

Right for the Right Reasons? Evaluating the Structural Fidelity of LLM Emotion Explanations

Hajime Hotta^{1*} and Minh-Tien Nguyen^{2*}

¹ Hajime Institute, Singapore.

hotta@hajime.institute

² Hung Yen University of Technology and Education, Hung Yen, Vietnam.

tiennm@utehy.edu.vn

Abstract

This paper introduces a framework for evaluating structural fidelity in emotion reasoning: the degree to which the causal structure of AI-generated emotion explanations aligns with human-curated explicit causal abstractions. Unlike traditional surface-level text overlap metrics, we propose two metrics: **Structural Causal Similarity (SCS)** for measuring static structural alignment, and **Dynamic Causal Sensitivity (DCS)** as a proxy for measuring dynamic sensitivity to counterfactual interventions, whose Agency component we validate against three-annotator human ratings of counterfactual shift (Spearman $\rho = +0.39$, $n = 152$). We evaluate these on **EmoReason-4k**, a new benchmark comprising two human-annotated datasets ($N = 2000$ each) from social media and conversational dialogues. Testing seven LLMs identifies consistent graph-generation behaviors across models and domains, such as “structural elaboration” (a tendency for models to explicate intermediate steps that humans leave implicit) and sensitivity to shifts in causal agency locus. However, models struggle to generate updated causal states in complex discourse. Ablation studies support that our metrics capture underlying topological properties rather than just parsing artifacts.

1 Introduction

Emotion reasoning is the capacity to infer and explain the causal origins of human affective states. It has become an evaluation target in its own right, beyond predicting which emotion is expressed (Yeo and Jaidka, 2025). As Large Language Models (LLMs) are increasingly deployed to generate explicit emotion explanations, whether their reasoning aligns with human cognitive expectations remains an open challenge (Jacovi and Goldberg, 2020). Prior metrics target endpoint classification accuracy (Ai et al., 2025; Wu et al., 2025) or

surface-level lexical overlap, such as Emotion Interpretation (EI) (Lin et al., 2025). String matching cannot separate a sound causal explanation from a plausible hallucination that shares its vocabulary.

Figure 1 makes this concrete on an EmpatheticDialogues example. A human connects the *Cue* (a passed deadline) to the *Emotion* (Anxiety) via an *Appraisal of Uncertainty*, leaving self-evident background knowledge implicit, whereas GPT-5 infers malicious “other-blame” to justify *Anger*. Both texts discuss deadlines and interpersonal responses, so they yield a high BLEU score while their causal structures diverge drastically, motivating a shift from linguistic plausibility to structured causal comparison.

We address this gap by evaluating the structural fidelity of AI-generated emotion explanations against human-verified canonical abstractions, formalizing emotion reasoning as a directed causal graph: *Cue* $\rightarrow$ *Prior Knowledge* $\rightarrow$ *Appraisal* $\rightarrow$ *Emotion*. Two complementary metrics instantiate it: **Structural Causal Similarity (SCS)**, our primary measure of static structural alignment, and **Dynamic Causal Sensitivity (DCS)**, a lightweight probe of a model’s directional consistency under controlled counterfactual edits, whose Agency component we ground in three-annotator human ratings and whose VAD component we retain as a diagnostic indicator (Section 5.2).

We benchmark these metrics on **EmoReason-4k** ($N = 4000$), a new resource combining GPT-seeded drafts with strict human verification across social media and conversational domains. Across seven prominent LLMs we identify consistent generative behaviors: over-explication of intermediate reasoning steps (“structural elaboration”), and a fragile sensitivity to shifts in causal agency within complex discourse. The result is a practical way to audit how well explicit AI explanations comply with formalized human reasoning abstractions.

*Corresponding Author.

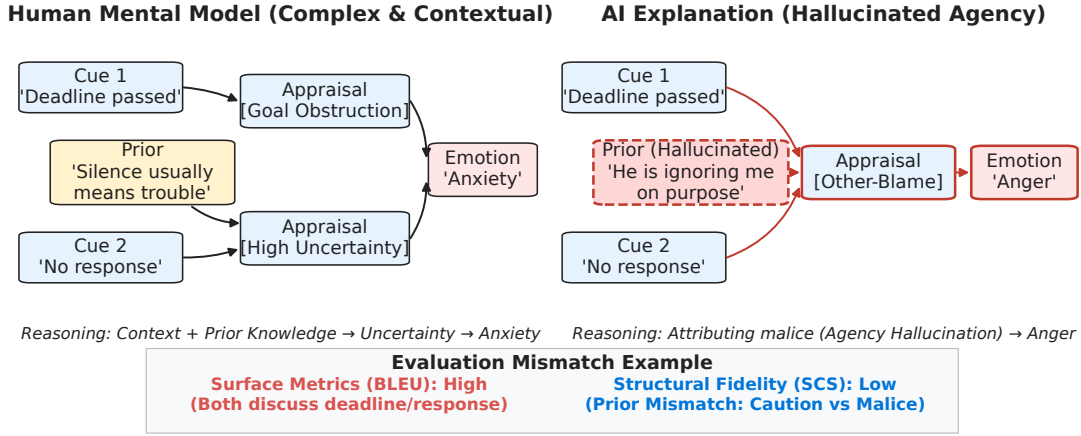


Figure 1: Structural Fidelity vs. Surface Similarity. This example exhibits an evaluation mismatch: the AI hallucinates malicious agency while the human infers uncertainty. SCS correctly assigns a low score despite surface text similarities.

2 Related Work

From emotion recognition to reasoning Early affective computing focused on classification. Work in that line covers knowledge-grounded empathetic dialogue (Li et al., 2022b), cross-modal transformers and hypergraph models (Zhu et al., 2024; Huang et al., 2024), and graph-based multimodal recognition (Shou et al., 2025; Ai et al., 2025). The field has since shifted toward LLM-based generation and analysis. Recent systems include emotional LLM suites (Liu et al., 2024), instruction-tuned and behavior-aware multimodal LLM frameworks (Cheng et al., 2024; Fu et al., 2025), LLM-supervised pre-training (Dutta and Ganapathy, 2025), and lightweight adapter plugins (Yang et al., 2025). Cognitive modeling with traits and distortions has also been applied to this task (Nguyen et al., 2026), though for label prediction rather than explanation structure. While recent metrics evaluate emotional explainability (e.g., EI (Lin et al., 2025)) and Theory of Mind (Yeo and Jaidka, 2025), they mostly assess surface-level text or classification performance. This misses the structural validity of the underlying causal reasoning. Concurrent work scores appraisal reasoning step by step against annotated inferential chains rather than at the emotion label (Sun et al., 2026). Similarly, causal emotion entailment methods enrich utterance-level causality with social commonsense (Li et al., 2022a) and knowledge-bridged interaction graphs (Zhao et al., 2023), and emotion-cause pair extraction advances through center-event reasoning (Wang et al., 2024), all of which model interactions between utterances but skip the internal cognitive derivation. Standard graph metrics

(e.g., Smatch for AMR or SHD for causal discovery) expect exact node overlap, which penalizes the natural paraphrasing and elaboration typical of LLM outputs.

Causal bounding and faithfulness We extend causal abstraction paradigms from mechanistic interpretability (Geiger et al., 2021; Wu et al., 2023) to evaluate the explicit structural fidelity of generated artifacts. Aligning with the shift toward faithfulness, from its definitional groundwork (Jacovi and Goldberg, 2020) through surveys of faithful explanation methods (Lyu et al., 2024) and dataset reviews (Wiegrefe and Marasović, 2021) to measurements of unfaithful chain-of-thought reasoning (Lanham et al., 2023; Turpin et al., 2023), we treat emotion as structured causal inference (Lombrozo, 2006; Barrett, 2017). Unlike NLP counterfactuals targeting prediction robustness (Kaushik et al., 2020; Gardner et al., 2020), our DCS metric dynamically evaluates the consistency of the internal reasoning state update.

