Arab World English Journal
INTERNATIONAL PEER REVIEWED JOURNAL ISSN: 2229-9327

Arab World English Journal (AWEJ) Special Issue on CALL Vol. 17, No.12. July 2026 Pp.484-499
DOI: https://dx.doi.org/10.24093/awej/call12.28

Simulated Empathy and Human Response: A Comparative Analysis of AI and Human Emotional Interaction

Abdullah Al Fraidan
Department of English Language,
College of Arts, King Faisal University, Al-Ahsa, Saudi Arabia
Corresponding Author**:** afridan@kfu.edu.sa

Jumana Waleed Buhaimed
Department of English Language,
College of Arts, King Faisal University, Al-Ahsa, Saudi Arabia
Email: 225001123@student.kfu.edu.sa

Received: 05/13/2026 Accepted:07/17/2026 Published:07/30/2026

Abstract
This qualitative comparative study examines how ChatGPT-4o simulates empathy when responding to emotionally charged English in an EFL context. It aims to compare artificial intelligence and human responses in emotional recognition, pragmatic tone, empathetic support, and linguistic authenticity. The study is significant because EFL interaction requires learners to interpret affective cues while selecting socially and culturally appropriate language. Fifty prompts generated 50 AI responses and 1,000 human responses from 20 advanced-level EFL learners; the data were organized into 50 prompt-level comparison sets and analyzed through qualitative content analysis and comparative discourse analysis. Two trained coders applied a hybrid framework and achieved substantial agreement (Cohen's $\kappa = .86$). ChatGPT recognized the intended emotion in 88% of its responses, used an appropriate tone in 84%, and displayed empathetic and pragmatically relevant support in 90%. Performance weakened with implicit, mixed, and culturally nuanced cues, while supportive language was sometimes formulaic or overly therapeutic. Human responses were more varied, culturally situated, and pragmatically flexible. The study recommends using ChatGPT as a teacher-mediated supplementary resource for emotional vocabulary and pragmatic practice, with explicit attention to cultural context, recurrent response formulas, and the distinction between simulated and human empathy.
Keywords: AI in language education, ChatGPT, EFL learners, emotional intelligence, emotional vocabulary, pragmatic competence

Cite as: Al Fraidan, A. & Buhaimed, J. W .(2026). Simulated Empathy and Human Response: A Comparative Analysis of AI and Human Emotional Interaction. *Arab World English Journal (AWEJ) Special Issue on CALL, Vol. 17.* (12)484-499. https://dx.doi.org/10.24093/awej/call12.28

Introduction

Conversational artificial intelligence has moved rapidly from information retrieval toward socially responsive interaction. Systems such as ChatGPT can identify affective wording, mirror a user's tone, validate d

istress, and offer advice in language that resembles interpersonal support. These behaviors are produced through probabilistic language generation rather than felt emotion, yet users may still interpret them as empathy. The central problem is therefore not whether an AI system experiences emotion, but how reliably its language recognizes emotion and performs contextually appropriate empathetic acts.

This problem is especially relevant to English as a Foreign Language (EFL) contexts. Emotional meaning is conveyed not only by vocabulary but also by intensity, indirectness, humor, culturally familiar expressions, and the relationship between speakers. EFL learners must interpret these signals while choosing language that is both grammatically intelligible and pragmatically suitable. A comparison between AI and advanced EFL responses can consequently reveal where machine-generated support is useful for language practice and where cultural or interpersonal nuance remains difficult to reproduce.

Recent educational research reports that conversational AI can support language practice, confidence, and access to feedback (Mohamed & Alian, 2023; Shikun et al., 2024). Research on AI-induced emotion and engagement also shows that affective experience is becoming central to L2 education (Yang & Zhao, 2024; Zhou & Hou, 2024). Other work suggests that speech-enabled AI may reduce foreign language anxiety by providing a low-risk practice environment (Bao, 2019). These studies establish pedagogical value, but learner perceptions and general proficiency gains do not demonstrate that an AI response is emotionally accurate, pragmatically appropriate, or comparable with a human response to the same stimulus. Output-level evidence is therefore needed.

