Entity Exchange in the Wild: A Diagnostic Study of LLMs for Conversational Entity Extraction

Soumya Jain, Ayush Kumar
{soumya.jain, ayush}@observe.ai
Observe.AI
Bangalore, India

Abstract

Entity extraction from spoken customer–agent conversations is increasingly driving automation in contact centers. Errors in extraction can trigger incorrect system actions such as database updates, verification failures, or workflow execution. While prior work has examined transcription noise and cross-turn reasoning, it has not systematically analyzed how entity-exchange phenomena shape extraction performance. We model these phenomena along three orthogonal axes: *Initiation* (how an entity becomes relevant), *Evolution* (how commitment to the value of an entity develops across turns), and *Articulation* (how the final value of an entity is expressed in surface form). We evaluate 16 large language models on 6,387 real-world conversations spanning 12 entity types across numeric, alphanumeric, temporal, and free-text categories. Performance varies by up to 50–60% within the same LLM depending on the entity-exchange phenomena. The most severe failures occur when the entity value is revised during the interaction and the model must contrast intermediate values with a committed value. Even when no revision occurs, digit-by-digit and encoded expressions remain a consistent source of error. Error-Aware prompting improves extraction across all three axes, yielding average gains of upto 6.4% across all LLMs. Together, this work provides a structured basis for benchmarking entity extraction in real-world deployments and isolating systematic failure modes.

1 Introduction

Spoken conversations exhibit structural dynamics absent in written text. Conversation analysis shows that speakers continuously engage in self-correction, clarification, and collaborative grounding to establish shared commitments over time (Schegloff et al., 1977). In dialogue, these conversational patterns shape when an entity becomes relevant, how its value is revised or negotiated and stabilized over time, and how the final committed value is articulated. Achieving correct extraction in such settings is challenging and depends on the system's ability to accurately understand such conversational dynamics (Figure 1). In customer service, sales, and support settings, extracted entities power downstream automation such as CRM updates, order tracking, verification, and workflow execution (Bandlamudi et al., 2024). With the increasing adoption of agentic conversational systems (Patel, 2025), extracted entities are no longer a static logging tasks but dynamic inputs to autonomous actions such as API calls and state updates. In this setting, extraction errors can propagate beyond textual misunderstanding and lead to incorrect or operationally consequential outcomes.

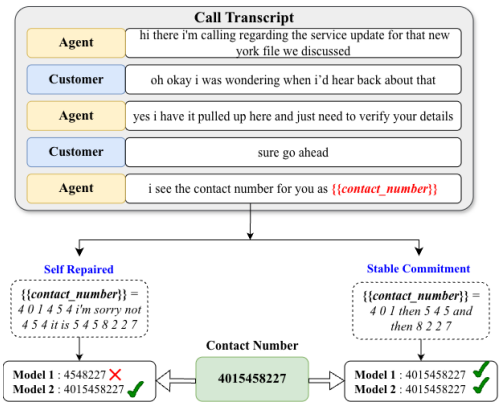

Figure 1: Illustrative example of entity exchange contrasting Stable Commitment and Self Repaired, showing the impact of self-repair on extraction accuracy.

Despite advances in automatic speech recognition (ASR) and large language models (LLMs), reliable entity extraction from spontaneous speech remains challenging. Szymański et al. (2023) show that transcription quality alone does not explain performance: Word Error Rate (WER) poorly predicts named entity recognition accuracy, and extraction failures persist even when no word-level ASR errors are present. Complementarily, Si et al. (2023) demonstrate that cross-turn slot reasoning remains difficult in spoken task-oriented dialogue.

We extend this line of research by reconceptualizing entity extraction as a function of conversational structure rather than transcription quality or cross-turn span distribution. We propose a set of entity-exchange phenomena that capture systematic variation in conversational structure and examine extraction performance across these dimensions to reveal distinct performance regimes. Specifically, we model entity exchange along three interactional dimensions: **Initiation**, describing how and when an entity becomes relevant; **Evolution**, describing how its value is negotiated and stabilized over multiple turns; and **Articulation**, describing how the final committed value is ultimately expressed. This framing elevates conversational structure to a central dimension of the entity extraction task, enabling analysis of how different conversational regimes yield systematically different outcomes. Our contributions are as follows:

- We propose a three axis taxonomy of entity exchange that characterizes entity exchange in conversations in the dimensions of initiation, evolution, and articulation.
- We conduct a large-scale zero-shot evaluation of LLM-based entity extraction, conditioning performance on interactional phenomena to quantify structure-sensitive degradation.
- We define an error taxonomy aligned with entity exchange structure and empirically map specific interactional phenomena to distinct extraction failure types.
- We derive six error aware principles from observed failure modes and integrate them into a unified prompt. We then evaluate the impact of this formulation on entity extraction performance.

2 Related Works

Entity extraction from spoken dialogue has typically been framed as a robustness problem, with industrial systems emphasizing mitigation of ASR-induced noise through domain-specific fine-tuning (Fu et al., 2022), phonetic resolution (Raghuvanshi et al., 2019), and multi-stage pipelines that apply post-hoc validation or correction using broader conversational context (Qamar et al., 2025a). Recent work has shifted toward the linguistic and interactional properties of spoken dialogue, with benchmarks like SpokenWOZ highlighting challenges such as cross-turn slot filling and reasoning over information distributed across multiple utterances (Si et al., 2023). Studies demonstrate that LLMs exhibit systematic weaknesses in conversational reasoning, particularly with turn management and backward-looking dialogue acts like self-correction (Qamar et al., 2025b), and that conversational irregularities degrade performance unless models are explicitly adapted (Mousavi et al., 2024). However, existing work has not explicitly analyzed how interactional structure affects entity extraction behavior.

Our work shifts focus away from ASR errors, transcript correction, or extraction model design, and instead studies entity exchange phenomena themselves as a systematic source of extraction difficulty. By modeling how entities become relevant, how their values evolve through interaction, and how final values are articulated, we provide a phenomenon-centric framework for analyzing and evaluating LLM-based entity extraction in spoken dialogues.

3 Study of Entity Exchange in Spoken Dialogue

In this section, we discuss the three-axis taxonomy of entity exchange phenomena in spoken contact center dialogue and formalize the entity extraction setup used in this study.

3.1 Three-Axis Taxonomy of Entity Exchange

We conceptualize entity exchange phenomena along three orthogonal axes: **Initiation**, **Evolution**, and **Articulation**. Each axis captures a distinct dimension of interactional structure observed in spoken contact center dialogue. Table 1 presents the granular categories within each axis along with illustrative examples.

Initiation. Initiation captures how an entity first becomes relevant to the state of the dialogue. An entity may enter through explicit solicitation, proactive disclosure, or implicit contextual triggers embedded in prior turns. These mechanisms differ in how clearly they mark the onset of entity discussion. When relevance is explicit, the onset of the entity episode is clearly marked; when implicit, the model must infer from discourse context that the conversation is establishing a value for the entity. Consequently, errors may arise from misidentifying the onset of the entity episode within the discourse.

Evolution. Evolution captures how commitment to an entity value unfolds across turns. Often en-

Phenomenon	Description and Example
INITIATION	
Elicited Entry	Entity enters in response to an explicit request. *Ex: "Agent: May I have your order number?" → "Customer: o nine five four"*
Volunteered Entry	Entity is introduced proactively without a prior request. *Ex: "Customer: My account ID is a zero five nine."*
Context-Triggered Entry	Entity becomes relevant implicitly through conversational or situational context. *Ex: "Agent: I see order o nine five four from your previous case."*
EVOLUTION	
Stable Commitment	A single value is introduced and remains unchanged. *Ex: "Customer: My order number is OD4481."*
Reinforced Commitment	The same value is repeated or confirmed without modification. *Ex: "Customer: Five four nine one." → "Agent: Five four nine one, got it."*
Non-Contributory Interleaving	Commitment unfolds across turns interleaved with acknowledgments or fillers that do not alter the value. *Ex: "Customer: Five five..." → "Agent: mm-hm" → "Customer: one two."*
Self-Revised Commitment	A speaker explicitly corrects or replaces their own prior value. *Ex: "Customer: Five five, sorry, five six seven."*
Other-Initiated Change	The interlocutor alters or corrects the proposed value. *Ex: "Agent: Is it AB12?" → "Customer: No, AB13."*
Implicit Commitment Drift	The value shifts without explicit repair markers. *Ex: "Customer: Pickup around ten or even better eleven should be great."*
Unresolved Commitment	The dialogue ends without a single stabilized value. *Ex: "Customer: It might be 452 or 454."*
ARTICULATION	
Uninterrupted Compositional Expression	Value appears in a single complete surface form. *Ex: "Agent: The total is five hundred."*
Fragmented Articulation	Value is distributed across multiple segments (digits, characters, or turns). *Ex: "Customer: a m i l i a, five four nine nine."*
Encoded Articulation	Value is expressed via alternative symbolic encoding. *Ex: "Agent: a as in apple z as in zebra"*
Partial / Implicit Articulation	Value is underspecified and requires contextual reconstruction. *Ex: "Agent: It should be same price as before."*
Ambiguous / Degraded Articulation	Surface form permits multiple plausible interpretations. *Ex: "Customer: Five three zero twenty four."*

Table 1: Three-axis taxonomy of entity exchange phenomena.

tities are incrementally specified, reiterated, corrected, or implicitly shifted before stabilizing. The final committed value is thus the outcome of a trajectory rather than a single mention. In interactions involving refinement or revision, intermediate values may remain locally coherent while ultimately being superseded. Correct extraction depends on identifying the stabilized endpoint of this trajectory rather than selecting an earlier candidate, making the structure of commitment development central to extraction behavior.

Articulation. Articulation captures how a committed semantic value is realized in surface form. Even when commitment is stable, realization may be canonical, fragmented across turns, symbolically encoded, partially stated, or ambiguous. Spoken dialogue frequently distributes digits, spells characters individually, or relies on contextual shorthand. As a result, the difficulty of extraction may arise not from uncertainty about the committed value, but from reconstructing that value from its linguistic encoding. The compositional structure of articulation therefore constitutes an independent source of variability in extraction performance.

