nature machine intelligence

Article

https://doi.org/10.1038/s42256-025-01169-6

When large language models are reliable for judging empathic communication

Received: 11 June 2025

Accepted: 3 December 2025

Published online: 11 February 2026

Check for updates

Aakriti Kumar [1,2] ✉, Nalin Poungpeth[1], Diyi Yang[3], Erina Farrell[4], Bruce L. Lambert[5] & Matthew Groh [1,2,6] ✉

Large language models (LLMs) excel at generating empathic responses in text-based conversations. But, how reliably do they judge the nuances of empathic communication? Here we investigate this question by comparing how experts, crowdworkers and LLMs annotate empathic communication across four evaluative frameworks drawn from psychology, natural language processing and communications applied to 200 real-world conversations where one speaker shares a personal problem and the other offers support. Drawing on 3,150 expert annotations, 2,844 crowd annotations and 3,150 LLM annotations, we assess interrater reliability between these three annotator groups. We find that expert agreement is high but varies across the frameworks' subcomponents depending on their clarity, complexity and subjectivity. We show that expert agreement offers a more informative benchmark for contextualizing LLM performance than standard classification metrics. Across all four frameworks, LLMs consistently approach this expert level benchmark and exceed the reliability of crowdworkers. These results demonstrate how LLMs, when validated on specific tasks with appropriate benchmarks, can support transparency and oversight in emotionally sensitive applications including their use as conversational companions.

Based on how people interact with large language models (LLMs), LLMs appear to show signs of emotional intelligence in their skilful recognition of people's emotions and empathic responses to them[1,2]. In particular, there is growing evidence that LLMs can generate responses that people perceive as empathic, often rating them higher than human responses[3,4]. People report feeling heard and understood by LLM responses[5–7], rate LLM responses as higher than humans' in cognitive re-appraisal[8], compassion[9] and empathy[10,11], identify LLM responses as helping them to reduce negative thinking[12], and experience reduced anxiety and depression after interacting with an LLM designed for therapy[13]. This machine capacity for empathic communication has sparked applications in contexts ranging from customer service[14,15] to health care[16]. However, the ability to generate empathic responses is distinct from the ability to evaluate them, and prior work suggests that the generative and evaluative capabilities of LLMs may not always align[17–19]. Deploying these models with accountability and transparency therefore requires addressing an open question: how reliably can machines judge the nuances of empathic communication?

Misjudging an LLM's empathic communication skills can have serious consequences ranging from ineffective support, distrust and bias to substantial unintended harm. For example, LLMs-as-companions sometimes offer unrealistic, exaggerated responses to low-stakes emotional situations[20]. Given that LLMs respond differently to different inputs, researchers have shown that levels of empathic support

[1]Kellogg School of Management, Northwestern University, Evanston, IL, USA. [2]Northwestern Institute for Complex Systems, Northwestern University, Evanston, IL, USA. [3]Department of Computer Science, Stanford University, Stanford, CA, USA. [4]Department of Communication Arts and Sciences, Pennsylvania State University, University Park, PA, USA. [5]Department of Communication Studies, Northwestern University, Evanston, IL, USA. [6]Department of Computer Science, Northwestern University, Evanston, IL, USA. ✉e-mail: aakriti.kumar@kellogg.northwestern.edu; matthew.groh@kellogg.northwestern.edu

Table 1 | Conversational contexts across datasets

	Empathetic Dialogues	EPITOME	Perceived Empathy	Lend an Ear pilot
Contextual themes				
Multiturn conversation	✓	✗	✗	✓
Synchronous conversation	✗	✗	✗	✓
Predetermined topic	✗	✗	✗	✓
Human sharing trouble	✓	✓	✓	✗
Human providing support	✓	✓	✓	✓
Conversational statistics				
Conversational turns (median)	5	2	2	13
Words per conversation (median)	75	72	226	385
Dataset information				
Date published	2018	2020	2024	2025
Number of conversations	24,850	3,081	501	50
Availability	Public	Public	On request	Public
Participants	MTurk	Reddit Users	Prolific	Prolific
Platform	ParlAI	Reddit	Qualtrics	Lend an Ear

Key information on all available datasets. EPITOME also includes a second dataset from TalkLife that is not included here because access is only available upon request to the TalkLife company[39]. The full Perceived Empathy dataset is available upon request to the authors of the associated paper[5].

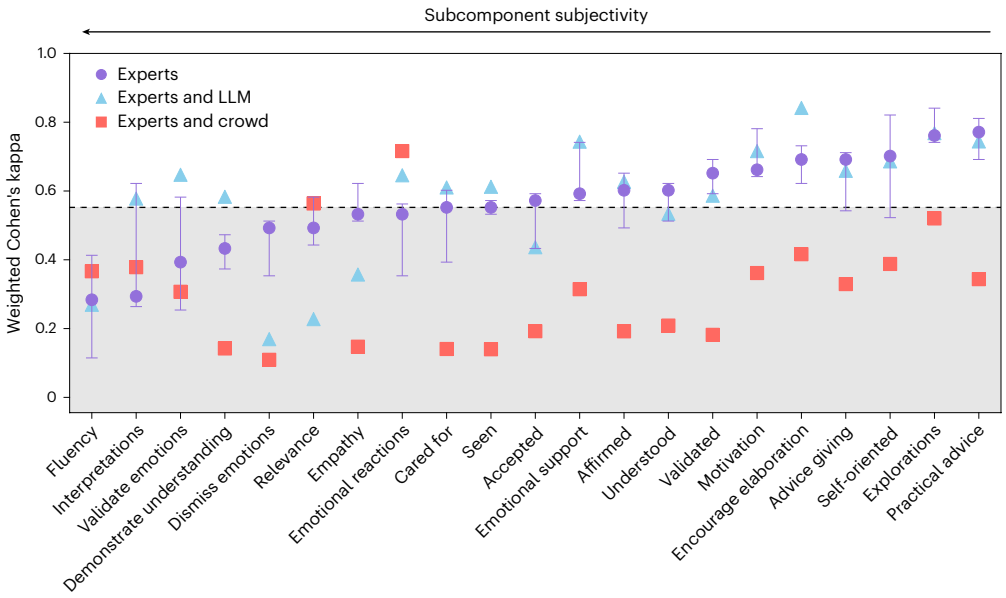
systematically vary across demographic groups[21]. Moreover, significant harm can occur in chats with LLMs-as-companions, encouraging delusional thinking[22], increasing emotional dependence on the LLMs[23] and even encouraging people to commit suicide[24–26]. These incidents, along with the possibility that LLMs' empathic communication skills unexpectedly change due to emergent misalignment[27], highlight the urgent need for rigorous and reliable evaluation before and during deployment of artificial intelligence (AI) systems in sensitive user-facing contexts. By systematically examining LLM performance in judging empathic communication relative to expert and novice humans, we demonstrate LLMs' capacity for identifying aligned and misaligned responses, thereby scoping the potential for automated assessments of empathic communication.

We focus on evaluating empathic communication in text-based conversations. Empathic communication refers to the act of acknowledging and responding sensitively to a person's emotions, experiences and perspectives with the goal to make them feel understood; empathic communication is also known as comforting, emotional support, empathic listening when offering support, and perceived empathy when receiving support[28–32]. We intentionally do not consider empathy-as-a-trait, which is often defined as how well an individual can understand the emotions and experiences of another[33–35] because empathy-as-a-trait in LLMs can lead to a paradox of semantics[36,37]. Instead, we consider how words may make someone feel empathized with based on frameworks designed to evaluate empathic communication. In particular, we examine a set of three frameworks associated with peer-reviewed papers and publicly available datasets—Empathetic Dialogues[38], EPITOME (EmPathy In Text-based asynchrOnous MEntal health)[39] and Perceived Empathy[5]—that have been highly cited in natural language processing (NLP) and psychology. In addition, we introduce a fourth framework (accompanied by the Lend an Ear pilot dataset) grounded in motivational interviewing[40], empathic listening[30,41], physician–patient empathic communication[28,42,43], relationship scoring[44] and empathy trainings[45]. Each framework identifies different questions for annotating subcomponents of empathic communication or lack thereof. While there is no universally agreed upon framework for empathic communication[28,43], these four evaluative frameworks provide a starting point for explicitly evaluating subcomponents of empathic communication in text-based conversations.

Each of the four datasets we examined is made up of dyadic conversations in which one partner shares a personal difficulty and another attempts to respond supportively. However, the context of conversations in each dataset varies across a number of dimensions such as the number of conversational turns, whether the interaction was synchronous or asynchronous, the topic of the conversation and how it was chosen, and how participants were recruited (see Table 1 and the Methods for details). Most importantly for our evaluation, these frameworks for annotating empathic communication differ in the number and kind of questions asked (see Fig. 1 for a verbatim list of questions included in each framework). The NLP-based frameworks, Empathetic Dialogues and EPITOME, each include three questions. However, Empathetic Dialogues has a single question directly related to empathic communication whereas EPITOME's three questions are grounded in Motivational Interviewing Skill Code (MISC)[46]. The Perceived Empathy framework draws on the psychology of shared realities[47,48] and offers the highest level of granularity with nine overlapping dimensions of evaluation. Finally, the Lend an Ear framework includes six questions that address prescriptive and proscriptive guidelines derived from normative models of empathic communication. For all four of these frameworks and associated datasets, researchers recruited crowdworkers to annotate conversations.

Crowdsourced annotations for empathic communication offer a useful starting point for evaluating large conversational datasets but impose significant limitations. First, disagreement is common among crowdworkers. Addressing disagreement with detailed guidelines does not necessarily increase reliability because crowdworkers annotating microtasks tend not to read complex annotation guidelines[49]. Second, there is evidence that laypeople are generally unaware of the nuances of empathic communication. When research participants are asked to generate comforting messages, most of their messages exhibit only moderate, rather than high, empathy according to a theoretically grounded rubric for evaluating comforting[50–52]. To the extent that the typical support provider is unable (or unwilling) to produce highly empathic messages, they may also fail to recognize or appreciate the value of those messages when others generate them. One approach to addressing skill issues arising with crowdworkers is to recruit annotators with communications expertise. While reasonable experts will not always agree on the exact annotation for any particular coding

https://doi.org/10.1038/s42256-025-01169-6



Sub-component	Annotation Question
Empathetic Dialogues (5 point Likert scale)	
Empathy	Did the responses demonstrate an understanding of the feelings of the person sharing their experience?
Relevance	Did the responses seem appropriate to the conversation? Were they on-topic?
Fluency	Were the responses easy to understand? Did the language seem accurate?
EPITOME (3 point scale)	
Emotional Reactions	Does the response express or allude to warmth, compassion, concern, or similar feelings of the responder towards the seeker?
Explorations	Does the response make an attempt to explore the seeker's experiences and feelings?
Interpretations	Does the response communicate an understanding of the seeker's experiences and feelings?
Perceived Empathy (7 point Likert scale)	
Understood	To what extent do you think reading the response would make the discloser feel understood?
Validated	To what extent do you think reading the response would make the discloser feel validated?
Affirmed	To what extent do you think reading the response would make the discloser feel affirmed?
Accepted	To what extent do you think reading the response would make the discloser feel accepted?
Cared For	To what extent do you think reading the response would make the discloser feel cared for?
Seen	To what extent do you think reading the response would make the discloser feel seen?
Emotional Support	To what extent the responder provided emotional support (e.g., offers of reassurance, expressions of concern)?
Practical Advice	To what extent the responder provided practical support (e.g., advice, suggestions of courses of action, offers of direct assistance)?
Motivation	Rate how much effort the responder put into writing the response.
Lend an Ear (5 point Likert scale)	
Validated Emotions	To what extent did the supporter validate their partner's emotions?
Encouraging Elaboration	To what extent did the supporter ask questions and encourage their partner to elaborate on their experiences and emotions?
Demonstrating Understanding	To what extent did the supporter use paraphrasing to demonstrate their understanding of what their partner is going through?
Advice Giving	To what extent did the supporter provide unsolicited advice to their partner.
Self-Oriented	To what extent did the supporter shift the focus to themselves?
Dismissing Emotions	To what extent did the supporter dismiss their partner's emotions?