Within Saudi higher education, conversational AI enters an already hybrid digital environment rather than replacing an exclusively face-to-face one. Saudi EFL learners have used mobile and social-media platforms for examination preparation, while newer work documents AI-supported writing, adaptive assessment, online assessment practices, and stakeholder readiness for immersive vocabulary learning (Al Fraidan, 2025a, 2025c; Al Fraidan & Al-Harazi, 2023; Al Fraidan & Alkuwaity, 2026; Al Fraidan & AlSalman, 2026). This continuity matters because learners' judgments of a chatbot are shaped by earlier experiences of automated feedback, platform-based communication, and institutionally mediated technology. The rapid adoption of digital tools also increases the practical importance of determining what kind of interaction a fluent system actually provides. If supportive language is taken as proof of understanding, learners may overestimate the system's cultural awareness; if all simulated support is dismissed as artificial, educators may overlook a useful source of immediate language practice. A comparative analysis can replace both assumptions with evidence about observable response patterns.

The emotional dimension of this problem is not peripheral to language learning. Anxiety, motivation, confidence, and perceived interpersonal support influence whether learners participate, persist, and risk producing unfamiliar language. Research on Saudi EFL writing shows that AI assistance can interact with uncertain motivation to improve argumentative performance, while task format and perceptions of teacher strictness can alter test anxiety (Al Fraidan, 2025a, 2025e). In a different applied setting, sustained emotionally supportive AI interaction was associated with engagement and behavioral consistency (Al Fraidan, 2025b). These findings do

not establish that AI possesses empathy, but they demonstrate that supportive discourse may have consequences for attention, willingness to communicate, and self-regulation. Evaluating the quality and limits of that discourse is therefore an applied-linguistic question as well as a technological one.

The present study aims to compare ChatGPT-4o and advanced EFL learners as responders to the same set of emotionally charged prompts. Its significance lies in connecting emotional vocabulary, pragmatic competence, and AI-mediated interaction through matched response data rather than attitude surveys alone. The study has four objectives: to evaluate emotional recognition, tone, and empathetic relevance in AI responses; to compare AI and human linguistic authenticity and nuance; to identify recurrent strengths and limitations in simulated empathy; and to derive evidence-based implications for teacher-mediated EFL practice.

The investigation addresses the following research questions:

1. How accurately does ChatGPT-4o recognize emotional vocabulary and affective cues in EFL-oriented prompts?
2. How appropriate are its tone, empathetic support, and pragmatic relevance?
3. How do AI and human responses differ in linguistic authenticity, variability, and cultural or contextual nuance?
4. What do the findings imply for the responsible use of conversational AI in EFL teaching and learning?

Literature Review

Emotional Vocabulary and Pragmatic Competence in EFL

Emotional vocabulary enables speakers to differentiate and communicate affective states with precision. Greater emotional granularity supports clearer interpretation and more differentiated regulation, whereas broad labels may obscure mixed or low-intensity states (Barrett et al., 2001; Gross, 2015). For EFL learners, lexical knowledge is only one part of this competence. A response may identify sadness correctly yet remain pragmatically unsuccessful if its advice is too direct, its reassurance is culturally incongruent, or its tone does not fit the relationship. Thomas's (1983) account of pragmatic failure is therefore important: grammatically acceptable language can still misfire when intended force or sociocultural expectations are misunderstood.

Research on emotional language technologies has often concentrated on detection. Sentiment and affective-computing systems can classify polarity or emotion categories from text (Cambria, 2013), while educational emotion-recognition models can infer affect from facial expressions (Salloum et al., 2025). Detection accuracy, however, is not equivalent to interactional adequacy. Recognizing anxiety does not determine whether the next utterance should validate, advise, ask a question, or simply remain present. The current study therefore treats emotional recognition as one dimension of a broader pragmatic response.