3.2 Problem Formulation

We study zero-shot entity extraction from spoken customer–agent conversations. Let $\mathcal{T} = \{T_1, \ldots, T_N\}$ denote a corpus of N transcripts, where each transcript $T_i = (t_{i,1}, \ldots, t_{i,n})$ is an ordered sequence of alternate agent-customer conversation turns. Given a transcript T_i and an entity type e, an extraction model M produces a prediction $\hat{v}_i = M(T_i, e)$, where $\hat{v}_i$ is either a structured value or `NULL`. Let v_i^* denote the final committed value for entity e in T_i, or `NULL` if no value is established. The extraction is correct if $\hat{v}_i \equiv v_i^*$

For each transcript T_i, we denote the interactional labels by $I_i \in \mathcal{C}_I$, $E_i \in \mathcal{C}_E$, and $A_i \in \mathcal{C}_A$, corresponding to Initiation, Evolution, and Articulation, respectively.

Axis-Specific Accuracy. To analyze performance as a function of interactional structure, we compute accuracy conditioned on each axis label. For any axis $X \in \{I, E, A\}$ and label $c \in \mathcal{C}_X$, we define:

$$\mathrm{Acc}_X(c) = \frac{\#\{i : X_i = c \wedge \hat{v}_i \equiv v_i^*\}}{\#\{i : X_i = c\}}.$$

Error Analysis. For incorrect predictions $\hat{v}_i \not\equiv v_i^*$, we further assign an error category from a predefined error taxonomy (Ref. Table 9). This allows us to analyze which failure types are most strongly associated with each interactional axis and phenomenon.

Details of the evaluation procedure, error categorization, and annotation reliability scores (κ) are provided in Section C.1.

4 Experimental Setup

4.1 Dataset

Real-World Conversational Benchmark. We construct a benchmark from 6,387 real-world English dyadic, multi-turn customer–agent conversations sampled from real contact center conversa-

tions[1] across operational verticals including logistics, edtech, debt collections, insurance, finance, healthcare, and retail. Conversations range from 25 to 180 turns and contain the naturally occurring spoken phenomena such as interruption, self-repair, negotiation, and cross-turn distribution of entity values. All data were de-identified prior to annotation to remove or mask personally identifiable information (PII) and payment card information (PCI). The benchmark spans 12 entity types across numeric, structured alphanumeric, temporal, and free-text categories (e.g., Order ID, Policy ID, Pickup Time, Store Address), with detailed dataset statistics and annotation breakdowns provided in Appendix A.

Each transcript is augmented with (i) the final committed value of the entity, or `NULL` when no commitment is reached; (ii) a single dominant label for each axis of the taxonomy—Initiation, Evolution, and Articulation; and (iii) an error category assigned to incorrect model predictions for structured diagnostic analysis.

Phenomena Annotation. Annotation followed an iterative, guideline-driven process with annotators engaged throughout taxonomy refinement. Initial axis definitions were applied to a pilot subset, and systematic disagreement analysis was used to sharpen decision boundaries, refine prioritization rules, and disambiguate closely related categories (e.g., progressive refinement versus explicit revision; fragmented versus encoded articulation). The guidelines were iteratively revised over three rounds, with each annotator labeling 50 instances per round.

Following guideline stabilization, labeling proceeded using a structured hybrid human–LLM workflow aligned with the finalized definitions. Inter-annotator agreement between human and LLM was computed on a stratified subset of 1,900 instances. **Cohen's** κ scores were $\kappa_I = \mathbf{0.82}$ for Initiation, $\kappa_E = \mathbf{0.62}$ for Evolution, and $\kappa_A = \mathbf{0.72}$ for Articulation, indicating fair to substantial agreement in the taxonomy. Each instance receives exactly one dominant label per axis under predefined prioritization rules, ensuring categorical consistency and enabling unambiguous conditional evaluation.

[1]Due to proprietary restrictions, the real-world dataset cannot be publicly released. To support future research, a synthetic dataset is available at https://github.com/Observeai-Research/entity-exchange-in-the-wild/.

4.2 Models and Evaluation

We evaluate 16 large language models (Appendix B) spanning both proprietary and open-weight families in a zero-shot extraction setting. Experiments are conducted using three variations of a fixed zero-shot prompt template (Figure 6), which specifies the target entity and instructs the model to extract the final committed value established in the conversation, or return `NULL` if no commitment is reached. Beyond zero-shot extraction, we further examine the impact of an Error-Aware prompting strategy designed to explicitly address common failure modes. The formulation and analysis of this approach are presented in Section 5.4.

We report overall extraction accuracy, performance across LLMs and entity-exchange phenomena, and the distribution of error categories. Details of the evaluation procedure, error categorization, and annotation reliability scores (κ) are provided in Section C.1.

5 Results

5.1 Performance Across Interactional Axes

Initiation: Initiation governs how an entity becomes relevant to the dialogue state. Accuracy is highest under *Elicited Entry* (approximately 74% for frontier systems), where adjacency cues explicitly signal relevance. The absence of these cues in *Volunteered Entry* reduces performance by 5–10 points across model families. The largest degradation appears under *Context-Triggered Entry*, where relevance must be inferred from discourse state; even strong systems drop into the mid-50% range, and smaller models fall below 40%. These results indicate that boundaries explicitly established through elicitation make it easier to extract the correct entity values, whereas implicitly introduced or context-triggered entries increase the difficulty of accurate extraction.

Evolution: The Evolution axis produces the largest dispersion in accuracy. Under *Stable Commitment*, leading models reach 75–79%. *Reinforced Commitment* produces only marginal change. Once commitment becomes dynamic, performance declines sharply.

Under *Self-Revised Commitment* and *Other-Initiated Change*, accuracy drops into the 30–50% range. The most severe degradation occurs under *Implicit Commitment Drift*, where performance frequently falls to 25–40%, representing a drop ex-

Model	All	I	E	A
nova-micro	40.2	40.3±1.1	30.5±11.3	34.9±13.9
nova-lite	51.8	50.1±3.1	41.8±12.3	40.7±14.2
nova-pro	42.0	42.7±1.9	33.1±10.9	39.9±9.7
nova-2-lite	24.7	22.1±3.3	19.1±6.9	22.4±6.6
llama-3.2-3b	14.1	16.2±2.8	10.3±4.7	14.6±12.1
llama-3.3-70b	56.5	55.3±2.0	44.0±14.4	50.3±17.9
llama-4-scout	36.3	37.6±1.9	27.9±9.8	37.4±18.6
llama-4-mav	59.4	57.1±3.3	46.1±15.2	51.8±12.6
claude-son	53.6	50.6±4.4	41.6±13.7	50.6±14.8
claude-hai	41.1	42.2±3.7	33.3±9.1	43.9±17.3
gpt-5.2	**70.4**	**65.1**±7.1	**57.9**±15.9	63.0±13.9
gpt-5-nano	55.1	49.0±7.9	44.2±13.5	50.8±8.2
gpt-5-mini	68.0	62.0±8.0	56.3±14.7	**64.7**±8.3
o4-mini	65.3	57.9±10.1	53.9±15.3	61.1±9.2
gem-flash	62.6	58.4±5.7	50.6±15.1	56.3±13.3
gem-lite	50.3	49.8±1.4	39.4±13.2	42.4±10.9
Mean	49.5	47.3±4.2	39.4±12.2	45.3±12.6

Table 2: Zero-shot accuracy (%) by model. Per-axis values show $\mu \pm \sigma$ across axis-specific categories.

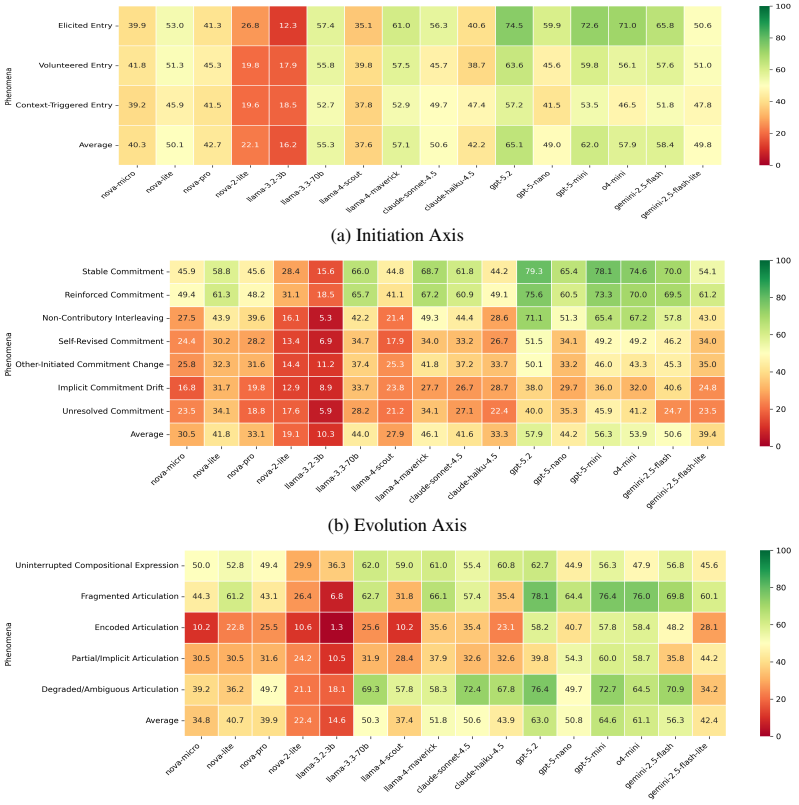

Figure 2: Zero-shot entity extraction performance across 16 LLMs. **Left**: Model accuracy with per-axis $\mu \pm \sigma$ across labels (**I**: Initiation, **E**: Evolution, **A**: Articulation). **Right**: Detailed accuracy heatmaps by phenomena class.

ceeding 50 points relative to stable cases. Under *Unresolved Commitment*, performance typically remains below 45%.

Articulation: Articulation isolates surface realization while holding commitment constant. Under *Uninterrupted Compositional Expression*, strong systems remain in the 55–63% range. However, surface variation alone induces substantial shifts.

Fragmented Articulation reveals pronounced capacity sensitivity, with gaps approaching 30 points between large and small models. *Encoded Articulation* is consistently the most difficult setting, with accuracy frequently dropping below 35% and, for some systems, below 15%.

