Empathy Applicability Modeling for General Health Queries

Anonymous ACL submission

Abstract

LLMs are increasingly being integrated into clinical workflows, yet they often lack clinical empathy, an essential aspect of effective doctor–patient communication. Existing NLP frameworks focus on reactively labeling empathy in doctors' responses but offer limited support for anticipatory modeling of empathy needs, especially in general health queries. We introduce the Empathy Applicability Framework (EAF), a theory-driven approach that classifies patient queries in terms of the applicability of emotional reactions and interpretations, based on clinical, contextual, and linguistic cues. We release a benchmark of real patient queries, dual-annotated by Humans and GPT-4o. In the subset with human consensus, we also observe substantial human–GPT alignment. To validate EAF, we train classifiers on human-labeled and GPT-only annotations to predict empathy applicability, achieving strong performance and outperforming the heuristic and zero-shot LLM baselines. Error analysis highlights persistent challenges: implicit distress, clinical-severity ambiguity, and contextual hardship, underscoring the need for multi-annotator modeling, clinician-in-the-loop calibration, and culturally diverse annotation. EAF provides a framework for identifying empathy needs *before* response generation, establishes a benchmark for anticipatory empathy modeling, and enables supporting empathetic communication in asynchronous healthcare.

1 Introduction

Clinical empathy integrates cognitive (understanding), emotional (resonating), and action-oriented (expressing) components (Guidi and Traversa, 2021). It is indispensable for clinical care, deepening therapeutic relationships and improving outcomes such as patient satisfaction, care effectiveness, reduced distress, and hospital length of stay (Guidi and Traversa, 2021; Olson, 1995; Hoffstädt et al., 2020); yet clinicians miss 70-90% of empathic opportunities during patient interactions (Morse et al., 2008; Hsu et al., 2012).

Large Language Models (LLMs) are increasingly integrated into healthcare workflows and patient interactions, with major EHR vendors such as EPIC adopting them for clinical messaging and nearly half of physicians reporting patients consult ChatGPT before visits (Antoniak et al., 2024; Sermo Team, 2025). While these trends highlight rapid adoption of LLMs in healthcare, they also raise concerns of lacking empathy crucial for asynchronous physician-patient interactions (Koranteng et al., 2023). However, effective empathy requires discernment, not just fluency. This highlights a critical, antecedent challenge: How can we systematically model the applicability of empathy, allowing systems to recognize the specific clinical and linguistic cues that necessitate an emotional response?

Modeling empathy in text is inherently difficult without non-verbal cues, and NLP research has historically over-weighted emotional aspects while overlooking cognitive empathy (Lahnala et al., 2022), even though both are vital in clinical care. To redress this imbalance, EPITOME (Sharma et al., 2020) captures the multidimensionality of empathy through emotional reactions, interpretations, and explorations, offering an empathetic lens on warmth, understanding, and curiosity in mental health support. Online Empathy (Chai et al., 2019) also addresses multidimensionality, classifying responses as Informational and Emotional. However, both EPITOME and Online Empathy assess empathy post hoc, labeling support-giver responses after they appear and thus offering no guidance while a clinician is composing a response to the patient query.

Lahnala et al. attempt to solve this particular problem with with an Appraisal Framework that annotates empathic opportunities and clinician elicitation and response as functions of (affect | judgment

| appreciation) in breaking-bad-news oncology dialogues (Lahnala et al., 2024). This discourse analysis lens excels at teaching stance shifts over multi-turn synchronous conversations, yet is not suited to single-turn, asynchronous general health queries: it classifies stance, not what the patient needs (cognitive clarification vs emotional warmth). Thus, it remains need-blind, providing little actionable help for one-off replies.

To address this gap, we propose the Empathy Applicability Framework (EAF), a theoretically grounded method to proactively identify when and what type of clinical empathy should be expressed in response to patient queries. EAF operationalizes empathy along two key dimensions: affective (emotional reactions) and cognitive (interpretations) – labeling each as *Applicable* or *Not Applicable* based on clinical, contextual, and linguistic cues within patient queries. Unlike prior work that evaluates empathy *in the response itself* (Chai et al., 2019; Sharma et al., 2020), EAF analyzes the patient's query before any response, enabling anticipatory reasoning. Evidence for the value of anticipation comes from Sibyl (Wang et al., 2025), which shows that anticipating user's emotional and contextual trajectory improves empathetic response generation. EAF can also mitigate limitations of reactive frameworks that overrely on surface lexical cues. For instance, EPITOME misclassified generic phrases as empathetic in nonsensical contexts, with false-positive rates exceeding 95% (Lee et al., 2023). By shifting empathy assessment from *post-hoc* response scoring to *pre-response* query analysis, EAF is methodologically significant: it enables anticipatory inference of emotional and interpretive support needs, helping providers and LLMs detect and act on empathic opportunities in general health queries.

We make three primary contributions: (i) **Framework Design**: we introduce and theoretically ground the EAF in clinical empathy literature, clearly differentiating our anticipatory model from prior post-hoc approaches; (ii) **Annotated and analyzed Benchmark**: a novel dataset of 1,300 patient queries annotated by humans and GPT-4o (included in the supplementary materials), demonstrating EAF's reliability and interpretability; and (iii) **Operationalization Challenges**: we identify and systematically analyze specific contexts where anticipatory empathy annotations diverge, highlighting opportunities for future research in multi-annotator modeling, clinician-in-the-loop systems, and culturally sensitive annotation strategies.

2 Empathy Applicability Framework and Theoretical Grounding

The EAF identifies empathetic needs proactively by assessing patient queries along two dimensions adapted from EPITOME (Sharma et al., 2020) and informed by Chai et al.'s distinction between emotional and informational support (Chai et al., 2019): *Emotional Reactions* and *Interpretations*. Table 1 summarizes the EAF, detailing applicable and non-applicable cues for each dimension.

To develop EAF, we performed inductive thematic coding on 300 randomly selected patient queries from the HealthcareMagic and iCliniq datasets (Li et al., 2023), identified themes, formed subcategories (cues), and finally, we iteratively refined to comprehensively and distinctly capture empathy applicability.

Additionally, we ground EAF cues in Patient-Centred Care (PCC), a widely used framework for clinician–patient communication, focusing on the PCC functions of responding to emotions and managing uncertainty to ensure clinically grounded expressions of empathy (Epstein and Street Jr, 2007; McCormack et al., 2011).

3 Methods

To determine whether EAF is reliably interpretable across a range of clinical queries and to identify any systematic challenges, we curated a diverse dataset of health-related queries and annotated them using the EAF, employing both human annotators and an LLM. To assess whether these annotations exhibit learnable patterns, indicating the internal consistency of EAF, we trained classifiers on the EAF-labeled data. The following subsections detail the annotation and modeling procedures.

3.1 Data Source

We sampled 9,500 patient queries from two publicly available datasets (HealthCareMagic and iCliniq) released by Li et al. (Li et al., 2023). We sampled 4,750 queries each from HealthCareMagic ($\approx 100k$ dialogues) and iCliniq ($\approx 10k$), to maximize linguistic and contextual diversity and avoid overfitting to a single source. As these datasets are publicly available and anonymized, our IRB determined that this study was exempt from human subjects review. The datasets do not carry an explicit license; therefore, we use them exclusively

Dimensions	Applicable cues	Not Applicable cues
Emotional Reactions Expressions of warmth, compassion, concern, or similar feelings conveyed by a doctor in response to a patient's query.	• **Severe Negative Emotion** • **Inferred Negative State** • **Seriousness of Symptoms** • **Concern for Relations** *Rationale:* Signals reflect distinct pathways of emotional distress, guiding when emotional reactions are warranted.	• **Routine Health Management** • **Purely Factual Medical Queries** • **Neutral Symptom Descriptions** • **Hypothetical Queries** *Rationale:* Signals no emotional content; omit reactions to maintain factual medical focus.
Interpretations Communication of an understanding of the patient's feelings (expressed or implied) and/or experiences (including contextual factors) inferred from the patient's query.	• **Expression of Feeling** • **Experiences or Context Affecting Emotional State** • **Symptoms with an Emotional Impact** • **Distressing Uncertainty About Health** *Rationale:* Signals lived burden, context, or uncertainty requiring interpretive acknowledgment.	• **Emotional-Reactions N/A cues +:** with absence of distressing contextual or experiential details. *Rationale:* Signals absence of both emotional and contextual cues, preventing over-empathizing and maintaining focus on informational needs.

Table 1: Empathy Applicability Framework (EAF). Each dimension lists cues for when an empathic dimension is *Applicable* or *Not Applicable*; Brief rationales explaining what each cue set captures follow the cues. Detailed description of the EAF and its cues with examples is provided in the Appendix A. Also, see Appendix Table 4 for concrete query scenarios illustrating cues usage and EAF operationalization.

for non-commercial research, in line with the authors' stated intention and public availability, and will release our de-identified EAF benchmark under the same non-commercial terms. To balance rigor and cost, 1,500 of the queries were earmarked for dual annotation by humans and GPT-4o to support reliability and error analyses, while the remaining 8,000 were annotated only by GPT-4o for predictive validity testing.

3.2 Annotation Task

The annotation task required using EAF to label patient queries as applicable or not applicable (see Table 1) on two dimensions of empathy: Emotional Reactions (EA) and Interpretations (IA). Human annotators were instructed to identify at least one best-fitting subcategory per dimension to justify their labels (they mostly selected a single best-fitting subcategory). The GPT annotations listed all relevant subcategories supporting labeling decisions.

3.2.1 Annotator Recruitment, Training and Calibration

Due to empathy annotation subjectivity, we prioritized consistency by avoiding crowdsourcing and instead recruited and trained two annotators from Pakistan with high English proficiency: HA1, a female with an MS in Linguistics, and HA2, a male with a BS in Computer Science. We recruited two annotators via departmental channels for about a one-month engagement. Informed consent to use the annotated dataset to train large language models was collected from the annotators prior to the start of the annotation process. Annotators were compensated US$360 (equivalent to a local monthly research salary). Annotators underwent three-stage training on 200 queries (50 + 50 + 100) from a subset of 1,500, with convergence meetings after each stage to clarify misunderstandings and align labeling. Training queries were excluded from later experiments. Annotators then independently labeled the remaining 1,300 queries following procedures in Section 3.2. Annotation instructions are detailed in Appendix B.

We *intentionally* employed lay annotators to capture the patient's perspective. Prior research shows that empathy is 'in the eye of the beholder' (Bernardo et al., 2018), and given that the empathy levels in the clinician's response will be perceived by the patient, lay annotators whose judgments reflect the patient/recipient experience are *better-suited* for this task. Additionally, prior studies show that clinicians often overlook empathic opportunities in favor of diagnostic focus (Hsu et al., 2012) and that patients' ratings of clinicians' empathy often diverge from clinicians' assessments (Bernardo et al., 2018; Hermans et al., 2018).

3.2.2 GPT Annotations

To scale the data set and enable comparison with human annotations, we used GPT-4o via the OpenAI API, prompted to act as an expert annotator using contrastive prompting (Gao and Das, 2024). The model was given definitions of EA and IA, subcategory descriptions with examples, and labels indicating whether each subcategory was Applicable or Not Applicable. Then it returned the matching subcategories, with the format inherently indicating the applicability class (annotation scripts included in the supplementary software). Complete prompt specifications, including with and without our framework are included in Appendix F.

For the 1,300 human-annotated queries, GPT-4o generated five annotation passes per query, with final labels determined by majority vote[1]. For the remaining 8,000 queries, a single-pass annotation was used due to cost constraints. This yielded two subsets: 1,300 queries labeled by both humans and GPT (with majority-voted GPT labels) and 8,000 labeled solely by GPT (single-pass annotation). **Note:** Throughout the remainder of this text, all references to GPT refer specifically to GPT-4o.

3.3 Modeling Task and Approach

We frame empathy applicability prediction as two independent binary classification tasks. Given a patient query P_i, the objective is to predict, for each empathy dimension $d \in \{\text{EA}, \text{IA}\}$, whether that dimension is *Applicable* (1) or *Not Applicable* (0), denoted A_{id}. For each dimension, we fine-tune a distinct RoBERTa-based classifier (Liu et al., 2019). Full architectural details, including the attention mechanism, the pooling operation, and the model diagram, are provided in the appendix E.

4 Evaluation Setup and Experiments

This section details the evaluation setup and model training configurations used in our experiments.

4.1 Annotator Agreement

We assessed human annotation reliability using raw agreement and Cohen's Kappa across the 1,300 independently labeled queries. For GPT-generated annotations, we compared majority-voted GPT labels with a subset of human-annotated queries: queries where both human annotators reached an agreement. This allows us to evaluate GPT performance without confounding disagreement over error or subjectivity.

4.2 Conceptual Alignment

To examine whether humans and GPT rely on similar rationales, we performed an UpSet plot analysis (Figure 1). This analysis was limited to queries where humans and GPT agreed on the overall applicability label, allowing us to assess alignment in subcategory reasoning rather than outcome. A match is coded as *Full* if GPT includes both subcategories selected by the two human annotators, *Partial* if GPT's subcategories overlap with only one human's subcategory label and *No match* if GPT matches neither human subcategory.

4.3 Divergence Bar and Qualitative Analysis

Given the subjective nature of empathy, we analyze mismatches as directional divergences rather than strict errors. To characterize disagreement, we use three-way divergence bars (Figure 2) that decompose label mismatches within each subcategory into *Annotator Spread* (one human labeled Applicable, the other Not), *LLM-Adds* (GPT labeled Applicable, humans Not), and *LLM-Omits* (GPT labeled Not, humans Applicable). Furthermore, we performed qualitative analysis on a subset of queries where GPT labeled differently, and identified thematic patterns that highlight the different labeling.