3 Emotional Reasoning & Data Construction

We adopt a reduced, theory-inspired appraisal abstraction (Scherer et al., 2001), mapping complex cognitive dimensions onto an operational 4-component theoretical framework: **Cue**, **Prior Knowledge**, **Appraisal**, and **Emotion**. While this framework defines the required *types* of causal information, we do not enforce a strict 4-node limit. When applied to natural language explanations, this theoretical structure maps onto a *variable-length causal graph*. Components may be rep-

resented by multiple nodes (e.g., several chained appraisals), or left entirely implicit by human annotators (e.g., omitting self-evident prior knowledge, yielding a simpler *Cue* $\rightarrow$ *Appraisal* $\rightarrow$ *Emotion* subgraph). For evaluating explicit emotion explanations, this flexible graph mapping of the 4-component theory serves as a principled constraint that isolates the core reasoning path, avoiding the noise of unconstrained naturalistic text. To measure dynamic state updates, we track shifts explicitly across a **5-dimensional VAD and Agency state space**: Valence, Arousal and Dominance (Mohammad, 2018), plus Self-Agency and Other-Agency. We include **Agency** because assigning responsibility (self vs. other) is key for distinguishing social emotions like guilt and anger. Although we omit some theoretical dimensions (like Novelty), this space offers a numeric proxy for tracking directional shifts across different contexts.

By conceptualizing emotion explanations as these bounded, multi-dimensional graphs, we can move beyond surface-level text comparison and mathematically evaluate whether an AI model’s underlying reasoning topology matches human cognitive expectations. To instantiate this evaluation, we developed two human-curated datasets, collectively called **EmoReason-4k** ($N = 4000$), designed specifically to capture human emotion reasoning as explicit, dynamic causal structures. We derived **EmoReason-GE** ($N = 2000$ base samples) representing standard social media text from the GoEmotions test split (Demszky et al., 2020), and **EmoReason-ED** ($N = 2000$ base samples) to capture complex interpersonal dynamics from EmpatheticDialogues (Rashkin et al., 2019). We filtered for structural complexity (length > 5 words, discourse connectives present) to ensure the presence of overt reasoning signals.

Each base sample yields a dynamic pair: (1) an **Original Graph** isolating the explicit sequence connecting the context Cue to the Emotion (“I wanted this feature to get rid of the beige skins. . .” $\rightarrow$ Unpleasantness $\rightarrow$ Annoyance), and (2) a **Counterfactual Graph** created via targeted manual text edits by human annotators to isolate and test dynamic variable sensitivity (adding “. . . who thought that looked good?” introduces Uncertainty and updates the emotion to Confusion).

Three expert annotators (linguistics/psychology) underwent 5-hour training on Appraisal Theory semantics. Node typing agreement reached Krippendorff’s $\alpha = 0.81$, with mean graph-level pairwise

structural alignment of 0.72. To accelerate annotation, GPT-5 drafted initial structures, but every graph underwent human verification. Human correction moved the drafts substantially. Initial draft versus final canonical SCS was 0.48, and 31–42% of samples were rejected for irreconcilable ambiguity, so the final graphs are human-curated explicit reasoning abstractions. While the high rejection rate indicates stringent quality filtering, it follows from the design goal of building a *diagnostic probe* over explicit structural reasoning rather than a sample of natural conversation.

4 Structural Fidelity Assessment

Following Scherer et al. (2001), we formalize emotion reasoning as constructing a directed causal graph $G = (V, E)$, with nodes $V \subseteq \{\text{Cue, Prior, Appraisal, Emotion}\}$ and causal edges E . The task is to evaluate the structural alignment between the AI-generated inference graph and the human-verified canonical graph.

4.1 Structural Causal Similarity (SCS)

To illustrate SCS, we return to the concept of structural fidelity (Figure 1 shows an inverse example where high lexical overlap masks severe structural dissimilarity). When vocabulary differs (e.g., concrete physiological vs. abstract psychological terms), two explanations can still follow the identical causal path: *Cue* $\rightarrow$ *Appraisal* $\rightarrow$ *Emotion*. SCS algorithmically measures this explicit structural alignment, achieving a valid score that reflects the shared reasoning topology despite surface-level lexical differences.

Structural Causal Similarity (SCS) SCS is a static structural match score between the canonical human-curated graph and the AI-generated graph. Unlike unweighted Graph Edit Distance (GED) formulations that penalize any deviation uniformly, SCS weights the core causal path while accommodating the natural “structural elaboration” typical of LLMs (e.g., intermediate deductive steps humans leave implicit). Optimal GED computation is NP-hard, but our graphs are bounded by the small set of semantic node types, so a greedy abstraction mapping remains viable for this constrained vocabulary.

Given an extracted graph G_E and a canonical graph G_C , SCS is defined as the geometric mean of semantic node overlap and structural edge alignment. We use the geometric mean so that a severe

failure in either dimension heavily penalizes the overall score, though a true zero requires an exact zero in one of the sub-components. This ensures high scores require *both* correct meaning and valid causal logic:

$$\text{SCS}(G_E, G_C) = \sqrt{\text{Sim}_{\text{semantic}} \times \text{Sim}_{\text{structural}}} \quad (1)$$

Semantic Node Similarity ($\text{Sim}_{\text{semantic}}$) We first map nodes between G_E and G_C based on identical types. The base substitution penalty relies on SBERT distance, while unmapped nodes incur asymmetric penalties. Omitting a canonical node costs more than inserting a novel elaboration, which accommodates the tendency of LLMs to generate verbose cognitive steps. Edges between matched nodes are checked for symmetry, and missing edges add mismatch penalties (c_{edge}). The total structural edit distance (d_{SemGED}) normalizes these operations by overall graph weight to yield a continuous $[0, 1]$ semantic similarity:

$$\text{Sim}_{\text{semantic}} = \exp \left(- \frac{d_{\text{SemGED}}(G_C, G_E)}{W(G_C) + W(G_E)} \right)$$

Specific normative parameters (e.g., insertion cost = 0.5, deletion cost = 1.0) set strict boundaries for valid elaborations. These weights instantiate the theoretical premise that completely omitting a core Appraisal node is a more severe cognitive failure than generating a superfluous but plausible intermediate reasoning step. This stability is further probed via sensitivity analysis in Appendix B.9. Node alignment utilizes greedy matching for efficiency, which empirically tracks optimal assignment algorithms (Appendix A.1).

Structural Consistency ($\text{Sim}_{\text{structural}}$) This dimension acts as a sanity check against egregious causal violations, ensuring that the explicit AI graph loosely follows valid deductive reasoning patterns independently of textual semantic overlap. We compute consistency as:

$$\text{Sim}_{\text{structural}} = 0.5 \cdot P_{\text{existence}} + 0.5 \cdot O_{\text{consistency}}$$

This 0.5/0.5 heuristic weighting assumes equal baseline importance between having a connected causal path and the directionality of that sequence.

Path existence ($P_{\text{existence}} \in \{0, 1\}$) checks for a valid path from Cue to Emotion, heavily penalizing fundamentally disconnected graphs. **Order consistency** ($O_{\text{consistency}}$) measures the fraction of

graph edges honoring valid temporal causal ranks: $r(\text{Cue}) = 0 < r(\text{Prior}) = 1 < r(\text{Appraisal}) = 2 < r(\text{Emotion}) = 3$. We allow forward flow and same-level connections to accommodate concurrent multi-appraisals, penalizing only explicit backward violations that contradict linear causality.

Ultimately, this dual design successfully prevents an AI model from scoring highly simply by generating a semantically correct "bag of keywords." By forcing semantic matching to route through valid topological channels, achieving a high SCS score demands both correct linguistic intent *and* a coherent generative structure. Section 5 reports the baseline validations behind this claim.

Concretely, for the cue *"The QB said they loved it"* (target emotion *Joy*), the canonical graph is the three-node chain $\text{Cue} \rightarrow \text{"achievement of desired goal"} \rightarrow \text{Joy}$, while a representative AI graph explicates two further steps: $\text{Cue} \rightarrow \text{"the speaker supports the QB's team"} \rightarrow \text{"the QB's approval indicates a positive outcome for the team"} \rightarrow \text{"achievement of desired goal"} \rightarrow \text{Joy}$. The asymmetric costs price the two inserted nodes at 0.5 each rather than as errors, and the pair scores $\text{SCS} = 0.82$. BLEU on the same pair is 0.00, so surface metrics sharply penalize this valid elaboration.

4.2 Dynamic Causal Sensitivity (DCS)

While SCS evaluates static structural alignment, **Dynamic Causal Sensitivity (DCS)** acts as a proxy to measure how dynamically the extracted reasoning shifts in response to contextual changes. Using the counterfactual pairs in EmoReason-4k, DCS evaluates whether a model updates its internal causal graph in the correct direction when the input semantic dimensions are perturbed.