5.2 Model-Level Analysis

Table 2 shows substantial dispersion across the 16 LLMs, with overall accuracy ranging from 14.1% (Llama-3.2-3B) to 70.4% (GPT-5.2). GPT-5.2 achieves the highest performance (70.4%), followed by GPT-5-mini (68.0%) and o4-mini (65.3%), while smaller models such as Llama-3.2-3B (14.1%) and Nova-2-lite (24.7%) perform markedly worse.

Performance generally improves with scale within families, except for Nova-Pro (42.0%) underperforming Nova-Lite (51.8%). In the Llama series, accuracy increases from 14.1% (3B) to 56.5% (70B) and 59.4% (Llama-4-Maverick). A similar progression holds in the GPT family (nano < mini < GPT-5.2), indicating that larger models better manage commitment tracking and symbolic reconstruction under interactional complexity.

However, variance within a single model across different category of entity-exchange often exceeds differences across adjacent tiers: even GPT-5.2 drops sharply under Implicit Commitment Drift and Encoded Articulation. Overall, scale improves robustness, but structural bottlenecks such as implicit revision and encoded articulation remain unresolved even for frontier LLMs.

5.3 Error Distribution

Each incorrect prediction is assigned exactly one primary error category corresponding to the earliest point of failure (E1–E8; Table 9). Conditioning errors on interactional phenomena reveals systematic alignment between conversational structure and failure mode. The overall distribution of errors across the axes is mentioned in Table 3.

Initiation. The initiation axis reveals a trade-off between explicit anchoring and symbolic fidelity. *Elicited Entry* achieves the highest accuracy (51.1%) and the lowest *Complete Recall Failure*

Phenomena	NE	E1	E2	E3	E4	E5	E6	E7	E8
Initiation Axis									
Elicited Entry	51.1	21.7	13.3	5.6	0.3	5.1	0.8	1.8	0.2
Volunteered Entry	46.7	14.9	12.4	9.0	1.0	10.1	1.9	3.9	0.2
Context-Triggered Entry	44.0	8.2	12.9	12.6	0.7	11.3	3.4	6.6	0.3
Evolution Axis									
Stable Commitment	56.3	15.2	15.3	4.5	0.6	3.2	1.5	3.1	0.2
Reinforced Commitment	56.4	14.2	12.9	7.5	0.5	4.6	1.1	2.5	0.2
Non-Contributory Interlea.	42.1	35.1	10.1	7.0	0.2	2.4	0.9	2.0	0.2
Self-Revised Commitment	32.1	28.3	9.9	7.8	0.1	19.1	0.8	1.7	0.1
Other-Initiated Commitmen.	34.0	23.4	11.3	11.1	0.4	15.9	1.2	2.6	0.1
Implicit Commitment Drift	27.0	8.7	12.6	14.3	0.9	30.7	1.7	3.7	0.4
Unresolved Commitment	27.7	17.2	11.7	11.3	1.6	10.3	6.1	12.3	1.8
Articulation Axis									
Uninterrupted Composition.	51.9	3.5	14.0	11.7	1.0	11.0	1.8	5.0	0.1
Fragmented Articulation	53.8	20.0	12.3	5.5	0.3	5.3	0.9	1.7	0.2
Encoded Articulation	30.7	42.9	13.4	4.4	0.3	5.0	1.0	2.1	0.3
Partial/Implicit Arti.	36.5	9.8	11.5	15.0	1.9	7.0	9.0	8.1	1.3
Degraded/Ambiguous Arti..	53.6	11.7	18.1	4.9	0.5	4.9	1.1	4.7	0.5

Table 3: Zero-Shot error distribution (%) by phenomena. **NE**: No Error; **E1**: Surface Realization; **E2**: Canonicalization; **E3**: Span Underspec.; **E4**: Slot Misassign.; **E5**: Discourse Misalign.; **E6**: Recall Failure; **E7**: Evidence Contradiction; **E8**: Ungrounded Generation.

(0.8%), reflecting the stabilizing effect of direct prompts. However, it exhibits the highest *Surface Realization Error* (21.7%), as these cases frequently involve complex, fragmented, or encoded articulations (~80%). *Context-Triggered Entry* forms a reasoning cliff (44.0% accuracy): when relevance must be inferred from dialogue state, models often fail to track the correct commitment, particularly when multiple candidate values are present, resulting in elevated recall failures despite canonical mentions (Table 11). *Volunteered Entry* lies between these extremes (46.7%), inheriting both surface vulnerabilities (14.9%) due to fragmented articulation and discourse misalignment risk (10.1%) in the absence of explicit anchoring.

Evolution. The evolution axis isolates commitment-tracking failures. *Conversational Interleaving*, even when non-contributory, fragments attention and produces the highest *Surface Realization Error* (35.1%). Revision phenomena further degrade robustness: *Implicit Drift* yields peak *Discourse State Misalignment* (30.7%) due to the absence of explicit repair markers, while *Self-Revised* exchanges remain vulnerable (19.1%) as models lock into earlier values and fail to update correctly. *Unresolved Commitments* induce 12.3% *Evidence Contradiction Errors*, reflecting a tendency to generate definitive values in the presence of unresolved ambiguity.

Articulation. Symbolic form strongly determines decoding failure. *Encoded* and *Fragmented Articulation* trigger the highest *Surface Realization Errors* (42.9% and 20.0%), indicating breakdowns when reconstructing values from phonetic encodings or turn-separated spans. *Partial/Implicit Expression* produces elevated *Complete Recall Failure* (9.0%), *Evidence Contradiction* (8.1%), and *Span Underspecification* (15.0%) as models struggle to ground underspecified intent. *Degraded Articulation* peaks in *Canonicalization Failure* (18.1%) under phonetic ambiguity. Notably, even *Canonical Expression* exhibits non-trivial *Discourse State Misalignment* (11.0%), largely driven by revision phenomena such as implicit drift, demonstrating that intact surface realization alone does not ensure correct commitment tracking.

5.4 Effectiveness of Error-Aware Prompting.

We introduce **Error-Aware (EA) Prompting** to explicitly ground task logic through six design principles (P_1–P_6, Appendix D), targeting known failures in symbolic fidelity and commitment tracking. As shown in Figure 4, improvements are *inversely proportional to interactional complexity*. In the Articulation axis, *Canonicalization Failure* (E_2) drops substantially from 13.9% to 3.2%, whereas *Surface Realization* (E_1) gains remain marginal (1–2% overall) and nearly unchanged under *Encoded Articulation* ($42.9\% \rightarrow 40.7\%$), indicating that prompting cannot repair breakdowns in symbolic remapping.

EA prompting improves *Discourse State Misalignment* (E_5) in both *Self-Revised Commitments* ($+4.7\%$) and *Implicit Commitment Drift* ($+4.8\%$), confirming that explicit grounding enhances synchronization with interactional state changes. However, stricter grounding induces conservative behavior: *Complete Recall Failure* (E_7) increases markedly, including an 18.9% recall drop for *Partial/Implicit Articulation*. Overall (Fig. 4), the model trades recall for higher symbolic and interactional integrity, preferring `NULL` over potentially ungrounded or corrupted extractions. Results are detailed in Table 10.

Despite overall gains, some error categories worsen under error-aware prompting. Most notably, E6: Recall Failure increases by +6.41% on average across all classes of entity-taxonomy, making it the dominant regression. This pattern suggests that while the prompt improves precision and normalization, it may also encourage more conservative extraction behavior, causing the model to omit valid entities. Smaller regressions are also observed for E7: Evidence Contradiction (+0.77%) and E4: Slot Misassignment (+0.19%), though these effects are comparatively minor. The increase in recall failures

is especially severe for phenomena already associated with ambiguity or weak articulation, such as Partial/Implicit Articulation and Unresolved Commitment, where omitted entities may reflect insufficient contextual grounding.

Taken together, these results suggest that error-aware prompting shifts model behavior toward a precision-oriented extraction regime. The model becomes better at generating canonical, well-grounded outputs, but sometimes at the cost of missing entities that require broader contextual inference. This tradeoff is desirable in settings where downstream systems prioritize correctness over coverage, but applications requiring maximal recall may benefit from combining error-aware prompting with a secondary recall-oriented recovery stage.

6 Synthetic Transcript Generation

To support future research, we constructed a phenomena-annotated synthetic dataset derived from 3,500 real contact center transcripts across nine industry verticals. The dataset is generated using a two-phase pipeline: the first phase produces raw synthetic transcripts that maintain structural and conversational fidelity, followed by a controlled entity insertion phase.

First, raw synthetic transcripts are generated from de-identified structural representations of real redacted transcripts, preserving conversational flow, speaker roles, domain context, and ASR-like spoken style while removing identifying content. Second, controlled entity exchanges are inserted into the synthetic transcripts so that each example has a known target entity value and a specified label on each axis of the phenomena taxonomy. This yields transcripts that are both privacy-preserving and suitable for controlled evaluation of entity extraction under diverse conversational phenomena. Full algorithmic details are in Appendix F.

6.1 Evaluation

We evaluate generated transcripts against their real counterparts along three dimensions: lexical divergence, semantic fidelity, and structural preservation (Table 4). Lexical metrics quantify surface overlap and serve as an indicator of memorization and re-identification risk; low BLEU-4 (0.076) and ROUGE-L (0.218) suggest limited copying of source wording. Semantic metrics assess whether the generated calls preserve the underlying intent and topic progression despite surface-level rewriting; high BERTScore F_1 (0.865) and sentence-embedding cosine similarity (0.903) indicate strong semantic alignment with the originals. Structural metrics measure preservation of conversation-level properties; a turn-count ratio close to one (0.975) and a small speaker-ratio delta (0.014) indicate that call length and agent/customer balance are largely maintained. Refer F.3 for details on evaluation metrics.

Dimension	Metric	Mean	Std
Lexical	BLEU-4	0.076	0.018
	ROUGE-L F_1	0.218	0.026
	Jaccard overlap	0.457	0.051
Semantic	TF-IDF cosine	0.213	0.035
	Sentence-emb. cosine	0.903	0.031
	BERTScore F_1	0.865	0.019
Structural	Turn-count ratio	0.975	0.296
	Speaker-ratio Δ	0.014	0.035
	Turn-length KS stat.	0.224	0.071

Table 4: Quality metrics for generated transcripts vs. their real counterparts.