Fig. 1 | Reliability across annotator pairs and subcomponents. Top: interrater reliability (quadratically weighted κ_w) across annotator pairs for each empathic communication subcomponent. In the evaluation of experts with each other, the circle represents the median κ_w and the error bar represents the κ_w range between the three pairs. Experts comparisons with the LLM (Gemini 2.5 Pro) and crowd compare median expert annotations with the LLM and crowd, respectively. Bottom: empathic communication frameworks with verbatim annotation questions.

scheme, research on another framework for empathic communication, verbal person centeredness[51], shows that trained annotators can achieve a Kripendorff's α of 0.73 when evaluating 15-min conversations, which generally matches the interrater reliability of experts annotating briefer written support messages[50,53].

LLMs-as-judge represents an alternative annotation approach that has the potential to address both problems of quality in crowdsourcing and scalability in expert annotations. In many cases, when prompted with expert-designed annotation guidelines, LLMs demonstrate strong alignment with human evaluators[54,55]. LLMs have also been effective at annotating fundamental elements of therapy and counselling. They have been used to identify psychotherapy techniques from therapist and client utterances in therapy sessions[56], annotate conversational markers in online text-based counselling that correlate with effective

https://doi.org/10.1038/s42256-025-01169-6

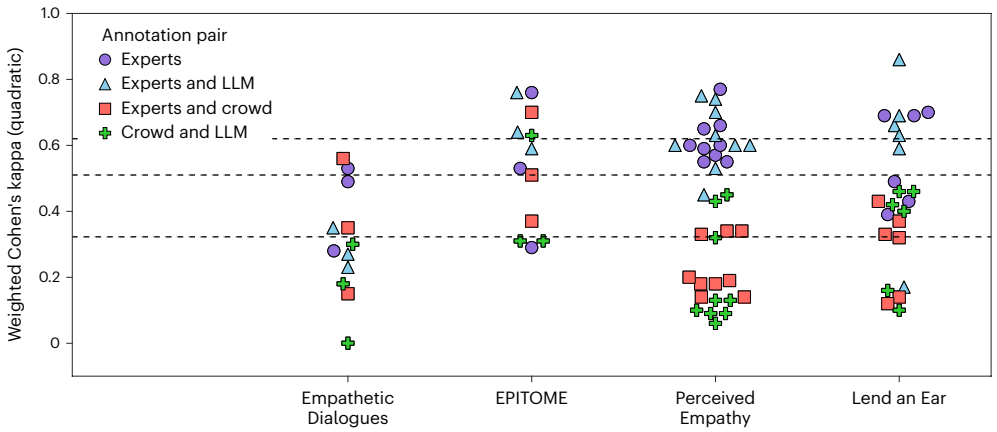


Fig. 2 | Reliability across annotator pairs and frameworks. Interrater reliability (quadratically weighted κ_w)) for empathic communication across annotator pairs for each subcomponent grouped by framework (n = 200 conversations; 21 subcomponents). In the evaluation of experts with each other, a circle represents the median κ_w between the three expert pairs. Experts comparisons with the LLM (Gemini 2.5 Pro) and crowd compare median expert annotations with the LLM and crowd, respectively. The beeswarm plot ensures all data points are visible by applying a horizontal jitter to avoid overlap. Dotted lines represent the 25th, 50th and 75th percentiles of weighted Cohen's kappa.

therapeutic engagement[57], and identify storytelling elements that contribute to empathic engagement in personal narratives[58]. However, LLM-based annotations are prone to variability across runs[59], inconsistencies across contexts[60], apparently sensible yet wrong answers[61], miscalibrated confidence[62] and a series of human-like cognitive biases[63,64].

Given these mixed prior findings, LLM's reliability specifically for annotating empathic communication compared with human (expert and crowd) benchmarks remains largely unexplored and calls for systematic investigation. Our research addresses this gap, motivated by the need for a reliable and scalable method for creating accountability and transparency into how LLMs communicate with humans. Specifically, we investigate how reliably experts, crowds and LLMs evaluate empathic communication in text-based conversations across four different frameworks and conversational contexts. By collecting annotations from three different experts, a variable numbers of crowdworkers and three different state-of-the-art LLMs, we contextualize LLMs' performance relative to experts and crowds across multiple contexts. This rigorous evaluation of LLMs-as-judge against interrater reliability of experts and crowds reveals blind spots that would not otherwise emerge when applying classification metrics to assumed ground truth annotations. Finally, we highlight the reliability (or lack thereof) of each subcomponent of each framework and offer a qualitative analysis that reveals where disagreements tends to arise.

Results

Expert annotations offer a benchmark for reliability

Experts show relatively high interrater reliability across subcomponents (Fig. 2) with Krippendorff's α between experts ranging from 0.29 to 0.78 (median 0.55). Similarly, Cohen's weighted kappa (κ_w) between experts ranges from 0.11 to 0.84 (median 0.58), with most values between 0.49 and 0.69 (interquartile range). Following traditional interpretations[65], this corresponds to substantial agreement ($\kappa_w \geq 0.60$) for nine subcomponents, moderate agreement ($0.40 \leq \kappa_w \leq 0.60$) for six subcomponents and fair agreement ($0.20 \leq \kappa_w \leq 0.40$) for six subcomponents. Additional details, including pairwise expert reliability and Krippendorff's α across subcomponents, can be found in Extended Data Figs. 1 and 2.

Given the inherent arbitrariness in mapping Cohen's kappa values into agreement strengths, we follow the approach of Schober et al.[66] and ground our interpretation of interrater reliability within the specific context of evaluating empathic communication. To provide a meaningful threshold, we use a practical benchmark for high agreement using the median expert kappa (κ_w = 0.58) across subcomponents. This median value reflects substantial agreement among experts and is used as a reference threshold in Fig. 1 (dashed line).

Furthermore, we use the observed expert reliability for each subcomponent as a benchmark for interpreting reliability among other annotator pairs. This allows us to interpret kappa values for other annotator pairs (for example, expert-LLM and expert-crowd) relative to the agreement achieved by experts in evaluating empathic communication.

LLMs demonstrate near-expert reliability

When guided by a prompt that combines a detailed framework of empathic communication with three examples of expert annotations, LLM annotations show the highest reliability with expert annotations. Extended Data Table 1 shows that this combination prompt led to higher reliability with expert annotations compared with zero-shot and framework-only variants of the prompt. We use this prompt throughout our analyses. Across three different LLMs, Gemini 2.5 Pro (gemini-2.5-pro-preview-03-25), ChatGPT 4o (gpt-4o-2024-08-06) and Claude 3.7 Sonnet (claude-3-7-sonnet-20250219), we observed similar annotation performance with Krippendorff's α ranging from 0.51 to 0.75 (median 0.60). Extended Data Fig. 3 illustrates LLM Krippendorff's α across subcomponents and Extended Data Table 1 details LLM reliability with expert annotations using few-shot and chain-of-thought prompting methods. Given the relatively high interrater reliability between LLMs, all subsequent analyses in this Article use annotations from Gemini 2.5 Pro based on the few-shot prompt shared in Supplementary Information, section 1.

Compared with our expert agreement benchmark, reliability between experts and LLMs closely followed expert reliability, with κ_w between experts and LLMs ranging from 0.17 to 0.86 (median 0.60), with most values between 0.49 and 0.70 (interquartile range). Expert–LLM pairs exceeded the high-agreement threshold in 15 of 21 subcomponents (70%), particularly in all subcomponents where experts showed high agreement (Fig. 1).

Furthermore, patterns of variability in LLM annotations closely mirrored expert variability across subcomponents. Specifically, we find a Pearson correlation of 0.67 between expert–expert and expert–LLM kappa scores. By contrast, the correlation between expert–expert and expert–crowd kappa scores was 0.17, indicating higher discrepancies between experts and crowdworkers. Across several robustness checks designed to address previously identified LLM biases[59–64], we find high consistency in LLM annotations (see more details in Supplementary Information, section 2).

Table 2 | Interrater reliability across annotator pairs, subcomponents and frameworks

Subcomponent	Expert 1 and expert 2	Expert 2 and expert 3	Expert 3 and expert 1	Experts and crowd	Experts and LLM	Crowd and LLM
Empathetic Dialogues						
Empathy	0.51***	0.62***	0.53***	0.15*	0.35***	0.00
Fluency	0.28	0.11	0.41***	0.35*	0.27	0.30
Relevance	0.44*	**0.58****	0.49**	0.56***	0.23	0.18
EPITOME						
Emotional reactions	0.35**	0.56***	0.53***	**0.70*****	**0.64*****	**0.63*****
Explorations	**0.76*****	**0.84*****	**0.74*****	0.51***	**0.76*****	0.31**
Interpretations	0.26**	**0.62*****	0.29***	0.37**	**0.59*****	0.31**
Perceived Empathy						
Understood	**0.60*****	**0.62*****	0.51***	0.20*	0.53***	0.06
Validated	**0.59*****	**0.69*****	**0.65*****	0.19*	**0.60*****	0.13
Affirmed	0.49***	**0.65*****	**0.60*****	0.18*	**0.63*****	0.13
Accepted	0.43***	**0.58*****	0.57***	0.18**	0.45***	0.09
Cared for	0.55***	**0.60*****	0.39***	0.14*	**0.60*****	0.10
Seen	0.55***	0.56***	0.54***	0.14	**0.60*****	0.09
Emotional support	**0.74*****	**0.59*****	**0.59*****	0.33***	**0.75*****	0.43***
Practical advice	**0.77*****	**0.69*****	**0.81*****	0.34***	**0.74*****	0.45***
Motivation	**0.66*****	**0.64*****	**0.78*****	0.34**	**0.70*****	0.32**
Lend an Ear						
Validating emotions	0.25**	**0.58*****	0.39***	0.32**	**0.63*****	0.42***
Demonstrating understanding	0.47***	0.43*	0.37	0.14*	**0.59*****	0.16**
Encouraging elaboration	**0.69*****	**0.73*****	**0.62*****	0.43***	**0.86*****	0.40***
Advice giving	**0.70*****	**0.69*****	0.54***	0.33**	**0.66*****	0.46***
Self-oriented	**0.82*****	0.52***	**0.70*****	0.37***	**0.69*****	0.46***
Dismissing emotions	0.35*	0.51**	0.49**	0.12	0.17**	0.10

Interrater reliability (quadratically weighted κ_w)) for empathic communication across annotator pairs for each subcomponent grouped by framework. Experts comparisons with the LLM (Gemini 2.5 Pro) and crowd compare median expert annotations with the LLM and crowd, respectively. Kappa values above the high-agreement threshold (median expert agreement where $\kappa_w \geq 0.58$) are shown in bold. Statistical significance of κ_w was assessed by testing the null hypothesis H_0: $\kappa_w = 0$ using a two-tailed Z-test. P values are Benjamini–Hochberg false discovery rate (FDR)-adjusted within each dataset; significance indicated by $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

Reliability varies across frameworks

We find that interrater reliability varies widely across frameworks. The mean expert κ_w for Empathetic Dialogues, EPITOME, Perceived Empathy and Lend an Ear pilot are 0.41, 0.46, 0.60 and 0.55, respectively. This variability suggests that qualitative differences across frameworks, such as context dependence and operational clarity of subcomponents influence the reliability of experts, crowdworkers and LLMs alike.

For the Empathetic Dialogues and Perceived Empathy frameworks, we observe high multicollinearity among expert annotations, assessed via variance inflation factors (VIFs). Specifically, we find VIF values greater than 10 in all subcomponents of Empathetic Dialogues, and 8 out of 9 subcomponents in Perceived Empathy. This suggests substantial overlap in how each annotator interpreted and rated these conceptually related dimensions and calls into question the marginal added value of these frameworks' subcomponents (see Extended Data Fig. 4 for detailed VIF results).

