EVIDENCE DRIFT: EVALUATING CONDITIONAL RELIABILITY IN GENERATIVE AFFECTIVE DIALOGUE UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

In generative affective dialogue understanding, a correct emotion label can accompany an incorrect sentiment label, cause judgment, or support strategy. Marginal field scores and schema checks do not isolate this conditional failure. We call it *evidence drift* and quantify it with the conditional evidence drift rate (CEDR). We study paired prediction on MELD, RECCON, and ESConv using Qwen2.5-7B-Instruct. Under supervised joint generation, CEDR is 3.1%, 13.3%, and 36.2% for sentiment, cause judgments, and support strategies, respectively. On ESConv, this risk coexists with a strategy macro-F1 only 0.45 percentage points below that of an evidence-only expert. ESConv CEDR is 73.4% under label-constrained zero-shot generation. The failure is also observed with Mistral-7B-Instruct-v0.3 on RECCON and ESConv. We introduce Conditional Evidence Verification and Deferral (CEVD), which scores the original evidence given the dialogue and predicted emotion. Using a threshold selected on development data, it withholds evidence with weak support while preserving the primary prediction. On the held-out ESConv test set with Mistral, CEVD retains 89.3% evidence coverage and lowers CEDR among retained predictions from 36.4% to 32.6%. It outperforms random deferral at the same coverage ($p = 0.0002$). These findings support reporting conditional evidence error alongside marginal field performance and evidence coverage.

1 INTRODUCTION

Affective dialogue understanding extends beyond single-label emotion classification. MELD supports contextual emotion and sentiment recognition in multi-party conversations (Poria et al., 2019). RECCON annotates whether an utterance causes a target emotion (Poria et al., 2021). ESConv associates emotional-support responses with interpretable strategy labels (Liu et al., 2021). These datasets provide complementary annotations for studying reliability beyond a single emotion label. We construct paired prediction tasks from these annotations and ask whether the associated field remains correct when the jointly predicted emotion is correct.

Large language models (LLMs) make such multi-field prediction increasingly natural. InstructERC reformulates conversational emotion recognition as instruction-following, retrieval-augmented generation (Lei et al., 2023). Speaker information has also been incorporated into LLM-based emotion recognition (Fu et al., 2025). AffectGPT further extends generative affect understanding to multimodal settings (Lian et al., 2025). A model can consequently return an emotion together with sentiment, cause, or strategy in a single structured response.

Existing structured-generation methods primarily address syntactic validity. Grammar-constrained decoding can enforce task-specific output formats (Geng et al., 2023), but valid structure does not establish whether each predicted field agrees with its reference label. Models may predict the correct emotion without relying on human-aligned triggers (Singh et al., 2024). Reasoning about support strategies also remains challenging in emotional-support systems (Zhang et al., 2024). Thus, primary-label correctness should not be treated as evidence that the associated field is reliable.

Figure 1 illustrates this failure: the model correctly predicts *anxiety* but produces *self-disclosure* instead of the reference strategy *affirmation and reassurance*. Both predictions are legal labels,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

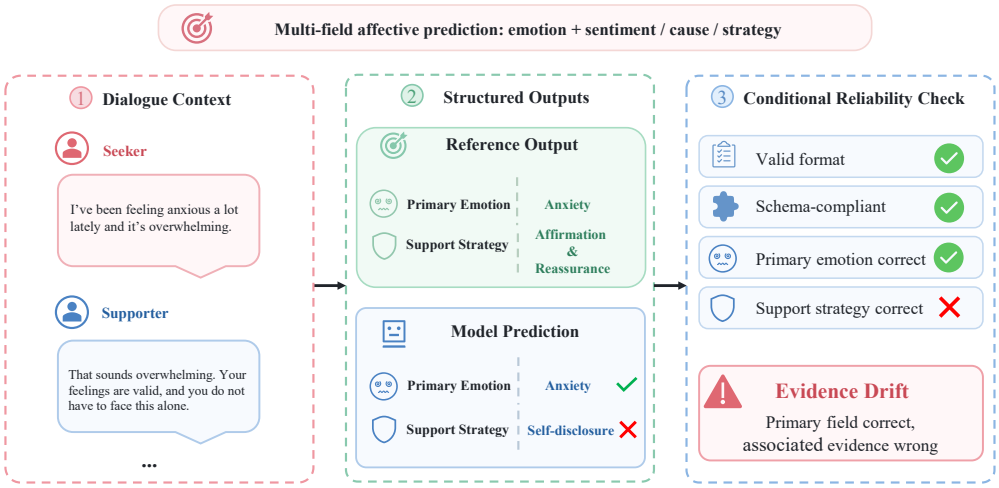

Figure 1: A schematic example of evidence drift in emotion–strategy prediction. The output is parseable and schema-compliant, and the primary emotion is correct, but the predicted support strategy differs from the reference strategy.

so structural validation alone cannot expose the error. Additional error analyses are provided in Appendix I.

We call this conditional prediction failure **evidence drift**: the primary affective field is correct, but its associated evidence field is incorrect. We use *evidence field* operationally for a structured field that contextualizes, explains, or characterizes the primary judgment. Sentiment contextualizes emotion in MELD, a cause judgment provides explanatory information in RECCON, and a strategy label characterizes the observed supporter response in ESConv. The term does not imply causal proof or access to the model's latent reasoning. Evidence error includes missing or illegal values; structural diagnostics distinguish these cases from incorrect predictions within the legal label space.

Marginal field scores summarize performance over the full test set; they do not locate evidence errors within the subset where the primary prediction is correct. Joint exact match measures complete correctness but does not express this conditional risk on its own. We report evidence drift rate (EDR) over all instances and conditional evidence drift rate (CEDR) within that subset. These metrics make the remaining evidence error explicit without replacing standard field evaluation.

We evaluate emotion–sentiment prediction on MELD, emotion–cause prediction on RECCON, and emotion–strategy prediction on ESConv. The evaluation covers label-constrained zero-shot generation, task-specific adaptation, expert composition, and two instruction-tuned model families. Under supervised joint generation, CEDR reaches 3.1%, 13.3%, and 36.2% on the three datasets, respectively. Evidence drift also persists across the other evaluated settings, showing that strong primary-field performance can coexist with substantial residual evidence error. Because the datasets differ in domain, label distribution, and annotation design, this cross-dataset ordering is descriptive rather than a controlled task comparison.

We further introduce **Conditional Evidence Verification and Deferral** (CEVD), a lightweight selective mechanism that preserves the primary prediction while withholding evidence with weak conditional support. On held-out Mistral–ESConv, CEVD retains 89.3% evidence coverage while reducing CEDR among retained predictions from 36.4% to 32.6%, outperforming matched random deferral.

Our main contributions are as follows:

- We define **evidence drift** as evidence error conditioned on a correct primary prediction and report structural diagnostics separately.
- We operationalize evidence drift with EDR and CEDR and evaluate it across three affective field pairs, multiple prediction settings, and two model families.

- We introduce **CEVD**, which preserves the primary prediction while selectively withholding unreliable evidence and lowers conditional error relative to matched random deferral.

2 Related Work

Generative affective dialogue understanding. Affective dialogue understanding was initially studied as emotion or sentiment prediction from conversational context. Recent work has moved this setting toward instruction following and generative modeling. InstructERC formulates emotion recognition in conversation as a retrieval-augmented generation task (Lei et al., 2023). Generated external knowledge can further improve conversational emotion recognition (Tu et al., 2023), while other studies directly examine the affect recognition ability of large language models (Feng et al., 2024). Speaker characteristics and explicit emotion clues provide additional context for LLM-based prediction (Fu et al., 2025; Shen et al., 2025b). AffectGPT extends this line to multimodal generative understanding (Lian et al., 2025). EmoBench evaluates emotional understanding and application (Sabour et al., 2024). Jafari et al. separately assess emotion understanding, generation, and consistency (Jafari et al., 2026). These studies show that emotion recognition alone does not characterize affective competence. Our question concerns a different evaluation unit: within a paired structured prediction, how often is the associated label wrong when the emotion label is correct? We examine this conditional error across sentiment, cause judgments, and support strategies.