7 Conclusion

We present a phenomenon centric framework for analyzing entity extraction in real world spoken dialogue through the lens of entity exchange. By modeling Initiation, Evolution, and Articulation as orthogonal interactional dimensions, we provide a structured account of how conversational dynamics govern entity behavior. Empirical evaluation across 16 LLMs reveals systematic and predictable failure patterns conditioned on these phenomena.

Limitations

Our study evaluates zero-shot extraction behavior and does not examine how supervised adaptation or interaction-aware fine-tuning may alter the observed failure patterns. The three-axis taxonomy assigns a single dominant label per axis to each episode, enabling systematic comparison but potentially compressing episodes that exhibit multiple or transitional interactional behaviors.

We do not claim that the proposed three-axis taxonomy is exhaustive or uniquely optimal. Alternative axes or additional subcategories could plausibly be defined, and different thought process may yield alternative structural decompositions of entity exchange. Our objective is not to assert exclusivity, but to provide a principled and operationally useful framework grounded in empirical observation. Our

proposed taxonomy was iteratively refined with five annotators and informed by real-world contact center data to balance informativeness, actionability, and annotation reliability. During development, more granular schemas introduced overlapping dimensions and increased subjectivity, while coarser formulations reduced explanatory specificity. The final design therefore reflects a deliberate trade-off between structural expressiveness and reproducible annotation.

The benchmark focuses on English contact-center dialogue, and interactional dynamics may vary across languages, domains, or multimodal settings. Finally, while we benchmark multiple LLM families and sizes, our results do not exhaust the space of architectures, prompting strategies, or decoding configurations.

Ethical Considerations

Data Source, Privacy, and Governance. This study analyzes real-world spoken contact center conversations to examine how interactional structure affects entity extraction. All conversations were processed under strict privacy, security, and data protection controls prior to research use. Personally identifiable information (PII), protected attributes, and organization-specific metadata were removed or irreversibly anonymized through automated redaction pipelines followed by manual verification. Raw audio, speaker identities, and account-level identifiers are not released. The benchmark consists solely of de-identified transcripts and derived annotations necessary for scientific analysis.

Access to the data was restricted to authorized researchers operating within secure environments. Transcripts were stored and processed on controlled infrastructure with audit mechanisms and access logging. The dataset was constructed exclusively for research purposes, and no attempt was made to re-identify individuals. Any example excerpts included in the paper are minimally edited or paraphrased to prevent residual identification while preserving the interactional phenomena under analysis.

Regulatory Compliance and Third-Party Model Use All experiments described in this paper were conducted in compliance with applicable privacy and data protection regulations. Interactions with third-party language models were governed by appropriate contractual and regulatory safeguards where required. These measures ensured that sensitive or regulated data were not disclosed for purposes beyond the immediate research use case. No data were retained, stored, or used by external service providers for model training, fine-tuning, or product improvement.

Annotator Roles and Oversight. Annotation and validation were conducted by a team with expertise in computational linguistics and dialogue systems. To ensure accountability and schema consistency, responsibilities were structured as follows:

1. A primary annotator verified entity values and normalization consistency across transcripts.

2. A secondary annotator reviewed interactional labels across the three axes (Initiation, Evolution, and Articulation) and evaluated alignment with the formal taxonomy definitions.

3. A senior reviewer adjudicated disagreements, refined labeling guidelines where necessary, and performed spot audits for quality assurance.

Disagreements were resolved through structured discussion to maintain consistency and minimize subjective drift in labeling decisions.

Risk and Downstream Impact. In real-world deployments, extracted entities may trigger downstream automated actions such as database updates, workflow execution, or verification procedures. Extraction errors can therefore propagate beyond textual misunderstanding and lead to operational consequences. This work explicitly analyzes systematic failure modes in large language models to better understand how conversational dynamics contribute to such risks. The objective is diagnostic: to identify predictable interactional failure patterns that can inform safer system design.

Limitations and Responsible Use. Improved extraction accuracy does not by itself guarantee safe automation. Entity extraction systems operate within broader socio-technical pipelines that include speech recognition, normalization layers, policy constraints, and human oversight. Responsible deployment requires layered safeguards, monitoring, and human-in-the-loop mechanisms. This work provides a structured analytical framework

for understanding interactional failure modes, supporting the development of more transparent, accountable, and reliable conversational AI systems.

References

Jayachandu Bandlamudi, Kushal Mukherjee, Prerna Agarwal, Ritwik Chaudhuri, Rakesh Pimplikar, Sampath Dechu, Alex Straley, Anbumunee Ponniah, and Renuka Sindhgatta. 2024. Building conversational artifacts to enable digital assistant for apis and rpas. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):22725–22733.

Xue-Yong Fu, Cheng Chen, Md Tahmid Rahman Laskar, Shashi Bhushan Tn, and Simon Corston-Oliver. 2022. An effective, performant named entity recognition system for noisy business telephone conversation transcripts. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 96–100, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Seyed Mahed Mousavi, Gabriel Roccabruna, Simone Alghisi, Massimo Rizzoli, Mirco Ravanelli, and Giuseppe Riccardi. 2024. Are llms robust for spoken dialogues? *Preprint*, arXiv:2401.02297.

Alpesh Patel. 2025. Transforming service with data-driven ai agents: The evolution of salesforce agentforce agents.

Ayesha Qamar, Arushi Raghuvanshi, Conal Sathi, and Youngseo Son. 2025a. Auto review: Second stage error detection for highly accurate information extraction from phone conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1308–1321, Vienna, Austria. Association for Computational Linguistics.

Ayesha Qamar, Jonathan Tong, and Ruihong Huang. 2025b. Do LLMs understand dialogues? a case study on dialogue acts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26219–26237, Vienna, Austria. Association for Computational Linguistics.

Arushi Raghuvanshi, Vijay Ramakrishnan, Varsha Embar, Lucien Carroll, and Karthik Raghunathan. 2019. Entity resolution for noisy ASR transcripts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 61–66, Hong Kong, China. Association for Computational Linguistics.

Emanuel A. Schegloff, Gail D. Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53:361 – 382.

Shuzheng Si, Wen-Cheng Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. In *Neural Information Processing Systems*.

Piotr Szymański, Lukasz Augustyniak, Mikolaj Morzy, Adrian Szymczak, Krzysztof Surdyk, and Piotr Żelasko. 2023. Why aren't we NER yet? artifacts of ASR errors in named entity recognition in spontaneous speech transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1746–1761, Toronto, Canada. Association for Computational Linguistics.

A Dataset Statistics and Label Distribution

A.1 Real-World Dataset

We construct a benchmark from real-world English dyadic, multi-turn conversations between human customer-support agents and customers across operational verticals including logistics, edtech, debt collections, insurance, finance, healthcare, and retail. The corpus comprises 6,387 de-identified transcripts sampled from live production contact center environments, ensuring that interactions reflect authentic service exchanges rather than scripted, simulated, or crowdsourced dialogue. Call lengths range from approximately 25 conversational turns ($\sim$2 minutes) to as many as 180 turns ($\sim$45 minutes), capturing substantial variation in conversational depth and structural complexity. The average turn length is 24 words, with the shortest turn consisting of a single word and the longest extending to 123 words.

All conversations originate from real contact center interactions and therefore preserve natural spoken phenomena, including interruption, self-repair, incremental specification, collaborative grounding, negotiation, hesitation, and cross-turn distribution of entity values. To mitigate transcript-led errors and reduce the impact of ASR artifacts on evaluation, the transcripts used for benchmarking were manually labeled, ensuring that entity annotations reflect the intended conversational commitments rather than surface transcription noise. Transcripts were de-identified prior to annotation. Prior to annotation, all transcripts were systematically de-identified to remove or mask personally identifiable information (PII) and payment card information (PCI). Sensitive attributes, including personal names, contact details, financial identifiers, and

account-linked information, were anonymized or replaced with type-consistent placeholders to prevent re-identification while preserving conversational structure. This de-identification process ensures compliance with data protection requirements and enables research use without exposing sensitive customer information.

The benchmark spans 12 entity types organized into four broad categories. Numeric entities include Helpline Number, Case Number, Order ID, and Order Total. Structured alphanumeric identifiers include Tracking Number, Policy ID, and Department ID. Temporal entities include Date of Registration and Pickup Time. Free-text entities include Agent Name, Store Address, and Support Email. Together, these entity types cover strictly numeric identifiers, structured codes, temporal expressions, and unconstrained lexical entities, providing a heterogeneous and deployment-realistic evaluation setting for entity extraction in spoken dialogue. We only include the conversations where the queried entity type is relevant to the interaction, ensuring evaluation measures extraction when the entity is within task scope rather than absent from the dialogue.

A.2 Label Distribution

The evaluation corpus consists of $N = 6,387$ annotated entity exchange instances extracted from diverse conversational transcripts. To ensure broad coverage of symbolic and structural challenges, we utilize 12 distinct entity types categorized into four primary semantic classes: *Numeric* (e.g., phone numbers, order id etc), *Alpha-Numeric* (e.g., Tracking Number, alphanumeric IDs), *Temporal* (e.g., registration date, pickup time) and *Free-text* (e.g., names, store addresses, support email). Each instance is triply annotated across our three interactional axes: Initiation, Evolution, and Articulation, providing a granular view of conversational entity dynamics.

Table 5 provides the comprehensive distribution of labels across the Initiation, Evolution, and Articulation axes. To analyze the interactional density of our benchmark, Figure 3 illustrates the **pairwise distribution** of labels. This matrix captures the co-occurrence patterns between distinct phenomena, such as the intersection of *Implicit Drift* (E_5) and *Encoded Articulation* (A_2). These pairwise mappings highlight the specific high-entropy regions within the dataset where models must simultaneously navigate discourse state changes and complex symbolic remapping. Beyond the interactional axes, we provide a detailed breakdown of the data by content type. Table 6 and Table 7 show the semantic class-wise distribution and entity-wise distribution, respectively

Axis	Phenomena Class (Code)	Count	%
Initiation	Elicited Entry (I_0)	4,451	69.7
	Volunteered Entry (I_1)	1,151	18.0
	Context-Triggered (I_2)	785	12.3
Evolution	Stable Commitment (E_0)	2,312	36.2
	Reinforced (E_1)	1,926	30.2
	Non-Contributory Interleaving (E_2)	702	11.0
	Self-Revised Commitment (E_3)	262	4.1
	Other-Initiated Commitment (E_4)	999	15.6
	Implicit Drift Commitment (E_5)	101	1.6
	Unresolved Commitment (E_6)	85	1.3
Articulation	Canonical Expression (A_0)	1,663	26.0
	Fragmented Articulation (A_1)	3,441	53.9
	Encoded Articulation (A_2)	989	15.5
	Partial/Implicit Articulation (A_3)	95	1.5
	Degraded/Ambiguous Articulation (A_4)	199	3.1

Table 5: Label distribution across interactional axes ($N = 6,387$). The corpus spans 12 entity types across numeric, alpha-numeric, and free-text categories.