4.4 Model Evaluation

We evaluated the performance of the classifiers trained to predict empathy applicability (Applicable vs. Not Applicable). Reported metrics include accuracy, weighted F1 score, and macro-averaged F1 score across both dimensions (EA and IA). To contextualize classifier performance, we compared results against four baselines: Random Guessing (assigns labels at random), Always Applicable, Always Not Applicable, and `o1`-Zero-Shot (based on OpenAI's reasoning model, without invoking empathy applicability framework). For the `o1` baseline, we provide only the definition of the target dimension (EA or IA) and prompt it to classify each patient query as 'Applicable' or 'Not Applicable', preserving the zero-shot setting without framework cues. These baselines help determine whether our trained models learn meaningful patterns beyond simple heuristics or zero-shot LLM reasoning. In

[1]Majority voting ensured consistency across passes. More than 94% of queries received the same label on the first pass and as the majority vote for both empathy dimensions, indicating minimal divergence. Hence, we report evaluation metrics only with the majority-voted labels.

Dimension	Human–Human κ (agree / disagree)	Human–GPT κ (agree/ disagree)
EA	0.521 (981 / 315)	0.617 (668 / 152)
IA	0.404 (898 / 398)	0.652 (678 / 142)

Table 2: Cohen's κ with agreement counts: human–human agreement on the full set and human–GPT alignment on the human-consensus subset.

addition, following best practices for sanity checking and overfitting control, we include *classical* text classifiers trained and evaluated on the Human Set (Section 4.5): Logistic Regression (LR) and Linear SVM with TF–IDF features(1–3 grams). These linear models provide a transparent reference for what is learnable from local lexical features.

4.5 Model Training and Training Sets

Each classifier for the EA and IA tasks is based on RoBERTa-base (≈125 M parameters) and was trained on two distinct datasets (data and scripts included in the supplementary material): **Human Set:** Contains only queries where both human annotators reach consensus on a label for a given dimension, serving as a high-fidelity benchmark aligned with human judgment. **Autonomous Set:** Consists of GPT-labeled data from the 8,000-query pool, with no human supervision. This tests whether models trained solely on GPT output can approximate human consensus.

For the Human Set, we split the data into subsets of training (75%), validation (5%), and test (20%). For the Autonomous Set, training was done entirely on GPT-labeled data, but testing used the same human-consensus test set as the Human Set to enable consistent evaluation relative to human agreement. Training used a single NVIDIA A40 GPU per run. A Human-Set run finished in ≈15 min GPU time, while an Autonomous-Set run took ≈40 min; thus the total compute budget per dimension is <1 GPU-hour. All models were trained for 10 epochs using a learning rate of 2×10^{-5} and a batch size of 8. To ensure comparability, all models shared the same architecture and hyperparameters.

5 Results

In this section, we present our findings related to the reliability of the EAF and the challenges in operationalizing it. Additional dataset characterization beyond agreement and modeling results is provided in Appendix G.

5.1 Reliability of the EAF

We evaluated reliability along three axes of Consistency, Predictive validity and Conceptual alignment.

Consistency. We first assess agreement on Applicable/Not Applicable labeling between human annotators across 1,300 queries, and between GPT-4o and the *human consensus* on a subset of 820[2] queries. As shown in Table 2, human annotators achieved moderate agreement on both empathy dimensions, with an overall Cohen's κ of 0.46. This falls within the typical range for empathy annotation tasks; for example, Sibyl (Wang et al., 2025) reported scores between 0.4 and 0.6. Notably, agreements outnumbered disagreements *by a factor of two to three*, suggesting that the EAF supports relatively consistent human labeling despite the inherent subjectivity of empathy.

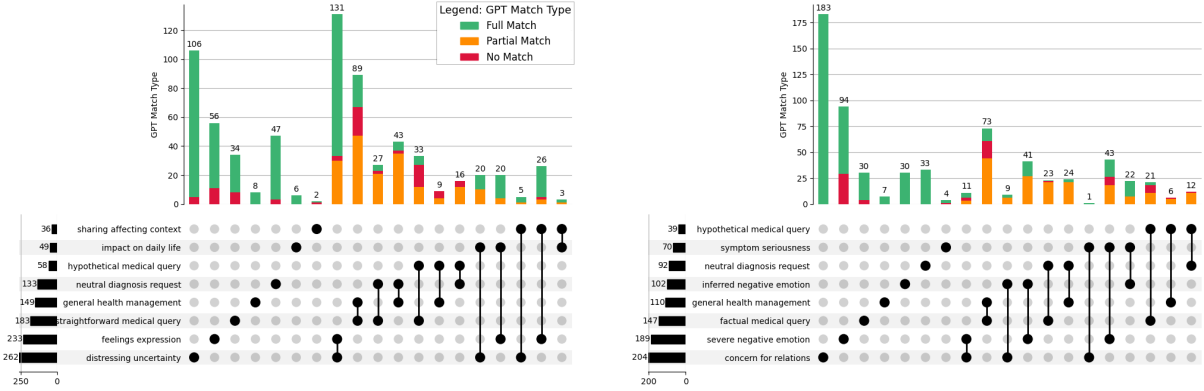
GPT aligned well with the *human consensus dataset*, queries where both humans agreed, achieving three-way agreement. For both EA and IA, Cohen's κ exceeded 0.6 and raw agreement was about 80% (Table 2). These results reflect agreement on human-aligned cases, demonstrating EAF's effectiveness in guiding GPT to anticipate empathy applicability in clearer contexts, excluding more ambiguous or complex queries (see section 5.2).

Predictive Validity. We next evaluated whether EAF annotations are machine-learnable. As shown in Table 3, classifiers trained on human consensus data achieved high performance: LR attains **0.84** Macro-F1 for EA and **0.80** for IA (SVM: 0.83/0.77), establishing a strong classical reference. Our transformer (RoBERTa-base) exceeds these values significantly (**for EA: (0.92 vs. 0.84); and for IA (0.87 vs. 0.80)**). Models trained on GPT-only annotations (the Autonomous set) also performed well, achieving around **0.85** for EA and **0.77** for IA on the same held-out human-consensus test set, reflecting expected loss from noisier labels or differences from human labeling. Our models also significantly outperformed the trivial baselines (random guessing, always applicable, always not applicable, and o1 Zero-Shot), which yielded substantially lower scores. McNemar's test (McNemar, 1947) confirmed statistical significance of the transformers over the trivial baselines (max $p < 10^{-4}$)

[2]For the full set, Appendix Table 5 shows comparable GPT agreement with each human annotator on affective EA, but substantially more variable agreement on cognitive IA.

Table 3: Classification results across training sets and baselines (single run on the human-consensus test set). Bold indicates best performance. Classical baselines (TF–IDF+LR/SVM) are trained on the Human Set only. Our transformer significantly outperforms all the baselines.

Training Set / Model	**EA**			**IA**		
	Acc	Macro-F1	Wtd-F1	Acc	Macro-F1	Wtd-F1
Random	0.47	0.47	0.47	0.44	0.43	0.44
Always Applicable	0.52	0.34	0.36	0.53	0.35	0.37
Always Not Applicable	0.48	0.32	0.31	0.47	0.32	0.30
o1 Zero-Shot	0.55	0.40	0.41	0.62	0.53	0.54
Human-supervised models (train and tested on human-consensus set)						
Logistic Regression	0.84	0.84	0.84	0.80	0.80	0.80
Linear SVM	0.83	0.83	0.83	0.77	0.77	0.77
Transformer (RoBERTa-base)	**0.92**	**0.92**	**0.92**	**0.87**	**0.87**	**0.87**
Autonomous-supervised model (train on GPT labels, test on human-consensus test set)						
Transformer (RoBERTa-base)	0.85	0.85	0.85	0.78	0.77	0.77



(a) Interpretation Applicability (IA) subcategory matches

(b) Emotional Applicability (EA) subcategory matches

Figure 1: UpSet plots comparing GPT and human rationales for **(a)** IA and **(b)** EA subcategories. For each query, each human annotator selects one best-fit subcategory for their rationale; thus the human set is either a *single-dot* combination (both humans chose the same subcategory) or a *two-dot* combination (humans chose different subcategories). Horizontal bars show how often each subcategory appears in the human annotation set across all queries. Each vertical bar shows the frequency of a unique human-combination and is split by GPT agreement: Full (GPT's subcategory set covers the entire human set), Partial (GPT matches only one of the two human subcategories), and No match (GPT matches neither human subcategory).

and over the classical baselines (max $p \leq 0.02$). Taken together, strong linear performance indicates consistent linguistic realizations of the constructs, while the transformer's margin suggests benefits from broader context rather than overfitting. Overall, these results show that EAF-labeled data encode structured and learnable patterns.

Conceptual Alignment. We further examined whether humans and GPT rely on similar reasoning when assigning EAF labels. UpSet plot analysis (Figure 1) shows strong conceptual alignment. In many cases, both human annotators independently selected the same subcategory and GPT matched it, especially for both applicability and non-applicability cues such as *Severe Emotion* or *Factual Queries*. These matches indicate that the EAF defines meaningful categories that are consistently identifiable by both humans and LLMs.

When annotators selected different subcategories for the same label, GPT often matched both. For example, in queries involving both *Expression of Feeling* and *Distressing Uncertainty*, GPT cited both reasons, suggesting that GPT can reconcile diverse human rationales and underscores the framework's breadth in conceptualizing clinical empathy. No-match cases are rare, and GPT typically overlaps with at least one human subcategory. Appendix G.4 quantifies and corroborates these trends, showing match rates (match vs. miss, conditioned on humans using that subcategory) above 80% for most subcategories.

Collectively, these results establish that EAF supports consistent human judgments, yields learnable patterns, and promotes interpretable reasoning

across both humans and LLMs, making it well suited for anticipatory empathy modeling in clinical settings.

5.2 Systematic Challenges in Operationalizing Anticipatory Empathy

Divergence bar analysis (Section 4) revealed that inter-human agreement is significantly lower for interpretations (IA) than for Emotional Reactions (EA) (Table 2), and that despite moderate overall human-GPT agreement (Table 2), there is divergence at the subcategory level. Subsequent qualitative analysis revealed three key challenges in applying the EAF, with implications for any clinical empathy framework in NLP.

5.2.1 Challenge 1: Subjectivity in Identifying Implied Distress

The categories *Inferred Negative State* (EA) and *Distressing Uncertainty* (IA) show substantial divergence in inter-human and human-GPT annotations (Figure 2).

A qualitative review of 50 randomly selected cases[3] (25 each for Distressing Uncertainty and Inferred Negative State)[4] by the first author acting as adjudicator revealed that in more than 50% of the queries, one could reasonably infer implied emotional distress *or* determine that the query is driven by factual intent. For instance, the female annotator labeled a pain-and-menstrual-cycle query as Distressing Uncertainty, while the male annotator treated it as a factual diagnostic request, illustrating the subjectivity of distress inference.

5.2.2 Challenge 2: Clinical-Severity Ambiguity

In the category *Serious Symptoms* (EA), GPT labeled 100 queries as requiring emotional reactions when humans did not (Figure 2). Qualitative analysis of 25 randomly selected cases[3] where only GPT had labeled empathy applicability revealed three patterns: (1) In 40% of the cases, GPT appropriately flagged empathy needed for patients with chronic or life-threatening conditions (e.g., post-liver transplant complications) that human annotators with no medical background had overlooked (2) borderline cases with reasonable disagreement (16%), such as prolonged low-grade fever after kidney stones, and (3) GPT overgeneralization of vivid but non-serious pain symptoms (44%) that did not meet the EAF criteria of chronic or life-threatening severity (for example, lip numbness after dental problems).

5.2.3 Challenge 3: Contextual Hardship

GPT frequently over-applied *Symptoms Emotional Impact* (SEI) and *Context Sharing* (CS) tags compared to humans (Figure 2). An analysis of 25 randomly selected[3] mismatched labels in SEI category, and all 20 mismatches in CS revealed that while GPT sometimes correctly identified complex distress signals humans missed (20-25% of the cases), it more often (75-80% of the cases) equated physical discomfort with emotional distress – potentially reflecting Western-centric training biases (Johnson et al., 2022; Cao et al., 2023).

These challenges, rooted in subjective inference, clinical ambiguity, and cultural variation, highlight the complexity of implementing clinical empathy. Addressing them requires moving beyond single-annotator consensus toward frameworks that embrace interpretive pluralism, clinical expertise, and cultural sensitivity.

