	llama-3.1-70b	llama-4-maverick	nova-micro	nova-11te	nova-pro	claude-3.5-haiku	claude-3.7-sonnet	claude-4-sonnet	deepseek-r1	gemin-1.5-flash-lite	gemin-1.5-flash	gpt-4o-mini	gpt-4o	gpt-4o-1-horzo	gpt-4o-1-mini	gpt-4o-1	o3-mini	o4-mini	Average
Fidelity Gap (TVD) (↓ better)																					
Turn Length	0.118	0.127	0.118	0.128	0.126	0.123	0.126	0.123	0.124	0.124	0.128	0.130	0.124	0.117	0.118	0.120	0.120	0.123	0.120	0.123	0.123
Speaker	0.144	0.156	0.144	0.149	0.161	0.152	0.153	0.142	0.134	0.134	0.150	0.171	0.164	0.135	0.143	0.150	0.141	0.143	0.149	0.147	0.148
Position	0.179	0.169	0.155	0.162	0.160	0.157	0.166	0.157	0.160	0.161	0.162	0.177	0.168	0.160	0.157	0.151	0.152	0.159	0.154	0.159	0.161
Urgency	0.140	0.143	0.138	0.146	0.146	0.144	0.142	0.148	0.147	0.152	0.159	0.151	0.150	0.139	0.139	0.142	0.145	0.144	0.134	0.142	0.144
Solution	0.182	0.158	0.145	0.153	0.146	0.141	0.155	0.152	0.165	0.159	0.156	0.167	0.172	0.148	0.141	0.134	0.140	0.141	0.136	0.142	0.152
Politeness	0.217	0.235	0.224	0.229	0.231	0.229	0.230	0.215	0.212	0.207	0.224	0.242	0.232	0.220	0.218	0.214	0.218	0.221	0.209	0.222	0.222
Language Complexity	0.199	0.203	0.191	0.204	0.201	0.198	0.204	0.195	0.196	0.197	0.204	0.221	0.222	0.185	0.189	0.189	0.192	0.192	0.188	0.202	0.199
Sentiment	0.212	0.167	0.156	0.181	0.187	0.181	0.195	0.204	0.230	0.242	0.216	0.216	0.230	0.154	0.173	0.161	0.173	0.186	0.188	0.215	0.193
Disfluency	0.180	0.188	0.180	0.186	0.180	0.183	0.187	0.185	0.184	0.189	0.191	0.191	0.194	0.177	0.178	0.177	0.181	0.184	0.180	0.186	0.184
Topic	0.239	0.237	0.227	0.235	0.235	0.235	0.239	0.240	0.242	0.243	0.250	0.254	0.258	0.223	0.232	0.229	0.230	0.230	0.221	0.232	0.237
Information Repetition	0.347	0.349	0.347	0.337	0.349	0.347	0.351	0.352	0.357	0.343	0.354	0.394	0.382	0.328	0.344	0.312	0.323	0.345	0.315	0.347	0.346
Emotion Shift	0.289	0.362	0.350	0.335	0.352	0.349	0.334	0.334	0.298	0.308	0.329	0.327	0.318	0.368	0.346	0.347	0.328	0.320	0.312	0.281	0.329
Entity Type	0.405	0.395	0.375	0.361	0.428	0.420	0.421	0.329	0.298	0.283	0.340	0.438	0.411	0.431	0.411	0.449	0.378	0.383	0.422	0.326	0.385
Agent Action	0.463	0.468	0.463	0.472	0.472	0.474	0.476	0.476	0.488	0.490	0.487	0.494	0.499	0.460	0.467	0.466	0.473	0.470	0.470	0.482	0.476
Temporal Sequence	0.759	0.717	0.688	0.710	0.745	0.730	0.749	0.722	0.715	0.702	0.701	0.761	0.766	0.744	0.749	0.706	0.672	0.681	0.709	0.701	0.729
Average	0.272	0.272	0.260	0.266	0.275	0.271	0.275	0.265	0.263	0.262	0.270	0.289	0.286	0.266	0.267	0.263	0.258	0.261	0.260	0.260	0.268
Coverage (%) (↑ better)																					
Turn Length	86.97	86.77	87.85	86.13	86.57	86.83	85.6	85.73	85.83	85.28	84.45	85.74	85.62	87.55	86.92	87.63	87.11	87.29	87.75	85.80	86.52
Speaker	99.16	97.85	98.18	97.69	98.18	98.02	98.18	98.02	98.18	97.85	97.85	97.52	97.62	98.35	98.51	97.85	98.02	97.85	98.02	98.18	98.05
Position	98.8	97.78	98.18	97.52	98.08	97.95	98.05	97.95	98.02	97.65	97.43	97.44	98.25	98.41	97.82	98.02	97.78	97.98	98.05	97.94	98.05
Urgency	92.16	92.00	93.78	91.93	92.63	92.23	92.27	92.33	92.67	91.28	91.67	91.65	93.11	93.07	93.02	92.88	93.21	92.47	93.68	92.25	92.51
Solution	80.04	85.07	86.46	84.91	85.48	86.57	84.49	83.80	82.50	82.97	84.05	83.14	81.39	85.46	86.14	87.35	86.45	85.85	86.98	85.40	84.74
Politeness	95.19	95.46	95.85	94.94	94.72	94.81	95.00	93.73	93.22	92.91	94.05	93.09	93.02	96.05	95.50	95.15	95.16	95.15	94.77	93.61	94.57
Language Complexity	82.65	83.41	84.69	83.31	83.51	83.03	82.25	82.91	83.02	82.81	83.01	81.08	80.22	84.70	84.01	84.52	83.18	83.82	84.46	83.21	83.19
Sentiment	79.54	83.86	86.18	83.36	82.64	84.46	82.04	81.57	79.92	79.04	80.66	79.17	79.54	86.69	84.32	85.51	85.65	83.58	83.39	80.40	82.58
Disfluency	67.92	68.23	70.15	68.09	69.37	68.57	67.40	69.15	68.33	68.03	67.90	65.91	64.94	69.96	69.43	70.45	69.63	69.36	70.57	67.54	68.55
Topic	75.67	79.23	81.14	79.51	78.27	78.95	76.73	76.17	74.67	72.98	75.80	73.26	71.05	81.64	79.73	80.47	79.25	79.69	79.07	75.31	77.45
Information Repetition	60.84	62.01	61.62	63.12	61.57	61.85	60.62	59.85	60.53	61.91	60.40	55.71	57.05	63.69	61.33	65.76	64.00	61.85	65.74	62.49	61.60
Entity Type	50.75	52.15	54.16	56.55	47.14	48.81	48.85	60.44	67.49	71.00	60.00	45.52	49.74	46.34	49.64	44.68	54.50	53.62	48.93	63.14	53.67
Agent Action	67.77	68.14	70.62	68.79	67.01	68.22	65.99	66.98	64.69	64.74	64.36	65.31	64.12	70.16	68.72	70.56	68.80	68.75	69.64	65.90	67.46
Average	79.83	80.92	82.22	81.22	80.40	80.79	79.81	80.66	80.70	80.65	80.22	78.04	78.10	81.68	81.21	81.59	81.77	81.31	81.61	80.87	80.68
LLM Judge Score	2.07	4.04	4.79	4.87	4.68	4.61	4.85	4.83	4.72	4.81	4.71	3.87	3.96	4.71	4.85	4.72	4.78	4.78	4.74	4.79	4.64
Compression Factor	10.98	18.83	17.23	20.75	27.44	25.29	31.2	22.86	19.05	17.29	21.87	41.11	38.66	26.37	27.73	29.19	20.84	17.68	20.13	21.25	23.79

Table 2: Main evaluation results for 20 LLMs on 15 bias dimensions in call summarization. Reported metrics include: *Fidelity Gap* (TVD) (0–1, ↓ better), *Coverage* % (0–100, ↑ better), LLM Judge Score (1–5, ↑ better), and Compression Factor. We highlight the best scores in green and worst scores in red for each row.

dimensions like *Temporal Sequence*, the reference distribution is defined as a one-hot vector representing the ideal label. Consequently, only *Fidelity Gap* is computed, as *Coverage* is not applicable.

Full workflow and implementation details are in Appendix B, and labeler prompts are in Appendix E.

3 Experimental Setup

Dataset and Models We evaluate on 2500 real contact-center transcripts¹ from 12 domains (e.g., FinTech, Healthcare), summarized by 20 LLMs under uniform prompting (details in Appendix C).

Evaluation Metrics Our evaluation pairs two metrics to quantify bias for each dimension. We use **Total Variation Distance (TVD)** to measure the distributional shift, which serves as a robust and symmetric measure of the *fidelity gap*. We also compute **Coverage %**: the percentage of source labels that appear in the summary. To contextualize these findings, we also report **LLM-Judge score** (1–5 scale; see Appendix B.3 for details) for overall summary quality and **Compression Factor** (transcript/summary tokens) to measure the degree of abstraction. Additional divergence metrics are in Appendix D.2.

¹The dataset cannot be released due to its proprietary nature.

4 Results

We evaluated 20 LLMs across 15 bias dimensions (Table 2) and highlight the key findings below.

Overall Model Performance The majority of evaluated models demonstrate similar performance, occupying a narrow range for both average TVD (0.26–0.28) and Coverage (80.40–82.22%). However, our analysis reveals three key observations. *First*, model performance is not solely determined by scale; top performers include both large models like claude-4-sonnet and llama-3.3-70b and smaller ones like gpt-4.1-mini. *Second*, gemini-2.0-flash and gemini-2.0-flash-lite are notable outliers, exhibiting higher average TVD (0.286 and 0.289). *Finally*, we observe modest but consistent improvements from intra-family scaling. In the Llama series, for example, TVD drops from 0.272 (1B) to 0.260 (70B) as Coverage increases by 2.39%. This pattern holds for other model families.

Analysis by Bias Dimension The results reveal two clear groups of bias dimensions: **Most Challenging Dimensions**: The preservation of *Temporal Sequence* presents the most significant challenge, with the highest average TVD (0.729) by a large margin. This indicates models frequently alter event chronology, obscuring cause-and-effect. Furthermore, dimensions requiring granular de-

tail show the lowest information retention. *Entity type* coverage is the lowest on average at 53.67%, meaning nearly half of all named entities are typically omitted. Models also struggle with *Repetition* (61.60% coverage) and *Agent Actions* (67.46% coverage), suggesting a difficulty in capturing the significance of repeated points and agent activities. **Most Robust Dimensions:** In contrast, models are highly effective at preserving high-level structural information. The *Speaker* and *Position* dimensions show minimal bias, with very low average TVD (0.148 and 0.161) and high coverage (98.05% and 97.94%, respectively). This suggests that while models can reliably attribute statements and identify general location in conversation, they fail to preserve fine-grained details within those structural boundaries.

Influence of Compression on Bias Bias increases with compression: Pearson correlation shows that TVD increases ($r = 0.72$) and coverage drops ($r = -0.62$) as compression increases. An exception is llama-3.2-1b, which has the lowest compression (10.98) but still a high bias.