B Models Evaluated

To assess the impact of interactional complexity across varying architectures and reasoning capabilities, we selected a diverse set of 16 large language models. Our selection spans multiple major model providers and open-source families, including Meta (Llama), Amazon (Nova), Anthropic (Claude), Google (Gemini), and OpenAI (GPT/o-series). Furthermore, we intentionally included models of varying scales and reasoning specializations (e.g., llama-3.2 3B vs. llama-3.3 70B; gpt-5.2 vs. gpt-5-mini). This approach allows us to analyze how model scale, training objectives, and specialized reasoning pathways influence the prevalence of errors. For full transparency and reproducibility, the specific model identifiers and their shorthand names are listed in Table 8.

C Error Taxonomy

We propose a comprehensive error taxonomy to evaluate entity extraction from conversational transcripts. Derived from a systematic analysis of model behavior across varying levels of interactional noise, this framework provides complete coverage of all observed failures. By categorizing errors into eight distinct classes, we isolate failures related to canonicalization, symbolic remapping,

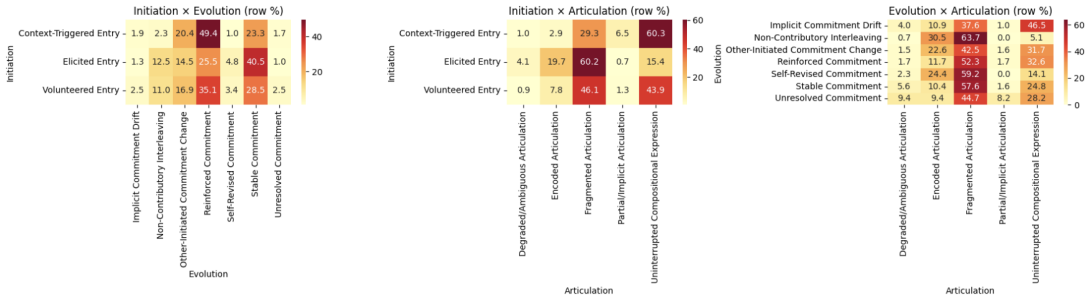

Figure 3: Pairwise Distribution

Entity Type	Count (n)	%
Numeric	1,981	31.01
Alpha-Numeric	1,841	32.8
Temporal	1,307	20.46
Free-text	1,253	19.62
Total	**6,387**	**100.0**

Table 6: Distribution of entity types by semantic category ($N = 6,387$). The corpus is heavily weighted toward symbolic sequences (75.5% total) to evaluate precision in decoding and commitment tracking.

Category	Entity Type	Count	%
Numeric	Helpline Number	495	7.8
($n = 1,981$)	Case Number	467	7.3
	Order ID	729	11.4
	Order Total	295	4.6
Alpha-Numeric	Tracking Number	665	10.4
($n = 1,841$)	Policy ID	602	9.4
	Department ID	574	9.0
Temporal	Date of Registration	566	8.9
($n = 1,307$)	Pickup Time	741	11.6
Free-text	Agent Name	709	11.1
($n = 1,253$)	Store Address	291	4.6
	Support Email	253	4.0
Total	—	**6,387**	**100.0**

Table 7: Entity-wise distribution across semantic categories ($N = 6,387$). The corpus balances high-precision numeric extraction with complex alphanumeric decoding and free-text resolution.

Short Name	Model ID
Nova Micro	amazon/nova-micro-v1
Nova Lite	amazon/nova-lite-v1
Nova Pro	amazon/nova-pro-v1
Nova 2 Lite	amazon/nova-2-lite-v1:0
Llama-3.2-3B	meta/llama3-2-3b-instruct-v1
Llama-3.3-70B	meta/llama3-3-70b-instruct-v1
Llama-4-Scout	meta/llama4-scout-17b-instruct-v1:0
Llama-4-Maverick	meta/llama4-maverick-17b-instruct-v1
Claude-4.5-Haiku	anthropic/claude-haiku-4-5-20251001-v1:0
Claude-4.5-Sonnet	anthropic/claude-sonnet-4-5-20250929-v1:0
Gemini-2.5-Flash-lite	google/gemini-2.5-flash-lite
Gemini-2.5-Flash	google/gemini-2.5-flash
GPT-5-nano	openai/gpt-5-nano
GPT-5-mini	openai/gpt-5-mini
GPT-5.2	openai/gpt-5.2
o4-mini	openai/o4-mini

Table 8: Identifiers of large language models (LLMs) used in this work.

discourse state synchronization, and evidence detailed as defined in Table 9.

C.1 Entity Extraction Evaluation and Error Classification

To evaluate entity extraction and identify the patterns of error, we employ a structured three-stage evaluation pipeline that combines deterministic matching with context-aware diagnostic analysis. Inter-annotator agreement was high for both entity value evaluation ($\kappa = 0.82$) and error category assignment ($\kappa = 0.71$), indicating reliable annotation and consistent error classification.

1. **Stage 1: Lexical Exact Match.** We first compare the normalized model prediction with the normalized ground truth using a deterministic string match. This step accounts for trivial formatting differences such as casing or whitespace. Predictions that match at this stage are labeled **NE (No–Error)** and require no further analysis. All remaining mismatches proceed to Stage 2.
2. **Stage 2: Context-Constrained Diagnostic.** For non-lexical matching cases, we apply an *Error Analysis Prompt* (see Appendix, Figs. 8 and 9) over a localized portion of the transcript. The prompt assigns exactly one category from the predefined taxonomy (E1–E8), with **NE** (No–Error) also available when semantic equivalence

is identified despite surface differences. This stage focuses on identifying errors related to normalization, span selection, slot assignment, and conversational state tracking.

3. **Stage 3: Grounding Failure Detection.** Finally, we examine the full transcript to verify whether the predicted value is consistent with the final committed value in the conversation. This step improves precision in identifying discourse-related errors, particularly cases involving revisions, superseded values, or unsupported generations.

This pipeline enables consistent and mutually exclusive error assignment while preserving alignment with the error taxonomy described in Appendix C.

D Error-Aware Entity Extraction

The Error-Aware (EA) Prompt integrates six core design principles (P_1–P_6), detailed in Prompt 7, specifically engineered to mitigate the reasoning failures identified in standard zero-shot extractions.

E Qualitative Error Analysis

Table 11 demonstrates representative examples of the qualitative failure modes observed during our evaluation.

E.1 Qualitative Illustration: Explicit Self-Repair

We revisit the explicit self-repair example introduced in 1 to examine commitment tracking under controlled conditions. The conversational context, entity type, and transcription quality are identical across models; only model behavior varies.

Three of four evaluated models fail to recover the final committed value. One retains a superseded digit and truncates the output. Another discards the pre-repair segment and repeats post-repair digits. A third produces a hybrid output combining invalidated and corrected segments. Only one model fully retracts the invalidated value and extracts the final commitment.

Importantly, the repair cue is explicit, adjacent, and linguistically unambiguous. The failure therefore does not stem from transcription noise or canonicalization difficulty but from misalignment with conversational repair logic. This example illustrates that explicit revision is not intrinsically unresolvable for large language models; rather, robustness varies sharply across architectures in their ability to update discourse state.

F Synthetic Transcript Generation: Algorithm

F.1 Phase 1: Raw Synthetic Transcript Generation

To retain conversational structure and entity distribution, each source transcript is decomposed into an ordered sequence of *topical segments*: contiguous turn groups that each complete a single communicative intent (e.g. an agent request followed by a customer response, or a self-correction sequence). For each segment, a language model extracts a de-identified summary of the conversational exchange, per-speaker sentiment, and an entity *type* inventory.

A second de-identification pass inspects every segment summary for residual leakage—quasi-identifiers such as brand names, geographic references, specific calendar years, or government service names—and rewrites any that remain.

In parallel, a separate model call extracts a de-identified call reason, which is likewise subjected to leakage correction.

Raw ASR-like synthetic transcripts are then generated by conditioning a generative model on the industry domain, approximate turn count, ordered anonymized topical segments, and deidentified call reason.

F.2 Phase 2: Controlled Entity Insertion

To introduce controlled entity exchanges, we mine approximately 60 named conversational patterns from a real annotated transcript corpus distributed across the 15 phenomena classes in our taxonomy. Each pattern describes a recurring conversational realization of a phenomena class; for example, *direct request* captures an agent explicitly asking for an entity value ("may I have your order number?"), while *NATO articulation* captures a speaker spelling a value using forms such as "B as in bravo."

For each pattern, we construct sets of positive and negative exemplars. Positive exemplars are real conversational snippets that strongly satisfy the pattern, evaluated against two criteria: the class boundary condition and the pattern-specific matching rule. For example, a positive exemplar for *agent readback* must include the agent repeating the actual entity value, not merely acknowledging it with

Code: Error Category	Formal Definition	Illustrative Example
E1: Surface Realization	Reproduction with character or digit corruption (substitution, deletion, or reordering) affecting literal fidelity while preserving the intended referent.	`12345` → `12354`
E2: Canonicalization Failure	Recognition of constituent components without successful normalization into the task-required atomic representation.	`23.5` → `twenty three five`
E3: Span Underspecification	Extraction of a proper subset of the entity that fails to provide sufficient information to uniquely identify the referent.	`12B Maple St` → `Maple St`
E4: Semantic Slot Misassignment	Assignment of a correctly grounded entity value to an incorrect entity type or semantic role relative to the task.	Extract User → `user@domain.com`
E5: Discourse Misalignment	Extraction of a value corresponding to an incorrect conversational state (e.g., a provisional or superseded value).	*"Was 54, now 51.5"* → `54`
E6: Complete Recall Failure	Failure to extract any entity value despite the presence of explicit and unambiguous evidence in the input.	*"Order A73921"* → `null`
E7: Evidence Contradiction	Output of an entity value that directly conflicts with explicit lexical evidence found in the input.	`12345` → `34350`
E8: Ungrounded Generation	Production of an entity value in the absence of any lexical or contextual evidence supporting it (pure fabrication).	No date in text → `March 10, 2024`

Table 9: Taxonomy of entity extraction and grounding errors. Categories are prioritized based on the earliest point of failure in the extraction pipeline, providing a mutually exclusive classification for performance analysis.

