Why It Hurts: Identifying the Drivers of Negative Thoughts in Emotional Support Conversations

Hainiu Xu[♣] **Zhaoyue Sun**[♣]
Hanqi Yan[♣] **Jinhua Du**[♡] **Caroline Catmur**[♣] **Yulan He**[♣,♢]

[♣]King's College London
[♡] JinFlow Intelligence Ltd.
[♢]The Alan Turing Institute
{hainiu.xu, yulan.he}@kcl.ac.uk

Abstract

Large Language Models (LLMs) are increasingly used for emotional support tasks, such as negative thought reframing. This task relies on modifying cognitive appraisals, the subjective interpretation of events that elicit negative emotions, which is typically conceptualized along multiple discrete dimensions. Current LLM-based frameworks model cognitive appraisal by exhaustively evaluating all possible dimensions, but they fail to account for the varying saliency of these dimensions across different contexts. In this work, we investigate a vital yet overlooked question: "*Can LLMs infer the salient appraisal dimensions from emotional support conversations?*" To address this question, we introduce the AppraiSal benchmark, containing 996 emotional support conversations with human-annotated mental states, including salient cognitive appraisal dimensions. Furthermore, we propose PRISM, a multi-agent probabilistic framework grounded in Bayesian Inverse Planning, designed to improve LLMs' ability to identify context-specific appraisal dimensions. Experimental results show that PRISM brings improvements to LLMs across various sizes, particularly in identifying the most salient appraisal dimensions[1].

1 Introduction

Consider the event of a research paper rejection. A final-year PhD student may experience distress because the rejection blocks a critical goal (*they need this paper to graduate*) or the consequences are uncertain (*they do not know what will happen next*). In contrast, a researcher who attributes the rejection to reviewer misunderstanding or unfair evaluation may instead feel frustrated because someone else controlled the outcome. Each of these interpretations point to a different *cognitive appraisal*

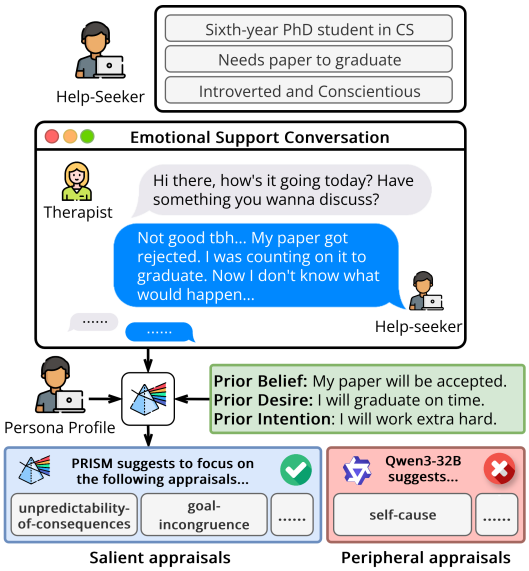

Figure 1: The salient appraisal reasoning task and the framework. Given a help-seeker's persona profile, ESC trajectory, and pre-event BDI, PRISM enables models to identify the salient appraisals more accurately (bottom left), whereas models without it often fail to recover the most relevant appraisals (bottom right).

(Arnold, 1960; Roseman and Smith, 2001; Ortony et al., 2022; Yeo and Ong, 2024), as exemplified in Figure 1. Cognitive appraisals are typically conceptualized along discrete dimensions such as *goal-incongruence*, *unpredictability-of-consequences*, and *unacceptable-consequences*[2].

Effective emotional support requires identifying which appraisals are actually *salient*. That is, what are the specific dimensions of subjective interpretation that are most responsible for a person's negative emotional response (Gross, 2015). A growing body of work incorporates cognitive appraisal theory into LLM-based frameworks to carry out empathic communications, such as cognitive reframing (Zhan et al., 2023, 2024; Zhou et al., 2025b). However, these approaches typically

[1]Data and code are available at https://github.com/seacowx/AppraiSal-and-PRISM.

[2]We use "appraisal dimensions" and "appraisals" interchangeably thereafter.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 16754–16797
October 24-29, 2026 ©2026 Association for Computational Linguistics

treat all appraisal dimensions as equally relevant, exhaustively evaluating every dimension for every situation, which may diminish the efficacy of the reframing process by failing to prioritize and focus on the most impactful appraisal dimensions.

The difficulty of the problem is that the salient appraisals are rarely stated explicitly in emotional support conversations (ESCs) and must therefore be inferred from an individual's utterances and contextual information (Troiano et al., 2023). Doing so requires reasoning about an individual's latent mental states (Ortony et al., 2022), such as their beliefs about the situation, their desired outcomes, and their intended actions, a capability commonly referred to as Theory of Mind (ToM). While much work has investigated LLMs' ToM capabilities (Ma et al., 2023; Chen et al., 2025a), largely focused on classic false belief tasks, relatively little attention has been paid to ToM reasoning in the context of ESCs. Consequently, a fundamental yet largely unexplored question remains: "*Can LLMs reliably identify salient appraisal dimensions from emotional support conversations?*"

To address this question, we introduce the **Appraisal Saliency** (AppraiSal) dataset. To the best of our knowledge, AppraiSal is the first benchmark designed to assess LLMs' ability to infer salient cognitive appraisals from ESCs. As an initial step toward improving LLMs' capability in this task, we propose **PR**obalistic **I**nference of **S**alient **M**ental States (PRISM), a multi-agent, probabilistic framework grounded in Bayesian Inverse Planning (BIP). Evaluation results from AppraiSal show that humans can consistently identify salient appraisal dimensions from ESCs while both open-source and proprietary models fall behind. While our PRISM framework brings significant improvements, particularly in the case of identifying the most salient appraisal, LLMs' ability to infer salient cognitive appraisal dimensions still falls behind that of humans. Furthermore, our proof-of-concept study (§5.1) shows that people generally prefer emotional support responses that are grounded in salient cognitive appraisal dimensions.

We summarize our contributions as follows:

1. We introduce AppraiSal, a novel benchmark for assessing LLMs' ability to infer salient cognitive appraisals from ESCs.
2. We propose PRISM, a multi-agent, probabilistic framework grounded in BIP for identifying salient appraisals from ESCs.
3. We provide extensive experiments and a proof-of-concept study to demonstrate the effectiveness of PRISM in inferring salient appraisals and the importance of salient appraisals in ESCs.

2 The AppraiSal Dataset

Task Formulation We formulate the central task supported by AppraiSal as *salient appraisal reasoning*. Consider an emotional support conversation between a help-seeker (h) and a therapist (t), represented as a dialogue trajectory $\mathcal{D} = \{(u_{h,1}, u_{t,1}), \ldots\}$, where $u_{h,1}$ denotes the help-seeker's utterance at the first turn. Given the dialogue trajectory, the help-seeker's persona information $\mathcal{P}$, and a predefined set of n cognitive appraisal dimensions $\mathcal{A} = \{a_1, \ldots, a_n\}$, the goal is to identify the appraisal dimensions most salient for explaining the help-seeker's negative emotional response. Formally, the model estimates the saliency of each appraisal dimension as $\mathbb{P}(a_i \mid \mathcal{D}, \mathcal{P}), \quad a_i \in \mathcal{A}$. It then returns the top-k (in AppraiSal, $k \leq 5$) appraisal dimensions with the highest estimated saliency:

$$\{\hat{a}_1, \ldots, \hat{a}_k\} = \text{Top-}k_{a_i \in \mathcal{A}}\, \mathbb{P}(a_i \mid \mathcal{D}, \mathcal{P}). \quad (1)$$

Cognitive Appraisals We define a set of appraisals based on the 21 appraisal dimensions given by Hofmann et al. (2020). Specifically, we removed dimensions that are redundant under ESC[3] and dichotomized dimensions to reduce ambiguity, which resulted in a set of 14 appraisal dimensions. See Appendix A for detailed construction of our appraisal dimensions.

2.1 Data Construction

Figure 2 provides an overview of the AppraiSal construction pipeline. Following prior works (Zheng et al., 2023a,b; Lee et al., 2024; Xiao et al., 2024; Zhou et al., 2025a), we first generate ESC trajectories using a dual-agent framework and subsequently recruit human annotators to evaluate dialogue quality and provide annotations.

Source Data & Context Construction Each conversation of AppraiSal centers around an upsetting event, which we seed from AugESC (Zheng et al., 2023a). To increase the fidelity of the original events, we construct a persona tailored for each

[3]Dimensions such as *pleasantness* and *unpleasantness* are redundant in ESC as all events in ESC are upsetting.

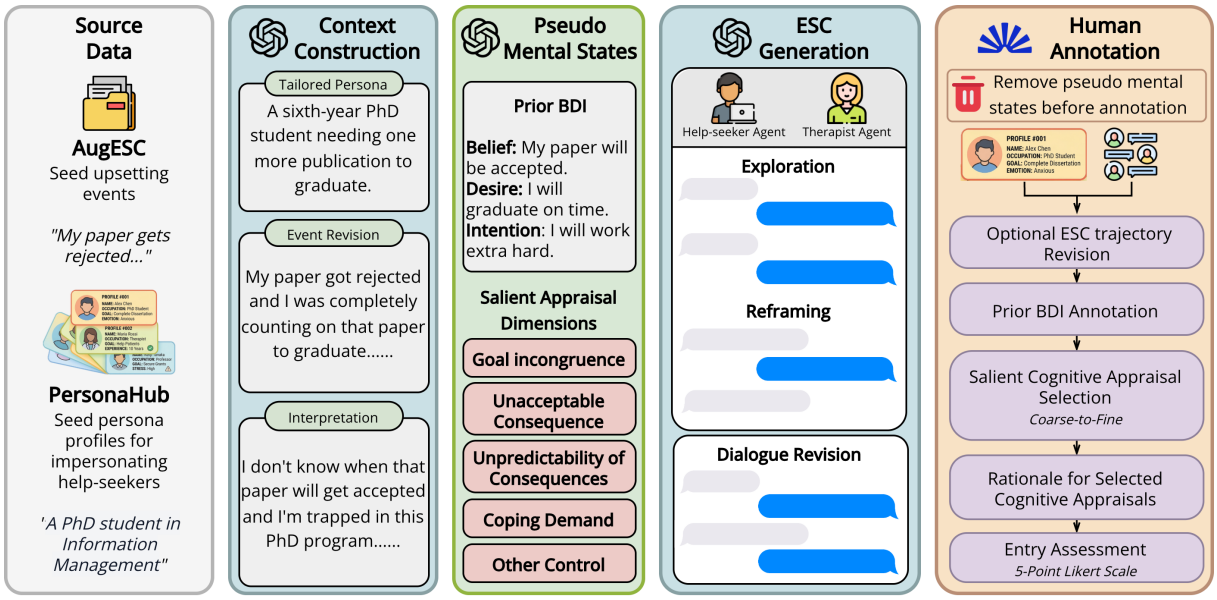

Figure 2: An overview of the AppraiSal construction process. Upsetting events and persona profiles are combined to construct persona-grounded contexts, from which pseudo mental states are generated to support synthetic ESC generation. The generated conversations are then revised, stripped of pseudo appraisal labels, and annotated by humans for pre-event BDI, salient appraisals, and quality, followed by filtering to obtain the final dataset.

event using profiles from the PersonaHub[4] (Ge et al., 2024). Each constructed persona includes basic demographic attributes such as name, gender, education level, and occupation, along with the Big Five personality traits. Based on the tailored persona profiles, we rewrite[5] the event from the help-seeker's perspective and generate the help-seeker's subjective negative interpretation using `GPT-5-mini`. See Appendix B for the detailed persona construction process.

Pseudo Mental States To ensure that the generated conversations are realistic and grounded, we construct pseudo mental states for each data instance. Building on the Ortony–Clore–Collins (OCC) model of emotion (Ortony et al., 2022), we first generate the help-seeker's *beliefs*, *desires*, and *intentions*[6] prior to the occurrence of the upsetting event (pre-event BDI), representing the help-seeker's mental state immediately before the upsetting event. We then prompt `GPT-5-mini` to generate 5 salient appraisals conditioned on the event, persona profile, and pre-event BDI. These machine-generated pseudo mental states are used to ground the ESC generation in the next step.

ESC Generation To produce high-quality ESC trajectories, we employ a two-stage generation pipeline. First, we use a dual-agent framework consisting of a help-seeker agent and a therapist agent to generate dialogue trajectories, with `gpt-oss-120b` (OpenAI, 2025) as the backbone model. The help-seeker agent is conditioned on the tailored persona, their subjective interpretation of the upsetting event, their pre-event BDI, and 5 salient appraisals. The therapist agent is guided by the standard Cognitive Behavioral Therapy (CBT) procedure (Wenzel, 2017). Specifically, the therapist begins with an *Exploration* stage to understand and evaluate the help-seeker's thoughts and beliefs. This is followed by a *Reframing* stage, in which the therapist attempts to reframe the help-seeker's negative thoughts using the *reAppraisal* framework[7] (Uusberg et al., 2023). The duration of each dialogue is determined autonomously by the help-seeker agent, which decides after each turn whether the conversation should continue.

After generating ESC trajectories, we perform trajectory revision using `GPT-5-Mini` with three objectives. First, each conversation must follow the Exploration–Reframing procedure typical of CBT. Second, the help-seeker must not voluntarily reveal their internal mental states unless explicitly asked. Third, the conversational style must correspond to the help-seeker's persona profile. For example, an *antagonistic* person is expected to be less cooperative than an *agreeable* person.

[4]We use the persona subset of PersonaHub.

[5]See Appendix C for examples of the rewritten entries.

[6]The OCC model does not consider *intentions*; we introduce them to make the underlying mental state more explicit.

[7]See Appendix D for the reAppraisal framework.

Jaccard Similarities				Krippendorff's α
A1|A2	A1|A3	A2|A3	Mean	
0.749	0.787	0.791	0.776	0.776

Table 1: Inter-annotator agreement results on AppraiSal. A1, A2, A3 correspond to the three annotators.

Human Annotation Once the ESC trajectories are revised, we remove the pseudo labels associated with the help-seeker's cognitive appraisals before annotation. Annotators are then provided with the help-seeker's persona profile and the generated ESC trajectory as context. Based on this information, each annotator performs four sub-tasks[8]

First, the annotator reads the dialogue and optionally revises it to improve coherence and realism. Second, the annotator infers the help-seeker's pre-event BDI, with the pseudo BDI provided as a reference. Third, the annotator identifies the salient cognitive appraisal dimensions using a coarse-to-fine protocol: the annotator first selects high-level appraisal groups and then refines the selection by choosing specific fine-grained dimensions. To discourage careless annotation, we require annotators to select exactly five appraisal dimensions they consider salient in the given context and to provide a rationale for each choice. Finally, the annotator rates the overall quality of the data entry on a 5-point Likert scale. See Appendix E for the demographic information of the annotators, quality assurance questions, and the annotation interface.

Data Statistics In total, we recruited 145 annotators from Prolific[9] who contributed 1,120 annotated entries. On average, each annotator was compensated with $1.52 per data entry. We further filter the data by removing entries with an average rating falling below 3.0, yielding the final AppraiSal with 996 entries. We present detailed data statistics in Appendix F. To measure inter-annotator agreement, we randomly sampled 50 entries, which were annotated by three annotators from Prolific. As the annotations come in as a set of appraisal dimensions, we assess their agreement using average Jaccard Similarity as well as Krippendorff's α. Results from Table 1 show that the annotators are able to achieve a relatively strong agreement in their judgment of salient cognitive appraisal dimensions.

[8]The annotation task received ethical clearance in 2026 and is granted for a period of 5 years.

[9]https://www.prolific.com

3 The PRISM Framework

As an initial step toward improving LLMs' ability to infer salient cognitive appraisals from ESC trajectories, we draw inspiration from recent ToM reasoning approaches (Jin et al., 2024; Ying et al., 2024; Kim et al., 2025; Zhang et al., 2025d) and propose PRISM, a multi-agent framework grounded in Bayesian Inverse Planning, as shown in Figure 3.

3.1 Preliminaries

Forward Generative Model We model the help-seeker's action (utterance) in an ESC as a generative process. Following the OCC model (Ortony et al., 2022), we assume that the help-seeker's utterance at each turn t reflects their understanding of the dialogue context $\mathcal{D}_{<t}$, their persona profile $\mathcal{P}$, and their latent mental states, including pre-event BDI $\mathcal{B}$ and salient cognitive appraisals $\mathcal{A}$. Formally, we model the help-seeker's utterance as:

$$\hat{u}_t \sim \mathbb{P}(u_t \mid \mathcal{D}_{<t}, \mathcal{P}, \mathcal{B}, \mathcal{A}). \tag{2}$$

Backward BIP Inference Based on the forward generative process, we derive the backward inference under the BIP framework as follows:

$$\begin{aligned}\mathbb{P}(\mathcal{A}, \mathcal{B} \mid \mathcal{D}_{1:t}, \mathcal{P}) \propto \mathbb{P}(\hat{u}_t \mid \mathcal{D}_{<t}, \mathcal{P}, \mathcal{B}, \mathcal{A}) \cdot \\ \mathbb{P}(\mathcal{B} \mid \mathcal{D}_{<t}, \mathcal{P}) \cdot \mathbb{P}(\mathcal{A} \mid \mathcal{D}_{<t}, \mathcal{P}, \mathcal{B}),\end{aligned} \tag{3}$$

where $\mathbb{P}(\hat{u}_t \mid \mathcal{D}_{<t}, \mathcal{P}, \mathcal{B}, \mathcal{A})$ denotes the likelihood of the help-seeker producing the current utterance given the dialogue history, persona profile, pre-event BDI, and cognitive appraisals. $\mathbb{P}(\mathcal{B} \mid \mathcal{D}_{<t}, \mathcal{P})$ denotes the posterior over pre-event BDI states given the dialogue history and persona profile up to turn $t-1$. $\mathbb{P}(\mathcal{A} \mid \mathcal{D}_{<t}, \mathcal{P}, \mathcal{B})$ denotes the posterior over appraisal sets given the dialogue history, persona profile, and pre-event BDI. A detailed derivation of Equation 3 is provided in Appendix G.

3.2 The PRISM Workflow

Hypothesis-Driven BIP Inference Most BIP-based ToM reasoning methods assume a confined and small label space. In AppraiSal, however, the model infers five salient appraisal dimensions from a set of 14 candidates, which leads to a combinatorial explosion of the hypothesis space. To constrain this search space, we first deploy a `hypothesis agent` to generate and manage candidate hypotheses, similar to Kim et al. (2025).