Lexical resources are especially consequential for EFL respondents because emotional precision depends on both receptive recognition and productive choice. Saudi EFL learners draw on strategic repertoires in both test taking and vocabulary learning, and these strategies shape how they approach academic tasks and unfamiliar words (Al Fraidan & AlSalman, 2023; Al Fraidan & Fakhli, 2024). Hybrid vocabulary assessment research likewise shows that task format can privilege recognition over active retrieval and may therefore provide an incomplete picture of usable lexical competence (Al Fraidan, 2025c). Emotional interaction raises the demand further: a responder must distinguish disappointment from anger, apprehension from fear, or relief from

uncomplicated happiness, then embed that interpretation in an appropriate interpersonal move. A system that supplies a sophisticated emotion word may model useful vocabulary, but learners still need to evaluate whether the word matches the prompt's intensity and whether the accompanying advice is pragmatically warranted. This distinction links vocabulary development to discourse-level judgment rather than treating emotion words as isolated items.

Pragmatic empathy is also culturally situated. Speech acts such as apology, consolation, reassurance, and advice are evaluated through expectations about status, solidarity, directness, and the seriousness of the event. Evidence from apology strategies in Saudi Arabia indicates that language choice, cultural sensitivity, gendered norms, and perceived power can influence whether a remedial act is accepted (Al Fraidan & Alhulaybi, 2025). The relevance to AI empathy is direct: an utterance can be semantically supportive but interactionally unsuccessful if it overlooks what the relationship permits or what the local context makes salient. A culturally neutral response may be safer across unknown users, yet neutrality can erase the idioms, religious references, indirectness, or shared assumptions through which human speakers display affiliation. Accordingly, the present analysis does not equate grammatical accuracy with authenticity. It examines whether the response performs an appropriate act and how its degree of contextual embedding compares with that of human EFL respondents.

Conversational AI in Language and Supportive Interaction

EFL studies generally report favorable learner attitudes toward chatbots and gains in accessibility, participation, or language practice. Mohamed and Alian (2023) found positive attitudes toward chatbot use, and Shikun et al. (2024) described benefits for English proficiency. Bao (2019) proposed that private, speech-enabled practice might mitigate foreign language anxiety. Yang and Zhao (2024) and Zhou and Hou (2024) further documented AI-related emotion and engagement in L2 education. Taken together, this literature shows that learners can value conversational AI; however, most studies emphasize perceptions, performance outcomes, or broad affective variables rather than matched analyses of the language through which empathy is performed.

Evidence from adjacent domains also requires qualification. Al Fraidan's (2025b) healthcare case illustrates the potential for emotionally responsive AI to support a user over time, while Haque and Rubya's (2023) review identifies promise as well as substantial limitations in mental-health chatbots. Liu et al.'s (2024) systematic review concludes that emotional support alone may not be enough to support EFL learning and should be integrated with cognitive support. These studies support attention to affective interaction, yet they do not establish that standardized supportive language is interchangeable with human empathy or culturally responsive communication.

Research from Saudi digital-learning contexts further shows why affective claims require close analysis of the interaction itself. Social-media platforms can support examination preparation and perceived achievement, but their value depends on how learners use particular channels and activities (Al Fraidan & Al-Harazi, 2023). AI-assisted argumentative writing can improve the clarity and structure of EFL texts when feedback is combined with motivational support (Al Fraidan, 2025a). Online assessment research, however, continues to identify differences among learner, instructor, and administrator perspectives, reminding educators that a tool's technical availability does not guarantee equivalent experiences or interpretations (Al Fraidan & AlSalman, 2026). These studies collectively position AI empathy within a broader ecology of tasks, users,

and institutional expectations. The present study adds output-level evidence by asking what supportive language the model produces before inferring what learners might gain from it.