"okay" or "correct." Negative exemplars are drawn from the nearest neighboring classes to sharpen decision boundaries; for instance, a backchannel-only exchange is a negative exemplar for reinforced commitment because it contains acknowledgment without value repetition. An applicability matrix specifies which entity–pattern combinations are supported by empirical evidence and can be sampled during generation.

For each synthetic transcript and target entity pair, a greedy coverage-balancing algorithm selects one class–pattern combination per axis. A lightweight model then generates a format-compliant canonical value. The target class, detailed phenomena descriptions, pattern articulations, boundary-disambiguation rules, and positive and negative exemplars are passed to a language model, which produces a structured *edit plan*: a sequence of turn-level insertions, replacements, and deletions at specified anchor points. The pipeline applies this plan deterministically, preserving the majority of the original transcript verbatim.

After insertion, a pattern-matching pass verifies the canonical value, and an independent judge classifier—without access to the generation target—validates phenomena labels. Any discrepancies are resolved through targeted rewrites, followed by re-validation.

F.3 Evaluation Metrics

All metrics are computed pairwise between each real source transcript T_i and its generated counterpart S_i. We evaluate across four dimensions:

- **Lexical** metrics measure surface-form overlap and serve as a proxy for re-identification risk—lower is safer. **BLEU-4** computes 4-gram precision of S_i against T_i over the full document. **ROUGE-L** F_1 measures the longest common subsequence between the two texts. **Jaccard overlap** is the token-set intersection over union: $|W_T \cap W_J| \,/\, |W_T \cup W_J|$.
- **Semantic** metrics assess whether the synthetic transcript preserves the conversational intent and topic structure of the original, regardless of surface wording. **TF-IDF cosine** computes document-level similarity using unigram and bigram term frequencies (full document, no truncation). **Sentence-embedding cosine** encodes

Phenomena	NE	E1	E2	E3	E4	E5	E6	E7	E8
Initiation Axis									
I_0: Elicited Entry	59.3	20.8	3.0	4.9	0.4	3.9	4.3	3.0	0.3
I_1: Volunteered Entry	52.5	13.8	4.2	7.5	1.5	8.6	7.3	4.2	0.5
I_2: Context-Triggered Entry	49.1	7.1	3.8	9.7	1.9	10.2	10.9	6.6	0.7
Evolution Axis									
E_0: Stable Commitment	64.2	14.9	2.9	3.7	0.7	2.4	6.6	4.3	0.3
E_1: Reinforced Commitment	63.4	13.6	3.8	5.9	0.9	4.1	4.5	3.3	0.4
E_2: Non-Contributory Interlea.	48.6	32.6	3.1	6.2	0.4	1.7	4.3	2.6	0.4
E_3: Self-Revised Commitment	41.1	26.1	3.2	7.1	0.4	14.4	5.2	2.1	0.3
E_4: Other-Initiated Commitmen.	40.4	21.7	3.6	9.8	1.0	14.3	5.2	3.3	0.6
E_5: Implicit Commitment Drift	32.4	8.7	3.8	16.1	0.8	25.9	6.7	5.0	0.5
E_6: Unresolved Commitment	30.4	15.1	3.7	7.2	1.0	8.0	22.1	10.6	2.0
Articulation Axis									
A_0: Uninterrupted Composition.	58.5	2.8	3.6	9.3	2.0	9.5	8.1	5.6	0.5
A_1: Fragmented Articulation	61.3	19.5	3.0	4.8	0.2	4.1	4.1	2.6	0.4
A_2: Encoded Articulation	38.2	40.7	4.7	4.5	0.4	3.9	4.9	2.5	0.3
A_3: Partial/Implicit Articula.	33.9	9.5	2.7	9.4	1.5	7.2	28.0	6.8	0.9
A_4: Degraded/Ambiguous Articu.	58.5	11.9	2.1	3.7	0.1	4.9	7.1	10.9	0.7

Table 10: Error distribution (%) of entity-extraction across entity-exchanges under the error-aware prompt. **NE**: No Error; **E1**: Surface Realization; **E2**: Canonicalization; **E3**: Span Underspec.; **E4**: Slot Misassign.; **E5**: Discourse Misalign.; **E6**: Recall Failure; **E7**: Evidence Contradiction; **E8**: Ungrounded Generation.

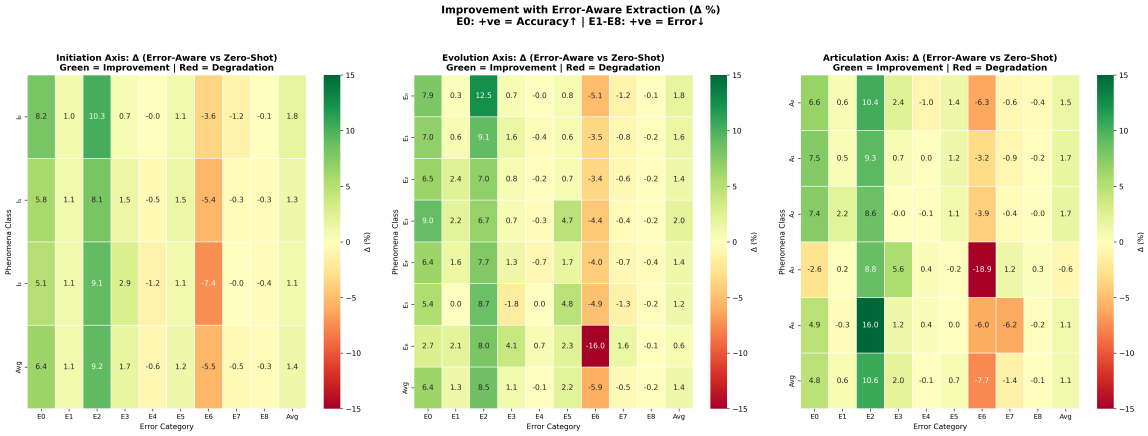

Figure 4: Distribution of errors after applying error-aware entity extraction ($\Delta\%$) across three entity-exchange phenomena axes. Positive values (green) indicate improvement: for *E0* (No error), accuracy increased; *E1–E8*, error rate reduced. Negative values (red) indicate degradation. Results averaged across all models.

Axis	Representative Utterance	GT	Pred	Error Mode	Interactional Reasoning
Initiation	*"This is in regards to the case number ending in **1860**, correct?.."*	1860	NULL	Complete Recall Failure (E_6)	Breakdown in **dialogue stack tracking**. Despite stable commitment, model misses relevance when entity enters via implicit task state.
Evolution	*"4 0 1 4 5 4... i'm sorry **not 4 5 4**, it is **5 4 5 8 6 6 9**."*	. . . 5458669	. . . 454066	Discourse State Misalignment (E_5)	Model fails to resolve explicit, adjacent repair logic, preserving invalidated segments as hybrid hallucination.
Articulation	*"Last 8 digits of Tracking Number... **J** as in Jack **E** as in Echo **A** as in Alpha 4 4 9 8 0."*	JEA44980	J**AE**44980	Surface Realization Error (E_1)	Failure in **symbolic remapping**. Under high cognitive load from phonetic encoding, model executes literal **character swap** (JEA → JAE) despite stable commitment.
Articulation	*"..and your registration date is eight twenty ninety five"*	8/20/95	eight twenty ninety five	Canonicalization Error (E_2)	Compositional date expression requires numeric normalization; model outputs verbatim surface form.

Table 11: Qualitative failure modes across axes

each transcript in 200-word chunks, mean-pools the chunk embeddings, and measures the co-

sine between the resulting document vectors. **BERTScore** F_1 aligns contextual token embeddings (RoBERTa-large) between T_i and S_i and reports the F_1 of the greedy alignment.

- **Structural** metrics verify that turn-taking dynamics and speaker balance are preserved. **Turn-count ratio** is $|S_i| \,/\, |T_i|$; values near 1.0 indicate the synthetic transcript matches the original length. **Speaker-ratio delta** is the absolute difference in agent-turn fraction between S_i and T_i; values near 0 indicate preserved speaker balance. **Turn-length KS statistic** is the Kolmogorov–Smirnov distance between the per-turn token-length distributions of T_i and S_i.

G Prompt Templates

Initiation Taxonomy Prompt

INITIATION TAXONOMY
Models how an entity first enters the conversation.
I0. Elicited Entry
Entity first enters the conversation in response to an explicit request/prompt.

- Example: "May I have your order number?" → "o nine five four"
- Example: "Do you want your case number?" → "three five seven six"

I1. Volunteered Entry
Entity is introduced proactively without a prior request/prompt.

- Example: "My account ID is a zero five nine.."

I2. Context-Triggered Entry
Entity becomes relevant implicitly through conversational or external context.

- Example: "I see order o nine five four from your previous case"

ASSIGNMENT RULE:

- Choose EXACTLY ONE initiation label based on how the entity first enters the conversation.
- Focus on the trigger that made the entity relevant, not how it was articulated.

RULE OF THUMB:
Ask: "What happened IMMEDIATELY BEFORE the entity first appeared?"

- → If there was a QUESTION or REQUEST for this entity → **I0 (Elicited Entry)**
  Key signals: "May I have..?", "Can you give me..?", "What is your..?", "Could you provide..?"
- → If the speaker OFFERED the entity WITHOUT being asked → **I1 (Volunteered Entry)**
  Key signals: Speaker introduces entity mid-conversation, often with "My X is...", "I have...", "Let me give you..."
- → If the entity was PULLED FROM CONTEXT (system, prior conversation, external source) → **I2 (Context-Triggered Entry)**
  Key signals: "I see...", "According to our records...", "From your previous case...", "The system shows..."

IMPORTANT:

- Ignore HOW the value was spoken (spelled out, fragmented, etc.) — that's Articulation.
- Ignore WHETHER the value changed later.
- Focus ONLY on the conversational trigger that brought the entity into discussion.