6 Discussion and Conclusion

Asynchronous patient communication requires *anticipatory* mechanisms that can signal empathic needs *before* a response is written. EAF addresses this gap by assigning applicability labels to patient queries, indicating whether empathy is warranted and which dimension (emotional versus interpretive) should be expressed. This framing complements recent advances in empathetic response generation, which enrich responses by reasoning about emotional causes, rhetorical framing, and likely future trajectories. For example, Lee et al. integrate figurative language and empathy-cause signals to generate responses that are both emotionally engaging and contextually grounded. Complementarily, Chen et al. use cause-aware chain-of-thought prompting: they extract the emotional cause from the dialogue and insert generated commonsense relations (xEffect, xReact, xIntent, xNeed, xWant) conditioned on the extracted cause into the prompt, which significantly improves empathy in human evaluations (Chen et al., 2024). Likewise, Sibyl (Wang et al., 2025) helps models anticipate the seeker's likely next move and improves empathetic generation. However, these systems are de-

[3] *Detailed patient queries, mis-aligned labels, and qualitative interpretations are included in the supplementary material as the* `misalignment_analysis.csv` *file*

[4] *a sample size consistent with prior clinical-NLP error analyses; (Hu et al., 2024)*

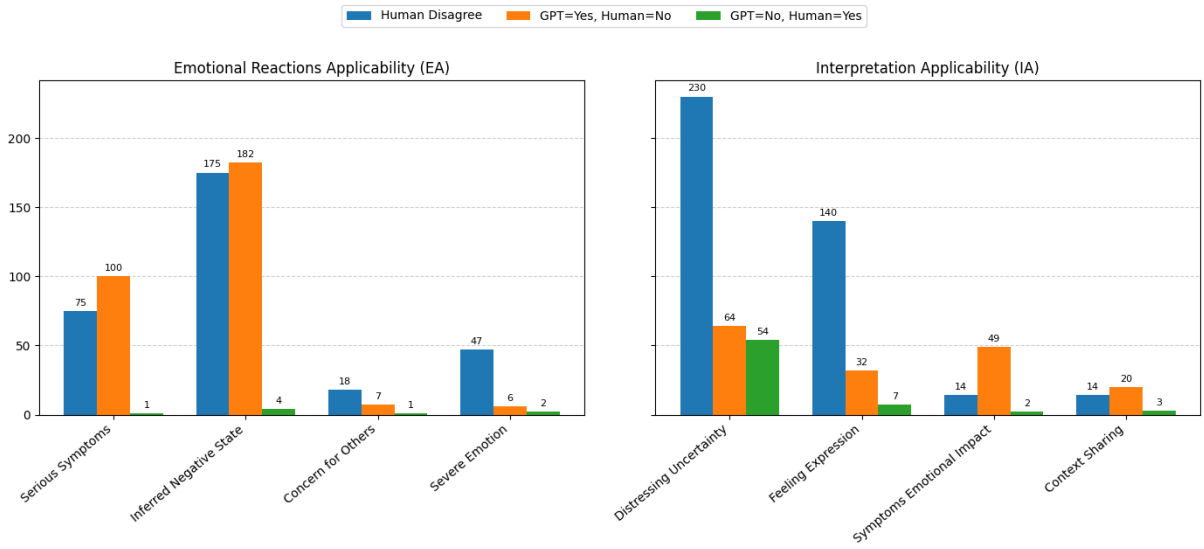


Figure 2: Three-way divergence for every subcategory. Orange = *Annotator Spread in Humans* (One Applicable, other not); Blue = *LLM-Adds Empathy Dimension* (GPT Applicable, Humans Not); Green = *LLM-Omits Empathy Dimension* (GPT Not, Humans Applicable).

signed primarily for open-domain emotional support and often assume the user's emotional state is explicit or readily inferable. In contrast, general health queries frequently contain ambiguous, subtle, or medically grounded cues about whether emotional reactions or interpretive support are appropriate. Our anticipatory assessment framework identifies the *empathetic needs embedded in the patient's query* prior to generation, and can be integrated with cause-aware, figurative-language, and commonsense-augmented methods to guide empathetic response generation in clinical and general-health settings.

However, EAF faces challenges from **subjective inference**, particularly when cues are implicit (e.g., *Inferred Negative Emotional State*). As appraisal theory suggests, divergent interpretations of distress often reflect genuine ambiguity rather than noise (Wondra and Ellsworth, 2015). Furthermore, these appraisals are shaped by **cultural priors of emotion**. Eichbaum et al. (2023) warn that Western-centric empathy models can misfire cross-culturally. Indeed, GPT-4o—trained on predominantly Western data (Johnson et al., 2022)—often labeled minor inconveniences as empathy-worthy where our South Asian annotators did not. This highlights a **cultural bias** inherent to LLMs that encode American norms (Cao et al., 2023).

The NLP community increasingly embraces this variability through multi-annotator models and annotator-aware representations that yield calibrated uncertainty estimates and capture interpretive styles (Davani et al., 2022; Mokhberian et al., 2023). Gordon et al.'s (Gordon et al., 2021) *jury learning* further shows how selecting annotator subsets aligned with demographic perspectives can preserve pluralism (Gordon et al., 2021). In clinical empathy contexts, retaining subjective variability can help anticipate diverse patient needs, including domain-specific perspectives (e.g., oncologists who prioritize emotional support as central to care) (Dekker et al., 2020). Building on this, we advocate extending disagreement-aware and annotator-aware frameworks toward **diversity-aware modeling** that explicitly accounts for culturally patterned differences in what counts as an empathy need, rather than collapsing them into a single consensus label.

This work makes three contributions to clinical empathy in NLP. First, we introduce the Empathy Applicability Framework (EAF), shifting from reactive to anticipatory modeling. Second, we establish a benchmark of 1,300 patient queries demonstrating reliable EAF labels. Third, our analysis identifies challenges, namely subjective inference, clinical-severity ambiguity, and contextual hardship, as opportunities to embrace interpretive pluralism via multi-annotator frameworks. By combining a practical framework with empirical operationalization, this work advances empathy modeling that respects interpretive complexity while remaining computationally tractable.

7 Limitations

Our study faces three key constraints, the first two mirroring limitations reported by Ali et al. (2025). First, we relied on only two human annotators, neither of whom had clinical training, which limited the range of perspectives represented; expanding the size, clinical expertise, and cultural diversity of the annotator pool would better capture the variability of empathy judgments. Second, all automatic annotations were produced with GPT-4o—selected for its widespread availability through ChatGPT—but this exclusive focus on the GPT series limits the generalization of our findings to other model architectures (e.g., Gemini, Claude, GPT reasoning models, or open-source alternatives). Third, human annotators selected a single most-salient subcategory per dimension, while GPT-4o returned multiple subcategories; this procedural mismatch hinders direct comparison of disagreement patterns, and aligning the guidelines would allow for more rigorous evaluation. Future work should therefore involve a more diverse set of human annotators, evaluate multiple LLM families trained under different specifications, and standardize annotation procedures between humans and models to obtain broader insights for improving empathy modeling in NLP for clinical contexts.

8 Ethical considerations

We developed the EAF to augment not replace clinician empathy judgments. Deploying EAF therefore requires close attention to several intertwined ethical risks that must be mitigated through thoughtful design and implementation.

A primary concern is the moral and social impact of artificial empathy. Because LLMs lack authentic emotional experience, we must ask whether the 'applicable emotional reactions' they generate can truly convey warmth or connection. If users perceive these reactions as hollow or manipulative, an *uncanny valley* effect could ensue, in which attempted comfort backfires by appearing inauthentic. Determining *whether, when,* and *how* automated empathy should be implemented, and addressing potential deception or user discomfort, requires a systematic study of user perceptions of authenticity versus artificiality.

A second mirror image danger arises from the same gap between simulated language and genuine feeling. As *Empathic AI Can't Get Under the Skin* discussed, LLMs lack the biological and psychological underpinnings that ground human empathy, yet their empathic language can evoke real emotional responses (Nature Machine Intelligence, 2024). Kirk et al. warn that users may form perceived emotional bonds with such systems, risking unhealthy attachment or disclosure of sensitive information (Nature Machine Intelligence, 2024). Thus, rejection born of perceived inauthenticity and devotion born of mistaken authenticity are twin failure modes rooted in the same ontological limitation.

For these reasons, we insist that the EAF be used strictly within a *human-in-the-loop* pipeline. Clinicians must retain final authority over how and when empathy is expressed, supported by transparent rationales and safeguards that guard against both deceptive alienation and false intimacy, thus protecting patients from the dual harms of artificial empathy.

References

Iqra Ali, Jesse Atuhurra, Hidetaka Kamigaito, and Taro Watanabe. 2025. Hlu: Human vs llm generated text detection dataset for urdu at multiple granularities. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3495–3510.

Maria Antoniak, Aakanksha Naik, Carla S Alvarado, Lucy Lu Wang, and Irene Y Chen. 2024. Nlp for maternal healthcare: Perspectives and guiding principles in the age of llms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1446–1463.

Monica Oliveira Bernardo, Dario Cecílio-Fernandes, Patrício Costa, Thelma A Quince, Manuel João Costa, and Marco Antonio Carvalho-Filho. 2018. Physicians' self-assessed empathy levels do not correlate with patients' assessments. *PloS one*, 13(5):e0198488.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.

Yibo Chai, Fengyang Wu, Rui Sun, Zhongliang Zhang, Jie Bao, Runxin Ma, Qizhou Peng, Danqin Wu, Yexing Wan, and Keyu Li. 2019. Predicting future alleviation of mental illness in social media: an empathy-based social network perspective. In *2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pages 1564–1571. IEEE.

Xinhao Chen, Chong Yang, Man Lan, Li Cai, Yang Chen, Tu Hu, Xinlin Zhuang, and Aimin Zhou. 2024. Cause-aware empathetic response generation via chain-of-thought fine-tuning. *arXiv preprint arXiv:2408.11599*.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Joost Dekker, Jeanet Karchoud, Annemarie MJ Braamse, Hilde Buiting, Inge RHM Konings, Myra E van Linde, Claudia SEW Schuurhuizen, Mirjam AG Sprangers, Aartjan TF Beekman, and Henk MW Verheul. 2020. Clinical management of emotions in patients with cancer: introducing the approach "emotional support and case finding". *Translational behavioral medicine*, 10(6):1399–1405.

Quentin Eichbaum, Charles-Antoine Barbeau-Meunier, Mary White, Revathi Ravi, Elizabeth Grant, Helen Riess, and Alan Bleakley. 2023. Empathy across cultures–one size does not fit all: from the ego-logical to the eco-logical of relational empathy. *Advances in Health Sciences Education*, 28(2):643–657.

Ronald M Epstein and Richard L Street Jr. 2007. Patient-centered communication in cancer care: promoting healing and reducing suffering.

Xiang Gao and Kamalika Das. 2024. Customizing language model responses with contrastive in-context learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18039–18046.

Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Clarissa Guidi and Chiara Traversa. 2021. Empathy in patient care: from 'clinical empathy'to 'empathic concern'. *Medicine, Health Care and Philosophy*, 24:573–585.

Lianne Hermans, Tim Olde Hartman, and Patrick W Dielissen. 2018. Differences between gp perception of delivered empathy and patient-perceived empathy: a cross-sectional study in primary care. *British Journal of General Practice*.

Hinke Hoffstädt, Jacqueline Stouthard, Maartje C Meijers, Janine Westendorp, Inge Henselmans, Peter Spreeuwenberg, Paul de Jong, Sandra van Dulmen, and Liesbeth M van Vliet. 2020. Patients' and clinicians' perceptions of clinician-expressed empathy in advanced cancer consultations and associations with patient outcomes. *Palliative Medicine Reports*, 1(1):76–83.

Ian Hsu, Somnath Saha, Phillip Todd Korthuis, Victoria Sharp, Jonathon Cohn, Richard D Moore, and Mary Catherine Beach. 2012. Providing support to patients in emotional encounters: a new perspective on missed empathic opportunities. *Patient education and counseling*, 88(3):436–442.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*.

Erica Koranteng, Arya Rao, Efren Flores, Michael Lev, Adam Landman, Keith Dreyer, and Marc Succi. 2023. Empathy and equity: Key considerations for large language model adoption in health care. *JMIR Medical Education*, 9:e51199.

Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022. A critical reflection and forward perspective on empathy and natural language processing. *arXiv preprint arXiv:2210.16604*.

Allison Claire Lahnala, Béla Neuendorf, Alexander Thomin, Charles Welch, Tina Stibane, and Lucie Flek. 2024. Appraisal framework for clinical empathy: A novel application to breaking bad news conversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1393–1407.

Andrew Lee, Jonathan K Kummerfeld, Larry An, and Rada Mihalcea. 2023. Empathy identification systems are not accurately accounting for context. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1686–1695.

Gyeongeun Lee, Zhu Wang, Sathya N Ravi, and Natalie Parde. 2025. From heart to words: Generating empathetic responses via integrated figurative language and semantic context signals. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4490–4502.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lauren A McCormack, Katherine Treiman, Douglas Rupert, Pamela Williams-Piehota, Eric Nadler, Neeraj K Arora, William Lawrence, and Richard L Street Jr. 2011. Measuring patient-centered communication in cancer care: a literature review and the development of a systematic approach. *Social science & medicine*, 72(7):1085–1095.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

N Mokhberian, MG Marmarelis, FR Hopp, V Basile, F Morstatter, and K Lerman. 2023. Capturing perspectives of crowdsourced annotators in subjective learning tasks. arxiv preprint. *arXiv preprint arXiv:2311.09743*.

Diane S Morse, Elizabeth A Edwardsen, and Howard S Gordon. 2008. Missed opportunities for interval empathy in lung cancer communication. *Archives of internal medicine*, 168(17):1853–1858.

Nature Machine Intelligence. 2024. Empathic ai can't get under the skin. *Nature Machine Intelligence*, 6:495.

Joanne K Olson. 1995. Relationships between nurse-expressed empathy, patient-perceived empathy and patient distress. *Image: The Journal of Nursing Scholarship*, 27(4):317–322.

Torkel Richert, Björn Johnson, and Bengt Svensson. 2018. Being a parent to an adult child with drug problems: Negative impacts on life situation, health, and emotions. *Journal of Family Issues*, 39(8):2311–2335.

