# 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 comprises three components: a cognitive component for understanding the patient's emotional and psychological state; an emotional component to resonate with the patient's feelings; and an action-oriented component to express this understanding through verbal and non-verbal cues (Guidi and Traversa, 2021). It is indispensable for clinical care, deepening therapeutic relationships and improving outcomes such as patient satisfaction, care effectiveness, and hospital length of stay (Guidi and Traversa, 2021). Research demonstrates empathy's clinical value through improved patient outcomes and reduced distress, yet clinicians miss 90% of empathic opportunities during patient interactions (Olson, 1995; Hoffstädt et al., 2020; 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 physician-patient interactions (Koranteng et al., 2023). This gap prompts an urgent question: How can we assess and improve LLMs' ability to convey empathy in general healthcare settings, particularly in drafting asynchronous empathetic responses?

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 reply. Lahnala et al. extend this line of work 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 these gaps, we propose the Empathy Applicability Framework (EAF), a theoretically grounded method to proactively identify when and how 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. This anticipatory approach enables providers and LLMs to better detect empathy opportunities for general health queries, potentially improving patient-provider communication.

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 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), iteratively refined to comprehensively and distinctly capture empathy applicability.

Additionally, we ground EAF cues in Patient-Centred Care (PCC) functions (Epstein and Street Jr, 2007) to ensure their alignment with clinically valid expressions of empathy. PCC's *Responding to Emotions* function — particularly the *reassurance* domain (McCormack et al., 2011) — is embodied in EAF's Emotional Reaction applicability cues, which capture both implicit and explicit expressions of distress, such as *Concern for Relations* and *Severe negative Emotion*. PCC's *emotion-validation domain* is similarly reflected in EAF's Interpretation applicability cues, including *Expression of feeling*. Finally, the PCC's *Managing uncertainty* function is represented in Interpretation applicability cues that address distressing uncertainty, the emotional impact of symptoms, and context sharing. By anchoring EAF's applicability cues in these PCC functions, we reinforce its foundation in patient-centered empathy and generate theory-informed signals that can be detected by language models.

## 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, 4,750 each from HealthCareMagic ($\approx 100k$ dialogues) and iCliniq ($\approx 10k$), both publicly released by Li et al. (Li et al., 2023), to maximize linguistic and contextual diversity and avoid overfitting to a single source. Because these anonymized datasets are publicly available, our Institutional Review Board determined that this research does not meet the criteria for human subjects research requiring IRB approval. The datasets carry no explicit licence; we therefore use them exclusively for non-commercial research, in line with the authors'

| Dimensions | Applicable cues | Not Applicable cues |
|---|---|---|
| **Emotional Reactions**<br>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. | • **Severe Negative Emotion**: Explicit expression of distress (e.g., "I'm terrified").<br>• **Inferred Negative State**: Implied distress via anxious or urgent language (e.g., repeated inquiries).<br>• **Seriousness of Symptoms**: Inherently distressing serious or life-threatening conditions requiring reassurance.<br>• **Concern for Relations**: Heightened concern for close relations necessitating emotional support.<br>*Rationale:* Signals reflect distinct pathways of emotional distress, guiding when emotional reactions are warranted. | • **Routine Health Management**: General health advice, routine follow-ups, or management recommendations without emotional involvement.<br>• **Purely Factual Medical Queries**: Medical term definitions, clarification of diagnostic procedures, or explanations of specific medical facts.<br>• **Neutral Symptom Descriptions**: Non-emotional diagnosis requests or symptom descriptions.<br>• **Hypothetical Queries**: Hypothetical medical scenarios without emotional urgency or concern.<br>*Rationale:* Signals no emotional content; omit reactions to maintain factual medical focus. |
| **Interpretations**<br>Communication of an understanding of the patient's feelings (expressed or implied) and/or experiences (including contextual factors) inferred from the patient's query. It involves recognizing and articulating what the patient is feeling and why, based on their situation, concerns, and history. | • **Expression of Feeling**: explicit or implied emotional distress (e.g., frustration, anxiety, hopelessness)<br>• **Experiences or Context Affecting Emotional State**: Non-medical social, environmental, or personal situations affecting emotional state, such as financial difficulties.<br>• **Symptoms with an Emotional Impact**: Symptoms affecting emotional well-being or daily life, with distress conveyed.<br>• **Distressing Uncertainty About Health**: Uncertainty, confusion, or mistrust about health, treatment, or future suggesting distress.<br>*Rationale:* Signals lived burden, context, or uncertainty requiring interpretive acknowledgment. | • **Emotional-Reactions N/A cues +:** with absence of distressing contextual or experiential details.<br>*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. See Appendix Table 4 for concrete query scenarios illustrating cues usage and EAF operationalization. Additional detailed description of the EAF is provided in the Appendix A.