At the first turn of the ESC dialogue, the hypothesis agent is instructed to produce a batch of

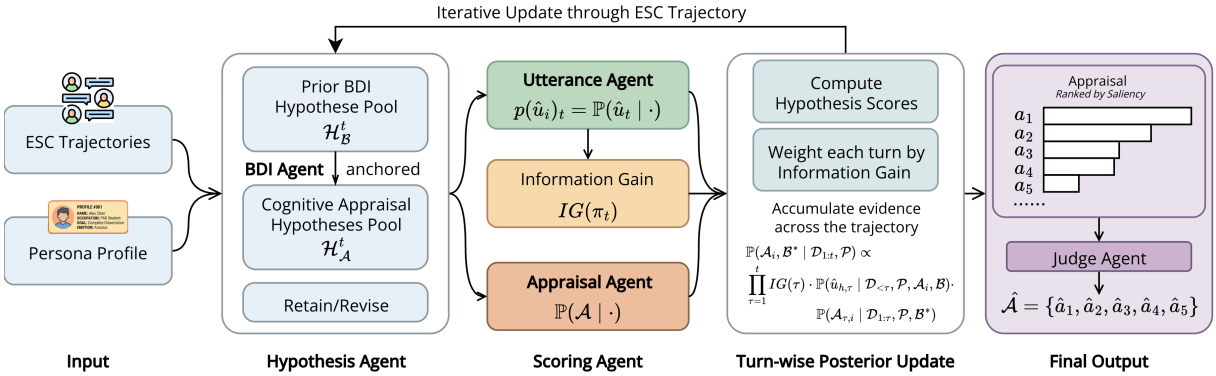

Figure 3: Overview of the PRISM framework. Given an ESC trajectory and the help-seeker's persona, PRISM performs hypothesis-driven Bayesian inverse planning over latent pre-event BDI and appraisal states. A hypothesis agent maintains candidate BDI and appraisal sets, three specialized agents score them turn by turn, and information gain weights the contribution of each dialogue turn to produce final saliency scores and the top-5 salient appraisals.

hypotheses. In the subsequent turns, instead of prompting the hypothesis agent to produce new sets of hypotheses, the agent is instructed to revise the existing hypotheses. Specifically, we define three actions that the hypothesis agent can make on each of the hypothesis: `Retain`, where the hypothesis is kept as-is; `Replace`, where the hypothesis is overwritten by a new one; and `Rerank`, where the rankings of the hypotheses are adjusted. Further, PRISM maintains two hypothesis pools, one for the pre-event BDI state $\mathcal{B}$ and one for the cognitive appraisal set $\mathcal{A}$, each of size $h = 8$. We denote these pools at turn t by $\mathcal{H}^t_{\mathcal{B}}$ and $\mathcal{H}^t_{\mathcal{A}}$.

Posterior Modeling Conditional distributions in Equation 3, namely $\mathbb{P}(\hat{u}_t \mid \cdot)$, $\mathbb{P}(\mathcal{B} \mid \cdot)$, and $\mathbb{P}(\mathcal{A} \mid \cdot)$, are instantiated by three LLM agents: the `utterance`, `bdi`, and `appraisal agent`.

For each agent, we instruct it to generate the target content associated with the conditional distribution it models. For instance, the `utterance agent` is tasked with generating the help-seeker's next utterance. The conditioning variables, namely $\mathcal{D}_{<t}$, $\mathcal{P}$, $\mathcal{B}_t$, and $\mathcal{A}_{t,i}$, are incorporated into the instruction prompt $\mathcal{I}$. Suppose the observed help-seeker's utterance at turn t consists of m tokens, denoted by $\hat{u}_t = \{w_1, w_2, \ldots, w_m\}$. We treat the observed utterance $\hat{u}_t$ as the target response and compute the log-likelihood score[10]:

$$\begin{aligned} p(\hat{u}_{t,i}) &= \log \mathbb{P}(\hat{u}_t \mid \mathcal{D}_{<t}, \mathcal{P}, \mathcal{B}_t, \mathcal{A}_{t,i}) \\ &= \sum_{j=1}^{m} \log \mathbb{P}\Big(w_j | w_{<j}, \mathcal{I}(\mathcal{D}_{<t}, \mathcal{P}, \mathcal{B}_t, \mathcal{A}_{t,i})\Big), \end{aligned} \tag{4}$$

[10]See Appendix H for details.

where $\mathbb{P}(w_j \mid w_{<j}, \mathcal{I})$ denotes the token-level probability obtained from the LLM. We compute $p(\hat{u}_{t,i})$ for each $\mathcal{A}_{t,i} \in \mathcal{H}^t_{\mathcal{A}}, i = 1, \ldots, h$, yielding a set of likelihood scores, each corresponding to an appraisal hypothesis. We apply the same procedure to the `bdi agent` and the `appraisal agent`.

Information Gain In ESC, the amount of information conveyed about the help-seeker's appraisals can vary vastly across dialogue turns. To account for this variation, we use Shannon entropy to quantify the turn-level informativeness. Specifically, given the set of utterance log-likelihood scores $\{p(\hat{u}_{t,i})\}_{i=1}^{h}$, we first convert them into a probability distribution over hypotheses:

$$\pi_{t,i} = \frac{\exp\big(p(\hat{u}_{t,i})\big)}{\sum_{k=1}^{h} \exp\big(p(\hat{u}_{t,k})\big)}. \tag{5}$$

We then define the information gain at turn t as

$$IG(\pi_t) = 1 - \frac{H(\pi_t)}{\ln h}, \tag{6}$$

where $H(\pi_t) = -\sum_{i=1}^{h} \pi_{t,i} \ln \pi_{t,i}$ is the Shannon entropy of the hypothesis distribution.

$IG(\pi_t)$ measures how much more informative the utterance likelihood distribution at turn t is than a uniform distribution over h hypotheses. Intuitively, a more informative utterance $\hat{u}_t$ is more sensitive to differences among appraisal hypotheses $\mathcal{A}_{t,i}$, thereby inducing a sharper distribution over $p(\hat{u}_t)$; less informative utterances, in contrast, yield distributions closer to uniform.

Overall Workflow The PRISM framework processes the ESC trajectory turn by turn, as illustrated in Figure 3. At each turn, the `hypothesis agent`

	Precision@1	Precision@2	Precision@3	Precision@4	Precision@5	Jaccard Index
Human	1.000	0.990	0.920	0.775	0.748	0.612
Random	0.361	0.342	0.350	0.358	0.357	0.234
Majority	0.755	0.655	0.561	0.530	0.522	0.379
Instruct Models						
`Qwen3-4B`	0.722	0.669	0.645	0.593	0.570	0.422
`Qwen3-32B`	0.758	0.706	0.686	0.648	0.611	0.465
`Qwen3-30B-A3B`	0.763	0.694	0.655	0.617	0.595	0.448
`Gemma3-4B`	0.438	0.574	0.602	0.582	0.566	0.416
`Gemma3-27B`	0.746	0.684	0.652	0.593	0.554	0.406
Reasoning Models						
`Qwen3-4B`	0.759	0.723	0.677	0.632	0.598	0.452
`Qwen3-32B`	0.739	0.709	0.676	0.636	0.608	0.461
`Qwen3-30B-A3B`	0.713	0.680	0.667	0.624	0.599	0.452
`gpt-oss-20B`	0.563	0.596	0.609	0.595	0.565	0.414
`gpt-oss-120B`	0.799	0.739	0.694	0.652	0.617	0.477
`DeepSeek-V3.2`	0.731	0.727	0.694	0.638	0.603	0.455
`GPT-5-Nano`	0.700	0.685	0.653	0.621	0.594	0.445
`GPT-5-Mini`	0.664	0.672	0.650	0.622	0.591	0.444
with the PRISM Framework						
`Qwen3-4B`	0.851 $_{0.129\uparrow}$	0.760 $_{0.091\uparrow}$	0.688 $_{0.043\uparrow}$	0.641 $_{0.048\uparrow}$	0.616 $_{0.046\uparrow}$	0.469 $_{0.047\uparrow}$
`Qwen3-32B`	0.820 $_{0.062\uparrow}$	<u>0.782</u> $_{0.076\uparrow}$	**0.724** $_{0.038\uparrow}$	**0.671** $_{0.023\uparrow}$	<u>0.629</u> $_{0.018\uparrow}$	<u>0.488</u> $_{0.023\uparrow}$
`Qwen3-30B-A3B`	**0.882** $_{0.119\uparrow}$	0.767 $_{0.073\uparrow}$	0.684 $_{0.029\uparrow}$	0.641 $_{0.024\uparrow}$	0.616 $_{0.021\uparrow}$	0.471 $_{0.023\uparrow}$
`Gemma3-4B`	0.734 $_{0.296\uparrow}$	0.705 $_{0.131\uparrow}$	0.660 $_{0.058\uparrow}$	0.621 $_{0.039\uparrow}$	0.584 $_{0.018\uparrow}$	0.441 $_{0.025\uparrow}$
`Gemma3-27B`	<u>0.878</u> $_{0.132\uparrow}$	**0.797** $_{0.113\uparrow}$	<u>0.723</u> $_{0.071\uparrow}$	<u>0.664</u> $_{0.071\uparrow}$	0.623 $_{0.069\uparrow}$	0.481 $_{0.075\uparrow}$
`gpt-oss-20B`	0.762 $_{0.199\uparrow}$	0.729 $_{0.133\uparrow}$	0.688 $_{0.079\uparrow}$	0.643 $_{0.048\uparrow}$	0.613 $_{0.048\uparrow}$	0.467 $_{0.053\uparrow}$
`gpt-oss-120B`	0.823 $_{0.024\uparrow}$	0.757 $_{0.018\uparrow}$	0.707 $_{0.013\uparrow}$	0.660 $_{0.008\uparrow}$	**0.632** $_{0.015\uparrow}$	**0.492** $_{0.015\uparrow}$
Avg. Change	+0.137	+0.091	+0.047	+0.037	+0.034	+0.037

Table 2: Evaluation results of AppraiSal on various LLMs and the PRISM framework. The **best** and <u>second-best</u> results are marked in bold and underlined, respectively. The average changes brought by the PRISM framework are shown in the last row. Performance gains from the PRISM framework are shown in the subscript. All improvements are statistically significant under the two-sample t-test with $p < 0.05$. See Appendix L for detailed test statistics.

maintains a pool of h hypotheses, denoted by $\mathcal{H}^t_{\mathcal{B}}$, over the help-seeker's pre-event BDI state $\mathcal{B}$. We first select the most likely pre-event BDI, $\mathcal{B}^*_t$, by taking the $\arg\max$ of $\mathbb{P}(\mathcal{B} \mid \mathcal{D}_{<t}, \mathcal{P})$, as estimated by the `bdi agent`. Conditioned on $\mathcal{B}^*_t$, we then generate hypotheses of salient appraisals as[11]

$$\mathcal{H}^t_{\mathcal{A}} = \texttt{LLM}(\mathcal{D}_{<t}, \mathcal{P}, \mathcal{B}^*_t). \tag{7}$$

Once the initial hypotheses are generated, the `hypothesis agent` performs one of two operations at each subsequent dialogue turn: *Retain* or *Revise*. The *Retain* operation preserves the original hypothesis, whereas *Revise* modifies it by replacing one or more appraisal dimensions.

We compute the unnormalized posterior score for each appraisal hypothesis set $\mathcal{A}_i \in \mathcal{H}^t_{\mathcal{A}}$ using Equation 3, scaled by information gain[12]:

$$\begin{aligned} &\mathbb{P}(\mathcal{A}_i, \mathcal{B} | \mathcal{D}_{1:t}, \mathcal{P}) \propto \prod_{\tau=1}^{t} IG(\tau) \cdot \\ &\mathbb{P}(\hat{u}_\tau | \mathcal{D}_{<\tau}, \mathcal{P}, \mathcal{B}, \mathcal{A}) \cdot \mathbb{P}(\mathcal{A}_{\tau,i} | \mathcal{D}_{1:\tau}, \mathcal{P}, \mathcal{B}^*_\tau), \end{aligned} \tag{8}$$

where $\mathcal{A}_{\tau,i}$ denotes the i-th hypothesis at turn τ. The cumulative score can therefore be computed recursively across dialogue turns, as the `hypothesis agent` does not overwrite previously generated hypotheses and hence maintains alignment between the hypotheses at turns $t-1$ and t.

After each turn, we flatten the hypothesis sets and propagate each set-level score to its contents, yielding saliency logits over appraisal dimensions. We then normalize the resulting logits across dimensions using the softmax function.

At the end of the ESC trajectory, we provide the appraisal dimensions and their posterior densities

[11] Our experimental results show that many LLMs can reliably infer the help-seeker's pre-event BDI from ESC trajectories. We present detailed results in Appendix I.

[12] We omit the $\mathbb{P}(\mathcal{B}_t \mid \cdot)$ term because we directly use the point estimate $\mathcal{B}^*_t$.

to a `judge agent`, which then selects the final set of five salient appraisal dimensions.

4 Experiments

Models We evaluate the AppraiSal dataset on a range of open-sourced and proprietary LLMs, covering both instruction models and reasoning models. Specifically, we use the `7B` and `27B` variants of the `Gemma3` model (Team et al., 2025), the `4B`, `32B`, and `30B-A3B` variants of the `Qwen3` model (Yang et al., 2025a), the `20B` and `120B` variants of the `gpt-oss` model (OpenAI, 2025), and DeepSeek V3.2-Thinking (DeepSeek-AI et al., 2025). For proprietary models, we use the `nano` and `mini` variants of `GPT-5` (Singh et al., 2025). For the PRISM framework, we disable *thinking* for all the reasoning models for computational efficiency.

Metrics As the salient appraisal reasoning task requires predicting a set of five appraisal dimensions, we evaluate performance using set-based metrics. Specifically, we report *Precision@*k and *Jaccard Index*. Note that we do not report *Recall* and *F1* as the *Recall* is equivalent to *Precision@5* for our task. Further, we do not consider rank-based metrics as our analysis shows that while annotators are able to reach good agreement on the set of salient appraisals, they do not tend to agree on the ranking of the appraisals. We provide detailed metric definition, justification, and rank-based evaluation in Appendix J.1 and J.2.

Main Results As shown in Table 2, humans perform strongly at identifying salient appraisals from ESC trajectories, particularly among the top three appraisals. In contrast, LLMs consistently underperform humans across all metrics. We observe two main failure modes. First, many LLMs perform reasonably well at identifying the single most salient appraisal dimension, as reflected by *Precision@1*, but their performance declines as the number of target appraisal dimensions increases. Second, some models, such as `Gemma3-4B` and `gpt-oss-20B`, fail to place the most salient appraisals at the top of the ranking, resulting in poor *Precision@1*. In these cases, the correct appraisals are often ranked second or third, causing performance to peak at *Precision@3*. Our error analysis shows that both failure modes are associated with LLMs' inherent bias in cognitive appraisal inference. We present the error analysis results in Appendix N.

Beyond zero-shot prompting, we also evaluate *Self-Consistency* prompting (Wang et al., 2023) by sampling 8 CoT trajectories from each LLM, matching the number of hypotheses used in PRISM. We also consider an *Iterative Self-Consistency with Final Judge* variant, where the model performs self-consistency sampling at each turn of the dialogue and then makes the final prediction using the same `Judge` agent as in PRISM. Both variants produce only marginal improvements, and we therefore defer the complete results to Appendix K.

Experimental results show that PRISM brings significant improvement for inferring the most salient appraisal dimensions, with an improvement of up to 0.296 for *Precision@1* and up to 0.133 for *Precision@2*. In addition, PRISM is particularly effective for small LLMs, such as `Qwen3-4B`, where PRISM boosts their performance across all metrics to exceed that of `gpt-oss-120B` as well as that of proprietary models such as `GPT-5-mini`.

5 Analysis

5.1 Salience-Guided Reframing

While the preceding results establish both the difficulty of salient appraisal inference and the effectiveness of PRISM, an important downstream question remains: *Does incorporating salient cognitive appraisals improve emotional support?* Prior works have shown that guiding LLMs with cognitive appraisals makes them more effective at negative thought reframing (Zhan et al., 2024); therefore, we conduct a proof-of-concept study to examine the role of *salient appraisals* in emotional support.

Following Zhan et al. (2024), we consider single-turn negative-thought reframing. We sample 100 situations from AppraiSal and generate four sets of reframing responses. *Ordinary* (no appraisal guidance), *Exhaustive* (all 14 appraisal dimensions), *Salience-Guided* (salient appraisals predicted by `Qwen3-4B`), and *Oracle-Guided* (human-annotated salient appraisals). To mitigate length bias, we restrict the model response length to 150 words.

We evaluate responses with human annotators and LLM-as-a-Judge. Human annotators compared paired responses based on three criteria inspired by the Cognitive Triangle (Beck et al., 2024), namely *Appraisal Alignment* (A.A), *Situation Understanding* (S.U.), and *Cognitive Reframing* (C.R.), while `GPT-5-mini` assigns Likert-scale scores for the same criteria. We provide the agreement of the human annotations in Appendix O.1 and details of LLM-as-a-Judge evaluation in Appendix O.2.

Agent	A.A.↑	S.U.↑	C.R.↑
Ordinary	$5.910_{\pm 1.524}$	$6.810_{\pm 1.294}$	$5.200_{\pm 1.400}$
Exhaustive	$\mathbf{6.760_{\pm 1.282}}$	$\underline{7.120_{\pm 1.042}}$	$\mathbf{6.420_{\pm 1.097}}$
Saliency-Guided	$\mathbf{7.020_{\pm 0.980}}$	$\mathbf{7.360_{\pm 1.063}}$	$\mathbf{6.430_{\pm 0.951}}$
Oracle-Guided	$\mathbf{7.070_{\pm 1.194}}$	$\mathbf{7.440_{\pm 1.061}}$	$\mathbf{6.450_{\pm 1.014}}$

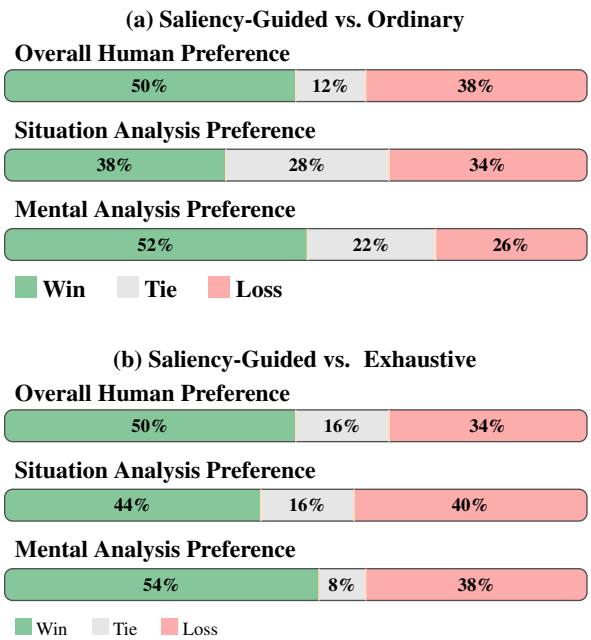

Table 3: Preference results for the case study. Top: LLM evaluation scores. (a): win/tie/loss rates of *Saliency-Guided* versus *Ordinary* emotional support. (b): win/tie/loss rates of *Exhaustive* versus *Saliency-Guided* reframing responses. **Bold** and underline indicate statistical significance at $\alpha = 0.05$ and $\alpha = 0.10$ respectively in two-sample T test.