Insufficiency of Quality Metrics Holistic metrics like LLM-Judge score weakly correlate with bias: Pearson coefficients show modest improvements in TVD ($r = -0.46$) and coverage ($r = 0.48$) as scores increase. However, high-scoring models like nova-pro (score = 4.85) can still exhibit severe *Temporal Sequence Bias* (TVD = 0.749), revealing that such LLM-Judge metrics evaluating quality of summaries overlook structural fidelity.

Over-rep	Short	Customer	Very Early	High	Diagnostic
Under-rep	Very Short, Long	Agent	Very Late, Late	Moderate	Directives
	Turn Length	Speaker	Position	Urgency	Solution
Over-rep	Elevated	Verbose Hedging	Negative	Repeat, Filled	Issue Explanation
Under-rep	Standard	Empathetic Softening	Positive	Interject	Info Gathering, Closing
	Politeness	Language Complexity	Sentiment	Disfluency	Topic
Over-rep	Agent - Self repetition	Amplified, Focused	People, Company	Give Information	Early Shift, Late Shift
Under-rep	Customer - Self repetition	Balanced	Date, Location, Monetary	Building Rapport, Ask Information,	In-order
	Information Repetition	Emotion Shift	Entity Type	Agent Activity	Temporal Sequence

Figure 3: Specific labels that are over- or under-represented consistently across all models.

Analysis of Representation Patterns Our fine-grained analysis reveals systematic biases (Figure 3). Models consistently over-represent labels like Negative sentiment and Early segments, while

under-representing labels like Building-Rapport and Directives. This indicates a model tendency to construct simplified, problem-focused narratives, sacrificing crucial interactional context.

4.1 Bias Mitigation via Targeted Prompting

To demonstrate our framework’s utility, we investigated bias mitigation by constructing a system prompt based on our analysis (see Appendix, Box F.1). This prompt was designed to directly counteract the specific operational biases, such as position bias towards early segments of transcript and the omission of rapport-building dialogue, that our framework identified. We evaluated this intervention on a representative set of nine models, including small and large variants from four major families and a reasoning model, to measure its effectiveness. The primary metrics were the change in Coverage %, Fidelity Gap (TVD), LLM-Judge score and summary conciseness (Compression Factor).

The results, shown in Table 3 and Figure 4, confirm that this targeted approach was broadly successful. All models exhibited a marked increase in Coverage, and all but one achieved a reduction in TVD. This demonstrates that the prompt effectively guided the models to produce summaries that were both more faithful to the source material and less biased. The improvements were particularly significant for more capable models. For example, claude-4-sonnet not only had the largest Coverage gain at +5.87% but also the greatest TVD reduction at -0.047. Other models like nova-pro and llama-4 also showed substantial gains, validating the efficacy of the intervention.

Beyond the overall improvements, our analysis revealed two more nuanced findings. First, we observed a distinct **scaling effect**: larger models consistently benefited more from the mitigation prompt than their smaller counterparts. For example, claude-4-sonnet’s TVD reduction was better than that of its smaller family member, claude-3.5-haiku. This suggests that more capable models are better equipped to understand and implement complex, corrective instructions. Second, this enhanced fidelity often introduced a fidelity-conciseness trade-off. To incorporate the additional required information, most models produced longer, less compressed summaries. Interestingly, even though the summaries demonstrably improved in terms of bias reduction, the

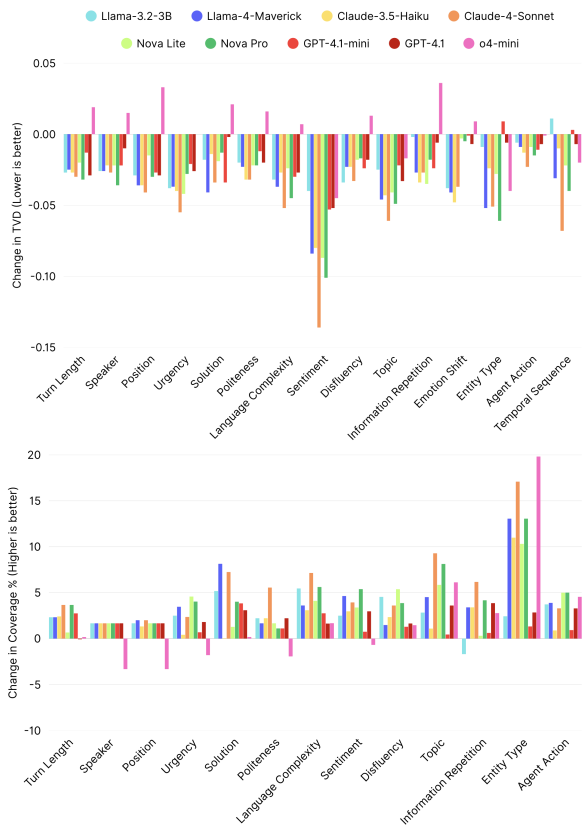


Figure 4: Change in TVD and Coverage % after applying the bias mitigation prompt. Negative change in TVD and positive change in Coverage % indicate improvement.

LLM-Judge scores declined. This suggests that automatic evaluators, while effective for coarse quality judgments, are often insensitive to nuanced fairness and coverage improvements. Examples comparing initial and improved summaries are provided in Appendix F.2.

Model	Δ Cov. (\uparrow better)	Δ TVD (\downarrow better)	Δ LLM Judge (\uparrow)	Δ Compr.
claude-3.5-haiku	+2.72	-0.032	-0.41	-6.21
claude-4-sonnet	+5.87	-0.047	-0.15	-9.97
llama-3.2-3b	+2.66	-0.022	-0.51	-12.77
llama-4-maverick	+4.34	-0.036	-0.37	-11.99
nova-lite	+3.70	-0.027	-0.24	-14.50
nova-pro	+4.97	-0.034	-0.29	-19.50
gpt-4.1-mini	+1.51	-0.019	-0.07	+0.12
gpt-4.1	+2.46	-0.019	-0.08	-0.05
o4-mini	+2.09	+0.003	-0.36	+12.82
Average Change	+3.59	-0.026	-0.28	-6.45

Table 3: Impact of the targeted mitigation prompt on model performance. The table shows absolute change (Δ) in average Coverage %, TVD, LLM Judge scores and Compression Factor. Negative TVD and positive Coverage % and LLM Judge score indicate improvement; negative Compression Factor means longer summaries.

While full mitigation is beyond the scope of this paper, this experiment validates that *BlindSpot* provides actionable feedback to tangibly improve model behavior. These insights are crucial for guiding future research toward developing more robust and reliable AI systems.

5 Conclusion

This work demonstrates that while LLMs produce fluent summaries of contact center conversations, they contain systematic operational biases. To address this, we introduce *BlindSpot*, a framework that quantifies these distortions across 15 contact center specific dimensions using divergence and coverage metrics. We show that the detailed analysis from *BlindSpot* is actionable; its findings enabled us to construct a targeted prompt that measurably reduces bias. This research provides a crucial toolset for building more transparent, trustworthy, and domain-aware summarization systems.

6 Limitations

While our framework systematically detects biases in LLM-generated summaries, it does not evaluate the harmfulness, user impact, or real-world consequences of these biases. The current metrics, Total

Variation Distance and Coverage, quantify distributional misalignments but do not capture how these biases affect user trust, business decisions, or fairness in downstream applications.

Our analysis is constrained to English-language contact center transcripts. Consequently, the framework’s applicability to multilingual contexts remains untested.

Finally, while the use of LLMs as zero-shot LLM labeler enables scalability, it introduces potential propagation of existing model biases, especially for subjective dimensions like *politeness*, into the annotations themselves, a limitation inherent in LLM-based evaluation pipelines.

7 Ethics Statement

This work focuses on identifying and quantifying biases in LLM-generated summaries of contact center transcripts. Our dataset consists of anonymized, real-world transcripts that do not contain personally identifiable information. All experiments were conducted using publicly available LLMs and datasets under appropriate usage terms.

Our goal is to improve transparency and accountability in language model behavior, not to assign blame to any specific model or provider. However, we acknowledge that exposing model biases, especially across dimensions like sentiment, speaker prominence, or topic selection—may influence deployment decisions and perceptions of fairness. We urge practitioners to interpret our findings within the methodological scope of this study and avoid overgeneralizing results beyond contact center summarization.

No human annotators were employed for labeling tasks; all labels were produced by LLMs, with validation on a small human-rated subset. There was no involvement of vulnerable populations. We believe our findings contribute positively to the responsible development and evaluation of language technologies.

References

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. [Summeval: Re-evaluating summarization evaluation](#). *Preprint*, arXiv:2007.12626.

Mingqi Gao and Xiaojun Wan. 2022. [DialSummEval: Revisiting summarization evaluation for dialogues](#). In *Proceedings of the 2022 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5693–5709, Seattle, United States. Association for Computational Linguistics.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. [Prometheus 2: An open source language model specialized in evaluating other language models](#). *Preprint*, arXiv:2405.01535.

Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020. [Leveraging graph to improve abstractive multi-document summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6232–6243, Online. Association for Computational Linguistics.

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

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](#). *Preprint*, arXiv:2303.16634.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. [StereoSet: Measuring stereotypical bias in pretrained language models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Olubusayo Olabisi and Ameeta Agrawal. 2024. [Understanding position bias effects on fairness in social multi-document summarization](#). *arXiv preprint*. Accepted at VarDial 2024; submitted May 3, 2024.

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

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. [Gender bias in coreference resolution](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Aashraya Sachdeva, Sai Nishanth Padala, Anup Pattnaik, Varun Nathan, Cijo George, Ayush Kumar, and Jithendra Vepa. 2023. Tailored real-time call summarization system for contact centers. In *Inter-speech 2023*, pages 5261–5262.

- Julius Steen and Katja Markert. 2024. [Bias in news summarization: Measures, pitfalls and corpora](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5962–5983, Bangkok, Thailand. Association for Computational Linguistics.
- David Thulke, Yingbo Gao, Rricha Jalota, Christian Dugast, and Hermann Ney. 2024. [Prompting and fine-tuning of small llms for length-controllable telephone call summarization](#). *Preprint*, arXiv:2410.18624.
- David Wan, Jesse Vig, Mohit Bansal, and Shafiq Joty. 2024. [On positional bias of faithfulness for long-form summarization](#). *Preprint*, arXiv:2410.23609.
- Chao Wang, Neo Wu, Lin Ning, Jiaying Wu, Luyang Liu, Jun Xie, Shawn O’Banion, and Bradley Green. 2024. [Usersumbench: A benchmark framework for evaluating user summarization approaches](#). *arXiv preprint*. V1 version released on August 30, 2024.
- Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023. [A causal view of entity bias in \(large\) language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1013–1025.
- Lin Yuan and Zhou Yu. 2019. [Abstractive dialog summarization with semantic scaffolds](#). *Preprint*, arXiv:1910.00825.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating text generation with BERT](#). In *International Conference on Learning Representations (ICLR)*.
- Karen Zhou and Chenhao Tan. 2023. [Entity-based evaluation of political bias in automatic summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10374–10386.
- [FirstName] Zhu and 1 others. 2024. [Quite good, but not enough: Nationality bias in large language models](#). In *Proceedings of LREC 2024*, pages 1180–1195.