Cause judgments and support strategies. Cause judgments and support strategies play different roles beyond the primary emotion label. RECCON frames conversational cause recognition as deciding whether a candidate utterance causes a target emotion (Poria et al., 2021). SemEval-2024 extends emotion-cause analysis to multimodal conversations (Wang et al., 2024), and ECC connects cause information with empathetic response generation (He et al., 2025). Trigger-focused analysis also shows that a correct emotion prediction need not rely on evidence aligned with human annotations (Singh et al., 2024). This work demonstrates the value of cause information, but cause accuracy is normally reported as a separate task score.

ESConv assigns a support-strategy label to each observed supporter response (Liu et al., 2021). ESCoT introduces an interpretable reasoning process for emotional support generation (Zhang et al., 2024). Other studies examine preference bias across strategies (Kang et al., 2024), evaluate the quality of LLM-based support conversations (Zhao et al., 2024), or model strategy prediction from mixed emotions and discourse dynamics (Wan et al., 2025). These approaches clarify how causes and strategies should be modeled or evaluated, yet they usually consider the fields in isolation or assess the response as a whole. They do not ask whether a cause judgment or strategy label is still correct when it is generated together with an already-correct primary field. We formulate that remaining error as evidence drift and evaluate it consistently across sentiment, cause, and strategy fields.

Structured-output reliability. A separate body of work focuses on whether generated outputs follow a requested structure. Grammar-constrained decoding enforces task-specific formats at inference time (Geng et al., 2023). SO-LM learns to generate structured objects that conform to complex schemas (Tavanaei et al., 2024). Format restrictions can also change task performance rather than only the surface form of an answer (Tam et al., 2024). More recent methods improve schema following through reinforcement learning (Lu et al., 2025), explicit output structuring (Shen et al., 2025a), or JSON-based instruction tuning (Gao et al., 2025). Causal analysis further suggests that the effect of a structured format depends on the task and instruction (Yuan et al., 2026). These methods address parseability, required fields, and legal values. Evidence drift concerns a different layer of reliability: an output may satisfy every structural requirement while pairing a correct primary field with incorrect evidence. Our evaluation therefore reports structural diagnostics separately from conditional semantic error.

Verification and selective prediction. Research on hallucination has established that fluent output may still be unsupported (Ji et al., 2023). SelfCheckGPT uses agreement across sampled responses to detect unstable content (Manakul et al., 2023), while Chain-of-Verification separates an initial answer from a later checking stage (Dhuliawala et al., 2024). These methods mainly verify free text

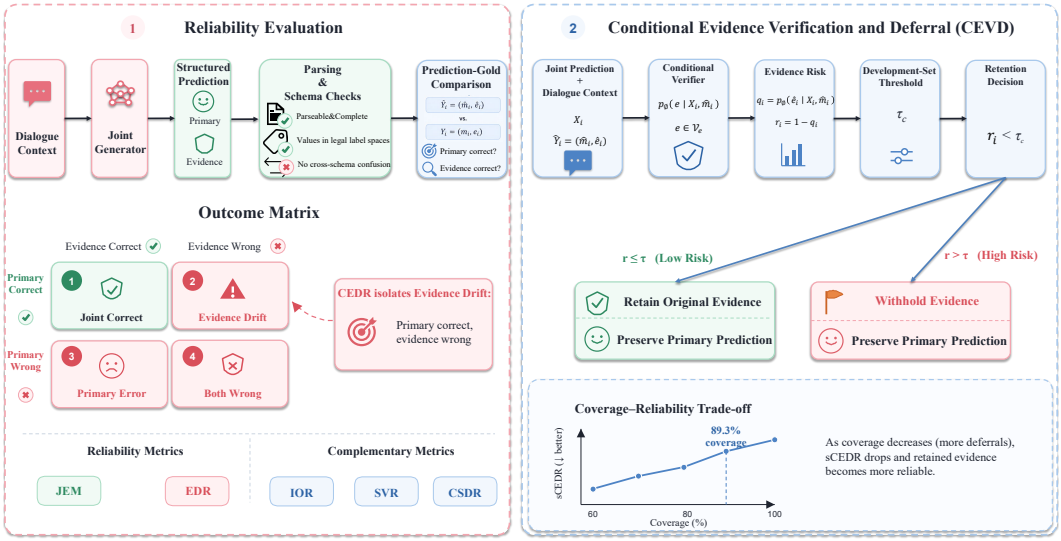

Figure 2: Overview of evidence-drift evaluation and CEVD. CEVD scores the original evidence prediction and either releases or withholds it using a threshold selected on development data. The primary prediction remains unchanged.

or factual claims. In our setting, the model returns fields with predefined labels, and the relevant error is conditioned on the correctness of another field.

Selective prediction addresses a complementary question: when should a system abstain instead of releasing a high-risk prediction? Selective classification formalizes the relationship between retained coverage and prediction risk (Geifman & El-Yaniv, 2017). SelectiveNet learns prediction and rejection jointly (Geifman & El-Yaniv, 2019), while later work provides objectives for multiclass abstention (Mao et al., 2024). Calibration research cautions that a neural confidence score is not automatically a probability of correctness (Guo et al., 2017). Selective generation applies abstention to language-model outputs (Lee et al., 2024). Partial abstention has also been studied in multilabel classification, where a model predicts some labels and abstains on others (Nguyen & Hüllermeier, 2020). Structured abstention allows predictions to be withheld for subsets of related outputs (Finocchiaro et al., 2025). CEVD studies a specific case: the primary prediction is fixed, the original evidence is scored using that prediction and the dialogue, and retained evidence is evaluated conditional on primary correctness. Our focus is the evidence error that remains after a correct primary prediction and how selective release changes that error.

3 Methodology

Figure 2 summarizes the two parts of our methodology. We first define evidence drift and the metrics used to measure conditional evidence error and identify associated structural failures. We then introduce Conditional Evidence Verification and Deferral (CEVD), which scores the evidence produced by a joint generator and withholds it when conditional support is weak. The primary prediction is never changed.

3.1 Task Formulation and Evidence Drift

For instance i, let X_i denote the dialogue context. The gold structured annotation and model prediction are

$$Y_i = (m_i, e_i), \qquad \hat{Y}_i = (\hat{m}_i, \hat{e}_i), \tag{1}$$

where $m_i \in \mathcal{V}_m$ is the primary affective label and $e_i \in \mathcal{V}_e$ is the associated evidence label. We use *evidence field* as an operational umbrella term rather than a claim that every field is causal evidence. Sentiment contextualizes the emotion in MELD, the binary cause judgment provides explanatory information in RECCON, and the support-strategy label describes the observed supporter response

in ESConv. Dataset-specific field semantics, complete label inventories, and JSON output schemas are provided in Appendices A–B.

All outputs are processed by the same deterministic parser and label normalizer. After this step, each predicted field belongs to its legal label space or takes the sentinel value $\perp$. The sentinel represents an unavailable or illegal field value arising from an unparsable output, a missing key, or a label-space violation. It is counted as incorrect when compared with a valid reference label. No instance is removed from evaluation. Evidence drift is defined by

$$d_i = \mathbb{I}(\hat{m}_i = m_i \wedge \hat{e}_i \neq e_i)\,, \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Thus, a missing or illegal evidence value counts as drift when the primary field is correct. If the primary field is $\perp$ or incorrect, the instance does not enter the CEDR conditioning set.

We additionally record two structural diagnostics. Invalid output rate (IOR) counts outputs whose JSON object or required fields cannot be recovered. Schema violation rate (SVR) counts parseable, field-complete outputs containing an illegal field value. Both rates use the full evaluation set. Under this convention, IOR and SVR are disjoint, but either may overlap with evidence drift when the primary prediction is correct. Cross-schema confusion is a subtype of SVR. Detailed accounting rules are provided in Appendix C.