stated intent 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 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. Recruitment was conducted via a flyer distributed through the lab's WhatsApp group. The flyer outlined the study's objective and indicated a workload of approximately one month. 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 received a lump sum of about US $360, equivalent to a one-month local research assistant salary, suitable to their qualifications and living costs. 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.

#### 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).

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

Following the annotation and modeling processes outlined in the Methods section, we designed evaluations and experiments to assess the reliability of the EAF and identify challenges in its use. This section details the evaluation setup and model training configurations used in our experiments.

### 4.1 Evaluation Setup

Our evaluations address four key aspects: annotation quality, conceptual alignment between annotators and LLMs, predictive performance of classifiers, and analyses of disagreement patterns. Each aspect is described in the following sections.

#### 4.1.1 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 human consensus labels on a subset of queries. This subset included only those where humans reached a clear agreement, allowing us to evaluate GPT performance without confounding disagreement over error or subjectivity.

#### 4.1.2 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 and *Partial* if only one overlaps.

#### 4.1.3 Divergence Bar and Qualitative Analysis

For identifying systematic challenges with the use of EAF, we construct three-way divergence bars (Figure 2) that partition each subcategory into: *Annotator Spread* (one human labeled Applicable, the other Not), *LLM-Adds* (GPT Applicable, humans Not) and *LLM-Omits* (GPT 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.

#### 4.1.4 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

[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.

(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.

### 4.2 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 $2 \times 10^{-5}$ 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.

### 5.1 Reliability of the EAF

We evaluated reliability along three axes: (a) Consistency, the degree of agreement between annotators, typically measured via inter-annotator agreement metrics like Cohen's $\kappa$ (Sun et al., 2025); (b) Predictive validity, whether annotation labels can be reliably learned by models, indicating a systematic signal rather than noise (Buechel et al., 2018; Richie et al., 2022); and (c) Conceptual alignment, evidence that annotators draw on similar rationales when assigning labels, supporting the construct validity (Herrewijnen et al., 2024). The evaluation setup details are in Section 4.1.

| Dimension | Human–Human $\kappa$ (agree / disagree) | Human–GPT $\kappa$ (agree/ disagree) |
|---|---|---|
| EA | 0.521 (981 / 315) | 0.617 (668 / 152) |
| IA | 0.404 (898 / 398) | 0.652 (678 / 142) |