Sermo Team. 2025. Can physicians and patients trust AI doctor apps like ChatGPT? `https://www.sermo.com/resources/ai-doctor-app/`. Blog post; accessed 22 July 2025.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

Lanrui Wang, Jiangnan Li, Chenxu Yang, Zheng Lin, Hongyin Tang, Huan Liu, Yanan Cao, Jingang Wang, and Weiping Wang. 2025. Sibyl: Empowering empathetic dialogue generation in large language models via sensible and visionary commonsense inference. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 123–140.

Joshua D Wondra and Phoebe C Ellsworth. 2015. An appraisal theory of empathy and other vicarious emotional experiences. *Psychological review*, 122(3):411.

A Empathy Applicability Framework Detail

A.1 Emotional Reactions in General Health Queries

A.1.1 Definition

Emotional Reactions refer to expressions of warmth, compassion, concern, or similar feelings conveyed by a doctor in response to a patient's query. These reactions aim to provide emotional support and reassurance to the patient.

A.1.2 Emotional Reactions Not Applicable (N/A)

Emotional reactions are not necessary or expected in the doctor's response when the patient's query is factual, neutral, or a simple advice request, without expressing emotional distress. Below are detailed categories reflecting when emotional reactions are not applicable:

1. Purely Factual Medical Queries Description: The patient requests specific medical information, including explanations of medical concepts, without emotional distress or underlying distressing uncertainty.

Examples:

- "What is the use of Tylenol?"
- "Is it possible to outgrow a seafood allergy?"

2. General Health Management Without Emotional Involvement Description: The patient seeks guidance on health management, follows up on prior advice, or requests basic guidance on minor health issues, without expressing emotional distress or underlying distressing uncertainty. Here the guidance is on what the patient should do.

Examples:

- "I'm managing diabetes with insulin. How often should I check my blood sugar levels?"
- "I have swelling in my ankle after a long walk. Should I be concerned?"
- "I had an X-ray for a fracture; should it be strapped or cast right away?"

3. Diagnosis Requests with Neutral Symptom Descriptions Description: The patient describes symptoms neutrally without expressing emotional distress or underlying distressing uncertainty. Here the request is about asking what the doctor thinks the issue is.

Examples:

- "I have intermittent knee pain from working out. How would I know if I tore cartilage?"

- "Hello. I am having pain in my jaw area, immediately in front of my left ear. The pain is random. My feeling is it is somehow related to sinus but that's just a gut feeling."

4. Hypothetical Medical Queries Without Emotional Concern Description: The patient inquires about hypothetical situations without emotional involvement.

Examples:

- "If someone has XYZ symptoms, what might be the cause?"
- "What would happen if a person skipped their medication?"

A.1.3 Emotional Reactions Applicable

Definition: Emotional reactions are necessary or expected in the doctor's response when:

- The patient expresses emotions like fear, worry, frustration, or distress.
- The patient implies emotional distress over symptoms affecting their well-being.
- The patient's tone suggests a need for reassurance or emotional support.
- The patient is expressing concern for a close relation (e.g., a child, spouse).

Below are detailed categories reflecting when emotional reactions are applicable:

1. Seriousness of Symptoms Definition: The patient describes symptoms that suggest a life-threatening or chronic health condition significantly impacting long-term health or quality of life. This includes diseases like cancer, heart disease, mental health issues, or chronic conditions leading to disability. The symptoms suggest a life-threatening or serious health condition that could significantly impact long-term health or quality of life.

Examples:

- "My father has been having severe chest pains and shortness of breath. Could it be a heart attack?"
- "I've been experiencing numbness and weakness in my limbs for months. Could this be multiple sclerosis?"
- "I'm 78 and have been told I have a floating hernia after bowel cancer surgery. Can it be cured?"

2. Severe Negative Emotion Expressed Definition: The patient explicitly states intense emotions such as fear, frustration, or anger regarding their health.

Examples:

- "I feel depressed and anxious like never before. I cannot sleep at night."
- "I am scared and plan on taking my son to the doctor. Should I be overly worried?"
- "I'm terrified about my recent diagnosis of cancer."

3. Underlying Negative Emotional State Inferred Definition: The patient implies emotional distress that isn't explicitly stated but can be inferred from their tone or descriptions, such as subtle signs of emotional worry, frustration, or distress about delays or uncertainties. Focus on emotional worry, not the medical concern.

Examples:

- "I am starting to get a little alarmed by this spotting after ovulation. Is this cause for concern?" (Worry inferred)
- "I have been trying to conceive, and the report does not look right to me. I just want to take a second opinion." (Anxiety inferred)
- "I need to be a bit more at ease after what I read about diabetic enteropathy. I was a bit scared if it might be fatal." (Fear inferred)

4. Concern Severity for Close Relations Definition: The patient is asking on behalf of someone with whom they share a close, protective relationship, implying heightened emotional concern.

Examples:

- "Hello, I am the mother of a five-year-old. He has a small lump that hasn't gone away. Should I take him to a dermatologist?"
- "My son recently started daycare and has gotten sick. His fever was 102.9. Should I take him to the hospital?"

A.2 Interpretations in General Health Queries

A.2.1 Definition

Interpretations refer to the communication of an understanding of the patient's feelings (expressed or implied) and/or experiences (contextual factors) inferred from the patient's query. It's about recognizing and articulating what the patient is feeling and why, based on their situation, concerns, and history.

A.2.2 Interpretations Applicable

Interpretations are necessary when the patient's query requires the doctor to communicate an un-

derstanding of the patient's feelings (expressed or implied) and/or experiences (contextual factors). This involves acknowledging emotions, underlying concerns, or contextual elements that influence the patient's emotional state. Below are detailed categories reflecting when interpretations are applicable:

1. Expression of Feelings (Explicit or Implicit)
Description:

The patient expresses emotions directly or implies them through language or tone. This includes feelings such as fear, anxiety, frustration, sadness, or hopelessness.

Examples:

- **Explicit Expression:**
  - "I'm really scared about these chest pains."
  - "I'm frustrated because my symptoms aren't improving."
  - "I have been in severe pain. It hurts so bad getting out of bed."
- **Implicit Expression:**
  - "I guess I have to accept this is how things will be now."
  - "Nothing seems to be helping."
  - "I don't know what to do anymore."

2. Sharing Experiences or Contextual Factors Affecting Emotional State and Well-being
Description:

The patient shares personal experiences, contextual factors, or circumstances that influence their health and emotional state. These include social, environmental, or personal situations beyond medical concerns that affect their emotional state.

Examples:

- "With my father's illness and financial stress, I'm feeling overwhelmed."
- "I've been under a lot of pressure at work, and now I'm having trouble sleeping."
- "Ever since the accident, I can't stop thinking about what happened."
- "I recently moved to a different state, haven't found a general practitioner, and haven't paid my high deductible for the year."

3. Expressions of Distressing Uncertainty About Health or Treatment
Description:

Uncertainties, confusion, or mistrust about their health status, treatment, or future are leading to emotional distress. This includes questions about prognosis, treatment effectiveness, or doubt about potential outcomes that indicate or imply underlying emotional distress. The focus should not be on uncertainty alone but specifically on uncertainty that reflects or suggests emotional distress in the patient.

Examples:

- "I'm not sure if this treatment is really working for me."
- "Do you think I should get a second opinion?"
- "Will chemo be fatal?"
- "Should my wife also get examined?"
- "Is this something that sounds like I should consider doing?"
- "I am wondering if I should see a doctor."

4. Symptoms Significantly Affecting Emotional Well-being or Daily Life
Description:

The patient describes symptoms that significantly impact their emotional well-being or daily functioning, and they express or imply emotional distress because of these symptoms. The key is the emotional impact of the symptoms, not just the symptoms themselves.

Examples:

- "My symptoms have been affecting my job for months."
- "I'm so tired all the time that I can't take care of my kids properly."
- "These migraines are making it impossible to enjoy my hobbies."
- "The pain is getting worse every day, and it's really wearing me down."

A.2.3 Interpretations Not Applicable

Interpretations are not necessary when the patient's query does not require the doctor to communicate an understanding of the patient's feelings or experiences. This occurs when:

- The query is straightforward, factual, or routine.
- There are no expressed or implied feelings needing acknowledgment.
- There are no contextual factors (experiences) or underlying uncertainty concerns leading to emotional distress that require understanding.

Below are detailed categories reflecting when interpretations are not applicable:

1. Straightforward Medical Queries Lacking Emotion, Distressing Uncertainty, and Context

Description: The patient requests specific medical information or explanations of medical concepts without expressing emotional distress, underlying distressful uncertainty, or providing context (social, environmental, or personal situations) implying an emotional state. These queries are strictly informational and lack emotional or experiential elements requiring interpretation.

Examples:

- "What is the use of Tylenol?"
- "Hello doctor, I would like to get an opinion regarding the attached chest radiograph. I wish to know if there are any abnormalities like scarring."

2. General Health Management Requests Without Emotion, Context, and Distressing Uncertainty

Description: The patient seeks guidance on health management, follows up on prior advice, or requests basic guidance on minor health issues without expressing emotional distress, underlying distressful uncertainty, or providing contextual factors (social, environmental, or personal situations) that imply an emotional state. Here the guidance is on what the patient should do.

Examples:

- "I'm managing diabetes with insulin. How often should I check my blood sugar levels?"
- "I have intermittent knee pain from working out. How would I know if I tore cartilage?"
- "I had an X-ray for a fracture; should it be strapped or cast right away?"

3. Diagnosis Requests with Neutral Symptom Descriptions Lacking Distressing Uncertainty and Context

Description:

The patient describes symptoms neutrally without expressing emotional distress or underlying distressful uncertainty. They provide necessary details without implying feelings or contextual factors (social, environmental, or personal situations) that need acknowledgment. These descriptions are straightforward and lack emotional or experiential content requiring interpretation. Here the request is about asking what the doctor thinks the issue is.

Examples:

- "I have swelling in my ankle after a long walk. Should I be concerned?"
- "Hello doctor, I am suffering from pain in my mouth. It feels like sensitivity pain. I cannot say it is pain exactly; it is irritating a lot. No pain in teeth. It feels like itching in my gums (middle of the teeth). Please tell me what I can do."

4. Hypothetical Medical Queries With No Emotions, Context, and Distressing Uncertainty

Description:

The patient inquires about hypothetical situations or general medical information without expressing or implying personal feelings or contextual factors (social, environmental, or personal situations) that need acknowledgment.

These queries are theoretical and lack emotional or experiential aspects requiring interpretation.

Examples:

- "If someone has XYZ symptoms, what might be the cause?"
- "What would happen if a person skipped their medication?"

B Annotation Instructions for Human Annotators

Annotators received an Excel workbook containing the patient queries and a fixed header with the instructions shown in Figure 3. For each `pat_query`, they assigned *Emotional Reactions* and *Interpretations* labels (`Applicable` / `Not Applicable`) and selected the justifying sub-category, as defined in Appendix A. The header also links to a Google Doc, reproduced verbatim in Appendix A, that provides the full framework details for reference during annotation.

For accessibility and clarity, we restate the instructions here in text form.

B.1 Instructions Given to Annotators

Instructions:

1. **Read the Document:** Access and thoroughly review the following document contaning the Framework Details: defined in Appendix A. Focus on understanding the details outlined below.
2. **Understand Emotional Reactions:**
   - **Emotional Reactions Definition:** Learn what emotional reactions are and their role in doctor-patient communication.
   - Understand when emotional reactions are applicable or not applicable by reviewing:

Sub-definitions, Subcategories and Examples that illustrate their use in the relevant scenarios.

3. **Classify Emotional Reactions:** For each patient query, follow these steps:
   - **Determine Emotional Reactions Applicability or Not Applicability:** Decide whether emotional reactions are applicable or not applicable in response to the patient query.
   - **Select a Subcategory:**
     – If applicable, choose the subcategory that best explains why emotional reactions are needed in response to the patient query.
     – If not applicable, select the subcategory that justifies why emotional reactions are not necessary in response to the patient query.

4. **Understand Interpretations:**
   - **Interpretations Definition:** Learn what interpretations are and their role in doctor-patient communication.
   - Understand when interpretations are applicable or not applicable by reviewing: Sub-definitions, Subcategories and Examples that illustrate their use in the relevant scenarios.

5. **Classify Interpretations:** For each patient query, follow these steps:
   - **Determine Interpretations Applicability or Not Applicability:** Decide whether interpretations are applicable or not applicable in response to the patient query.
   - **Select a Subcategory:**
     – If applicable, choose the subcategory that best explains why interpretations are needed in response to the patient query.
     – If not applicable, select the subcategory that justifies why interpretations are not necessary in response to the patient query.

B.2 Additional Verbal Clarifications

During training sessions, annotators received the following clarifications:

- If they could not understand whether the symptoms or medical issue is **severe**, they were allowed to briefly search online (e.g., Google) to check whether the condition is typically serious.
- If they were unsure whether the query was emotionally significant for the patient, they were encouraged to **go with what they felt** and believe their own judgment.
- For each dimension (EA and IA), annotators were instructed to:
  1. Read the patient query.
  2. While annotating a dimension, first read the *Applicable* definition. If they believe it fits, go through the Applicable subcategories one by one and tag at least the one they think fits best.
  3. If the Applicable definition does not feel like it fits, they should still briefly review the Applicable subcategories to verify this.
  4. Then move to the *Not Applicable* definition and repeat the same process with the Not Applicable subcategories.

B.3 Boundary Cases: Subjectivity and Lack of Medical Expertise

Empathy applicability judgments are inherently subjective, and some patient queries lie at the boundary between emotional and informational intent. Such disagreements often reflect **legitimate interpretive variability** rather than simple annotation error. These challenges are explored in detail in Section 5.2 (Systematic Challenges in Operationalizing Anticipatory Empathy); here, we briefly revisit representative cases to make these boundary conditions more transparent.