3.2 Reliability-Oriented Evaluation

We report primary-field accuracy and evidence macro-F1. These marginal scores are complemented by three metrics for joint and conditional reliability:

$$\mathrm{JEM} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{m}_i = m_i \wedge \hat{e}_i = e_i), \qquad \mathrm{EDR} = \frac{1}{N}\sum_{i=1}^{N} d_i, \qquad \mathrm{CEDR} = \frac{\sum_{i=1}^{N} d_i}{\sum_{i=1}^{N}\mathbb{I}(\hat{m}_i = m_i)}. \tag{3}$$

JEM measures complete structured correctness. EDR measures the prevalence of evidence drift over the full test set. CEDR measures the remaining evidence error within the primary-correct subset and is the main diagnostic metric in this study.

Every prediction with a correct primary label is either jointly correct or an evidence-drift case. Thus, $\mathrm{Acc}_m = \mathrm{JEM} + \mathrm{EDR}$ and, for $\mathrm{Acc}_m > 0$, $\mathrm{CEDR} = 1 - \mathrm{JEM}/\mathrm{Acc}_m$. CEDR is therefore a conditional summary of joint correctness. We report it with primary accuracy and the number of conditioning instances. We also report $\Delta F1_e = F1_e^{\mathrm{expert}} - F1_e^{\mathrm{joint}}$. This compares marginal evidence performance between models; it is not the same quantity as conditional evidence error.

3.3 Conditional Evidence Verification and Deferral

Selection objective. Given a fixed joint generator, CEVD chooses which evidence predictions to release without changing the primary output. The goal is to reduce evidence error within the retained subset at a useful coverage. This follows selective prediction (Geifman & El-Yaniv, 2017), but the conditioning on primary correctness is used only for evaluation. CEVD does not observe test reference labels. Its release decision uses the dialogue, the predicted labels, and a threshold selected on the development set.

Conditional verifier. Let $s_\phi(X, m, e)$ be a verifier score for evidence candidate e conditioned on dialogue context X and primary label m. For every legal candidate $e \in \mathcal{V}_e$, CEVD forms the listwise support distribution

$$p_\phi(e \mid X, m) = \frac{\exp s_\phi(X, m, e)}{\sum_{e' \in \mathcal{V}_e} \exp s_\phi(X, m, e')}. \tag{4}$$

The verifier is trained on the training split with

$$\mathcal{L}_{\mathrm{ver}} = -\frac{1}{N_{\mathrm{tr}}}\sum_{i=1}^{N_{\mathrm{tr}}} \log p_\phi(e_i \mid X_i, m_i). \tag{5}$$

The listwise objective compares all legal evidence labels for the same context. Training uses the gold primary label m_i, whereas inference uses the generator prediction $\hat{m}_i$. These labels coincide

on the primary-correct subset used to compute selective CEDR. Candidate construction, verifier inputs, implementation details, and leakage controls are provided in Appendix D.

Risk score and evidence retention. Let $g_i = \mathbb{I}(\hat{m}_i \in \mathcal{V}_m \wedge \hat{e}_i \in \mathcal{V}_e)$ indicate whether both generated labels are legal. For eligible predictions, CEVD scores the original evidence. Otherwise, support is set to zero:

$$q_i = \begin{cases} p_\phi(\hat{e}_i \mid X_i, \hat{m}_i), & g_i = 1, \\ 0, & g_i = 0, \end{cases} \qquad r_i = 1 - q_i. \tag{6}$$

The retention decision is $a_i(\tau) = g_i \mathbb{I}(r_i \leq \tau)$. The primary prediction is never revised. Evidence is released only when $a_i(\tau) = 1$ and is otherwise withheld. Withholding is an abstention action, not a label in $\mathcal{V}_e$. CEVD does not replace the original evidence with the verifier's highest-scoring alternative.

Threshold selection. For a target coverage c, we set τ_c to the empirical c-quantile of development risks and apply the same threshold to the test set. We use $c = 0.90$ as the main setting and $c = 0.80$ as an additional setting. Because of score ties and differences between the development and test distributions, realized coverage need not equal c. We therefore report actual test coverage. This procedure selects a threshold; it neither calibrates scores as correctness probabilities nor guarantees a specified test error rate.

Coverage-aware evaluation. For a fixed threshold, we report overall evidence coverage, coverage within the primary-correct subset, and selective CEDR:

$$\begin{gathered} \mathrm{Cov}(\tau) = \frac{1}{N} \sum_i a_i(\tau), \\ \mathrm{PCov}(\tau) = \frac{\sum_i \mathbb{I}(\hat{m}_i = m_i) a_i(\tau)}{\sum_i \mathbb{I}(\hat{m}_i = m_i)}, \qquad \mathrm{sCEDR}(\tau) = \frac{\sum_i d_i a_i(\tau)}{\sum_i \mathbb{I}(\hat{m}_i = m_i) a_i(\tau)}. \end{gathered} \tag{7}$$

We also report drift capture and deferral precision:

$$\mathrm{Capture}(\tau) = \frac{\sum_i d_i [1 - a_i(\tau)]}{\sum_i d_i}, \qquad \mathrm{DPrec}(\tau) = \frac{\sum_i d_i [1 - a_i(\tau)]}{\sum_i \mathbb{I}(\hat{m}_i = m_i)[1 - a_i(\tau)]}. \tag{8}$$

Any rate with a zero denominator is undefined. Capture measures the fraction of all drift cases withheld by CEVD. DPrec measures the drift rate among withheld primary-correct instances. To separate targeted selection from the mechanical effect of lower coverage, we compare CEVD with random deferral at exactly the same number of retained test predictions. The matched-random protocol, Monte Carlo intervals, and empirical p-value calculation are described in Appendix G. Additional operating points and risk-score comparisons are provided in Appendix H.

4 Experiments

The experiments examine conditional evidence error, its variation across adaptation, composition, and model settings, and whether CEVD improves selective evidence release over random deferral.

4.1 Experimental Setup

Datasets and field pairs. We evaluate 11,643 test instances from three public datasets. MELD provides utterance-level emotion and sentiment annotations and contains 2,610 test instances (Poria et al., 2019). RECCON pairs a target emotion with a binary cause judgment for a candidate utterance. We evaluate 7,224 candidate instances from its DailyDialog portion (Poria et al., 2021). ESConv pairs the seeker's emotion with the strategy label assigned to the observed supporter response and contains 1,809 test instances (Liu et al., 2021). These datasets differ in domain, annotation protocol, label distribution, and task difficulty. We therefore interpret their CEDR differences descriptively rather than as causal effects of field type.

Table 1: Main evidence-drift results under Qwen2.5-7B Joint-LoRA. Expert Evid. F1 is obtained from the evidence-only expert, and $\Delta F1_e$ is Expert F1 minus Joint F1. CEDR measures evidence error conditioned on a correct primary prediction.

Dataset	Field Pair	Primary Acc.	Joint Evid. F1	Expert Evid. F1	$\Delta F1_e$	JEM	EDR	CEDR [95% CI]
MELD	emotion–sentiment	0.7019	0.7382	0.7462	0.0080	0.6805	0.0215	0.0306 [0.0228, 0.0383]
ESConv	emotion–strategy	0.6368	0.5796	0.5841	0.0045	0.4063	0.2305	0.3620 [0.3336, 0.3899]
RECCON	emotion–cause	0.8779	0.8081	0.8118	0.0037	0.7615	0.1164	0.1326 [0.1247, 0.1407]

Models and diagnostic settings. We use Qwen2.5-7B-Instruct (Qwen et al., 2024) as the primary backbone and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) for cross-family comparison on ESConv and RECCON. Joint-LoRA generates the primary and evidence fields in one JSON object. Primary-only and evidence-only experts provide field-wise reference results. EP-Compose combines the predictions of both experts. Hybrid-Compose keeps the joint model's primary prediction and replaces only its evidence prediction with the output of the evidence-only expert. We also evaluate label-constrained zero-shot generation with the complete legal label sets and the required JSON schema.