A Taxonomy of Bias Dimensions

This appendix provides a comprehensive description of the 15 dimensions of bias evaluated in our study. The taxonomy is organized into five classes, each targeting a distinct aspect of summary fidelity. For each dimension, we provide its rationale, a description of its labels, and its operational significance.

A.1 Rationale for Bias Classes

The five classes provide a structured approach to understanding different facets of summary quality and potential bias.

1. Class: Content & Information Fidelity

Core Purpose: To ensure the summary is a factually accurate and actionable record of the conversation's substance. Biases in this class directly compromise the summary's primary function as a reliable source of truth.

Dimensions within this Class:

- **Entity Type:** This dimension tracks the presence of key named entities. Its operational importance is paramount; the omission of a single key identifier such as a case number, product ID, or callback number, can render a summary useless for follow-up actions and break continuity in the customer journey.
- **Topic:** This dimension ensures the summary reflects the primary purpose and subject matter of the call. A summary with topic bias might over-represent a brief mention of a billing issue in a call that was primarily about technical support, leading to mis-categorization and flawed business intelligence.
- **Solution:** This dimension is crucial for accurately tracking resolution success and agent effectiveness. Misrepresenting a partial fix as a full resolution directly inflates metrics like First Call Resolution (FCR). Furthermore, providing flawed data about which solutions work (or don't work) undermines product and service improvement efforts.
- **Information Repetition:** This dimension captures the nuanced handling of repeated statements. Repetition in a dialogue is not redundant; it is a rich signal often lost in summarization. We identify several key patterns:

- **Customer Self-Repetition:** A customer repeating their issue multiple times is a strong indicator of rising frustration, a feeling of not being heard, or confusion about the agent's response.
- **Agent Repeating Customer:** An agent paraphrasing or repeating a customer's statement is a standard technique for active listening and confirming understanding. Capturing this is vital for evaluating agent soft skills.
- **Customer Repeating Agent:** A customer repeating an agent's instructions or confirmation number indicates their attempt to verify information, which is a critical part of the interaction.
- **Agent Self-Repetition:** An agent repeating a compliance script or a key piece of information is often a matter of procedural record and must be documented.

A summary that simply collapses these repeated instances into a single mention loses this critical interactional context. Furthermore, due to ASR (Automatic Speech Recognition) errors, repeated content can sometimes appear contradictory or slightly different in the transcript. How a model handles these near-duplicates, whether it omits them, averages them, or correctly identifies the most likely intent is a key test of its robustness.

2. Class: Conversational Structure & Flow

Core Purpose: To assess the summary's narrative integrity, ensuring the chronological and causal sequence of events is preserved. The "story" of the call is often as important as its individual facts.

Dimensions within this Class:

- **Position:** This dimension addresses the well-known "lead bias," where models favor information from the beginning of a text. In a contact center context, this is operationally dangerous because crucial resolution steps, escalation decisions, and final confirmations are typically found in the middle and late stages of a conversation and are thus prone to omission.
- **Turn Length** This dimension measures how summary fidelity varies based on the length and complexity of individual turns. Conversations are composed of a mix of utterance

types: short, functional turns (e.g., “Yes,” “Okay,” a case number) and long, narrative turns (e.g., a customer explaining their entire problem history). A key challenge for summarization is to correctly weigh the importance of these different turn types. A model might over-represent short, declarative turns while failing to extract the crucial details embedded within a single long, information-dense monologue. This dimension, therefore, measures the model’s robustness in handling turns of varying complexity and its ability to avoid being biased towards either terse or verbose utterances.

- **Temporal Sequence:** This dimension measures whether the chronology of key events is preserved. A summary that misorders events, for example, by placing a customer’s expression of frustration *after* a proposed solution, fundamentally breaks the cause-and-effect narrative and can lead to unfair assessments of agent performance.

3. Class: Speaker & Role Representation

Core Purpose: To focus on the fair and accurate attribution of utterances and actions to the conversational participants. This is essential for accountability and performance evaluation.

Dimensions within this Class:

- **Speaker:** This dimension reflects the balance in prominence between the customer and agent voices. A summary with speaker bias might over-represent the agent’s turns, making them seem domineering, or under-represent them, making them appear passive. Both scenarios lead to a distorted picture of the interaction.
- **Agent Action:** This dimension tracks whether key agent behaviors are captured. QA scorecards are built around discrete agent actions like *questioning*, *informing*, *empathizing*, and *building rapport*. A summary that omits these actions provides an incomplete record for performance assessment and coaching. (Note: Customer activity is not separately modeled, as customer turns are typically reactive and lack the standardized operational roles of an agent).

4. Class: Linguistic & Stylistic Dimensions

Core Purpose: To target distortions in the *manner*

and *tone* of the conversation. These stylistic features carry significant diagnostic information about the customer experience and agent professionalism that is lost if a summary only reports literal content.

Dimensions within this Class:

- **Language Complexity:** This dimension addresses the simplification or complication of language. A summary that over-simplifies technical language may fail to document an agent’s expertise. Conversely, a summary that fails to capture the simplicity of an agent’s explanation may miss an example of excellent customer communication.
- **Disfluency:** This dimension tracks the presence of hesitations, false starts, and repetitions. While often considered “noise,” disfluencies are a rich source of information. Removing a customer’s hesitations can erase crucial evidence of their uncertainty or confusion, misrepresenting the true customer experience and an agent’s effectiveness in providing clarity.
- **Politeness:** This dimension measures the representation of social niceties. An agent’s demeanor is a core metric for QA. A summary that “sanitizes” a rude interaction or makes a professional agent seem curt eliminates vital data for performance reviews and coaching.

5. Class: Affective & Pragmatic Interpretation

Core Purpose: To address the emotional and intentional subtext of the conversation, which is often more critical for business outcomes than the raw facts.

Dimensions within this Class:

- **Sentiment:** This dimension captures the emotional valence of the interaction. Its importance for risk management cannot be overstated. A summary that minimizes genuine customer frustration by labeling it as neutral “unhappiness” or “dissatisfaction” can cause a high-priority churn risk to be overlooked by downstream systems and human reviewers.
- **Emotion Shift:** This dimension identifies more nuanced changes in emotional representation, such as amplification (making a neutral comment sound negative) or attenuation (weakening a strong emotion). These

shifts affect the perceived severity of an issue and can lead to misprioritization in customer retention workflows.

- **Urgency:** This dimension measures the representation of time-sensitivity. Failing to flag a high-urgency request—such as “I need to cancel this fraudulent transaction *right now!*”—represents a direct and immediate failure in customer service with potentially significant financial and reputational consequences.

A.2 Detailed Descriptions of Bias Dimensions and their Labels

The following table 4 provides a complete list of the 15 bias dimensions, their corresponding labels used for classification, and a brief description. The source of the annotation (LLM-annotated, computed, or derived) is also indicated. Dimensions marked with (Multiselect) allow for the assignment of multiple labels per turn or proposition.

Bias Dimension	Labels	Operational Significance
1. Content & Information Fidelity		
Entity Type	People, Identifiers, Phone Number, Email, Time Info, Date, Location Info, Products/Services, Monetary, Company/Organization, Other	Over/underrepresentation of key factual data required for action.
Topic	Greeting/Introductions, Identity Verification, Issue, Information Gathering, Product/Service Inquiry, Diagnosis/Troubleshooting, Solution, Action, Transaction, Offers/Upgrades, Sales, Resolution Confirmation, Next Steps, Closure, Empathy, Complaint, Policy Explanation, Feedback, Scheduling, Billing, Compliance, Miscellaneous	Over-focus or neglect of certain topical segments, skewing the perceived purpose of the call.
Solution	Diagnosis, Advisory, Root Cause, Directive/-Command, Preventive Measure, Escalate, Self-Help, Partial Fix, Rejected Fix, Follow-up, Set Expectation, Reassure, No Solution	Omission or distortion of resolutions, impacting FCR and product insights.
Information Repetition	No Repetition, Customer Self-Repetition, Agent Self-Repetition, Customer Repeats Agent, Agent Repeats Customer	Loss of context regarding participant frustration or confirmation loops.
2. Conversational Structure & Flow		
Position (computed)	Very Early, Early, Mid, Late, Very Late	Preference for information from specific segments of the conversation.
Turn Length (computed)	Very Short, Short, Mid, Long, Very Long	Variation in summary fidelity across dialogues of different length.
Temporal Sequence (derived)	In-order, Early-shift, Late-shift, Omitted, Added	Distortion of the chronological order of events, breaking causal chains.
3. Speaker & Role Representation		
Speaker (computed)	Agent, Customer	Unequal representation of agent vs. customer voice and contribution.
Agent Action	Request Information, Provide Information, Confirm Understanding, Build Rapport, Acknowledge, Escalate, Compliance, Other	Misrepresentation of agent actions, impacting performance evaluation.
4. Linguistic & Stylistic Dimensions		
Language Complexity (Multiselect)	Simple/Clear, Declarative, Long/Multi-Clause, Technical, Jargon, Abbreviations, Dense, Wordy/Vague, Formal, Informal, Empathic, Blunt, Slang, Passive Voice	Disproportionate simplification or complication of the original language style.
Disfluency (Multiselect)	Filled Pause, Repetition, False Start, Repair, Prolongation, Stutter, Discourse Marker, Interjection, Cutoff	Selective omission of speech imperfections that signal user confusion.
Politeness	Impolite, Standard, Minimal, Elevated	Neutralization or exaggeration of politeness, masking agent/customer demeanor.
5. Affective & Pragmatic Interpretation		
Sentiment	Very Positive, Positive, Neutral, Negative, Very Negative	Divergence in emotional tone, masking customer satisfaction or churn risk.
Emotion Shift (derived)	Balanced, Amplified, Attenuated, Inverted, Spurious	How the summary distorts, drops, or fabricates emotional nuance.
Urgency	None, Low, Moderate, High, Critical	Failure to represent time-sensitive requests, leading to service failures.

Table 4: The full taxonomy of 15 bias dimensions, organized by class. For each dimension, we provide its corresponding labels and operational significance. Labeling sources are noted in parentheses.

Code	Label	Description
very_early	Very Early	Tokens in the first 20% of the transcript.
early	Early	Tokens in the next 20% (20%–40%).
mid	Mid	Tokens in the middle segment (40%–60%).
late	Late	Tokens in the following 20% (60%–80%).
very_late	Very Late	Tokens in the final 20% of the transcript (80%–100%).

Table 5: The label set for the Position bias dimension, including descriptions and short codes used for labeling.