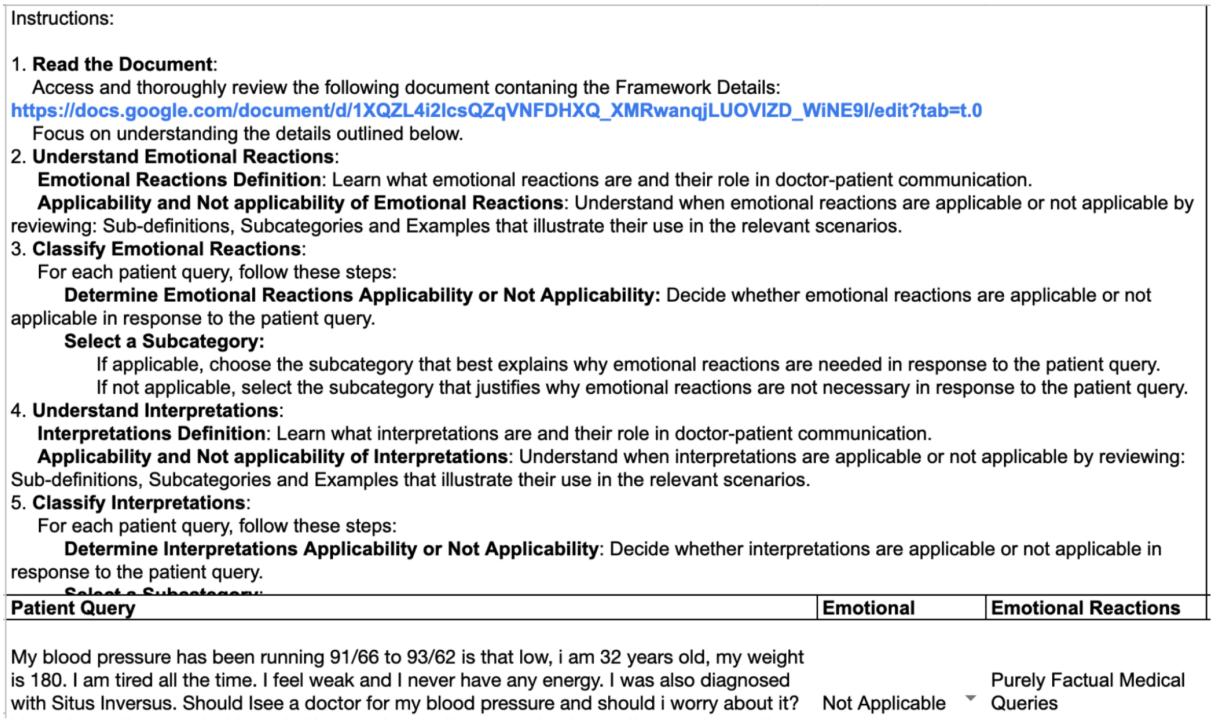
To make these cases transparent, we provide a supplementary file (`misalignment_analysis.csv`) listing detailed patient queries, mis-aligned labels, and qualitative interpretations. The queries highlighted as exhibiting *Divergent Interpretation* correspond to *Reasonable Disagreement* around both symptom severity and whether emotional content is present.

C Illustrative Scenarios for EAF Operationalization

See Table 4 for illustrative scenarios demonstrating the operationalization of the EAF.

D Appendix: Human-GPT Agreement Analysis

Table 5 presents pairwise agreement between GPT and each human annotator. “Agreed” and “Disagreed” columns denote the number of queries


Instructions:

1. **Read the Document**:
Access and thoroughly review the following document contaning the Framework Details:
https://docs.google.com/document/d/1XQZL4i2lcsQZqVNFDHXQ_XMRwanqjLUOVlZD_WiNE9I/edit?tab=t.0
Focus on understanding the details outlined below.
2. **Understand Emotional Reactions**:
Emotional Reactions Definition: Learn what emotional reactions are and their role in doctor-patient communication.
Applicability and Not applicability of Emotional Reactions: Understand when emotional reactions are applicable or not applicable by reviewing: Sub-definitions, Subcategories and Examples that illustrate their use in the relevant scenarios.
3. **Classify Emotional Reactions**:
For each patient query, follow these steps:
Determine Emotional Reactions Applicability or Not Applicability: Decide whether emotional reactions are applicable or not applicable in response to the patient query.
Select a Subcategory:
If applicable, choose the subcategory that best explains why emotional reactions are needed in response to the patient query.
If not applicable, select the subcategory that justifies why emotional reactions are not necessary in response to the patient query.
4. **Understand Interpretations**:
Interpretations Definition: Learn what interpretations are and their role in doctor-patient communication.
Applicability and Not applicability of Interpretations: Understand when interpretations are applicable or not applicable by reviewing: Sub-definitions, Subcategories and Examples that illustrate their use in the relevant scenarios.
5. **Classify Interpretations**:
For each patient query, follow these steps:
Determine Interpretations Applicability or Not Applicability: Decide whether interpretations are applicable or not applicable in response to the patient query.

Patient Query	Emotional	Emotional Reactions
My blood pressure has been running 91/66 to 93/62 is that low, i am 32 years old, my weight is 180. I am tired all the time. I feel weak and I never have any energy. I was also diagnosed with Situs Inversus. Should Isee a doctor for my blood pressure and should i worry about it?	Not Applicable	Purely Factual Medical Queries

Figure 3: Screenshot of the annotation spreadsheet provided to annotators. The header shows the instructions and links to the framework document.

where both annotators assigned the same or different labels of Applicable or Not Applicable, respectively.

E Model Architecture Details

Each empathy dimension—Emotional Reactions (EA) and Interpretations (IA)—is modeled independently. We fine-tune a pretrained RoBERTa-based model (Liu et al., 2019) separately for each dimension, while maintaining the same overall architecture. "Independently" means each classifier learns to predict the applicability of one dimension without sharing parameters or optimization across tasks. For fine-tuning, we incorporate an attention mechanism based on a feed-forward network. The model architecture is illustrated in Figure 4.

The model follows an attention-based pooling approach built on top of a pretrained RoBERTa encoder. The encoder converts patient queries into contextualized token embeddings, capturing the meaning of each word based on its surrounding context. When a sentence is processed by RoBERTa, it generates a hidden representation for each token, reflecting its contextual meaning. Unlike traditional methods that rely solely on the [CLS] token or an average of all embeddings, this model applies a learned attention mechanism to identify the most relevant tokens for classification.

Specifically, the model uses a feed-forward neural network to compute attention scores for each token. A linear transformation first maps each token embedding to a scalar score, which then passes through a Tanh activation to constrain values between [1,1] and avoid extremes. Since not all tokens contribute equally to classification, the model converts these raw scores into attention weights using a softmax function across the sequence. This normalization ensures that important words receive higher weights, while less relevant words are assigned lower importance.

After computing attention weights, the model performs a weighted sum of token embeddings. Tokens with higher attention scores contribute more significantly to the final pooled representation, highlighting the most relevant parts of the query. This pooled vector is then passed through a classification-linear layer, which outputs logits representing the likelihood of belonging to either the "Not Applicable" or "Applicable" class. During training, the model optimizes both the attention mechanism and the classification layer via cross-entropy loss, thereby improving accuracy in empathy classification.

Training separate models for EA and IA avoids crosstalk between tasks. Each classifier learns

Empathy Dimension	Scenario Type	Scenario	Applicability	Explanation
Emotional Reaction	Explicit Need	*"Hello doctor, I am having constant eye floaters, low back and hip pain, and also my rib cage hurts. I feel depressed and anxious like never before. I cannot sleep at night. An MRI of my brain shows a tiny flare, but radiologists say it's nothing to worry about. What should I do?"*	Applicable	The patient explicitly expresses intense negative emotions—feeling depressed and anxious—and states an inability to sleep. An emotional reaction from the doctor is necessary to provide support and reassurance.
Emotional Reaction	Implicit Need	*"Hello doctor, my son has been experiencing frequent headaches over the past week. We've tried over-the-counter medications, but there's no improvement. What should we do?"*	Applicable	Emotional reactions are applicable here because, as Richert et al. (Richert et al., 2018) find, parents of children with health (drug) issues often experience significant distress and negative mental health effects. The mother may be experiencing worry and anxiety about her child's well-being, even if she doesn't explicitly express it.
Emotional Reaction	Not Needed	*"Hello doctor, I was suffering from an infection in my tonsil for the past four days. I went to an ENT specialist who prescribed antibiotics. Now my tonsil pain has subsided, but I still feel something stuck on the left side of my throat where the pain was. I have no problem swallowing. Kindly advise me on what to do next."*	Not Applicable	The patient provides a neutral description of symptoms without expressing emotional concern or distress. The primary need is factual medical advice. An emotional reaction from the doctor is not necessary in this case.
Interpretation	Explicit Need	*"Hello doctor, I am feeling extremely anxious about my upcoming surgery. I can't stop worrying about the possible complications."*	Applicable	The patient explicitly expresses feelings of anxiety and worry. The doctor should communicate an understanding of these feelings, acknowledging the patient's emotional state and providing appropriate support.
Interpretation	Implicit Need	*"Hello doctor, I've been taking the medication as prescribed, but I'm not seeing any improvement. Is there something I'm doing wrong?"*	Applicable	The patient implies feelings of frustration and possibly self-blame. The doctor should interpret and acknowledge these underlying feelings, demonstrating understanding and support.
Interpretation	Not Needed	*"I was playing with my sister s boyfriends brother and I swung to hit him like I said we were playing around and I my wrist hit his elbow really hard when it happened my hand got really numb and my vein was hurting really bad and it s 6 hours later and my vein still hurts what should I do"*	Not Applicable	The query is a straightforward request for diagnosis with neutral symptom descriptions. It does not express emotions or distressing contextual factors that require acknowledgment. The doctor's response should focus solely on providing a factual diagnosis.

Table 4: Empathy Dimensions, Scenarios, Applicability, and Explanations

dimension-specific patterns from the data, resulting in a simple and modular approach that enables focused analysis of empathy applicability in patient queries.

F Prompt Design for LLM Annotations

For more detail on the prompt design used for LLM (GPT-4o, o1) based annotations, we provide here the exact prompts used in our experiments.

We used two styles of prompts:

- **With-framework (contrastive) prompts**: the LLM received the full Empathy Applicability Framework for a given dimension (Emotional Reactions or Interpretations), including both Applicable and Not Applicable subcategories with examples for each. This creates a contrastive in-context signal: the model must decide between multiple subcategories across both classes.
- **Without-framework prompts**: the LLM only received a short task definition (dimension defini-

Table 5: Cohen's κ agreement scores and confusion matrix counts between GPT-4o and each human annotator for Emotional Reactions (EA) and Interpretations (IA)

Annotator 1	Annotator 2	Kappa EA	Kappa IA	Agreed EA	Disagreed EA	Agreed IA	Disagreed IA
HA2	GPT	0.4402	0.5306	917	379	988	308
HA1	GPT	0.4096	0.3612	940	356	890	406

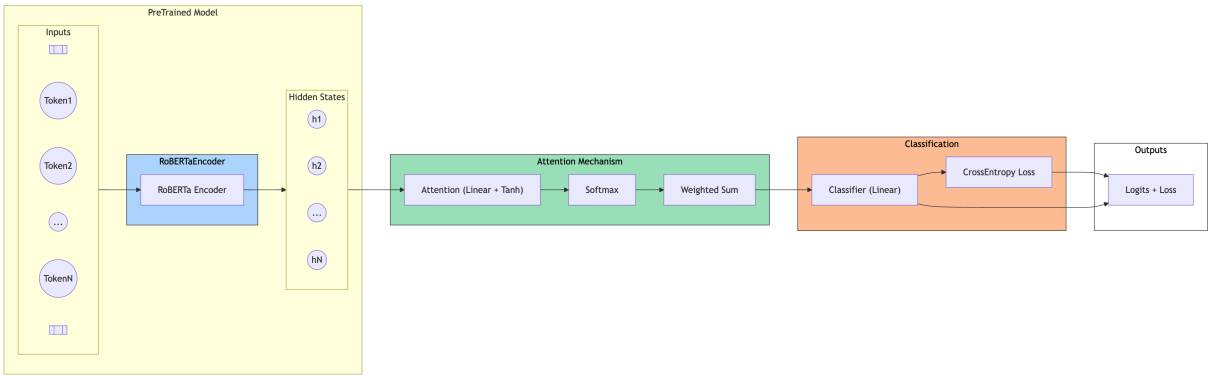


Figure 4: Empathy Dimension Applicability Model Architecture

tion + binary Applicability decision), without any subcategories or examples. This approximates a generic zero-shot setup without our framework.