Training and evaluation protocol. The generative models are adapted with LoRA (Hu et al., 2022) under 4-bit quantization following the QLoRA setting (Dettmers et al., 2023). Each adapter is trained for one epoch, and inference uses deterministic decoding. Expert composition uses two separately trained adapters and is a diagnostic comparison, not a comparison at equal total computational cost. We normalize casing, whitespace, and underscores but perform no semantic label remapping. The CEVD verifier is initialized from DeBERTaV3-base (He et al., 2023) and trained only on the training split; its threshold is selected only on the development split.

We report primary accuracy, evidence macro-F1, joint exact match (JEM), evidence drift rate (EDR), and conditional evidence drift rate (CEDR). Invalid-output and schema-violation statistics are monitored separately. Diagnostic intervals use 1,000 instance-level bootstrap resamples. CEVD is compared with 5,000 random-deferral trials at exactly matched test coverage. We additionally compare support, entropy, and disagreement using AUROC and AUPRC within the subset with correct primary predictions, with evidence drift as the positive class. For the two development coverage targets, we report actual test coverage alongside sCEDR. Prompts and normalization rules are provided in Appendix B, implementation details in Appendix D, and statistical procedures in Appendix G.

4.2 Evidence Drift Across Affective Field Pairs

Table 1 reports CEDR of 3.1%, 13.3%, and 36.2% on MELD, RECCON, and ESConv, respectively. Figure 3(a) shows instance-level bootstrap intervals, which do not account for within-dialogue dependence. The cross-dataset ordering is descriptive because the tasks are not controlled. Counts and additional intervals are provided in Appendix E.

Parsing does not establish label correctness. All supervised Qwen Joint-LoRA outputs in this comparison are parseable and field-complete (IOR is zero), so parsing failures and missing keys do not explain the drift counts. This does not establish label-space compliance; accounting rules and IOR counts are in Appendix C.

Marginal and conditional evaluation. On ESConv, joint and expert strategy macro-F1 are 57.96% and 58.41%, a gap of 0.45 percentage points. Joint CEDR is 36.2%. The F1 gap compares marginal performance over all test instances; CEDR measures error within the primary-correct subset. A small marginal gap therefore does not establish low conditional error.

4.3 Adaptation, Composition, and Model Family

Effect of task-specific adaptation. On ESConv, CEDR is 73.44% under label-constrained zero-shot generation and 36.20% after Joint-LoRA adaptation. On RECCON, the corresponding rates are

Table 2: Comparison of adaptation settings and model families. Primary accuracy and evidence macro-F1 are measured over the full test set. Each model's CEDR is computed on its own subset of instances with correct primary predictions. Zero-shot prompts include all legal labels and the required JSON schema.

Dataset	Setting	Primary Acc.	Evid. F1	CEDR
ESConv	Qwen zero-shot	0.3101	0.2068	0.7344
	Qwen Joint-LoRA	0.6368	0.5796	0.3620
	Mistral Joint-LoRA	0.6247	0.5789	0.3637
RECCON	Qwen zero-shot	0.7119	0.5521	0.2487
	Qwen Joint-LoRA	0.8779	0.8081	0.1326
	Mistral Joint-LoRA	0.8331	0.7751	0.1554

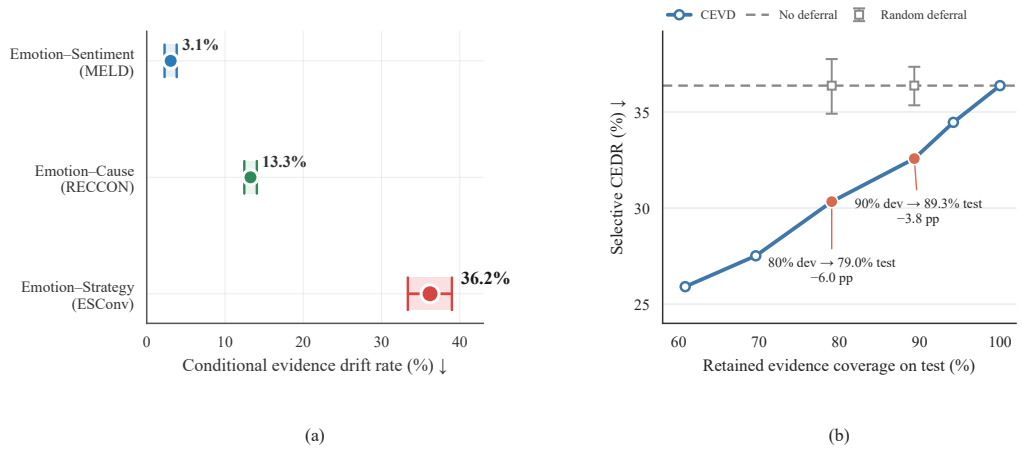

Figure 3: Conditional evidence risk and selective control. (a) CEDR under Qwen Joint-LoRA with 95% instance-level bootstrap intervals. (b) Selective CEDR versus evidence coverage on ESConv with Mistral. The dashed line denotes no deferral. Gray squares show matched random deferral, with error bars indicating central 95% intervals across random trials.

24.87% and 13.26%. Adapted models therefore have lower conditional error in these settings. Each rate uses the instances for which that model predicts the primary field correctly. The comparison reflects both evidence predictions and changes in the composition of that subset, rather than a paired estimate on identical instances.

Backbone comparison. Qwen and Mistral obtain similar CEDR values on ESConv: 0.3620 and 0.3637, respectively. On RECCON, their CEDR values are 0.1326 and 0.1554. Evidence drift is therefore observed with both evaluated model families, although the comparison is limited to two 7B instruction-tuned backbones.

Evidence replacement with the primary prediction fixed. Hybrid-Compose retains Qwen Joint-LoRA's primary predictions, so both settings are evaluated on the same 1,152 ESConv instances with a correct primary label. Replacing the original evidence with the evidence expert changes CEDR from 36.20% to 36.46%, approximately 0.26 percentage points higher. The point estimates therefore do not show an improvement from unconditional replacement. EP-Compose yields CEDR of 36.77% on ESConv and 13.74% on RECCON. It uses the primary expert, so its conditioning subset need not match Joint-LoRA's. The reported composition results are in Appendix F. We next examine selective release, which withholds some evidence without replacing its predicted label.

Table 3: CEVD results with Mistral-7B on the held-out ESConv test set. Cov. is overall evidence coverage, and P-Cov. is coverage within the primary-correct subset. sCEDR is computed over retained primary-correct predictions, and ΔsCEDR is its absolute change from no deferral. Capture is the proportion of drift cases withheld by CEVD, and Def. Prec. is the drift rate among withheld primary-correct instances. Subscripts denote development-set coverage targets.

Policy	Cov.	P-Cov.	sCEDR	ΔsCEDR	Capture	Def. Prec.
No deferral	1.000	1.000	0.3637	–	–	–
CEVD_{90}	0.893	0.894	0.3257	−0.0380	0.1995	0.6833
CEVD_{80}	0.790	0.796	0.3033	−0.0604	0.3358	0.6000

4.4 Selective Evidence Risk Control

We evaluate CEVD on Mistral–ESConv as a focused proof of concept. ESConv has the highest supervised CEDR and a fine-grained strategy space, providing a stringent test of selective evidence release while the primary prediction remains unchanged.

Coverage and conditional error. At the 90% development target, CEVD retains 89.33% of all evidence predictions and 89.38% within the primary-correct subset, lowering sCEDR from 36.37% to 32.57% (Table 3). At the 80% target, coverage is 79.05% and sCEDR is 30.33%. Withheld primary-correct predictions have error rates of 68.33% and 60.00%, respectively, approximately 1.88 and 1.65 times baseline CEDR. Capture and enrichment calculations are provided in Appendix H.

Comparison with random deferral. CEVD outperforms random deferral at both matched retained counts (5,000 trials; one-sided empirical $p = 0.0002$). Figure 3(b) shows the coverage and conditional error tradeoff; Appendix G details the test.

