Webside Manner: A Dataset and Analysis of Empathy Strategies in Verified Clinician Responses

Anonymous ACL submission

Abstract

Online medical advice often combines clinical explanation with empathy-oriented communication, but existing computational empathy frameworks are not tailored to this setting. We study fine-grained empathy strategy detection in verified health-worker responses from *r/AskDocs*. We introduce a domain-adapted multi-label annotation scheme and compare prompting, supervised fine-tuning, and layer-wise probing across multiple language models. Our results show that prompting underperforms probing, suggesting that empathy-relevant information can be recovered from hidden representations even when models do not reliably express it via direct generation. Fine-tuning improves output-level prediction and makes empathy strategies more linearly recoverable, especially for macro-level performance. We further analyze the predicted use of empathy strategies over time, across healthcare-related professions, and across medical subjects. Overall, our findings show that representation-level analysis is a useful complement to output-based modeling for studying empathy in online medical advice. Code and data are availabe.[1]

1 Introduction

Empathy is central to effective communication, especially in healthcare, where empathic responses can build trust and improve outcomes (Decety and Jackson, 2004; Chen et al., 2025a). In online medical advice forums such as *r/AskDocs*, users often ask health questions while experiencing uncertainty, fear, or distress. Verified healthcare workers must therefore do more than provide information. Their responses may also reassure, validate concerns, explain symptoms, ask follow-up questions, or offer concrete guidance. This makes online medical advice an important setting for studying empathy in NLP.

[1] https://anonymous.4open.science/r/webside_manner-C447/

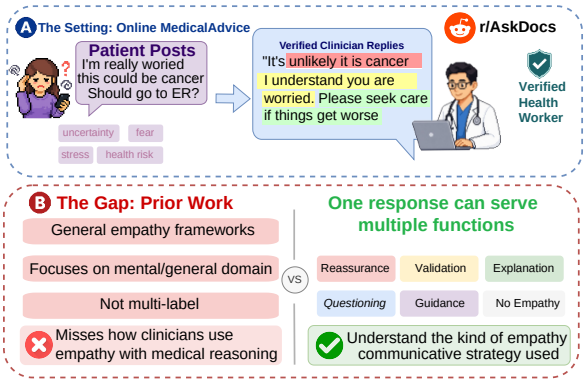

Figure 1: Overview and positioning of our study.

However, there is limited work on how empathy is used by clinicians online. It is not only a matter of whether a response sounds warm or supportive. A clinician may express empathy by reducing uncertainty, explaining a possible cause, normalizing a concern, asking for missing context, or giving concrete next steps. This creates a measurement problem: to study empathy in online medical advice at scale, we need models that can detect not only whether a response is empathic, but what communicative strategy it uses.

Prior work provides useful starting points for this problem. EPITOME defines expressed empathy through Emotional Reactions, Interpretations, and Explorations (Sharma et al., 2020), while recent work such as SENSE-7 proposes user-centered taxonomies for empathy in human–AI interaction (Suh et al., 2025). Other studies have introduced datasets and models for empathetic dialogue and emotional support (Rashkin et al., 2019; Sharma et al., 2020; Chen et al., 2025b; He et al., 2025; Ye et al., 2025). However, most work either treats empathy as a single overall response quality or focuses on mental health and open-domain support. Less is known about fine-grained empathy strategies in online medical advice, where empathic communication is often tied to clinical explanation, uncertainty reduction, and practical guidance.

This paper studies fine-grained empathy strategy detection in verified health-worker responses from Reddit's *r/AskDocs* community (See Figure 1). We formulate the task as multi-label classification using a domain-adapted empathy schema derived from EPITOME and SENSE-7. The schema includes six labels: no identifiable empathy strategy, affective response and reassurance, validation, perspective-taking and explanation, exploratory questioning, and supportive guidance. We frame this task as a measurement prerequisite: before analyzing empathy at scale in online medical advice, it is necessary to define the relevant empathy strategies and evaluate whether models can detect them reliably. We ask:

- **RQ1:** How accurately can language models detect fine-grained empathy strategies in online medical advice?
- **RQ2:** To what extent are empathy strategies linearly recoverable from the hidden representations of base and fine-tuned language models?
- **RQ3:** Given a trained detector, what descriptive patterns emerge in predicted empathy strategies over time, across healthcare professions, and across medical topics in *r/AskDocs*?

To answer these questions, we compare prompting, supervised fine-tuning, and layer-wise probing across multiple language models. Prompting and fine-tuning evaluate whether models can directly produce empathy strategy labels, while probing tests whether these strategies are linearly recoverable from hidden representations under supervised classification. We then apply the strongest detector to the larger verified *r/AskDocs* corpus to study large-scale descriptive patterns.

Our contributions are: **(1.)** We introduce a domain-adapted, multi-label annotation scheme and dataset for measuring empathy strategies in verified health-worker responses from *r/AskDocs*. **(2.)** We evaluate prompting, supervised fine-tuning, and layer-wise probing as alternative measurement approaches, showing that task-specific supervision substantially improves detection and that empathy strategies are more recoverable from hidden representations than from direct prompting. **(3.)** We use the strongest detector as a descriptive measurement tool to analyze large-scale variation in predicted empathy strategies over time, across healthcare-related professions, and across medical subjects.

2 Related Work

Online Medical Advice on Social Media. Online health communities provide large-scale records of how people seek medical information, describe symptoms, and receive support outside traditional clinical encounters (Eysenbach et al., 2004; Moorhead et al., 2013; De Choudhury and De, 2014). Compared with private clinical communication, social media health discourse is public, asynchronous, and text-based, producing conversational traces while also raising concerns about information quality, context loss, uneven expertise, and privacy (Moorhead et al., 2013; De Choudhury and De, 2014; Wadhwa et al., 2023). Reddit has been especially useful for health NLP because it contains topic-specific communities, conversational context, and time-stamped interactions. Prior work has used Reddit and related platforms to study mental-health disclosure and social support (De Choudhury and De, 2014; Cohan et al., 2018; MacAvaney et al., 2018), public-health and symptom surveillance (Paul and Dredze, 2021; Coppersmith et al., 2014), medical questions, personal experiences, health-related claims, and evidence retrieval (Wadhwa et al., 2023), and health misinformation (Broniatowski et al., 2018). Ask-the-doctor communities are especially relevant because users receive responses from peers and healthcare professionals (Nobles et al., 2020; Ayers et al., 2023). However, most computational work in this space focuses on medical content, answer quality, information reliability, or verification. We instead study the interpersonal strategies verified health workers use when combining medical advice with reassurance, validation, explanation, questioning, and guidance.

Computational Empathy. Computational empathy research has increasingly moved beyond treating empathy as a single affective score. Early work modeled empathic concern and personal distress in text (Buechel et al., 2018) and introduced datasets for open-domain empathetic response generation (Rashkin et al., 2019). Later work showed that empathy is not equivalent to sentiment or emotion, since empathic communication also involves understanding, validation, social appropriateness, and supportive action (Lahnala et al., 2022). Fine-grained frameworks such as EPITOME decompose expressed empathy into emotional reactions, interpretations, and explorations (Sharma et al., 2020), while other work studies empathetic response intents, figurative language, user-centered empathy

taxonomies, emotion-cause-aware response generation, and personalized emotional support (Welivita and Pu, 2020; Lee et al., 2024; Suh et al., 2025; He et al., 2025; Ye et al., 2025). Relatedly, Zhan et al. (2026) shows that LLMs can produce highly rated empathetic responses while relying on repetitive discourse moves, suggesting that empathy should be evaluated beyond surface warmth. We build on these frameworks, especially EPITOME and SENSE-7, but adapt them to online medical advice, where empathy is often expressed through clinically grounded explanation, uncertainty reduction, follow-up questioning, and concrete next steps.

Measuring Communicative Strategies at Scale. Large online communities make it possible to study how language and interaction patterns vary over time and across groups (Danescu-Niculescu-Mizil et al., 2013; Kulkarni et al., 2015; Hamilton et al., 2016; Del Tredici and Fernández, 2017; Mensah et al., 2020; Jaidka et al., 2018). Prior work has used such data to analyze semantic change, community evolution, sentiment, topics, and other forms of linguistic variation (Nguyen et al., 2016; Del Tredici and Fernández, 2017; Jaidka et al., 2018; Mensah et al., 2020; Stewart and Eisenstein, 2018; Zhang et al., 2017; Zhu and Jurgens, 2021). Our setting requires measuring a different kind of variation: the interpersonal strategies clinicians use when responding to health concerns. These strategies are not directly observed, so model-based detection becomes a measurement step before any large-scale analysis. We therefore first evaluate prompting, fine-tuning, and probing as detectors of empathy strategies, and only then use the strongest model to describe variation over time, across healthcare-related professions, and across medical subjects. These analyses are descriptive rather than causal, especially as LLMs become more common in emotionally sensitive and health-related communication (Ayers et al., 2023; Chen et al., 2025a; McBain et al., 2025; de Wynter, 2025; De Freitas et al., 2026; Zhan et al., 2026).

3 Data

Data Collection. The data come from the Reddit subreddit *r/AskDocs*, a medical advice forum where users post health-related questions and self-identified healthcare professionals respond. The raw data consists of 960,831 comments. The date ranges from Jul 2013 to Dec 2024. We removed deleted posts and comments, and dropped all comments authored by the "AutoModerator" bot. To focus on medically reliable responses, we filter by the author's verification status provided in the subreddit metadata and exclude the non-professionals. After this filtering, the resulting corpus contains 443,699 comments. Detailed descriptive statistics are provided in Table 1.

Corpus / label support	N	Posts		Subm.	
		Avg	Med	Avg	Med
All Data	443,699	45.0	28.0	200.9	156.0
Human-labeled AskDocs subset by empathy strategy					
NE	379	15.0	11.0	150.8	116.0
ER	471	59.1	37.0	198.6	155.0
VN	113	103.8	74.0	316.3	216.0
PE	1,105	66.4	48.0	212.3	161.0
EQ	384	46.6	35.0	205.8	153.0
SG	630	68.3	49.0	222.6	170.5

Table 1: Dataset statistics. "Submission" refers to the original Reddit post, computed as title plus body; "post" refers to the verified AskDocs comment. For the human-labeled subset, N denotes the multilabel support for each empathy strategy, rather than the mutually exclusive class counts. NE = no empathy, ER = emotional reaction/reassurance, VN = validation/normalization, PE = perspective-taking/explanation, EQ = exploratory questioning, and SG = supportive guidance.

Annotation Scheme and Task Description. We develop a domain-adapted annotation scheme for empathy strategies in online medical advice. The scheme draws on prior computational empathy frameworks, especially EPITOME's categories of emotional reaction, interpretation, and exploration (Sharma et al., 2020), and adapts them to clinical and online health-support settings, where empathy often co-occurs with reassurance, validation, explanation, questioning, and guidance (Epstein et al., 2007; Hashim, 2017; Tietbohl, 2022; Biyani et al., 2014). Detailed label definitions, decision rules, examples, and a picture of our annotation platform are provided in Appendix A.