Trust is another condition of educational usefulness. Research on AI-mediated language assessment in Saudi universities shows that procedural transparency and awareness of legal rights are associated with lower perceptions of algorithmic bias (Al Fraidan, 2025d). Although empathetic chat differs from automated scoring, the underlying principle remains relevant: users should understand what the system is doing, what evidence underlies its outputs, and where human oversight remains necessary. A conversational response may feel personalized while being generated from statistical patterns, and culturally incomplete outputs can appear authoritative because they are fluent. Responsible classroom use should therefore combine emotional-pragmatic evaluation with disclosure of the system's limitations, opportunities to contest or rewrite its response, and protection of sensitive personal information.

Theoretical Framework

Four complementary perspectives guide the analysis. First, the computers-as-social-actors framework predicts that people apply interpersonal expectations to responsive technologies even when they know those technologies are computational (Nass et al., 1994). Al Lily et al. (2023) extend this concern by describing the increasingly human-like role of generative AI in social life. In the present study, these perspectives explain why lexical validation, acknowledgment, and follow-up questions can be received as empathy despite the absence of subjective feeling.

Second, speech act theory and cross-cultural pragmatics distinguish what an utterance says from what it does (Searle, 1969; Thomas, 1983). Empathetic responses can perform acknowledgment, reassurance, advice, apology support, or invitations to elaborate. Their success depends on timing, degree of directness, and sociocultural expectations. This perspective supports the coding of pragmatic relevance and clarifies why an emotionally accurate label may still produce an unsuitable response.

Third, computer-mediated communication and social-presence theory explain both the accessibility and the limitations of text-based support. Reduced physical cues can constrain contextual interpretation, while carefully composed language can intensify perceived responsiveness or intimacy (Short et al., 1976; Walther, 1996). AI-generated fluency may therefore heighten social presence but also encourage polished, recurrent formulas that differ from spontaneous peer interaction.

Fourth, emotion-regulation theory treats language as a resource for naming, reframing, and managing affect (Gross, 2015). Together, the four perspectives define simulated empathy as an observable linguistic performance: the system must recognize affective cues, select an appropriate speech act and tone, and sustain plausible social presence. The unresolved gap is empirical. Few EFL studies compare AI and human responses to the same prompts while attending to implicit emotion, cultural nuance, and the difference between linguistic support and human emotional experience. This study addresses that gap.

These perspectives are complementary rather than competing. The computers-as-social-actors framework explains why users may respond socially; social-presence theory describes how textual design can intensify that impression; speech act and pragmatic theories evaluate what the response accomplishes in context; and emotion-regulation theory clarifies the possible function of naming or reframing affect. Together they also prevent two analytical errors. The first is to infer genuine feeling from socially fluent language. The second is to dismiss every machine-generated

response because it lacks subjective experience. The analytic focus instead falls on the quality, recurrence, and contextual fit of observable linguistic moves. This integrated framework guided both the coding dimensions and the interpretation of differences between standardized AI output and situated human response.

Methodology

Research Design

The study used a qualitative comparative design combining directed qualitative content analysis with comparative discourse analysis. Content analysis was appropriate because the research questions required systematic coding of recurring emotional and pragmatic features across a bounded corpus (Mayring, 2021). Comparative discourse analysis was used to examine how AI and human responders differed in wording, directness, interpersonal stance, and culturally situated meaning. The combination provided both transparent category counts and contextual interpretation; neither a frequency-only design nor an unstructured reading would have answered all four questions.

Participants

The human sample consisted of 20 advanced-level EFL learners enrolled in a university English program at King Faisal University. Purposive sampling selected learners who were classified at the advanced program level, could produce written responses in English, and agreed to complete the emotional-response task. Participation followed institutional ethical approval and informed consent procedures. The retained analytic record does not document distributions for gender, age, or years of English study, and it does not identify a separate standardized proficiency test administered for this research. Accordingly, "advanced" refers to the university program classification rather than an independently verified study score. These omissions limit demographic characterization and are not reconstructed retrospectively.