For each graph G , we calculate a 5-dimensional continuous state vector in the VAD–Agency space based on the textual content of its nodes:

$$\mathbf{S}(G) = [\mathcal{V}, \mathcal{A}, \mathcal{D}, \delta_{\text{self}}, \delta_{\text{other}}] \quad (2)$$

where $\mathcal{V}, \mathcal{A}, \mathcal{D} \in [0, 1]$ are Valence, Arousal, and Dominance scores averaged from the NRC-VAD lexicon (Mohammad, 2018). δ_{self} and δ_{other} denote the locus of agency. Because Agency is inherently categorical in traditional psychological models, we map it into this continuous space heuristically: if self-agency is explicitly attributed, $\delta_{\text{self}} = 1.0, \delta_{\text{other}} = 0.0$. If other-agency is explicitly attributed, $\delta_{\text{self}} = 0.0, \delta_{\text{other}} = 1.0$. If no explicit agency is identifiable (or neutral), we assign

$\delta_{\text{self}} = 0.5, \delta_{\text{other}} = 0.5$ to place the vector at the geometric center of the agency sub-space. While mapping high-dimensional emotional nuances into a rigid 5D space necessarily entails some loss of semantic fidelity, we adopt this projection not to capture every linguistic subtlety, but as the most theoretically established and computationally tractable proxy for quantifying continuous, dynamic causal shifts.

For a base text T_{base} and counterfactual T_{cf} , we extract graphs G_E^{base} and G_E^{cf} . We compute the model’s update vector $\Delta \mathbf{S}_E = \mathbf{S}(G_E^{\text{cf}}) - \mathbf{S}(G_E^{\text{base}})$, and the human canonical update vector $\Delta \mathbf{S}_C = \mathbf{S}(G_C^{\text{cf}}) - \mathbf{S}(G_C^{\text{base}})$. The primary multi-dimensional DCS formulation is the cosine similarity between these update vectors:

$$\text{DCS}_{5D}(\Delta \mathbf{S}_E, \Delta \mathbf{S}_C) = \frac{\Delta \mathbf{S}_E \cdot \Delta \mathbf{S}_C}{\|\Delta \mathbf{S}_E\| \|\Delta \mathbf{S}_C\|}$$

If both vectors perfectly align in direction, $\text{DCS} = 1.0$; opposite directions yield -1.0 . If $\Delta \mathbf{S}_C = \mathbf{0}$ and $\Delta \mathbf{S}_E = \mathbf{0}$ (both human and AI correctly identify that no shift occurred), we define $\text{DCS} = 1.0$. However, to avoid over-rewarding models that conservatively generate static graphs regardless of input (a "no-change" bias), we analyze performance across different shift types.

The Agency subspace shifts discretely while the $\mathcal{V}, \mathcal{A}, \mathcal{D}$ subspace shifts continuously, so the raw 5-D cosine can be dominated by the discrete component and we do not interpret it in isolation. We instead condition DCS on the gold intervention type of each counterfactual pair (VAD-shift, combined VAD+Agency, agency-isolated, or no-change; distribution in Appendix B.2), and treat the binary *zero-update rate* (the fraction of pairs whose extracted state registers no change at all) as the primary indicator of update stubbornness, with the continuous DCS as a secondary indicator of directional quality given an update. Both quantities inherit the detection envelope of the lexical state proxy, so we report them jointly with the proxy’s signal coverage, which we audit in Appendix B.5.

5 Validation of the Framework

Before applying our metrics to evaluate LLMs, we establish the validity of the structural abstraction itself. We must assess whether SCS and DCS are sensitive to meaningful structural differences rather than merely echoing textual overlap, prompt sensitivity, or the specific dialect of the extraction parser.

Metric	GE		ED	
	ρ	p	ρ	p
ROUGE-L	+0.182	<0.001	+0.093	<0.001
BERTScore	+0.164	<0.001	+0.082	<0.001
BLEU	+0.106	<0.001	+0.040	<0.001
EI	+0.063	<0.001	+0.023	0.043

Table 1: Spearman correlation (ρ) between SCS and baseline metrics. All $|\rho| < 0.2$, confirming SCS captures distinct structural variations.

5.1 Controlled Baselines and Human Alignment

SCS effectively distinguishes valid structures (Identity graphs scoring 1.0) from Random graphs (structurally random graphs with sampled labels, 0.248), as detailed in Table 7 (Appendix B). The Shuffled labels baseline (canonical structure, randomly permuted node labels) scores 0.570, because topological similarity is preserved and SBERT embeddings retain residual semantic proximity even across shuffled emotional terms. On GE this control sits inside the models’ own score range (Table 3), so SCS separates models only weakly there. On ED, where canonical graphs are longer, it falls to 0.390 and every model clears it under both prompting conditions (Appendix B.8). DCS reliably discriminates Aligned Change (perfectly matching human directional vectors, 1.0) from Random Change (randomly sampled VADA update vectors, 0.005).

To further validate SCS, we conducted a human evaluation on a random subset of 500 samples. Two annotators, independent of the three who built the canonical graphs, rated the “structural faithfulness” of AI explanations on a 5-point Likert scale by examining both text and extracted graphs. SCS correlates with these human structural judgments at $\rho = 0.65$ (95% CI [0.60, 0.70], $p < 0.001$). It scores higher than a GPT-5 LLM-as-a-Judge baseline ($\rho = 0.58$), BERTScore ($\rho = 0.25$), and BLEU ($\rho = 0.18$).

We also compared SCS against four established NLP metrics computing text explanations derived from graph nodes (Table 1). All correlations are weak ($|\rho| < 0.2$), confirming that SCS measures a fundamentally different dimension from textual similarity. The stronger correlations on GE than ED reflect the structural complexity of dialogue contexts. An exact-match overlap baseline (strict lexical identity on nodes/edges) yielded a mean of 0.081 vs. SCS’s 0.498 ($r = 0.269$, $p < 0.001$), confirming that SCS’s sentence-embedding inte-

gration is essential for capturing valid semantic variations.

5.2 Human Validation of DCS Interventions

Three expert annotators independently rated perceived shifts for the 152-pair EmoReason-ED counterfactual pilot on 5-point Likert scales (VAD shift, Agency shift, overall change), blind to intervention labels and to proxy outputs. The design is balanced across the four intervention types (38 each). Agreement is high on the Agency dimension (interval Krippendorff $\alpha = 0.914$; pairwise weighted $\kappa \in [0.883, 0.936]$), and the mean human Agency rating correlates with the proxy’s agency signal at Spearman $\rho = +0.391$ ($p < 0.001$, $n = 152$). A proxy-free manipulation check separates the two channels: raters score agency-manipulated pairs above the remaining pairs (Vargha–Delaney $A = 0.698$), whereas the corresponding VAD check does not discriminate ($A = 0.379$), and the VAD rating-versus-proxy correlation is null ($\rho = -0.131$, $p = 0.11$). We therefore scope the human-grounded claim to the Agency dimension of DCS, retain VAD-DCS as a diagnostic indicator, and name a pairwise forced-choice elicitation as the replacement instrument for VAD grounding in Section 8. The pilot’s bound (one dataset, 152 pairs, three raters) is restated in the Limitations. The full rating instructions, the per-rater CSV and the analysis script that reproduces every number in this subsection ship with the release.

5.3 Multi-Extractor Robustness

A potential vulnerability in structural evaluation is metric circularity. Because GPT-5 generates data and acts as the initial evaluator, SCS might capture alignment with a “GPT-5 structural dialect.” To test this empirically, we re-extracted graphs for the entire dataset using Claude 4.5 Sonnet. As detailed in Table 2, per-model SCS under the two extractors is strongly correlated (Pearson $r = 0.91$; Spearman $\rho = 0.79$ over the seven models), and the largest per-model shift is 0.016 SCS, which bounds any GPT-5-dialect advantage. SCS therefore measures *relative structural divergence* through a shared evaluation lens. It does not measure absolute distance to an unobservable cognitive ground truth.

This agreement holds at the aggregate level and not per item. On a stratified counterfactual subsample, the binary zero-update indicator agrees across the two extractors at Cohen $\kappa = 0.14$ on agency-isolated pairs ($n = 140$), about what a bi-

nary read-out of a stochastic pipeline supports. We therefore claim only that aggregate patterns survive extractor substitution, and make no claim of per-item diagnostic validity, for which a coreference-aware extractor is the prerequisite (Section 8). Appendix B.5 reports a complementary audit of the DCS state proxy itself, which covers keyword matching, signal coverage, and undefined-delta semantics of the lexical agency channel.