F.1 Emotional Reactions: With-Framework Contrastive Prompt

The core schema passed to the LLM for Emotional Reactions with the full framework was:

```
ANNOTATION_SCHEMA = {
  "instruction": "Annotate emotional reactions in general
       health queries based on the following schema. For each
        query, return the matching subcategories. Think
       logically and ensure to revisit your annotation for
       each query",
  "definitions": {
     "Emotional Reactions": {
        "description": "Expressions of warmth, compassion,
             concern, or similar feelings conveyed by a
             doctor in response to a patient's query.",
        "categories": [
           {
              "name": "Purely Factual Medical Queries",
              "description": "The patient requests specific
                    medical information, including
                   explanations of medical concepts,
                   without emotional distress or
                   underlying distressing uncertainty.",
              "examples": ["What is the use of Tylenol?", "
                   Is it possible to outgrow a seafood
                   allergy?"],
              "class" : "Emotional Reactions Not Applicable
                   "
           },
           {
              "name": "General Health Management Without
                   Emotional Involvement",
              "description": "The patient seeks guidance on
                    health management, follows up on prior
                    advice, or requests basic guidance on
                   minor health issues, without expressing
                    emotional distress or underlying
                   distressing uncertainty. Here the
                   guidance is on what the patient should
                   do.",
              "examples": ["I'm managing diabetes with
                   insulin. How often should I check my
                   blood sugar levels?", "I have swelling
                   in my ankle after a long walk. Should I
                    be concerned?"],
              "class" : "Emotional Reactions Not Applicable
                   "
           },
           {
              "name": "Diagnosis Requests with Neutral
                   Symptom Descriptions",
              "description": "The patient describes
                   symptoms neutrally without expressing
                   emotional distress or underlying
                   distressing uncertainty. Here the
                   request is about asking what the doctor
                    thinks the issue is.",
              "examples": ["I have intermittent knee pain
                   from working out. How would I know if I
                    tore cartilage?", "Hello. I am having
                   pain in my jaw area, immediately in
                   front of my left ear."],
              "class" : "Emotional Reactions Not Applicable
                   "
           },
           {
              "name": "Hypothetical Medical Queries Without
                    Emotional Concern",
              "description": "The patient inquires about
                   hypothetical situations without
                   emotional involvement.",
              "examples": ["If someone has XYZ symptoms,
                   what might be the cause?", "What would
                   happen if a person skipped their
                   medication?"],
              "class" : "Emotional Reactions Not Applicable
                   "
           },
           {
              "name": "Seriousness of Symptoms",
              "description": "The patient describes
                   symptoms that suggest a life-
                   threatening or chronic health condition
                    significantly impacting long-term
                   health or quality of life. This
                   includes diseases like cancer, heart
                   disease, mental health issues, or
                   chronic conditions leading to
                   disability.",
              "examples": ["My father has been having
                   severe chest pains and shortness of
                   breath. Could it be a heart attack?", "
                   I've been experiencing numbness and
                   weakness in my limbs for months."],
              "class" : "Emotional Reactions Applicable"
           },
```

```
            {
                "name": "Severe Negative Emotion Expressed",
                "description": "The patient explicitly states
                      intense emotions such as fear,
                     frustration, or anger regarding their
                     health.",
                "examples": ["I feel depressed and anxious
                     like never before. I cannot sleep at
                     night.", "I'm terrified about my recent
                      diagnosis of cancer."],
                "class" : "Emotional Reactions Applicable"
            },
            {
                "name": "Underlying Negative Emotional State
                     Inferred",
                "description": "The patient implies emotional
                      distress that isn't explicitly stated
                     but can be inferred from their tone or
                     descriptions, such as subtle signs of
                     emotional worry, frustration, or
                     distress about delays or underlying
                     distressing uncertainties. Focus on
                     emotional worry, not the medical
                     concern.",
                "examples": ["I am starting to get a little
                     alarmed by this spotting after
                     ovulation. Is this cause for concern?",
                      "I need to be a bit more at ease after
                      what I read about diabetic enteropathy
                     ."],
                "class" : "Emotional Reactions Applicable"
            },
            {
                "name": "Concern Severity for Close Relations
                     ",
                "description": "The patient is asking on
                     behalf of someone with whom they share
                     a close, protective relationship,
                     implying heightened emotional concern.",

                "examples": ["Hello, I am the mother of a
                     five-year-old. He has a small lump that
                      hasn't gone away.", "My son recently
                     started daycare and has gotten sick.
                     His fever was 102.9. Should I take him
                     to the hospital?"],
                "class" : "Emotional Reactions Applicable"
            }
        ]
    }
  },
  "output_format": "json",
  "example_query": {
    "query": "I'm scared and plan on taking my son to the
         doctor. Should I be overly worried?",
    "annotations": [
      {
        "subcategories": [
          {"name": "Severe Negative Emotion Expressed"},
          {"name": "Concern Severity for Close Relations"}
        ],
        "class": "Emotional Reactions Applicable",
        "reason": "The patient explicitly expresses fear
             regarding their son's health and shows
             heightened emotional concern for a close
             relation."
      }
    ]
  }
}
```

F.2 Interpretations: With-Framework Contrastive Prompt

The corresponding schema for Interpretations (IA) with the full framework was:

```
ANNOTATION_SCHEMA = {
  "instruction": "Annotate interpretations in general health
       queries based on the following schema. For each query,
        return the matching subcategories. Think logically
       and ensure to revisit your annotation for each query",
  "definitions": {
    "Interpretations": {
      "description": "Interpretations refer to the
           communication of an understanding of the
           patients feelings (explicit or implied) and/or
            experiences (contextual factors) inferred
           from their query. It's about recognizing and
           articulating what the patient is feeling and
           why, based on their situation, concerns, and
           history.",
      "categories": [
        {
          "name": "Expression of Feelings (Explicit or
               Implicit)",
          "description": "The patient expresses
               emotions directly or implies them
               through language or tone. This includes
                feelings such as fear, anxiety,
               frustration, sadness, or hopelessness.",

          "examples": [
            "I'm really scared about these chest
                 pains.",
            "I'm frustrated because my symptoms aren'
                 t improving.",
            "I guess I have to accept this is how
                 things will be now.",
            "Nothing seems to be helping.",
            "I don't know what to do anymore."
          ],
          "class": "Interpretations Applicable"
        },
        {
          "name": "Sharing of Experiences or Contextual
                Factors Affecting Emotional State and
               Well being",
          "description": "The patient shares personal
               experiences, contextual factors, or
               circumstances that influence their
               health and emotional state. These
               include social, environmental, or
               personal situations, beyond medical
               concerns, that affect their emotional
               state.",
          "examples": [
            "With my fathers illness and financial
                 stress, Im feeling overwhelmed.",
            "I've been under a lot of pressure at
                 work, and now I'm having trouble
                 sleeping.",
            "Ever since the accident, I can't stop
                 thinking about what happened.",
            "I recently moved to a different state,
                 haven't found a general practitioner
                 , and haven't paid my high
                 deductible for the year."
          ],
          "class": "Interpretations Applicable"
        },
        {
          "name": "Expressions of Distressing
               Uncertainty About Health or Treatment",
          "description": "Uncertainties, confusion, or
               mistrust expressed by a patient about
               their health status, treatment, or
               future that lead to significant
               emotional distress. This includes
               statements involving questions about
               prognosis, treatment effectiveness, or
               doubt about potential outcomes,
               specifically when accompanied by
               explicit or implicit signs of emotional
                distress.",
          "examples": [
            "Im not sure if this treatment is really
                 working for me, and its making me
                 anxious.",
            "Is this something that sounds like I
                 should consider doing? Im so
                 confused about whats right.",
            "I feel lost. Should my wife also get
                 examined?",
            "Do you think theres any hope for me
                 after trying this?"
          ],
          "class": "Interpretations Applicable"
        },
        {
```

```
            "name": "Symptoms Significantly Affecting
                 Emotional Well-being or Daily Life",
            "description": "The patient describes
                 symptoms that significantly impact
                 their emotional well-being or daily
                 functioning, and they express or imply
                 emotional distress because of these
                 symptoms. The key is the emotional
                 impact of the symptoms, not just the
                 symptoms themselves.",
            "examples": [
                "These migraines are making it impossible
                     to enjoy my hobbies.",
                "I'm so tired all the time that I can't
                     take care of my kids properly.",
                "My symptoms have been affecting my job
                     for months.",
                "The pain is getting worse every day, and
                     it's really wearing me down."
            ],
            "class": "Interpretations Applicable"
        },
        {
            "name": "Straightforward Medical Queries
                 Lacking Emotion, Distressing
                 Uncertainty, and Context",
            "description": "The patient requests specific
                 medical information or explanations of
                 medical concepts without expressing
                 emotional distress, underlying
                 distressful uncertainty or providing
                 context (social, environmental, or
                 personal situations) implying an
                 emotional state. These queries are
                 strictly informational and lack
                 emotional or experiential elements
                 requiring interpretation.",
            "examples": [
                "What is the use of Tylenol?",
                "Hello doctor, I would like to get an
                     opinion regarding the attached chest
                     radiograph. I wish to know if there
                     are any abnormalities like scarring
                     ."
            ],
            "class": "Interpretations Not Applicable"
        },
        {
            "name": "General health management requests
                 Without Emotion, Context, and
                 Distressing Uncertainty",
            "description": "The patient seeks guidance on
                 health management, follows up on prior
                 advice, or requests basic guidance on
                 minor health issues without expressing
                 emotional distress, underlying
                 distressful uncertainty, or providing
                 contextual factors that imply an
                 emotional state. Additionally, they may
                 include personal medical context, such
                 as test results, medications taken,
                 and previous medical consultations.
                 Here the guidance is on what the
                 patient should do.",
            "examples": [
                "I have intermittent knee pain from
                     working out. How would I know if I
                     tore cartilage?",
                "I had an X-ray for a fracture; should it
                     be strapped or cast right away?",
                "Hi, my husband is 39, and his SGPT and
                     SGOT levels in a recent test were
                     101 and 98 respectively. His
                     triglycerides are 280, which is high
                     . His height is 168 cm and weight is
                     79 kg. What does a rise in these
                     values indicate? What precautions
                     should he take?"
            ],
            "class": "Interpretations Not Applicable"
        },
        {
            "name": "Diagnosis Requests with Neutral
                 Symptom Descriptions Lacking
                 Distressing Uncertainty and Context",
            "description": "The patient describes
                 symptoms neutrally without expressing
                 emotional distress or underlying
                 distressful uncertainty. They provide
                 necessary details without implying
                 feelings or contextual factors. These
                 descriptions are straightforward and
                 lack emotional or experiential content
                 requiring interpretation. Here the
                 request is about asking what the doctor
                 thinks the issue is.",
            "examples": [
                "I have swelling in my ankle after a long
                     walk. Should I be concerned?",
                "Hello doctor, I am suffering from pain
                     in my mouth. It feels like
                     sensitivity pain. I cannot say it is
                     pain exactly; it is irritating a
                     lot. No pain in teeth. It feels like
                     itching in my gums (middle of the
                     teeth). Please tell me what I can do
                     ."
            ],
            "class": "Interpretations Not Applicable"
        },
        {
            "name": "Hypothetical Medical Queries with no
                 Emotions, Context, and Distressing
                 Uncertainty",
            "description": "The patient inquires about
                 hypothetical situations or general
                 medical information without expressing
                 or implying personal feelings or
                 contextual factors that need
                 acknowledgment. These queries are
                 theoretical and lack emotional or
                 experiential aspects requiring
                 interpretation.",
            "examples": [
                "If someone has XYZ symptoms, what might
                     be the cause?",
                "What would happen if a person skipped
                     their medication?"
            ],
            "class": "Interpretations Not Applicable"
        }
    ]
  }
 },
 "output_format": "json",
 "example_query": {
  "query": "I'm not sure if this treatment is really
       working for me.",
  "annotations": [
   {
    "subcategories": [
     {
      "name": "Expressions of Distressing
           Uncertainty About Health or
           Treatment"
     }
    ],
    "class": "Interpretations Applicable",
    "reason": "The patient explicitly expresses doubt
         about the effectiveness of the treatment,
         which requires interpretation."
   }
  ]
 }
}
```

F.3 Prompts Without the Framework (Definition-Only)

For the *without-framework* condition, the LLM received only a short task description and the names of the Applicability labels. No subcategories or examples were provided.

F.3.1 Emotional Reactions (without framework).

```
ANNOTATION_SCHEMA = {
```

```
    "instruction": "Read the patient query and decide whether
         emotional reactions are necessary in the response. "
                "Emotional reactions refer to the expressions
                     of warmth, compassion, concern, or
                     similar feelings "
                "conveyed by a doctor in response to a patient
                     's query. "
                "If emotional reactions are necessary, mark it
                      as 'Emotional Reactions Applicable'. "
                "If not, mark it as 'Emotional Reactions Not
                     Applicable'. Think carefully and be
                     consistent.",
    "output_format": "json"
}
```

F.3.2 Interpretations (without framework).

```
ANNOTATION_SCHEMA = {
    "instruction": "Read the patient query and decide whether
         interpretations are necessary in the response. "
                "Interpretations refer to the communication of
                      an understanding of the patients
                     feelings "
                "(explicit or implied) and/or experiences (
                     contextual factors) inferred from their
                     query. "
                "If interpretations are necessary, mark it as
                     'Interpretations Applicable'. "
                "If not, mark it as 'Interpretations Not
                     Applicable'. Think carefully and be
                     consistent.",
    "output_format": "json"
}
```

Together, these listings document the exact prompts used in both the with-framework (contrastive) and without-framework settings for LLM annotations.

G Dataset Analyses

To characterize the released benchmark beyond agreement and modeling results, Figure 5 summarizes label base rates (H1/H2/GPT), EA–IA co-applicability patterns, and the heavy-tailed distribution of query lengths, providing a high-level view of dataset variability. In the following sections, we present additional analyses of annotation consistency, framework coherence, and stability across the labeling process. Specifically, we analyze: (i) subcategory usage and co-occurrence patterns across human annotators and GPT rationales; (ii) length effects on applicability with confidence intervals; (iii) run-order drift checks to assess stability over the labeling sequence; and (iv) match vs. miss portions of subcategory rationales when humans and GPT agree on the overall applicability label, to assess alignment in rationales.