Materials and Data Sources

The stimulus set comprised 50 short English prompts representing sadness, anger, anxiety, joy, and contextually ambiguous affect. The prompts included direct emotional labels, indirect cues, mixed signals, culturally familiar expressions, and situations in which intent had to be inferred. EFL educators and researchers reviewed the set for linguistic accessibility, range of affect, and relevance to everyday interaction.

The prompt set was designed to elicit response decisions rather than emotion labels alone. Direct items contained an identifiable affective word or a familiar physical sign, whereas indirect items required inference from an event, contradiction, understatement, or interpersonal expectation. Mixed and ambiguous items allowed more than one plausible reading so that the analysis could examine calibration rather than force a single decontextualized answer. The situations varied in whether a supportive responder might acknowledge, reassure, advise, apologize, praise, or invite elaboration. This range was important for EFL analysis because learners must move from recognizing vocabulary to selecting a speech act and degree of directness. The prompts were short enough to limit extraneous background, but this control also means that responders sometimes lacked relationship information that would ordinarily guide human interaction.

ChatGPT-4o was the AI system examined. Each prompt was submitted individually through the official OpenAI platform using a standardized instruction requesting a natural, empathetic English response. No follow-up prompting or response rewriting was applied before analysis. The AI corpus therefore contained 50 responses. Each of the 20 participants responded in writing to the same 50 prompts, producing 1,000 human responses. For comparison, the corpus was organized as 50 prompt-level sets, each containing one AI response and the 20 corresponding human responses. This organization corrects the distinction between individual responses and paired stimulus-level comparison units.

Procedure and Analysis

After collection, all responses were anonymized and arranged by prompt. A hybrid coding framework combined categories specified by the research questions with themes developed through repeated reading (Braun & Clarke, 2006). The directed categories were emotional recognition, tone appropriateness, empathetic and pragmatic relevance, and linguistic authenticity. Emergent codes captured indirectness, humor, cultural reference, formulaic validation, therapeutic framing, advice, and invitations to elaborate.

The operational definitions were applied at response level. Emotional recognition concerned whether the response named or clearly oriented to the affect licensed by the prompt. Tone appropriateness concerned the match among emotional intensity, register, and interpersonal stance. Empathetic-pragmatic relevance required an intelligible supportive act that addressed the situation rather than offering generic positivity. Linguistic authenticity was interpreted comparatively through variability, idiomaticity, culturally grounded reference, and adaptation to inferred relationships; it was not treated as a claim that one response was more sincere. Formulaic language was coded when substantially similar validation–reassurance–invitation sequences recurred across different prompts. Therapeutic framing referred to counseling-like metaphors, regulation instructions, or diagnostic-sounding interpretations that exceeded the apparent severity or informality of the stimulus. These definitions separated observable text from assumptions about the responder's internal state.

Two trained coders independently coded the full corpus after calibrating the framework on a subset. For the AI responses, a positive judgment was recorded when the output (a) identified or accurately inferred the intended emotion, (b) adopted a tone compatible with the stimulus, and (c) performed a contextually relevant supportive act. Counts were calculated over the 50 AI responses. Human responses were analyzed for the range and character of their discourse strategies rather than reduced to a single normative score. Disagreements were discussed after independent coding, and the finalized coding achieved Cohen's $\kappa = .86$.

Trustworthiness was strengthened through coder comparison, peer debriefing, an audit trail linking codes to prompt-level records, and checks of interpretations against the response sets (Lincoln & Guba, 1985). These procedures supported consistency while preserving qualitative differences that a purely automated sentiment score would have obscured.

Ethical Considerations

The project received approval from the relevant King Faisal University ethics committee (No. 2025-JanETHIC986). Participation was voluntary, responses were anonymized for analysis, and no personally identifying information is reported.

Results

Overall AI Performance

Across the 50 prompts, ChatGPT-4o accurately recognized the intended emotional vocabulary or affective cue in 44 responses (88%). It used a tone judged appropriate to the situation in 42 responses (84%) and provided empathetic, pragmatically relevant support in 45 responses (90%). These results indicate strong performance with explicit cues, alongside identifiable weaknesses when emotion was implicit, mixed, or culturally embedded.