Our final schema contains six response-level labels: **NE** for no identifiable empathy strategy, **ER** for emotional reaction and reassurance, **VN** for validation and normalization, **PE** for perspective-taking and explanation, **EQ** for exploratory questioning, and **SG** for supportive guidance. Because a single clinician's response may involve multiple strategies, we formulate the task as a multi-label classification problem. Given a verified health worker's comment c replying to a patient post p on *r/AskDocs*, the task is to predict a label set $Y \subseteq \{\text{NE}, \text{ER}, \text{VN}, \text{PE}, \text{EQ}, \text{SG}\}$. **NE** is mutually

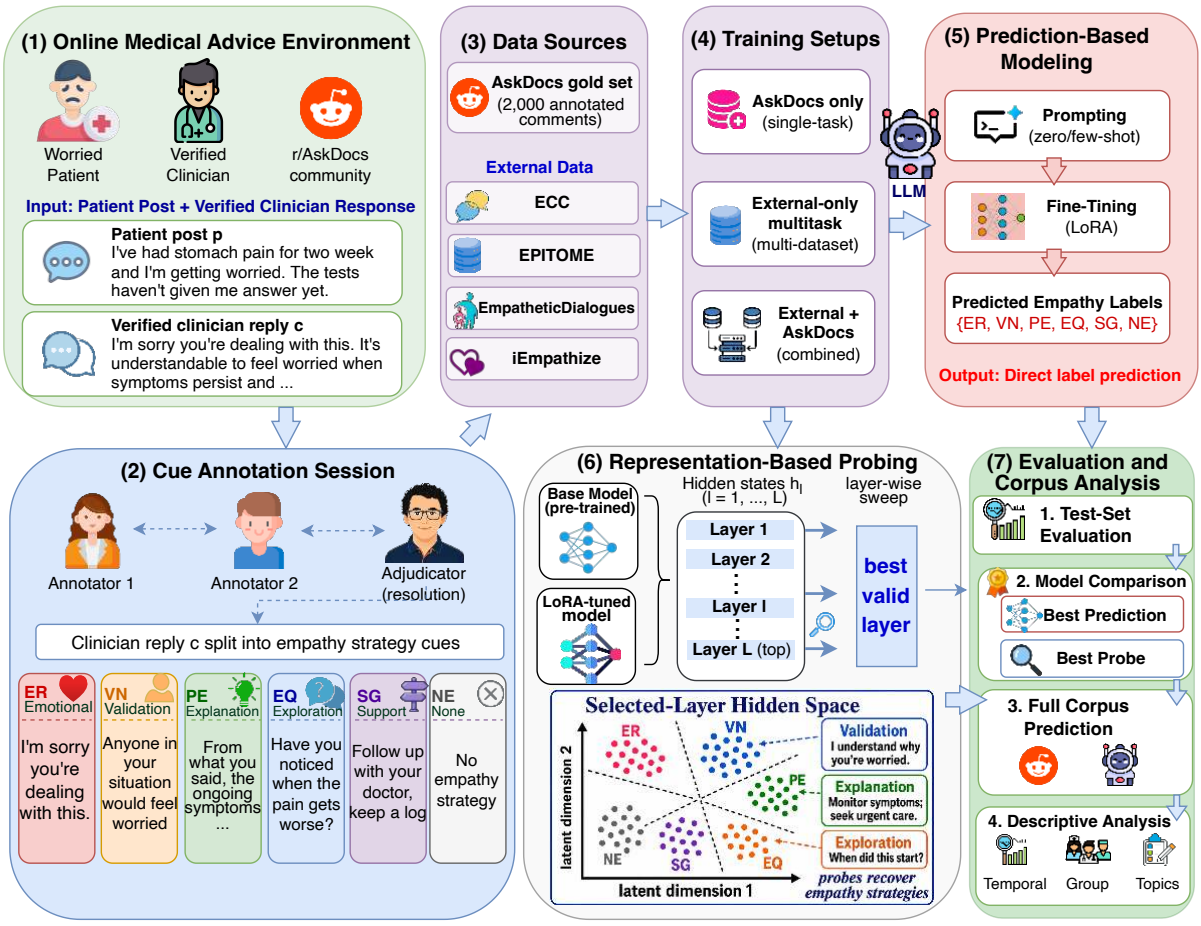

Figure 2: Overview of our framework for empathy strategy detection in online medical advice.

exclusive with the empathy-strategy labels, while ER, VN, PE, EQ, and SG may co-occur.

Two annotators independently labeled a subset of 2,000 comments following the annotation guidelines. They used a custom annotation software shown in Figure A.1 of the Appendix. We treat each annotation as a set of labels and measure inter-annotator agreement using Krippendorff's α with Jaccard distance between label sets (Krippendorff, 2019; Passonneau, 2006). After guideline refinement and calibration, the two annotators achieved substantial agreement ($\alpha = 0.81$). Disagreements were subsequently adjudicated by a third annotator to produce the final gold labels used in our experiments.

4 Experiments Settings

Baselines. We compare against four categories of baselines. 1) Simple label-distribution baselines to establish lower-bound performance: a stratified one-vs-rest dummy classifier, a most-frequent one-vs-rest dummy classifier, and an always-NE baseline that predicts no identifiable empathy strategy for every comment. These baselines help assess whether models learn beyond label imbalance. 2) Classical lexical classifiers using TF-IDF features, including logistic regression and LinearSVC with grid search. These models provide strong non-neural baselines for determining how much of the task can be solved from surface lexical cues alone. 3) Fine-tuned pretrained encoder models, including BERT-base (Devlin et al., 2019) and RoBERTa-base (Liu et al., 2019), as supervised multi-label classifiers. 4) GPT-4o-mini (OpenAI, 2022) as a prompted LLM baseline.

Models and Implementation Details. We evaluate five open-weight LLM backbones: Qwen2.5-7B, Qwen2.5-14B, Qwen3-8B, Gemma-3-4B, and Gemma-3-12B (Qwen Team, 2025; Yang et al., 2025; Gemma Team et al., 2025). For each model, we compare output-level predictions via prompting, LoRA-supervised fine-tuning (Hu et al., 2022), and linear probes trained on hidden-state representations extracted from the models. Prompting tests whether models can directly apply our empathy-strategy schema from instructions alone, while LoRA fine-tuning tests whether task-specific supervision improves generated label predictions.

Probing evaluates whether empathy-strategy information is linearly recoverable from model representations, both before and after fine-tuning; for each probing setting, we report test performance using the validation-selected layer. Intuitively, we hope to understand whether LLMs also represent empathy concepts in their hidden states that can be predicted, and how these representations change with fine-tuning.

We conduct LLM experiments using Hugging Face `transformers` and `peft` for LoRA fine-tuning. The encoder and classical baselines are implemented using Hugging Face and scikit-learn. We report micro-F1 and macro-F1 for all models. Additional details, including model checkpoints, LoRA fine-tuning, hidden-state extraction, probe training, hyperparameters, and prompt templates are provided in Appendices B and D.

To examine whether broader empathy supervision transfers to online medical advice, we also fine-tune selected models under two multitask settings. We use four external empathy-related datasets: EmpatheticDialogues (Rashkin et al., 2019), EPITOME (Sharma et al., 2020), IEMPATHIZE (Hosseini and Caragea, 2021), and ECC (He et al., 2025). In the *external-only* setting, models are trained only on these external datasets; in the *external+AskDocs* setting, these datasets are combined with the AskDocs training split. These variants allow us to test whether general empathy datasets provide useful signal for our domain, either alone or as complementary supervision. Details of each dataset and prompt templates are provided in Appendix D.

5 Results

Comparisons with baselines. Table 2 reports the performance of the baselines. The dummy baselines (random and most-frequent label prediction) perform poorly, indicating that the task cannot be solved by exploiting label priors alone. Classical TF-IDF classifiers improve over these lower-bound baselines, but remain substantially below supervised neural models. Among the baseline methods, RoBERTa-base achieves the strongest performance, with .7629 micro-F1 and .6812 macro-F1, followed closely by BERT-base. This suggests that task-specific supervised training is important for detecting fine-grained empathy strategies in online medical advice. The prompted GPT-4o-mini baseline performs better than classical lexical models in macro-F1, but remains below fine-tuned encoder models, indicating that general instruction-following ability alone is insufficient for reliably applying our domain-adapted empathy schema.

Method	miF1	mF1
Always S0	.1427	.0514
Dummy OvR (most frequent)	.4242	.1172
Dummy OvR (stratified)	.3428	.2133
Logistic Regression	.5379	.4456
LinearSVC	.5413	.4609
GPT-4o-mini Prompt	.5641	.5506
Gemma-3-12B Prompt	.6558	.5941
BERT-base FT	.7535	.6713
RoBERTa-base FT	.7629	.6812
Main Methods		
Gemma-3-12B Base Probe	.7633	.6939
Qwen3-8B LoRA Probe	.7816	.7278
Qwen2.5-14B LoRA	**.7941**	**.7404**

Table 2: Comparison with baseline models on the AskDocs empathy strategy test set. For main methods, we report the best-performing AskDocs-only model.

Output- and Representation-Level Comparison. Table 3 compares output-level prediction, through prompting and LoRA fine-tuning, with representation-level prediction, through Base Probe and LoRA Probe. Before fine-tuning, Base Probe substantially outperforms Prompt across model families, with macro-F1 gains ranging from 8.5 points for Qwen3-8B to 20.9 points for Gemma-3-4B and an average gain of 12.9 points. For example, Gemma-3-4B improves from .4678 macro-F1 with prompting to .6763 with Base Probe, while Qwen2.5-7B improves from .5428 to .6824. This suggests that empathy-strategy information is already encoded in model representations, even when models do not reliably express it through direct prompted prediction.