Code	Label	Description
agent	Agent	Utterances spoken by the service agent.
customer	Customer	Utterances spoken by the customer.

Table 6: The label set for the Speaker bias dimension, including descriptions and short codes used for labeling.

Code	Label	Description
people	People	Named individuals.
identifiers	Identifiers	IDs like account numbers.
phone_number	Phone Number	Telephone numbers.
email	Email	Email addresses.
time_info	Time Info	Time-related entities (e.g., 3 PM).
date	Date	Dates and calendar references.
location_info	Location Info	Geographical references.
product	Products/Services	Product or service mentions.
monetary	Monetary	Currency and financial references.
company_organization	Company/Organization	Business or organization names.
other	Others	Named entities not in predefined types.

Table 7: The label set for the Entity Type bias dimension, including descriptions and short codes used for labeling.

Code	Label	Description
very_pos	Very Positive	Strongly positive tone
pos	Positive	Moderately positive tone
neg	Negative	Moderately negative tone
very_neg	Very Negative	Strongly negative tone
info	Informational	Information content or presence of factual tokens (dates, names, IDs) — high priority over 'neutral'
neutral	Neutral	Does not have information and contains explicit neutral-emotion cues (e.g., “okay,” “fine,” “so-so,” “not sure”)

Table 8: The label set for the Sentiment bias dimension, including descriptions and short codes used for labeling.

Code	Label	Description
greet	Greetings/Introductions	Greetings, introductions
id_verif	ID Verification	ID or account verification
issue	Issue/Problem Statement	Customer's reason for contact
info_gath	Information Gathering	Agent probing/investigating
prod_inq	Product Inquiry	Product or service questions
diag	Diagnosis	Diagnosis or troubleshooting
soln	Solution	Proposing a solution
action	Action	Performing an action
transact	Transaction	Payments, refunds, orders
offers	Offers	Service offers or upgrades
sales	Sales	Sales, upselling, persuasion
resolve_conf	Resolution Confirmation	Confirming issue is resolved
next	Next Steps	Next steps, follow-ups
close	Closure	Farewell, call closure
empathy	Empathy	Expressing care or rapport
complaint	Complaint Handling	Handling complaints/escalation
policy	Policy Explanation	Explaining rules or terms
feedback	Feedback Request	Requesting feedback or surveys
sched	Scheduling	Appointments, scheduling
billing	Billing Issues	Billing/payment issues
compliance	Compliance	Compliance or regulations
misc	Miscellaneous	Miscellaneous

Table 9: The label set for the Topic bias dimension, including descriptions and short codes used for labeling.

Code	Label	Description
filled	Filled Pause	"uh", "um", etc.
silent	Silent Pause	Silent pauses
repeat	Repetition	Word/phrase repetition
false_start	False Start	Incomplete start
repair	Repair	Self-correction
prolong	Prolongation	Stretched sounds
stutter	Stutter	Repeated syllables
marker	Discourse Marker	Discourse filler ("like", "you know")
interject	Interjection	"oh!", "hmm"
cutoff	Cutoff	Abandoned utterance
placeholder	Placeholder	"sort of", "you know what I mean"
overlap	Overlap	Overlapping talk

Table 10: The label set for the Disfluency bias dimension, including descriptions and short codes used for labeling.

Code	Label	Description
ask_info	Request Information	Asking for details or clarification (e.g., "Could you confirm your order number?").
give_info	Provide Information	Supplying facts, context, or background not tied to a solution.
check_under	Confirm Understanding	Verifying if the other party comprehends or observes the same thing (e.g., "Do you see the change on your end?").
rapport	Build Rapport	Expressions of empathy, politeness, friendliness, or gratitude.
backchannel	Acknowledgement / Cue	Verbal cues like "Uh-huh", "Okay", or "Got it" to show active listening.
escalate	Escalate / Transfer Action	Referring or handing over to another party or department.
compliance	Compliance / Verification	Fulfilling identity, legal, or policy requirements.
idle	Passive / No-Op Response	Moments of silence or minimal interaction without progress.
other	Other Conversational Act	Any conversational act not covered above, such as small talk.

Table 11: The label set for the Agent Action bias dimension, including descriptions and short codes used for labeling.

Code	Label	Description
very_short	Very Short	Dialogues with 0-5 tokens
short	Short	Dialogues with 5-15 tokens
mid	Mid	Dialogues with 15-50 tokens
long	Long	Dialogues with 50-100 tokens
very_long	Very Long	Dialogues with more than 100 tokens

Table 12: The label set for the Turn Length bias dimension, including descriptions and short codes used for labeling.

Code	Label	Description
inorder	Correct Order	Events appear in the same order as in the original call.
early-shift	Shifted Earlier	An event appears earlier in the summary than in the original call.
late-shift	Shifted Later	An event appears later in the summary than in the original call.
omitted	Omitted Event	A key event from the original call is missing in the summary.
added	Added Event	The summary introduces an event not present in the original call.

Table 13: The label set for the Temporal Sequence bias dimension, including descriptions and short codes used for labeling.

Code	Label	Description
balanced	Emotion Preserved	Summary preserves the exact sentiment(s) and intensity(s) of the transcript.
amplified	Emotion Amplified	Summary intensifies existing sentiment(s): stronger valence or added emphasis beyond transcript.
attenuated	Emotion Attenuated	Summary weakens or omits sentiment: reduces intensity or drops emotion to neutral/informational.
inverted	Emotion Inverted	Summary flips polarity: presents the opposite emotion to what the transcript expressed.
spurious	Emotion Introduced	Summary introduces emotion where transcript was purely factual or neutral.
focused	Emotion Narrowed	Transcript had multiple distinct emotions but summary reports only one (loss of nuance).

Table 14: The label set for the Emotion Shift bias dimension, including descriptions and short codes used for labeling.

Code	Label	Description
no_rep	No Repetition	No repetition occurred.
cust_self	Customer Self-Repetition	Customer repeats their own words.
agent_self	Agent Self-Repetition	Agent repeats themselves.
cust_echo	Customer Repeats Agent	Customer echoes agent.
agent_echo	Agent Repeats Customer	Agent echoes customer.

Table 15: The label set for the Information Repetition bias dimension, including descriptions and short codes used for labeling.

Code	Label	Description
standard_clear	Clear	Clear, direct, and easily understood language.
simple_syntax	Simple Syntax	Predominantly short, declarative sentences.
complex_syntax	Complex Syntax	Long, multi-clause, or convoluted sentences.
technical_terms	Technical Terms	Specialized terms related to a specific domain.
industry_jargon	Industry Jargon	Terms/phrases specific to an industry or company.
acronyms_abbreviations	Abbreviations	Use of shortened forms of words or phrases.
info_dense	Information Dense	Highly concise; packed with specific information.
verbose_hedging	Verbose / Hedging	Wordy, uses fillers, qualifiers, or vague language.
formal_register	Formal Register	Polished, professional, and structured tone.
informal_colloquial	Informal / Colloquial	Conversational, casual, everyday language.
empathetic_softening	Empathetic	Language used to show understanding or soften news.
abrupt_blunt	Blunt	Overly direct, lacking typical softeners or politeness.
idioms_slang	Idioms / Slang	Figurative expressions or informal slang.
passive_voice_prominent	Passive Voice	Significant use of passive voice constructions.

Table 16: The label set for the Language Complexity bias dimension, including descriptions and short codes used for labeling.

Code	Label	Description
diag_expl	Diagnostic Explanation	Identifying the nature of the issue.
advisory	General Advice	Offering advice or suggestions.
root_cause	Root Cause	Explaining the underlying reason for the issue.
directive	Directive / Commands	Concrete steps or commands to take.
preventive	Preventive	Preventing future issues from occurring.
escalate	Escalation	Escalation or transfer to another team.
self_help	Self-Help	Do-it-yourself instructions.
partial	Partial Fix	Incomplete or partial resolution.
rejected	Rejected	Solution was offered but not applied.
followup	Follow-Up	Future action or check-in is promised.
expect	Set Expectations	Sets realistic timelines or expectations.
reassure	Reassurance	Provides emotional closure or comfort.
no_soln	No Solution	No resolution was provided.

Table 17: The label set for the Solution bias dimension, including descriptions and short codes used for labeling.

Code	Label	Description
none	None	No politeness cues (no please/thank you/etc.)
minimal	Minimal	One-off courtesy (“thank you”, “please”)
standard	Standard	Expected level (“please let me know”, “thanks for waiting”)
elevated	Elevated	Multiple markers + honorifics (“sir/madam”, “kindly”)
impolite	Impolite	Impoliteness cues

Table 18: The label set for the Politeness bias dimension, including descriptions and short codes used for labeling.

Code	Label	Description
none	None	No urgency language
low	Low	Mild timeframe hints (“when you can”, “at your convenience”)
moderate	Moderate	Moderate urgency (“soon”, “shortly”)
high	High	Strong urgency (“ASAP”, “urgent”)
critical	Critical	Extreme immediacy (“immediately”, “right now”, “without delay”)

Table 19: The label set for the Urgency bias dimension, including descriptions and short codes used for labeling.

B Framework Methodology and Implementation

This section provides a detailed description of the *BlindSpot* framework’s methodology, including the end-to-end data processing workflow, the validation of our LLM Labeler, and a guide to interpreting the final bias metrics.

B.1 Detailed Workflow of *BlindSpot*

The framework’s core function is to quantify bias by comparing the distributional properties of a source transcript and its generated summary. The process, illustrated in Figure 2 of the main paper, is composed of three main stages: (1) creating a reference distribution from the transcript, (2) creating a summary distribution, and (3) calculating bias metrics.

B.1.1 Stage 1: Transcript Pipeline (Generating Reference Distribution P_d)

The objective of this pipeline is to establish a reference label distribution, P_d , for each of the 15 bias dimensions.

1. Transcript Segmentation To manage long contexts and ensure consistent JSON output from the LLM Labeler, each transcript T is first partitioned into sequential, non-overlapping segments $\{S_1, \dots, S_k\}$ of 50 turns each. This segmentation mitigates potential performance degradation and out-of-spec responses when processing very long transcripts in a single pass.

2. Turn-level Annotation We employ a hybrid approach to annotate every turn in the transcript across all bias dimensions. The annotation source depends on the nature of the dimension:

- **LLM-Annotated (Semantic Dimensions):** For dimensions requiring semantic understanding, we use our LLM Labeler (\mathcal{L}) to process each segment and assign labels. These include *Sentiment*, *Topic*, *Solution*, *Information Repetition*, *Language Complexity*, *Disfluency*, *Politeness*, *Urgency*, *Entity Type*, and *Agent Action*.
- **Computed (Structural Dimensions):** For dimensions based on the transcript’s structure, labels are computed algorithmically. *Speaker* is extracted directly from conversation metadata. *Position* is calculated by normalizing a turn’s index into one of five quintiles (‘Very

Early’, ‘Early’, ‘Mid’, ‘Late’, ‘Very Late’). *Turn Length* is determined by the token count of the turn, categorized into discrete length buckets.