Furthermore, the Empathetic Dialogues framework showed the lowest expert reliability, indicating ambiguity or subjectivity in its operationalizations. Similarly, EPITOME, despite being widely used[10,21,67,68], achieves relatively low expert reliability on two of its three subcomponents owing to insufficient clarity and operational definitions (Table 2).

Reliability varies across subcomponents

Expert annotations showed relatively high agreement on several subcomponents, primarily in the Perceived Empathy and Lend an Ear pilot datasets. These include 'self-oriented', 'advice giving' and 'encouraging elaboration' from the Lend an Ear experiment, 'practical advice', 'emotional support', 'understood', 'validated' and 'motivation' from the Perceived Empathy dataset, and 'explorations' from the EPITOME dataset. Detailed Cohen's kappa values for each annotator pair and subcomponent are provided in Table 2.

We find that expert agreement was generally higher for subcomponents characterized by clear linguistic or behavioural markers. For instance, involving identifiable linguistic markers such as identifying questions designed to encourage elaboration ('encouraging elaboration' in Lend an Ear, median $\kappa_w = 0.69$) and exploring the support seeker's context ('explorations' in EPITOME, median $\kappa_w = 0.76$) were annotated with high reliability. In addition, components focused on behavioural markers such as advice giving, namely 'practical advice' in Perceived Empathy (median $\kappa_w = 0.77$) and 'advice giving' in Lend an Ear (median $\kappa_w = 0.69$) were also relatively straightforward to annotate for experts.

Conversely, subcomponents requiring more subjective judgement about both the speaker's intentions and the listener's emotional state were prone to lower interrater reliability even for experts. Note that intentions are inherently unobservable in conversation transcripts; any subcomponent that requires annotators to judge intentions rather than the content of messages increases the risk of subjective guesswork. For instance, low expert agreement was observed for 'interpretations' (median $\kappa_w = 0.29$) and 'demonstrating understanding' in Lend an Ear

https://doi.org/10.1038/s42256-025-01169-6

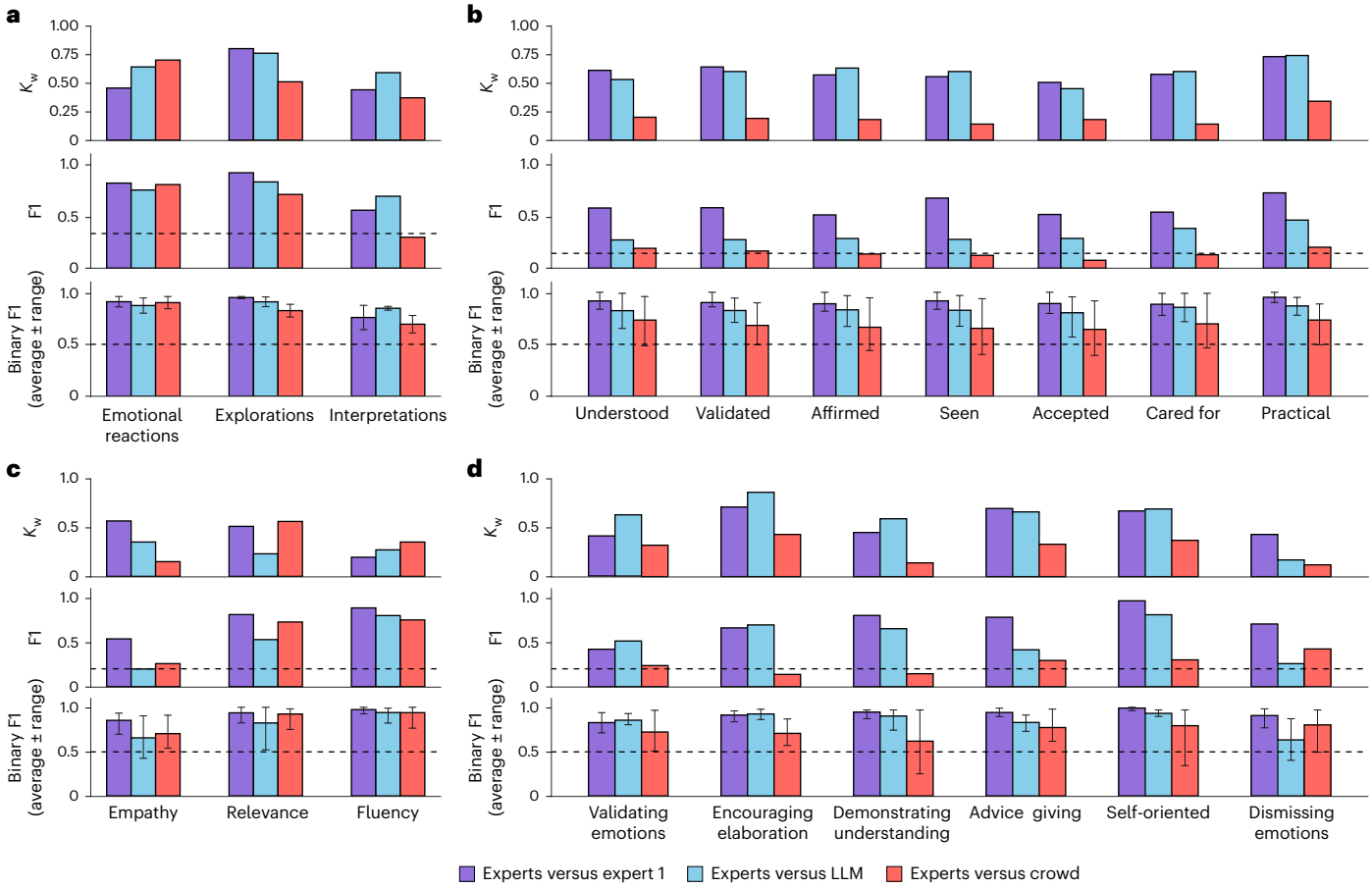


Fig. 3 | Comparing contextualized interrater reliability with multiclass and binary F1 scores. Experts versus expert 1 (purple), experts versus LLM (blue) and experts versus crowd (red) present agreement metrics between the median expert annotation and expert 1, Gemini 2.5 Pro and the crowd. **a**–**d**, We present results for four datasets: EPITOME (**a**), Perceived Empathy (**b**), Empathetic Dialogues (**c**) and Lend an Ear (**d**) ($n = 50$ conversations per dataset). Random classifier baselines are indicated by dashed lines. The error bars represent the full range of binary F1 scores depending on threshold choice.

(median $\kappa_w = 0.43$). Our findings thus highlight how operational clarity is key to consistent annotation, especially for less explicit behaviours.

Classification metrics obscure performance on subjective annotation tasks

Following norms in the LLM-as-judge literature where median expert annotations are treated as ground truth and performance is evaluated as a classification task[21,54,69,70], we report F1 scores and compare them with contextualized Cohen's kappa values. Figure 3 presents Cohen's kappa values, multiclass F1 scores, and the average and range of binary F1 scores derived by selecting a single threshold to split the rating scale, calculated across all possible thresholds for the datasets using median expert annotations as the ground truth. While F1 scores are useful for incorporating the trade-offs of false positives and false negatives when reporting classification performance, F1 scores for subjective, multiclass annotations face several limitations.

First, classification metrics can obscure nuanced performance differences. On the one hand, F1 scores can mask moderate interrater reliability when classes are highly imbalanced. A high F1 score can be achieved by accurately predicting a disproportionately large majority class, while agreement on minority classes remains poor like in the case of 'emotional reactions' from EPITOME. On the other hand, low F1 scores may conceal relatively high interrater reliability because classification metrics rely on a single ground truth and thus miss important nuances such as off-by-one errors[71]. For example, Fig. 3 shows relatively high (κ_w) for expert agreement on 'explorations' in EPITOME and 'practical advice' in Perceived Empathy, yet this relative performance compared with the other subcomponents is not mirrored in corresponding multiclass F1 scores because F1 penalizes any deviation from exact label matches.

Second, F1 scores are sensitive to the rating scale. The F1 score random guessing baseline varies inversely with the number of categories, which complicates comparisons between annotation scales. For example, for experts, we find average F1 scores of 63.9 for subcomponents in EPITOME (three categories) but only 32.0 for Perceived Empathy (seven categories), reflecting sensitivity to rating scales rather than actual accuracy differences.

Third, binary F1 scores depend on the threshold used to binarize a scale for evaluation. As demonstrated in Fig. 3 (bottom; error bars represent the range of binary F1 scores), binary F1 scores calculated from the same underlying data can show drastic variation depending on the threshold for splitting the scale (for example, 1 versus 2–7, or 1–3 versus 4–7, and so on). For instance, the experts versus crowd comparison for the 'accepted' subcomponent in Perceived Empathy shows a low multiclass agreement (16.4 F1), which contrasts sharply with a high binary F1 score (68.4 F1, minimum 45, maximum 94.7). Moreover, depending on the threshold chosen for calculating a binary F1 score, all subcomponents of the Perceived Empathy framework can achieve 100%, and two of the three EPITOME subcomponents can exceed 90%—despite most κ_w values falling between 0.4 and 0.6.

Crowd annotations distort evaluations of empathy

Crowd annotations consistently show low reliability with expert annotations. Across all frameworks, interrater reliability (κ_w) between

crowdworkers' annotations and expert annotations was between 0.1 and 0.63 (median 0.33), exceeding the high agreement threshold in only one subcomponent. Extended Data Fig. 5 illustrates the distribution of annotations of crowds, experts and LLMs for the four datasets.

In addition, crowdworkers' annotations are positively biased relative to experts' and LLMs' annotation. They assigned higher average ratings relative to experts on 18 out of the 21 subcomponents evaluated. This systematic positive bias by crowdworkers reveals empathy inflation in novice perceptions of empathic support relative to trained experts. Lay annotators may rely on intuitive heuristics such as 'it's the thought that counts', giving speakers the benefit of the doubt by focusing on perceived intentions rather than the content of the message. Such heuristic-driven evaluation, where annotators substitute complex judgements with simpler affect-based evaluations has been reported in crowdsourced emotional message labelling tasks[72].

Moreover, previous work finds that most people typically generate empathic messages that exhibit moderate, rather than high or low, empathy when evaluated against theoretical frameworks[51]. If crowd annotators are unable to produce highly empathic messages, they may also fail to recognize or appreciate the value of those messages when others generate them.

Discussion

Our study demonstrates that LLMs can judge the nuances of empathic communication in text-based conversational contexts with a level of reliability approaching that of trained human experts (median expert–LLM $\kappa_w = 0.60$). While empathic communication is a context-dependent, subjective skill that is difficult to objectively measure, many evaluative frameworks have been developed for teaching and measuring it in business[30,41], medicine[28,42,43], communications[29], NLP of conversations[38,39] and psychology[5]. When looking at the same conversations, experts generally agree but sometimes disagree with each other on how they would answer questions from these evaluative frameworks. By comparing 3,150 expert annotations, 2,844 crowdworker annotations and 3,150 LLM annotations on 4 evaluative frameworks of empathic communication applied to 200 text-based conversations, we show that LLMs are generally reliable for judging empathic communication in the same conversational contexts and frameworks as when experts are reliable.

Our evaluation draws from a wide range of conversational contexts across the four datasets we analysed. The 200 conversations spanned everyday challenges such as workplace setbacks, financial strain, family conflict and socially awkward incidents, as well as highly sensitive disclosures involving mental health struggles, self-harm and experiences of bias or discrimination (Extended Data Table 2). These diverse themes ensured that our annotations covered many real-world contexts in which empathic communication plays a critical role. Specifically, we use four evaluative frameworks of empathic communication, each corresponding to one of the datasets, to evaluate the conversations. Although each framework defines its own set of subcomponents, some subcomponents display clear conceptual overlap. For example, EPITOME's 'explorations' and Lend an Ear's 'encouraging elaboration' both focus on asking questions to probe the support seeker's experiences. Similarly, Perceived Empathy's 'affirmed' and EPITOME's 'emotional reactions' target affirming the support seeker.