After LoRA fine-tuning, the output–representation gap largely disappears. LoRA and LoRA Probe achieve closely matched macro-F1 across model families, with an average absolute difference below one point. For instance, Gemma-3-12B obtains .6966 macro-F1 with LoRA and .6969 with LoRA Probe, while Qwen3-8B obtains .7185 with LoRA and .7278 with LoRA Probe. *This suggests that fine-tuning aligns the model's output behavior with empathy-relevant distinctions that are already separable in its internal representations.*

Fine-tuning also improves representation-level recoverability. LoRA Probe outperforms Base

Method	Qwen2.5-7B		Gemma-3-4B		Gemma-3-12B		Qwen3-8B		Qwen2.5-14B	
	miF1	mF1	miF1	mF1	miF1	mF1	miF1	mF1	miF1	mF1
Panel A: AskDocs-only comparison										
Prompt	.5963	.5428	.5921	.4678	.6558	.5941	.6108	.5428	.6092	.5652
Base Probe	.7483	.6824	.7422	.6763	.7633	.6939	.7074	.6277	.7560	.6763
LoRA	**.7801**	**.7015**	**.7871**	.7017	**.7827**	.6966	**.7913**	.7185	**.7941**	**.7404**
LoRA Probe	.7781	.6923	.7807	**.7144**	.7768	**.6969**	.7816	**.7278**	.7815	.7229
Panel B: Multitask fine-tuning variants										
External-only LoRA	.4491	.4629	.6073	.5192	.5898	.5100	.5835	.5511	.4492	.4442
External-only LoRA Probe	.7625	.6872	.7602	.6718	.7611	.6881	.7577	.6805	.7759	.7144
External+AskDocs LoRA	.7705	.6946	.7820	.6924	**.7934**	**.7464**	**.7952**	**.7257**	**.7894**	**.7346**
External+AskDocs LoRA Probe	**.7827**	**.7186**	**.7981**	**.7313**	.7802	.7161	.7756	.7094	.7879	.7263

Table 3: Main results on the AskDocs empathy strategy test set. Panel A compares prompting, hidden-state probes, AskDocs-only LoRA tuning, and probes over LoRA-tuned states. Panel B reports multitask variants using external empathy datasets. Bold marks the best comparable result within each model family and panel.

Probe for every model, with the largest gain for Qwen3-8B, where macro-F1 increases from .6277 to .7278. The strongest result is achieved by Qwen2.5-14B LoRA, which reaches .7941 micro-F1 and .7404 macro-F1. Overall, probing reveals empathy-related information that prompting does not fully access, while fine-tuning narrows the gap between internal representational capacity and structured output behavior.

Effects of Training Data Composition. We next test whether external empathy-related supervision improves AskDocs empathy strategy classification, reporting these results separately in Panel B of Table 3. External-only LoRA transfers weakly at the output level and performs substantially below AskDocs-only LoRA across models. This suggests that general empathy datasets do not by themselves capture the domain- and taxonomy-specific distinctions needed for online medical advice, where models must map empathic behavior onto categories such as reassurance, validation, questioning, and guidance.

The picture is different at the representation level. Probes trained on hidden states from external-only fine-tuned models substantially outperform the corresponding generated LoRA outputs. This is not zero-shot transfer, since the probes still use in-domain AskDocs labels. Instead, it suggests that external empathy supervision can induce partially transferable internal representations, even when generated labels remain misaligned with the AskDocs taxonomy.

Combining external data with AskDocs supervision gives the strongest multitask results. External+AskDocs LoRA improves over external-only LoRA for all models, with the best output-level result from Gemma-3-12B (.7934 micro-F1, .7464 macro-F1), while the strongest probe result comes from Gemma-3-4B External+AskDocs LoRA Probe (.7981 micro-F1, .7313 macro-F1). Overall, external empathy data is most useful when anchored by in-domain examples, and its benefits may appear either in output predictions or in the separability of hidden representations.

Label	LoRA Output			LoRA Probe		
	P	R	F1	P	R	F1
NE	.60	.55	.57	.66	.52	.58
ER	.73	.76	.74	.77	.73	.75
VN	.73	.30	.42	.50	.30	.37
PE	.84	.92	.88	.82	.90	.86
EQ	.90	.96	.93	.94	.95	.94
SG	.72	.84	.77	.77	.83	.80

Table 4: Per-label performance of the Qwen3-8B LoRA output model and the Qwen3-8B LoRA probe.

Error Analysis. For error analysis, we compare the output-level (fine-tuned) and representation-level methods (probing). Both models are based on Qwen3-8B with similar patterns. Table 4 shows substantial variation across empathy categories, with VN being the most difficult label. Although VN has relatively high precision (.73), its recall is much lower (.30), yielding an F1 score of .42. Among 27 gold VN instances, the model identifies only 8 and misses 19. Most of these false negatives are still assigned to related categories, where 16 are predicted as PE and 7 as ER. This suggests that validation and normalization are often subsumed by medical explanation or reassurance, especially when normalizing language is embedded in clinical reasoning. The low recall may also reflect VN's

smaller support in the annotated data and its conceptual overlap with reassurance and perspective-taking. In one test-set case where the poster expressed fear about colorectal cancer, the clinician replied: *"A lot of people seem to freak out about colon cancer today. Your symptoms do not sound worrisome, and colon cancer is extremely rare at your age."* The gold labels are ER, VN, and PE, but the model predicts only ER and PE. The model captures reassurance through the low-risk framing and explanation through the age-based medical reasoning, but misses VN, which is expressed by normalizing the poster's worry as a common concern. This example suggests that the model detects more explicit reassuring and explanatory functions while missing subtler normalization when multiple empathy strategies appear in the same response.

The model also performs relatively weakly on NE, with a precision of .60, a recall of .55, and an F1 of .57. One common error is that the model overgeneralizes from question-like or information-seeking language. For example, in one test-set case, the clinician wrote: *"If you could condense your question with data into 2–3 sentences, I'd love to help."* The gold label is NE, but the model predicts EQ. Although the response requests additional input, it does not ask for clinically relevant context, explore the poster's concern, or acknowledge the poster's situation. Instead, it serves primarily as a procedural request that the poster reformat the question. A similar pattern appears in rhetorical questions that challenge the poster's judgment rather than invite elaboration. These errors suggest that the model sometimes treats surface-level question forms as exploratory empathy, even when the pragmatic function is not patient-centered.

6 Analysis

To better understand how empathy strategies appear in large-scale verified *r/AskDocs* responses, we apply the best-performing output-level model to the verified corpus and aggregate the predicted labels by time and professional group. Since these analyses are based on model predictions rather than human annotations, we interpret the results as descriptive patterns of predicted use of empathy strategies, not as direct measures of individual clinicians' empathy.

Temporal Trends. Figure 3 shows 12-month rolling averages of predicted empathy strategies in verified AskDocs responses. PE is the most common strategy throughout the period, staying near 50–60%, which suggests that verified clinicians most often express empathy through explanation and perspective-taking. SG is the next most common strategy, but it gradually declines after the mid-2010s. ER remains relatively stable with a small increase over time, while VN stays rare across all years. EQ declines from the early period and remains lower in later years. NE is shown in the figure for comparison and increases in later years, suggesting more responses are predicted as neutral or non-empathic.

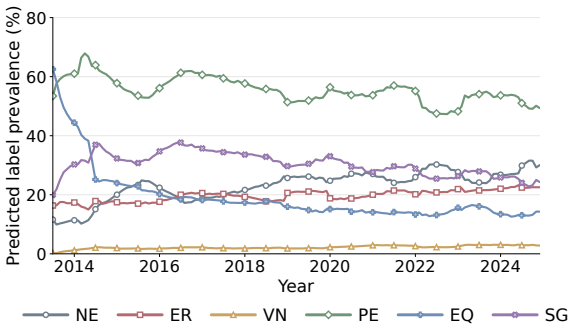

Figure 3: Temporal trends in predicted empathy strategy prevalence in verified AskDocs responses. Lines show 12-month rolling averages of monthly predicted prevalence; NE is omitted. ER = emotional reaction/reassurance, VN = validation/normalization, PE = perspective-taking/explanation, EQ = exploratory questioning, and SG = supportive guidance.

We also run an exploratory interrupted time-series analysis using December 2022, the public release of ChatGPT, as the breakpoint (OpenAI, 2022). This test is descriptive, not causal, because we do not observe clinician AI use and do not have an untreated control group. Most labels do not show a clear break at this point. The main exception is EQ, which increases immediately after December 2022 by 3.87 percentage points ($p < .001$), followed by a small negative slope change (-0.10 percentage points per month, $p = .036$). This suggests a short-term increase in exploratory questioning, rather than a sustained post-ChatGPT increase. Full results are reported in Appendix C.2.

Professional Variation. Figure 4 shows predicted empathy strategy prevalence across verified professional groups. The main pattern is that physician responses have the highest predicted NE rate and comparatively lower rates of several empathy-strategy labels because 25.3% of physician responses are predicted as NE, compared with 17.9% among non-physician verified responders. Physicians also show lower predicted rates of EQ and

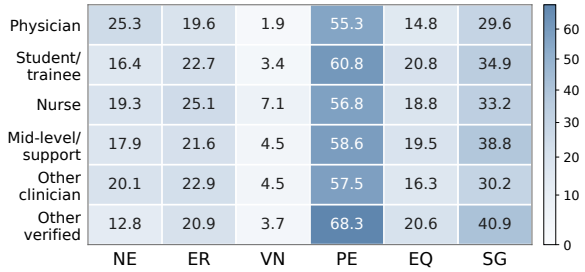

Figure 4: Predicted empathy strategy prevalence by professional group, shown as the percentage of verified responses assigned to each label.

SG than groups such as students/trainees and mid-level or support roles. In contrast, nurses have the highest predicted ER and VN rates, suggesting more frequent use of reassurance, emotional acknowledgment, and normalization. Mid-level and support roles show the highest SG rate, while student/trainee responses show relatively high rates of PE, EQ, and SG.

These are just observations. Professional groups may respond to different types of posts, and topic, severity, urgency, and diagnostic complexity may shape which empathy strategies are appropriate or feasible (Hashim, 2017; Biyani et al., 2014). For example, physician responses may be more concentrated in diagnostic or risk-assessment contexts, where concise explanation is prioritized, whereas nurses, trainees, or support roles may more often provide reassurance, normalization, or practical next steps. These patterns motivate our qualitative analysis of how clinical context shapes the use of predicted empathy strategies.

Qualitative Analysis. We use BERTopic (Grootendorst, 2022) to identify themes from each submission's title and body. We manually review the 75 BERTopic topics and group them into 15 broader clinical themes, such as skin or visible symptoms, sexual and reproductive health, and healthcare access. Details on the topic modeling pipeline, topic-count selection, full codebook, and corpus coverage are in Appendix C.3.

Figure 5 shows that perspective-taking and explanation (PE) is the most common strategy across most clinical themes. This suggests that verified clinicians often express empathy by explaining the likely cause of symptoms or next steps, rather than using explicit emotional language. This is expected because many AskDocs posts are seeking medical explanations. Other strategies vary by clinical context. Mental health and anxiety posts show higher supportive guidance (SG), suggesting that clinicians often pair explanation with coping-oriented advice. Cancer anxiety and cardiopulmonary concerns show higher emotional reaction and reassurance (ER), likely because these posts often involve fear of serious illness and invite more direct reassurance or compassionate acknowledgment (Epstein et al., 2007). Skin and visible symptom posts show the highest prevalence of NE and lower rates of affective strategies. These cases may depend more on images or missing clinical details, leading clinicians to give brief assessments, ask for more information, or note the limits of online advice. Healthcare access and referral posts show higher PE and SG, reflecting responses that explain the care process while also providing supports.