- **Derived (Relational Dimensions):** Two dimensions are inferred from the primary labels. *Emotion Shift* is derived by comparing the sentiment of a proposition to its source turns, and *Temporal Sequence* is derived from the mapping between chronologically ordered summary propositions and their source turn indices.

3. Reference Distribution (P_d) Generation The turn-level annotations are aggregated across the entire transcript to form a normalized categorical distribution P_d for each dimension d , which serves as our reference or “ground truth.”

B.1.2 Stage 2: Summary Pipeline (Generating Summary Distribution Q_d)

This pipeline generates a corresponding distribution, Q_d , from the LLM-generated summary.

1. Summary Generation The summarization model under evaluation, \mathcal{M} , generates an abstractive summary S from the full, unsegmented transcript T . This mirrors real-world usage where the model processes the entire conversation at once.

2. Proposition Extraction To enable fine-grained, sentence-level analysis, the generated summary S is decomposed into a set of minimal semantic units, or **propositions** $\{p_1, \dots, p_m\}$. This is performed by an LLM instructed to isolate each atomic fact or claim, creating a standardized unit of analysis.

3. Proposition Labeling and Mapping Each proposition is then labeled using the same hybrid methodology as the transcript turns. For turn-dependent dimensions like *Position* or *Urgency*, a crucial mapping step is performed where the LLM Labeler identifies the set of source turn indices that each proposition summarizes (a one-to-many mapping). The labels from these source turns are then projected onto the proposition.

4. Summary Distribution (Q_d) Generation The proposition-level labels are aggregated to form the summary’s categorical distribution Q_d for each dimension d .

B.1.3 Stage 3: Bias Quantification and Interpretation

1. Metric Calculation With both P_d and Q_d computed, we quantify bias using two complementary metrics:

- **Fidelity Gap:** We use Total Variation Distance between P_d and Q_d to measure the overall distributional distortion. A score of 0 indicates identical distributions, while higher values indicate greater divergence.
- **Coverage:** We calculate the percentage of labels present in the transcript (where $P_d(c) > 0$) that are also present in the summary (where $Q_d(c) > 0$). This directly measures the omission of information.

For derived dimensions like *Temporal Sequence*, the reference distribution P_d is a one-hot distribution representing the ideal chronological order, allowing TVD to directly measure any reordering.

2. Interpreting Results The combination of our two metrics provides a nuanced view of a summary’s faithfulness. For each dimension, we interpret the pair as follows:

- **Low Fidelity Gap & High Coverage:** A faithful summary that retains nearly all source labels and preserves their original proportions.
- **Low Fidelity Gap & Low Coverage:** A selectively faithful summary that accurately represents the distribution of the labels it includes but omits other labels entirely.
- **High Fidelity Gap & High Coverage:** A distorting summary that mentions information from all source labels but skews their relative importance, leading to misrepresentation.
- **High Fidelity Gap & Low Coverage:** The worst-case scenario; a summary that both ignores entire sets of labels and misrepresents those it chooses to include.

B.2 LLM Labeler Validation

The integrity of our framework hinges on the reliability of our LLM Labeler (\mathcal{L}), GPT-4o. To validate its performance, we conducted a rigorous human annotation study.

1. Dataset Creation: We randomly sampled 1,000 turn-proposition pairs from our dataset, ensuring coverage across all 15 bias dimensions. A human annotator trained in contact center analytics and familiar with the operational context, independently validated each label assigned by the LLM Labeler according to detailed annotation guidelines.

2. Evaluation: The LLM Labeler (\mathcal{L}) achieved an accuracy of 93.68% against human annotation. As expected, performance varied slightly by dimension, with higher accuracy on objective dimensions like *Entity Type* and slightly lower, yet still high, accuracy on more subjective dimensions like *Politeness*. This result gave us confidence in using the LLM as a scalable and reliable tool for our large-scale analysis.

B.3 LLM-Judge for Holistic Quality Assessment

To contextualize our fine-grained bias findings, we also measure the overall, holistic quality of each summary using an “LLM-as-a-Judge” approach. The goal is to establish a baseline quality score against which we can compare our bias metrics. This allows us to investigate a central question of our work: can summaries that are perceived as high-quality by a powerful LLM still harbor operational biases?

Implementation Details For each of the 50,000 transcript-summary pairs (2500 transcripts \times 20 models), we prompt a powerful arbitrator LLM (GPT-4o) to act as an impartial judge. As detailed in Box B.3, the judge is tasked with assigning an integer score from 1 (poor) to 5 (excellent) based on three explicit criteria derived from standard summarization quality dimensions:

- 1. Factual Consistency:** This criterion ensures that all claims, facts, and events mentioned in the summary are factually supported by the source transcript. It penalizes any hallucinations or contradictions.
- 2. Completeness:** This assesses whether the summary includes all critical information from the conversation without significant omissions of key events, decisions, or outcomes.

3. **Succinctness and Relevance:** This criterion, framed in the prompt as “Presence of irrelevant information,” penalizes summaries that include extraneous details, conversational filler, or other information not directly relevant to the core purpose of the interaction.

The judge is instructed to output both the numerical score and a brief textual justification for its reasoning. For our quantitative analysis, we use the numerical score, which we refer to as the **LLM Judge Score**.

Prompt for LLM Judge Score

You are provided with an input call transcript and its abstractive summary. Your task is to evaluate the quality of the summary according to the transcript.

Assign an integer score between 1 and 5 (higher the score, better the response quality).

Evaluate the response using the following criteria:

1. **Factual Consistency** - Are the facts and claims in the summary correct?
2. **Completeness** - Is all necessary information included?
3. **Presence of irrelevant information** - Does the summary stay focused on the task?

Output Format: Score: [1-5] Reason: [Feedback on prompt]

Acknowledged Limitations While scalable and effective for capturing general quality, we acknowledge the known limitations of the LLM-as-a-Judge paradigm. These include potential agreement bias (a tendency to favor summaries stylistically similar to its own training data), positional bias (overweighting information at the beginning or end of the summary), and an inability to detect subtle but operationally critical omissions that our *BlindSpot* framework is designed to find. Therefore, we use this score not as an absolute measure of truth, but as a proxy for a summary’s perceived holistic quality. The potential for this high-level score to mask the fine-grained biases we investigate is a central motivation for our work.

C Experimental Configuration

This section provides a detailed overview of the experimental configuration used in our study, including model generation parameters, dataset statistics, and the full list of evaluated models.

C.1 Generation Parameters

To ensure a fair and reproducible comparison, we employed a standardized set of generation parameters for all summarization tasks. The specific settings were chosen to elicit factual and deterministic outputs while accommodating different model types. For the majority of models, we set the temperature to 0 to minimize randomness and produce the most likely, consistent summary for a given transcript. For reasoning models, we used a temperature of 1 and set `reasoning_effort` to low. Other key parameters, such as `top_p`, `frequency_penalty`, and `presence_penalty`, were set to neutral values to avoid confounding the results and to observe the models’ inherent summarization behaviors. The maximum output length was capped at 1000 tokens, which was sufficient for all summaries in our corpus. All large language model (LLM) inferences were performed through API endpoints using a pay-as-you-go method.

Parameter	Value
Temperature (non-reasoning models)	0
Temperature (reasoning models)	1
Top-p	1.0
Max Tokens	1000
Frequency Penalty	0.0
Presence Penalty	0.0
Stop	None
Seed	None
Reasoning Effort (reasoning models)	low

Table 20: LLM generation parameters for summarization.

For the LLM Labeler, which performs the labeling tasks in our framework, we used GPT-4o with slightly different parameters to balance consistency with nuanced classification. A low temperature of 0.1 was chosen to ensure high reproducibility and determinism while allowing for minimal flexibility.

C.2 Models Evaluated

To conduct a comprehensive audit of bias, we selected a diverse set of 20 large language models. Our selection spans multiple major

Parameter	Value
LLM	GPT-4o
Temperature	0.1
Top-p	1.0
Max Tokens	None
Frequency Penalty	0.0
Presence Penalty	0.0
Stop	None
Seed	None

Table 21: LLM generation parameters for LLM Labeler.

model providers and open-source families, including Meta (Llama), Amazon (Nova), Anthropic (Claude), Google (Gemini), and OpenAI (GPT). Furthermore, we intentionally included models of varying scales within the same family (e.g., Llama-3.2 1B vs. Llama-3.3 70B; GPT-4.1-nano vs. GPT-4.1). This approach allows us to analyze the influence of both model architecture and parameter scale on the prevalence and nature of biases. For full transparency and reproducibility, the specific model identifiers used in our experiments are listed in Table 22.

Short Name	Model ID
Llama-3.2-1B	meta/llama3-2-1b-instruct-v1
Llama-3.2-3B	meta/llama3-2-3b-instruct-v1
Llama-3.3-70B	meta/llama3-3-70b-instruct-v1
Llama-4-Maverick	meta/llama4-maverick-17b-instruct-v1
Nova Micro	amazon/nova-micro-v11
Nova Lite	amazon/nova-lite-v1
Nova Pro	amazon/nova-pro-v1
Claude-3.5-Haiku	anthropic/claude-3-5-haiku-20241022-v1
Claude-3.7-Sonnet	anthropic/claude-3-7-sonnet-20250219-v1
Claude-4-Sonnet	anthropic/claude-sonnet-4-20250514-v1
Deepseek-R1	deepseek/r1-v1
Gemini-2.0-Flash	google/gemini-2.0-flash
Gemini-2.0-Flash-lite	google/gemini-2.0-flash-lite
GPT-4o-mini	openai/gpt-4o-mini-2024-07-18
GPT-4o	openai/gpt-4o-2024-08-06
GPT-4.1-nano	openai/gpt-4.1-nano-2025-04-14
GPT-4.1-mini	openai/gpt-4.1-mini-2025-04-14
GPT-4.1	openai/gpt-4.1-2025-04-14
o3-mini	openai/o3-mini-2025-01-31
o4-mini	openai/o4-mini-2025-04-16

Table 22: Identifiers of LLMs used in evaluation.

C.3 Transcript Statistics

Our evaluation was conducted on a corpus of 2,500 real, anonymized contact center transcripts from 12 distinct domains. As shown in Table 23, the conversations are substantial and highly variable in length. The average transcript contains approximately 317 turns and over 5,000 tokens, with the longest conversation extending to 548 turns and over 11,000 tokens. This significant variation in length and content provides a robust testbed for evaluating the models’ summarization capabilities across a

range of complexities, from brief, straightforward interactions to long, multi-issue dialogues. The distribution is slightly right-skewed, with the median length (290 turns) and call duration (37 minutes and 31 seconds) being lower than the mean, which is typical for such datasets. All agent–customer turn-by-turn conversations were processed through our in-house ASR system, which achieved a word error rate (WER) of 11.2%.