Comparison of risk scores. Support, entropy, and disagreement yield AUROC point estimates ranging from 0.7078 to 0.7093. Their different test coverages preclude claims of superiority at matched coverage (Appendix H, Tables 18–20).

Error characteristics. ESConv drift includes confusions between legal strategy labels. In RECCON, missed causes account for more observed drift errors than false cause assignments. Appendix I provides the error breakdown, and Appendix C details structural accounting.

Scope and limitations. Dataset differences preclude causal conclusions about field type. CEVD is tested only on ESConv with Mistral and one verifier. The score comparison neither isolates the primary label's contribution nor establishes superiority at matched coverage. Bootstrap intervals exclude dialogue dependence and training-seed variation. Results measure agreement with benchmark annotations, not clinical safety or whether a strategy is uniquely appropriate.

5 Conclusion

We study evidence drift: an incorrect associated field despite a correct primary prediction. Experiments on MELD, RECCON, and ESConv show that marginal evidence scores do not directly characterize this conditional error. On ESConv with Mistral, CEVD retains 89.3% of evidence predictions and reduces sCEDR from 36.4% to 32.6%, outperforming matched random deferral without changing primary predictions. These findings support reporting conditional error alongside field performance and evidence coverage. Broader validation across tasks and model families remains necessary.

AI use statement

We used ChatGPT to help refine the conceptual framing and metric definitions, provide feedback on methodology and experimental design, and assist with interpreting reported results. It also assisted with drafting and revising text and code, literature search, reference formatting, schematic figures,

and LaTeX preparation. The authors checked the cited sources and mathematical definitions, tested the AI-assisted code used in the experiments, and verified the reported numerical results against the corresponding saved experimental outputs. Figure 1 is illustrative rather than a dataset record. The authors take responsibility for the final manuscript and associated artifacts, including material prepared with AI assistance.

Ethics statement

This study evaluates predictions against annotations in existing dialogue datasets. Emotion labels are not clinical diagnoses, and agreement with a support-strategy annotation does not establish that a response is safe or uniquely appropriate. CEVD withholds selected evidence predictions; it is not a validated clinical intervention or a guarantee of safe emotional support. The motivating dialogue example is explicitly schematic.

Reproducibility statement

Section 3 defines the evaluation metrics and CEVD. Appendices A–D describe the task construction, output conventions, and training protocol. Appendices E–H provide the reported results and statistical procedures. The resampling analyses use fixed predictions and do not measure variation across independent training runs.

References

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115, 2023. doi: 10.52202/075280-0441. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.212. URL https://aclanthology.org/2024.findings-acl.212/.

Shutong Feng, Guangzhi Sun, Nurul Lubis, Wen Wu, Chao Zhang, and Milica Gašić. Affect recognition in conversations using large language models. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 259–273, Kyoto, Japan, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.sigdial-1.23. URL https://aclanthology.org/2024.sigdial-1.23/.

Jessie Finocchiaro, Rafael Frongillo, and Enrique Nueve. Structured prediction with abstention via the Lovász hinge. arXiv preprint arXiv:2505.06446, 2025. URL https://arxiv.org/abs/2505.06446.

Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Lili Shan, Yulin Wu, and Bingquan Liu. LaERC-S: Improving LLM-based emotion recognition in conversation with speaker characteristics. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 6748–6761, Abu Dhabi, UAE, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.451/.

Chang Gao, Wenxuan Zhang, Guizhen Chen, and Wai Lam. JsonTuning: Towards generalizable, robust, and controllable instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24029–24055, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.1232. URL https://aclanthology.org/2025.findings-acl.1232/.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10932–10952, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.674. URL https://aclanthology.org/2023.emnlp-main.674/.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

Yuanyuan He, Yongsen Pan, Wei Li, Jiali You, Jiawen Deng, and Fuji Ren. ECC: An emotion-cause conversation dataset for empathy response. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6000–6017, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.306. URL https://aclanthology.org/2025.emnlp-main.306/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Sadegh Jafari, Els Lefever, and Veronique Hoste. Disentangling emotion understanding and generation in large language models. In *The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis (WASSA 2026)*, pp. 161–171, Rabat, Morocco, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.wassa-1.14. URL https://aclanthology.org/2026.wassa-1.14/.

Ziwei Ji, Nayeon Lee, Rita Maria Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730. URL https://dl.acm.org/doi/10.1145/3571730.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. arXiv preprint arXiv:2310.06825, 2023. URL https://arxiv.org/abs/2310.06825.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15232–15261, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.813. URL https://aclanthology.org/2024.acl-long.813/.

Minjae Lee, Kyungmin Kim, Taesoo Kim, and Sangdon Park. Selective generation for controllable language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 50494–50527. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5a6815122f533193a022cbc41786c1cc-Abstract-Conference.html.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. InstructERC: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. arXiv preprint arXiv:2309.11911, 2023. URL https://arxiv.org/abs/2309.11911.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36993–37014. PMLR, 2025. URL https://proceedings.mlr.press/v267/lian25a.html.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. 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)*, pp. 3469–3483, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.269. URL https://aclanthology.org/2021.acl-long.269/.

Yaxi Lu, Haolun Li, Xin Cong, Zhong Zhang, Yesai Wu, Yankai Lin, Zhiyuan Liu, Fangming Liu, and Maosong Sun. Learning to generate structured output with schema reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4905–4918, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.243. URL https://aclanthology.org/2025.acl-long.243/.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Theoretically grounded loss functions and algorithms for score-based multi-class abstention. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4753–4761. PMLR, 2024.

Vu-Linh Nguyen and Eyke Hüllermeier. Reliable multilabel classification: Prediction with partial abstention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5264–5271, 2020. doi: 10.1609/aaai.v34i04.5972. URL https://ojs.aaai.org/index.php/AAAI/article/view/5972.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050/.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, Alexander Gelbukh, and Rada Mihalcea. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332, 2021. doi: 10.1007/s12559-021-09925-7. URL https://doi.org/10.1007/s12559-021-09925-7.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024. URL https://arxiv.org/abs/2412.15115.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5986–6004, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.326. URL https://aclanthology.org/2024.acl-long.326/.

Zhengyuan Shen, Darren Yow-Bang Wang, Soumya Smruti Mishra, Zhichao Xu, Yifei Teng, and Haibo Ding. SLOT: Structuring the output of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 472–491, Suzhou (China), 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-industry.32. URL https://aclanthology.org/2025.emnlp-industry.32/.

Zhiyu Shen, Yunhe Pang, Yanghui Rao, and Jianxing Yu. CoE: A clue of emotion framework for emotion recognition in conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23548–23563, Vienna, Austria, 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1148. URL https://aclanthology.org/2025.acl-long.1148/.

Smriti Singh, Cornelia Caragea, and Junyi Jessy Li. Language models (mostly) do not consider emotion triggers when predicting emotion. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 603–614, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.51. URL https://aclanthology.org/2024.naacl-short.51/.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1218–1236, Miami, Florida, US, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.91. URL https://aclanthology.org/2024.emnlp-industry.91/.

Amir Tavanaei, Kee Kiat Koo, Hayreddin Ceker, Shaobai Jiang, Qi Li, Julien Han, and Karim Bouyarmane. Structured object language modeling (SO-LM): Native structured objects generation conforming to complex schemas with self-supervised denoising. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 821–828, Miami, Florida, US, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.62. URL https://aclanthology.org/2024.emnlp-industry.62/.

Geng Tu, Bin Liang, Bing Qin, Kam-Fai Wong, and Ruifeng Xu. An empirical study on multiple knowledge from ChatGPT for emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12160–12173, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.813. URL https://aclanthology.org/2023.findings-emnlp.813/.

Chenwei Wan, Matthieu Labeau, and Chloé Clavel. EmoDynamiX: Emotional support dialogue strategy prediction by modelling mixed emotions and discourse dynamics. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1678–1695, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.81. URL https://aclanthology.org/2025.naacl-long.81/.