Model	GPT-5 Ext.	Claude 4.5 Ext.	Δ
Claude 4.5 Sonnet	0.603	0.588	−0.015
GPT-5	0.586	0.575	−0.011
GPT-5-mini	0.586	0.570	−0.016
GPT-5-nano	0.579	0.563	−0.016
Gemini 2.5 Flash	0.576	0.572	−0.004
Llama 3.1 8B	0.571	0.566	−0.005
Qwen3 30B	0.568	0.559	−0.009

Table 2: Extraction robustness. Per-model SCS under the GPT-5 versus Claude 4.5 Sonnet extractors (Pearson $r = 0.91$; Spearman $\rho = 0.79$; largest per-model shift 0.016).

6 Experimental Setup

We evaluated seven LLMs: **GPT-5** (normal, mini, nano), **Gemini 2.5 Flash**, **Claude 4.5 Sonnet**, **Llama 3.1 8B Instruct**, and **Qwen3 30B A3B**. Temperature was fixed to 0.0 wherever the API exposes it and to the lowest deterministic setting otherwise. Explanations were converted to graphs using GPT-5 (temperature=0.0) as a fixed extraction model to isolate reasoning from parsing variance.

To test instruction effects, we compared two formats: (i) **Direct Prompting** (straightforward appraisal request), and (ii) **Scaffolded Cognitive Prompting** (step-by-step schema derivation enforcing the 4-component theoretical constraints). For DCS, counterfactuals were created via controlled edits isolating cognitive variable shifts.

7 Model Evaluation & Findings

Having validated the metrics, we apply them to benchmark the seven LLMs on EmoReason-4k.

Schema compliance vs. prompt strategy For each sample, we computed SCS by comparing the model’s extracted graph to the human canonical graph. Results (Table 3) show that Scaffolded prompting consistently outperforms Direct prompting. Scaffolded guidance regularizes the output format and forces compliance with the requested

4-component framework, which directly raises the SCS structural alignment score. The gain therefore measures *schema compliance* and does not establish improved latent reasoning.

Gemini 2.5 Flash achieved the highest structural alignment (SCS = 0.578), followed by Claude 4.5 Sonnet (SCS = 0.570). Open-source models also performed competitively: Llama 3.1 8B (0.556) and Qwen3 30B A3B (0.549) achieved scores comparable to the GPT-5 family (0.542–0.557). Paired t-tests confirm that Scaffolded prompting significantly outperforms Direct prompting across all tested models ($p < 0.001$).

The two DCS columns condition on the gold intervention type: VAD-shift averages over VAD-only pairs ($n=723$ –1,617 per model) and VAD+Agency over combined VAD+Agency pairs (the N column). The lower N for the GPT models reflects safety refusals by those models when generating counterfactual explanations. Per-model 95% bootstrap CIs from the multi-seed counterfactual regeneration (an EmoReason-ED subsample) are given in Table 8; CI half-widths of ≈ 0.08 caution against over-reading inter-model DCS gaps below ~ 0.10 .

On EmoReason-GE, every model tracks the human update direction under both conditions, with DCS between 0.595 and 0.701 against a random-change floor of 0.005 (Table 7). The binary zero-update indicator never fires here. Across the seven benchmarked models, 0 of 10,888 counterfactual pairs register a frozen state, so the freezing reported below is specific to multi-turn dialogue. Open-source models (Llama, Qwen) reach DCS on combined VAD+Agency shifts in a similar range to proprietary models, indicating that directional update tracking under combined interventions is present across training architectures on straightforward single-author text. Because EmoReason-GE contains no agency-isolated counterfactuals (Appendix B.2), this table licenses no architecture-level conclusion about the agency channel on its own.

However, this competency proves remarkably fragile. Moving from the relatively standard social media assertions in EmoReason-GE to the complex, multi-turn interpersonal dynamics of EmoReason-ED reveals a suggestive pattern we characterize as *Agency Stubbornness*. Directional tracking degrades on ED across the board. DCS on VAD-shift counterfactuals drops from GE to ED for all seven models (Table 4). On the agency channel, a single-model probe (GPT-5-mini, $N=1,000$ pairs;

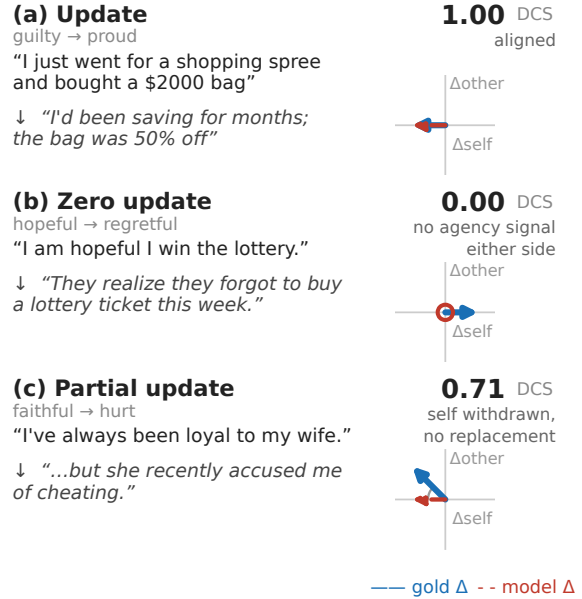


Figure 2: Agency-channel worked examples from the EmoReason-ED single-model probe (GPT-5-mini). DCS on this channel is the cosine between the gold agency update Δ and the model’s, in the (self, other) accountability plane.

Appendix B.8) registers exactly zero agency update on 75.7% of pairs. An instrument audit (Appendix B.5) attributes the bulk of this zero-mass to the detection envelope of the lexical state proxy rather than to model behavior alone, so we read the per-model magnitude of the effect as instrument-bounded rather than as a purely behavioral rate.

For example, when “My colleague ruined the presentation” is changed to “I ruined the presentation,” the model stubbornly extracts ‘Agent=Other’ for both and does not revise causal agency despite the clear syntactic shift (Appendix B.10). Agency Stubbornness is a composite *operational failure mode*. It combines a genuine difficulty in updating the multi-turn locus of control with parser noise on ambiguous pronouns. The instrument audit bounds the second component without removing it (Appendix B.5).

Figure 2 works three real pairs from that probe through the agency channel, one per regime. The *update* case rewrites “I just went for a shopping spree and bought a \$2000 bag” (*guilty*) so that the purchase was planned and discounted. The extracted state then withdraws the self-attribution exactly as the gold graph does. The *zero update* case is the one to read carefully. The model’s state does not move because the lexical proxy detects no agency on either side, so the pair joins the zero-

Model	SCS (Static Alignment)		DCS (Dynamic Sensitivity)		
	Direct	Scaffolded	VAD-shift	VAD+Agency	N
Claude 4.5 Sonnet	0.522	0.570	0.699	0.692	381
Llama 3.1 8B [†]	0.535	0.556	0.701	0.648	380
Qwen3 30B [†]	0.514	0.549	0.692	0.652	382
GPT-5-mini	0.505	0.557	0.668	0.626	242
GPT-5	0.500	0.545	0.672	0.611	240
Gemini 2.5 Flash	0.530	0.578	0.650	0.609	382
GPT-5-nano	0.500	0.542	0.620	0.595	240

Table 3: Model performance on EmoReason-GE. **SCS** measures static structural alignment under two prompt strategies; **DCS** measures dynamic sensitivity to counterfactual causal shifts, conditioned on the gold intervention type. [†]: open-source models.