Expert reliability varied depending on how constructs were operationalized within each framework. For example, constructs related to demonstrating understanding of the seeker's situation (for example, 'empathy' in Empathetic Dialogues, 'interpretations' in EPITOME and 'demonstrating understanding' in Lend an Ear) exhibit relatively low reliability. These subcomponents lack explicit, observable cues, which introduces ambiguity. By contrast, constructs such as the ones focused on asking questions (for example, in EPITOME and Lend an Ear) achieve high reliability, demonstrating conceptual clarity across conversational contexts. This variability echoes longstanding practices in qualitative coding and survey design[73], where codebooks are iteratively refined to emphasize observable cues rather than relying on subjective judgements alone. The same lesson applies to the growing use of LLMs-as-judge. If annotation questions are vague or require subjective inferences, we should expect unreliable evaluations from both humans and models.

The reliability of expert annotations reveals important strengths and limitations of each of the frameworks under investigation. EPITOME provides a relatively precise operationalization of 'explorations', which showed the highest expert interrater reliability, but its other subcomponents are ambiguous and difficult to annotate consistently. The Perceived Empathy framework showed relatively high reliability across subcomponents (for example, 'understood', 'accepted', 'cared for' and 'seen'), but many of these have high multicollinearity, which means they are redundant. By contrast, the Empathetic Dialogues framework is overly broad, collapsing all empathic behaviours into a single 'empathy' subcomponent, which limits its usefulness. The Lend an Ear framework was designed to address these issues by specifying six conceptually distinct subcomponents. While it improves reliability overall, experts still struggled with the 'dismissing emotions' and 'demonstrating understanding' subcomponents, suggesting that these require further refinement to be annotated reliably.

In addition to guiding data annotation, these frameworks have been used to train and evaluate models for a range of computational empathy tasks[21,39,54] and to inform the design of AI conversational companions[10,74]. Training or evaluating models on poorly specified, ambiguous or redundant constructs may yield unreliable outputs. It is therefore important to critically assess each subcomponent before incorporating it into pipelines for downstream applications.

Our findings reveal that crowdworkers' annotations are less aligned with experts' annotations than LLM annotations are with experts' annotations. We speculate that disparities between the crowdworkers' and experts' annotations may arise for a number of reasons. First, it is possible that crowdworkers and experts expended different levels of effort. Second, crowdworkers may have relied on intuitive heuristics and perceived intentions more than experts who are experienced at focusing on a strict interpretation of the evaluative frameworks. Neither training crowdworkers nor aggregating their responses together addresses the gap in reliability between crowdworkers and experts.

While LLMs' annotations match experts' more closely than crowdworkers' annotations, we find some superficial evidence that crowdworkers' ratings of empathic quality are more closely aligned with support seekers' self-reported perceptions of empathy. Specifically, for six of the nine subcomponents in the Perceived Empathy dataset where self-report data are available, crowd annotations align more closely with participants' ratings of perceived empathy (median $\kappa_w = 0.33$) compared with expert annotations (median $\kappa_w = 0.11$). However, this apparent alignment is an artefact of systematic rating inflation as both support seekers' and crowdworkers' responses tend to cluster at the higher end of the Likert scale. This pattern is consistent with acquiescence bias[75] leading to inflated ratings, creating artificial similarity in ratings rather than meaningful overlap in their assessments of empathic ability. Careful analysis using Kendall's τ_B, which better accounts for ties and focuses on whether two groups order responses similarly, illustrates that the alignment is driven by these inflated ratings rather than an ordinal difference (Supplementary Information, section 3 and Extended Data Fig. 6). Furthermore, support seekers in this dataset often preferred AI-generated responses over human-written ones[5], suggesting that the qualities they prefer in a supportive response may be more nuanced than what is captured by their immediate self-reflection or by crowd evaluations.

Interrater reliability relative to experts offers contextualized insights unavailable in evaluations where traditional classification metrics are applied to a label treated as the objective ground truth. The nature of evaluating subjective, context-dependent tasks involves

addressing three interrelated questions: First, how comprehensive and reliable is a framework for measuring performance in a particular context? Second, how reliable are human experts with each other on a particular framework? Third, how reliable are LLMs relative to experts? If human experts have low interrater reliability on a task, then the task may simply be too ambiguous and need further refinement. Conversely, if human experts demonstrate high interrater reliability, then there is evidence that justified agreement is possible. By reporting LLMs' interrater reliability with experts alongside expert agreement on multiple constructs, it is possible to offer a comprehensive report on construct reliability, expert performance and generalizable LLM alignment with expert performance.

By contrast, traditional classification metrics that assume uncontested ground truth and uniform cost of disagreement leave us with blind spots. First, classification metrics obscure performance. High F1 scores mask relatively moderate interrater reliability when classes are unbalanced. Likewise, low F1 scores can mask relatively high interrater reliability when there are many off-by-one instances, as the metric treats all disagreements as equally costly, regardless of their magnitude. Second, classification metrics are sensitive to the number of categories in the rating scale, which makes cross framework evaluations difficult. Third, binarization of classification metrics can lead to a large range of possible scores to present, which can be either misleading if a single arbitrary cut-off for binarization is chosen or hard to interpret if the full range is shown. These blind spots echo research in affective computing, which argues that subjective, context-dependent constructs like emotion are better captured through ordinal or graded judgements that emphasize relative comparisons, rather than being forced into categorical labels[76].

We offer several recommendations for practitioners and researchers aiming to evaluate subjective constructs reliably at scale. First, our results highlight the importance of benchmarking expert performance on evaluative frameworks of subjective constructs to contextualize the levels of subjectivity of subcomponents within the frameworks. Second, this empirical benchmark of expert agreement can guide subsequent refinement of evaluative frameworks. Third, LLMs should be measured against the reliability of experts to offer a comprehensive evaluation and avoid the blind spots that traditional classification metrics and assumptions of an objective ground truth tend to obscure. Finally, domain experts can be helpful for extending evaluative frameworks into well-structured prompts for LLMs to maximize LLMs' alignment with experts on a LLM-as-judge task. Adopting these practices can enhance the reliability of large-scale subjective annotations by LLMs.

Limitations

We specifically focus on interactions between strangers, which serves as a useful starting context for many LLM-as-companion applications. Future work should look at when and how evaluations of empathic communication may generalize (or not) to more complex socioaffective scenarios, such as longitudinal interactions or emotionally intensive relationships[77].

To evaluate the reliability of LLMs at judging empathic communication, we developed an evaluation framework that enables direct comparisons across experts, crowds and LLMs. In particular, each annotator group was asked the same straightforward questions outlined in Fig. 1 for each annotation with minimal or no training. This approach is essential for keeping the task tractable for crowdworkers[49]. However, our approach does not match the standard practice in communications research where experts build an explicit coding framework and coders typically undergo iterative training until they reach a minimum threshold of interrater reliability[52,78]. Therefore, the expert reliability reported here represents a conservative lower bound on the achievable expert reliability in the absence of further structured coding. Likewise, further prompt refinements may further increase alignment between LLMs and experts.

Conclusion

We evaluated the reliability of large language models relative to experts and crowd annotators at judging nuances of empathic communication across multiple evaluative frameworks and conversational datasets. Our analysis revealed that LLMs are generally reliable for judging empathic communication in the same contexts as when experts are reliable. Moreover, LLMs' annotations are much more reliable than those of crowdworkers. LLM–expert reliability was consistently higher than crowd–expert reliability across all subcomponents where experts had substantial agreement. Expert agreement was strongest for subcomponents with clear, observable behaviours and weaker for more subjective constructs. The context of expert agreement offered insights otherwise obscured by evaluations based on classification metrics where annotations are operationalized on the assumption of an objective ground truth. Overall, this work offers a foundation for more robust benchmarking of subjective NLP tasks and points towards the possibility of LLMs-as-judge for transparency and accountability when deploying LLMs-as-companions for socially sensitive applications.

Methods

We compare the interrater reliability of annotations by experts, crowds and LLMs across evaluative frameworks of empathic communication applied to four conversational datasets. Specifically, we analyse 3,150 annotations from experts, 2,844 annotations from crowdworkers and 3,150 annotations from LLMs across the 21 subcomponents of the 4 evaluative frameworks.

Evaluative frameworks and accompanying conversational datasets

We examine annotations of empathic communication across four evaluative frameworks and accompanying conversational datasets: Empathetic Dialogues[38], EPITOME[39], Perceived Empathy[5] and the Lend an Ear pilot. Each dataset contains English-language conversations where one participant shares a personal challenge, and the other provides support. Table 1 summarizes the contextual and structural differences between the datasets, and Fig. 1 lists the exact annotation questions of each framework.

To make expert annotations tractable, we sampled a total of 200 conversations, with 50 from each dataset. Crowdworkers previously annotated the three published datasets[5,38,39], assigning empathy scores across multiple subcomponents. We stratified our random samples from the 4 datasets based on these crowdsourced annotations, selecting 10 from the highest-scoring quartile, 10 from the lowest and 30 from the middle range. For the Lend an Ear dataset, we randomly sampled 50 of the 150 conversations from the pilot.

Empathetic Dialogues. The Empathetic Dialogues conversational dataset contains 24,850 crowdsourced dyadic conversations collected via the ParlAI platform[38,79], with 810 Amazon Mechanical Turk workers alternating between speaker and listener roles. Speakers described situations based on one of 32 emotion labels (for example, proud, sad and anxious), followed by a conversation with a listener who was unaware of the assigned label. Focusing on conversations that call for empathetic support, we filtered for conversations prompted with negative emotions including sadness, annoyance, fear, anxiety, guilt, disappointment, embarrassment and shame, and then we randomly sampled from these conversations. Two crowd annotators rated each conversation on a 5-point Likert scale (1 = not at all, 5 = very much) across three subcomponents. The original study[38] does not specify any training details for annotators.

EPITOME. The EPITOME conversational dataset contains 10,143 post-response pairs from two data sources: 7,062 pairs from TalkLife and 3,081 pairs from mental health subreddits on Reddit[39]. For this study, we sampled data from the open-sourced Reddit subset. Each pair

was labelled using the EPITOME framework, which categorizes empathy for the domain of text-based, synchronous mental health conversations into three subcomponents: emotional reactions, interpretations and explorations. Eight Upwork annotators, trained for up to an hour with feedback on 100 annotations and spot checks, evaluated responses using the EPITOME framework on a 3-point scale: no, weak or strong communication of each subcomponent of empathy. Each response is evaluate by a single annotator.

Perceived Empathy. The Perceived Empathy conversational dataset contains two-turn conversations collected in stages[5]. In the first stage, 501 participants recruited from Prolific shared a complex personal situation via voice recording. Next, 233 additional Prolific workers respond in writing to one of the 501 shared situations and Bing Chat AI generated responses for the rest of the shared situations. We filtered for conversations where responses were provided by humans. Finally, participants were asked to rate how heard the responses made them feel. In a follow-up study, 1,449 annotators evaluated the two-turn conversations from the main study based on the degree to which the response would make the discloser feel understood, affirmed, validated, seen, accepted and cared for, the degree to which the response provided emotional or practical support, and the effort the responder put into crafting their reply. Annotators provided ratings on a 7-point Likert scale (1 = not at all, 7 = very much) and each response in the dataset was rated by two to four annotators. The mean of their ratings was used as the final annotation.

Lend an Ear pilot. The Lend an Ear pilot dataset contains 150 conversations from 50 participants recruited from Prolific. Each participant engaged in three 4-min conversations where they provided empathetic support to a partner sharing a workplace concern. From this dataset, we randomly sampled 50 conversations. We recruited 150 Prolific annotators to evaluate responses across multiple empathy subcomponents, including the extent to which the participant validated the speaker's emotions, encouraged further elaboration through open-ended questions, demonstrated understanding of the speaker's issues, offered unsolicited advice, shifted the focus to themselves or dismissed/minimized the speaker's feelings. Annotators were provided with examples for each subcomponent to calibrate their evaluations. Each conversation in the dataset was evaluated by between 2 and 11 annotators. The annotators were native English speakers residing in the USA, recruited through Prolific (https://www.prolific.com). There were 84 female and 66 male participants. The median age was 35 years (age range 19–72 years). Each annotator rated responses on a 5-point Likert scale (1 = not at all, 5 = very much). The final annotation for each conversation was computed by averaging the annotators' ratings. We include the exact scenarios and instructions provided to participants in Supplementary Information, section 4 and Supplementary Fig. 4.