Statistic	num_turns	token_count	call_duration (mm:ss)
Count	549	605	09:09
Mean	317	5110	36:58
Std	128	2180	13:52
Min	55	244	10:01
25%	214	3003	28:20
50%	290	5048	37:31
75%	429	6840	44:32
Max	548	11348	92:32

Table 23: Summary statistics of number of turns and token counts across transcripts.

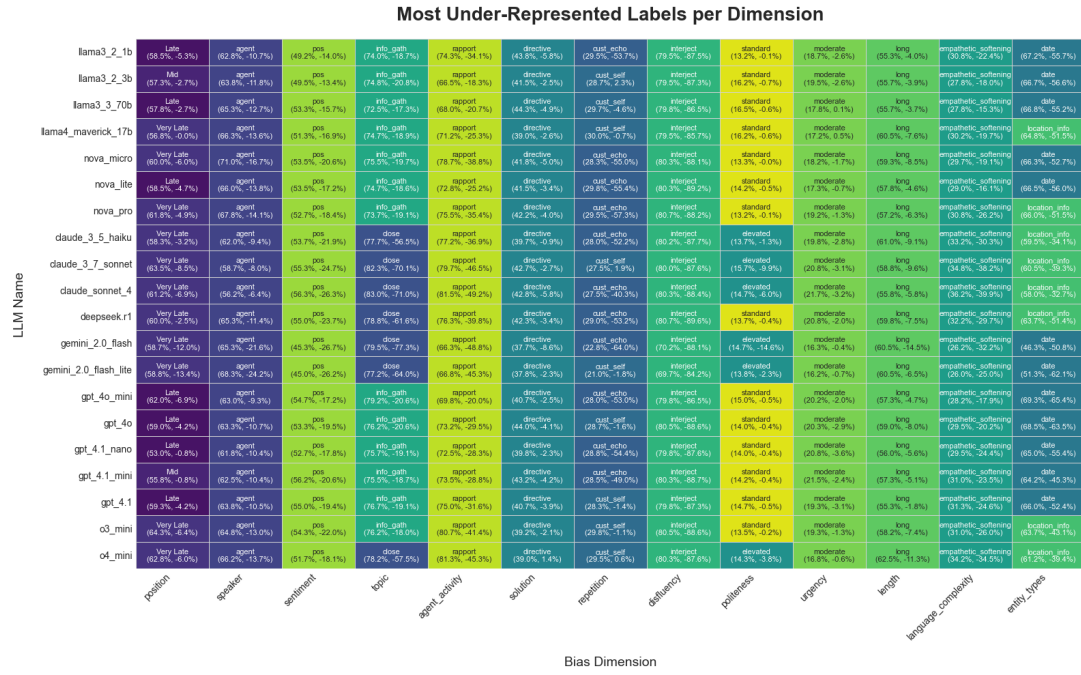
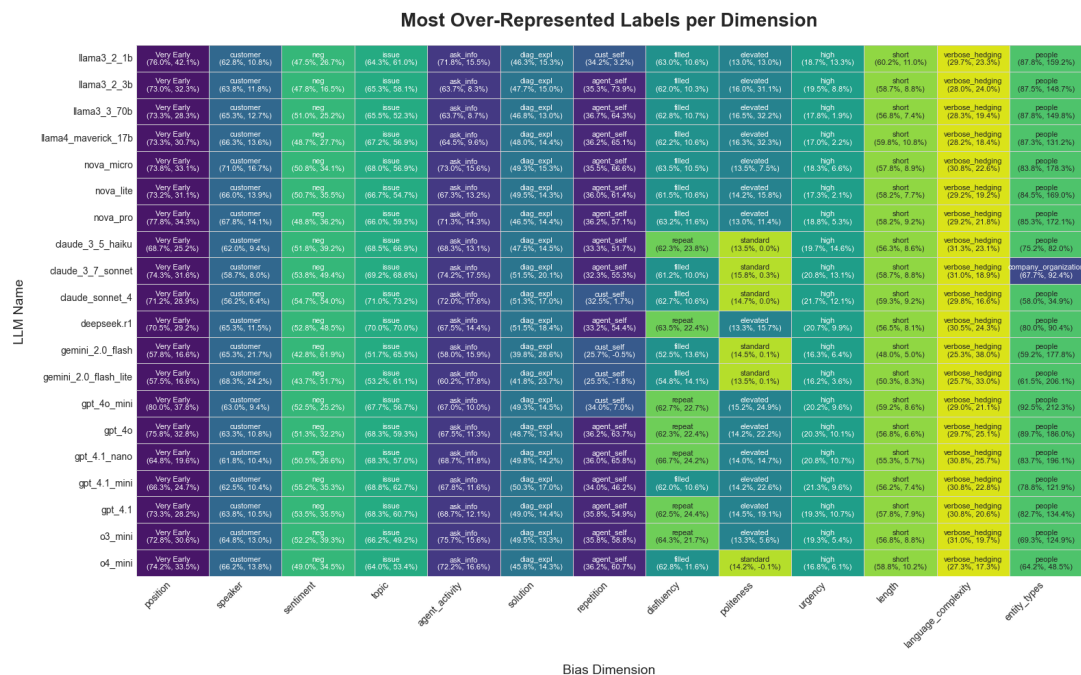


Figure 5: Systematic over- and under-representation of specific labels across all 20 LLMs. The top plot depicts over-represented labels; the bottom plot depicts under-represented labels. Each label is annotated with two percentages indicating (1) the proportion of summaries in which it is over- or under-represented and (2) the average magnitude of that deviation. The results reveal a consistent, cross-model tendency to construct simplified, problem-centric narratives while underrepresenting interactional and resolution-focused content.

E Prompts

Prompt for Summarization

System Prompt: You are a helpful assistant designed to summarize text.

User Prompt Templates:

1. Summarize the following dialog. `<dialog> {transcript} </dialog>`
2. Please provide a summary of the contact-center conversation transcript. `<transcript> {transcript} </transcript>`
3. Generate a summary of the conversation. `<conversation> {transcript} </conversation>`

Prompt for Semantic Proposition Extractor

You are a semantic analysis assistant. Your task is to decompose the given paragraph into atomic semantic propositions. Each proposition must preserve a minimal, standalone unit of meaning and reflect a single assertion or fact conveyed by the text.

Guidelines:

1. Use the original words where possible; do not paraphrase unnecessarily.
2. Resolve pronouns if possible.
3. A proposition should typically follow the (subject; predicate; object/modifier) structure.
4. Include time, place, and recipient details as separate propositions when appropriate.
5. Do not explain or justify. Just return the list of propositions.

Next, extract entities from the summary and categorize them into the following predefined types:

- **people:** Agent name, Customer name, 3rd parties
- **identifiers:** Ticket ID, Account No., Policy No.
- **phone_number:** Phone numbers
- **email:** Email addresses
- **time_info:** Time, Duration, Deadlines
- **date:** Dates
- **location_info:** Address, City, Branch
- **products_services:** Items discussed or complained about
- **monetary:** Price, Refund, Discount
- **company_organization:** Mentioned institutions
- **others:** Miscellaneous/Unclassified entities

Return a JSON object with two keys: propositions and entities.

- **propositions:** an object where keys are sequential numeric strings (e.g., "1", "2") and values are the proposition texts.

- entities: an object with the exact keys listed above, each containing a list of extracted entities (even if empty).
- "The overall JSON structure should be: "propositions": "1": "John filed a complaint.", "2": "The issue occurred yesterday at 10 AM." , "entities": "people": ["John"], "identifiers": ["WR123X62"], "phone_number": [], "email": [], "time_info": ["10 AM"], "date": ["yesterday"], "location_info": [], "products_services": ["mobile insurance"], "monetary": ["\$112"], "company_organization": ["accolade"], "others": []

User Prompt:

Process the following summary to extract semantic propositions and entities and provide the output in JSON format:\n\nSummary: \n"

Prompt for Transcript Labeling

You are a transcript analysis assistant. Your task is to annotate each turn in a conversation transcript using a fixed set of linguistic and conversational dimensions, and separately extract entities mentioned across the entire transcript.

Each turn begins like: "1: Speaker: ..." Analyze each turn independently.

Dimensions (Fixed Order with Short Labels)

Each turn must be annotated in the following order. Always include all dimensions. Empty lists are allowed where applicable.

Key	Dimension Name	Type
sent	Sentiment	single value
topic	Topic Category	single value
agent	Agent Action	single value
sol	Solution Type	list
rep	Information Repetition	single value
disf	Disfluency Patterns	list
lang	Language Complexity Patterns	list
polite	Politeness	single value
urgency	Urgency	single value

Output Format (JSON)

"map": [[1, "neutral", "greet", "ask_info", [], "no_rep", [], [], "minimal", "low"], [2, "pos", "diag", "escalate", ["diag_expl"], "cust_self", ["filled"], ["plain", "formal"], "standard", "high"]], "entity": "people": ["Alex"], "phone_number": [9512384859], "monetary": ["\$100"], ... (and other entity categories)

map: List of arrays — one for each turn. Each array must contain 10 elements: [turn_number, sent, topic, agent, sol, rep, disf, lang, polite, urgency]

entity: Dictionary of extracted entities. Entity extraction is a separate task — do not confuse with turn-level annotation.

Allowed Values and Glossary:

1. sent - Sentiment

Code	Meaning
very_pos	Strongly positive tone
pos	Moderately positive tone
neg	Moderately negative tone
very_neg	Strongly negative tone
info	Information content or presence of factual tokens (dates, names, IDs) – high priority over neutral
neutral	Does not have information and contains explicit neutral-emotion cues (e.g., “okay,” “fine,” “so-so,” “not sure”)

2. topic - Topic Category

Code	Description
greet	Greetings, introductions
id_verif	ID or account verification
issue	Customer’s reason for contact
info_gath	Agent probing/investigating
prod_inq	Product or service questions
diag	Diagnosis or troubleshooting
soln	Proposing a solution
action	Performing an action
transact	Payments, refunds, orders
offers	Service offers or upgrades
sales	Sales, upselling, persuasion
resolve_conf	Confirming issue is resolved
next	Next steps, follow-ups
close	Farewell, call closure
empathy	Expressing care or rapport
complaint	Handling complaints/escalation
policy	Explaining rules or terms
feedback	Requesting feedback or surveys
sched	Appointments, scheduling
billing	Billing/payment issues
compliance	Compliance or regulations
misc	Miscellaneous

3.agent - Agent Action:

Code	Category	Notes
ask_info	Request Information	“Could you confirm your order?”
give_info	Provide Information	Facts or explanations not tied to a fix
check_under	Confirm Understanding	“Do you see the change on your end?”
rapport	Build Rapport	Empathy, friendliness, thank-you
backchannel	Acknowledgement / Cue	“Uh-huh,” “Okay,” “Got it.”
escalate	Escalate / Transfer Action	“I’m connecting you to billing.”
compliance	Compliance / Verification	Identity, policy, legal checks
idle	Passive / No-Op Response	Silence gaps marked or minimal reply
other	Other Conversational Act	Anything else (e.g., small talk)