Figure 5: Initiation Taxonomy Prompt

Evolution Taxonomy Prompt

EVOLUTION TAXONOMY
Models how commitment to an entity is constructed, reinforced, revised, or fails to resolve through conversational interaction.
DEFINITION OF COMMITMENT: A commitment is an explicit verbal act in which a speaker proposes, asserts, or accepts a specific value (or portion of a value) as the intended content for the target entity. A commitment: binds the speaker to that value until revised or rejected; can be partial (e.g., first three digits) or complete; can be proposed, reinforced, revised, or left unresolved; is distinct from backchannels (e.g., "okay", "mm-hm") which acknowledge but do not commit to value content.
E0. Stable Commitment
Def: A single commitment is introduced and remains unchanged with no subsequent interactional engagement affecting its content.
Key: Single commitment proposed and implicitly accepted; No repetition, confirmation, or interleaving; No revisions, repairs, or competing alternatives; Value stated once and conversation moves on; Always happens in a single turn.
Ex: "My order number is OD4481." → (conversation continues)
E1. Reinforced Commitment
Def: The commitment is reinforced through explicit repetition of the value content by the same speaker or another participant, confirming without adding, removing, or modifying content.
Key: Value is explicitly REPEATED or RESTATED (by self or other); Repetition confirms/reinforces the commitment; No change to value content; Distinct from E2: E1 requires value repetition, E2 is backchannel only.
Ex: Customer: "Five four nine one." → Agent: "Five four nine one, got it."
E2. Non-Contributory Interleaving
Def: The commitment is temporally interleaved with interactional behavior that does NOT contribute to or modify commitment content.
Key: Interleaving includes: backchannels, fillers, overlaps, acknowledgements; No contribution to value construction; No rejection or modification; Commitment remains intact after interleaving.
Ex: "Five five five..." → "Okay, go ahead" → "one two three four"
E3. Self-Revised Commitment
Def: The SAME speaker explicitly revises, corrects, or rejects their OWN prior commitment.
Key: Explicit repair or rejection markers present (e.g., "sorry", "actually", "no wait", "I mean"); Replacement or modification of earlier commitment; Final commitment differs from earlier one; Self-initiated repair.
Ex: "Five five... sorry, five six seven."
E4. Other-Initiated Commitment Change
Def: The OTHER participant alters, corrects, or substantively contributes to the commitment content, resulting in a different final value than originally proposed.
Key: Other speaker CHANGES the value (not just confirms it); Includes: correction, rejection, collaborative contribution, negotiation; Final value differs from or extends original; Distinct from E1: E1 = same value confirmed, E4 = value changed by other.
Ex: Agent: "Is it AB1243?" → Customer: "No, AB1234."
E5. Implicit Commitment Drift
Def: Commitment changes without explicit linguistic markers of repair, rejection, or correction. The shift happens SILENTLY.
Key: No explicit repair cues (no "sorry", "no", "actually", "wait"); Value silently shifts to a DIFFERENT value; Conversation proceeds as if new commitment was always intended; Commitment IS resolved (just changed silently).
Ex: "Pickup around ten... eleven works." (shifts 10→11, no marker)
E6. Unresolved Commitment
Def: No stable commitment is established by the end of the interaction.
Key: Multiple competing commitments remain possible; Explicit uncertainty or lack of resolution; No final accepted commitment.
Ex: "It could be 452 or 454." / "Let me check and get back to you."

PRIORITY RULES (Highest → Lowest):
E6 > E5 > E4 > E3 > E2 > E1 > E0
PRIORITY RATIONALE: This ordering reflects increasing complexity in commitment tracking. Unresolved episodes (E6) dominate all others since no extractable value exists. Implicit drift (E5) outranks explicit changes (E3/E4) because silent shifts lack interactional cues, making them harder to track. Other-initiated changes (E4) supersede self-repairs (E3) as they involve cross-speaker coordination. Self-revision (E3) outranks non-contributory interleaving (E2) because it introduces competing commitments. Interleaved delivery (E2) subsumes reinforcement (E1) since multi-turn construction is more complex than simple confirmation. Reinforcement (E1) outranks stability (E0) as any interaction adds complexity over a single clean commitment.
ASSIGNMENT RULE: Return EXACTLY ONE label per entity episode. In case of multiple applicable classes, assign the highest-priority label.

RULE OF THUMB - Decision Tree:
Ask: "How did commitment change through interaction?"
Step 1: Was a FINAL commitment reached?
→ NO, multiple values still possible, uncertainty remains → **E6**
Step 2: Did the value CHANGE at some point?
→ NO change at all, single value proposed and accepted → Step 3
→ YES, value changed → Step 4
Step 3: Was there any INTERACTION around the commitment?
→ NO interaction, just stated and done → **E0**
→ YES, confirmed/repeated by same or other party → **E1**
→ YES, but only non-contributory (backchannels) → **E2**
Step 4: HOW did the value change?
→ Changed SILENTLY without any marker → **E5**
→ Changed with EXPLICIT marker → Step 5
Step 5: WHO initiated the change?
→ SAME speaker corrected themselves → **E3**
→ OTHER speaker corrected/changed it → **E4**

KEY DISTINCTIONS:
E0 vs E1: E0 = No interaction after commitment (value stated, conversation moves on); E1 = Value explicitly REPEATED (by self or other) to confirm
E1 vs E2: E1 = Value content REPEATED ("Five four nine one, got it"); E2 = Backchannel only, value NOT repeated ("yes", "okay", "mm-hm")
E1 vs E4: E1 = SAME value confirmed via repetition; E4 = Value CHANGED by other participant
E3 vs E4: E3 = SELF corrects own value; E4 = OTHER corrects/changes value
E3/E4 vs E5: E3/E4 = EXPLICIT marker present (sorry, no, actually); E5 = NO marker, value shifts silently
E5 vs E6: E5 = Value changed but RESOLVED; E6 = NOT RESOLVED (uncertainty remains)

Articulation Taxonomy Prompt

ARTICULATION TAXONOMY
Models how the entity value is expressed on the surface.
CORE PRINCIPLE: Articulation is assigned with respect to the final committed semantic value. When multiple surface realizations of this value occur within an entity episode, assign the articulation class corresponding to the MOST COMPLEX realization, independent of speaker or turn order.

A0. Uninterrupted Compositional Expression
Def: The final value is articulated continuously and cohesively, using a single COMPOSITIONAL numeric or lexical grammar, with exactly one valid interpretation.
ALL must hold: Articulated in a SINGLE TURN (not spread across multiple turns); Single, continuous utterance or uninterrupted span; COMPOSITIONAL grammar (e.g., "two thousand one hundred" NOT "two one zero zero"); Exactly one valid parse; NO symbol-by-symbol or digit-by-digit enumeration (that is A1!); No encoding, masking, or ambiguity; No interruptions, backchannels, confirmations, or speaker changes during articulation.
Ex: "The total is two thousand one hundred and forty five."
A1. Fragmented Articulation
Def: The final value is articulated as a sequence of canonical symbols or substrings distributed across multiple segments (tokens, pauses, or turns), OR as digit-by-digit/character-by-character enumeration, without introducing an alternative encoding.
Key: Symbols themselves are canonical (directly spoken); No indirect encoding or masking; Includes: digit-by-digit enumeration, character spelling, chunked delivery.
Ex: "Five five five one two three four" (digit-by-digit); "a m i l i a" (character-by-character); "two one zero zero" (digit enumeration); Customer: "five five five" → Agent: "okay" → Customer: "one two three four"
A2. Encoded Articulation
Def: The final value is articulated via an indirect but deterministic encoding, where symbols are expressed through descriptions, labels, or mnemonics rather than directly.
Key: Symbols are not spoken canonically; Requires decoding or interpretation; Mapping is unambiguous once decoded.
Ex: "A as in Apple, B as in Bravo"; "Double seven"; "Twenty-three POINT five" → "point" is encoding for decimal
A3. Partial or Implicit Articulation
Def: The final value is not fully articulated, and must be inferred using conversational or external context.
Key: Surface form is incomplete; Correct value depends on prior context or shared state; Multiple completions may exist without context.
Ex: "The card ends in four five one two."; "Same number as before."
A4. Degraded or Ambiguous Articulation
Def: The surface realization permits multiple plausible parses, even when taken in isolation.
Key: Ambiguity is intrinsic to the surface form; Multiple numeric or lexical interpretations are valid; Ambiguity persists even with full local context.
Ex: "five three zero twenty four" → {53024, 530204, 500024}

ASSIGNMENT RULES:
1. Return EXACTLY ONE label.
2. When the value appears multiple times, classify based on the MOST COMPLEX realization.
3. In case of conflict, apply priority: A4 > A3 > A2 > A1 > A0

CRITICAL DISTINCTION - A0 vs A1:
A0 requires COMPOSITIONAL grammar: "two thousand" (compositional) vs "two zero zero zero" (enumeration). If digits/characters are spoken one-by-one or in small chunks → A1 (Fragmented). If value uses standard numeric/lexical composition as a single phrase → A0 (Compositional).

RULE OF THUMB:
Ask: "What is the MOST COMPLEX way the final value was articulated?"
→ Is it AMBIGUOUS with multiple valid parses even in isolation? → **A4**
Signals: Intrinsic ambiguity, multiple numeric/lexical interpretations, unclear boundaries
→ Is it INCOMPLETE and requires CONTEXT to complete? → **A3**
Signals: "ends in...", "same as before", masked digits, depends on prior state
→ Does it use INDIRECT ENCODING (phonetic, mnemonics, descriptions)? → **A2**
Signals: "A as in Apple", "double seven", "point five", deterministic decoding needed
→ Is it DIGIT-BY-DIGIT, CHARACTER-BY-CHARACTER, or FRAGMENTED? → **A1**
Signals: "five five five one two three", "a m i l i a", enumeration not composition
→ Is it COMPOSITIONAL, CONTINUOUS, and UNAMBIGUOUS in a SINGLE TURN? → **A0**
Signals: "two thousand one hundred", standard grammar, one valid parse, no enumeration
PRIORITY CHECK: If multiple seem applicable, pick the HIGHEST priority (A4 > A3 > A2 > A1 > A0).