Model	Overall	VAD-shift	VAD+Agency	N	Model	Total Δ	Appr. Δ	Prior Δ
Claude 4.5 Sonnet	0.478	0.632	0.581	2000	GPT-5	+3.23	+2.65	+0.58
GPT-5-mini	0.459	0.619	0.595	1387	GPT-5-nano	+3.11	+2.48	+0.64
Gemini 2.5 Flash	0.441	0.577	0.538	1999	GPT-5-mini	+2.80	+2.21	+0.59
GPT-5	0.436	0.605	0.547	1379	Claude 4.5 Sonnet	+2.35	+1.82	+0.54
Qwen3 30B A3B [†]	0.433	0.594	0.552	2000	Qwen3 30B [†]	+2.30	+1.71	+0.59
Llama 3.1 8B [†]	0.422	0.578	0.537	2000	Llama 3.1 8B [†]	+2.25	+1.60	+0.66
GPT-5-nano	0.352	0.531	0.501	1384	Gemini 2.5 Flash	+1.89	+1.31	+0.57

Table 4: DCS on EmoReason-ED, conditioned on the gold intervention type. **Overall** averages over all four types including no-change pairs (the N column). The other two columns condition on the gold label (n per model: 410–591 and 250–371). Agency-isolated estimates are withheld because the lexical proxy’s gold-side delta is undefined on 52% of those pairs (Appendix B.5). [†]: open-source.

update mass without licensing any claim about the model’s reasoning (Appendix B.5). The *partial update* case shows what the continuous score buys over the binary indicator: withdrawing *self* without naming a replacement sits 45° off the gold update and scores 0.71.

Structural elaboration Because our chosen insertion penalty (0.5) is asymmetric, SCS explicitly accommodates the tendency of models to output more nodes overall than humans (Total Diff > 0), a trend we refer to as “structural elaboration.” Node-distribution analysis confirms the pattern across all models (Table 5), where LLMs systematically externalize more nodes. Specifically, 77% of these additional nodes were of the Appraisal type.

We interpret this as an *annotation granularity mismatch*, in which LLMs overly explicate logical intermediate steps that human annotators, trained to compress explanations into a highly canonicalized schema, naturally leave implicit (e.g., explicitly reasoning about team affiliation when a Quarterback is mentioned, as detailed in the “QB” example in

Table 5: Structural Elaboration (EmoReason-GE, Scaffolded Prompting). Mean difference in nodes (AI – Human). The Appraisal Diff shows most over-structuring occurs in cognitive steps. [†]: open-source.

Appendix B.1).

All models showed significant structural elaboration on EmoReason-GE (Table 5), with GPT-5 showing the largest increase (+3.23 nodes), followed by GPT-5-nano and GPT-5-mini. Gemini 2.5 Flash showed a smaller increase, while open-source models (Qwen, Llama) fell between the proprietary families.

One-sample t -tests confirm significant structural elaboration for all seven tested models ($p < 0.001$). This suggests that over-structuring may be partially inherent to autoregressive text generation, though the severity may be amplified by downstream alignment artifacts. The pattern also varies by domain, since on EmoReason-ED elaboration is reduced for some models (Appendix B.8), indicating that the degree of over-structuring interacts with the discourse characteristics of the source text. In general, we observe that LLMs tend to explicate intermediate cognitive steps that humans leave implicit.

Graph interventions as behavioral evidence To investigate whether explicit graphs represent active behavioral dependencies rather than post-hoc rationalizations, we conducted a graph intervention experiment (Table 6). We provided only the structural text representation of the graph (withholding

Intervention Type	GPT-5	Llama 3.1 8B
Unfairness	+56.60	+31.80
Coping Deficit	+40.40	+27.00
Self-Agency	+35.60	+23.20
Average (targeted)	+44.20	+27.33
Random appraisal (control)	+21.60	+14.40

Table 6: Emotion Shift Rate (%) after graph intervention on Appraisal nodes ($N = 500$).

the original context) to models, intervening solely on the Appraisal node’s string.

Targeted substitutions shifted predictions at roughly twice the random-control rate for both models (‘Unfairness’: 56.6% on GPT-5).

8 Discussion

Prompting schema compliance and open-source competitiveness Because the scaffolding gain measures schema compliance rather than latent reasoning, the SCS ranking tracks how readily a model adopts an imposed causal format. Open-source models match the proprietary families on that ranking and on combined-shift DCS, which suggests that structural schema compliance and basic causal sensitivity are not exclusive properties of massive closed-weight architectures but generalizable properties of instruction-tuned models.

Graph efficacy versus post-hoc rationalization

A recurring question in explainable AI is whether explicit reasoning traces act as functional drivers of the final generated output, or merely as post-hoc linguistic rationalizations. The intervention above (Appendix B.7) bears on it, within limits. This behavioral manipulation does not establish mechanistic causality in the latent weights, and it does not rule out post-hoc fabrication during unconstrained generation. It does show that models *can* and *do* use explicit causal graphs as active conditional cues for forward prediction, which supports our graph abstraction as a functionally relevant diagnostic proxy.

The diagnostic sufficiency of the VADA proxy

Scherer’s appraisal model includes many cognitive dimensions (novelty, goal relevance, and others), and tracking all of them across varying domains introduces excessive noise in generative text. The compressed **VADA space**, which isolates Agency on its own explicit axis, trades exhaustive tracking for a simplified, stable proxy. Our empirical results show this compression is viable on the affective

axes. VAD-shift DCS is computable at scale and separates domains cleanly, while the agency axes expose the limits of lexical proxies in multi-turn discourse. Appendix B.5 quantifies that instrument-level boundary, which we treat as itself diagnostic. Because the pilot’s Likert elicitation did not discriminate VAD-manipulated pairs from the remaining pairs (Section 5.2), we name a pairwise forced-choice comparison as the replacement elicitation instrument for future human grounding of the VAD channel.

Boundary cases and next steps Two follow-ups fall directly out of the framework’s stated boundaries. First, the 31–42% of candidates rejected during construction are predominantly cases whose causal chain is left implicit, either because the emotion or its trigger is never verbalized or because the connective between them is carried by context rather than an overt discourse marker, so no explicit canonical graph can be fixed without speculative annotation (a worked rejected example appears in Appendix B.10). Extending SCS and DCS to this implicit regime is the most natural next benchmark, with a salience-weighted parser prior as the concrete first step. Second, on the measurement side, the audit in Appendix B.5 makes a coreference-aware extractor the most important upgrade, since pronoun resolution in multi-turn dialogue is precisely where the lexical agency channel loses coverage; a dedicated human-validation study of DCS is the corresponding metric-side step.

9 Conclusion

We introduce a framework to evaluate Structural Fidelity over explicit emotion explanations using **SCS** (static alignment) and **DCS** (dynamic sensitivity). SCS is human-grounded, tracking human structural judgments at $\rho = 0.65$, ahead of a GPT-5-as-judge baseline. The Agency component of DCS carries a three-rater validation ($\alpha = 0.914$, $\rho = +0.39$). Across seven LLMs, every model over-structures relative to human baselines on EmoReason-GE (+1.9 to +3.2 nodes), an effect that attenuates on ED (Appendix B.8), and sensitivity to shifts in causal agency proves fragile in multi-turn dialogue, a magnitude we bound with an explicit audit of the lexical proxy (Appendix B.5). This shifts the evaluation focus from surface-level textual similarity to structured causal logic, with validated and released tools for auditing the explicit reasoning alignment of affective AI.

Limitations

Our methodology necessitates several explicit scope boundaries. First, **Data Selection**: Our stringent annotation filtering for overt reasoning resulted in a 31–42% sample rejection rate. That rate is deliberate, since EmoReason-4k targets explicitly formalizable reasoning and omits highly latent implicit sequences to keep the benchmark a clean diagnostic probe. Consequently, our metrics measure fidelity to an explicit structural abstraction. They do not reach the fluid, latent cognitive mechanisms of humans. The benchmark makes no claim about implicit or inferential emotion attribution, which is the majority case in unconstrained text.

Second, **Extraction Pipeline**: We evaluate reasoning via an LLM pipeline extraction. SCS therefore measures *relative structural divergence* through a shared evaluation lens. It does not recover absolute distance to an unobservable ground truth. Findings like “structural elaboration” partially reflect an annotation granularity mismatch, where LLMs deviate from human compression conventions, and need not indicate a definitive cognitive flaw.

Third, **Metric Heuristics**: SCS utilizes theory-driven normative parameters (e.g., omission penalties), and DCS collapses high-dimensional dynamic states into a computationally tractable VADA proxy. Both trade semantic nuance for evaluations that are interpretable and stable, and both are choices we could have made differently.

Fourth, **Agency Measurement**: the lexical agency proxy has limited coverage in multi-turn dialogue (audited in Appendix B.5), so per-model agency-channel DCS on EmoReason-ED is instrument-bounded and is not reported in isolation. The DCS human-grounding pilot is likewise bounded: one dataset, 152 pairs, three raters.