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. SemEval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pp. 2039–2050, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.semeval-1.277. URL https://aclanthology.org/2024.semeval-1.277/.

Han Yuan, Yue Zhao, Li Zhang, Wuqiong Luo, and Zheng Ma. Quantifying the impact of structured output format on large language models through causal inference. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 1771–1795, Rabat, Morocco, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-eacl.91. URL https://aclanthology.org/2026.findings-eacl.91/.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. ESCoT: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13395–13412, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.723. URL https://aclanthology.org/2024.acl-long.723/.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, Zhixu Li, Yan Teng, Yanghua Xiao, and Yingchun Wang. ESC-eval: Evaluating emotion support conversations in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15785–15810, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.883. URL https://aclanthology.org/2024.emnlp-main.883/.

APPENDIX

A DATASET CONSTRUCTION AND FIELD SEMANTICS

A.1 EVALUATED FIELD PAIRS

We study three closed-set prediction settings. Each output contains a primary affective field and one associated evidence field. Table 4 summarizes the construction used in the experiments.

Table 4: Task construction. The evidence roles are contextual, explanatory, and response-oriented, respectively.

Dataset	Primary	Evidence	Test N
MELD	emotion	sentiment polarity	2,610
RECCON	emotion	binary cause judgment	7,224
ESConv	seeker emotion	strategy label	1,809

MELD. Each instance contains the dialogue context and a target utterance. The model jointly predicts the target emotion and sentiment polarity (Poria et al., 2019). The normalized primary inventory is {anger, disgust, fear, joy, neutral, sadness, surprise}; the sentiment inventory is {negative, neutral, positive}. Sentiment is treated as contextual affective information, not as a free-text rationale.

RECCON. We use 7,224 candidate-cause instances from the DailyDialog portion of RECCON (Poria et al., 2021). Each instance contains the dialogue context, a target emotion-bearing utterance, and a candidate utterance. The evidence target is the binary label {false, true}, indicating whether the candidate utterance causes the target emotion. The primary inventory follows the seven DailyDialog emotion categories used by the released RECCON split. The task is therefore candidate-cause classification rather than free-form cause generation.

ESConv. The input contains the dialogue context and the observed supporter response (Liu et al., 2021). The primary inventory is {anxiety, depression, sadness, anger, fear, disgust, shame}. The evidence inventory contains eight strategy labels: {question, restatement or paraphrasing, reflection of feelings, self-disclosure, affirmation and reassurance, providing suggestions, information, others}. The strategy label characterizes the observed response. It is therefore called *response-oriented*; the task does not predict the next dialogue action.

A.2 Instance construction

MELD instances are defined at the target-utterance level. RECCON instances pair a target emotion-bearing utterance with one candidate utterance, which explains why its evaluation unit is a candidate judgment rather than a unique target emotion. ESConv instances pair the seeker-level emotion annotation with the strategy attached to an observed supporter response. The resulting test sizes are those shown in Table 4 and sum to 11,643.

A.3 Scope of the comparison

The three datasets were selected to cover contextual, explanatory, and response-oriented evidence roles. They do not form a controlled intervention on evidence type. Differences in CEDR can also reflect domain, label imbalance, context construction, annotation policy, and task difficulty. No causal claim is made from the cross-dataset ordering.

B Output Schemas, Prompts, and Normalization

B.1 Required JSON schemas

All supervised and label-constrained zero-shot generators are instructed to return one JSON object and no additional prose. The required schemas are

$$\begin{aligned} \text{MELD: } &\{\texttt{"emotion"} : m,\ \texttt{"sentiment"} : e\}, \\ \text{RECCON: } &\{\texttt{"emotion"} : m,\ \texttt{"is_cause"} : e\}, \\ \text{ESConv: } &\{\texttt{"emotion_type"} : m,\ \texttt{"strategy"} : e\}. \end{aligned}$$

The zero-shot prompts enumerate the complete legal labels for both fields and use the same schemas as the supervised models.

B.2 Joint-generation prompt template

The dataset-specific prompt is instantiated from the following template.

```
System: You are an affective dialogue classifier.
Return one JSON object only.
Do not add explanations or extra keys.
User: Dialogue context: [CONTEXT]
Target utterance or response: [TARGET]
Primary field: [PRIMARY FIELD]
Legal primary labels: [PRIMARY LABELS]
Evidence field: [EVIDENCE FIELD]
Legal evidence labels: [EVIDENCE LABELS]
Return exactly the required two-key JSON object.
```

For RECCON, the prompt separately marks the target emotion-bearing utterance and the candidate cause utterance. For ESConv, the target response is the observed supporter utterance whose strategy is classified.

B.3 Single-field experts and composition

The primary-only and evidence-only experts receive the same dialogue input and legal label inventories as the joint generator. Each expert returns only its assigned key. EP-Compose combines the two expert predictions,

$$\hat{Y}_i^{\mathrm{EP}} = \left(\hat{m}_i^{\mathrm{expert}}, \hat{e}_i^{\mathrm{expert}}\right), \tag{9}$$

whereas Hybrid-Compose retains the joint model's primary field and substitutes the evidence-only prediction,

$$\hat{Y}_i^{\mathrm{Hybrid}} = \left(\hat{m}_i^{\mathrm{joint}}, \hat{e}_i^{\mathrm{expert}}\right). \tag{10}$$

These settings are diagnostic baselines; neither performs conditional selection.

B.4 Normalization and exact matching

Before comparison, string values are lowercased, surrounding whitespace is removed, repeated whitespace is collapsed, and underscores are converted to spaces. Boolean values are normalized to `true` or `false`. No synonym matching, semantic label remapping, embedding similarity, or manual correction is applied. The same deterministic procedure is used for all models and settings.

B.5 Prediction freezing

Each generator produces one prediction per instance with deterministic decoding. Prediction files are frozen before expert composition, bootstrap resampling, verifier scoring, threshold selection, and CEVD evaluation. This separates model generation from downstream diagnostic resampling.

C Parsing and Structured-Output Accounting

C.1 Definitions

A prediction is *parseable* if a JSON object can be recovered. It is *field-complete* if both required keys are present. It is *schema-compliant* if it is parseable and field-complete and both normalized values belong to their legal label spaces.

Let P_i indicate parseability and F_i indicate field completeness, with $F_i = 0$ when no object is recovered. Let $L_{i,m}$ and $L_{i,e}$ indicate legal normalized values for the two fields; an unavailable value has indicator zero. We define

$$\mathrm{IOR} = \frac{1}{N}\sum_i (1 - P_i F_i), \tag{11}$$

$$\mathrm{SVR} = \frac{1}{N}\sum_i P_i F_i (1 - L_{i,m} L_{i,e}). \tag{12}$$

Both rates use the full evaluation size N as the denominator. IOR and SVR are disjoint under this convention, but either may overlap with evidence drift. A cross-schema confusion is an SVR case in which an illegal value for one field belongs to the other field's legal inventory.

C.2 No silent exclusion

Missing or illegal fields are represented by $\perp$ for correctness scoring. A completely unparsable output receives $\perp$ in both fields. Every instance remains in the evaluation set. Correct primary predictions paired with missing or illegal evidence count as evidence drift. Unavailable or incorrect primary predictions do not enter the CEDR conditioning set.

Table 5: Representative output cases under the stated accounting convention. Structural diagnostics and label correctness are evaluated separately.

Prediction case	Drift	JEM	IOR	SVR
No recoverable JSON object	no	0	1	0
Correct primary; evidence key missing	yes	0	1	0
Both keys present; correct primary; illegal evidence	yes	0	0	1
Primary key missing	no	0	1	0
Both keys present; illegal primary value	no	0	0	1
Correct primary; legal but incorrect evidence	yes	0	0	0
Both fields legal and correct	no	1	0	0

C.3 Confirmed parseability result

All supervised Joint-LoRA predictions used for the main diagnostic comparison were parseable and field-complete. Thus, IOR is 0 on MELD, RECCON, and ESConv (0 of 11,643 test instances in

total). Schema-space violations and cross-schema events are retained under the rules above; they are not removed before field or evidence-drift scoring. The headline CEDR results therefore do not arise from dropping malformed predictions.