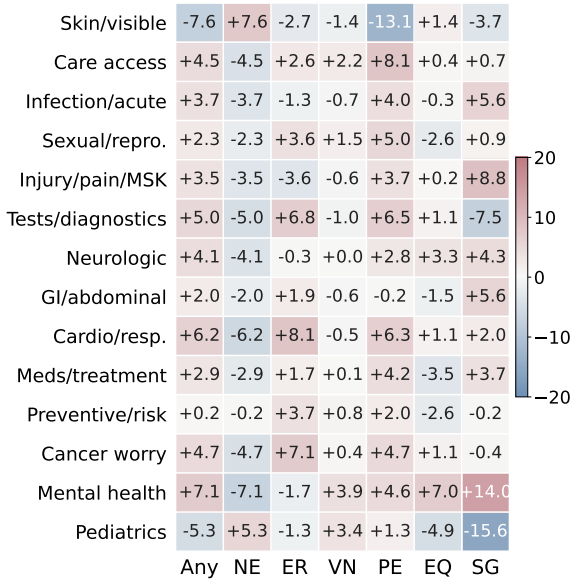

Figure 5: Predicted empathy strategy differences by clinical theme. Cells show percentage-point deviations from the corpus-wide average; positive values indicate higher-than-average use.

7 Conclusion

In this paper, we study multi-label empathy strategy classification in online medical support responses. We compare prompting, supervised fine-tuning, and hidden-state probing across multiple language models. Our results show that fine-tuning improves output-level prediction, while probing fine-tuned representations provides additional gains, especially for macro-level performance. Exploratory analyses further reveal temporal and professional variation in predicted empathy strategies. These findings suggest that representation-level analysis is a useful complement to output-based modeling for studying empathy in online medical advice.

Limitations

This study has several limitations. First, our experiments focus on a specific empathy classification setting in medical support responses. Although this domain is important because empathy plays a central role in how people seek and receive support online, the findings may not directly generalize to other forms of empathic communication, such as mental health counseling, peer support communities, or everyday interpersonal conversations. Future work should evaluate whether the same output- and representation-level patterns hold across broader domains and interaction contexts.

Second, although our task formulation allows responses to express multiple empathy categories, the boundaries among empathy functions remain inherently overlapping. A single response may simultaneously validate the user's feelings, take the user's perspective, ask follow-up questions, and offer coping suggestions. While the multi-label setting better reflects this complexity than a single-label formulation, it also introduces challenges for annotation and evaluation. In particular, different empathy categories may be highly correlated, and standard aggregate metrics may not fully capture partial correctness or the hierarchical relationships among affective, cognitive, and motivational forms of empathy. Future work could further examine label dependencies and develop evaluation methods that better reflect the layered structure of empathic communication.

Third, while probing provides useful evidence about what information is linearly recoverable from model representations, it does not fully explain how models use that information during generation or classification. A probe can show that empathy-related distinctions are encoded in hidden states, but it cannot establish a causal relationship between those representations and the model's final predictions. Future work could combine probing with intervention-based methods to more directly test whether specific representational features influence empathy judgments.

Finally, our analysis is limited to the model families, layers, and fine-tuning settings considered in this study. Different model architectures, training objectives, prompt formats, or larger-scale fine-tuning strategies may produce different representation-level patterns. We therefore view our findings as evidence that fine-tuned representations can improve empathy classification, rather than as a universal claim about all language models.

Ethical Considerations

This work studies automated empathy classification in medical support contexts, where model errors may have sensitive implications. We emphasize that the models examined in this paper are intended for research on language understanding, not for replacing human judgment in clinical, medical, or counseling settings. Empathy predictions should not be used to evaluate individual patients, caregivers, clinicians, or community members without careful validation and human oversight.

We also recognize the privacy concerns associated with analyzing online medical support data. Even when data are publicly available, users may discuss sensitive health experiences, emotional distress, or personal circumstances. Researchers should minimize the collection and reproduction of identifiable information, avoid quoting examples that could make users searchable, and follow appropriate data protection practices. In our qualitative analysis, examples are either paraphrased or anonymized to reduce the risk of re-identification.

Finally, this work should be understood as supporting better measurement and analysis of empathic communication, rather than as encouraging fully automated moderation or evaluation. Used responsibly, empathy classification tools may help researchers better understand patterns of support in online communities. However, deployment in real-world settings would require additional validation, bias auditing, transparency, and mechanisms for human review.

References

John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, and Zechariah Zhu. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6):589–596.

Anthony L. Back, Robert M. Arnold, Walter F. Baile, James A. Tulsky, and Kelly Fryer-Edwards. 2005. Approaching difficult communication tasks in oncology. *CA: A Cancer Journal for Clinicians*, 55(3):164–177.

Prakhar Biyani, Cornelia Caragea, Prasenjit Mitra, and John Yen. 2014. Identifying emotional and informational support in online health communities. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 827–836, Dublin, Ireland. Dublin

City University and Association for Computational Linguistics.

David A. Broniatowski, Amelia M. Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C. Quinn, and Mark Dredze. 2018. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American Journal of Public Health*, 108(10):1378–1384.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

David Chen, Kabir Chauhan, Rod Parsa, Zhihui Amy Liu, Fei-Fei Liu, Ernie Mak, Lawson Eng, Breffni Louise Hannon, Jennifer Croke, Andrew Hope, Nazanin Fallah-Rad, Phillip Wong, and Srinivas Raman. 2025a. Patient perceptions of empathy in physician and Artificial Intelligence chatbot responses to patient questions about cancer. *npj Digital Medicine*, 8(1):275.

Francine Chen, Scott Carter, Tatiana Lau, Nayeli Suseth Bravo, Sumanta Bhattacharyya, Kate Sieck, and Charlene C. Wu. 2025b. Empathy prediction from diverse perspectives. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria. Association for Computational Linguistics.

Julie W. Childers, Hailey Bulls, and Robert Arnold. 2023. Beyond the nurse acronym: The functions of empathy in serious illness conversations. *Journal of Pain and Symptom Management*, 65(4):e375–e379.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: user lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 307–318, New York, NY, USA. Association for Computing Machinery.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):71–80.

Julian De Freitas, Zeliha Oğuz-Uğuralp, Ahmet Kaan Uğuralp, and Stefano Puntoni. 2026. AI companions reduce loneliness. *Journal of Consumer Research*, 52(6):1126–1148.

Adrian de Wynter. 2025. If eleanor rigby had met ChatGPT: A study on loneliness in a post-LLM world. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19898–19913, Vienna, Austria. Association for Computational Linguistics.

Jean Decety and Philip L. Jackson. 2004. The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews*, 3(2):71–100.

Marco Del Tredici and Raquel Fernández. 2017. Semantic variation in online communities of practice. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Long papers*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ronald M. Epstein, Tanveer Hadee, Jennifer Carroll, Sarah C. Meldrum, Jean Lardner, and Cleveland G. Shields. 2007. "Could This Be Something Serious?" Reassurance, Uncertainty, and Empathy in Response to Patients' Expressions of Worry. *Journal of General Internal Medicine*, 22(12):1731–1739.

Gunther Eysenbach, John Powell, Marina Englesakis, Carlos Rizo, and Anita Stern. 2004. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *BMJ*, 328(7449):1166.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Rame, Morgane Riviere, and 1 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *Preprint*, arXiv:2203.05794.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

M. Jawad Hashim. 2017. Patient-centered communication: Basic skills. *American Family Physician*, 95(1):29–34.

Yuanyuan He, Yongsen Pan, Wei Li, Jiali You, Jiawen Deng, and Fuji Ren. 2025. ECC: An emotion-cause conversation dataset for empathy response. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China. Association for Computational Linguistics.

Mahshid Hosseini and Cornelia Caragea. 2021. It takes two to empathize: One to seek and one to provide. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):13018–13026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kokil Jaidka, Niyati Chhaya, and Lyle Ungar. 2018. Diachronic degradation of language models: Insights from social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200, Melbourne, Australia. Association for Computational Linguistics.

Ching-Yun Ko and Pin-Yu Chen. 2026. vllm hook v0: A plug-in for programming model internals on vllm. *arXiv preprint arXiv:2603.06588*.

Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*, fourth edition edition. SAGE Publications, Inc., Thousand Oaks, California.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 625–635, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022. A critical reflection and forward perspective on empathy and natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2139–2158, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gyeongeun Lee, Christina Wong, Meghan Guo, and Natalie Parde. 2024. Pouring your heart out: Investigating the role of figurative language in online expressions of empathy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 519–529, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Sean MacAvaney, Bart Desmet, Arman Cohan, Luca Soldaini, Andrew Yates, Ayah Zirikly, and Nazli Goharian. 2018. RSDD-time: Temporal annotation of self-reported mental health diagnoses. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 168–173, New Orleans, LA. Association for Computational Linguistics.

Ryan K. McBain, Robert Bozick, Melissa Diliberti, and 1 others. 2025. Use of generative AI for mental health advice among US adolescents and young adults. *JAMA Network Open*, 8(11):e2542281.

Humphrey Mensah, Lu Xiao, and Sucheta Soundarajan. 2020. Characterizing the evolution of communities on Reddit. In *Proceedings of the International Conference on Social Media and Society*, pages 58–64. Association for Computing Machinery.

S Anne Moorhead, Diane E Hazlett, Laura Harrison, Jennifer K Carroll, Anthea Irwin, and Ciska Hoving. 2013. A new dimension of health care: Systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of Medical Internet Research*, 15(4).

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Comput. Linguist.*, 42(3):537–593.

Alicia L. Nobles, Eric C. Leas, Mark Dredze, and John W. Ayers. 2020. Examining peer-to-peer and patient-provider interactions on a social media community facilitating ask the doctor services. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):464–475.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/index/chatgpt/`. Accessed: 2026-05-21.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Michael Paul and Mark Dredze. 2021. You are what you tweet: Analyzing twitter for public health. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):265–272.

Qwen Team. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Ian Stewart and Jacob Eisenstein. 2018. Making "fetch" happen: The influence of social and linguistic context on nonstandard word growth and decline. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4370, Brussels, Belgium. Association for Computational Linguistics.

Jina Suh, Lindy Le, Erfan Shayegani, Gonzalo Ramos, Judith Amores, Desmond C. Ong, Mary Czerwinski, and Javier Hernandez. 2025. SENSE-7: Taxonomy and dataset for measuring user perceptions of empathy in sustained human-AI conversations. *Preprint*, arXiv:2509.16437.

Caroline K. Tietbohl. 2022. Empathic validation in physician–patient communication: An approach to conveying empathy for problems with uncertain solutions. *Qualitative Health Research*, 32(3):413–425. PMID: 34894864.

Somin Wadhwa, Vivek Khetan, Silvio Amir, and Byron Wallace. 2023. RedHOT: A corpus of annotated medical questions, experiences, and claims on social media. In *Findings of the Association for Computational Linguistics: EACL 2023*, Dubrovnik, Croatia. Association for Computational Linguistics.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, and 1 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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

Hongli Zhan, Emma S. Gueorguieva, Javier Hernandez, Jina Suh, Desmond C. Ong, and Junyi Jessy Li. 2026. Discourse diversity in multi-turn empathic dialogue. *arXiv preprint arXiv:2604.11742*.