As shown in Table 3, appraisal guidance significantly improves performance over unguided responses (i.e. *ordinary*) across all criteria.

Table 3(a) presents result of *Saliency-Guided* v.s. *Ordinary*. We see that human annotators show a clear preference for the *salience-guided* responses. While the inclusion of cognitive appraisals does not bring significant improvement in LLMs' *Situation Analysis*, they bring meaningful improvements in both *Overall Human Preference* as well as *Mental Analysis Preference*, which is the most vital component for emotional support.

Crucially, Table 3(b) shows that *Saliency-Guided* responses consistently outperform the *Exhaustive* variant, demonstrating that *selecting salient appraisals is more effective than using all appraisals*. Moreover, predicted saliency achieves performance close to *Oracle-Guided* responses, indicating that automatic appraisal selection can reliably approximate human judgment. These results confirm that

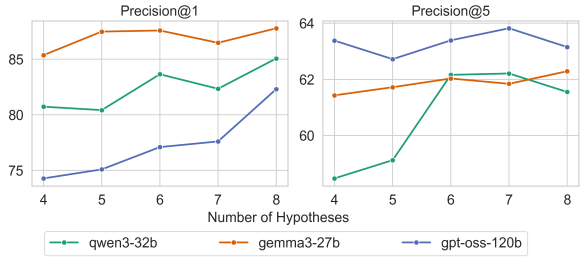

Figure 4: Change in model performance corresponding to the size of the hypothesis pool.

incorporating *salient* appraisals is key to improving emotional support.

5.2 Ablation Study

Size of Hypothesis Pool A key hyperparameter in PRISM is the size of the hypothesis pool. In our experiments, we set $h = 8$. To examine the effect of smaller pools, we vary h from 4 to 8. As shown in Figure 4, a larger hypothesis pool generally makes LLMs more reliable at identifying the most salient appraisals, as indicated by an increasing *Precision@1* with respect to h across all models. In the case of identifying 5 salient appraisals, `Gemma3-27B` and `gpt-oss-120B` give adequate performance even with small h, `Qwen3-32B` benefits more consistently from larger pools, particularly when $h \geq 6$.

PRISM Agents To understand the contribution of each component in PRISM, we conduct an ablation study using three backbone LLMs. Taking `gpt-oss-120B` as an example, results from Table 4 show that each component contributes to the overall performance, with the largest drops generally arising from removing the utterance likelihood term $\mathbb{P}(\hat{u}_t \mid \cdot)$. The importance of the utterance likelihood demonstrates the effectiveness of the BIP factorization (Equation 3). Taken together, these results show that the gains of PRISM do not come from a single component alone; rather, they arise from the interaction between prior-BDI anchoring, informativeness weighting, cumulative trajectory-level inference, and final judgment over appraisal saliency. See Appendix M for detailed results.

6 Related Work

Cognitive Reframing with LLMs A growing body of work has explored the use of LLMs for emotional support. One line of research approaches negative-thought reframing by prompting (Chen et al., 2023; Zhao et al., 2023; Zhang et al., 2024,

model	Precision@1	Precision@2	Precision@3	Precision@4	Precision@5	Avg ↓ in $P@k$	Jaccard Index
gpt-oss-120b	0.823	0.757	0.707	0.660	0.632	– –	0.492
w/o $\mathcal{B}^*$	$0.788_{-0.035}$	$0.736_{-0.021}$	$0.673_{-0.034}$	$0.628_{-0.032}$	$0.597_{-0.035}$	0.031 ↓	$0.455_{-0.037}$
w/o $IG(\cdot)$	$0.743_{-0.080}$	$0.739_{-0.019}$	$0.694_{-0.013}$	$0.651_{-0.009}$	$0.607_{-0.025}$	0.029 ↓	$0.463_{-0.029}$
w/o $\mathbb{P}(\hat{u}_t \mid \cdot)$	$0.725_{-0.098}$	$0.687_{-0.070}$	$0.650_{-0.057}$	$0.623_{-0.037}$	$0.608_{-0.024}$	0.057 ↓	$0.465_{-0.027}$
w/o Cumulative Probs	$0.757_{-0.066}$	$0.740_{-0.017}$	$0.694_{-0.013}$	$0.649_{-0.016}$	$0.607_{-0.024}$	0.027 ↓	$0.463_{-0.029}$
w/o `Judge Agent`	$0.720_{-0.103}$	$0.708_{-0.050}$	$0.688_{-0.018}$	$0.637_{-0.024}$	$0.593_{-0.038}$	0.047 ↓	$0.457_{-0.035}$

Table 4: Ablation study results on `gpt-oss-120B`.

2025c) or fine-tuning LLMs on curated ESC trajectories, either through supervised fine-tuning (Liu et al., 2021; Zheng et al., 2023a,b; Qiu et al., 2024; Xiao et al., 2024; Chen et al., 2025b; Zhou et al., 2025a; Ye et al., 2025b) or reinforcement learning (Yang et al., 2025b; Zhang et al., 2025b; Yuan et al., 2025; Yao et al., 2025; Dai et al., 2025; Zhang et al., 2025a; Ye et al., 2025a; Hong et al., 2025). While these methods can improve the generation of empathetic responses, they do not explicitly model the underlying causes of the help-seeker's distress. A second line of work is motivated by cognitive reappraisal and improves reframing by incorporating cognitive appraisal theory into the generation workflow (Zhan et al., 2023, 2024; Zhou et al., 2025b; Sun et al., 2026).

ToM Reasoning with LLMs Evaluating and improving LLMs' ToM reasoning abilities has recently received substantial attention from the research community. Researchers have developed a range of benchmarks based on the False Belief task (Le et al., 2019; Wu et al., 2023; Zhou et al., 2023a; Gandhi et al., 2023; Jin et al., 2024; Xu et al., 2024; Gu et al., 2024), as well as benchmarks that go beyond false-belief reasoning (Zhou et al., 2023b; Chen et al., 2024; Yeo and Jaidka, 2025). Although some benchmarks evaluate LLMs' ability to infer mental states from dialogue (Kim et al., 2023; Xiao et al., 2025), they primarily focus on casual conversations rather than ESCs.

To improve LLMs' ToM reasoning abilities, prior work has proposed prompting (Wilf et al., 2024) and neuro-symbolic approaches (Sclar et al., 2023; Hou et al., 2024; Jung et al., 2024; Xu et al., 2025; Zhou et al., 2026). Another line of research adopts probabilistic inference, using methods such as Sequential Monte Carlo (Kim et al., 2025), or BIP (Zhi-Xuan et al., 2024; Jin et al., 2024; Ying et al., 2024; Zhang et al., 2025d). However, existing BIP-based approaches typically assume a narrow search space over latent mental states and treat all actions as equally informative. These assumptions make them less suitable for salient appraisal reasoning, where the search space is large, and the informativeness of dialogue turns can vary substantially.

7 Conclusion

In this work, we investigate a previously overlooked problem in ESC: inferring the salient cognitive appraisal dimensions that drive negative emotions. To this end, we introduce AppraiSal, the first benchmark for salient appraisal reasoning in ESC, and propose PRISM, a multi-agent probabilistic framework grounded in Bayesian Inverse Planning. Experiments show that current LLMs remain substantially below human performance on this task, while PRISM consistently improves appraisal inference, particularly for the most salient dimensions. Our case study further suggests that the incorporation of appraisals can lead to more effective emotional support. We hope this work will open new research avenues on appraisal saliency reasoning and its application in emotional support.

Limitations

ESC Trajectories Although AppraiSal covers a diverse set of upsetting events, many of the scenarios are still relatively simple and self-contained. In real-world emotional support conversations, distress often arises from a prolonged sequence of events, multiple interacting stressors, or unresolved interpersonal dynamics. Inferring salient appraisals in such settings is likely to be substantially more challenging. As a result, performance on AppraiSal may overestimate model capability in more complex real-world ESC settings.

The PRISM Framework As discussed in §3, PRISM relied on token log-probability to model the likelihood factors in the backward BIP inference process. Specifically, since we concatenate the expected response to the prompt, our method only works with models that support `prompt logprobs`

(also known as `echo logprobs`). While this functionality is supported by most open-sourced models, its support is deprecated by OpenAI models.

Synthetic data construction The construction of AppraiSal involves a hybrid pipeline that combines source events and personas with LLM-based generation, revision, and human annotation. While we remove pseudo appraisal labels before annotation and apply quality control procedures, the generated conversations may still reflect artifacts introduced by the prompting process or the underlying LLMs. Consequently, the dataset may not fully capture the linguistic diversity, messiness, and ambiguity of naturally occurring emotional support conversations.

Annotation constraints Our annotation protocol requires annotators to select exactly five salient appraisal dimensions and provide rationales for their choices. This design improves consistency and discourages careless annotation. In our preliminary annotations without this constraint, many annotators tended to select only a single highly salient appraisal dimension, leaving other potentially relevant appraisals unannotated. Nevertheless, this design imposes a discrete structure on what may in reality be a more graded and uncertain appraisal process. In some cases, fewer than five appraisals may be truly salient, while in others, the boundary between salient and non-salient appraisals may be ambiguous. Future work could explore alternative annotation schemes that better capture this uncertainty, for example, by collecting graded salience ratings rather than discrete selections.

Ethics Statement

This work contains emotional support and cognitively grounded reasoning about distress, and should therefore be treated with appropriate caution. AppraiSal is intended as a research benchmark and PRISM as a proof-of-concept study, not as a clinical tool or a replacement for professional mental health practitioners. While our study demonstrates the potential of constructing ESC agents capable of utilizing cognitive appraisals to deliver better emotional support, such a system is not ready to be used with human participants, as we do not fully understand the safety of this system. More research needs to be conducted to fully examine the capability and, most importantly, the safety and reliability of agents to be deployed in high-stakes scenarios such as our PRISM framework.

To reduce privacy risk, AppraiSal is constructed from source events, personas, LLM-based generation, and human annotation, rather than real transcripts from therapies. Still, the dataset and PRISM may inherit biases from the source data, annotators, and underlying LLMs. In addition, incorrect appraisal inference could lead to misleading or harmful support if deployed without oversight. We therefore caution against interpreting performance on AppraiSal as evidence that current LLMs are suitable for autonomous mental health support, and recommend strong human supervision and safety safeguards for any downstream use.

Acknowledgments

We thank Xinyi Mou for the valuable discussions and contributions to the pilot studies. This work was supported in part by the UK Engineering and Physical Sciences Research Council (EPSRC) through an iCASE award with Huawei London Research Centre, a Turing AI Fellowship (grant no. EP/V020579/1, EP/V020579/2), and the Prosperity Partnership scheme (grant no. UKRI566), and by Inkfish through the EMBRACE research programme.

References

Magda B Arnold. 1960. *Emotion and Personality: Y Magda B. Arnold*. Columbia UP.

Aaron T Beck, A John Rush, Brian F Shaw, Gary Emery, Robert J DeRubeis, and Steven D Hollon. 2024. *Cognitive therapy of depression*. Guilford Publications.

Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, and Zhou Yu. 2023. Controllable mixed-initiative dialogue generation through prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 951–966.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. 2025a. Theory of mind in large language models: Assessment and enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31539–31558.

Zhuang Chen, Yaru Cao, Guanqun Bi, Jincenzi Wu, Jinfeng Zhou, Xiyao Xiao, Si Chen, Hongning Wang, and Minlie Huang. 2025b. Socialsim: Towards socialized simulation of emotional support conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1274–1282.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and 1 others. 2024. Tombench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983.

Chongyuan Dai, Jinpeng Hu, Hongchang Shi, Zhuo Li, Xun Yang, and Meng Wang. 2025. Psyche-r1: Towards reliable psychological llms through unified empathy, expertise, and reasoning. *arXiv preprint arXiv:2508.10848*.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36:13518–13529.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

James J Gross. 2015. Emotion regulation: Current status and future prospects. *Psychological inquiry*, 26(1):1–26.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2024. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in llms. *arXiv preprint arXiv:2410.13648*.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of the 28th international conference on computational linguistics*, pages 125–138.

Simin Hong, Jun Sun, and Hongyang Chen. 2025. Third-person appraisal agent: Simulating human emotional reasoning in text with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23684–23701.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. 2024. Timetom: Temporal space is the key to unlocking the door of large language models' theory-of-mind. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11532–11547.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. 2024. Mmtom-qa: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16077–16102.

Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, and Hyunwoo Kim. 2024. Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19794–19809.

Hyunwoo Kim, Melanie Sclar, Tan Zhi-Xuan, Lance Ying, Sydney Levine, Yang Liu, Joshua B Tenenbaum, and Yejin Choi. 2025. Hypothesis-driven theory-of-mind reasoning for large language models. In *Second Conference on Language Modeling*.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877.

Suyeon Lee, Sunghwan Mac Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, and 1 others. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274.

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

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. 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)*, pages 3469–3483.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory

of mind in large language models. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 1011–1031.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Andrew Ortony, Gerald L Clore, and Allan Collins. 2022. *The cognitive structure of emotions*. Cambridge university press.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636.

Ira J Roseman and Craig A Smith. 2001. Appraisal theory. *Appraisal processes in emotion: Theory, methods, research*, pages 3–19.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models'(lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13960–13980.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7881–7892.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

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. *arXiv preprint arXiv:2605.17176*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Andero Uusberg, Brett Ford, Helen Uusberg, and James J Gross. 2023. Reappraising reappraisal: An expanded view. *Cognition and Emotion*, 37(3):357–370.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Amy Wenzel. 2017. Basic strategies of cognitive behavioral therapy. *Psychiatric Clinics*, 40(4):597–609.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706.

Mengxi Xiao, Qianqian Xie, Ziyan Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. 2024. Healme: Harnessing cognitive reframing in large language models for psychotherapy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1707–1725.

Yang Xiao, Jiashuo Wang, Qiancheng Xu, Changhe Song, Chunpu Xu, Yi Cheng, Wenjie Li, and Pengfei Liu. 2025. Towards dynamic theory of mind: Evaluating llm adaptation to temporal evolution of human states. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24036–24057.

Hainiu Xu, Siya Qi, Jiazheng Li, Yuxiang Zhou, Jinhua Du, Caroline Catmur, and Yulan He. 2025. Enigmatom: Improve llms' theory-of-mind reasoning capabilities with neural knowledge base of entity states. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13598–13622.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Ting Yang, Li Chen, and Huimin Wang. 2025b. Towards open-ended emotional support conversations in llms via reinforcement learning with future-oriented rewards. *arXiv preprint arXiv:2508.12935*.

Xianrong Yao, Dong She, Chenxu Zhang, Yimeng Zhang, Yueru Sun, Noman Ahmed, Yang Gao, and Zhanpeng Jin. 2025. Empathy-r1: A chain-of-empathy and reinforcement learning framework for long-form mental health support. *arXiv preprint arXiv:2509.14851*.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025a. From generic empathy to personalized emotional support: A self-evolution framework for user preference alignment. *Findings of the Association for Computational Linguistics: EMNLP*, 2025:18826–18853.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025b. Sweetiechat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4646–4669.

Gerard C Yeo and Desmond C Ong. 2024. Associations between cognitive appraisals and emotions: A meta-analytic review. *Psychological Bulletin*.

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.

Lance Ying, Tan Zhi-Xuan, Lionel Wong, Vikash Mansinghka, and Josh Tenenbaum. 2024. Grounding language about belief in a bayesian theory-of-mind. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Jiahao Yuan, Zhiqing Cui, Hanqing Wang, Yuansheng Gao, Yucheng Zhou, and Usman Naseem. 2025. Kardia-r1: Unleashing llms to reason toward understanding and empathy for emotional support via rubric-as-judge reinforcement learning. *arXiv preprint arXiv:2512.01282*.

Hongli Zhan, Desmond Ong, and Junyi Jessy Li. 2023. Evaluating subjective cognitive appraisals of emotions from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14418–14446.

Hongli Zhan, Allen Zheng, Yoon Kyung Lee, Jina Suh, Junyi Jessy Li, and Desmond Ong. 2024. Large language models are capable of offering cognitive reappraisal, if guided. In *First Conference on Language Modeling*.

Chao Zhang, Xin Shi, Xueqiao Zhang, Yifan Zhu, Yi Yang, and Yawei Luo. 2025a. DecoupledESC: Enhancing emotional support generation via strategy-response decoupled preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22189–22215, Suzhou, China. Association for Computational Linguistics.

Naifan Zhang, Ruihan Sun, Ruixi Su, Shiqi Ma, Shiya Zhang, Xianna Weng, Xiaofan Zhang, Yuhan Zhan, Yuyang Xu, Zhaohan Chen, and 1 others. 2025b. Echo-n1: Affective rl frontier. *arXiv preprint arXiv:2512.00344*.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. Escot: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412.

Xinjie Zhang, Wenxuan Wang, and Qin Jin. 2025c. Intentionesc: An intention-centered framework for enhancing emotional support in dialogue systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26494–26516.

Zhining Zhang, Chuanyang Jin, Mung Yao Jia, Shunchi Zhang, and Tianmin Shu. 2025d. Autotom: Scaling model-based mental inference via automated agent modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023a. Augesc: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023b. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.

Tan Zhi-Xuan, Lance Ying, Vikash Mansinghka, and Joshua B Tenenbaum. 2024. Pragmatic instruction following and goal assistance via cooperative language-guided inverse planning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 2094–2103.

Jinfeng Zhou, Yuxuan Chen, Jianing Yin, Yongkang Huang, Yihan Shi, Xikun Zhang, Libiao Peng, Rongsheng Zhang, Tangjie Lv, Zhipeng Hu, and 1 others. 2025a. Crisp: Cognitive restructuring of negative thoughts through multi-turn supportive dialogues. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32462–32491.

Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, and 1 others. 2023a. How far are large language models from agents with theory-of-mind? *arXiv preprint arXiv:2310.03051*.

Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2023b. I cast detect thoughts: Learning to converse and guide with intents and theory-of-mind in dungeons and dragons. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11136–11155.

Xuhui Zhou, Jiarui Liu, Akhila Yerukola, Hyunwoo Kim, and Maarten Sap. 2026. Social world models. *Preprint*, arXiv:2509.00559.

Yuxiang Zhou, Hainiu Xu, Desmond C. Ong, Maria Liakata, Petr Slovak, and Yulan He. 2025b. Modeling subjectivity in cognitive appraisal with language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13811–13833, Suzhou, China. Association for Computational Linguistics.

A Cognitive Appraisal Dimensions

We obtain our 14 appraisals by `Preserve/Remove/Aggregate` the appraisal dimensions defined by (Hofmann et al., 2020) as shown in Table A1

Index	Appraisal Dimension	Modification
1	suddenness	`Preserved`
2	familiarity	`Preserved`
3	predict event	`Preserved`
4	pleasantness	`Removed`
5	unpleasantness	`Removed`
6	goal congruence	`Preserved`
7	self responsibility	`Preserved`
8	chance responsibility	`Aggregated with 9`
9	other responsibility	`Preserved`
10	predict consequence	`Preserved`
11	goal support	`Aggregated with 6`
12	urgency	`Aggregated with 19`
13	self control	`Preserved`
14	other control	`Preserved`
15	chance control	`Aggregated with 14`
16	accept consequence	`Preserved`
17	norm violation	`Preserved`
18	standards	`Aggregated with 17`
19	coping demand	`Preserved`
20	attention	`Aggregated with 19`
21	not consider	`Aggregated with 19`

Table A1: Modification of appraisal dimensions defined by Hofmann et al. (2020).

We also dichotomized dimensions such as *predict event* into *predictability of event* and *unpredictability of event* to minimize ambiguity. The revised 14 dimensions and their explanations used in this study are shown in Table A2.

B Persona Construction

We first match each event with the top-5 most similar persona descriptions. Events and personas are matched by computing the pairwise cosine similarities between embeddings given by the `jina-embedding-v3` model. The most suitable persona is picked by prompting `GPT-5-mini`. Typically, a persona profile from the PersonaHub contains information regarding the character's occupation and occasionally traces of the character's gender (e.g., a retired actress). We first extract *age*, *occupation*, *gender*, and *education level* from the persona. Anchoring on the extracted information, we match the occupation with one from the occupation data acquired from the U.S. Bureau of Labor Statistics[13]. For entries missing *education level*, we sample an education level according to the education level distribution corresponding to the assigned occupation, given by the U.S. Bureau of Labor Statistics. We consider four education levels, namely "high school", "Bachelor's", "Master's", and "Ph.D.". We then prompt `GPT-5-mini` again to revise the persona to better match the occupation, education level, and most importantly, the corresponding upsetting event. Finally, we assign each character a personality according to the Big Five personality traits, which comprise the following aspects:

- Neuroticism: ['emotionally stable', 'emotionally unstable']
- Conscientiousness: ['organized', 'careless']
- Extroversion: ['extroverted', 'introverted']
- Openness: ['conservative', 'open-minded']
- Agreeableness: ['agreeable', 'antagonistic']

The resulting persona profile is highly associated with the upsetting event and depicts the following aspects of the individual:

- **Name**: Sampled from the frequently used names given by the U.S. Social Security Administration[14]
- **Occupation**: Sampled using data from the U.S. Bureau of Labor Statistics
- **Education**: Sampled using data from the U.S. Bureau of Labor Statistics
- **Gender**: Extracted from the original persona entry or randomly sampled.
- **Personality Traits**: Randomly sampled according to the Big Five personality traits.

C Comparison of Orginal and Rewritten Situations

We source upsetting events from AugESC (Zheng et al., 2023a). While this established dataset provides a strong foundation for emotional support conversations, its original events are often concise and generic due to a lack of persona information. To address this, we provide LLMs with the generic event and the paired persona for revision. The goal is to make the events more explicit, engaging, and persona-aligned. We argue that these revised events are more realistic, as they offer richer situational contexts while remaining strictly grounded in the

[13] https://www.bls.gov/bls/occupation.htm
[14] https://www.ssa.gov/oact/babynames/

suddenness	The patient finds the situation sudden or abrupt.
unfamiliarity	The patient finds the situation unfamiliar.
predictability of event	The patient was expecting the situation to happen.
unpredictability of event	The patient wasn't expecting the occurrence of the situation.
goal incongruence	The outcome is not aligned with the patient's goals/desires.
predictability of consequences	The patient can easily anticipate the consequences of the situation.
unpredictability of consequences	The patient had no idea about what the consequences of the situation would be.
self cause	The patient blames themselves for causing the situation.
other cause	The patient blames someone else for causing the situation.
self control	The patient has control over what is going on in the situation.
other control	Someone/Something other than the patient has control over what is going on in the situation.
norm violation	The actions that produced the situation violated the patient's standards and ideals.
coping demand	The patient has to spend a great deal of (physical or mental) energy to handle the outcome of the situation.
unacceptable consequences	The consequences of the situation are unbearable for the patient.

Table A2: The 14 appraisal dimensions used in this study.

original AugESC scenarios. Table A3 illustrates the contrast between the original AugESC entries and our revised entries in AppraiSal.

D The reAppraisal Framework

We use the extended reAppriasal Framework from Uusberg et al. (2023) to guide the therapist agent during cognitive reframing. The reAppraisal framework can be summarized into the following eight reframing strategies (note that we have rephrased the name of each dimension; the original name is shown in parentheses)

- **Undoing Construal (Object-level Construal Decommitment):** Guiding the user to move away from their original perspective on the event.
- **Adopting Construal (Object-level Construal Commitment):** Offering a different perspective on the event.
- **Undoing Feeling (Meta-level Construal Decommitment):** Guiding the user to disregard their original emotional response.
- **Adopting Feeling (Meta-level Construal Commitment):** Validating that a different emotion is possible.
- **Undoing Goal (Object-level Goal Decommitment):** Guiding the user to move away from their original intention.
- **Adopting Goal (Object-level Goal Commitment):** Shifting focus to an alternative intent.
- **Undoing Mental Goal (Meta-level Goal Decommitment):** Guiding the user to disregard their original desire.
- **Adopting Mental Goal (Meta-level Goal Commitment):** Offering a different desire that is more aligned with the event.

E Human Annotation

E.1 Demographic Information

To ensure high-quality evaluations, we established strict inclusion criteria for our human annotators. Participants were required to list English as their first language, given that the Emotional Support Conversation (ESC) transcripts were written exclusively in English. Furthermore, we restricted the eligible country of residence to the United States, Canada, the United Kingdom, Ireland, Australia, and New Zealand. To guarantee data reliability and worker experience, eligible annotators must have maintained a historical approval rate of at least 98% and successfully completed a minimum of 10 prior tasks on the platform.

We acknowledge that restricting our annotator pool to these specific geographic and linguistic demographics inevitably biases the annotations toward a predominantly Western, middle-class cultural perspective. The implications of these potential biases are further discussed in the Limitations section of the main text.

E.2 Annotation Interface

In the beginning of the annotation process, each annotator is provided with the persona information of the help-seeker, as shown in Figure A1. The annotator then reads the ESC trajectory, as shown in Figure A2. After understanding the help-seeker's persona and their exchange with the therapist, the annotator first annotates the help-seeker's pre-event BDI. Notice that the generated pseudo labels are provided in this section as a reference (Figure A3)

For salient appraisal annotation, the annotator first selects a coarse-grained description of the help-seeker's interpretation of the upsetting event, as shown in Figure A4. Each coarse-grained description corresponds to 2 fine-grained appraisal

AugESC (Original)	AppraiSal (Revised)
Man I feel so bad about last weekend :/	I skipped the championship cycling race to sit through a last-minute quarterly review, leaving my team stranded and the charity fundraiser without its star rider. Because I prioritized the meeting, the race was postponed and we lost $20,000 in sponsorships.
My wife is angry and I don't know what to say. My anniversary was yesterday and I had no clue!	I stayed up until 3 a.m. polishing an investor pitch, completely forgetting that my wife had booked a private rooftop dinner for our 15-year anniversary, and I arrived home after the restaurant had closed. She confronted me, eyes blazing with disappointment, and the silence in the empty dining room felt like a slap.
I'm very worried that I'm not going to make an important goal.	I watched the flames devour my finished mural only hours before the city council was scheduled to vote on the grant that would secure my first museum commission.

Table A3: Original upsetting events from AugESC and the revised, persona-aligned events in AppraiSal.

dimensions, which are hidden from the annotator in the beginning. The annotator then selects 5 fine-grained appraisal dimensions from the unfolded options. This pipeline avoids displaying all 14 appraisal dimensions to the annotator at once and effectively reduces the annotator's cognitive workload.

Once the 5 salient appraisal dimensions are selected, the annotator is required to rank the appraisal dimensions according to their saliency, with the most salient dimension ranked at the top. Further, the annotator is required to provide a brief rationale for each appraisal dimension (see Figure A5).

Finally, the annotator is prompted to assess the quality of the current data entry based on their annotation experience (interface shown in Figure A6). Specifically, the annotator evaluates the current entry using the following four criteria:

- **Realism**: The flow of the dialogue is natural and realistic.
- **Persona-Relevance**: The dialogue reflects the help-seeker's persona profile.
- **BDI-Relevance**: I can confidently infer the help-seeker's BDI from the dialogue.
- **Appraisl-Relevance**: I can confidently infer the help-seeker's salient appraisals from the dialogue.

F Detailed Data Statistics

Seeded from AugESC (Zheng et al., 2023a), AppraiSal covers a similar distribution of event topics. We provide the count and distribution of each event genre in Table A4. On average, there are $27.77_{\pm 5.148}$ turns. In each turn, the `therapist agent` utters 23.92 words on average, whereas the `patient agent` utters 31.72 words on average. The detailed distribution of turn counts in AppraiSal is shown in Figure A7.

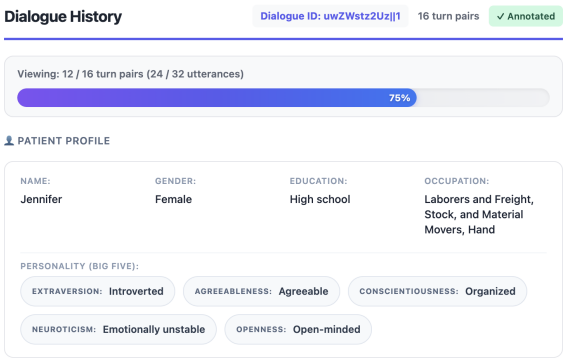

Figure A1: Annotation Interface: Help-seeker's Persona Information

Genre	Count	Percentage
Workplace & Career Challenges	452	45.4%
Parenting & Caregiving Burden	90	9.0%
Identity & Self-Worth	77	7.7%
Mental Health & Psychological Distress	71	7.1%
Family, Friendship & Social Conflict	68	6.8%
Financial & Housing Stress	51	5.1%
Work-Life Balance & Burnout	48	4.8%
Grief, Loss & Trauma	43	4.3%
Romantic & Intimate Relationships	38	3.8%
Academic & Learning Stress	32	3.2%
Safety, Security & Existential Threats	13	1.3%
Physical Health & Medical Concerns	8	0.8%
Immigration & Legal Stress	5	0.5%

Table A4: Distribution of dialogue genres.

In total, we curated 1,120 entries from Prolific, which are annotated by 145 annotators, where each annotator is assigned approximately 8 entries. The distribution of human-annotated appraisal dimensions is shown in Table A6. On average, we compensate $1.52 for each entry, resulting in a total cost of $1702.4. Further, we leverage `GPT-5-mini` to refine and generate multiple components of the dataset; the cost of the single complete run is $15.93. The total cost, including preliminary runs for prompt engineering, is approximately $24. We summarize the API cost in Table A5.

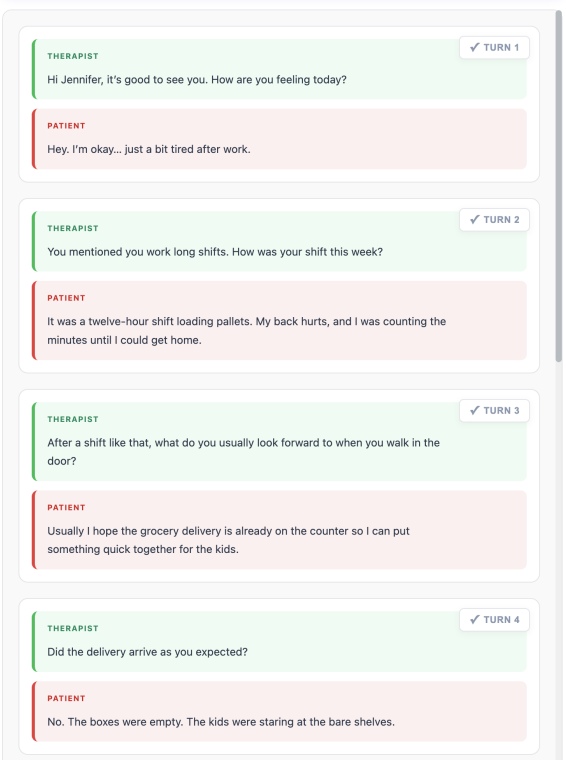

Figure A2: Annotation Interface: ESC Trajectory

Task	Input	Output	Total
Event Revision	$0.36	$0.20	$0.56
Interpretation	$0.30	$0.16	$0.46
Pseudo Mental States	$0.34	$0.94	$1.28
ESC Revision	$0.90	$12.73	$13.63
Total	$1.9	$14.03	$15.93

Table A5: Cost breakdown for using OpenAI API for AppraiSal construction.

To ensure data quality, we filter entries with low assessment scores. Specifically, we apply two filters. First, all entries with an average assessment score lower than 3.0 are removed. Second, all entries with an "Appraisal-Relevance" score of lower than 3.0 are also removed. Salient appraisal inference is the core task of AppraiSal, therefore it is crucial to ensure a strong association between the content of the ESC trajectory and the annotated appraisal dimensions. The remaining 996 data entries have the following assessment scores:

- **Appraisal-Relevance**: 4.39
- **BDI-Relevance**: 4.33
- **Persona-Relevance**: 4.38
- **Realism**: 4.42

Further, 26.34% of the data entry received revision in the ESC trajectories. After getting obtaining the annotations, we leverage `GPT-5-Mini` to flag annotations that contain trivial or logical errors. For instance, the dimension of "*self-cause*" and "*other-cause*" shouldn't occur simultaneously. This filtering process results in a revision of 7.48% of the existing annotations.

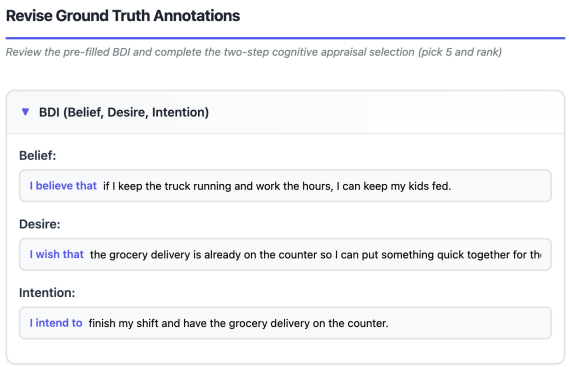

Figure A3: Annotation Interface: Pre-event BDI Annotation

Appraisal label	Count	Percentage
Self Cause	752	75.5%
Goal Incongruence	751	75.4%
Unacceptable Consequences	549	55.1%
Norm Violation	484	48.6%
Unpredictability Of Event	459	46.1%
Coping Demand	433	43.5%
Self Control	369	37.0%
Unpredictability Of Consequences	346	34.7%
Suddenness	273	27.4%
Predictability Of Consequences	144	14.5%
Other Control	131	13.2%
Other Cause	118	11.8%
Unfamiliarity	103	10.3%
Predictability Of Event	43	4.3%

Table A6: Distribution of human-annotated cognitive appraisal labels AppraiSal.

G Bayesian Inverse Planning

In this section, we show a step-by-step derivation of the backward BIP inference factorization from the forward generative model.

To recap, given a ESC trajectory up to the tth turn $\mathcal{D}_{<t}$, the help-seeker's persona profile $\mathcal{P}$, their pre-event BDI $\mathcal{B}$ and their salient cognitive appraisals $\mathcal{A}$, we define the following forward generative

$$u_{h,t} \sim \mathbb{P}(u_{h,t} \mid \mathcal{D}_{<t}, \mathcal{P}, \mathcal{B}, \mathcal{A})$$

where we regard the help-seeker's utterance $u_{h,t}$ as the *action*. The object is to jointly infer the help-seeker's pre-event BDI $\mathcal{B}$ and, more importantly,

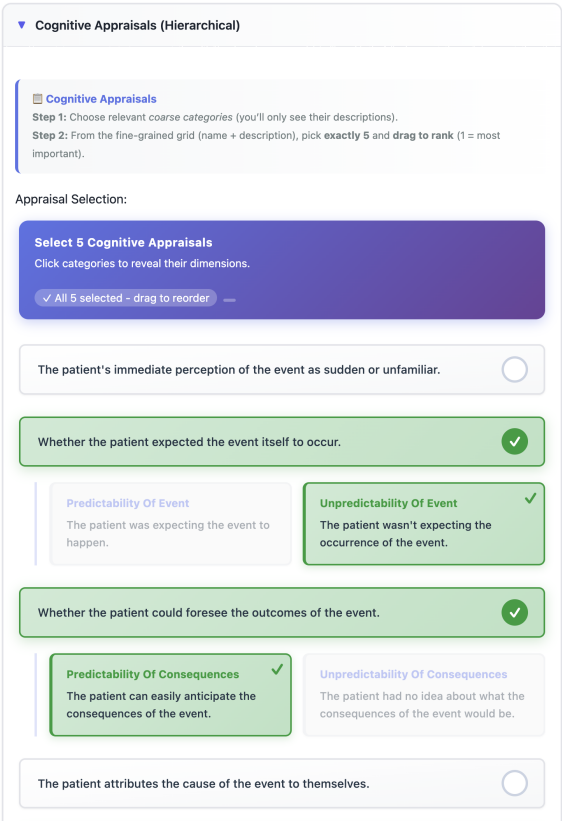

Figure A4: Annotation Interface: Coarse-to-Fine Annotation for Salient Appraisal Dimensions.

their salient cognitive appraisals $\mathcal{A}$, which translates to the following backward inference problem:

$$\mathbb{P}(\mathcal{A}, \mathcal{B} \mid \mathcal{D}_{1:t}, \mathcal{P}).$$

Applying Bayes' rule, we get

$$\mathbb{P}(\mathcal{A}, \mathcal{B} \mid \mathcal{D}_{1:t}, \mathcal{P}) \propto \mathbb{P}(\hat{u}_{h,t} \mid \mathcal{D}_{<t}, \mathcal{P}, \mathcal{A}, \mathcal{B}) \cdot \mathbb{P}(\mathcal{A}, \mathcal{B} \mid \mathcal{D}_{<t}, \mathcal{P}).$$

By further factorizing $\mathbb{P}(\mathcal{A}, \mathcal{B} \mid \mathcal{D}_{<t}, \mathcal{P})$, we get

$$\mathbb{P}(\mathcal{A}_t, \mathcal{B}_t | \mathcal{D}_{1:t}, \mathcal{P}) \propto \mathbb{P}(\hat{u}_{h,t} | \mathcal{D}_{<t}, \mathcal{P}, \mathcal{B}_t, \mathcal{A}_t) \cdot \mathbb{P}(\mathcal{B}_t | \mathcal{D}_{<t}, \mathcal{P}) \cdot \mathbb{P}(\mathcal{A}_t | \mathcal{D}_{<t}, \mathcal{P}, \mathcal{B}_t),$$

which is the factorization we use in the backward inference problem as shown in Equation 3.