Table 6: Confirmed invalid-output counts for supervised Qwen Joint-LoRA.

Dataset	Test N	IOR
MELD	2,610	0
RECCON	7,224	0
ESConv	1,809	0
Total	11,643	0

D Models, Training, and Leakage Controls

D.1 Generators and diagnostic settings

Qwen2.5-7B-Instruct is the primary generator (Qwen et al., 2024). Mistral-7B-Instruct-v0.3 is used for a cross-family comparison on ESConv and RECCON (Jiang et al., 2023). Within a dataset and backbone, the joint, primary-only, and evidence-only models use the same train-development-test partition. The joint model returns both fields in one JSON object. Label-constrained zero-shot generation uses the same legal labels and required schema but no task-specific adaptation.

D.2 Shared generator protocol

The generators are adapted with LoRA under four-bit quantization (Hu et al., 2022; Dettmers et al., 2023). Each reported adapter is trained for one epoch, and inference uses deterministic decoding. The joint model and each single-field expert are trained separately. Expert composition is therefore a diagnostic comparison, not a comparison at equal total computational cost.

Table 7: Shared settings used by the reported experiments.

Component	Reported setting
Primary backbone	Qwen2.5-7B-Instruct
Replication backbone	Mistral-7B-Instruct-v0.3
Adaptation	LoRA with four-bit quantization
Training duration	one epoch per adapter
Generation	deterministic; one JSON object
Normalization	case, whitespace, underscore only
Verifier backbone	DeBERTaV3-base
Development coverage targets	0.90 and 0.80
Bootstrap resamples	1,000
Matched-random trials	5,000

D.3 Conditional verifier

The reported CEVD experiment uses a DeBERTaV3-base verifier (He et al., 2023) for the ESConv strategy field. For dialogue context X, primary label m, and candidate evidence e, the verifier produces $s_\phi(X, m, e)$. Scores for all legal evidence candidates are normalized listwise. Training minimizes the negative log likelihood of the gold evidence candidate. At inference, the verifier receives the joint model's predicted primary label and scores the joint model's original evidence prediction.

The verifier is trained with the gold primary label and receives the predicted primary label at inference. These labels coincide on the primary-correct subset used to compute sCEDR. The ESConv candidate set contains all eight strategy labels listed in Section A; no candidate is generated open-endedly.

D.4 Data separation and threshold selection

Generator training uses only the training split. Verifier training also uses only training data. The development split is used to select the numerical risk threshold for a target evidence coverage. Test labels, test confusion patterns, and test risk scores are not used for model fitting, risk-score selection, or threshold selection. The selected threshold is transferred unchanged to the held-out test set.

D.5 Run and uncertainty scope

Each table entry is computed from one frozen generator run for the stated setting. The 1,000 bootstrap resamples and 5,000 matched-random trials are evaluation procedures and are not independent model-training runs. Reported bootstrap intervals condition on the frozen models and do not include variation across fine-tuning seeds.

E Complete Diagnostic Results

E.1 Main Joint-LoRA results

Table 8: Primary accuracy and marginal evidence macro-F1 under Qwen2.5-7B Joint-LoRA.

Dataset	Primary Acc.	Joint F1	Expert F1	$\Delta F1_e$
MELD	0.7019	0.7382	0.7462	0.0080
RECCON	0.8779	0.8081	0.8118	0.0037
ESConv	0.6368	0.5796	0.5841	0.0045

Table 9: Joint and conditional reliability under Qwen2.5-7B Joint-LoRA.

Dataset	JEM	EDR	CEDR
MELD	0.6805	0.0215	0.0306
RECCON	0.7615	0.1164	0.1326
ESConv	0.4063	0.2305	0.3620

For Qwen2.5-7B Joint-LoRA, let $N_m = \sum_i \mathbb{I}(\hat{m}_i = m_i)$ denote the number of instances with a correct primary prediction, and let $N_d = \sum_i d_i$ denote the number of evidence-drift cases. The number of jointly correct predictions is $N_m - N_d$. For $N_m > 0$, evidence accuracy within the primary-correct subset is $(N_m - N_d)/N_m = 1 - \mathrm{CEDR}$.

Table 10 reports this decomposition for each dataset. Joint-correct counts and conditional evidence accuracies are calculated from the reported primary-correct and drift counts. Conditional evidence accuracy refers only to this subset, not to the full test set.

Table 10: Decomposition of the primary-correct subset under Qwen2.5-7B Joint-LoRA. Joint correct is primary correct minus drift. Conditional evidence accuracy is $1 - \mathrm{CEDR}$, not evidence accuracy over the full test set. Rates are proportions rounded to four decimal places from the reported counts.

Dataset	Primary correct	Drift	Joint correct	CEDR	Conditional evidence acc.
MELD	1,832	56	1,776	0.0306	0.9694
RECCON	6,342	841	5,501	0.1326	0.8674
ESConv	1,152	417	735	0.3620	0.6380

Table 11: Percentile intervals from 1,000 instance-level bootstrap resamples of fixed predictions. They do not model dialogue dependence or training-seed variation.

Dataset	Metric	Point	2.5%	97.5%
MELD	JEM	0.6805	0.6621	0.6981
	EDR	0.0215	0.0157	0.0268
	CEDR	0.0306	0.0228	0.0383
RECCON	JEM	0.7615	0.7515	0.7709
	EDR	0.1164	0.1098	0.1238
	CEDR	0.1326	0.1247	0.1407
ESConv	JEM	0.4063	0.3842	0.4273
	EDR	0.2305	0.2106	0.2504
	CEDR	0.3620	0.3336	0.3899

Table 12: Effects of adaptation and model family. Zero-shot prompts contain all legal labels and the required schema.

Dataset	Setting	Primary	Evid. F1	CEDR
ESConv	Qwen zero-shot	0.3101	0.2068	0.7344
	Qwen Joint-LoRA	0.6368	0.5796	0.3620
	Mistral Joint-LoRA	0.6247	0.5789	0.3637
RECCON	Qwen zero-shot	0.7119	0.5521	0.2487
	Qwen Joint-LoRA	0.8779	0.8081	0.1326
	Mistral Joint-LoRA	0.8331	0.7751	0.1554

E.2 Instance-level bootstrap intervals

E.3 Adaptation and model-family comparison

Task-specific adaptation substantially lowers CEDR but does not eliminate it. The ESConv result is nearly identical for the two evaluated model families, and the RECCON result remains non-zero for both. These comparisons show that the observed failure is not confined to one backbone, but they do not support a claim about model families beyond the two evaluated 7B instruction-tuned models.

F Expert-Composition Diagnostics

Table 13: CEDR under unconditional field composition. A dash denotes a setting not included in the reported comparison.

Dataset	Joint-LoRA	EP-Compose	Hybrid-Compose
RECCON	0.1326	0.1374	–
ESConv	0.3620	0.3677	0.3646

For ESConv, Hybrid-Compose preserves Joint-LoRA's primary predictions, so both CEDR values use the same 1,152 instances. The reported rates, 0.3646 and 0.3620, differ by approximately 0.26 percentage points. The point estimates do not show an improvement from unconditional replacement. EP-Compose uses a different source for the primary prediction and need not have the same conditioning subset. The aggregate rates do not identify how many predictions were corrected or made incorrect; those counts require aligned predictions for individual instances.

G Statistical Procedures

G.1 Instance-level bootstrap

For a dataset with N evaluation records, each bootstrap replicate draws N records with replacement from the frozen prediction file and recomputes JEM, EDR, and CEDR. The generator, verifier, and threshold are not retrained. The procedure is repeated 1,000 times, and the empirical 2.5th and 97.5th percentiles are reported. Because the evaluation record is the resampling unit, the intervals do not capture within-dialogue dependence.