Justine Zhang, William Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):377–386.

Jian Zhu and David Jurgens. 2021. The structure of online social networks modulates the rate of lexical change. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2201–2218, Online. Association for Computational Linguistics.

Appendix

A Annotation Scheme and Criteria

A.1 Label Schema and Task Definition

We develop a domain-adapted annotation scheme for empathy strategies in online medical advice. The scheme draws on prior computational empathy frameworks, especially EPITOME (Sharma et al., 2020). Because online medical advice differs from general emotional support, we further adapt these categories using prior work on patient-centered clinical communication, reassurance, empathic validation, and online health support. In clinical communication, empathy is often operationalized through observable practices such as eliciting patients' concerns, using open-ended questions, understanding the patient's perspective, expressing empathy, and offering support (Back et al., 2005; Hashim, 2017; Childers et al., 2023). Reassurance is also a central clinical response to patient worry and uncertainty (Epstein et al., 2007), while empathic validation has been described as a distinct practice for demonstrating shared understanding, supporting the patient's position, and normalizing health-related concerns (Tietbohl, 2022). In online health settings, social support research further distinguishes emotional support from informational support, suggesting that medical advice often combines affective responses with explanation, advice, and concrete guidance (Biyani et al., 2014). These perspectives motivate a schema that captures both socioemotional and medically grounded forms of empathy in verified health-worker responses.

Our final schema contains six response-level empathy-strategy labels. Because a single medical comment may express multiple forms of empathy, we formulate the task as a multi-label classification problem. To avoid implying an ordinal empathy scale, we refer to labels using mnemonic category codes rather than numeric indices. **NE** denotes no identifiable empathy strategy and is assigned when a response provides no clear emotional, cognitive, exploratory, or supportive uptake of the poster's situation. **ER** captures emotional reaction and reassurance, including emotional acknowledgment and

attempts to reduce worry or uncertainty (Sharma et al., 2020; Epstein et al., 2007; Childers et al., 2023). **VN** captures validation and normalization of the poster's feelings, concerns, or experience, reflecting clinical practices that communicate shared understanding, legitimize patient concerns, or normalize health-related experiences (Tietbohl, 2022; Hashim, 2017; Childers et al., 2023). **PE** captures perspective-taking and explanation, extending EPITOME's interpretation category to medical advice settings where clinicians explain symptoms, risks, or mechanisms in ways that demonstrate understanding of the poster's situation (Sharma et al., 2020; Hashim, 2017). **EQ** captures exploratory questioning through relevant follow-up questions, aligning with both EPITOME's exploration category and patient-centered communication practices that elicit patients' concerns, context, expectations, and missing information (Sharma et al., 2020; Hashim, 2017; Childers et al., 2023). **SG** captures supportive guidance, including motivational support, coping guidance, or concrete next steps, reflecting support-oriented clinical communication as well as online health support categories such as informational support and guidance (Back et al., 2005; Childers et al., 2023; Biyani et al., 2014). The same label definitions and decision criteria are also used to construct the instructions for prompted LLM baselines, ensuring that human annotation and model prediction are grounded in a consistent operationalization of empathy strategies.

A.2 Label Code Mapping

The main text uses mnemonic label codes to avoid implying that empathy strategies form an ordinal scale. During model development and prompting, we used the original internal labels S0–S5. Table A.1 maps the internal labels to the paper-facing labels.

Internal	Paper	Definition
S0	NE	No identifiable empathy strategy
S1	ER	Emotional reaction and reassurance
S2	VN	Validation and normalization
S3	PE	Perspective-taking and explanation
S4	EQ	Exploratory questioning
S5	SG	Supportive guidance

Table A.1: Mapping between internal experimental labels and paper-facing empathy strategy codes.

B Experimental Details

B.1 Model Checkpoints

Table B.1 summarizes the public checkpoints used in our experiments. Encoder baselines are fine-tuned as multi-label classifiers, while larger language models are evaluated through prompting, LoRA fine-tuning, hidden-state probing, or their combinations. We used two servers, one server with four NVidia H100s and another with two NVidia A6000 GPUs.

Model group	Checkpoints and use
Encoder baselines	`bert-base-uncased` and `roberta-base`, used for supervised encoder fine-tuning baselines.
Instruction-tuned LLMs	`google/gemma-3-4b-it`, `google/gemma-3-12b-it`, `Qwen/Qwen2.5-7B-Instruct`, `Qwen/Qwen2.5-14B-Instruct`, and `Qwen/Qwen3-8B`, used for prompting, LoRA fine-tuning, and hidden-state probing where applicable.
API baseline	`gpt-4o-mini`, used as a prompted LLM baseline.

Table B.1: Model checkpoints used in the experiments.

B.2 TF–IDF and Encoder Baselines

The TF–IDF baselines use the post title, post body, and clinician response as input. We evaluate one-vs-rest logistic regression and linear SVM classifiers, as well as dummy baselines. For the tuned linear models, hyperparameters are selected with three-fold grid search using macro-F1 on the training split. The TF–IDF search space includes unigram and unigram–bigram features, maximum vocabulary sizes of 50,000 and 100,000, minimum document frequency values of 1 and 2, and classifier regularization strengths $C \in \{0.5, 1, 2, 4\}$ with either balanced or unweighted class weights.

We also fine-tune BERT-base and RoBERTa-base with a sigmoid multi-label classification head. Both encoder models use 1,400 training examples, 200 validation examples, and 400 test examples. We use a per-device training batch size of 8, per-device evaluation batch size of 16, warmup ratio of 0.06, no positive-class reweighting, and keep-tail truncation for long post bodies. Runs are ranked by

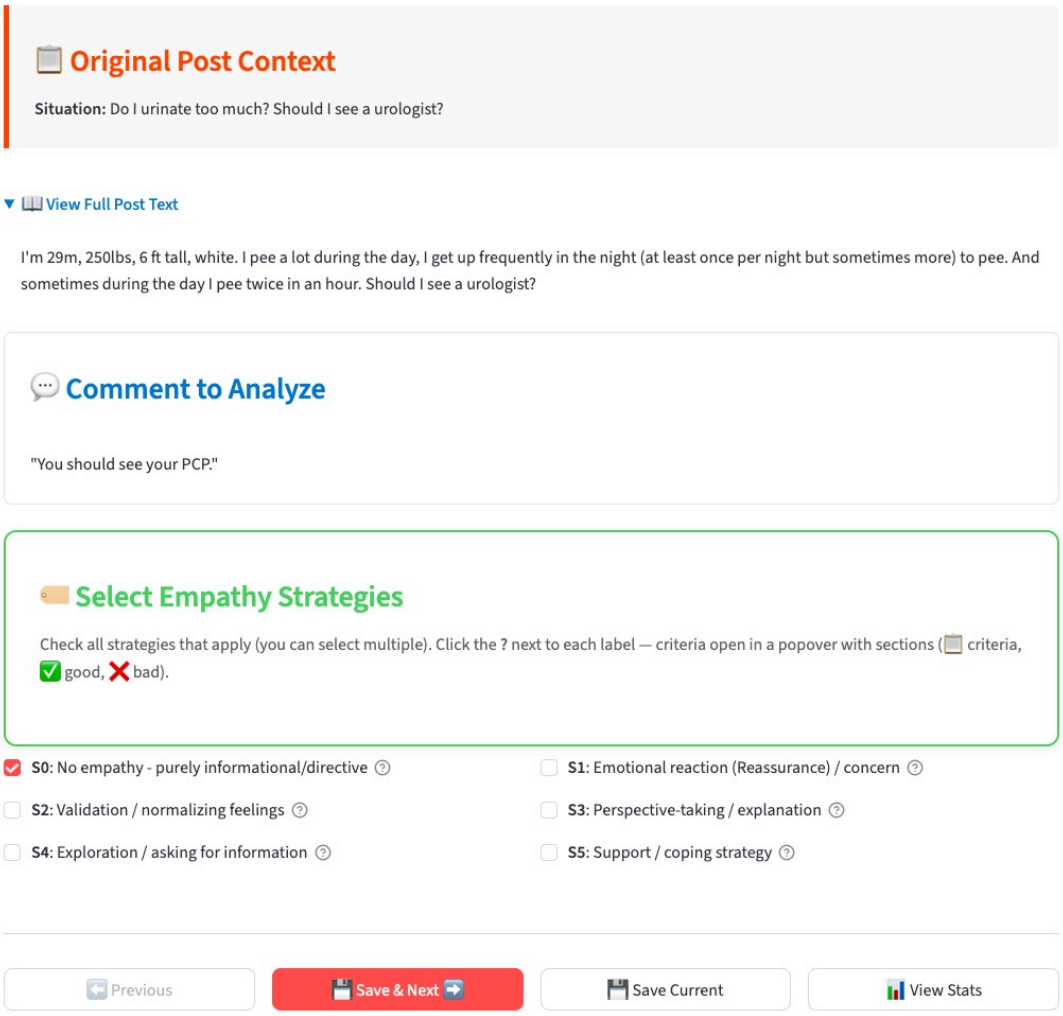

Figure A.1: An screenshot of our annotation app.

Model	NE	ER	VN	PE	EQ	SG
BERT-base Encoder FT	.35	.35	.20	.40	.35	.50
RoBERTa-base Encoder FT	.20	.25	.15	.60	.25	.45

Table B.2: Validation-tuned decision thresholds for encoder fine-tuning baselines.

validation macro-F1, checkpoints are selected by evaluation loss, and per-label decision thresholds are tuned on the validation split.

B.3 LoRA Supervised Fine-Tuning

We fine-tune instruction-tuned LLMs using LoRA adapters. All LoRA runs use the same adapter configuration and differ primarily in the base model and training data composition. We use a learning rate of 2×10^{-4}, weight decay of 0.0, and a warmup ratio of 0.03. The per-device training batch size is 1 with 16 gradient accumulation steps, giving an effective batch size of 16. The maximum sequence length is 2048, and long inputs are handled using keep-tail truncation. Training uses bfloat16 precision with NF4 4-bit quantization and double quantization, with gradient checkpointing enabled to reduce memory use.

The LoRA adapter uses rank $r = 16$, scaling parameter $\alpha = 32$, and dropout of 0.05. Adapters are applied to the attention and feed-forward projection modules: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. During training, per-device evaluation batch size is 1; during final SFT evaluation, it is 4. We use the last training checkpoint for evaluation rather than selecting checkpoints based on test-set performance. Final predictions are evaluated on the held-out AskDocs test set.

B.4 Hidden-State Probing

We train multilabel linear probes on hidden representations extracted from the LLMs. For each example, the input consists of the post title, post body, and target clinician response, formatted using the same empathy-labeling prompt templates and S0–S5 criteria as in the output-level experiments. Gold annotations are converted into six-dimensional multi-hot vectors over S0–S5. During prediction post-processing, we enforce the mutual exclusivity constraint between S0 and S1–S5.