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

**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 $\kappa$ 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.

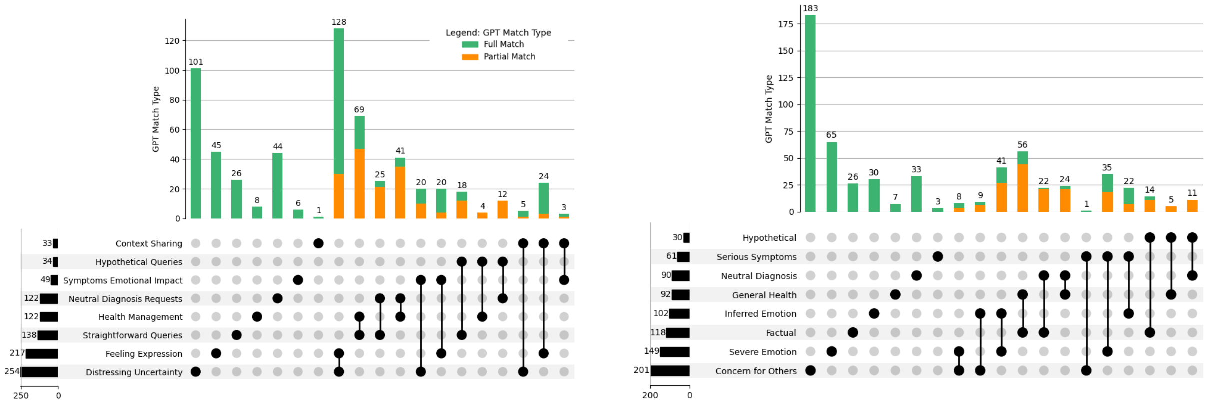
GPT aligned well with the *human consensus dataset*, queries where both humans agreed, achieving three-way agreement. For both EA and IA, Cohen's $\kappa$ exceeded 0.6 and raw agreement was about 80% (Table 2). These results reflect agreement on human-aligned cases, demonstrating EAF's effectiveness in guiding GPT to anticipate empathy applicability in clearer contexts, excluding more ambiguous or complex queries (see section 5.1).

**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, with F1 scores exceeding 90% for EA and approximately 87% for IA. Models trained on GPT-only annotations (the Autonomous set) also performed well, achieving around 85% for EA and 77% for IA. All models significantly outperformed the baselines (random

[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 the training sets and baselines, reported from a single run on the test set. Bold indicates best performance. Models outperform baselines with McNemar's largest $p$-value $\approx 10^{-4}$.

| Training Set | 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** | **0.92** | **0.92** | **0.92** | **0.87** | **0.87** | **0.87** |
| Autonomous | 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 showing agreement between GPT and human annotators for **(a)** IA and **(b)** EA subcategories. Vertical bars represent unique combinations of subcategories matched by GPT, split by match type (green = Full subset match; orange = GPT match only one out of two subcategories stated by humans). Horizontal bars show total frequency of each subcategory across all matches.

guessing, always applicable, always not applicable and o1 Zero-Shot), which yielded accuracies below 62% and substantially lower F1 scores. McNemar's test (McNemar, 1947) confirmed statistical significance over baselines ($p < 10^{-4}$). These results suggest that the 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 are reflected in the green single-dot bars, indicating 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, as shown in multi-dot full-match cases. For example, in queries involving both *Expression of Feeling* and *Distressing Uncertainty*, GPT cited both reasons. This suggests that the model can reconcile diverse human rationales and also underscores the framework's breadth in conceptualizing clinical empathy.

Collectively, these results establish EAF as a reliable framework for capturing clinical empathy in NLP. It supports consistent human judgments, facilitates learnable patterns, and promotes interpretive reasoning across humans and LLMs, making it well suited for anticipatory empathy modeling in the clinical context.

### 5.2 Systematic Challenges in Operationalizing Anticipatory Empathy

Divergence bar analysis (Section 4.1) 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.