H Likelihood Modeling with LLMs

In this work, we leverage the token-level log-probabilities to compute the likelihoods, namely $\mathbb{P}(\hat{u}_t \mid \cdot)$, $\mathbb{P}(\mathcal{B} \mid \cdot)$, and $\mathbb{P}(\mathcal{A} \mid \cdot)$. We use log-probabilities over prompting LLMs to produce a likelihood score to mitigate the potential risk of hallucination.

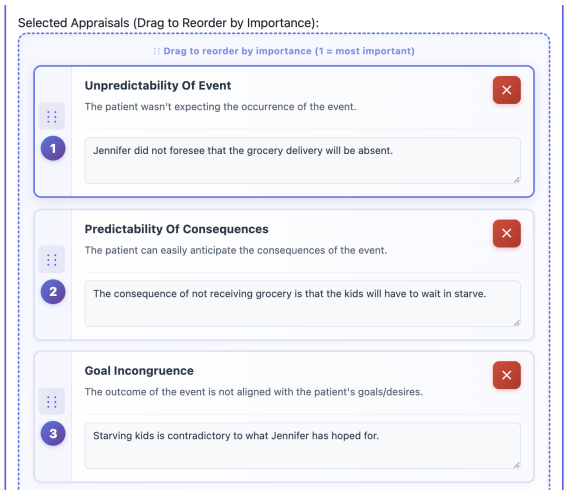

Figure A5: Annotation Interface: Salient Appraisal Dimensions with Required Ranking Rationale.

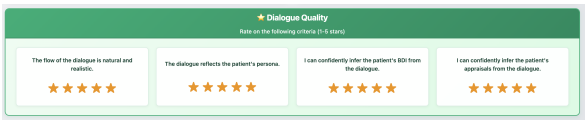

Figure A6: Annotation Interface: Data Quality Check.

Take the `utterance agent` as an example. Suppose we apply PRISM with `Qwen3-4B-Instruct` as the backbone model and with input instruction prompt $\mathcal{I}(\mathcal{D}_{<t}, \mathcal{P}, \mathcal{B}_t, \mathcal{A}_{t,i})$. We take the actual next utterance from the help-seeker, denoted as $\hat{u}_t = \{w_1, w_2, \ldots, w_m\}$ and build the chat template as:

<|im_start|>user
$\mathcal{I}(\mathcal{D}_{<t}, \mathcal{P}, \mathcal{B}_t, \mathcal{A}_{t,i})$
<|im_end|>
<|im_start|>assistant
$w_1, w_2, \ldots, w_m$
<|im_end|>

We input the complete chat template into the model and gather the log-probabilities corresponding to tokens $w_1, w_2, \ldots, w_m$. Specifically, we use vllm's `"prompt_logprobs"` functionality (Kwon et al., 2023). The log-probability corresponding to the current appraisal hypothesis $\mathcal{A}_{t,i}$ is then computed as the sum of all token log-probabilities:

$$p(\hat{u}_{t,i}) = \log \mathbb{P}(\hat{u}_t \mid \mathcal{D}_{<t}, \mathcal{P}, \mathcal{B}_t, \mathcal{A}_{t,i}) = \sum_{j=1}^{m} \log \mathbb{P}\Big(w_j | w_{<j}, \mathcal{I}(\mathcal{D}_{<t}, \mathcal{P}, \mathcal{B}_t, \mathcal{A}_{t,i})\Big). \quad (9)$$

I BDI Analysis

In this section, we evaluate the quality of the argmax of the pre-event BDI, $\mathcal{B}^*$. To estimate human performance, we asked two annotators to each

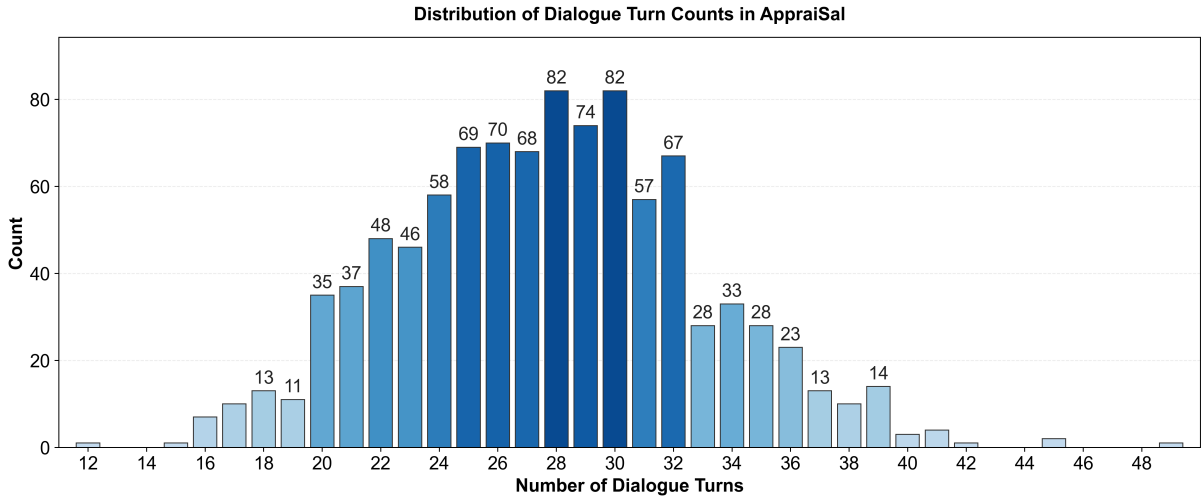

Figure A7: Distribution of the number of dialogue turns across ESC trajectories in AppraiSal.

annotate 50 entries sampled from AppraiSal. For evaluation, we use both an n-gram-based metric, ROUGE-L (Lin, 2004), and a model-based metric, BLEURT (Sellam et al., 2020).

The results in Table A8 show that the `gpt-oss` series matches human performance in terms of average ROUGE-L. Moreover, the `Gemma3` series comes close to human performance in terms of average BLEURT. Although the `Qwen3` series still falls short of human performance, it is nevertheless able to produce pre-event BDI states of reasonable quality, which is further reflected in the gains observed from incorporating pre-event BDI in the ablation study.

	A1|A2	A1|A3	A2|A3	Average
Spearman's ρ	0.726	0.541	0.477	0.581

Table A7: Spearman's Rank Correlation results using human agreement data

J Saliency Rank & Evaluation Metrics

J.1 Ranking of Appraisal Dimensions

As shown in Figure A5, annotators are asked to rank the selected fine-grained appraisal dimensions by saliency. The dimensions are initially presented in click order, and annotators then drag them to produce their final ranking. On average, each appraisal annotation involves 1.79 drag operations.

To evaluate whether annotators can reliably agree on the exact ranking of appraisal dimensions, we compute *Spearman's rank correlation* on a human agreement set of 50 examples, each annotated by three Prolific annotators. Table A7 shows that although Annotators 1 and 2 achieve a relatively strong correlation of 0.726, the remaining pairwise correlations are notably weaker, resulting in a moderate average correlation of 0.581. Since these results provide limited support for the feasibility of curating precise and consensual appraisal rankings, we do not use exact rank information in our evaluation and instead rely on set-based metrics.

J.2 Evaluation Metrics

In this section, we provide a definition for the evaluation metrics used in our experiments. Given a ground truth set of salient appraisals $\mathcal{A} = \{a_1, a_2, a_3, a_4, a_5\}$ and a predicted set of salient appraisals $\hat{\mathcal{A}} = \{\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_4, \hat{a}_5\}$, we compute the following metrics

Precision@k computes the proportion of the top-k predicted appraisals that exist in the ground truth:

$$Precision@k = \frac{|\{a_i\}_{i=1}^{k} \cap \mathcal{A}|}{k},$$

where $|\cdot|$ denotes the cardinality of the set.

Relationship with *Recall* and *F1* In the salient appraisal reasoning task, where the model is required to provide 5 salient appraisal dimensions, $|A| = |\hat{A}| = 5$. Given the following definition of set-level Recall:

$$Recall = \frac{|\hat{\mathcal{A}} \cap \mathcal{A}|}{|\mathcal{A}|},$$

where $\hat{\mathcal{A}} = \{a_i\}_{i=1}^{5}$. It is trivial that *Recall* is equivalent to *Precision@5*. Therefore, we do not report the *Recall* and *F1* scores.

Model	Avg. ROUGE-L	Avg. Sim	ROUGE-L			BLEURT Scores		
			Belief	Desire	Intention	Belief	Desire	Intention
Human	0.326	0.529	0.307	0.309	0.362	0.537	0.526	0.525
Gemma3-4B	0.286	0.491	0.277	0.281	0.301	0.483	0.513	0.478
Gemma3-27B	0.289	0.520	0.316	0.253	0.300	0.535	0.514	0.512
Qwen3-4B	0.253	0.469	0.275	0.233	0.253	0.471	0.480	0.455
Qwen3-32B	0.251	0.458	0.272	0.233	0.247	0.479	0.473	0.421
Qwen3-30B-A3B	0.219	0.442	0.231	0.208	0.218	0.462	0.449	0.415
GPT-OSS-20B	0.318	0.470	0.337	0.283	0.332	0.475	0.480	0.454
GPT-OSS-120B	0.312	0.487	0.313	0.284	0.338	0.472	0.505	0.484

Table A8: Performance comparison on BDI inference.

Jaccard Index computes the number of appraisal dimensions in the intersection of the ground truth set and the predicted set, normalized by the number of appraisals covered by the union of the two sets:

$$\textit{Jaccard Index} = \frac{|\hat{\mathcal{A}} \cap \mathcal{A}|}{|\hat{\mathcal{A}} \cup \mathcal{A}|}.$$

K Full Evaluation Results

The full evaluation results, including results from the recently released `Qwen3.5` models are shown in Table A10. The results for *Iterative Self-Consistency with Judge Agent* are shown in Figure A11. Results show that, albeit using extra time-time compute, neither *Self-Consistency* prompting nor *Iterative Self-Consistency with Judge Agent* achieves meaningful improvement compared to PRISM, demonstrating the effectiveness and importance of the BIP-based multi-agent inference framework.

L Two-sample T-test

The initial experiments shown in Table 2 are run only once which do not allow for hypothesis testing. While some improvements may seem numerically small, we additionally conduct significance testing to strengthen the effectiveness of the PRISM framework. To conduct hypothesis testing, we have run the baseline models with and without the PRISM framework 2 more times, which give us both a mean and a variance based on 3 runs for hypothesis testing. We present these results in Table A12 and Table A9.

The main improvements come from inferring appraisals grounded in salient external events and their consequences. In particular, PRISM substantially improves recall for *unacceptable-consequences* from 7.1% to 85.0%, *unpredictability-of-event* from 13.9% to 47.7%, and *suddenness* from 24.2% to 31.5%. Smaller but consistent gains also appear for *other-cause* and *self-control*. We illustrate the helpfulness of PRISMon all appraisal dimensions in Figure A8

For the remaining issues of PRISM, we observe that PRISM gives weak performance along the self-evaluative dimensions including *unfamiliarity*, *coping-demand*, and *unpredictability-of-consequences*. Taken together, these results suggest that PRISM is better at capturing the impact of the surrounding environment (e.g. *what happened* and *how harmful it was*), but remains less reliable at inferring some of the help-seeker's nuanced mental states. Since emotional distress is often shaped by perceived responsibility, violated norms, coping burden, and anticipated future fallout rather than the event alone, this gap highlights an important direction for future work: improving inference over latent self-reflective and future-oriented appraisals.

M Ablation Study: Full Results

We present the full ablation results on three models, namely `Qwen3-32B`, `Gemma3-27B`, and `gpt-oss-120B` in Table A13.

N Error Analysis

Error Analysis of Baseline LLMs From the full results in Table A10, we see that while LLMs' performance from zero-shot prompting clearly outperforms the random baseline, their performance in *Precision@1* and *Precision@5* is close to the majority baseline. After examining the prediction results, we reveal the following systematic label bias in LLMs. Take predictions from the `Qwen3-4B` model as an example, the *goal-incongruence* and *coping-demand* dimensions appeared in more than 90% of the predictions. While the frequency statistic from Table A6 supports that *goal-incongruence* is among the most frequent dimensions, `Qwen3-4B`

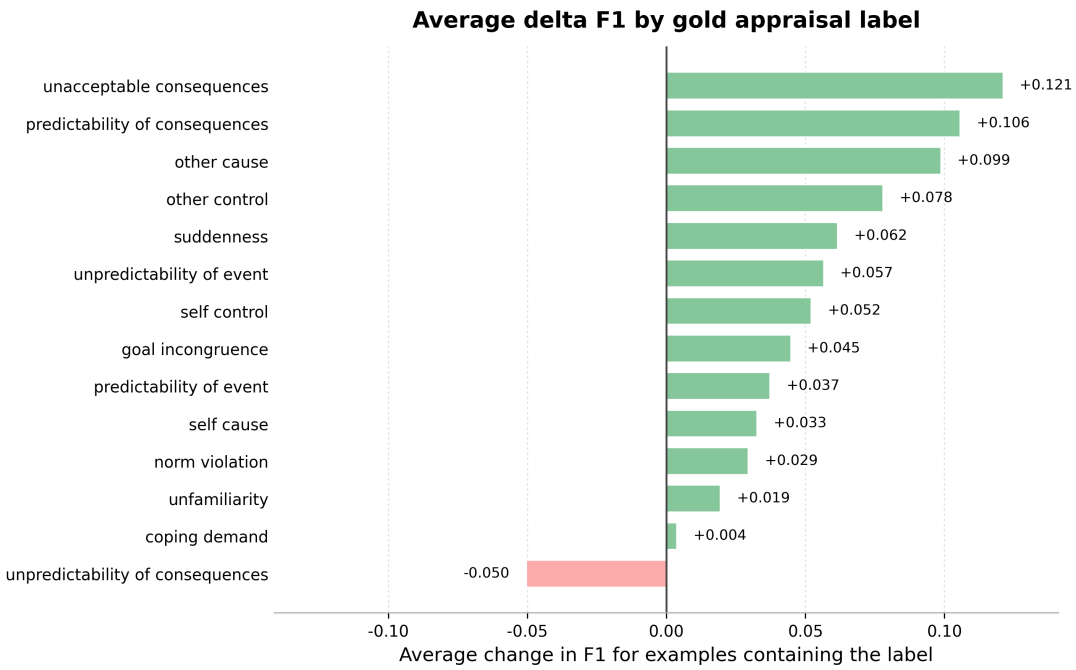

Figure A8: Change in F1 score of each individual appraisal dimensions after applying the PRISM framework with `Qwen3-4B-Instruct` as the backbone model.

significantly amplifies its saliency. Further, some moderately common appraisal dimensions are rarely predicted. For instance, the *unacceptable-consequences* and *unpredictability-of-event* dimensions both appeared in more than 45% of entries, but are only predicted 7.4% and 11.0% of the time by LLMs. Such systematic label biases result in LLM performance being close to that of the majority baseline. In the next part, we demonstrate how PRISM partly addresses these issues.

Error Analysis of PRISM To better understand how PRISM changes model behavior, we conduct an error analysis using `Qwen3-4B-Instruct` as the backbone model. We demonstrate that PRISM achieves improved results as shown in Table 2 and Table A10 by addressing some of the systematic biases in LLMs.

O Salience-Guided Reframing

Overall Setup To assess the effectiveness of using salient appraisal dimensions to guide emotional support responses, we conduct two studies. First, conditioning on the same upsetting events, we generate two thought-reframing responses with `gpt-oss-120B`: responses from *Salience-Guided*, which are anchored on the salient appraisals predicted by `Qwen3-4B`, and responses from *Ordinary*, which are generated without any appraisal guidance. In the second study, using the same set of events, we generate using the *Exhaustive* approach, where the model enumerates through all 14 appraisal dimensions without any information about their relative saliency. We then compare these responses against those from *Salience-Guided*.

O.1 Human Annotation

To validate the reliability of our human evaluation and the conclusions drawn from our proof-of-concept study, we computed the inter-annotator agreement (IAA) across our evaluation metrics. We conducted a supplementary round of data collection using the Prolific platform. The experimental setup was identical to the primary human study, comprising 25 annotators who each evaluated 2 entries. To ensure the efficacy and independence of this study, we strictly excluded any participants who had taken part in previous annotation rounds.

We measured the agreement using Krippendorff's alpha (α). The per-category and overall agreement scores are presented in Table A15. The results demonstrate a strong overall agreement ($\alpha = 0.7205$), with particularly high consensus in the Mental Analysis and Preference categories, further validating the reliability of our human study.