Hidden states are extracted using `vLLM-Hook` (Ko and Chen, 2026). We perform teacher-forced activation extraction: the complete prompt and target clinician response are provided to the model, and hidden states are collected from the resulting forward pass rather than from free-form decoding. This ensures that the extracted representations correspond to the fixed clinician response for each example and are not affected by sampling or other generation-time variability.

For each selected transformer layer, we use the hidden state at the final token position as the representation for probing. When multiple layers are analyzed, hidden states are extracted for each layer under the same procedure, and probes are trained according to the layer-specific configuration used in the experiment. We report probing results based on these layer-wise hidden representations. The same extraction pipeline is used for both base models and LoRA-adapted models. For LoRA-adapted runs, the corresponding adapters are loaded through vLLM's LoRA request mechanism during hidden-state extraction.

The probe is a multilabel one-vs-rest logistic regression classifier implemented as `StandardScaler` followed by `OneVsRestClassifier(LogisticRegression)`. The scaler is fit only on the corresponding training data, and logistic regression uses an L2 penalty. Hyperparameters are selected using `RandomizedSearchCV` over $C \sim \log U(10^{-4}, 10^{3})$, class-weighting options including balanced and unweighted settings, and `max_iter` values of 1000, 2000, and 4000, with `lbfgs` as the solver.

For probe training, the original training split is further divided into `train_fit` and validation subsets using a 9:1 split, grouped by `comment_id` so that examples associated with the same original comment remain in the same fold. Hyperparameter tuning is performed on `train_fit` using five-fold cross-validation. After selecting the logistic regression configuration, per-label decision thresholds are tuned using out-of-fold probabilities from `train_fit`. Thresholds are optimized by coordinate ascent over $\{.30, .35, .40, .45, .50, .55, .60\}$, using either macro-F1 or micro-F1 as the tuning objective.

Layer selection is conducted in a separate sweep phase. Candidate layers are evaluated on the held-out validation split, with test evaluation and artifact saving disabled. The best layer is selected as the layer with the highest macro-F1 score on the held-out validation split. After selecting the layer, we rerun the probe, refit the scaler and classifier on the full training split, and evaluate the final model on the held-out test split. Figures B.1 and B.2 show representative layer-sweeping curves for Qwen3-8B and Gemma-3-12B.

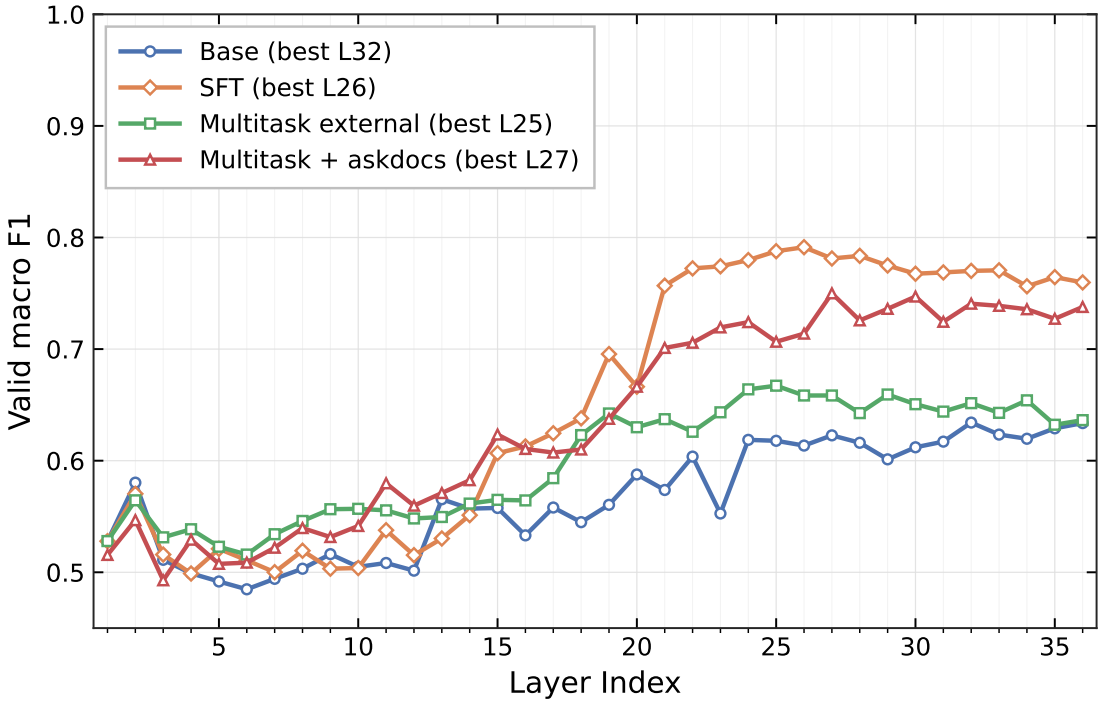

Figure B.1: Qwen3-8B Layer Sweeping and Selection

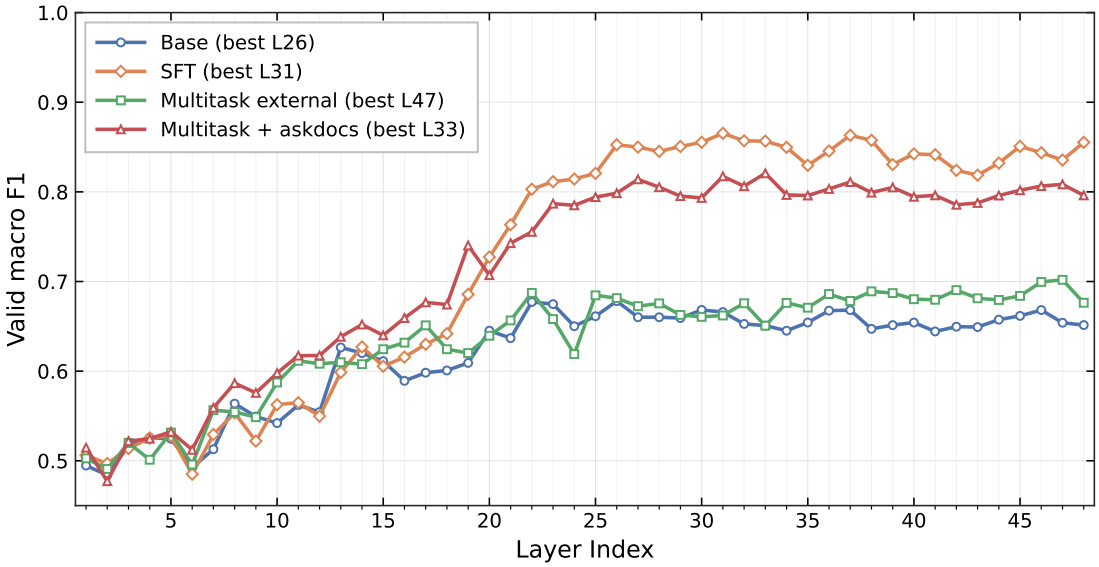

Figure B.2: Gemma-3-12B Layer Sweeping and Selection

C In-Depth Analysis

C.1 NE Change over Time

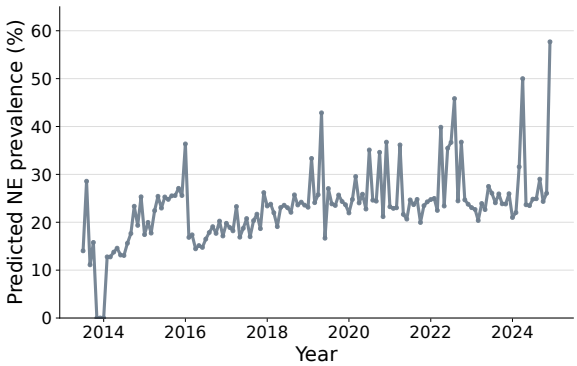

Figure C.1: Monthly predicted prevalence of NE in verified AskDocs responses. Lines show three-month rolling averages. NE denotes responses with no identifiable empathy strategy.

C.2 Exploratory Interrupted Time-Series Analysis

We conduct an exploratory interrupted time-series analysis to examine whether predicted empathy strategy prevalence changed around the public release of ChatGPT. Because ChatGPT was released on November 30, 2022, we use December 2022 as the first post-release month. This analysis is descriptive rather than causal: the data do not include an untreated control group, and we do not observe whether individual clinicians used generative AI tools.

We aggregate comment-level predictions to the month level. For each verified clinician response i, month t, and empathy strategy label $k \in \{\text{NE}, \text{ER}, \text{VN}, \text{PE}, \text{EQ}, \text{SG}\}$, let $y_{ikt} = 1$ if the response is predicted to contain label k, and 0 otherwise. Monthly predicted prevalence is defined as

$$p_{kt} = 100 \times \frac{\sum_{i \in t} y_{ikt}}{n_t},$$

where n_t is the number of verified clinician responses in month t. The outcome is therefore measured in percentage points. Because the task is multilabel, the prevalence rates for ER, VN, PE, EQ, and SG are not mutually exclusive and may sum to more than 100%. NE is mutually exclusive with the empathy-strategy labels during prediction post-processing.

For each label k, we estimate the following segmented time-series model at the monthly level:

$$p_{kt} = \alpha_k + \beta_{1k} T_t + \beta_{2k} D_t + \beta_{3k} A_t + \sum_{m=2}^{12} \gamma_{km} M_{mt} + \varepsilon_{kt}.$$

Here, T_t is a continuous month index, D_t is an indicator equal to one for December 2022 and later months, and A_t is a post-breakpoint trend variable equal to zero before December 2022 and increasing by one in each month after the breakpoint. M_{mt} denotes month-of-year indicators, with January omitted as the reference month. Models are estimated separately for each label using monthly observations. We weight observations by the number of verified responses in each month, n_t, so that months with more responses contribute more information to the estimated trend. Standard errors are adjusted using Newey–West heteroskedasticity- and autocorrelation-consistent standard errors with a three-month lag.

The coefficient β_{2k} captures the immediate level shift after the breakpoint, while β_{3k} captures the change in monthly trend after the breakpoint. Both coefficients are measured in percentage points. Because this analysis is exploratory and repeated across six labels, we interpret the results as descriptive evidence of temporal patterning rather than causal evidence of ChatGPT's effect on clinician responses.

Label	Pre-trend		Level shift		Slope change		Post-trend	
	Est.	p	Est.	p	Est.	p	Est.	p
NE	0.05	.006	-1.49	.119	-0.01	.811	0.04	.446
ER	0.04	$<$ .001	-0.14	.810	0.01	.709	0.05	.074
VN	0.01	$<$ .001	-0.04	.815	-0.01	.254	0.00	.805
PE	-0.01	.528	-1.76	.083	0.11	.104	0.10	.131
EQ	-0.09	$<$ .001	3.87	$<$ .001	-0.10	.036	-0.20	$<$ .001
SG	-0.05	.008	-1.40	.115	0.07	.106	0.02	.647

Table C.1: Exploratory interrupted time-series results using December 2022 as the breakpoint. The outcome is monthly predicted label prevalence, measured in percentage points. Models are estimated separately by label, weighted by the number of verified responses per month, include month-of-year indicators, and use Newey–West standard errors with a three-month lag.