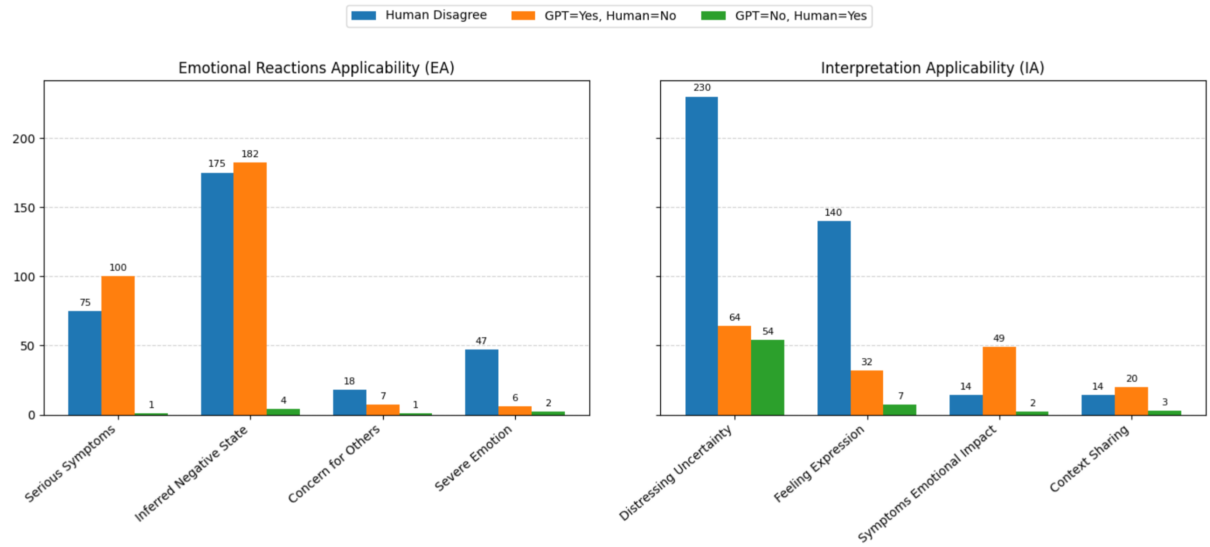



Figure 2: Three-way divergence for every subcategory. Orange = *Annotator Spread in Humans* (One Applicable, other not); Blue = *LLM-Adds Empathy Dimension* (GPT Applicable, Humans Not); Green = *LLM-Omits Empathy Dimension* (GPT Not, Humans Applicable).

#### 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). *Distressing Uncertainty* or confusion, mistrust, or uncertainty about the health condition that leads to emotional distress exhibits the highest variability in inter-human annotation, helping to explain the relatively low Cohen's $\kappa$ score for IA (Table 2) between human annotators. 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. An illustrative example is a query labeled as Distressing Uncertainty by the female annotator: *"Five days ago I started experiencing extreme sharp pain in my back below my rib cage... I have started my menstrual cycle today. Could this have caused such extreme pain?* The male annotator interpreted this as a factual diagnostic request, highlighting how experiential differences shape the interpretation of distress. **Future direction:** Multi-annotator modeling and disagreement-aware approaches (Davani et al., 2022; Gordon et al., 2021) can preserve interpretive diversity.

#### 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 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). **Future direction:** A clinician-in-the-loop annotation pipeline with severity taxonomies and GPT-based verification can calibrate judgments while minimizing expert burden.

#### 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 mismatched labels in SEI category, and all 20 mismatches in CS revealed that while

[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)*

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). **Future direction:** Culturally diverse annotation pools and localized hardship taxonomies can improve cross-cultural empathy modeling, ensuring contextually appropriate responses across patient populations.

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

Conventional reactive empathy models in NLP, post hoc classifiers that label responses (to health queries) as empathetic *after* they are expressed (Sharma et al., 2020; Chai et al., 2019) are misaligned with clinical needs. Even clinicians miss 90% of empathic opportunities, acknowledging only 10% of patient-distress cues during lung cancer visits (Morse et al., 2008; Hsu et al., 2012). Reactive models cannot guide clinicians at the critical moment of deciding how to respond empathetically, making them less equipped for asynchronous patient communication (Antoniak et al., 2024). Although the Empathic Opportunity Perception and Distinction frameworks have recently shown success in synchronous Narrative Medicine by alerting physicians to real-time empathic opportunities (Charon, 2001; Ma et al., 2025), asynchronous contexts such as portal messaging require anticipatory mechanisms. EAF fills this gap by assigning applicability labels to patient queries before response generation, signaling whether empathy is warranted and which dimension (emotional versus interpretive) should be expressed. This proactive approach aligns with recent advances in NLP that demonstrate that empathy ratings improve by inferring users' emotions through cause-aware prompting and predicting psychological needs via the Sibyl paradigm (Chen et al., 2024; Wang et al., 2025).