Ethics Statement

This research involved the use of public datasets (GoEmotions and EmpatheticDialogues) and the generation of graph structures via zero-shot prompts. While we did not conduct psychological experiments on human subjects, we employed expert annotators for data verification. All annotators were compensated fairly at a professional hourly rate and provided informed consent for the data creation task. Our results show that directional tracking degrades in multi-turn dialogue, which is

the setting mental health and social advice applications actually occupy. Models may misattribute responsibility precisely where the exchange is most involved. Structural evaluation metrics like DCS can audit such systems before deployment.

Acknowledgments

This work is part of the Cognition-aware AI project by Hajime Institute. We thank the anonymous reviewers for their insightful comments.

Data Release and Licensing

The EmoReason-4k annotations and scripts will be released under CC-BY-4.0 to facilitate research reproducibility. The datasets include human-annotated causal graphs (2,000 samples each for GE and ED) with canonical structures, counterfactual pairs (2,000 pairs each) with human-authored edits and expert-validated intervention labels, and source text from GoEmotions (Apache 2.0) and EmpatheticDialogues, whose derived files retain the source CC BY-NC 4.0 license. The release further includes the 152-pair three-rater DCS pilot (ratings CSV and annotator instructions), the multi-seed counterfactual-regeneration run behind Table 8, the cross-extractor runs of Section 5.3, the figure-regeneration scripts, the analysis scripts, the state-proxy audit scripts of Appendix B.5, and a one-command reproduction of Table 3.

All source text is publicly available. We have ensured no personally identifiable information or toxic content is propagated in our human-annotated artifacts. The datasets will be hosted on a public repository (<https://github.com/Hajime-Institute/emoreason-4k>) with full documentation.

References

- Wei Ai, Fuchen Zhang, Yuntao Shou, Tao Meng, Haowen Chen, and Keqin Li. 2025. [Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(11):11418–11426.
- Lisa Feldman Barrett. 2017. [The theory of constructed emotion: an active inference account of interoception and categorization](#). *Social cognitive and affective neuroscience*, 12(1):1–23.
- Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang

- Peng, and Alexander Hauptmann. 2024. [Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning](#). *Advances in Neural Information Processing Systems*, 37:110805–110853.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. [Goemotions: A dataset of fine-grained emotions](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.
- Soumya Dutta and Sriram Ganapathy. 2025. [LLM supervised pre-training for multimodal emotion recognition in conversations](#). In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Yulin Wu, and Bingquan Liu. 2025. [Be-MERC: Behavior-aware MLLM-based framework for multimodal emotion recognition in conversation](#). *Preprint*, arXiv:2503.23990.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, and 7 others. 2020. [Evaluating models’ local decision boundaries via contrast sets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323.
- Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. [Causal abstractions of neural networks](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586.
- Jian Huang, Yuanyuan Pu, Dongming Zhou, Jinde Cao, Jinjing Gu, Zhengpeng Zhao, and Dan Xu. 2024. [Dynamic hypergraph convolutional network for multimodal sentiment analysis](#). *Neurocomputing*, 565:126992.
- Alon Jacovi and Yoav Goldberg. 2020. [Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205.
- Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2020. [Learning the difference that makes a difference with counterfactually-augmented data](#). In *International Conference on Learning Representations*.
- Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiuotė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. [Measuring faithfulness in chain-of-thought reasoning](#). *Preprint*, arXiv:2307.13702.
- Jiangnan Li, Fandong Meng, Zheng Lin, Rui Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. 2022a. [Neutral utterances are also causes: Enhancing conversational causal emotion entailment with social commonsense knowledge](#). In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, pages 4209–4215.
- Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumeng Chen. 2022b. [Knowledge bridging for empathetic dialogue generation](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10993–11001.
- Yuxiang Lin, Jingdong Sun, Zhi-Qi Cheng, Jue Wang, Haomin Liang, Zebang Cheng, Yifei Dong, Jun-Yan He, Xiaojiang Peng, and Xian-Sheng Hua. 2025. [Why we feel: Breaking boundaries in emotional reasoning with multimodal large language models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5196–5206.
- Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. [EmoLLMs: A series of emotional large language models and annotation tools for comprehensive affective analysis](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487–5496.
- Tania Lombrozo. 2006. [The structure and function of explanations](#). *Trends in cognitive sciences*, 10(10):464–470.
- Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. [Towards faithful model explanation in NLP: A survey](#). *Computational Linguistics*, 50(2):657–723.
- Saif Mohammad. 2018. [Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words](#). In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 174–184.
- Minh-Tien Nguyen, Huu-Loi Le, Manh-Cuong Phan, and Hajime Hotta. 2026. [CMTD: Cognitive modeling with traits and distortions for multimodal emotion recognition in conversations](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 839–854, San Diego, California, United States. Association for Computational Linguistics.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. [Towards empathetic open-domain conversation models: A new benchmark and dataset](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381. ACL.
- Klaus R. Scherer, Angela Schorr, and Tom Johnstone, editors. 2001. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.

Yuntao Shou, Tao Meng, Wei Ai, and Keqin Li. 2025. [Dynamic graph neural ODE network for multi-modal emotion recognition in conversation](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 256–268.

Zhaoyue Sun, Hainiu Xu, Andero Uusberg, James J. Gross, Petr Slovak, and Yulan He. 2026. [CAREBench: Evaluating LLMs’ emotion understanding by assessing cognitive appraisal reasoning](#). *Preprint*, arXiv:2605.17176.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. [Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965.

Botao Wang, Keke Tang, and Peican Zhu. 2024. [Enhancing emotion-cause pair extraction in conversations via center event detection and reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10773–10783.

Sarah Wiegrefe and Ana Marasović. 2021. [Teach me to explain: A review of datasets for explainable natural language processing](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Chengyan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. 2025. [Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6257–6274.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah D. Goodman. 2023. [Interpretability at scale: Identifying causal mechanisms in alpaca](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 78205–78226.

Yang Yang, Xunde Dong, and Yupeng Qiang. 2025. [MSE-adapter: A lightweight plugin endowing LLMs with the capability to perform multimodal sentiment analysis and emotion recognition](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25642–25650.

Gerard Christopher Yeo and Kokil Jaidka. 2025. [Beyond context to cognitive appraisal: Emotion reasoning as a theory of mind benchmark for large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26517–26525.

Weixiang Zhao, Yanyan Zhao, Zhuojun Li, and Bing Qin. 2023. [Knowledge-bridged causal interaction network for causal emotion entailment](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):14020–14028.

Xiaofei Zhu, Jiawei Cheng, Zhou Yang, Zhuo Chen, Qingyang Wang, and Jianfeng Yao. 2024. [CMATH](#):

[Cross-modality augmented transformer with hierarchical variational distillation for multimodal emotion recognition in conversation](#). *Preprint*, arXiv:2411.10060.

A Additional Experimental Details

A.1 Greedy vs. Hungarian Matching Comparison

To validate our choice of greedy matching for node alignment, we compared it against the optimal Hungarian assignment algorithm on a subset of 200 graph pairs from our dataset. The Hungarian algorithm theoretically guarantees optimal assignment (minimum total cost) for bipartite matching problems. Our comparison showed that greedy matching achieved nearly identical results to Hungarian matching: mean cost difference = 0.0032 (Greedy: 2.8080, Hungarian: 2.8048), with no statistically significant difference ($t = 1.44, p = 0.17$). The effect size (Cohen’s $d = -0.0064$) indicates a negligible practical difference. SCS scores were also nearly identical (Greedy: 0.7920, Hungarian: 0.7922, mean difference = 0.0002, $p = 0.17$).

This result validates that greedy matching provides an excellent approximation to the optimal solution for our small graphs (3-7 nodes). The semantic similarity constraints (SentenceBERT embeddings) naturally guide matching quality, reducing the need for global optimization. We chose greedy matching for computational efficiency ($O(n^2 \log n)$ vs. $O(n^3)$) and implementation simplicity, with the empirical validation that it achieves results statistically indistinguishable from the optimal Hungarian algorithm.

A.2 Graph Extraction Details

All explanations were converted to causal graphs using a fixed extraction model (GPT-5, temperature=0.0) with a standardized parsing prompt. This ensures that differences in SCS scores reflect reasoning differences rather than extraction artifacts. The extraction prompt follows a structured format that requires nodes of types {Cue, Prior, Appraisal, Emotion} and edges representing causal relations. The extraction process includes automatic validation for schema consistency and cycle detection.