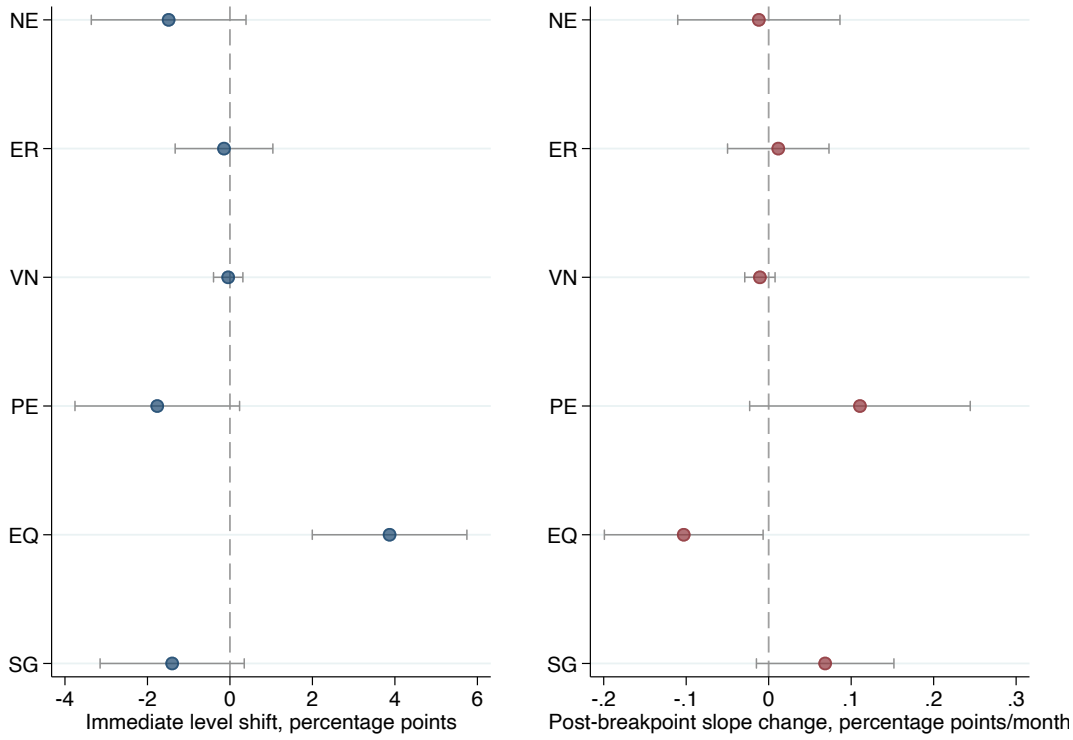

Figure C.2: Interrupted time-series coefficient estimates for predicted empathy strategy prevalence. Points show coefficient estimates and bars show 95% confidence intervals. The left panel reports the immediate level shift after the December 2022 breakpoint, and the right panel reports the post-breakpoint slope change.

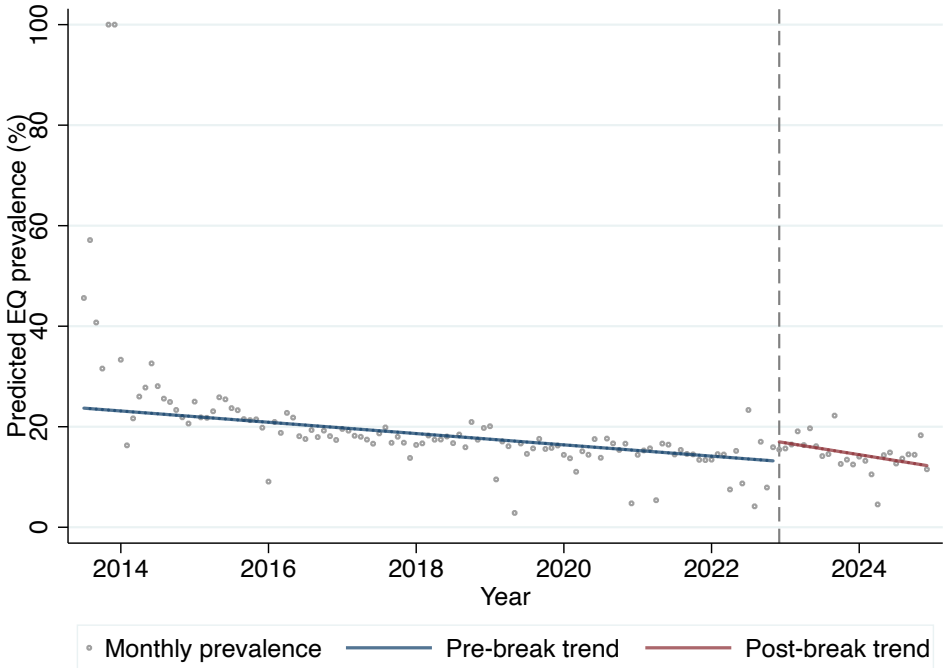

Figure C.3: Interrupted time-series visualization for predicted exploratory questioning (EQ) prevalence. Points show monthly predicted EQ prevalence, and lines show fitted pre- and post-breakpoint trends using December 2022 as the breakpoint.

C.3 BERTopic Implementation Details

We use BERTopic to identify recurring clinical contexts in the full verified AskDocs corpus. Topic modeling is performed at the post level, therefore each document is constructed by concatenating the Reddit submission title and body. Clinician replies are not used as topic-modeling input; instead, topic assignments are linked back to clinician responses after inference.

The input file contains 441,699 verified AskDocs rows. Before embedding, we normalize text by replacing non-breaking spaces, collapsing repeated whitespace, stripping leading and trailing whitespace, and truncating each document to 6,000 characters. Documents are embedded using `sentence-transformers/all-MiniLM-L6-v2`. To make full-corpus topic modeling computationally feasible, BERTopic is fit on a random sample of 50,000 documents using `random_state=42`. The fitted model is then used to assign topics to all 441,699 documents in batches of 10,000.

The BERTopic pipeline uses UMAP for dimensionality reduction and HDBSCAN for clustering. UMAP is configured with `n_neighbors=15`, `n_components=5`, `min_dist=0.0`, `metric=cosine`, and `random_state=42`. HDBSCAN is configured with `min_cluster_size=50`, `metric=euclidean`, `cluster_selection_method=eom`, and `prediction_data=True`. Topic representations are generated using BERTopic's class-based TF-IDF representation. The vectorizer removes English stop words, uses unigrams and bigrams, sets `min_df=1`, and limits the vocabulary to 50,000 features.

To choose the final number of topics, we evaluate several topic-reduction settings on the 50,000-document fitting sample. Specifically, we compare candidate topic counts using topic coherence, measured by c_v, and topic diversity. As shown in Figure C.4, coherence increases substantially when moving from very small topic counts to more fine-grained solutions, but the marginal coherence gains become smaller after approximately 75 topics. At the same time, topic diversity begins to decline as the number of topics increases further. We therefore select 75 topics as a balance between semantic coherence, diversity, and manual interpretability.

The resulting topic-model outputs include topic sizes, top class-based TF-IDF terms, representative documents, and row-level topic assignments. For qualitative interpretation, we construct a topic review table containing each topic ID, topic size, top words, representative post titles, representative documents, and fields for manual topic naming. We then manually group semantically similar fine-grained BERTopic topics into broader clinical themes, including cancer anxiety, medication and side effects, skin or visible symptoms, sexual and reproductive health, mental health and anxiety, injury or musculoskeletal symptoms, lab result interpretation, and infection or acute symptoms. The outlier topic -1 is treated as a mixed/noise category rather than as a substantive clinical theme.

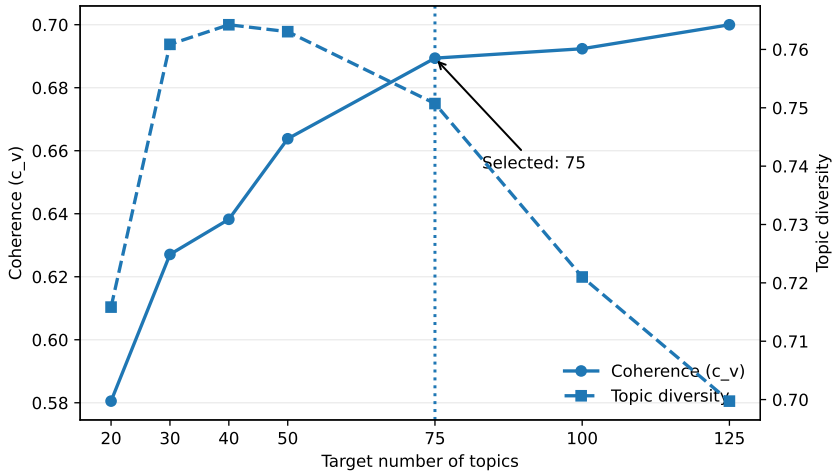

Figure C.4: Topic-count selection diagnostics for BERTopic. We select 75 topics as a balance between topic coherence and topic diversity.

Broad theme	N	**Share**	**Any**	**NE**	**ER**	**VN**	**PE**	**EQ**	**SG**
Other / mixed / unclear	150,938	34.2	73.7	26.3	18.4	2.6	53.6	15.8	28.3
Skin / visible symptoms	53,482	12.1	69.1	30.9	17.9	1.2	43.2	17.5	27.3
Healthcare access, referrals, and care-seeking	38,595	8.7	81.2	18.8	23.2	4.7	64.4	16.5	31.6
Infection / acute symptoms	34,408	7.8	80.4	19.6	19.3	1.9	60.3	15.7	36.6
Sexual and reproductive health	30,553	6.9	79.0	21.0	24.2	4.1	61.3	15.3	31.8
Injury / pain / musculoskeletal symptoms	25,539	5.8	80.2	19.8	17.0	2.0	60.0	16.2	39.7
Diagnostic uncertainty, test results, and interpretation	21,001	4.8	81.7	18.3	27.4	1.6	62.8	17.1	23.5
Neurological symptoms	20,262	4.6	80.8	19.2	20.2	2.6	59.1	19.3	35.2
Gastrointestinal and abdominal symptoms	16,267	3.7	78.7	21.3	22.5	2.0	56.1	14.5	36.6
Cardiovascular and respiratory symptoms	12,814	2.9	82.9	17.1	28.7	2.1	62.6	17.1	33.0
Medication, treatment, procedures, and side effects	12,780	2.9	79.6	20.4	22.2	2.7	60.5	12.6	34.6
Preventive care, lifestyle, and exposure risk	12,109	2.7	76.9	23.1	24.3	3.4	58.3	15.5	30.7
Cancer anxiety / serious disease worry	8,890	2.0	81.4	18.6	27.6	3.0	61.0	17.2	30.5
Mental health and anxiety	3,270	0.7	83.8	16.2	18.8	6.5	60.9	23.0	45.0
Pediatrics and child health	791	0.2	71.4	28.6	19.2	5.9	57.6	11.1	15.3

Table C.2: Predicted empathy strategy prevalence by broad clinical theme. N reports the number of verified clinician responses assigned to each theme, and Share reports the percentage of the full verified corpus assigned to that theme. Any indicates that a response is predicted to contain at least one empathy strategy other than NE. Because the task is multi-label, ER, VN, PE, EQ, and SG are not mutually exclusive and may sum to more than 100%.