G.2 Matched random deferral

At each CEVD operating point, random deferral retains exactly the same number of test predictions as CEVD. Let R_b be the random baseline's sCEDR in trial b, and let R_{CEVD} be the observed CEVD result. Since lower sCEDR is better, the one-sided empirical p-value is

$$p = \frac{1 + \sum_{b=1}^{B} \mathbb{I}(R_b \le R_{\mathrm{CEVD}})}{B + 1}, \qquad B = 5{,}000. \tag{13}$$

The plus-one correction prevents a zero Monte Carlo p-value.

Table 14: Matched-random deferral at the two reported operating points.

Policy	Cov.	Rand. mean	Random central 95% interval	p
CEVD_{90}	0.8933	0.3637	[0.3535, 0.3735]	0.0002
CEVD_{80}	0.7905	0.3637	[0.3491, 0.3775]	0.0002

At both operating points, none of the 5,000 matched-random trials achieved sCEDR less than or equal to the CEVD result. Therefore, $p = (1 + 0)/(5{,}000 + 1) = 0.00019996$, reported as $p = 0.0002$.

G.3 Interpretation of uncertainty

The bootstrap intervals quantify sampling uncertainty conditional on frozen predictions. The matched-random intervals quantify variability induced by randomly choosing which evidence predictions to retain at fixed coverage. Neither procedure estimates uncertainty from retraining the generator or verifier with different seeds.

H Additional CEVD Analysis

H.1 Risk and retention

We use the scoring and eligibility rule defined in Section 3. Evidence is eligible for release only when both generated labels are legal. For eligible predictions, $q_i = p_\phi(\hat{e}_i \mid X_i, \hat{m}_i)$ and $r_i = 1 - q_i$. Otherwise, $q_i = 0$ and the eligibility gate is zero. At the development-selected threshold τ_c, evidence is retained only if it is eligible and $r_i \le \tau_c$. The primary prediction is never changed.

H.2 SELECTIVE METRICS

For a fixed threshold, the reported metrics are

$$\mathrm{Cov}(\tau) = \frac{1}{N}\sum_i a_i(\tau), \tag{14}$$

$$\mathrm{PCov}(\tau) = \frac{\sum_i \mathbb{I}(\hat{m}_i = m_i)a_i(\tau)}{\sum_i \mathbb{I}(\hat{m}_i = m_i)}, \tag{15}$$

$$\mathrm{sCEDR}(\tau) = \frac{\sum_i d_i a_i(\tau)}{\sum_i \mathbb{I}(\hat{m}_i = m_i)a_i(\tau)}, \tag{16}$$

$$\mathrm{Capture}(\tau) = \frac{\sum_i d_i[1 - a_i(\tau)]}{\sum_i d_i}, \tag{17}$$

$$\mathrm{DPrec}(\tau) = \frac{\sum_i d_i[1 - a_i(\tau)]}{\sum_i \mathbb{I}(\hat{m}_i = m_i)[1 - a_i(\tau)]}. \tag{18}$$

PCov is evidence coverage within the primary-correct subset. Capture is the fraction of drift cases withheld, and DPrec is the drift rate among withheld primary-correct instances.

H.3 HELD-OUT OPERATING POINTS

Table 15: Held-out Mistral-7B results on ESConv. Δ is the absolute sCEDR change from no deferral.

Policy	Test Cov.	P-Cov.	sCEDR	Δ
No deferral	1.0000	1.0000	0.3637	–
CEVD_{90}	0.8933	0.8938	0.3257	-0.0380
CEVD_{80}	0.7905	0.7965	0.3033	-0.0604

H.4 ERROR CONCENTRATION AMONG WITHHELD EVIDENCE

Deferral precision is the evidence error rate among withheld instances with a correct primary prediction. We express error concentration relative to baseline CEDR as $\mathrm{DPrec}(\tau)/\mathrm{CEDR}$ when the ratio is defined. This is a transformation of existing metrics, not an independent experimental result.

Using the reported rounded values, $0.6833/0.3637 \approx 1.88$ and $0.6000/0.3637 \approx 1.65$.

Table 16: Error concentration at the two development-selected operating points. Enrichment is DPrec divided by baseline CEDR (0.3637), calculated from the reported rounded values.

Policy	sCEDR	Capture	DPrec	Enrichment
CEVD_{90}	0.3257	0.1995	0.6833	1.88
CEVD_{80}	0.3033	0.3358	0.6000	1.65

H.5 FULL COVERAGE CURVE

Table 17: Points used for the CEVD coverage-error curve in the main paper.

Retained evidence coverage	sCEDR
0.6081	0.2591
0.6960	0.2752
0.7905	0.3033
0.8933	0.3257
0.9420	0.3446
1.0000	0.3637

H.6 RISK-SCORE COMPARISON

Table 18: Risk-score comparison at the 90% development target.

Score (90%)	Cov.	sCEDR	Δ	Capture	DPrec
Support	0.8933	0.3257	−0.0380	0.1995	0.6833
Entropy	0.8878	0.3320	−0.0317	0.1849	0.6281
Disagreement	0.8983	0.3297	−0.0340	0.1849	0.6667

Table 19: Risk-score comparison at the 80% development target.

Score (80%)	Cov.	sCEDR	Δ	Capture	DPrec
Support	0.7905	0.3033	−0.0604	0.3358	0.6000
Entropy	0.7828	0.3041	−0.0597	0.3431	0.5826
Disagreement	0.7927	0.3037	−0.0600	0.3358	0.5974

Table 20: Risk ranking among primary-correct Mistral-ESConv predictions.

Score	AUROC	AUPRC
Original support	0.7093	0.5600
Entropy	0.7078	0.5526
Disagreement	0.7081	0.5565

CEVD should be interpreted as selective risk reduction rather than automatic correction. Lower sCEDR is achieved by withholding a subset of evidence predictions while preserving all primary predictions.

I ERROR ANALYSIS AND QUALITATIVE-CASE PROTOCOL

I.1 ESCONV STRATEGY CONFUSIONS

ESConv drift often occurs between legal strategy labels that are semantically related but serve different conversational functions. Representative observed directions are shown in Table 21.

Table 21: Representative ESConv evidence-drift directions. All values are legal strategy labels, so the cases are semantic rather than schema errors.

Reference	Prediction	Functional distinction
Information	Providing suggestions	Information is replaced by advice.
Reflection of feelings	Affirmation and reassurance	Reflection is replaced by validation.
Others	Question	A residual label is replaced by information seeking.

The directions illustrate why schema compliance does not imply evidence correctness: both the reference and predicted values belong to the legal strategy inventory, but they assign different conversational functions to the same observed response.

I.2 RECCON CAUSE-DIRECTION ASYMMETRY

Among the 841 RECCON instances in which the primary emotion is correct but the cause judgment is incorrect, 599 are true-cause→non-cause errors and 242 are non-cause→cause errors.

Missed causes account for 71.2% of the observed evidence-drift cases, compared with 28.8% for false cause assignments. These are shares of errors, not class-specific false-negative and false-positive rates. They do not establish a higher probability of one error type without the corresponding class denominators.

Table 22: Direction of RECCON evidence-drift errors.

Direction	Count	Share of RECCON drift
True cause → non-cause	599	71.2%
Non-cause → cause	242	28.8%

I.3 Status of the motivating example

Figure 1 is explicitly schematic and is not presented as an ESConv test record. No manually authored example is labeled as a real dataset instance in this supplement. This distinction prevents an illustrative example from being mistaken for evidence about a specific frozen prediction.

I.4 Protocol for any released qualitative cases

Any real qualitative case included in an anonymous artifact must be extracted from the frozen test predictions and identified only by an anonymous sample ID. A complete case record should contain the dialogue excerpt, gold primary and evidence fields, original joint prediction, verifier support q_i, risk r_i, selected threshold τ_c, and retain-or-withhold decision. Text may be shortened only for display, with omissions marked; labels and predictions must not be manually rewritten.