G.1 Subcategory Prevalence and EA×IA Co-occurrence (Humans vs. GPT)

Table 6 and Table 7 report the prevalence of subcategories along with the **Applicable** or **Not Applicable** classification of each subcategory under the EAF. This makes it possible to assess the *intuitive coherence* of the framework: subcategories that encode similar affective cues or uncertainty (Applicable) should align with each other in all dimensions, while informational or routine requests (Not Applicable) should be clustered separately.

Because each query receives exactly one subcategory from each human annotator, human counts sum to $N = 1296$ per dimension. In contrast, GPT may assign multiple subcategories per query; thus GPT totals exceed N (EA: 2100; IA: 2572), corresponding to an average of 1.62 EA subcategories and 1.98 IA subcategories per query.

The prevalence distributions show both stable structure and meaningful variability in how cues are operationalized. For EA, H1 frequently uses the Not Applicable subcategory *Factual Medical Query* (31.8%), while H2 rarely uses it (2.9%) and instead assigns both Applicable and Not Applicable categories such as *Concern for Relations* (20.6%) and *General Health Management* (19.4%). A similar pattern appears in IA: H1 frequently assigns the Not Applicable subcategory *Straightforward Medical Query* (33.3%), whereas H2 more often assigns the Applicable subcategory *Distressing Uncertainty* (40.8%). GPT assigns Applicable EA categories at higher rates, especially *Inferred Negative Emotion* (59.4%) and *Symptom Seriousness* (27.5%), reflecting broader recall for empathic-need cues. Across dimensions, GPT again assigns multiple IA cues per query and frequently marks Applicable interpretive cues (*Distressing Uncertainty*, *Feelings Expression*, *Impact on Daily Life*, *Sharing Affecting Context*), consistent with its broader recall of applicability signals.

Figure 6 further supports the framework's *intuitive coherence* when interpreted through the Applicable vs. Not Applicable split of subcategories (Tables 6–7). Not Applicable request-types tend to pair with Not Applicable interpretations (Not×Not), while affective/uncertainty cues (Applicable) more often co-occur with Applicable interpretive signals (App×App). Concretely, in the human heatmaps, EA *Factual Medical Query* aligns strongly with IA *Straightforward Medical Query* (H1: 30%), and *Neutral Diagnosis Request* co-occurs with its IA counterpart (H1: 19%; H2: 9%), both canonical Not×Not pairings. In contrast, Applicable EA cues align with Applicable IA cues: *Severe Negative Emotion* co-occurs with *Feelings Expression* (H1: 15%), and *Concern for Relations* frequently pairs with *Distressing Uncertainty* (H2: 16%), reflecting clinically intuitive links between expressed (or inferred) affect and corresponding interpretive needs.

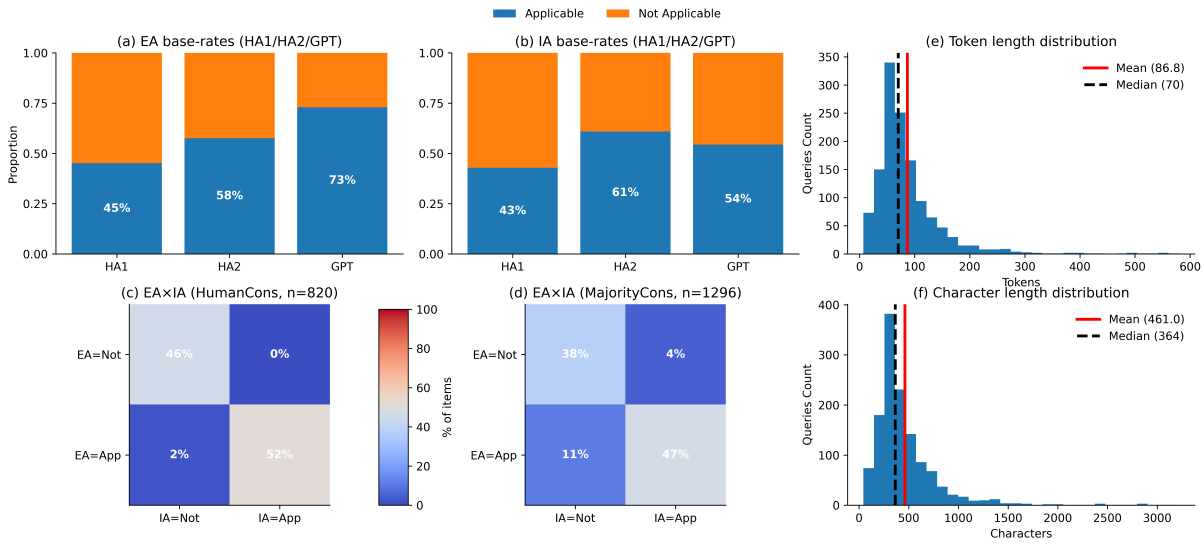


Figure 5: **Dataset overview panel: base rates, EA–IA coupling, and query length distributions.** Panels (a–b) report the binary label base rates for *Emotional Reactions (EA)* and *Interpretations (IA)* from Human Annotator 1 (HA1), Human Annotator 2 (HA2), and GPT. Bars are shown as stacked proportions of *Applicable* vs. *Not Applicable*. Panels (c–d) show EA×IA co-occurrence as 2×2 heatmaps for (c) *Human consensus* (only items where HA1 and HA2 agree on both EA and IA) and (d) *Majority consensus* (majority vote over HA1, HA2, and GPT), with each cell annotated by the percentage of items in that consensus subset; Panels (e–f) summarize query length: (e) token-count histogram (simple word tokenization) and (f) character-count histogram; vertical reference lines mark the mean (solid red) and median (dashed black). Together, the figure summarizes label prevalence, the empirical coupling between EA and IA decisions, and the distribution of textual input lengths in patient queries.

GPT exhibits broader App×App cross-pairings, most prominently EA *Inferred Negative Emotion* co-occurring with multiple Applicable IA cues including *Distressing Uncertainty* (38%), *Feelings Expression* (27%), *Impact on Daily Life* (24%), and *Sharing Affecting Context* (20%), which is expected given GPT's multi-label rationale annotations. Overall, these structured pairings indicate that cross-dimensional co-occurrence is not arbitrary: it aligns with the EAF's applicability semantics while also highlighting how multi-cue rationales (GPT) differ from single-label human assignments.

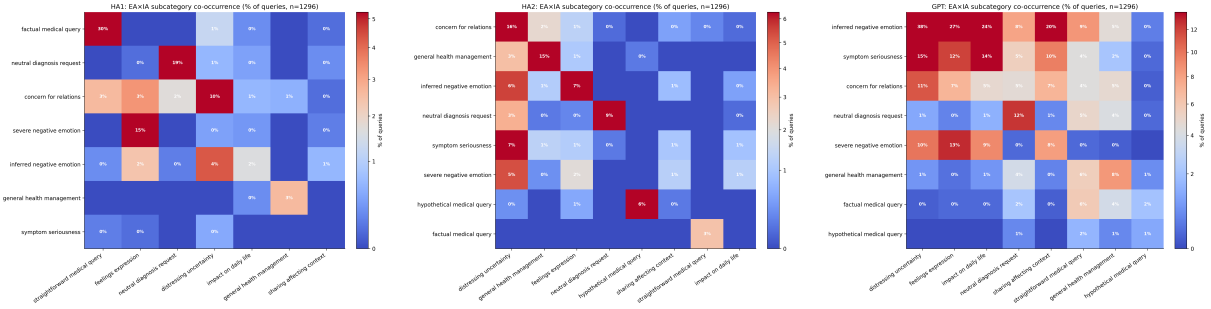
Additionally, these prevalence and co-occurrence results provide actionable signals for refining the EAF and improving annotation practice. First, the strong Not×Not and App×App structure suggests the framework's applicability split is broadly coherent, but the sharp annotator skews in how "routine/informational" vs. "affective/uncertainty" cues are operationalized (e.g., heavier use of *Factual Medical Query* and *Straightforward Medical Query* versus greater use of *Concern for Relations* and *Distressing Uncertainty*) highlight subcategories that may be very broad or boundary-sensitive. These are prime candidates for guideline refinement (clearer decision rules, additional contrastive examples, or merging/splitting categories), while consistently rare categories can be reconsidered for consolidation if they contribute limited discriminative value. Second, the co-occurrence matrices can be leveraged to diagnose systematic annotation patterns and potential drift: stable, clinically intuitive pairings (e.g., *Severe Negative Emotion* with *Feelings Expression*) indicate consistent interpretation, whereas unexpected or diffuse pairings can reveal where annotators may be diverging due to inconsistent application rather than subjectivity, and targeted review could mitigate those differences. In this sense, co-occurrence structure is not only a validation of the EAF semantics, but also a practical tool for targeted adjudication, annotator training, and future modeling choices.

G.2 Length Effects on Applicability (with Confidence Intervals)

To assess whether query length is associated with applicability judgments, we stratify items into token-length deciles and plot EA/IA applicability rates for each label source (HA1, HA2, GPT, Hu-

EA subcategory	Shortname	Class	HA1	HA2	GPT
Underlying Negative Emotional State Inferred	inferred negative emotion	App	121 (9.3)	197 (15.2)	770 (59.4)
Concern Severity For Close Relations	concern for relations	App	246 (19.0)	267 (20.6)	277 (21.4)
Severe Negative Emotion Expressed	severe negative emotion	App	209 (16.1)	132 (10.2)	177 (13.7)
Seriousness Of Symptoms	symptom seriousness	App	9 (0.7)	150 (11.6)	357 (27.5)
Diagnosis Requests With Neutral Symptom Descriptions	neutral diagnosis request	Not	262 (20.2)	165 (12.7)	226 (17.4)
Purely Factual Medical Queries	factual medical query	Not	412 (31.8)	38 (2.9)	108 (8.3)
General Health Management Without Emotional Involvement	general health management	Not	37 (2.9)	252 (19.4)	156 (12.0)
Hypothetical Medical Queries Without Emotional Concern	hypothetical medical query	Not	0 (0.0)	95 (7.3)	29 (2.2)

Table 6: EA subcategory prevalence with applicability class (App vs. Not). Values are count (percent of $N = 1296$). Humans assign one subcategory per query (counts sum to N). GPT may assign multiple subcategories per query; thus GPT counts can exceed N.



(a) HA1: EA×IA co-occurrence. (b) HA2: EA×IA co-occurrence. (c) GPT: EA×IA co-occurrence (multi-label IA/EA allowed).

Figure 6: EA×IA subcategory co-occurrence (percent of queries, $N = 1296$). Humans provide a single EA and IA subcategory per query, yielding sharper pairings; GPT may assign multiple subcategories per query, producing broader co-occurrence patterns.

manCons, MajorityCons). The x-axis reports the mean token length per decile (μ), and we annotate the number of items per decile (n). Error bars denote confidence intervals for each applicability estimate.

Across both dimensions, applicability increases with length: shorter queries (D1–D2) receive substantially lower applicability rates, while longer queries (D9–D10) show consistently higher applicability across all sources. Although absolute rates differ by source (e.g., GPT tends to assign applicability more frequently than humans), the upward trend is shared, suggesting that length (and the additional context typically present in longer queries) is systematically associated with applicability rather than reflecting annotator idiosyncrasies alone.

G.3 Run-order Drift Checks

Finally, we evaluate whether labeling behavior drifts over the annotation sequence (e.g., fatigue or order effects). We plot the rolling mean of pairwise differences in applicability rates (window size $= 32$) across run index for each pair (H1–H2, H1–GPT, H2–GPT), and apply ADWIN change detection ($\delta = 0.002$) to flag potential change points.

Across $N = 1296$ items, ADWIN identifies only a small number of localized change points in each rolling-difference stream. For EA, ADWIN flags 3 points for H1–H2 (at indices 767, 1023, 1279), 2 for H1–GPT (223, 1151), and 3 for H2–GPT (191, 479, 959). For IA, ADWIN flags 3 points for H1–H2 (415, 639, 1279), 3 for H1–GPT (191, 863, 1151), and 4 for H2–GPT (223, 415, 607, 1087). Because these indices correspond to positions in

IA subcategory	Shortname	Class	HA1	HA2	GPT
Expressions Of Distressing Uncertainty About Health Or Treatment	distressing uncertainty	App	222 (17.1)	529 (40.8)	556 (42.9)
Expression Of Feelings (Explicit Or Implicit)	feelings expression	App	274 (21.1)	176 (13.6)	423 (32.6)
Symptoms Significantly Affecting Emotional Well-Being Or Daily Life	impact on daily life	App	44 (3.4)	36 (2.8)	382 (29.5)
Sharing Experiences Or Contextual Factors Affecting Emotional State And Well Being	sharing affecting context	App	15 (1.2)	48 (3.7)	308 (23.8)
Straightforward Medical Queries Lacking Emotion, Distressing Uncertainty, And Context	straightforward medical query	Not	431 (33.3)	41 (3.2)	285 (22.0)
Diagnosis Requests With Neutral Symptom Descriptions Lacking Distressing Uncertainty And Context	neutral diagnosis request	Not	269 (20.8)	118 (9.1)	339 (26.2)
General Health Management Requests Without Emotion, Context, And Distressing Uncertainty	general health management	Not	41 (3.2)	264 (20.4)	242 (18.7)
Hypothetical Medical Queries With No Emotions, Context, And Distressing Uncertainty	hypothetical medical query	Not	0 (0.0)	84 (6.5)	37 (2.9)

Table 7: IA subcategory prevalence with applicability class (App vs. Not). Values are count (percent of $N = 1296$). Humans assign one subcategory per query (counts sum to N). GPT may assign multiple subcategories per query; thus GPT counts can exceed N.

the rolling stream, each detection reflects a local shift over the surrounding ∼32-item neighborhood rather than a single item. Overall, the sparse detections and the absence of sustained shifts in the rolling trajectories suggest broadly stable annotation behavior over the run, with only occasional local fluctuations in pairwise disagreement.