However, **subjective inference** poses a challenge when affective cues are implicit, as in our Inferred Negative Emotional State and Distressing Uncertainty categories. Annotators rely on personal appraisals to interpret distress, leading to divergent judgments. Appraisal theory formalizes this variability, suggesting emotions toward others depend on individual evaluations of circumstances (Wondra and Ellsworth, 2015). Such disagreements reflect genuine interpretive differences, not noise. The NLP community increasingly embraces this through multi-annotator models treating annotator decisions separately to generate calibrated uncertainty estimates (Davani et al., 2022), or via Annotator-Aware Representations embedding annotators' interpretive styles (Mokhberian et al., 2023). Gordon et al.'s (Gordon et al., 2021) *jury learning* exemplifies this, selecting annotator subsets aligned with demographic perspectives. In clinical empathy contexts, preserving subjective variability helps models anticipate diverse patient needs, e.g., cancer-related queries annotated by oncologists prioritizing emotional support as central to clinical care (Dekker et al., 2020).

This work makes three significant contributions to clinical empathy modeling in NLP. First, we introduce the Empathy Applicability Framework, shifting from reactive to anticipatory empathy modeling, essential in asynchronous communication where clinicians need to proactively craft empathetic responses. Second, we establish a benchmark of 1,300 real patient queries demonstrating reliable and learnable EAF labels, providing foundations for future research. Third, our analysis identifies empathy modeling challenges — subjective inference, clinical-severity ambiguity, and contextual hardship — as opportunities to embrace interpretive pluralism via multi-annotator frameworks and domain-specific perspectives. By offering both a practical framework and empirical insights into operationalization, this work advances clinical empathy modeling that respects interpretive complexity while remaining computationally tractable. The EAF thus represents a key step toward AI systems supporting clinicians in delivering empathetic, patient-centered care across diverse contexts.

## 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.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. *arXiv preprint arXiv:1808.10399*.

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.

Rita Charon. 2001. Narrative medicine: a model for empathy, reflection, profession, and trust. *jama*, 286(15):1897–1902.

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.

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.

Elize Herrewijnen, Dong Nguyen, Floris Bex, and Kees van Deemter. 2024. Human-annotated rationales and explainable text classification: a survey. *Frontiers in Artificial Intelligence*, 7:1260952.

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.

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*.

Hua Ma, Effie Lai-Chong Law, Xu Sun, Weili Yang, Xiangjian He, Glyn Lawson, Huizhong Zheng, Qingfeng Wang, Qiang Li, and Xiaoru Yuan. 2025. Towards empathic medical conversation in narrative medicine: A visualization approach based on intelligence augmentation. *International Journal of Human-Computer Studies*, 199:103506.

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.

Russell Richie, Sachin Grover, and Fuchiang Rich Tsui. 2022. Inter-annotator agreement is not the ceiling of machine learning performance: Evidence from a comprehensive set of simulations. In *Proceedings of the 21st workshop on biomedical language processing*, pages 275–284.

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*.

Yinglun Sun, Jose Zavala, Shuju Shi, Rachel Finegold, Roxana Girju, and Jeffrey Moore. 2025. Medical-care: building and annotating an empathy-rich corpus. *Language Resources and Evaluation*, pages 1–36.

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 understanding 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.