To validate extraction robustness, we built an independent secondary pipeline using Claude 4.5 Sonnet and conducted a multi-extractor ablation. We found that instance-level SCS scores exhibited a highly significant cross-extractor correlation

(Pearson $r = 0.687, p < 0.001$). Claude 4.5 Sonnet extracted the logic with a slightly higher average SCS (0.540) compared to the GPT-5 baseline (0.498). Agreement was also high in the top and bottom quartile rankings. This demonstrates that SCS remains robust across structurally different proprietary parsers and does not introduce a localized systematic bias.

A.3 VAD Mapping Table

Emotion labels from both GoEmotions and EmpatheticDialogues were mapped to VAD coordinates using the NRC VAD Lexicon (Mohammad, 2018). For labels not directly in the lexicon (approximately 18% of cases across both datasets), we used theoretical mappings based on Appraisal Theory, validated by two psychology experts (agreement rate = 0.85). The full mapping table is provided in the supplementary materials. Key mappings include: *joy* $\rightarrow$ [V=0.95, A=0.85, D=0.80], *anger* $\rightarrow$ [V=0.15, A=0.90, D=0.80], *sadness* $\rightarrow$ [V=0.20, A=0.35, D=0.30], *fear* $\rightarrow$ [V=0.20, A=0.90, D=0.20].

The theoretical mapping process followed these principles: (i) Valence reflects pleasantness (positive emotions: $V > 0.5$, negative: $V < 0.5$), (ii) Arousal reflects activation level (high-arousal emotions like anger/fear: $A > 0.8$, low-arousal like sadness: $A < 0.4$), and (iii) Dominance reflects control/power (high-dominance emotions like anger/pride: $D > 0.7$, low-dominance like fear/embarrassment: $D < 0.3$). Complex labels (e.g., "Anger toward her") were normalized to primary emotions ("anger") before mapping.

B Extended Results and Analyses

Baseline	SCS	DCS
Identity ¹	1.000 / 0.996	–
Shuffled Labels	0.570	–
Random Graph	0.248	–
Aligned Change	–	1.000
Random Change	–	0.005

Table 7: Metric validation. SCS and DCS scores on controlled baselines.

B.1 Qualitative Analysis: Structural Elaboration and Errors

We present a qualitative comparison illustrating “structural elaboration” (verbosity mismatch) and hallucination errors.

In many instances where LLMs generated significantly more nodes than human annotators (e.g., +6 nodes in Total Diff), the additional nodes were often semantically plausible but over-explicated steps that humans considered implicit. For example, given the cue “*The QB said they loved it*” (Target emotion: *Joy*), the human canonical graph concisely maps the cue to the appraisal “*achievement of desired goal*.” Conversely, the LLM extracted a chain of intermediate inferences: *Cue* $\rightarrow$ “*The speaker supports the QB’s team*” $\rightarrow$ “*The QB’s approval indicates a positive outcome for the team*” $\rightarrow$ “*achievement of desired goal*” $\rightarrow$ *Joy*.

While logically coherent (resulting in high Emotion Interpretation scores), this represents a *verbosity mismatch* relative to normal human cognitive reporting. Importantly, SCS penalizes this behavior equally alongside more severe failure cases (Reasoning Hallucination), where models occasionally hallucinated non-existent Prior Knowledge nodes to justify an incorrect emotion prediction, or misassigned Agency (e.g., attributing internal self-agency to an external meteorological event) to force compliance with a generated appraisal schema.

B.2 Counterfactual Quality

We verified the quality of our counterfactual pairs through human validation. Our expert annotators independently confirmed that 96% of pairs represented valid causal interventions, where validity was defined as: (i) text remains grammatically and semantically coherent, and (ii) the graph update correctly reflects the intended causal variable change (e.g., agency shift). Quantitative analysis on EmoReason-GE showed that 79% of pairs were VAD-only (modified affective states without altering agency) and 21% were VAD+Agency, showing precise variable control. On EmoReason-ED, the distribution differed: 30% VAD-only, 18% VAD+Agency, 15% Agency-only, and 37% No-change, reflecting the different discourse characteristics of empathetic dialogues.

B.3 Multi-Seed Confidence Intervals

To characterize the irreducible noise floor of counterfactual-based DCS, we regenerated counterfactual explanations with multiple seeds on an EmoReason-ED subsample and recomputed DCS per seed ($n = 148$ DCS values per model). Table 8 reports per-model means with bootstrap 95% CIs. CI half-widths of 0.076–0.090 imply that per-model DCS gaps below ~ 0.10 lie within the run-

to-run noise band; we accordingly read the DCS columns of Table 3 at the level of coarse tiers rather than strict rankings.

Model	Mean DCS	95% CI
Claude 4.5 Sonnet	0.508	[0.427, 0.585]
Gemini 2.5 Flash	0.503	[0.422, 0.585]
GPT-5	0.471	[0.390, 0.551]
GPT-5-mini	0.471	[0.386, 0.550]
Qwen3 30B A3B [†]	0.438	[0.361, 0.515]
Llama 3.1 8B [†]	0.437	[0.360, 0.512]
GPT-5-nano	0.436	[0.345, 0.526]

Table 8: Multi-seed counterfactual regeneration on EmoReason-ED ($n = 148$ DCS values per model): over-all DCS with bootstrap 95% CIs. [†]: open-source.

B.4 DCS Distribution Analysis

On EmoReason-GE, mean DCS values (0.61–0.70) are significantly above zero ($p < 0.001$), with standard deviations of 0.37–0.42 reflecting substantial per-sample variation. On EmoReason-ED, DCS values are lower (0.35–0.48) but still significantly positive. The higher standard deviations on ED reflect the greater diversity of counterfactual types in dialogue contexts. Across both datasets, 83–87% of pairs exhibit positive directional agreement ($DCS > 0$), while negative DCS values (3–7% on GE, 5–10% on ED) indicate directionally inverted updates that may reflect systematic reasoning errors in specific contexts.

B.5 Auditing the DCS State Proxy

We audited the lexical state proxy that produces $S(G)$, with three findings. **(a) Keyword matching.** The agency detector matches unbounded substrings, so function words can trigger spurious agency flags (e.g., *the* contains *he*). On EmoReason-ED gold graphs, 60.8% of other-agency flags and 88.1% of self-agency flags have no genuine pronoun or accountability marker behind them, and 69–81% of model-extracted graphs saturate to a constant other-agency flag; a constant flag cancels in the orig→CF delta and mutes the channel. **(b) Coverage.** On the single-model ED probe (GPT-5-mini, $N=1,000$ pairs), the agency channel carries a detectable signal on 37% of pairs (VAD: 100%). Of the pairs registering zero agency update (75.7%), 83.2% carry no detectable agency signal on either side, and restricting to signal-bearing pairs reverses the VAD-versus-agency ordering (34.3% vs. 49.7% zero-update). The zero-update spike in Figure 3 therefore visualizes the

proxy’s detection envelope in multi-turn discourse at least as much as model behavior, consistent with the operational-failure-mode reading in the main text. **(c) Undefined-delta semantics.** When either update vector is $\mathbf{0}$, the 5-D cosine is undefined; the implementation returns 1.0 when both are zero and 0.0 otherwise. On agency-isolated ED pairs, 52% of gold-side deltas are zero under the lexical detector, so roughly half of those DCS values are forced constants rather than measured cosines. For these reasons we do not report per-model agency-channel DCS estimates in isolation; Table 4 conditions on intervention types whose gold-side deltas are well-defined. The audit scripts ship with the data release.

B.6 VAD Mapping Robustness

To validate the robustness of our VAD (Valence, Arousal, Dominance) mapping for DCS calculation, we compared the NRC VAD Lexicon values and Russell’s circumplex model. We evaluated 424 emotion labels from our dataset and found strong correlations across all VAD dimensions: Valence ($r = 0.996, p < 0.001$), Arousal ($r = 0.992, p < 0.001$), and Dominance ($r = 0.986, p < 0.001$). These high correlations confirm that our VAD mapping is robust to alternative theoretical assignments, suggesting that DCS scores are not overly sensitive to the specific VAD values chosen from the lexicon.