G.4 Match vs. Miss by subcategory

In this section, we quantify subcategory-level rationale overlap in cases where GPT agrees with the human consensus label (Applicable / Not Applicable) for the given dimension. For each query, each human provides one best-fit subcategory; the human rationale set is therefore either a singleton (both humans chose the same subcategory) or a size-two set (humans chose different subcategories). We then compare this human set to GPT's provided subcategories as rationales for labeling the query as Applicable or not. For each human-selected subcategory occurrence, we count a **Match** if GPT includes that same subcategory, and a **Miss** otherwise. Figure 9 reports, for each subcategory, the % Match vs. % Miss, along with the total number of such occurrences (N).

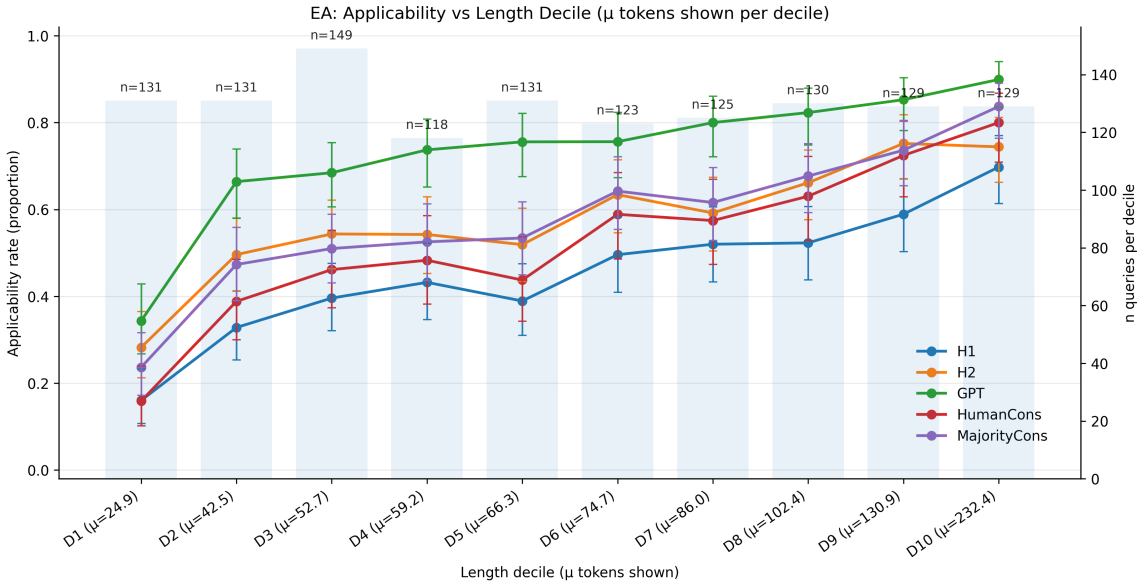
Across both EA and IA, match rates are generally high, indicating that when GPT agrees with humans on the binary applicability label, it often also identifies the same underlying rationale cues. However, several subcategories exhibit higher miss rates, suggesting systematic differences in how GPT justifies an agreed-upon label.

For **EA** (Figure 9a), GPT shows near-complete alignment for *Concern for Relations* ($N = 204$) and *Inferred Negative Emotion* ($N = 102$), and very strong alignment for *Neutral Diagnosis Request* ($N = 92$) and *Symptom Seriousness* ($N = 70$). In contrast, *Hypothetical Medical Query* exhibits the largest miss portion ($N = 39$), and both *Severe Negative Emotion* ($N = 189$) and *Factual Medical Query* ($N = 147$) show notable misses, indicating that GPT sometimes agrees on EA applicability while grounding its justification in different cues than the humans.

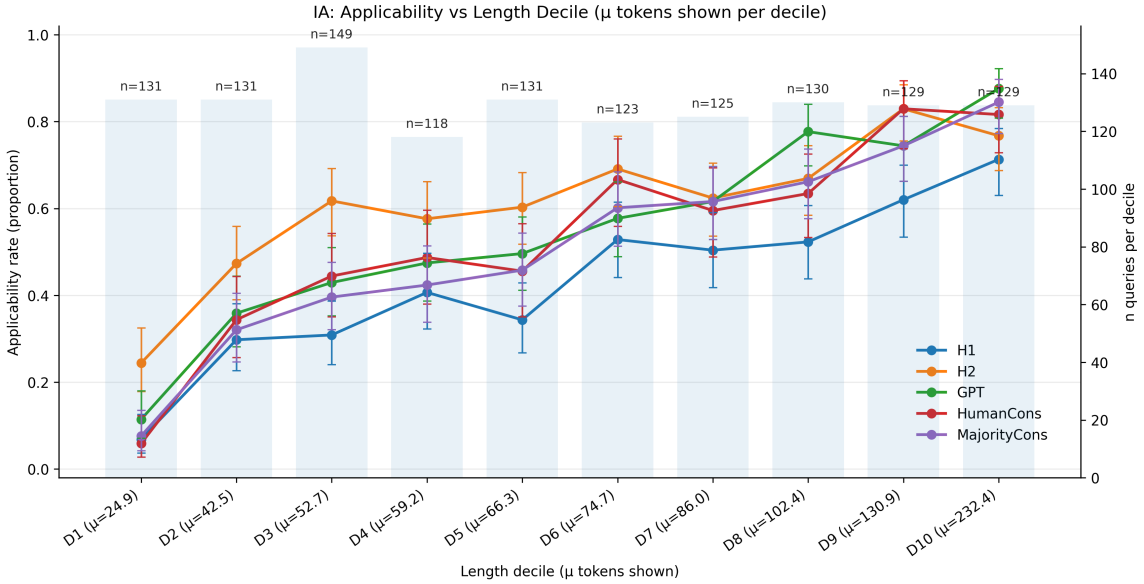
For **IA** (Figure 9b), alignment is strongest for *Distressing Uncertainty* ($N = 262$) and *Feelings Expression* ($N = 233$), and is also high for *Neutral Diagnosis Request* ($N = 133$). *Impact on Daily Life* shows a perfect match in this subset ($N = 49$), though this category is comparatively smaller. The most challenging IA category is again *Hypothetical Medical Query* ($N = 58$), which has the largest miss fraction; *Straightforward Medical Query* ($N = 183$) also shows a higher miss rate than the more affective/uncertainty categories.

Together, these patterns indicate substantial ra-

tionale overlap when GPT agrees with the human consensus applicability label, but alignment varies by subcategory. Misses occur not only for hypothetical and some informational categories, but also for certain affective categories (e.g., *Severe Negative Emotion*). This suggests that mismatches reflect less a single cue-type split and more the prevalence of multi-cue queries, where multiple plausible rationales can support the same applicability judgment.

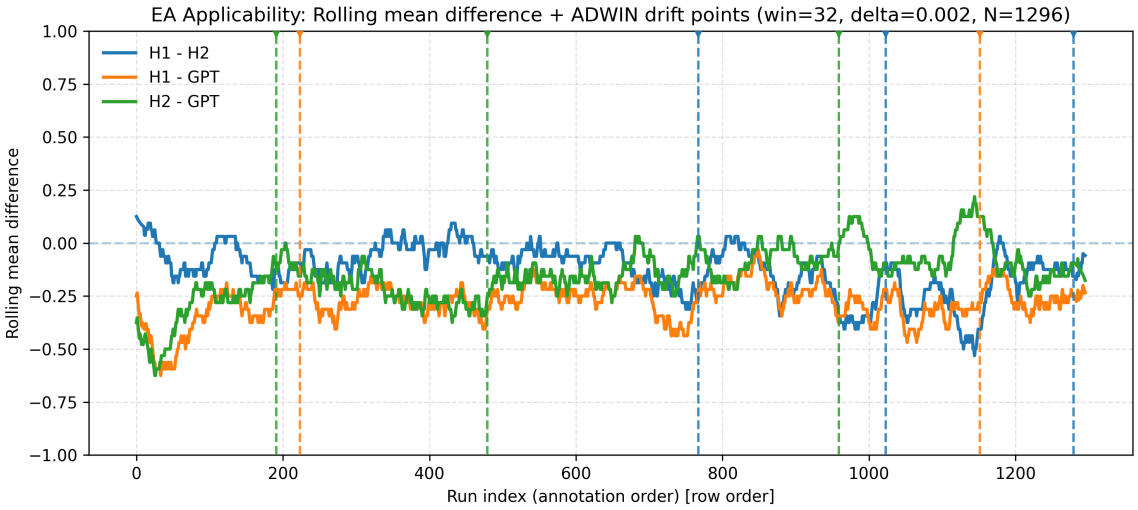


(a) EA: Applicability vs. token-length decile (μ tokens shown per decile; error bars are confidence intervals).

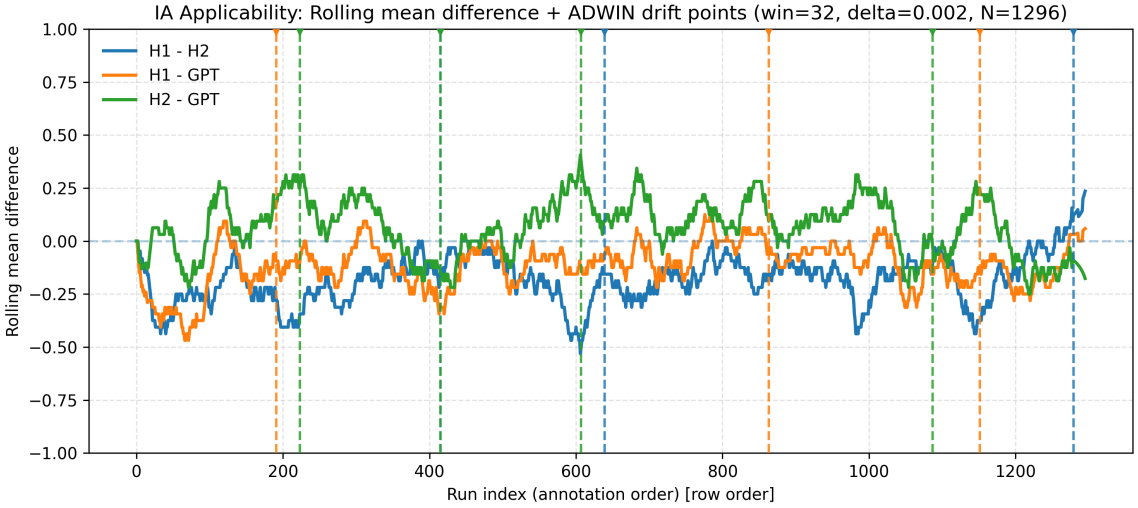


(b) IA: Applicability vs. token-length decile (μ tokens shown per decile; error bars are confidence intervals).

Figure 7: Applicability rates by query-length decile. Bars indicate the number of queries per decile (n). Error bars denote confidence intervals.



(a) EA: rolling mean pairwise difference + ADWIN drift points (win=32, $\delta = 0.002$).



(b) IA: rolling mean pairwise difference + ADWIN drift points (win=32, $\delta = 0.002$).

Figure 8: Run-order drift diagnostics using rolling mean differences in applicability rates and ADWIN change detection. Vertical dashed lines indicate detected change points.

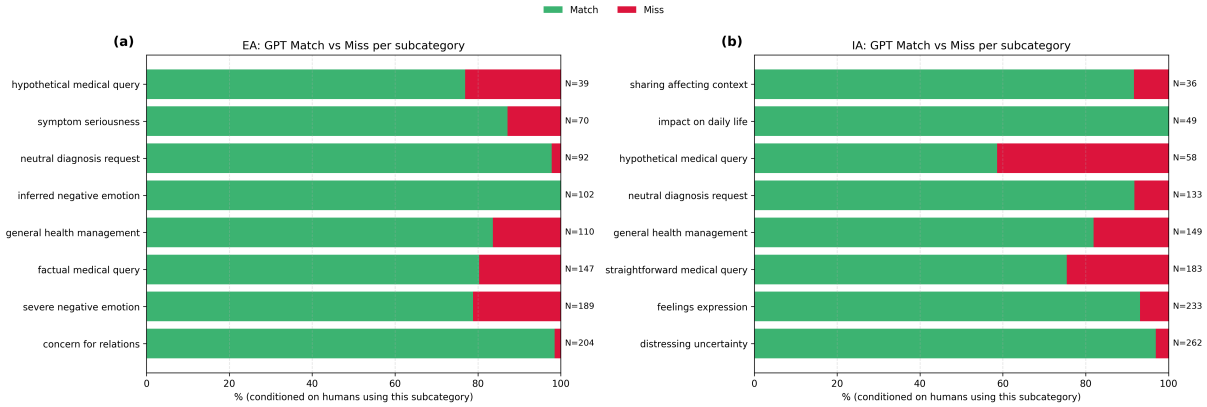


Figure 9: **Rationale Match vs. Miss by subcategory** Stacked bars show, for each subcategory, the percentage of **Match** (green) vs. **Miss** (red) between GPT and the human-selected subcategory rationale, computed only on queries where GPT agrees with the human consensus applicability label. The N label denotes the number of times the subcategory appears in the human rationale set within this agreement subset. Panels show **(a)** EA and **(b)** IA.