D Prompt Templates

The prompts below preserve the original internal labels used during model development. In the main text, we report the same categories using mnemonic label codes: S0 = NE, S1 = ER, S2 = VN, S3 = PE, S4 = EQ, and S5 = SG.

D.1 AskDocs Empathy Classification Prompt

LLM prompting, LoRA supervised fine-tuning, LoRA evaluation, and full-data prediction use the same AskDocs empathy classification template. The complete template is shown in Figures D.1 and D.1. The first part defines the task, label set, general labeling rules, and criteria for S0–S2; the second part continues the criteria for S3–S5 and specifies the required JSON output format.

D.2 External Empathy Dataset Prompts

For multitask LoRA fine-tuning, we convert the external empathy datasets into FLAN-style instruction-following examples. Each dataset uses a short system instruction together with a task-specific user template. The templates for EmpatheticDialogues, EPITOME, IEMPATHIZE, and ECC are shown in Figures D.2, D.3, D.4, and E.1, respectively.

D.2.1 EmpatheticDialogues Response Generation

EmpatheticDialogues is an open-domain dialogue dataset in which conversations are grounded in speakers' emotional situations, making it useful for response-generation supervision in affectively salient contexts (Rashkin et al., 2019).

D.2.2 Empathy-Mental-Health Mechanism Rating

EPITOME focuses on empathy expressed in asynchronous text-based mental health support and operationalizes empathy through communication mechanisms such as emotional reactions, interpretations, and explorations (Sharma et al., 2020).

D.2.3 IEMPATHIZE Direction Classification

IEMPATHIZE studies empathy in an online cancer support community and distinguishes the direction of empathy in user messages, including empathy seeking and empathy providing (Hosseini and Caragea, 2021).

D.2.4 ECC Emotion-Cause Identification

ECC provides empathetic conversations annotated with emotions and their underlying causes, supporting instruction examples that require identifying emotion–cause relations before producing or evaluating empathetic responses (He et al., 2025).

E Use of AI Assistants

AI-based tools were used to assist with language refinement and clarity. All ideas, analyses, and conclusions presented in this work are solely those of the authors.

Prompt Template: Empathy Labeling (Part 1)

You are an expert empathy labeler.
Your job: read the post context and the reply comment, then assign empathy labels for the COMMENT.
Follow the label definitions and criteria strictly.
Label definitions:

- S0: No empathy / irrelevant / purely informational without emotional component.
- S1: Emotional reactions, concern, or reassurance.
- S2: Validation / normalizing the person's feelings or experience.
- S3: Perspective-taking / explanation / reframing.
- S4: Exploration / asking clarifying questions.
- S5: Support / coping suggestions / next-step guidance.

Rules:

- Labels S1-S5 may co-occur.
- If any of S1-S5 applies, do not assign S0.

Empathy criteria:
S0

- Dismissive: The comment dismisses, minimizes, or trivializes the OP's experience, or is emotionally unresponsive to it.
- Hostile/Abusive: The comment is critical, blaming, judgmental, degrading, or disrespectful toward the OP or the OP's experience.
- Hostile/Abusive: The comment is hostile, abusive, malicious, or expresses ill will toward the OP, including ridicule, insults, mockery, or wishing harm.
- Hostile/Abusive: The comment undermines, attacks, or invalidates the OP.
- Factual: The comment is purely factual, procedural, or informational without any emotional acknowledgment, understanding, or supportive intent.
- Factual: The comment focuses only on correction, contradiction, or instruction without any empathic component.
- Non-responsive: The comment is irrelevant or self-focused rather than responsive to the OP's situation.
- Irrelevant: The comment is irrelevant to the OP's core emotional or medical concern.
- Skeptical: The comment focuses on questioning the authenticity of the OP's situation.

S1

- Emotional Reactions: The comment expresses sympathy or sorrow in response to the OP's distress.
- Emotional Reactions: The comment expresses care, concern, or compassion toward the OP.
- Emotional Reactions: The comment expresses positive affect, such as relief, happiness, comfort, soothing, or warmth, in response to the OP's situation.
- Emotional Reactions: The comment acknowledges the emotional pain or distress in the OP's experience by recognizing it as sad, painful, upsetting, or difficult.
- Reassurance: The comment attempts to reduce the OP's fear, anxiety, or distress by emphasizing safety, normalcy, hope, or likely positive outcomes.
- Reassurance: The comment tells the OP not to panic, not to worry excessively, or suggests that the issue may not be as serious as feared in an empathetic tone.
- Reassurance: The comment emphasizes commonness, manageability, or likely improvement in order to comfort the OP.
- Reassurance: The comment's main function is soothing or calming distress rather than validating the legitimacy of the OP's feelings.

S2

- Validation: The comment affirms that the OP's feelings, reactions, concerns, symptoms, or experiences are understandable, reasonable, or not abnormal.
- Validation: The comment explicitly states that the OP's reaction or concern makes sense.
- Validation: The comment normalizes the OP's experience by suggesting that it is common, not unusual, or not something the OP is alone in experiencing.
- Validation: The comment's main function is to affirm the legitimacy or normality of the OP's internal experience rather than to explain causes or provide action guidance.

Figure D.1: Prompt template for assigning empathy labels to a reply comment given the post context, part 1.

Prompt Template: Empathy Labeling (Part 2)

S3

- Perspective-taking: The comment interprets how the OP may be feeling or what aspect of the situation may be especially difficult or distressing for them.
- Perspective-taking: The comment reflects cognitive understanding of the OP's subjective experience rather than only expressing emotion, validation, or reassurance.
- Perspective-taking: The comment helps articulate the OP's perspective, concerns, or emotional experience in a more fully understood and interpretable way.
- Explanation: The comment provides interpretive or informational understanding of what may be happening in the OP's situation.
- Explanation: The comment explains possible causes, mechanisms, consequences, or medical context related to the OP's symptoms or concerns.
- Explanation: The comment offers an alternative interpretation of the OP's condition, symptoms, or situation.
- Explanation: The comment clarifies broader medical norms, context, or the limits of interpretation in the absence of sufficient information.
- Explanation: The comment's main function is to help the OP understand the issue better with current available information.

S4

- Exploration: The comment asks for more specificity, indicating a desire to clarify the OP's situation.
- Exploration: By requesting additional details, the commenter seeks to better understand the OP's concerns.
- Exploration: By asking relevant and specific questions about symptoms or the reason for the lab tests, the commenter is actively seeking to clarify the OP's situation.
- Exploration: The comment asks follow-up questions in order to gather more information about the OP's symptoms, circumstances, or experience.

S5

- Support: The comment shows intention to help, guide, or encourage action that benefits the OP.
- Support: The comment encourages the OP to seek help, take care of themselves, or pursue a helpful course of action.
- Coping Strategies: The comment offers concrete next steps, resources, or actionable suggestions to help the OP manage, cope with, or respond to the issue.
- Coping Strategies: The comment recommends specific behaviors, strategies, or routines that may help the OP address the concern.
- Coping Strategies: The comment provides practical advice or resource-oriented guidance tailored to the OP's situation.

Assign empathy labels for the reply comment.
Output requirements:

- Output exactly one JSON object and nothing else.
- Format: `{"labels": ["S0""S1"|"S2"|"S3"|"S4"|"S5"]|`
- Labels must be unique and sorted ascending (S0..S5).
- Labels S1-S5 may co-occur.
- If any of S1-S5 applies, do not assign S0.

Post context
Title:
{post_title}
Body:
{post_body}
Comment to label (the reply)
{comment}
{Expected output format.}

`{"labels": ["S0"|"S1"|"S2"|"S3"|"S4"|"S5"]}`

Figure D.1: Prompt template for assigning empathy labels to a reply comment given the post context, continued. The placeholders `{post_title}`, `{post_body}`, and `{comment}` are replaced with the corresponding dataset fields at inference time.

Prompt Template: Empathetic Dialogue Response Generation

You are doing empathy-related instruction tuning.
Follow the task instructions exactly and output only the requested JSON object.

Task: Generate the next empathetic dialogue response.

Emotion/situation label:
{context}

Original situation prompt:
{prompt}

Dialogue so far:
{history}

Output JSON format:

```
{"response": "..."}
```

Figure D.2: Prompt template for generating the next empathetic dialogue response. The placeholders `{context}`, `{prompt}`, and `{history}` are replaced with the corresponding dialogue fields at inference time.

Prompt Template: Empathy Communication Mechanism Rating

You are doing empathy-related instruction tuning.
Follow the task instructions exactly and output only the requested JSON object.

Task: Rate one empathy communication mechanism in a mental-health support reply.

Mechanism:
{mechanism_description}

Level meanings:

- 0 = no communication of this mechanism.
- 1 = weak or generic communication.
- 2 = strong communication.

Support seeker post:
{seeker_post}

Support response:
{response_post}

Output JSON format:

```
{"mechanism": "...", "level": 0|1|2, "rationales": ["..."]}
```

Figure D.3: Prompt template for rating one empathy communication mechanism in a mental-health support reply. The placeholders `{mechanism_description}`, `{seeker_post}`, and `{response_post}` are replaced with the corresponding input fields at inference time.

Prompt Template: Empathy Direction Classification

You are doing empathy-related instruction tuning.
Follow the task instructions exactly and output only the requested JSON object.

Task: Classify the empathy direction expressed in the text.

Labels:

- none
- seeking_empathy
- providing_empathy

Text:
{text}

Output JSON format:

```
{"empathy_direction": "none|seeking_empathy|providing_empathy"}
```

Figure D.4: Prompt template for classifying the empathy direction expressed in a text. The placeholder `{text}` is replaced with the corresponding input text at inference time.

Prompt Template: Emotion-Cause Identification

You are doing empathy-related instruction tuning.
Follow the task instructions exactly and output only the requested JSON object.

Task: Identify the emotion and its cause in an empathy-related dialogue.

Dialogue:
{cumulative_dialogue}

Output JSON format:

```
{"emotion_label": "...", "emotion_utterance": "...", "cause_utterance": "..."}
```

Figure E.1: Prompt template for identifying the emotion and its cause in an empathy-related dialogue. The placeholder `{cumulative_dialogue}` is replaced with the corresponding dialogue context at inference time.