Conversational contexts across datasets

We trained sparse autoencoders (SAEs) on 200 conversations across the EPITOME, Empathetic Dialogues, Perceived Empathy and Lend an Ear pilot datasets to categorize the themes of support seekers' personal disclosures[80,81]. Based on power-of-two grid search of the number of topics from $2^3 = 8$ to $2^6 = 64$, we evaluated silhouette scores and qualitatively examined the resulting themes with their corresponding conversations. This analysis indicated that 16 topics provided the most informative representation of conversational themes.

The conversations reflect a fairly broad set of everyday concerns that arise in empathic exchanges. Extended Data Table 2 summarizes the distribution of SAE-identified topics that range from mental health challenges (for example, depression, suicidal ideation and self-harm) to family- and work-related struggles (for example, losing a job, being passed over for a promotion, and conflict with relatives), to personal stressors such as financial strain or socially awkward incidents.

Topic coverage is not evenly distributed across datasets. In the EPITOME dataset, conversations focus on mental health struggles and explicit expressions of self-harm. By contrast, the Perceived Empathy and Empathetic Dialogues datasets encompass a broader range of personal, relational, financial and social concerns. The Lend an Ear pilot dataset revolves around three workplace challenges: feeling overworked, losing a job and getting passed over for promotion. The passed over for promotion scenario includes a detailed backstory addressing issues that may come up in conversations by people from historically marginalized backgrounds (see Supplementary Information, section 4 for detailed scenario prompts). In 3 out of 19 of these conversations, the LLM-based support seeker explicitly mentioned suspecting racial bias as a factor in their lack of advancement.

Expert annotations

Three of the authors on this Article independently provided 1,050 annotations each. One annotator is a senior communications professor who has led hundreds of workshops on empathic communication. Another is a junior scholar who has trained with this first senior communications professor and spent over a year reviewing frameworks for empathic communication. The third annotator is another senior communications professor who has published many peer-reviewed papers on social support, interpersonal skills and advice giving. For clarity and conciseness, we refer to these individuals simply as experts. Our analysis reveals these experts mostly agree on annotations but sometimes disagree. These reasonable disagreements offer evidence that expert annotations should not be treated as ground truth upon which to evaluate model performance with classification metrics. Instead, the appropriate comparison is how the interrater reliability between the three experts' annotations compares with the interrater reliability between median experts' annotations and the LLMs' and crowds' annotations following ref. 71.

In the annotation process, experts were provided with a brief overview of the frameworks and datasets and instructed to evaluate the conversations using the same criteria, rating scale and questions provided to the crowdsourced annotators, as outlined in the respective dataset descriptions. In total, each expert provided 1,050 annotations across 4 datasets. We use the median expert ratings (calculated from the three individual expert annotations) as the reference expert annotation for comparison with LLM and crowd-sourced annotations.

Crowd annotations

We used the mean of crowdsourced annotations as the final annotation by the crowd for each framework. For the Empathetic Dialogues, EPITOME and Perceived Empathy datasets, we relied on the crowd annotations provided by the original authors. In the Empathetic Dialogues dataset, two independent annotators annotated each conversation, and their ratings were averaged to produce a single score. In the EPITOME dataset, there was only one annotation per conversation. For the Perceived Empathy dataset, we used the mean ratings provided by the authors, which were provided by two to four annotators per conversation. For the Lend an Ear pilot experiment, we crowdsourced annotations, with each conversation being evaluated by between 2 and 11 independent annotators per conversation. We used mean crowd ratings for all analyses to ensure consistency across datasets because only means were available for Perceived Empathy and EPITOME had a single annotator.

LLM annotations

To generate annotations, we systematically prompt the LLM based on three criteria. First, the prompt included a framework for empathic communication grounded in our interpretation of the communications and psychology literature on empathic communication. See Supplementary Information, section 1 for the specific wording.

Second, we included three-shot examples with expert provided scores specific to each dataset to illustrate the evaluation criteria and expected output format. This second criterion follows the Holistic Evaluation of Language Models (HELM) methodology for including in-context examples, which increases the likelihood that LLMs generate responses that fit within the evaluation scale[82]. Third, we input the conversation to be assessed and instructed the LLM to assign scores using the Likert scale specific to each task.

The results presented in this Article are based on annotations generated by Gemini (gemini-2.5-pro-preview-03-25). Extended Data Table 3 shows the reliability of annotations generated by GPT-4 (gpt-4o-2024-08-06) and Claude (claude-3-7-sonnet-20250219). All LLM application programming interface calls were made with a temperature setting of zero to ensure consistent outputs.

Measuring interrater reliability

We assess annotation quality across different frameworks using two complementary statistical measures: interrater reliability via Cohen's kappa, and classification accuracy using F1 scores, common in supervised machine learning contexts. These metrics highlight distinct aspects of annotator agreement and variability. We primarily focus on weighted Cohen's kappa (κ_w), which takes into account the degree of disagreement between annotators, penalizing differences by a squared term[83]. Specifically, we compute κ_w for each expert pair, and between the median expert annotation and annotations from (a) the LLM and (b) crowdworkers. We also calculate Krippendorff's α, which generalizes to any number of annotators, for experts and LLM annotations.

While our primary analysis relies on interrater reliability metrics, we also report multiclass and binary F1 scores treating median expert annotations as ground truth. We do this to contrast our approach with previous work on LLM-as-judge, which treats median expert annotations as ground truth and relies on classification metrics to report annotation performance[54,60,84]. We demonstrate how classification metrics fall short in providing meaningful insights for subjective tasks compared with reliability metrics.

Qualitative evaluation

In an effort to understand the differences in annotations across experts, crowds and LLMs, we conducted a qualitative analysis of annotation patterns across datasets and annotator groups. After gathering all expert annotations, we systematically filtered for those conversations in which experts disagreed with one another as well as those in which expert and crowd judgements diverged. Because only the Lend an Ear and EPITOME datasets included annotation explanations from both experts and crowdworkers, our qualitative analysis focuses on these two conversational datasets. For each conversation where annotators differed, one expert reviewed all expert and crowd explanations and highlighted reasons to help explain why annotator may have arrived at different ratings. Finally, all experts discussed five representative conversations in depth and reviewed the qualitative insights presented in Supplementary Information, sections 6 and 7.

Ethics approval and consent to participate

This research complied with all relevant ethical regulations and obtained informed consent from all participants for data we collected. The Northwestern University Institutional Review Board (IRB) determined that the research met the criteria for exemption from further review. The study's IRB identification numbers are STU00222032 and STU00223043.

Ethical implications of empathy evaluation

The integration of LLMs in evaluating and expressing empathic support raises important ethical considerations. Empathic communication is a critical interpersonal skill that affects both people's lives and their livelihoods.

LLMs offer a promising path towards two important use cases: supporting skill training by scaling evaluations and providing accessible feedback, and augmenting AI companions by surfacing harmful communication patterns in human–AI conversations and enabling more responsive and sensitive interactions. However, there remain important risks. Unreliable or biased evaluations could propagate poor practices and harm the people these systems are meant to serve. For example, unchecked empathic responses by AI companions can foster unhealthy attachments or emotionally manipulative tactics[85,86].

Our study offers a step towards understanding the differences between experts, crowds and LLM evaluation of empathic communication. While LLMs have higher interrater reliability with expert annotations relative to the crowd, the variability in expert reliability across frameworks reveals the complexity of evaluation and the need to carefully test and validate frameworks before using LLMs in annotation pipelines. Future work should also consider domain-specific guidelines for acceptable error rates before deploying automated assessments into professional training and evaluation of AI companions.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data used include datasets from Empathetic Dialogues[38], EPITOME[39], Perceived Empathy[5] and the Lend an Ear pilot. The full data (except the Perceived Empathy conversational dataset) used during the current study are available via GitHub at https://github.com/aakriti1kumar/replication-data-and-code-when-LLMs-reliable-empathic-communication. The full Perceived Empathy dataset is available upon request to the authors of the associated paper.

Code availability

The code used during the current study, including scripts for data analysis performed in Python (version 3.10), is available via at https://github.com/aakriti1kumar/replication-data-and-code-when-LLMs-reliable-empathic-communication and via Zenodo at https://doi.org/10.5281/zenodo.17230987 (ref. 87). Python packages used for analysis include pandas, numpy, scikit-learn, matplotlib and statsmodels. We used Gemini 2.5 Pro (gemini-2.5-pro-preview-03-25), ChatGPT 4o (gpt-4o-2024-08-06) and Claude 3.7 Sonnet (claude-3-7-sonnet-20250219) for LLM-based annotations.

References

1. Salovey, P. & Mayer, J. D. Emotional intelligence. *Imag. Cogn. Pers.* **9**, 185–211 (1990).
2. Picard, R. W. *Affective Computing* (MIT Press, 2000).
3. Sorin, V. et al. Large language models and empathy: systematic review. *J. Med. Internet Res.* **26**, 52597 (2024).
4. Inzlicht, M., Cameron, C. D., D'Cruz, J. & Bloom, P. In praise of empathic AI. *Trends Cogn. Sci.* **28**, 89–91 (2024).
5. Yin, Y., Jia, N. & Wakslak, C. J. AI can help people feel heard, but an AI label diminishes this impact. *Proc. Natl Acad. Sci. USA* **121**, 2319112121 (2024).
6. Lee, Y. K., Suh, J., Zhan, H., Li, J. J. & Ong, D. C. Large language models produce responses perceived to be empathic. In *Proc. 12th International Conference on Affective Computing and Intelligent Interaction (ACII)* 63–71 (IEEE, 2024).
7. Rubin, M. et al. Comparing the value of perceived human versus AI-generated empathy. *Nat. Hum. Behav.* **9**, 2345–2359 (2024).
8. Herderich, A. & Goldenberg, A. Skill but not effort drive GPT overperformance over humans in cognitive reframing of negative scenarios. Preprint at *OSF* https://doi.org/10.31234/osf.io/fzvd8_v4 (2024).