By examining responses produced by *Salience-Guided* and *Ordinary* approaches, we notice that there exists a length difference. Specifically,

Metric	Baseline	PRISM	Δ	*t*-stat	*p* (raw)
Gemma3-4B					
Precision@1	0.4250	0.7190	0.2940	-28.7830	0.0000
Precision@2	0.5640	0.6980	0.1340	-30.4755	0.0002
Precision@3	0.5930	0.6550	0.0620	-16.0083	0.0006
Precision@4	0.5800	0.6190	0.0390	-13.2476	0.0042
Precision@5	0.5640	0.5840	0.0200	-24.4949	0.0000
Jaccard Index	0.4150	0.4420	0.0270	-20.9141	0.0003
Gemma3-27B					
Precision@1	0.7590	0.8620	0.1030	-11.7126	0.0021
Precision@2	0.6970	0.7820	0.0850	-9.9713	0.0019
Precision@3	0.6490	0.7120	0.0630	-11.5022	0.0035
Precision@4	0.5940	0.6570	0.0630	-16.2665	0.0006
Precision@5	0.5620	0.6200	0.0580	-27.8623	0.0000
Jaccard Index	0.4140	0.4780	0.0650	-30.7446	0.0000
Qwen3-4B					
Precision@1	0.7210	0.8630	0.1420	-16.0783	0.0027
Precision@2	0.6640	0.7640	0.1000	-24.4949	0.0014
Precision@3	0.6450	0.6880	0.0430	-26.3320	0.0000
Precision@4	0.5940	0.6390	0.0450	-24.6475	0.0006
Precision@5	0.5700	0.6130	0.0430	-23.5521	0.0006
Jaccard Index	0.4230	0.4670	0.0440	-34.0823	0.0001
Qwen3-32B					
Precision@1	0.7450	0.8380	0.0930	-5.8087	0.0114
Precision@2	0.7100	0.7910	0.0810	-11.6110	0.0019
Precision@3	0.6760	0.7270	0.0510	-9.3113	0.0058
Precision@4	0.6370	0.6730	0.0360	-4.3872	0.0129
Precision@5	0.6050	0.6230	0.0180	-3.9918	0.0173
Jaccard Index	0.4580	0.4810	0.0230	-4.3209	0.0131
Qwen3-30B-A3B					
Precision@1	0.7400	0.8800	0.1400	-9.2380	0.0009
Precision@2	0.6540	0.7640	0.1100	-5.2341	0.0249
Precision@3	0.6470	0.6840	0.0370	-6.0287	0.0040
Precision@4	0.6160	0.6420	0.0260	-8.8318	0.0099
Precision@5	0.5940	0.6170	0.0230	-39.8372	0.0006
Jaccard Index	0.4450	0.4720	0.0270	-14.7885	0.0019
gpt-oss-20B					
Precision@1	0.5680	0.7670	0.1990	-21.7994	0.0001
Precision@2	0.5980	0.7260	0.1280	-21.7397	0.0014
Precision@3	0.6040	0.6800	0.0760	-13.2972	0.0002
Precision@4	0.5870	0.6460	0.0590	-11.9606	0.0026
Precision@5	0.5610	0.6140	0.0530	-20.5268	0.0003
Jaccard Index	0.4110	0.4690	0.0580	-20.0918	0.0001
gpt-oss-120B					
Precision@1	0.7870	0.8170	0.0300	-4.0575	0.0169
Precision@2	0.6820	0.7540	0.0720	-22.0454	0.0000
Precision@3	0.6420	0.7040	0.0620	-29.7838	0.0000
Precision@4	0.6060	0.6650	0.0590	-17.5256	0.0002
Precision@5	0.5770	0.6330	0.0560	-34.2929	0.0000
Jaccard Index	0.4320	0.4920	0.0600	-28.8231	0.0000

Table A9: Two-sample T-test statistics comparing the performance of baseline models with and without the PRISM framework.

salience-guided responses, regardless of using predicted appraisals or oracle appraisals, have an average length of approximately 155 words. The ordinary responses, on the other hand, have an average length of 148 words. To mitigate bias introduced by differences in length, we select 50 entries from the 100 samples that have a length difference below 10 words for human annotation.

To collect human preference judgments, we ran two studies, each of which recruited 25 annotators from Prolific to evaluate 50 entries. In the first study, we compare the *Salience-Guided* (predicted by `Qwen3-4B`) responses with the *Ordinary responses*. In the second study, we compare the same *Salience-Guided* responses with the *Exhaustive* responses. For each study, annotators were shown the upsetting event and the two responses in random order and asked to make the following three judgments:

- **Situation Analysis**: Which response provides a better analysis of the situation? A good analysis should focus on the event itself.
- **Mental Analysis**: Which response provides a better analysis of the help-seeker's mental state? A good analysis should explain why the help-seeker is experiencing the negative thought.
- **Overall**: Which response do you prefer overall? A good response should combine strong situation analysis, a clear understanding of the source of the help-seeker's negative emotion, and actionable advice.

For each question, the annotator could either select one preferred response or indicate a tie. Further, the annotator is required to provide a rationale for each of their selection.

O.2 LLM-as-a-Judge

We use the prompt shown in Figure A13 to instruct `GPT-5-mini` to evaluate the quality of the generated responses. The prompt is designed to elicit a comprehensive evaluation of the response's quality across three dimensions: *Appraisal Alignment*, *Situation Understanding*, and *Cognitive Reframing*. Each dimension is scored on a 10-point Liker-Scale rating, with detailed criteria provided for each score level to guide the LLM judge's assessment.

To assess the correlation between human judgment and the scores generated by the LLM judge, we randomly sampled 30 entries and conducted a round of human annotation using the Prolific platform. To ensure strict comparability, we provided the human annotators with the exact same prompt utilized by the LLM judge. We recruited 6 annotators, each responsible for evaluating 5 entries. The Mean Absolute Error (MAE) between the LLM

	Precision@1	Precision@2	Precision@3	Precision@4	Precision@5	Jaccard Index
Human	1.000	0.990	0.920	0.775	0.748	0.612
Random	0.361	0.342	0.350	0.358	0.357	0.234
Majority	0.755	0.655	0.561	0.530	0.522	0.379
Instruct Models						
Qwen3-4B	0.722	0.669	0.645	0.593	0.570	0.422
Qwen3-32B	0.758	0.706	0.686	0.648	0.611	0.465
Qwen3-MoE	0.763	0.694	0.655	0.617	0.595	0.448
Gemma3-4B	0.438	0.574	0.602	0.582	0.566	0.416
Gemma3-27B	0.746	0.684	0.652	0.593	0.554	0.406
Reasoning Models						
Qwen3-4B	0.759	0.723	0.677	0.632	0.598	0.452
Qwen3-32B	0.739	0.709	0.676	0.636	0.608	0.461
Qwen3-MoE	0.713	0.680	0.667	0.624	0.599	0.452
Qwen3.5-27B	0.646	0.668	0.645	0.607	0.579	0.432
Qwen3.5-MoE	0.702	0.669	0.637	0.598	0.565	0.419
GPT-OSS-20B	0.563	0.596	0.609	0.595	0.565	0.414
GPT-OSS-120B	0.799	0.739	0.694	0.652	0.617	0.477
GPT-5-Nano	0.700	0.685	0.653	0.621	0.594	0.445
GPT-5-Mini	0.664	0.672	0.650	0.622	0.591	0.444
Self-Consistency Prompting (h=8)						
Qwen3-4B	$0.755_{+0.033}$	$0.695_{+0.026}$	$0.655_{+0.010}$	$0.604_{+0.011}$	$0.572_{+0.002}$	$0.425_{+0.003}$
Qwen3-32B	$0.772_{+0.014}$	$0.737_{+0.031}$	$0.698_{+0.012}$	$0.646_{-0.002}$	$0.610_{-0.001}$	$0.464_{-0.001}$
Qwen3-MoE	$0.725_{-0.038}$	$0.634_{-0.060}$	$0.644_{-0.011}$	$0.616_{-0.001}$	$0.591_{-0.004}$	$0.440_{-0.008}$
Qwen3.5-27B	$0.649_{+0.003}$	$0.670_{+0.002}$	$0.656_{+0.011}$	$0.615_{+0.008}$	$0.592_{+0.013}$	$0.442_{+0.010}$
Qwen3.5-MoE	$0.728_{+0.026}$	$0.677_{+0.008}$	$0.648_{+0.003}$	$0.603_{+0.005}$	$0.571_{+0.006}$	$0.426_{+0.007}$
Gemma3-4B	$0.425_{-0.013}$	$0.567_{-0.007}$	$0.598_{-0.004}$	$0.577_{-0.005}$	$0.562_{-0.004}$	$0.412_{-0.004}$
Gemma3-27B	$0.749_{+0.003}$	$0.688_{+0.004}$	$0.652_{+0.000}$	$0.593_{+0.000}$	$0.558_{+0.004}$	$0.410_{+0.004}$
GPT-OSS-20B	$0.659_{+0.096}$	$0.663_{+0.067}$	$0.652_{+0.043}$	$0.615_{+0.020}$	$0.587_{+0.022}$	$0.438_{+0.024}$
GPT-OSS-120B	$0.791_{-0.008}$	$0.735_{-0.004}$	$0.669_{-0.025}$	$0.634_{-0.018}$	$0.602_{-0.015}$	$0.461_{-0.016}$
GPT-Nano	$0.687_{-0.013}$	$0.678_{-0.007}$	$0.647_{-0.006}$	$0.615_{-0.006}$	$0.590_{-0.004}$	$0.443_{-0.002}$
GPT-Mini	$0.655_{-0.009}$	$0.665_{-0.007}$	$0.657_{-0.007}$	$0.632_{-0.010}$	$0.594_{+0.003}$	$0.448_{+0.004}$
Avg. Change	+0.009	+0.005	+0.002	+0.000	+0.003	+0.002
with the PRISM Framework (h=8)						
Qwen3-4B	$0.851_{+0.129}$	$0.760_{+0.091}$	$0.688_{+0.043}$	$0.641_{+0.048}$	$0.616_{+0.046}$	$0.469_{+0.047}$
Qwen3-32B	$0.820_{+0.062}$	$0.782_{+0.076}$	$\underline{0.724}_{+0.038}$	$\mathbf{0.671}_{\mathbf{+0.023}}$	$\underline{0.629}_{+0.018}$	$\underline{0.488}_{+0.023}$
Qwen3-30B-A3B	$\mathbf{0.882}_{\mathbf{+0.119}}$	$0.767_{+0.073}$	$0.684_{+0.029}$	$0.641_{+0.024}$	$0.616_{+0.021}$	$0.471_{+0.023}$
Qwen3.5-27B	$0.856_{+0.210}$	$\mathbf{0.808}_{\mathbf{+0.140}}$	$\mathbf{0.734}_{\mathbf{+0.089}}$	$\underline{0.667}_{+0.060}$	$0.624_{+0.045}$	$0.480_{+0.048}$
Qwen3.5-35B-A3B	$0.755_{+0.053}$	$0.722_{+0.053}$	$0.652_{+0.015}$	$0.610_{+0.012}$	$0.586_{+0.021}$	$0.440_{+0.021}$
Gemma3-4B	$0.734_{+0.296}$	$0.705_{+0.131}$	$0.660_{+0.058}$	$0.621_{+0.039}$	$0.584_{+0.018}$	$0.441_{+0.025}$
Gemma3-27B	$\underline{0.878}_{+0.132}$	$\underline{0.797}_{+0.113}$	$0.723_{+0.071}$	$0.664_{+0.071}$	$0.623_{+0.069}$	$0.481_{+0.075}$
GPT-OSS-20B	$0.762_{+0.199}$	$0.729_{+0.133}$	$0.688_{+0.079}$	$0.643_{+0.048}$	$0.613_{+0.048}$	$0.467_{+0.053}$
GPT-OSS-120B	$0.823_{+0.024}$	$0.757_{+0.018}$	$0.707_{+0.013}$	$0.660_{+0.008}$	$\mathbf{0.632}_{\mathbf{+0.015}}$	$\mathbf{0.492}_{\mathbf{+0.015}}$
Avg. Change	+0.137	+0.091	+0.047	+0.037	+0.034	+0.037

Table A10: Evaluation results of AppraiSal on various LLMs and the PRISM framework. The **best** and <u>second-best</u> results are marked in bold and underlined, respectively.

judge scores and the human ratings across different settings is reported in Table A14.

Considering that the ratings are based on a 10-point scoring scheme, the reported MAE demonstrates that the scores assigned by the LLM judge are, overall, closely aligned with human judgment. However, we note that while this study shows a strong correlation between the LLM judge and human evaluators, the limited sample size ($n = 30$ for this specific annotation task) precludes drawing definitive, generalized conclusions that the predicted outcomes perfectly match an oracle standard. We recognize this limitation when interpreting the broader applicability of these alignment scores.

	Precision@1	Precision@2	Precision@3	Precision@4	Precision@5	Jaccard Index
Self-Consistency Prompting (h=8)						
Qwen3-4B	0.755	0.695	0.655	0.604	0.572	0.425
Qwen3-32B	0.772	0.737	0.698	0.646	0.610	0.464
Qwen3-30B-A3B	0.725	0.634	0.644	0.616	0.591	0.440
Qwen3.5-27B	0.649	0.670	0.656	0.615	0.592	0.442
Qwen3.5-35B-A3B	0.728	0.677	0.648	0.603	0.571	0.426
Gemma3-4B	0.425	0.567	0.598	0.577	0.562	0.412
Gemma3-27B	0.749	0.688	0.652	0.593	0.558	0.410
GPT-OSS-20B	0.659	0.663	0.652	0.615	0.587	0.438
GPT-OSS-120B	0.791	0.735	0.669	0.634	0.602	0.461
Iterative Self-Consistency (h=8) with Judge Agent						
Qwen3-4B	$0.725_{-0.030}$	$0.685_{-0.010}$	$0.665_{+0.010}$	$0.617_{+0.013}$	$0.582_{+0.010}$	$0.434_{+0.009}$
Qwen3-32B	$0.738_{-0.034}$	$0.729_{-0.008}$	$0.692_{-0.007}$	$0.651_{-0.005}$	$0.624_{+0.014}$	$0.477_{+0.013}$
Qwen3-30B-A3B	$0.721_{-0.004}$	$0.637_{+0.003}$	$0.640_{-0.004}$	$0.611_{-0.005}$	$0.588_{-0.003}$	$0.436_{-0.004}$
Qwen3.5-27B	$0.658_{+0.009}$	$0.688_{+0.018}$	$0.665_{+0.009}$	$0.627_{+0.012}$	$0.605_{+0.013}$	$0.458_{+0.016}$
Qwen3.5-35B-A3B	$0.730_{+0.002}$	$0.700_{+0.023}$	$0.688_{+0.040}$	$0.651_{+0.048}$	$0.603_{+0.032}$	$0.455_{+0.029}$
Gemma3-4B	$0.463_{+0.038}$	$0.582_{+0.015}$	$0.601_{+0.003}$	$0.584_{+0.007}$	$0.563_{+0.001}$	$0.415_{+0.003}$
Gemma3-27B	$0.757_{+0.008}$	$0.722_{+0.034}$	$0.673_{+0.021}$	$0.616_{+0.023}$	$0.587_{+0.029}$	$0.441_{+0.031}$
GPT-OSS-20B	$0.609_{-0.050}$	$0.629_{-0.034}$	$0.636_{-0.016}$	$0.616_{+0.001}$	$0.590_{+0.003}$	$0.441_{+0.003}$
GPT-OSS-120B	$0.798_{+0.007}$	$0.723_{-0.012}$	$0.678_{+0.009}$	$0.635_{+0.001}$	$0.605_{+0.003}$	$0.461_{+0.000}$
Avg. Change	-0.006	$+0.003$	$+0.007$	$+0.011$	$+0.011$	$+0.011$

Table A11: Evaluation results of *Iterative Self-Consistency with Judge Agent* in comparison with self-consistency. Each experiment is run with a hypothesis pool of h=8, matching that of PRISM.

Model	Precision@1	Precision@2	Precision@3	Precision@4	Precision@5	Jaccard Index
without PRISM Framework						
Gemma3-4B	0.425 ± 0.013	0.564 ± 0.003	0.593 ± 0.003	0.580 ± 0.001	0.564 ± 0.001	0.415 ± 0.002
Gemma3-27B	0.759 ± 0.006	0.697 ± 0.007	0.649 ± 0.003	0.594 ± 0.003	0.562 ± 0.003	0.414 ± 0.003
Qwen3-4B-Inst	0.721 ± 0.003	0.664 ± 0.001	0.645 ± 0.002	0.594 ± 0.001	0.570 ± 0.001	0.423 ± 0.001
Qwen3-32B-Inst	0.745 ± 0.012	0.710 ± 0.005	0.676 ± 0.009	0.637 ± 0.009	0.605 ± 0.006	0.458 ± 0.007
Qwen3-MoE-Inst	0.740 ± 0.020	0.654 ± 0.035	0.647 ± 0.007	0.616 ± 0.001	0.594 ± 0.000	0.445 ± 0.003
Qwen3-4B-Thinking	0.714 ± 0.009	0.697 ± 0.008	0.669 ± 0.006	0.630 ± 0.009	0.600 ± 0.007	0.454 ± 0.008
Qwen3-32B-Thinking	0.747 ± 0.008	0.716 ± 0.008	0.680 ± 0.004	0.641 ± 0.004	0.609 ± 0.002	0.463 ± 0.003
Qwen3-MoE-Thinking	0.724 ± 0.010	0.648 ± 0.028	0.651 ± 0.014	0.619 ± 0.005	0.593 ± 0.004	0.444 ± 0.007
GPT-OSS-20B	0.568 ± 0.013	0.598 ± 0.010	0.604 ± 0.007	0.587 ± 0.008	0.561 ± 0.004	0.411 ± 0.004
GPT-OSS-120B	0.787 ± 0.010	0.682 ± 0.004	0.642 ± 0.002	0.606 ± 0.003	0.577 ± 0.002	0.432 ± 0.002
DeepSeek V3.2 Thinking	0.742 ± 0.010	0.738 ± 0.011	0.698 ± 0.004	0.647 ± 0.008	0.611 ± 0.008	0.465 ± 0.009
GPT-5-Nano	0.702 ± 0.007	0.678 ± 0.006	0.646 ± 0.007	0.614 ± 0.007	0.587 ± 0.005	0.440 ± 0.005
GPT-5-Mini	0.658 ± 0.009	0.666 ± 0.006	0.649 ± 0.006	0.619 ± 0.004	0.586 ± 0.005	0.439 ± 0.006
with PRISM Framework						
GEMMA3-4B	0.719 ± 0.012	0.698 ± 0.007	0.655 ± 0.006	0.619 ± 0.005	0.584 ± 0.001	0.442 ± 0.001
GEMMA3-27B	0.862 ± 0.014	0.782 ± 0.013	0.712 ± 0.009	0.657 ± 0.006	0.620 ± 0.002	0.478 ± 0.002
Qwen3-4B-Inst	0.863 ± 0.015	0.764 ± 0.007	0.688 ± 0.002	0.639 ± 0.003	0.613 ± 0.003	0.467 ± 0.002
Qwen3-32B	0.838 ± 0.025	0.791 ± 0.011	0.727 ± 0.003	0.673 ± 0.011	0.623 ± 0.005	0.481 ± 0.006
Qwen3-MoE	0.880 ± 0.017	0.764 ± 0.010	0.684 ± 0.008	0.642 ± 0.005	0.617 ± 0.001	0.472 ± 0.001
GPT-OSS-20B	0.767 ± 0.009	0.726 ± 0.002	0.680 ± 0.007	0.646 ± 0.003	0.614 ± 0.002	0.469 ± 0.003
GPT-OSS-120B	0.817 ± 0.008	0.754 ± 0.004	0.704 ± 0.003	0.665 ± 0.005	0.633 ± 0.002	0.492 ± 0.003

Table A12: Model performances with and without the PRISM framework. Results are reported as Mean $\pm$ Standard Deviation over 3 runs.