4.sol - Solution Type (multi-select)

Code	Description
diag_expl	Diagnostic explanation
advisory	General advice
root_cause	Explaining root cause
directive	Concrete steps or commands
preventive	Prevent future issues
escalate	Escalation or transfer
self_help	Do-it-yourself instructions
partial	Incomplete or partial fix
rejected	Offered but not applied
followup	Future action promised
expect	Sets realistic timelines
reassure	Emotional closure
no_soln	No solution given

5. rep - Repetition

Code	Description
no_rep	No repetition present
cust_self	Customer repeats self
agent_self	Agent repeats self
cust_echo	Customer echoes agent
agent_echo	Agent echoes customer

6.disf - Disfluencies (multi-select)

Code	Description
filled	“uh”, “um”, etc.
silent	Silent pauses
repeat	Word/phrase repetition
false_start	Incomplete start
repair	Self-correction
prolong	Stretched sounds
stutter	Repeated syllables
marker	Discourse filler (“like”, “you know”)
interject	“oh!”, “hmm”
cutoff	Abandoned utterance
placeholder	“sort of”, “you know what I mean”
overlap	Overlapping talk

7. lang - Language Complexity (multi-select)

Code	Description	Example
standard_clear	Clear, direct, and easily understood language.	The default if no other specific complexities are prominently featured.
simple_syntax	Predominantly short, declarative sentences.	"I can help. What is your name? The account is open."
complex_syntax	Long, multi-clause, or convoluted sentences.	"Given the information you've provided, and after checking the system, it appears that the issue, which started last Tuesday, will require a technician to resolve it."
technical_terms	Specialized terms related to a specific domain.	"Modem," "IP address," "deductible," "API endpoint."
industry_jargon	Terms/phrases specific to an industry/company.	"Tier 2 escalation," "SKU," "churn rate," "SOP."
acronyms_abbreviations	Use of shortened forms of words or phrases.	"ASAP," "ID," "ETA," "KYC."
info_dense	Highly concise; packed with specific information.	"Policy AX47 requires form B2, due COB Friday for Q3 processing."
verbose_hedging	Wordy, uses fillers, qualifiers, or vague language.	"Well, you know, it's sort of like, I guess maybe we could perhaps try to see..."
formal_register	Polished, professional, often more structured.	"We wish to inform you...", "It is imperative that..."
informal_colloquial	Conversational, casual, everyday language.	"No worries!", "Gonna check that for ya.", "Awesome!"
empathetic_softening	Language used to show understanding or soften news.	"I understand this must be frustrating...", "Unfortunately...", "I'm afraid..."
abrupt_blunt	Overly direct, lacking typical softeners/politeness.	"No. Can't do that. Next." (Extreme example)
idioms_slang	Figurative expressions or informal slang.	"Bite the bullet", "cool", "spill the beans."
passive_voice_prominent	Significant use of passive voice construction.	"The account was accessed", "A decision will be made." (When frequent)

8. polite - Politeness

Code	Description
none	No politeness cues (no please/thank you/etc.)
minimal	One-off courtesy ("thank you", "please")
standard	Expected level ("please let me know", "thanks for waiting")
elevated	Multiple markers + honorifics ("sir/madam", "kindly")
impolite	Impoliteness cues)

9. urgency - Urgency

Code	Description
none	No urgency language
low	Mild timeframe hints (“when you can”, “at your convenience”)
moderate	Moderate urgency (“soon”, “shortly”)
high	Strong urgency (“ASAP”, “urgent”)
critical	Extreme immediacy (“immediately”, “right now”, “without delay”)

Entity Extraction (Separate Task)

Extract entities from the full transcript, not turn-by-turn. Group into these categories (keys in entity block):

- people
- identifiers
- phone_number
- email
- time_info
- date
- location_info
- products_services
- monetary
- company_organization
- others

User Prompt:

Analyze the following transcript segment:\n<transcript>\n{segment_turns}</transcript>

Prompt for Turn to Proposition Mapping

Your task is to map each turn in a transcript to the summary propositions it expresses.

You will receive:

1. A set of numbered summary propositions.
2. A transcript segment containing turns, each starting with a turn number like "X: Speaker: . . .", where X is the turn number.

Your Task:

- For each turn, identify which summary propositions (by their original number) are semantically expressed in that turn.
- A proposition matches a turn if the information in the proposition is present in the turn or can be reasonably inferred from it.
- Focus only on semantic content matching, not other analysis.

Output Requirements:

- Produce a JSON object where:
 - Keys are turn numbers (e.g., "1", "2").
 - Values are lists of 0-based indices of matched summary propositions.
 - If no matches are found for a turn, do not include that turn in the output.
- A proposition can match multiple turns. If so, include its index in each relevant turn.

JSON Format Example:

```
{
  "0": [0, 2],
  "2": [1]
}
```

Example Input:

- Summary Propositions:
 0. Agent name is Sarah.
 1. The sky is blue.
 2. The grass looks dead.
- Transcript:

0: Agent: Hi, I am Sarah. Beautiful blue sky today!
 1: Customer: The grass looks dead.

Example Output:

```
{
  "0": [0, 1],
  "1": [2]
}
```

User Prompt:

Map the following dialogue turns to the summary propositions:\n<propositions>{summary_proposition_string}</propositions>\n<transcript>{segment_turns}</transcript>

Prompt for Summary Labeling

Your task is to annotate each proposition (atomic-unit of summary) using a fixed set of conversational and linguistic dimensions.
 Each proposition is about either the agent or the customer, and may express actions, emotions, or procedural events.
 ### Dimensions (Fixed Order with Short Labels)
 Each proposition must be annotated in the following order. Always include all dimensions for each proposition. Use [] for empty values in list-type fields.

Key	Dimension Name	Type
sent	Sentiment	single value
spk	Speaker	single value
topic	Topic Category	single value
agent	Agent Action	single value
sol	Solution Type(s)	list
lang	Language Complexity Patterns	list
polite	Politeness	single value
urgency	Urgency	single value

Output Format

Return a compact JSON object with:

Keys: Proposition index as a string

Values: List of 8 values in ****fixed order****: [sent, spk, topic, agent, sol, lang, polite, urgency]

Example

"0": ["very_pos", "customer", "empathy", "ask_info", [], ["simple_syntax"], minimal, high],

"1": ["neutral", "agent", "offers", "give_info", ["advisory"], ["info_dense"], standard, none],

"2": ["very_neg", "agent", "diag", "check_under", ["diag_expl"], ["standard_clear"], elevated, low]

Allowed Values and Glossary

1. sent - Sentiment

Code	Meaning
very_pos	Strongly positive tone
pos	Moderately positive tone
neg	Moderately negative tone
very_neg	Strongly negative tone
info	Information content or presence of factual tokens (dates, names, IDs) – high priority to this over neutral
neutral	Does not have information and contains explicit neutral-emotion cues

2. spk - Speaker

agent, customer, misc

3. topic - Topic Category

Code	Description
greet	Greetings, introductions
id_verif	ID or account verification
issue	Customer's reason for contact
info_gath	Agent probing/investigating
prod_inq	Product or service questions
diag	Diagnosis or troubleshooting
soln	Proposing a solution
action	Performing an action
transact	Payments, refunds, orders
offers	Service offers or upgrades
sales	Sales, upselling, persuasion
resolve_conf	Confirming issue is resolved
next	Next steps, follow-ups
close	Farewell, call closure
empathy	Expressing care or rapport
complaint	Handling complaints/escalation
policy	Explaining rules or terms
feedback	Requesting feedback or surveys
sched	Appointments, scheduling
billing	Billing/payment issues
compliance	Compliance or regulations
misc	Miscellaneous

4. agent - Agent Action

Code	Category	Notes
ask_info	Request Information	"Could you confirm your order?"
give_info	Provide Information	Facts or explanations not tied to a fix
check_under	Confirm Understanding	"Do you see the change on your end?"
rapport	Build Rapport	Empathy, friendliness, thank-you
backchannel	Acknowledgement / Cue	"Uh-huh," "Okay," "Got it."
escalate	Escalate / Transfer Action	"I'm connecting you to billing."
compliance	Compliance / Verification	Identity, policy, legal checks
idle	Passive / No-Op Response	Silence gaps marked or minimal reply
other	Other Conversational Act	Anything else (e.g., small talk)

5. sol - Solution Type (multi-select)

Code	Description
diag_expl	Diagnostic explanation
advisory	General advice
root_cause	Explaining root cause
directive	Concrete steps or commands
preventive	Prevent future issues
escalate	Escalation or transfer
self_help	Do-it-yourself instructions
partial	Incomplete or partial fix
rejected	Offered but not applied
followup	Future action promised
expect	Sets realistic timelines
reassure	Emotional closure
no_soln	No solution given

6. lang - Language Complexity (multi-select)

Code	Description
standard_clear	Clear, direct, and easily understood language.
simple_syntax	Predominantly short, declarative sentences.
complex_syntax	Long, multi-clause, or convoluted sentences.
technical_terms	Specialized terms related to a specific domain.
industry_jargon	Terms/phrases specific to an industry/company.
acronyms_abbreviations	Use of shortened forms of words or phrases.
info_dense	Highly concise; packed with specific information.
verbose_hedging	Wordy, uses fillers, qualifiers, or vague language.
formal_register	Polished, professional, often more structured.
informal_colloquial	Conversational, casual, everyday language.
empathetic_softening	Language used to show understanding or soften news.
abrupt_blunt	Overly direct, lacking typical softeners/politeness.
idioms_slang	Figurative expressions or informal slang.
passive_voice_prominent	Significant use of passive voice construction.

7. polite - Politeness

Code	Description
none	No politeness cues (no please/thank you/etc.)
minimal	One-off courtesy (“thank you”, “please”)
standard	Expected level (“please let me know”, “thanks for waiting”)
elevated	Multiple markers + honorifics (“sir/madam”, “kindly”)
impolite	Impoliteness cues

8. urgency - Urgency

Code	Description
none	No urgency language
low	Mild timeframe hints (“when you can”, “at your convenience”)
moderate	Moderate urgency (“soon”, “shortly”)
high	Strong urgency (“ASAP”, “urgent”)
critical	Extreme immediacy (“immediately”, “right now”, “without delay”)

Important Instructions

* Always include all 8 fields per proposition in the exact order: sent, spk, topic, agent, sol, lang, polite, urgency

- * For sol and lang, output a list of applicable codes or an empty list ([]) if none apply.
- * Use only the short-form codes provided above.

+

IMPORTANT: You must analyze ALL `{len(summary_propositions)}` propositions in the list. Do not skip any propositions. " Output a JSON object where keys are proposition indices (0-based, from 0 to `{len(summary_propositions)-1}`) and values are objects containing: You must include entries for indices 0 through `{len(summary_propositions)-1}`.