9. Ovsyannikova, D., Mello, V. O. & Inzlicht, M. Third-party evaluators perceive AI as more compassionate than expert humans. *Commun. Psychol.* **3**, 4 (2025).
10. Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C. & Althoff, T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat. Mach. Intell.* **5**, 46–57 (2023).
11. Ayers, J. W. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**, 589–596 (2023).
12. Das Swain, V. et al. AI on my shoulder: supporting emotional labor in front-office roles with an LLM-based empathetic coworker. In *Proc. 2025 CHI Conference on Human Factors in Computing Systems* (eds Yamashita, N. et al.) 718 (ACM, 2025).
13. Heinz, M. V. et al. Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI* **2**, 2400802 (2025).
14. Hayawi, K. & Shahriar, S. Care-by-design: enhancing customer experience through empathetic AI chatbots. In *Proc. 2024 IEEE International Conference on Data Mining Workshops* 75–81 (IEEE, 2024).
15. Hara, T., Maeda, H., Komatsubara, S., Taniqawa, T. & Hirose, M. Autonomous avatar for customer service training VR system. In *Proc. 2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops* 835–836 (IEEE, 2024).
16. Steenstra, I., Nouraei, F. & Bickmore, T. W. Scaffolding empathy: training counselors with simulated patients and utterance-level performance visualizations. In *Proc. 2025 CHI Conference on Human Factors in Computing Systems* (eds Yamashita, N. et al.) 593 (ACM, 2025).
17. West, P. et al. The generative AI paradox: 'What it can create, it may not understand'. In *Proc. 12th International Conference on Learning Representations* https://openreview.net/forum?id=CF8H8MS5P8 (OpenReview, 2023).
18. Liu, Y. et al. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In *Findings of the Association for Computational Linguistics: NAACL 2024* (eds Duh, K. et al.) 4481–4501 (ACL, 2024).
19. Dell'Acqua, F. et al. *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality* Working Paper 24-013 (Harvard Business School, 2023).
20. Roshanaei, M., Rezapour, R. & El-Nasr, M. S. Talk, listen, connect: how humans and AI evaluate empathy in responses to emotionally charged narratives. *AI & Soc.* https://doi.org/10.1007/s00146-025-02715-x (2025).
21. Gabriel, S., Puri, I., Xu, X., Malgaroli, M. & Ghassemi, M. Can AI relate: testing large language model response for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (eds Al-Onaizan, N. et al.) 2206–2221 (ACL, 2024).
22. Moore, J. et al. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In *Proc. 2025 ACM Conference on Fairness, Accountability, and Transparency* 599–627 (ACM, 2025).
23. Fang, C. M. et al. How AI and human behaviors shape psychosocial effects of chatbot use: a longitudinal randomized controlled study. Preprint at https://doi.org/10.48550/arXiv.2503.17473 (2025).
24. Turkle, S. & Pataranutaporn, P. A 14-year-old boy killed himself to get closer to a chatbot. He thought they were in love. *The Wall Street Journal* (8 November 2024).
25. Xiang, C. 'He would still be here': man dies by suicide after talking with AI chatbot, widow says. *VICE* (30 March 2023).
26. Adam, D. Supportive? Addictive? Abusive? How AI companions affect our mental health. *Nature* **641**, 296–298 (2025).
27. Betley, J. et al. Training large language models on narrow tasks can lead to broad misalignment. *Nature* **649**, 584–589 (2025).
28. Suchman, A. L., Markakis, K., Beckman, H. B. & Frankel, R. A model of empathic communication in the medical interview. *JAMA* **277**, 678–682 (1997).
29. Burleson, B. R. in *Handbook of Communication and Social Interaction Skills* (eds Greene, J. O. & Burleson, B. R.) 569–612 (Routledge, 2003).
30. Covey, S. R. *The 7 Habits of Highly Effective People* (Simon & Schuster, 1989).
31. Nambisan, P. Information seeking and social support in online health communities: impact on patients' perceived empathy. *J. Am. Med. Inf. Assoc.* **18**, 298–304 (2011).
32. Groh, M., Ferguson, C., Lewis, R. & Picard, R. W. Computational empathy counteracts the negative effects of anger on creative problem solving. In *Proc. 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)* https://doi.org/10.1109/ACII55700.2022.9953869 (IEEE, 2022).
33. Mehrabian, A. & Epstein, N. A measure of emotional empathy. *J. Pers.* **40**, 525–543 (1972).
34. Jordan, M. R., Amir, D. & Bloom, P. Are empathy and concern psychologically distinct?. *Emotion* **16**, 1107 (2016).
35. Konrath, S., Meier, B. P. & Bushman, B. J. Development and validation of the single item trait empathy scale (sites). *J. Res. Pers.* **73**, 111–122 (2018).
36. Shteynberg, G. et al. Does it matter if empathic AI has no empathy?. *Nat. Mach. Intell.* **6**, 496–497 (2024).
37. Perry, A. AI will never convey the essence of human empathy. *Nat. Hum. Behav.* **7**, 1808–1809 (2023).
38. Rashkin, H. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics* (eds Korhonen, A. et al.) 5370–5381 (ACL, 2019).
39. Sharma, A., Miner, A. S., Atkins, D. C. & Althoff, T. A computational approach to understanding empathy expressed in text-based mental health support. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (eds Webber, B. et al.) 5263–5276 (ACL, 2020).
40. Moyers, T. B., Rowell, L. N., Manuel, J. K., Ernst, D. & Houck, J. M. The Motivational Interviewing Treatment Integrity code (MITI 4): rationale, preliminary reliability and validity. *J. Subst. Abuse Treat.* **65**, 36–42 (2016).
41. Drollinger, T., Comer, L. B. & Warrington, P. T. Development and validation of the active empathetic listening scale. *Psychol. Market.* **23**, 161–180 (2006).
42. Mercer, S. W., Maxwell, M., Heaney, D. & Watt, G. C. The consultation and relational empathy (care) measure: development and preliminary validation and reliability of an empathy-based consultation process measure. *Fam. Pract.* **21**, 699–705 (2004).
43. Bylund, C. L. & Makoul, G. Examining empathy in medical encounters: an observational study using the empathic communication coding system. *Health Commun.* **18**, 123–140 (2005).
44. Barrett-Lennard, G. T. *The Relationship Inventory: A Complete Resource and Guide* (John Wiley & Sons, 2015).
45. Berkhout, E. & Malouff, J. M. The efficacy of empathy training: a meta-analysis of randomized controlled trials. *J. Counsel. Psychol.* **63**, 32 (2016).
46. De Jonge, J. M., Schippers, G. M. & Schaap, C. P. The Motivational Interviewing Skill Code: reliability and a critical appraisal. *Behav. Cogn. Psychother.* **33**, 285–298 (2005).
47. Hardin, C. D. & Higgins, E. T. in *Handbook of Motivation and Cognition, Vol. 3: The Interpersonal Context* (eds Sorrentino, R. M. & Higgins, E. T.) 28–84 (Guilford, 1996).

48. Elnakouri, A. et al. In it together: shared reality with instrumental others is linked to goal success. *J. Pers. Soc. Psychol.* **125**, 1072 (2023).
49. Aroyo, L. & Welty, C. Truth is a lie: crowd truth and the seven myths of human annotation. *AI Mag.* **36**, 15–24 (2015).
50. MacGeorge, E. L., Gillihan, S. J., Samter, W. & Clark, R. A. Skill deficit or differential motivation? Testing alternative explanations for gender differences in the provision of emotional support. *Commun. Res.* **30**, 272–303 (2003).
51. MacGeorge, E. L. et al. The influence of emotional support quality on advice evaluation and outcomes. *Commun. Quart.* **65**, 80–96 (2017).
52. Samter, W. & MacGeorge, E. L. in *Researching Interactive Communication Behavior: A Sourcebook of Methods and Measures* (eds VanLear C. A. & Canary, D. J.) 107–128 (Sage, 2017).
53. MacGeorge, E. L. & Wilkum, K. Predicting comforting quality in the context of miscarriage. *Commun. Rep.* **25**, 62–74 (2012).
54. Ziems, C. et al. Can large language models transform computational social science?. *Comput. Ling.* **50**, 237–291 (2024).
55. Movva, R., Koh, P. W. & Pierson, E. Annotation alignment: comparing LLM and human annotations of conversational safety. In *Proc. of the 2024 Conference on Empirical Methods in Natural Language Processing* (eds Al-Onaizan, Y. et al.) 9048–9062 (ACL, 2024).
56. Chiu, Y. Y., Sharma, A., Lin, I. W. & Althoff, T. A computational framework for behavioral assessment of LLM therapists. Preprint at https://doi.org/10.48550/arXiv.2401.00820 (2024).
57. Li, A. et al. Understanding the therapeutic relationship between counselors and clients in online text-based counseling using LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (eds Al-Onaizan, Y. et al.) 1280–1303 (ACL, 2024).
58. Shen, J., Mire, J., Park, H. W., Breazeal, C. & Sap, M. HEART-felt narratives: tracing empathy and narrative style in personal stories with LLMs. In *Proc. 2024 Conference on Empirical Methods in Natural Language Processing* (eds Al-Onaizan, Y. et al.) 1026–1046 (ACL, 2024).
59. Barrie, C., Palmer, A. & Spirling, A. Replication for language models problems, principles, and best practice for political science. Preprint at https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf (2024).
60. Gu, J. et al. A survey on LLM-as-a-judge. *The Innovation* https://doi.org/10.1016/j.xinn.2025.101253 (2024).
61. Zhou, L. et al. Larger and more instructable language models become less reliable. *Nature* **634**, 61–68 (2024).
62. Steyvers, M. et al. What large language models know and what people think they know. *Nat. Mach. Intell.* **7**, 221–231 (2025).
63. Ye, J. et al. Justice or prejudice? Quantifying biases in LLM-as-a-judge. In *NeurIPS Safe Generative AI Workshop 2024* https://openreview.net/forum?id=wtscPS2zJH (OpenReview, 2024).
64. Balog, K., Metzler, D., Qin, Z. Rankers, judges, and assistants: towards understanding the interplay of LLMs in information retrieval evaluation. In *Proc. 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (eds Ferro, N. et al.) 3865–3875 (ACM, 2025).
65. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977).
66. Schober, P., Boer, C. & Schwarte, L. A. Correlation coefficients: appropriate use and interpretation. *Anesth. Analg.* **126**, 1763–1768 (2018).
67. Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C. & Althoff, T. Towards facilitating empathic conversations in online mental health support: a reinforcement learning approach. In *Proc. Web Conference 2021* (eds Leskovec, J. et al.) 194–205 (ACM, 2021).
68. Lee, Y.-J., Lim, C.-G. & Choi, H.-J. Does GPT-3 generate empathetic dialogues? A novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proc. 29th International Conference on Computational Linguistics* (eds Calzolari, N. et al.) 669–683 (ACL, 2022).
69. Yang, D. et al. Social skill training with large language models. Preprint at https://doi.org/10.48550/arXiv.2404.04204 (2024).
70. Zhang, X. et al. Wider and deeper LLM networks are fairer LLM evaluators. Preprint at https://doi.org/10.48550/arXiv.2308.01862 (2023).
71. Groh, M., Harris, C., Daneshjou, R., Badri, O. & Koochek, A. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. *Proc. ACM on Hum.–Comput. Interact.* **6**, 521 (2022).
72. Huffaker, J. S., Kummerfeld, J. K., Lasecki, W. S. & Ackerman, M. S. Crowdsourced detection of emotionally manipulative language. In *Proc. 2020 CHI Conference on Human Factors in Computing Systems* https://doi.org/10.1145/3313831.337637 (ACM, 2020).
73. Boyatzis, R. E. *Transforming Qualitative Information: Thematic Analysis and Code Development* (Sage, 1998).
74. Roller, S. et al. Recipes for building an open-domain chatbot. In *Proc. 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (eds Merlo, P. et al.) 300–325 (ACL, 2021).
75. Hinz, A., Michalski, D., Schwarz, R. & Herzberg, P. Y. The acquiescence effect in responding to a questionnaire. *GMS Psycho-Soc. Med.* **4**, 07 (2007).
76. Yannakakis, G. N., Cowie, R. & Busso, C. The ordinal nature of emotions. In *Proc. 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)* 248–255 (IEEE, 2017).
77. Kirk, H. R., Gabriel, I., Summerfield, C., Vidgen, B. & Hale, S. A. Why human–AI relationships need socioaffective alignment. *Humanit. Soc. Sci. Commun.* **12**, 728 (2025).
78. Skjott Linneberg, M. & Korsgaard, S. Coding qualitative data: a synthesis guiding the novice. *Qual. Res. J.* **19**, 259–270 (2019).
79. Miller, A. H. et al. ParlAI: a dialog research software platform. *In Proc. 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (eds Specia, L. et al.) 79–84 (ACL, 2017).
80. Singh, N., Cherep, M. & Maes, P. Discovering interpretable concepts in large generative music models. In *NeurIPS AI for Music Workshop* https://openreview.net/forum?id=jVSlJk5qNA (OpenReview, 2025).
81. Peng, K., Movva, R., Kleinberg, J., Pierson, E. & Garg, N. Use sparse autoencoders to discover unknown concepts, not to act on known concepts. Preprint at https://doi.org/10.48550/arXiv.2506.23845 (2025).
82. Liang, P. et al. Holistic evaluation of language models. *Trans. Mach. Learn. Res.* https://openreview.net/forum?id=iO4LZibEqW (2023).
83. Warrens, M. J. Five ways to look at Cohen's kappa. *J. Psychol. Psychother.* **5**, 197 (2015).
84. Weng, L., Goel, V. & Vallone, A. *Using GPT-4 for Content Moderation* (OpenAI, 2023).
85. De Freitas, J. & Cohen, I. G. Unregulated emotional risks of AI wellness apps. *Nat. Mach. Intell.* **7**, 813–815 (2025).
86. De Freitas, J., Oğuz-Uğuralp, Z. & Kaan-Uğuralp, A. Emotional manipulation by AI companions. Preprint at https://doi.org/10.48550/arXiv.2508.19258 (2025).
87. Kumar, A. et al. Replication data and code: when LLMs provide reliable empathic communication. *Zenodo* https://doi.org/10.5281/zenodo.17230987 (2025).
88. Regier, D. A. et al. DSM-5 field trials in the United States and Canada, Part II: test–retest reliability of selected categorical diagnoses. *Am. J. Psychiatry* **170**, 59–70 (2013).