## C Illustrative Scenarios for EAF Operationalization

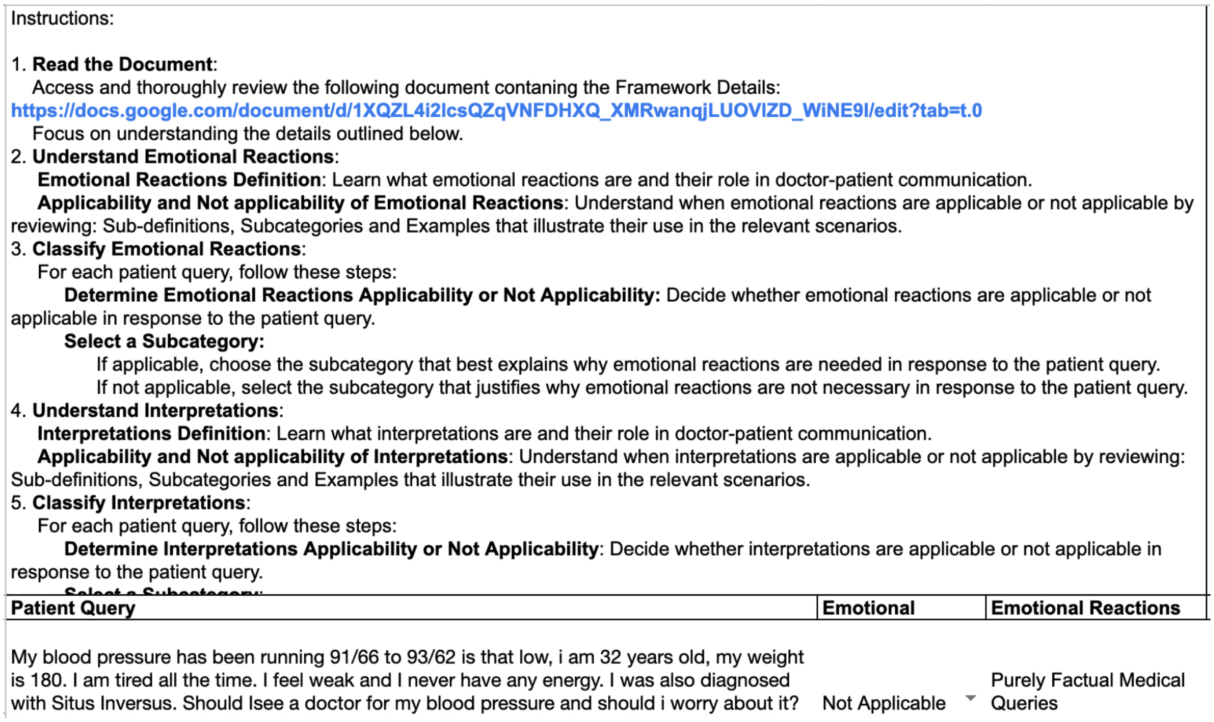
See Table 4 for illustrative scenarios demonstrating the operationalization of the EAF.

## 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 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.



Instructions:

1. **Read the Document**:
   Access and thoroughly review the following document contaning the Framework Details:
https://docs.google.com/document/d/1XQZL4i2lcsQZqVNFDHXQ_XMRwanqjLUOVIZD_WiNE9l/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.

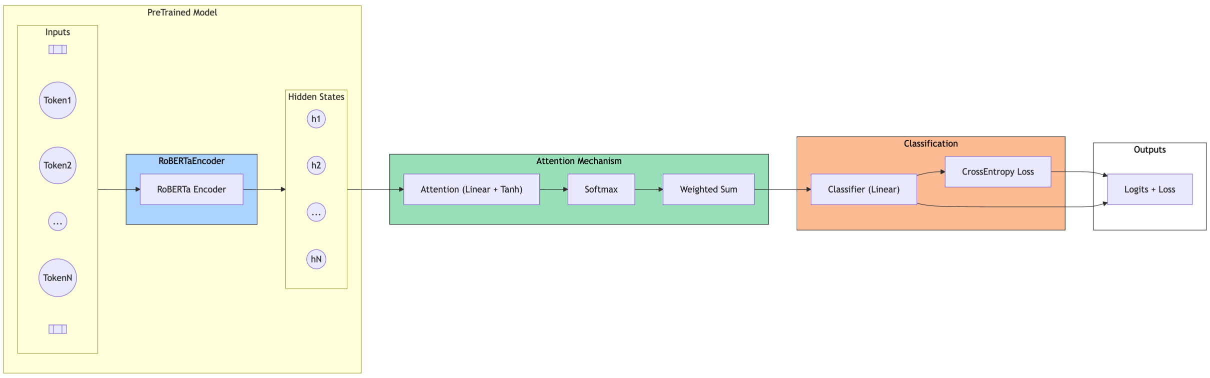



Figure 4: Empathy Dimension Applicability Model Architecture

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

| **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

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

Table 5: Cohen's $\kappa$ 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 |