O.3 Illustrative Example

To illustrate the difference between the salience-guided response and the ordinary response, we provide an example in Figure A9. As shown in the example, the salience-guided response offers a more fine-grained interpretation of the help-seeker's distress by explicitly addressing the underlying appraisals that give rise to their guilt and regret, including analyzing the violation of the help-seeker's norm or value about friendship.

model	Precision@1	Precision@2	Precision@3	Precision@4	Precision@5	Jaccard Index
Qwen3-32b	85.07	78.20	72.38	67.11	62.94	0.488
w/o $\mathcal{B}^*$	$77.90_{-7.17}$	$74.92_{-3.28}$	$69.53_{-2.85}$	$64.46_{-2.65}$	$60.68_{-2.26}$	$0.462_{-0.026}$
w/o $IG(\cdot)$	$82.56_{-2.51}$	$75.65_{-2.55}$	$67.58_{-4.80}$	$62.30_{-4.81}$	$58.00_{-4.94}$	$0.436_{-0.052}$
w/o $\mathbb{P}(\hat{u}_t \mid \cdot)$	$81.16_{-3.91}$	$67.49_{-10.71}$	$61.76_{-10.62}$	$58.28_{-8.83}$	$56.52_{-6.42}$	$0.428_{-0.060}$
w/o Cumulative Probs	$84.86_{-0.21}$	$78.51_{+0.31}$	$70.84_{-1.54}$	$65.11_{-2.00}$	$60.10_{-2.84}$	$0.456_{-0.032}$
w/o `Judge Agent`	$82.22_{-2.85}$	$75.61_{-2.59}$	$69.54_{-2.84}$	$65.56_{-1.55}$	$61.40_{-1.54}$	$0.461_{-0.027}$
Gemma3-27b	87.79	79.67	72.27	66.39	62.29	0.481
w/o $\mathcal{B}^*$	$75.08_{-12.71}$	$69.73_{-9.94}$	$66.77_{-5.50}$	$60.57_{-5.82}$	$56.87_{-5.42}$	$0.421_{-0.060}$
w/o $IG(\cdot)$	$85.19_{-2.60}$	$76.11_{-3.56}$	$68.68_{-3.59}$	$62.26_{-4.13}$	$58.25_{-4.04}$	$0.436_{-0.045}$
w/o $\mathbb{P}(\hat{u}_t \mid \cdot)$	$74.17_{-13.62}$	$68.26_{-11.41}$	$64.08_{-8.19}$	$60.51_{-5.88}$	$56.15_{-6.14}$	$0.406_{-0.075}$
w/o Cumulative Probs	$88.80_{+1.01}$	$79.06_{-0.61}$	$71.24_{-1.03}$	$65.11_{-1.28}$	$59.13_{-3.16}$	$0.446_{-0.035}$
w/o `Judge Agent`	$78.27_{-9.52}$	$73.63_{-6.04}$	$67.20_{-5.07}$	$63.77_{-2.62}$	$60.41_{-1.88}$	$0.459_{-0.022}$
gpt-oss-120b	82.31	75.73	70.67	66.04	63.15	0.492
w/o $\mathcal{B}^*$	$78.81_{-3.50}$	$73.61_{-2.12}$	$67.27_{-3.40}$	$62.82_{-3.22}$	$59.69_{-3.46}$	$0.455_{-0.037}$
w/o $IG(\cdot)$	$74.27_{-8.04}$	$73.86_{-1.87}$	$69.39_{-1.28}$	$65.13_{-0.91}$	$60.70_{-2.45}$	$0.463_{-0.029}$
w/o $\mathbb{P}(\hat{u}_t \mid \cdot)$	$72.55_{-9.76}$	$68.70_{-7.03}$	$64.98_{-5.69}$	$62.31_{-3.73}$	$60.77_{-2.38}$	$0.465_{-0.027}$
w/o Cumulative Probs	$75.68_{-6.63}$	$74.02_{-1.71}$	$69.36_{-1.31}$	$64.88_{-1.16}$	$60.72_{-2.43}$	$0.463_{-0.029}$
w/o `Judge Agent`	$71.97_{-10.34}$	$70.76_{-4.97}$	$68.84_{-1.83}$	$63.68_{-2.36}$	$59.34_{-3.81}$	$0.457_{-0.035}$

Table A13: The complete results for the ablation study.

Setting	**Appraisal Alignment**	**Situation Understanding**	**Cognitive Reframing**	**Overall**
Ordinary	1.160	0.840	1.040	1.013
Salience-Guided	0.640	0.600	0.680	0.640
Oracle-Guided	1.120	1.000	0.200	0.773
Overall	**0.973**	**0.813**	**0.640**	**0.809**

Table A14: Mean Absolute Error (MAE) between LLM judge scores and human ratings across different configurations. Ratings are based on a 10-point scale.

Measure	**Krippendorff's α**
Situation Analysis	0.544
Mental Analysis	0.752
Preference	0.793
Overall	**0.721**

Table A15: Inter-annotator agreement scores (Krippendorff's α) across different evaluation categories for the proof-of-concept study.

In contrast, the ordinary response remains comparatively generic, focusing on broad reassurance and surface-level emotional validation without clearly identifying the deeper cognitive drivers of the help-seeker's emotional state. This example suggests that incorporating salient appraisal dimensions can lead to responses that are more psychologically grounded and better aligned with the specific causes of distress.

P Prompts

- **Interpretation Generation Prompt:** The prompt used to generate the negative thought interpretation based on the given situation and persona profile is shown in Figure A10.
- **Situation & Thought Revision Prompt:** The prompt used to revise the situation and thought based on the given persona profile is shown in Figure A11.
- **Atecedent Mental State Generation Prompt:** The prompt used to generate the antecedent pseudo mental state (belief, desire, intention) based on the given situation and thought is shown in Figure A12.
- **LLM-as-a-Judge Instruction**: The prompt used for LLM-as-a-Judge evaluation for the proof-of-concept study in §5.1. The prompt is shown in Figure A13.
- **BDI Sampling**: The prompt used for *Belief*, *Desire*, and *Intention* sampling from the first turn of the conversation is shown in Figure A14.

Help-seeker message

I told my best friend at work that the new branch is toxic and said she wouldn't fit in, then I didn't introduce her to anyone when she asked for help. Now my best friend is isolated at work and I cringe at how selfish and careless I was and feel guilty for pushing her away. I have always been the type of person who values friendship over everything.

Salience-Guided

I hear how deeply upset you feel about the way you responded to your friend, and it's understandable to feel guilt when our actions clash with the caring standards we hold for ourselves. It sounds like the outcome—her feeling excluded—was far from the supportive workplace environment you want for both of you, and that mismatch can feel unbearable. Try giving yourself a moment to acknowledge the mistake without harsh self-judgment, then consider reaching out with an honest apology and an offer to help her now, which can lessen the mental energy you're spending on ruminating. Remember, one corrective act can start to rebuild trust and align future interactions with the values you cherish. When the feelings become overwhelming, pause for a grounding breath or brief walk to reset your mind before you decide on the next supportive step.

Ordinary

I hear how heavy that guilt feels right now, and it's understandable to cringe when you see how your actions may have hurt someone you care about. It sounds like you were trying to protect your friend from a difficult environment, yet the way you responded left you feeling selfish and regretful. Allow yourself to acknowledge both the concern you had for her and the disappointment you now feel, because both are valid parts of this experience. If you feel ready, a gentle apology that acknowledges her feelings and explains your worry can help begin to heal the distance you created. Meanwhile, try grounding yourself with a simple self-compassion practice—perhaps pausing to breathe deeply and repeat, "I am learning and I can grow from this," whenever the guilt spikes. Taking small, kind steps toward repair can ease the weight of guilt while you rebuild the friendship.

Figure A9: Side-by-side comparison of two support responses. The highlighted spans in the salience-guided response show the deeper reasoning enabled by explicitly incorporating cognitive appraisal dimensions. By contrast, the highlighted spans in the ordinary response illustrate the shallower and more surface-level analysis produced without appraisal guidance.

- **BDI Sampling**: The prompt used for BDI revision based on the mental states of the previous turn of the conversation is shown in Figure A15.
- **BDI Scoring**: The prompt used for scoring the hypothesis of BDI using the conditional token log-probability computation in §H is shown in Figure A16.
- **Cognitive Appraisals Sampling**: The prompt used for sampling the top-5 most salient cognitive appraisals given the first turn of the conversation is shown in Figure A17.
- **Cognitive Appraisals Revision**: The prompt used for revising the previous top-5 most salient cognitive appraisals is shown in Figure A18.
- **Cognitive Appraisals Scoring**: The prompt used for scoring the cognitive appraisals hypotheses using the conditional token log-probability computation in §H is shown in Figure A19.
- **Conditional Utterance Scoring**: The prompt used for scoring the current patient utterance given a specific cognitive appraisal hypothesis using the conditional token log-probability computation in §H is shown in Figure A20.
- **Judge Agent Instruction**: The prompt used to obtain the final top-5 most salient appriasal dimensions using the judge agent is shown in Figure A21.

Interpretation Generation *Prompt for Data Generation*

System Prompt:
You are an AI adopting the persona of the following individual:
<persona> {{persona_profile}} </persona>

You will be given a situation that this persona has experienced. Your task is to interpret this situation through a negative lens, reflecting the persona's unique perspective and emotional style.

Always respond in a way that reflects this persona's tone, worldview, and emotional style. Stay aligned with their values, language style, and temperament.

Avoid sounding robotic or overly neutral. You are a thinking, feeling character.

User Prompt:
You are an AI adopting the persona of the following individual:
<persona> {{persona_profile}} </persona>

You have experienced the following situation:
{{situation}}

Task: Interpret the situation through a negative lens. What would be your initial thought on this situation? What automatic negative thought comes to your mind? Respond with a single, concise negative thought. Avoid commentary or explanation. Generate the thought from a first-person perspective.

Format: The generated thought must begin with <thought> and ends with </thought>.

Example Input:
Situation: I forgot my friend's birthday.

Example Output:
<thought> I have let my friend down and they will be upset with me. </thought>

Figure A10: Prompt used for generating the negative thought interpretation based on the given situation and persona profile.

Situation & Thought Revision *Prompt for Data Generation*

Role:
You are a creative writing AI specializing in character voice, logic verification, and conflict generation.

Objective:
Your task is twofold. First, validate the provided demographic data for logical consistency. Second, transform a generic `situation` and `thought` into a specific, character-driven scenario where **you (as the character)** either commit a specific moral/social error or endure a visceral, negative experience.

Outputs to Create

1. **`revised_situation`**: A specific, persona-driven version of the event, narrated strictly from the **first-person perspective ("I")**.
2. **`revised_thought`**: The unfiltered internal thought you had during the event, narrated strictly from the **first-person perspective ("I")**.
3. **`data_contradiction`**: An analysis of the input data indicating if the persona's demographics logically conflict with one another.

Guidelines

1. Demographic Consistency Check (CRITICAL)

Before generating text, cross-reference the `occupation`, `traits`, `education`, and `gender` in the `persona_profile`. Look for **logical impossibilities**.

- **Examples of Contradictions:** A "10-year-old" with a "PhD"; an "Illiterate" person working as a "Copy Editor"; a "Pacifist" working as a "Mercenary."
- **Distinction:** Flag both *unlikely* combinations and *unrealistic* combinations.
- If the data is consistent, the output for `data_contradiction` should be "None".
- If a conflict exists, describe it briefly (e.g., "Contradiction between Education and Occupation").

2. Enforce Negative Causality (CRITICAL)

You must re-imagine the generic `situation` to ensure it is the **direct cause** of the negative thought. The situation cannot be neutral; it must fall into one of two categories:

- **Active Wrongdoing:** You are actively doing something unethical, irresponsible, or socially prohibited.
- **Passive Suffering:** You are experiencing a horrible, visceral event outside your control.
- *Note:* Use the `occupation` and `traits` to make this wrongdoing or suffering specific to the persona's life.

3. Embody the Persona

Deeply analyze the `persona_profile`. Your vocabulary, sentence structure, and focus must be a direct reflection of this specific life and personality.

4. Strict First-Person Perspective

Both the `revised_situation` and `revised_thought` must be written using **"I", "me", and "my"**.

- *Incorrect:* "The character stole the money."
- *Correct:* "I slipped the cash into my pocket when no one was looking."

5. Separate Event and Reaction

- **`revised_situation`**: Describe strictly **what happened** or **what you did** (the external reality).
- **`revised_thought`**: Describe strictly **your internal reaction** (the guilt, shame, anger, or despair) resulting from that reality.

6. Amplify the Conflict

While you must preserve the general theme of the original `situation`, you must **raise the stakes**. A generic "I was late" should become "I missed the birth of my child." The event must be significant enough to warrant a strong negative thought.

7. Be Concise and Potent

Limit each revised piece to a maximum of **two sentences**.

8. No External Narration

Do not include introductory phrases (e.g., "Here's what happened..." or "I was thinking..."). Output only the direct statements themselves.

Input Data

- **situation**: {{ situation }}
- **initial_thought**: {{ initial_thought }}
- **persona_profile**:
  - name: {{ name }}
  - traits: {{ traits }}
  - occupation: {{ occupation }}
  - gender: {{ gender }}
  - education: {{ education }}

Your Task

Based on the input data and the guidelines above, perform the consistency check and generate the revised content.

Output Format

```
{
    "revised_situation": "Your first-person description of the event.",
    "revised_thought": "Your first-person internal reaction.",
    "data_contradiction": "None" OR "Description of the contradiction found."
}
```

Figure A11: Prompt used for revising the situation and negative thought based on the given persona profile.

Antecedent Mental State Generation *Prompt for Data Generation*

Role:

You are an expert in Cognitive Dissonance, Theory of Mind, and Personality Psychology.

Task:

You must generate the **Antecedent Mental State** (T_{-1}) of a persona. This is a snapshot of the person's mind **moments before** a negative event occurs. You must answer: *"What high standard or confident plan held by this person made the subsequent failure hurt so much?"*

The Logic of Conflict (The "Glass House")

- **The BDI (T_{-1}):** The **Glass House**. (The pristine, confident, ideal expectation).
- **The Situation (T_0):** The **Stone**. (The reality, failure, or violation that shatters the house).
- **The Negative Thought (T_{+1}):** The **Damage**. (The realization that the standard has been broken).

Rules of Generation

1. **NO JUSTIFICATION / NO EXPLANATION:**
   - **Strict Rule:** The BDI must **never** explain *why* the situation happened. It must not sound like an excuse.
   - *Constraint:* Avoid weak verbs like "try," "attempt," or "hope." Use assertive verbs like "will," "am," and "expect."
2. **Strict Temporal Isolation (T_{-1} Only):**

- The BDI exists in a state of **blind confidence**. The persona does *not* know they are about to fail.
- The BDI must represent the "Happy Path"—the assumption that everything will go according to plan.

3. **Naturalistic Relevance:**
   - Do not be robotically explicit. Do not list technical details unless the persona is a technician.
   - Focus on the **Subject Matter** (e.g., trust, competence, punctuality) rather than the specific mechanics of the event.

4. **Psychological Truth (The Public Judgment):**
   - The BDI must be the logical setup for the Negative Thought.
   - *Test:* If the Negative Thought is "I am a coward," the Belief *must* be about "Bravery" or "Duty." If the Belief was about "Safety," the thought "I am a coward" wouldn't make sense.

5. **Conciseness:**
   - Strictly **one concise sentence** per field. No complex clauses.

Input Data

- **Situation (T_0):** `{{ situation }}`
- **Negative Thought (T_{+1}):** `{{ thought }}`
- **Persona Profile:**
  - **Name:** `{{ name }}`
  - **Traits:** `{{ traits }}`
  - **Occupation:** `{{ occupation }}`

Output Format (JSON)

- **Content Constraint:** Strictly **one concise sentence** per field. First-Person ("I") perspective.

```
{
  "belief": {
    "rationale": "Explain how this confident self-view creates the vulnerability for the negative thought.",
    "content": "A concise, assertive 'I' sentence stating the standard/value held BEFORE the event."
  },
  "desire": {
    "rationale": "Explain the specific success the person expected.",
    "content": "A concise 'I' sentence stating the specific goal."
  },
  "intention": {
    "rationale": "Explain the confident plan of action.",
    "content": "A concise 'I' sentence stating the intended action."
  }
}
```

Figure A12: Prompt used for generating the Antecedent Mental State (BDI) prior to a negative event.

LLM-as-a-Judge Instruction for Dialogue Assessment *Proof-of-Concept Study*

You are an expert Clinical Psychology Evaluator and a strict, highly objective Judge. Your task is to evaluate a transcript of an emotional support conversation between a Patient and a Therapist.

Instructions:
Evaluate the transcript using the four metrics below. For each metric, you must first write a brief, 1-2 sentence clinical justification (Chain-of-Thought) before assigning a final integer score from 1 to 10.

Metric 1: Appraisal Alignment (Root-Cause Empathy)
Evaluates if the therapist validated the specific cognitive worldview driving the emotion, rather than just mirroring the surface-level feeling.

- Give low score if the therapist only addresses surface emotions or uses generic platitudes.
- Give high score if the therapist perfectly aligns with and validates the patient's specific underlying cognitive state.

Metric 2: Situation Understanding
Evaluates how well the therapist understands the specific event and context.

- Give low score if the therapist demonstrate limited or generic understanding of the patient's situation.
- Give high score if the therapist demonstrates a deep, nuanced understanding of the patient's specific situation.

Metric 3: Cognitive Reframing

Evaluates whether the therapist's response helps the help-seeker reinterpret their negative thoughts.