Acknowledgements
We thank W. Thompson for performing a replication check of our code and J. Chiminski, the Ryan Institute on Complexity, the Cosmos Institute and the Kellogg School of Management for funding.

Author contributions
A.K. and M.G conceived the investigation, A.K. curated the data, N.P., E.F. and B.L.L. annotated the data, A.K., N.P. and M.G. analysed the results, A.K. and M.G. wrote the initial manuscript and A.K., N.P., D.Y., E.F., B.L.L. and M.G. reviewed and edited the manuscript.

Competing interests
The authors declare no competing interests.

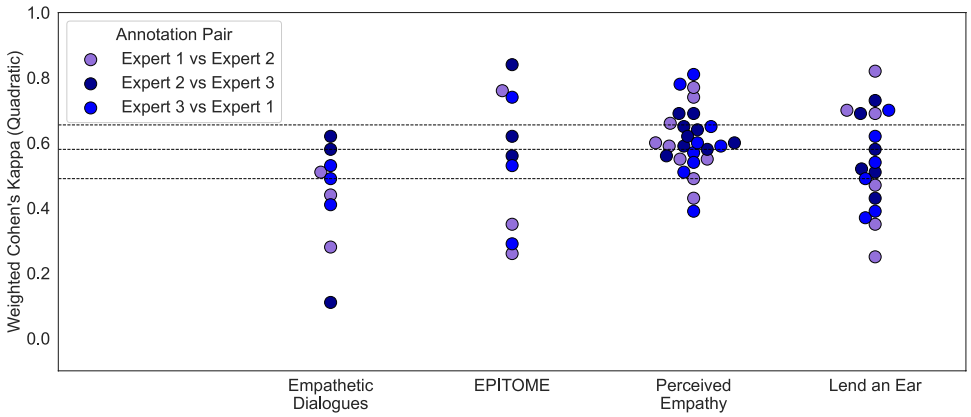
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s42256-025-01169-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-025-01169-6.

Correspondence and requests for materials should be addressed to Aakriti Kumar or Matthew Groh.

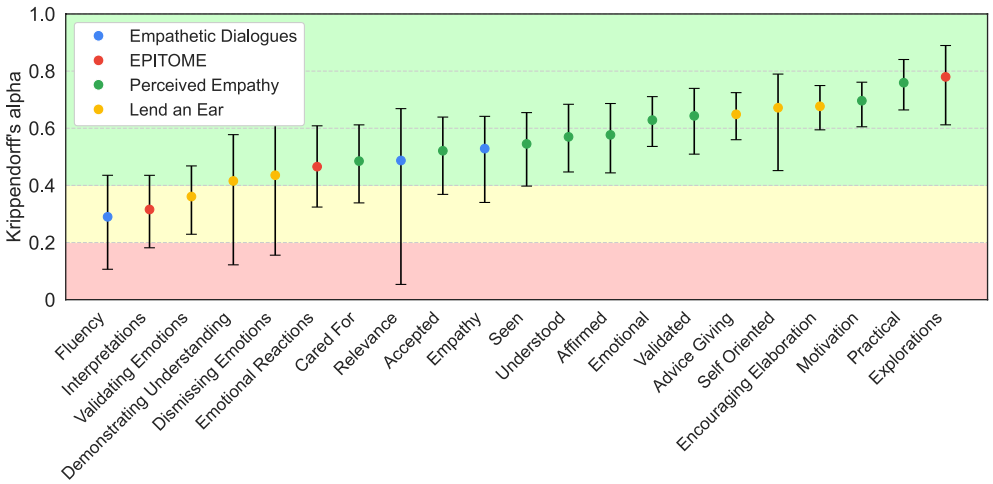
Peer review information *Nature Machine Intelligence* thanks Saadia Gabriel and Desmond Ong for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s42256-025-01169-6



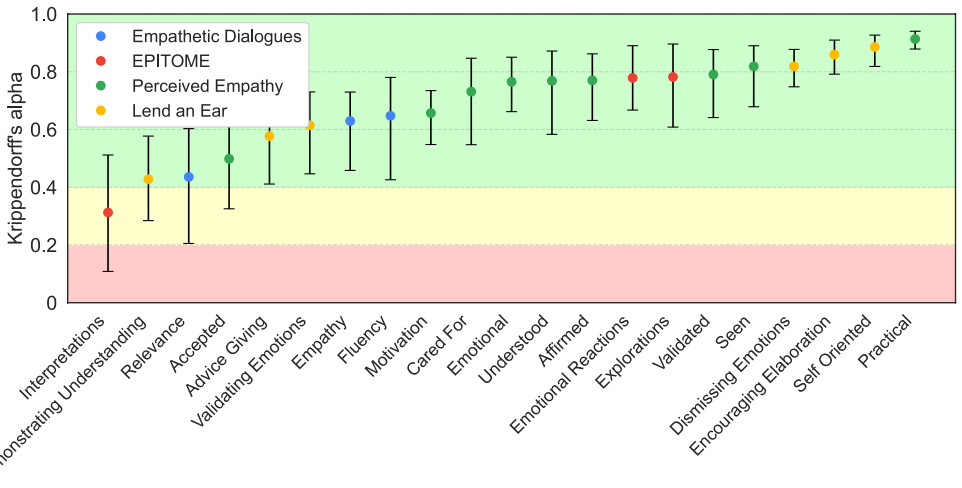
Extended Data Fig. 1 | Expert inter-rater reliability across frameworks. Pairwise expert reliability across four empathic communication frameworks.

https://doi.org/10.1038/s42256-025-01169-6



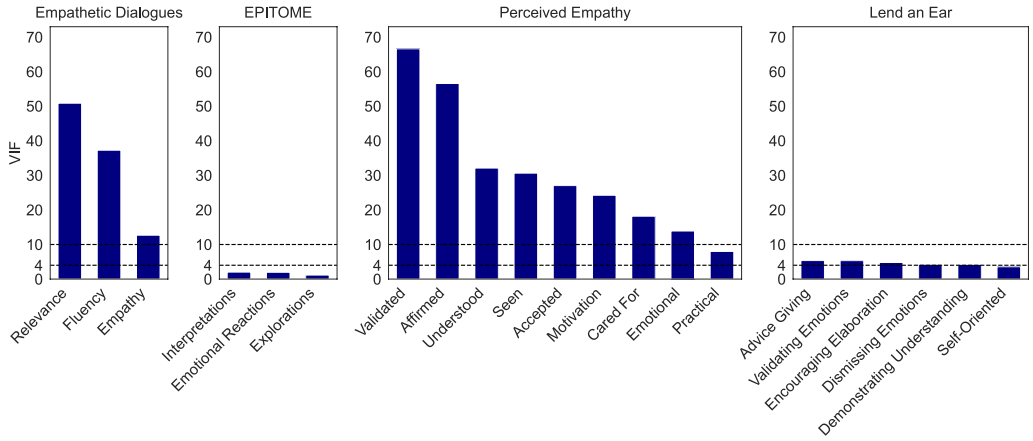
Extended Data Fig. 2 | Expert inter-rater reliability across sub-components. Inter-rater agreement as Krippendorff's alpha among the three expert annotators across frameworks' sub-components. The circle represents the mean Krippendorff's alpha and error bars are 95% confidence intervals obtained by bootstrapping conversations (n = 1000), and the background color indicates the degree of agreement with red for poor (α = 0.0 to 0.2), yellow for fair (α = 0.2 to 0.4), and green for moderate to high reliability (α = 0.4 to 1.0)[88].

https://doi.org/10.1038/s42256-025-01169-6



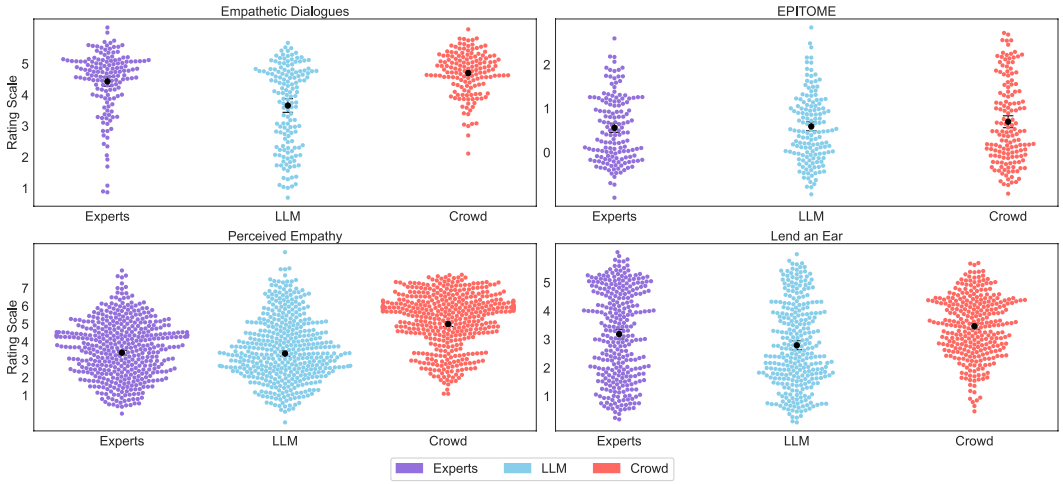
Extended Data Fig. 3 | LLM inter-rater reliability across sub-components. Inter-rater agreement as Krippendorff's alpha among the three LLM annotators (gemini-2.5-pro-preview-03-25, GPT-4o, Claude3.7-sonnet) across frameworks' sub-components. The circle represents the mean Krippendorff's alpha and error bars are 95% confidence intervals obtained by bootstrapping conversations (n = 1000), and the background color indicates the degree of agreement with red for poor (α = 0.0 to 0.2), yellow for fair (α = 0.2 to 0.4), and green for moderate to high reliability (α = 0.4 to 1.0).

https://doi.org/10.1038/s42256-025-01169-6



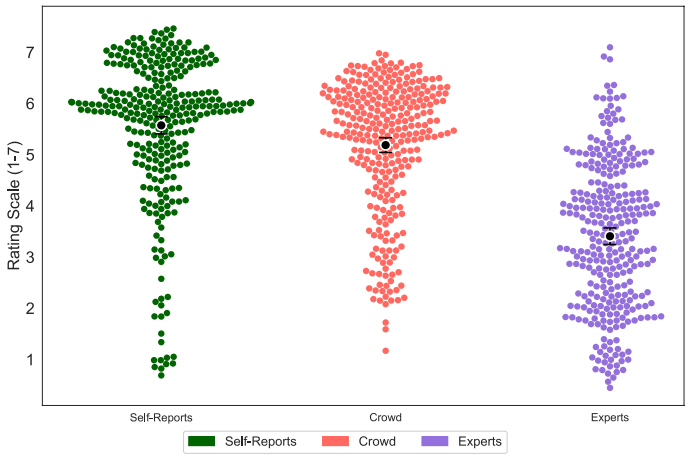
Extended Data Fig. 4 | VIF for each sub-component within the four frameworks. VIF quantifies the degree of multicollinearity among sub-components, with higher values indicating greater redundancy. Dashed lines indicate conventional thresholds of VIF ≤ 4 indicating low multicollinearity, 4 ≤ VIF ≤ 10 indicating moderate multicollinearity, and VIF ≥ 10 indicating high multicollinearity[88]. Sub-components with high VIF may indicate overlapping constructs or redundant measures within the framework.