Initiation System Prompt

You are an expert classifier for entity exchange behavior in customer-agent conversations.
Your task is to classify HOW an entity first enters the conversation (Initiation).
You must:

- Base decisions purely on the provided episode
- Not hallucinate or infer beyond what is given
- Return ONLY valid JSON in the required format

Initiation Classification User Prompt

Classify how the entity first enters the conversation.
TAXONOMY:
`{taxonomy}`
INPUT EPISODE:
`{episode}`
Return ONLY valid JSON in this format:

```
{
  "initiation": {
    "label": "<one of: I0, I1, I2>",
    "reason": "<brief explanation>"
  }
}
```

Evolution System Prompt

You are an expert classifier for entity exchange behavior in customer-agent conversations.
Your task is to classify HOW commitment to an entity value is constructed, reinforced, revised, or fails to resolve through conversational interaction (Evolution).
A commitment is an explicit verbal act proposing, asserting, or accepting a specific value as the intended content for the target entity.
You must:

- Base decisions purely on the provided episode
- Not hallucinate or infer beyond what is given
- Return ONLY valid JSON in the required format

Evolution Classification User Prompt

Classify how commitment to an entity evolves through interaction.
RULES:

1. Return EXACTLY ONE label. Priority: E6 > E5 > E4 > E3 > E2 > E1 > E0

LABELS:

- E0 = Stable (single commitment, no interaction)
- E1 = Reinforced (value REPEATED by self or other to confirm)
- E2 = Interleaved (backchannels only: “okay”, “mm-hm” — value NOT repeated)
- E3 = Self-Revised (self corrects with explicit marker)
- E4 = Other-Initiated Change (other CHANGES/CORRECTS value, not just confirms)
- E5 = Implicit Drift (value changes silently, no marker)
- E6 = Unresolved (no final commitment)

KEY: E1 vs E4 — E1 confirms SAME value, E4 CHANGES value.
TAXONOMY:
`{taxonomy}`
INPUT EPISODE:
`{episode}`
Return ONLY valid JSON:

```
{
  "evolution": {
    "label": "<E0|E1|E2|E3|E4|E5|E6>",
    "reason": "<brief explanation>"
  }
}
```

Articulation System Prompt

You are an expert classifier for entity exchange behavior in customer-agent conversations.
Your task is to classify HOW the entity value is expressed on the surface (Articulation).
You must:

- Base decisions purely on the provided episode
- Not hallucinate or infer beyond what is given
- Return ONLY valid JSON in the required format

Articulation Classification User Prompt

Classify how the entity value is expressed on the surface.
IMPORTANT RULES:

1. Return EXACTLY ONE label (not multiple).
2. When the value appears multiple times, classify based on the MOST COMPLEX realization (not necessarily the first).
3. If multiple labels seem applicable, apply priority: A4 > A3 > A2 > A1 > A0

TAXONOMY:
`{taxonomy}`
INPUT EPISODE:
`{episode}`
Return ONLY valid JSON in this format:

```
{
  "articulation": {
    "label": "<one of: A0, A1, A2, A3, A4>",
    "reason": "<brief explanation>"
  }
}
```

Zero Shot Entity Extraction Prompt

System Prompt:

You are an information extraction system.
Your task is to extract the value of the specified entity from the given conversation transcript. Return only the extracted value, without any explanation or additional text.
If the value is not present in the transcript, return "No evidence found".

User Prompt:

ENTITY:
`{entity}`

INSTRUCTION:
`{instruction}`

TRANSCRIPT:
`{transcript}`

Figure 6: Zero-shot entity extraction prompt. We run the experiment setup over 3 variations of $\{INSTRUCTION\}$. These variations are: "*Extract the entity from the provide conversation.*", "*What is the entity in this provided conversation?*", "*From the following conversation, find the entity and provide its value.*".

Error-Aware Entity Extraction System Prompt

You are a precise entity extraction system for conversational transcripts. Your task is to extract precise entity values from the transcript following the extraction principles below.
Important: Just return the entity value, do not include any additional text or explanation.

EXTRACTION PRINCIPLES
P1. Canonical Normalization
Transform all values to standardized forms:

- Numbers: digit form ("2259" not "two two five nine")
- Dates: YYYY-MM-DD or "DD Mon YYYY"
- Phone: digits only ("9876543210")
- Email: complete lowercase ("user@domain.com")
- Addresses: structured with all stated components

P2. Symbol Fidelity
Preserve exact character sequences—no substitutions, deletions, insertions, or transpositions. Verify numeric/alphanumeric identifiers character-by-character.
P3. Complete Span Coverage
Extract full entity boundaries. Partial fragments (incomplete addresses, truncated IDs, partial names) are invalid.
P4. Semantic Slot Accuracy
Assign values to correct entity types. Distinguish between similar slots (order_id vs customer_id, phone vs fax) using conversational context.
P5. Discourse State Resolution
When values are revised during conversation, extract only the FINAL committed value. Recognize correction markers: "actually", "sorry, I meant", "no wait", "changed to".
P6. Evidence Grounding
Extract only explicitly stated values. Return null for absent entities—never infer or fabricate.

OUTPUT
Return the canonical form of the entity, or null if not found.

Figure 7: Error-Aware Entity Extraction System Prompt

Error-Aware Entity Extraction User Prompt

Extract the entity from the transcript below.
Instruction:
`{instruction}`
Transcript:
`{transcript}`
Instructions:

- Return canonical form (digits for numbers, YYYY-MM-DD for dates)
- If value was corrected mid-conversation, return only the final version
- Extract complete span (full address/name/ID)
- Verify symbol accuracy before responding
- Return null if entity not mentioned
- If entity is not mentioned in the transcript, return "No evidence found"

Output Format:
Final entity value or "No evidence found". Do not include any additional text, explanation or reasoning.

Error Analysis System Prompt

You are an expert analyst evaluating errors made by entity extraction systems on conversational transcripts.
Your task is to assign exactly ONE error category to each prediction by comparing the model prediction against the ground truth and the provided conversation context.
You must strictly follow the provided error taxonomy, key properties, examples, and disambiguation rules.

- Do not invent new categories.
- Do not assign multiple labels.
- Select the most fundamental error that explains the failure.
- Return your decision in the specified JSON format.

Figure 8: Error Analysis System Prompt

Error Analysis User Prompt

You are given an entity extraction result. Your task is to analyze the ground truth and predicted entity value and assign an error category to the predicted entity value.
Target Entity: `{entity}`
Ground Truth Value: `{ground_truth}`
Model Prediction: `{prediction}`
Entity Context Snippet: `{context}`
ERROR TAXONOMY
E0. No Error (Canonical Match)
Key: Prediction is semantically equivalent to GT after normalization; Differences only in formatting; No symbols missing, added, or corrupted.
Ex: GT: 5/08/2025 → Pred: 5th aug 2025; GT: 9876543210 → Pred: 987-654-3210
E1. Surface Realization Error (Corruption)
Key: One or more symbols substituted, deleted, inserted, or reordered; Literal fidelity is broken.
Ex: GT: 12345 → Pred: 12354 (swap); GT: 324324 → Pred: 32432 (deletion)
E2. Canonicalization Failure
Key: All required symbols present; Not composed into canonical atomic form; Error in normalization, not symbol identity.
Ex: GT: 2259 → Pred: two two five nine; GT: 45.69 → Pred: 4569
E3. Span Underspecification
Key: Extracted span is strict subset of required entity; Insufficient to uniquely identify referent; Meaningful semantic subset, not corruption.
Ex: GT: 12B Maple Street, Bangalore → Pred: Maple Street
E4. Semantic Slot Misassignment
Key: Extracted value is correct itself; Assigned to wrong entity type or semantic role.
Ex: Task: extract client_id → Pred: phone number
E5. Discourse State Misalignment
Key: Value appears in dialogue; Belongs to intermediate/obsolete/superseded state; Final committed value is different; Includes rejected/corrected/repaired values.
Ex: "It was 54 earlier. Final is 51.5." → Pred: 54
E6. Complete Recall Failure
Key: No value extracted; Explicit and sufficient evidence present.
Ex: "My order number is A73921." → Pred: null
E7. Evidence Contradiction Error
Key: Explicit evidence exists for FINAL value; Prediction directly contradicts it; Predicted value does NOT appear in context.
Ex: Context: 12345 → Pred: 34350
E8. Ungrounded Entity Generation
Key: No lexical, semantic, or contextual evidence supports prediction; Value fabricated not misinterpreted; No similar entity in context.
Ex: No date mentioned → Pred: March 10, 2024

DISAMBIGUATION RULES
Rule 1: E7 vs E8
If explicit evidence exists and prediction contradicts → E7. If no evidence at all (fabricated) → E8. For proper nouns: if similar-sounding entity in context → E7; only E8 if no lexical/phonetic similarity.
Rule 2: E0 vs E2
If deterministic canonicalizer yields same value → E0. If symbols uncollapsed/unnormalized → E2. Formatting differences (hyphens, 2-digit vs 4-digit year) → E0.
Rule 3: E1 vs E2
Symbols substituted/deleted/inserted/reordered → E1. All symbols present but different format → E2. Date format with same symbols → E2. Decimal handling → E2. Digit-to-word → E2.
Rule 4: E1 vs E3
Numeric entities with missing digits → E1 (corruption). Meaningful semantic subset → E3. Key: E3 = valid but incomplete; E1 = corrupted.
Rule 5: E4 vs E5
Wrong entity type → E4. Correct type but wrong conversational state → E5. Value rejected/corrected/updated and model extracts old → E5.
Rule 6: E5 vs E7
Predicted value explicitly REJECTED/SUPERSEDED → E5. E7 applies when prediction contradicts FINAL value AND never appeared in transcript.
Rule 7: Address Handling
Minor word additions/omissions preserving addressability → E0. Reordering preserving meaning → E0. Missing critical components → E3. Corrupted words → E1. Unnormalized → E2.
Rule 8: GT Quality
Leading zeros missing in GT but present in prediction → E0. Verify GT correctness before E7.
NOTES
If multiple categories plausible, select EARLIEST FAILURE in pipeline. If insufficient evidence for E4/E7/E8 → "Can't determine". Dates: DD/MM/YYYY. Amounts: include decimal (45.69 not 4569).
OUTPUT FORMAT

```
{
  "error_category": "E#",
  "justification": "Reasoning with
    evidence from context."
}
```

Figure 9: Error Analysis System Prompt