B.7 Graph Intervention Details

As summarized in Section 7, we sampled $N = 500$ ground-truth graphs from EmoReason-GE. We provided structural representations (nodes and edges) to models, entirely withholding the original text. We intervened by fixing all nodes/edges except the Appraisal node, replacing its string to reflect orthogonal psychological dimensions like *Unfairness* (“perceived injustice”), *Coping Deficit* (“inability to manage”), or *Self-Agency* (“personal responsibility”). Prediction from the graph alone recovers the canonical emotion on 82.8% (GPT-5) and 77.4% (Llama 3.1 8B) of samples, and the intervention results (Table 6) indicate models actively process these abstract causal structures.

B.8 Cross-Domain Validation (EmoReason-ED)

To validate that our findings generalize beyond social media text, we replicated all experiments on EmoReason-ED (EmpatheticDialogues). Key results confirm the robustness of our framework.

Metric validity SCS discriminates correctly on EmoReason-ED: Identity = 0.996², Shuffled = 0.390, Random = 0.189 (all $p < 0.001$). DCS: Aligned = 1.000, Random = 0.076 (≈ 0). In a blind generation condition (emotion label hidden), the evaluation pipeline reconstructed the reasoning structures of the ground-truth dialogues with a high average SCS of 0.716 ($N = 1,000$ paired graphs), confirming stable inferential structuring in conversational contexts.

Model	Mean SCS		SD	
	Direct	Scaffolded	Direct	Scaffolded
Gemini 2.5 Flash	0.4436	0.4678	0.1042	0.1095
Claude 4.5 Sonnet	0.4377	0.4633	0.1031	0.1056
Llama 3.1 8B [†]	0.4431	0.4600	0.0994	0.1064
Qwen3 30B A3B [†]	0.4275	0.4585	0.1002	0.1069
GPT-5-mini	0.4177	0.4433	0.0884	0.0967
GPT-5	0.4150	0.4428	0.0803	0.0977
GPT-5-nano	0.4164	0.4417	0.0896	0.0935

Table 9: Model performance on EmoReason-ED. [†]: open-source.

Model performance (SCS) The model ranking is consistent with EmoReason-GE: Gemini models lead, followed by Claude, open-source models, and GPT-5 family. Scaffolded outperforms Direct for all models ($p < 0.001$). Overall SCS scores are lower on ED (~ 0.42 – 0.48) than GE (~ 0.50 – 0.59), likely because empathetic dialogues involve more complex interpersonal dynamics.

Dynamic sensitivity by intervention type (DCS) Overall DCS exceeds 0.3 for every model (one-sample t -test, $p < 0.001$; the ED random-change floor is 0.076), and DCS is lower on ED than GE across models and conditions, consistent with the greater complexity of interpersonal emotion reasoning in dialogues. We do not report a per-model agency-channel estimate; Appendix B.5 explains why the lexical proxy cannot support one on this corpus.

Agency Stubbornness vs. Pronoun Heuristics On the single-model probe (GPT-5-mini, $N=1,000$ pairs), the mean sensitivity magnitudes for VAD (0.2551) and Agency (0.2550) appear nearly identical, yet their distributions diverge (Figure 3; Mann-Whitney U , $p < 0.001$): 75.7% of pairs register zero agency update versus 49.7% for VAD. The

²The marginal 0.004 slippage from a perfect 1.0 is due to edge mapping threshold variances in the greedy alignment algorithm for highly disjointed dialogue topologies.

instrument audit (Appendix B.5) attributes the bulk of this zero-mass to detection coverage rather than behavior: 83.2% of zero-agency pairs carry no detectable agency signal on either side (channel coverage 37% vs. 100% for VAD), and restricting to signal-bearing pairs reverses the ordering (34.3% vs. 49.7%). The spike is therefore read as the boundary of the lexical proxy in multi-turn discourse, entangled with any genuine model-side difficulty in updating locus of control: exactly the composite *operational failure mode* discussed in the main text, now quantified. This is also why the zero-update indicator is reported jointly with signal coverage (Section 4.2).

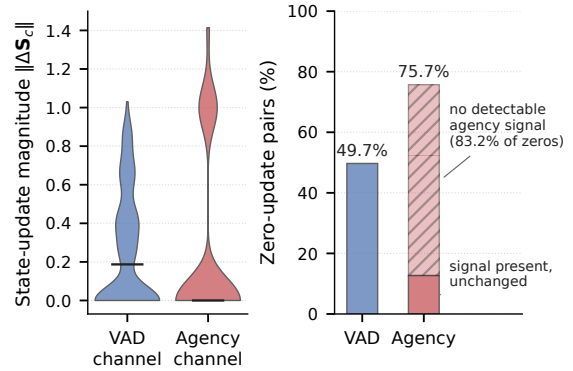


Figure 3: Channel state-update magnitudes $\|\Delta S_c\|$ on the EmoReason-ED single-model probe (GPT-5-mini, $N=1,000$): distributions with medians (left) and the zero-update share per channel (right). Mean magnitudes are nearly identical (≈ 0.255), but the Agency channel registers zero update on 75.7% of pairs versus 49.7% for VAD. The audit in Appendix B.5 attributes 83.2% of the agency zero-mass to pairs with no detectable agency signal on either side (hatched), so the spike visualizes the lexical proxy’s detection envelope as much as “Agency Stubbornness” itself.

Structural elaboration A notable cross-domain difference emerges in structural elaboration. While all models show positive Total Diff on EmoReason-GE (+1.89 to +3.23), several models show near-zero or *negative* Total Diff on EmoReason-ED (Gemini 2.5 Flash: -0.22 , Llama 3.1 8B: -0.01). GPT-5 family models still show positive elaboration on ED (+0.89 to +1.27), but much reduced from GE. This suggests that over-structuring is not a universal model property but depends on the interaction between model architecture and source text discourse structure.

B.9 Sensitivity Analysis of SCS Parameters

Because SCS utilizes several *a priori* normative hyperparameters for semantic alignment (e.g., omission cost = 1.0, elaboration cost = 0.5, edge mismatch cost = 0.5), we ran a sensitivity analysis to check that the model rankings carry interpretive weight.

We systematically perturbed all four structural parameters by $\pm 20\%$ using a grid search (e.g., varying insertion cost from 0.4 to 0.6, defining 81 independent metric configurations). Across all $\pm 20\%$ parameter perturbations, the relative ranking of models remained overwhelmingly stable (mean Spearman’s $\rho = 0.982$, min $\rho = 0.957$). While the absolute numerical SCS scores naturally shift when penalties are heightened or softened, the positional order of models (e.g., Gemini 2.5 Flash > Claude 4.5 Sonnet > Llama 3.1 8B > GPT-5 family) never reverses under any bounded perturbation. This confirms that our findings and the empirical gaps between models are robust to normative design choices.

B.10 Qualitative Error Analysis

Rejected candidates Construction rejected candidates at two stages: an automated screening pass over the candidate pool and the human verification pass whose 31–42% rejection rate Section 3 reports. Screening rationales are dominated by label–text mismatch, and a recurring secondary class comprises texts whose causal attribution requires inference beyond what is stated, the implicit class that the Limitations section scopes out. A representative rejected candidate reads “*I tried to write a cheque for a fare when I was a teenager, but the driver said it was fine and accepted it*”: the candidate graph attributed *Other-Accountability*, but that attribution rests on an unstated social reaction, so no explicit canonical graph can be fixed without speculative annotation. The screening records with per-candidate rationales ship with the release.

To better understand the reasoning limitations of the models, we conducted a qualitative failure analysis of the extracted graphs. We identified three primary categories of failure:

(a) Hallucinated Causal Edges: Models often generate causal links between nodes that are contextually proximate but logically independent. For instance, in EmoReason-ED, when a user describes feeling nervous about a job interview because of the rain, the model might incorrectly draw a direct

causal edge from ‘Rain’ to ‘Competence Appraisal’ rather than attributing the nervousness to the interview itself.

(b) Misassigned Agency Attribution: Across both datasets, LLMs frequently struggled to map causal responsibility correctly. In situations involving passive actions (e.g., “The package was delayed”), models display a strong prior toward self-agency or neutral-agency, failing to recognize external-other accountability, linking back to our observations of Agency Stubbornness.

(c) High SCS with Semantically Vacuous Reasoning: We observed cases where the model adheres perfectly to the Cue $\rightarrow$ Appraisal $\rightarrow$ Emotion schema (yielding a high Structural Consistency score), but the nodes themselves consist of generic boilerplates. For example, Appraisal nodes formulated as “Evaluation of the situation” rather than specific cognitive dimensions (e.g., “Goal Obstruction”). This highlights the value of human validation and embedding quality when scaling automated evaluation.