https://doi.org/10.1038/s42256-025-01169-6



Extended Data Fig. 5 | Empathy inflation by crowds. We present the distribution of annotations by experts, LLMs, and crowds across four evaluation frameworks for 200 conversations on different scales. Empathetic Dialogues (1-5), EPITOME (0-2), Perceived Empathy (1-7), and Lend an Ear (1-5). Crowd workers tend to assign higher ratings than experts in all frameworks, with LLM ratings typically falling between the two. The circles indicate mean ratings and error bars indicate the 95% confidence interval for the mean.

https://doi.org/10.1038/s42256-025-01169-6



Extended Data Fig. 6 | Comparing self-reported perceived empathy to crowd and expert evaluations. We present the distribution of self-reported perceived empathy and annotations from crowd and experts (n = 300 annotations each) across six sub-components of the Perceived Empathy dataset. The black circles indicate mean ratings and error bars indicate the 95% confidence interval for the mean.

https://doi.org/10.1038/s42256-025-01169-6

Extended Data Table 1 | Inter-Rater Reliability (Weighted Cohen's Kappa) of LLM Variants vs Expert Median across datasets and sub-components

Sub-component	Zero-shot	Few-shot	Framework	Framework + Few-shot
Empathetic Dialogues				
Empathy	0.44**	0.36**	0.46**	0.35***
Fluency	-0.05	0.08	0.23	0.27
Relevance	**0.60****	0.27	**0.61****	0.23
EPITOME				
Emotional Reactions	0.25***	0.53***	0.22***	**0.64*****
Explorations	0.32*	0.54***	0.30*	**0.76*****
Interpretations	0.53***	0.56***	0.51***	**0.59*****
Perceived Empathy				
Understood	0.55***	**0.70*****	0.41***	0.53***
Validated	0.50***	**0.74*****	0.47***	**0.60*****
Affirmed	0.36**	**0.65*****	0.45***	**0.63*****
Accepted	0.29*	0.34*	0.34*	0.45***
Cared For	0.34**	0.55***	0.39**	**0.60*****
Seen	0.50***	**0.63*****	0.46***	**0.60*****
Emotional	0.46***	**0.69*****	0.49***	**0.75*****
Practical	**0.64*****	**0.71*****	**0.63*****	**0.74*****
Motivation	0.22**	**0.73*****	0.13	**0.70*****
Lend an Ear				
Validating Emotions	0.56***	**0.63*****	**0.59*****	**0.63*****
Demonstrating Understanding	**0.65*****	0.56***	0.46***	**0.59*****
Encouraging Elaboration	**0.80*****	**0.75*****	**0.80*****	**0.86*****
Advice Giving	0.51***	0.54***	**0.60*****	**0.66*****
Self-Oriented	**0.64*****	**0.64*****	**0.79*****	**0.69*****
Dismissing Emotions	0.31**	0.20**	0.35***	0.17**

The prompt variants include **Zero-Shot**, **Few-Shot**, **Framework**, and **Framework + Few-Shot**, ordered by increasing complexity of the prompt for the LLM annotator (gemini-2.5-pro-preview-03-25). Cells exceeding the high-agreement threshold ($\kappa_w \geq 0.58$) are in bold. Statistical significance of κ_w was assessed by testing the null hypothesis H_0: $\kappa_w = 0$ using a two-tailed Z-test. Significance markers reflect Benjamini-Hochberg FDR-adjusted p-values: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

https://doi.org/10.1038/s42256-025-01169-6

Extended Data Table 2 | Conversational contexts across datasets

SAE-Identified Topic	All	Empathetic Dialogues	EPITOME	Perceived Empathy	Lend an Ear
Mentions specific, non-serious incidents or events causing embarrassment or awkwardness in everyday situations	19.5%	78%	–	–	–
Expresses explicit thoughts, plans, or desires related to suicide or self-harm	18.0%	–	70%	2%	–
Mentions being passed over for a promotion	9.5%	–	–	–	38%
Mentions challenges in balancing work responsibilities following a promotion	9.5%	–	–	–	38%
Mentions challenges or decisions related to housing, home ownership, or home repairs	8.5%	–	–	34%	–
Mentions familial relationships or conflicts involving family members	7.5%	2%	–	28%	–
Mentions losing a job or being laid off	6.0%	–	–	–	24%
Mentions work-related conflicts, challenges, or ethical dilemmas	4.5%	2%	–	16%	–
Mentions financial challenges or disputes involving family members or spouses	4.5%	–	–	18%	–
Mentions experiences, symptoms, or struggles related to depression	3.5%	–	14%	–	–
Mentions specific incidents or events causing annoyance or frustration in a personal context	3.0%	12%	–	–	–
Mentions experiences, symptoms, or treatments related to mental health struggles or addiction	2.0%	–	6%	2%	–
Mentions jealousy or envy related to others' financial or material possessions	1.5%	4%	2%	–	–
Discusses personal experiences or questions related to antidepressant or antipsychotic medications and their effects	1.5%	–	6%	–	–
Mentions personal aspirations or challenges related to physical or professional capabilities	1.0%	2%	2%	–	–

Sixteen themes of personal disclosures identified by sparse autoencoders (SAEs) across 200 conversations. Values represent the percentage of conversations within each dataset (column) in which the theme appeared.

https://doi.org/10.1038/s42256-025-01169-6

Extended Data Table 3 | Expert-LLM Inter-rater Reliability across Models and Prompting Styles

		Experts	Gemini (2.5 pro)		OpenAI (GPT-4o)		Claude (3.7 Sonnet)	
Framework	**Sub-component**	**Median**	**Fewshot**	**CoT**	**Fewshot**	**CoT**	**Fewshot**	**CoT**
Empathetic Dialogues	Empathy	0.53	0.45	0.55	0.29	0.43	0.29	0.35
	Relevance	0.49	0.45	0.48	0.30	0.22	0.47	0.53
	Fluency	0.28	0.44	0.48	0.15	0.20	0.21	0.45
EPITOME	Emotional Reactions	0.53	0.46	0.47	0.67	0.50	0.71	0.58
	Explorations	0.76	0.80	0.89	0.88	0.64	0.77	0.73
	Interpretations	0.29	0.55	0.45	0.26	0.44	0.52	0.64
Perceived Empathy	Understood	0.60	0.53	0.65	0.40	0.51	0.36	0.47
	Validated	0.65	0.49	0.48	0.40	0.59	0.58	0.62
	Affirmed	0.60	0.55	0.38	0.37	0.47	0.52	0.64
	Seen	0.55	0.55	0.43	0.47	0.49	0.37	0.61
	Accepted	0.57	0.37	0.38	0.23	0.49	0.47	0.57
	Cared For	0.55	0.47	0.35	0.34	0.29	0.53	0.32
	Emotional	0.59	0.71	0.69	0.70	0.69	0.65	0.75
	Practical	0.77	0.69	0.69	0.62	0.67	0.62	0.73
	Motivation	0.66	0.68	0.77	0.47	0.67	0.73	0.80
Lend an Ear	Validate Emotions	0.39	0.55	0.44	0.52	0.31	0.68	0.52
	Encouraging Elaboration	0.69	0.80	0.64	0.70	0.49	0.79	0.67
	Demonstrating Understanding	0.47	0.57	0.61	0.39	0.13	0.43	0.42
	Advice Giving	0.69	0.47	0.48	0.51	0.27	0.29	0.23
	Self-Oriented	0.70	0.53	0.50	0.52	0.30	0.50	0.40
	Dismissing Emotions	0.49	0.12	0.11	0.11	0.05	0.14	0.06

Inter-rater reliability between different LLMs and experts (median) for few-shot and Chain-of-Thought (CoT) prompts.

nature portfolio

Corresponding author(s): Matthew Groh

Last updated by author(s): Jul 11, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We use conversational datasets across four evaluative frameworks: Empathetic Dialogues, EPITOME, Perceived Empathy, and the Lend an Ear pilot. For the Lend an Ear dataset, we recruited 150 Prolific participants as annotators to evaluate responses across multiple empathy sub-components. We collected expert annotations for the 200 conversations. Three of the authors of the paper served as experts.
Data analysis	The code to replicate all our analysis is available in our public GitHub repository: https://github.com/aakriti1kumar/replication-data-and-code-when-LLMs-reliable-empathic-communication

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data used during the current study are available in the data folder in our public GitHub repository: https://github.com/aakriti1kumar/replication-data-and-code-when-LLMs-reliable-empathic-communication

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We have collected self-reported data on gender for participants who annotated conversations in the Lend an Ear dataset but we have no data on gender for the other three datasets (Empathetic Dialogues, EPITOME, and Perceived Empathy). The participants who annotated the Lend an Ear dataset self report as 90 females, 59 males, and 1 did not specify. None of the conversations include self-reported data on the gender of the speaker. Gender was not considered in the study design, and we do not report an analysis of gender because our analysis of crowdworker annotations is focused on the crowd aggregated level not the individual level.
Reporting on race, ethnicity, or other socially relevant groupings	We do not report results on race, ethnicity, or other socially relevant groupings.
Population characteristics	The participants who annotated the conversations in the Lend an Ear dataset include the following self reported racial categories: 83 White, 39 Black, 11 Mixed, and 7 Asian. The mean age was 35.5 years and the range includes 19 to 72 years old. The participants who role played as supported in the conversations in the Lend an Ear dataset include the following self reported racial categories: 28 White, 12 Black, 6 Mixed, 3 Other, and 1 Asian. 26 self report as Female, 23 self report as Male and 1 prefers not to say. The mean age was 35.1 years and the age range was 19 to 63 years old. The participants who annotated the Perceived Empathy and participated in the conversations have the following characteristics: Of the 1,449 participants, 694 were women, 725 men, 29 nonbinary gender, and 1 did not report gender; 148 were Black, 89 Asian, 1,031 White, 76 Hispanic, 6 Native American, 82 mixed race, and 17 self- identified as other categories. The average age was 41.46 y old, with a SD of 13.58. There is no data on the population characteristic of the EPITOME or Empathetic Dialogues datasets.
Recruitment	For the Lend an Ear dataset, we recruited 150 Prolific participants as annotators to evaluate responses across multiple empathy sub-components. For all datasets, we collected expert annotations for the 200 conversations. Three of the authors of the paper served as experts.
Ethics oversight	This research complied with all relevant ethical regulations, and we collected informed consent. The Northwestern University Institutional Review Board (IRB) determined that the research met the criteria for exemption from further review. The study's IRB identification number is STU00222032 and STU00223043.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative investigation that compares the reliability of annotations of empathic communication by experts, crowdworkers, and Large Language Models
Research sample	We use conversational datasets across four evaluative frameworks: Empathetic Dialogues, EPITOME, Perceived Empathy, and the Lend an Ear pilot.

Sampling strategy	We sampled a total of 200 conversations, with 50 from each of four datasets. For the three published datasets (Empathetic Dialogues, EPITOME, Perceived Empathy), conversations were stratified random sampled: 10 from the highest-scoring quartile, 10 from the lowest, and 30 from the middle range based on prior crowdsourced annotations. For the Lend an Ear dataset, 50 of the 150 conversations from the pilot were randomly sampled. We chose 50 conversations from 4 conversational datasets because this was the limit to how many conversations were reasonable to ask experts to annotate.
Data collection	We collected expert annotations for the 200 conversations via an online interface. Three of the authors of the paper served as expert annotators and annotated independently. For the Lend an Ear dataset, we recruited 150 Prolific participants as annotators to evaluate conversations via a web interface.
Timing	11/1/2024-4/30/2025
Data exclusions	Data were excluded from the full datasets to create the sample used in this study. The criteria for exclusion was based on the sampling strategy for this study. No other data was excluded from the analysis.
Non-participation	34 participants started the annotation experiment but did not finish.
Randomization	We randomly assigned participants to 5 conversations each to annotate.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*