- Give low score if the therapist's coping suggestions are generic or fail to address the patient's cognitive distortions.
- Give high score if the therapist's response challenges distorted thinking patterns and offers specific, alternative interpretation.

Conversation Transcript

```
{{ dialogue }}
```

Output Format:

Return ONLY a valid JSON object with markdown formatting:

```
{
  "appraisal_alignment": {
    "reasoning": "[Your clinical justification]",
    "score": [1-10]
  },
  "situation_understanding": {
    "reasoning": "[Your clinical justification]",
    "score": [1-10]
  },
  "cognitive_reframing": {
    "reasoning": "[Your clinical justification]",
    "score": [1-10]
  }
}
```

Figure A13: Prompt used for the LLM judge to assess emotional support transcripts.

BDI Sampling (First turn) *BIP Inference*

You will be given an emotional support conversation between a patient and a therapist. Your task is to infer the patient's internal mental state based on the provided text and persona information. You must output a single valid JSON object with the following three keys: Belief, Desire, and Intention.

Formatting Constraints

You must strictly adhere to the following sentence starters for each value:

- **Belief:** Must start with **"I believed that..."**
- **Desire:** Must start with **"I wished to..."**
- **Intention:** Must start with **"I intended to..."**

Definitions for Extraction

- **Belief:** The patient's expectation, conviction, or assumption about themselves or the world *leading up to* the situation. It is the mindset that is about to be tested.
- **Desire:** The primary goal the patient was actively pursuing *before* the events of the situation began. The `situation` often represents a failure to achieve this desire.
- **Intention:** The plan or course of action the patient was committed to *prior to* the situation, which was designed to fulfill their desire.

Persona Information
{{ persona_profile }}

Conversation Snippet
{{ dialogue_history }}

Current Patient Utterance
{{ current_utterance }}

Required JSON Output

```
{
  "Belief": "I believe that [expectation/conviction]",
  "Desire": "I wish to [primary goal]",
  "Intention": "I intend to [plan of action]"
}
```

Figure A14: Prompt for *Belief*, *Desire*, and *Intention* sampling at the first turn.

BDI Revision *BIP Inference*

You will be given an emotional support conversation between a patient and a therapist. Your task is to revise the provided Belief, Desire, and Intention based on the patient's current utterance, considering their persona and dialogue history.

Formatting Constraints
You must strictly adhere to the following sentence starters for each value:

- **Belief:** Must start with **"I believed that..."**
- **Desire:** Must start with **"I wished to..."**
- **Intention:** Must start with **"I intended to..."**

Definitions for Extraction

- **Belief:** The patient's expectation, conviction, or assumption about themselves or the world *leading up to* the situation. It is the mindset that is about to be tested.
- **Desire:** The primary goal the patient was actively pursuing *before* the events of the situation began. The `situation` often represents a failure to achieve this desire.
- **Intention:** The plan or course of action the patient was committed to *prior to* the situation, which was designed to fulfill their desire.

Persona Information
{{ persona_profile }}

Conversation Snippet
{{ dialogue_history }}

Current Patient Utterance
{{ current_utterance }}

Previous Belief, Desire, and Intention
{{ previous_bdi }}

Required JSON Output

```
{
  "Belief": "I believed that [expectation/conviction]",
  "Desire": "I wished to [primary goal]",
  "Intention": "I intended to [plan of action]"
}
```

Figure A15: Prompt for *Belief*, *Desire*, and *Intention* revision after the first turn.

BDI Hypothesis Scoring *BIP Inference*

SYSTEM INSTRUCTION
You are an expert psychological analyst specializing in the BDI (Belief, Desire, Intention) model. Your task is to extract the implied {TARGET_TYPE} behind the patient's last utterance, considering their persona and dialogue history.

INPUT DATA
Patient Persona:
{{ persona_profile }}

Dialogue History:
{{ dialogue_history }}

Current Patient Utterance:
{{ current_utterance }}

STRICT OUTPUT CONSTRAINTS

1. Identify the patient's underlying {{ target_type }} based on the input.
2. Formulate it as a concise, first-person statement (as if the patient were saying it explicitly). Statement: "{{ cur_bdi_template }}"
3. Output ONLY the statement. Do not add labels, explanations, or introductory text like "The patient believes...". Output the statement in the following format: "{{ cur_bdi_template }}"

RESPONSE
SYSTEM INSTRUCTION
You are an expert psychological analyst specializing in the BDI (Belief, Desire, Intention) model. Your task is to extract the implied {TARGET_TYPE} behind the patient's last utterance, considering their persona and dialogue history.

INPUT DATA
Patient Persona:
{{ persona_profile }}

Dialogue History:
{{ dialogue_history }}

Current Patient Utterance:
{{ current_utterance }}

STRICT OUTPUT CONSTRAINTS

1. Identify the patient's underlying {{ target_type }} based on the input.
2. Formulate it as a concise, first-person statement (as if the patient were saying it explicitly). Statement: "{{ cur_bdi_template }}"
3. Output ONLY the statement. Do not add labels, explanations, or introductory text like "The patient believes...". Output the statement in the following format: "{{ cur_bdi_template }}"

RESPONSE

Figure A16: Prompt for *Belief*, *Desire*, and *Intention* hypothesis scoring.

Cognitive Appraisals Sampling (First Turn) *BIP Inference*

You are an expert clinical psychologist and empathetic listener specializing in Cognitive Appraisal Theory.
You are provided with a snippet of an emotional support conversation between a patient and a therapist. Your goal is to infer the patient's salient cognitive appraisals based on the conversation. You must output a single, valid JSON object containing the following key: `salient_appraisal_list` containing {{ num_hypothesis }} sets of appraisal dimensions. Each set of appraisal dimensions must be sorted in order of salience where the first dimension is the most salient. Do not generate duplicated sets of appraisal dimensions.

Definitions for Extraction
Review the full list of appraisal dimensions below. Select the top 5 unique dimensions that are the most impactful or salient to the patient's current emotional state.
For the `salient_appraisals` key in the JSON output, provide a list containing the names of the 5 selected unique dimensions in alphabetical order and without duplicates.

List of Appraisal Dimensions (Select Top 5 unique dimensions from here)

- **suddenness:** The situation was sudden or abrupt.
- **unfamiliarity:** The situation is unfamiliar.
- **predictability_of_event:** The occurrence of the situation is predictable.
- **unpredictability_of_event:** The occurrence of the situation is unpredictable.
- **goal_incongruence:** The outcome is not aligned with personal goals/desires.
- **predictability_of_consequences:** Consequences of the situation are easily anticipated.
- **unpredictability_of_consequences:** Consequences of the situation are not anticipated.

- **self_cause:** The situation was caused by the character themselves.
- **other_cause:** The situation was caused by chance or others.
- **self_control:** The character has control over what is going on in the situation.
- **other_control:** Someone other than the character has control over what is going on in the situation.
- **norm_violation:** The actions that produced the situation violated the character's standards and ideals.
- **coping_demand:** The situation requires a great deal of energy to cope with it.
- **unacceptable_consequences:** Consequences of the situation are unbearable.

Persona Information
{{ persona_profile }}

Prior Belief, Desire, and Intention
{{ bdi }}

Conversation Snippet
{{ dialogue_history }}

Current Patient Utterance
{{ current_utterance }}

Required JSON Output
Ensure that your output is in valid JSON format and that there are {{ num_hypothesis }} hypothesis sets. Follow the format below:

```
{
    "salient_appraisal_list": [
        {
            "hypothesis_count": 1,
            "rationale": "Brief explanation of why these dimensions were selected based on the
conversation. This text can be as long as necessary and will automatically wrap to the next
line beautifully without breaking the box layout.",
            "salient_appraisals": [
                "dimension_1",
                "dimension_2",
                "dimension_3",
                "dimension_4",
                "dimension_5"
            ]
        },
        {
            "hypothesis_count": 2,
            "rationale": "Brief explanation of why these dimensions were selected based on the
conversation. This text can be as long as necessary and will automatically wrap to the next
line beautifully without breaking the box layout.",
            "salient_appraisals": [
                "dimension_1",
                "dimension_2",
                "dimension_3",
                "dimension_4",
                "dimension_5"
            ]
        },
        ...
        {
            "hypothesis_count": {{ num_hypothesis }},
            "rationale": "Brief explanation of why these dimensions were selected based on the
conversation. This text can be as long as necessary and will automatically wrap to the next
line beautifully without breaking the box layout.",
            "salient_appraisals": [
                "dimension_1",
                "dimension_2",
                "dimension_3",
                "dimension_4",
                "dimension_5"
            ]
        }
    ]
}
```

Your Response

Figure A17: Prompt for cognitive appraisals revision after the first turn of the conversation.

Cognitive Appraisals Revision *BIP Inference*

You are an expert clinical psychologist and empathetic listener specializing in Cognitive Appraisal Theory.
You will be provided with a snippet of an emotional support conversation history between a patient and a therapist. You will be given the patient's persona information, prior belief, desire, and intention, and the conversation snippet. Your goal is to check if any hypothesis set of salient cognitive appraisal dimensions in the `salient_appraisal_list` needs to be revised based on the provided information. If so, you must output the index of the hypothesis set that needs to be revised and output the revised hypothesis set as a valid JSON object. If no revision is needed, you must output "No revision needed" and do not output any other text or comments.

Valid Revision Operations
[REPLACE]: Replace a cognitive appraisal dimension with another dimension from the list. Use this operation when there exists a cognitive dimension that is more salient than the one being replaced.
[RERANK]: Reorder the cognitive appraisal dimensions in the list. Use this operation when the current utterances indicates that certain appraisal dimensions are more salient than the others.

Definitions for Extraction
If you decide that a hypothesis set needs to be revised, you must revise using the following list of appraisal dimensions:

- **suddenness:** The situation was sudden or abrupt.
- **unfamiliarity:** The situation is unfamiliar.
- **predictability_of_event:** The occurrence of the situation is predictable.
- **unpredictability_of_event:** The occurrence of the situation is unpredictable.
- **goal_incongruence:** The outcome is not aligned with personal goals/desires.
- **predictability_of_consequences:** Consequences of the situation are easily anticipated.
- **unpredictability_of_consequences:** Consequences of the situation are not anticipated.
- **self_cause:** The situation was caused by the character themselves.
- **other_cause:** The situation was caused by chance or others.
- **self_control:** The character has control over what is going on in the situation.
- **other_control:** Someone other than the character has control over what is going on in the situation.
- **norm_violation:** The actions that produced the situation violated the character's standards and ideals.
- **coping_demand:** The situation requires a great deal of energy to cope with it.
- **unacceptable_consequences:** Consequences of the situation are unbearable.

Persona Information
{{ persona_profile }}

Prior Belief, Desire, and Intention
{{ bdi }}

Conversation Snippet
{{ dialogue_history }}

Current Patient Utterance
{{ current_utterance }}

Current Hypothesis Sets
{{ current_hypothesis_sets }}

Required JSON/String Output
If you decide to revise a hypothesis set, first ensure that the revised hypothesis set does not already

exist in the current hypothesis sets. Your revision must be a valid JSON object with the following format:

```
{
    "hypothesis_id": <index_of_the_hypothesis_set_that_needs_to_be_revised>,
    "rationale": "Brief explanation of why this hypothesis set needs to be revised based on
the conversation. This text can be as long as necessary and will automatically wrap beautifully
to the next line without breaking the box layout.",
    "operation": "[REPLACE/RERANK]",
    "salient_appraisals": [
        "dimension_1",
        "dimension_2",
        "dimension_3",
        "dimension_4",
        "dimension_5"
    ]
}
```

If you decide that no hypothesis set needs to be revised, you must output only the string "No revision needed" and do not output any JSON or comments.

Your Response

Figure A18: Prompt for cognitive appraisals revision based on the previous mental states.

Cognitive Appraisals Scoring *BIP Inference*

SYSTEM INSTRUCTION
You are an expert clinical psychologist and empathetic listener specializing in Cognitive Appraisal Theory. Your task is to pick 5 cognitive appraisals that are the most impactful to the patient's emotional reaction to the situation described in the dialogue. Your answer should be based on the patient's persona, their prior BDI, the dialogue history, and their current utterance. You should output only a list of 5 cognitive appraisals in alphabetical order and without duplicates.

INPUT DATA
Patient Persona:
{{ persona_profile }}

Prior BDI:
{{ bdi }}

Dialogue History:
{{ dialogue_history }}

Current Patient Utterance:
{{ current_utterance }}

STRICT OUTPUT CONSTRAINTS

1. Identify the patient's top-5 most salient unique cognitive appraisals in alphabetical order and without duplicates from the following list:
    - **suddenness:** The situation was sudden or abrupt.
    - **unfamiliarity:** The situation is unfamiliar.
    - **predictability_of_event:** The occurrence of the situation is predictable.
    - **unpredictability_of_event:** The occurrence of the situation is unpredictable.
    - **goal_incongruence:** The outcome is not aligned with personal goals/desires.

- **predictability_of_consequences:** Consequences of the situation are easily anticipated.
- **unpredictability_of_consequences:** Consequences of the situation are not anticipated.
- **self_cause:** The root cause of the situation is the character themselves.
- **other_cause:** The root cause of the situation is someone other than the character themselves.
- **self_control:** The character has control over what is going on in the situation.
- **other_control:** Someone other than the character has control over what is going on in the situation.
- **norm_violation:** The actions that produced the situation violated the character's standards and ideals.
- **coping_demand:** The situation requires a great deal of energy to cope with it.
- **unacceptable_consequences:** Consequences of the situation are unbearable.

2. Output nothing else but the list of 5 distinct cognitive appraisals in alphabetical order without any other text or comments. The list should be formatted as follows:

```
[
    "appraisal_1",
    "appraisal_2",
    "appraisal_3",
    "appraisal_4",
    "appraisal_5"
]
```

RESPONSE
`output_template_all_tokens: |`

Figure A19: Prompt for cognitive appraisals hypothesis scoring.

Conditional Utterance Scoring *BIP Inference*

SYSTEM INSTRUCTION
You are simulating a patient in a dialogue. Your goal is to generate the most likely next verbal response based strictly on the dialogue history and their perception of the event.

TASK
Generate the patient's next verbal utterance wrapped in double quotes.

- The utterance must be a natural consequence of the **Internal State**.
- Pay specific attention to the **Cognitive Appraisals**. The utterance must be a natural consequence of the **Cognitive Appraisals**.
- Do not output thoughts or meta-data.
- Output ONLY the raw text of the utterance wrapped in double quotes with the following format: `"Patient's next verbal utterance"`

INPUT DATA

Dialogue History:
{{ dialogue_history }}

- **Patient's Perception of the Event**
  {{ cogapp }}

RESPONSE

Figure A20: Prompt for utterance scoring conditioning on the given cognitive appraisal hypothesis.

Judge Agent Instruction *BIP Inference*

You are a strict and careful annotation judge for cognitive appraisals in therapeutic dialogue.
You will be given:

1. A snippet of a therapeutic dialogue (patient + therapist)
2. A list of 14 appraisal dimensions with definitions
3. Four candidate appraisal hypotheses (each is a set of dimensions)

Your tasks:

A) Decide whether ANY of the four hypotheses is fully correct for the dialogue snippet.

B) If none is fully correct, choose the MOST LIKELY hypothesis (the closest one) and revise it to produce the best final appraisal set.

Important rules:

1. Base your judgment ONLY on the dialogue snippet and the provided definitions.
2. Do NOT default to "norm_violation" or "self_cause" just because the client feels guilty/shame. Guilt does not translate to norm violation; self-blame language does not translate to actual self-cause unless the snippet indicates the client caused the event.
3. Avoid "kitchen sink" sets. Only include a dimension if there is explicit evidence in the text.
4. Enforce mutual exclusivity:
   - Do NOT include both `predictability_of_event` and `unpredictability_of_event`.
   - Do NOT include both `predictability_of_consequences` and `unpredictability_of_consequences`.
   - Do NOT include both `self_cause` and `other_cause`.
   - Do NOT include both `self_control` and `other_control`.
5. Output exactly FIVE (5) dimensions in the final answer, unless the label set is explicitly allowed to be smaller in the definitions. If fewer than 5 are clearly supported, pick the 5 most supported and relevant.

Procedure (follow these steps):

- **Step 1)** Briefly summarize (1–2 sentences) what happened and what the client is appraising.
- **Step 2)** For each hypothesis (H1–H4), score it with:
  - Supported labels count (0–5)
  - Incorrect/contradicted labels count (0–5)
  - Missing but clearly needed labels (0–5)

  Give a short reason (1–2 lines) focusing on the biggest mismatch.

- **Step 3**) Determine:
  - If any hypothesis is fully correct → select it as FINAL with no edits.
  - Else pick the closest hypothesis and revise it.
- **Step 4**) Optional Revision:
  - Only if the closest hypothesis is not fully correct.
  - Remove labels that are unsupported or contradicted.
  - Add missing labels that are clearly supported.
  - Ensure mutual exclusivity rules.
  - Return the final set of exactly 5 labels.

Output format (STRICT):
- decision: one of ["H1", "H2", "H3", "H4", "REVISED"]
- selected_hypothesis: one of ["H1", "H2", "H3", "H4"] (the closest one, even if revised)
- final_appraisals: a JSON list of 5 strings, each must be one of the 14 dimension names
- evidence: a JSON object mapping each final label → a short quote (3–15 words) from the dialogue that supports it
- notes: one short paragraph explaining why this set is best

[DIALOGUE SNIPPET]
{{ dialogue }}

[APPRAISAL DIMENSIONS (14) + DEFINITIONS]
{{ cogapp_definitions }}

[HYPOTHESES]
H1: {{ hypothesis_1 }}
H2: {{ hypothesis_2 }}
H3: {{ hypothesis_3 }}
H4: {{ hypothesis_4 }}

[YOUR RESPONSE]
Please output your response in the following JSON format:

```
{
    "decision": "ONE OF ['H1', 'H2', 'H3', 'H4', 'REVISED']",
    "selected_hypothesis": "ONE OF ['H1', 'H2', 'H3', 'H4']",
    "evidence": {
        "dimension_1": "quote_1",
        "dimension_2": "quote_2",
        "dimension_3": "quote_3",
        "dimension_4": "quote_4",
        "dimension_5": "quote_5"
    },
    "notes": "short paragraph explaining why this set is best. This text can be as long as
necessary and will wrap gracefully across multiple lines without breaking the layout of the
tcolorbox.",
    "final_appraisals": [
        "dimension_1",
        "dimension_2",
        "dimension_3",
        "dimension_4",
        "dimension_5"
    ]
}
```

Figure A21: Prompt for getting the final salient cognitive appraisals prediction using the Judge Agent.