User Prompt: 'Please analyze the sentiment and determine the speaker for ALL `{len(summary_propositions)}` propositions below. '

'Make sure to include entries for indices 0 through `{len(summary_propositions)-1}`:\n\n`{summary_propositions}`'

F Bias Mitigation

To demonstrate that the fine-grained analysis provided by our *BlindSpot* framework is actionable, we conducted a preliminary experiment in bias mitigation. The goal was to use the specific, systemic biases identified in our main analysis to construct a targeted system prompt and then measure its impact on model behavior.

F.1 Constructing a Targeted System Prompt

Our main findings revealed consistent patterns of bias across most models, such as over-representing negative sentiment while under-representing agent rapport-building and resolution steps from the middle of the conversation. Based on these insights, we constructed a single, detailed system prompt designed to explicitly counteract these observed tendencies.

The prompt, shown in full in Box F.1, instructs the model to focus on high-fidelity, balanced summarization and provides a checklist of “Correction and Balancing Guidelines.” These guidelines directly map to the bias dimensions where we observed the most significant issues, such as *Sentiment Balance*, *Positional Coverage*, and *Topic and Activity Coverage*. By making the model explicitly aware of its potential blind spots, we hypothesized that we could steer its summarization process towards a more faithful representation of the source transcript.

Constructed System Prompt for Bias Mitigation

Your task is to summarize the following dialog with a focus on high fidelity and balance. Based on an analysis of previous outputs, apply the following corrections to ensure a more accurate and balanced summary.

Correction and Balancing Guidelines

1. Sentiment Balance:

- Ensure both positive and negative sentiments are represented if they appear in the transcript.
- **Specifically Include:** Positive sentiments expressed by the customer, especially those related to agreement or satisfaction with a solution.

2. Speaker Representation:

- Provide a balanced representation of contributions from both the customer and the agent.
- **Specifically Include:** Key agent responses, clarifying questions, and de-escalation efforts.

3. Positional Coverage:

- Draw information equitably from all parts of the conversation.
- **Specifically Include:** Details from the Mid, Late, and Very Late segments of the conversation, which often contain resolution steps and final agreements.

4. Topic and Activity Coverage:

- Broaden the scope of topics and activities included in the summary.
- **Topics to Include:** Information gathering/probing by agent, Call closure, ID verification, and Expression of empathy.
- **Agent Activities to Include:** Rapport-building, Asking for information, and Checking for understanding.

5. Solution and Repetition Types:

- **Solution Types:** Ensure representation of directive solutions (concrete, actionable steps).
- **Repetition Types:** Include all forms of significant repetition, such as:
 - customer repeating self,
 - agent repeating customer, and
 - customer repeating agent.

6. Linguistic and Structural Elements:

- **Disfluencies:** Include meaningful interjections (oh!, hmm) and incomplete starts if they indicate hesitation or a change of thought.
- **Turn Length:** Represent information from both very long and very short conversational turns if they are relevant.
- **Chronological Order:** Narrate events in the sequence they occurred in the transcript. Do not reorder them.

7. Factual and Emotional Fidelity:

- **Entity Representation:** Include a wider range of entities beyond people and organizations.
- **Specifically Include:** Dates, Locations, Product/Case IDs, Monetary values, Times, Phone numbers, and Emails.
- **Emotional Tone:** Reflect the emotional state of the speakers accurately. Avoid amplifying, attenuating, or neutralizing emotions expressed in the transcript.

Final Instruction

Produce a summary of the following dialog that strictly adheres to all the guidelines above. The final output should be a balanced, factually accurate, and structurally faithful representation of the original conversation.

F.2 Examples

Summary from nova-pro:

Original Summary

In the conversation, the customer, [CUSTOMER NAME], discusses issues with a purchased product, specifically unit number [UNIT NUMBER]. [CUSTOMER] highlights problems such as leakage, mismatched parts indicating prior damage, and concerns about the product's condition upon pickup. [AGENT NAME], the agent, acknowledges the issues and assures [CUSTOMER] that he will escalate the matter to his boss for a quicker resolution. They also touch on another unit, [UNIT NUMBER], which is ready but won't be paid for until the current issue is resolved. [AGENT] promises to call [CUSTOMER] back after discussing the matter with his boss.

Improved Summary

In the conversation between [CUSTOMER NAME] (customer) and [AGENT_NAME] (agent), [CUSTOMER] initially expressed satisfaction with [AGENT]'s assistance regarding a \$[MONEY] issue. [AGENT] confirmed the unit number [UNIT NUMBER] and inquired about its current location and the post-sale service status, which [CUSTOMER] confirmed as in his yard and already handled.

Regarding the second email, [CUSTOMER_NAME] elaborated on a product issue. [AGENT_NAME] empathized with [CUSTOMER_NAME]'s situation, especially his frustration with the service provided. [CUSTOMER_NAME] detailed his concerns about the product's condition, specifically mentioning a leak and mismatched parts indicating prior damage.

[AGENT_NAME] assured [CUSTOMER_NAME] that body issues like mismatched colors were not significant but acknowledged [CUSTOMER_NAME]'s dissatisfaction with the service experience. [AGENT_NAME] committed to scheduling a pickup for the product and addressing the unresolved issues with higher priority. He assured [CUSTOMER_NAME] that he would follow up after consulting with his boss, [MANAGER_NAME]. [CUSTOMER_NAME] agreed to wait for [AGENT_NAME]'s call, expressing gratitude for the efforts to resolve the problem. The conversation ended with mutual acknowledgment of the steps to be taken, highlighting [AGENT_NAME]'s efforts to de-escalate and resolve [CUSTOMER_NAME]'s concerns.

Summary from claude-4-sonnet:

Original Summary

Customer Issue: [CUSTOMER] called because she was locked out of her [ORG] account. Her phone automatically logged her into an old account from a previous employer ([ORG]) instead of her current employer's account ([ORG]), and she couldn't reset her password.

Key Problems Identified: - Customer has two [ORG] accounts (old and current employer) - Current account has her work phone number on file instead of personal number - App is requesting facial recognition instead of her usual PIN - Cannot access funds and needs money urgently

Account Details: - Current employer: [ORG] - Email: [EMAIL] - Personal phone: [PHONE] - Work phone on file: [PHONE] (business number, inaccessible after hours) - Debit card ending.

Resolution Process: Agent [AGENT] explained that to update the phone number and regain access, [CUSTOMER] must: 1. Provide recent pay stub and government ID photo via email 2. Wait up to 12 hours for the verification department to process 3. Cannot access account until verification is complete

Outcome: Email sent to customer with document requirements. Customer frustrated about the delay but agreed to provide requested documentation to regain account access.

Improved Summary

Customer Issue: [CUSTOMER] called [ORG] because she was locked out of her account and unable to reset her password to access her money for immediate needs like gas.

Account Verification Process: Agent Christian initially verified [CUSTOMER]'s identity using her full name ([CUSTOMER]), email ([EMAIL]), and phone number ([PHONE]). However, this information pulled up an old, cancelled account with her former employer Empress Healthcare.

Multiple Account Confusion: [CUSTOMER] explained she has two [ORG] accounts - one from a previous employer and a current active account with her new employer "Take" (started in [DATE]). She was frustrated because her phone had automatically logged her into the old account, and she couldn't access her current account.

Current Account Identification: After some confusion and [CUSTOMER] expressing frustration ("you're have to like i'm calling you for help dude"), the agent located her active account using: - Alternative email: [EMAIL] - Last four digits of debit card. - Employer: [ORG]

Core Problem Discovered: The active account had [CUSTOMER]'s work phone number ([PHONE]) on file instead of her personal number ([PHONE]). Since the work number belonged to a closed business, she couldn't receive password reset texts.

Customer Emotional State: [CUSTOMER] became increasingly panicked and frustrated throughout the call, stating "i'm panicking," "this is ridiculous," and expressing urgent need for money access. The agent apologized multiple times and reassured her: "it's okay take your time" and "i'm so sorry."

Resolution Process: Since [CUSTOMER] couldn't access the work phone number for verification, the agent explained she would need to follow a document verification process: 1. Agent would email [CUSTOMER] requesting two documents 2. [CUSTOMER] must provide a recent pay stub and government ID photo 3. Processing time: 12 hours after submission 4. Documents can be submitted as pictures

Call Conclusion: The agent sent the verification email to [EMAIL] during the call. [CUSTOMER] confirmed receipt and understood she needed to reply with photos of her pay stub and ID. The agent confirmed all information was clear and that [CUSTOMER]'s questions were answered.

Technical Note: [CUSTOMER] also mentioned the app was requesting facial recognition instead of her usual PIN, but the agent explained this was a device-level issue they couldn't directly address.

Summary from o4-mini:

Original Summary

The customer, [CUSTOMER NAME], called [ORG] because she was locked out of her current employer's account and kept being routed to an old account tied to a previous employer. After providing her name, two different email addresses, phone number, and the last four digits of her new debit card, the agent confirmed she was logged into the canceled (old) account. [CUSTOMER] wanted a simple password reset but couldn't access the reset link because the account's phone number on file belonged to her former employer. The agent explained that to change the phone number and regain access, [CUSTOMER] must email back two documents—a recent pay stub and a photo of her government ID—which the agent just sent instructions for. Once [CUSTOMER] replies with those documents, the account-update team will process her request (within about 12 hours), at which point she can reset her password, log back into her active account, and access her funds.

Improved Summary

The call begins with the agent, [AGENT NAME], greeting [CUSTOMER NAME] and requesting her full name for account verification. [CUSTOMER NAME] provides her name, email ([EMAIL]), and phone number ([PHONE NUMBER]), but then explains she has two [ORG] accounts—one linked to an old employer ([ORG]) and one to her current employer—and that she's inadvertently been logged into the wrong one.

Mid-call, the agent confirms [CUSTOMER]'s details match the [ORG] account, which [CUSTOMER] no longer uses. She asks if the agent can locate her new account by email or phone; the agent searches but finds only the old account's phone number (a business line ending in [PHONE NUMBER]). [CUSTOMER] grows frustrated ("I'm panicking. . . I don't understand your system") when she cannot reset her password by SMS or facial recognition. The agent apologizes for her difficulty and empathy is expressed, but explains that policy requires removal of the old phone number before updating to her current one.

To resolve this, the agent directs [CUSTOMER] to complete identity verification via email. He will send a secure request to [EMAIL] asking for: 1) a government-issued photo ID and 2) a recent pay stub from her new employer. Once she replies, the back-office team will update her phone number and unlock her password reset flow. [CUSTOMER] confirms receipt of the email, asks about timing, and the agent notes the update can take up to 12 hours after her submission. [CUSTOMER] accepts this plan and thanks the agent.

Throughout the call, the agent probes for the correct employer name, employee ID, and last four digits of her bank debit card to confirm the right account, practices de-escalation ("take your time," "I'm sorry you're frustrated"), and provides clear, directive next steps. [CUSTOMER] expresses both frustration at system limitations and relief at having a concrete path to regain access to her wages. The call closes with the agent asking if her questions were answered and [CUSTOMER] replying, "Yes, thank you."