Table 1. *ChatGPT-4o performance across coded dimensions*

Evaluation dimension	Accurate or appropriate	Not accurate or appropriate	Percentage
Emotional recognition	44	6	88%
Tone appropriateness	42	8	84%
Empathetic and pragmatic relevance	45	5	90%

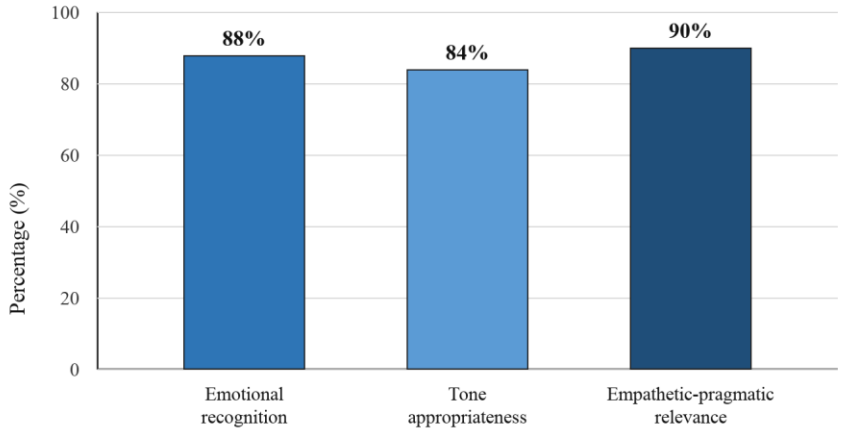


Figure 1. Distribution of ChatGPT-4o accuracy, tone appropriateness, and empathetic-pragmatic relevance

Emotional Recognition and Tone

Explicit affect was usually recognized without difficulty. In response to "I have a presentation tomorrow, and my hands won't stop shaking," the AI wrote, "It's completely normal to feel this way before a presentation—your body's reacting to stress. Right now, try deep breathing exercises or sip cold water." The response correctly connected a bodily cue with anxiety and combined normalization with a concrete coping suggestion. Human responses also recognized anxiety, but the coded sets contained a wider range of concise reassurance, direct advice, and peer-like encouragement.

The AI also recognized explicit sadness in "Today just feels empty, like nothing is going right," replying, "I'm really sorry you're feeling that way today. Sometimes it hits out of nowhere—that heavy, empty feeling where even small things feel off." The mirroring of "empty" and the invitation implied by its supportive stance produced an appropriate tone. Errors were concentrated in prompts with irony, culturally conventional understatement, or more than one plausible emotion; in those cases, the model sometimes selected the dominant lexical cue and missed the interpersonal implication.

Empathetic-Pragmatic Response Strategies

The model frequently used validation followed by an offer to continue. For "She promised she would be there, but she never showed up," it responded, "It hurts when someone breaks a promise, especially when you were counting on them. Want to talk about what happened?" This sequence acknowledged disappointment before requesting elaboration. Human response sets displayed more variation in speech acts, including sympathy, criticism of the absent person, practical advice, and culturally situated expressions of loyalty or disappointment. The AI's response was coherent and safe, but less variable.

Formulaic support was especially visible when the stimulus invited praise or self-evaluation. To "I've worked so hard, and no one even notices," the AI replied, "That sounds really heavy. You deserve recognition for your effort and dedication. Sometimes the world is quiet when it should be clapping loudest. Do you want to tell me more?" The response combined validation, figurative language, and a follow-up question. Across the corpus, however, similar sequences—acknowledgment, reassurance, advice, invitation—recurred often enough to make some outputs sound standardized.

A related pattern was therapeutic framing. In response to "I'm always trapped in the past or worried sick about what will happen," the AI wrote, "Being stuck in the past can feel like a chain, and worrying about the future can feel like drowning in what-ifs. Your power is in the present." The metaphor was emotionally coherent, but its counseling register exceeded the informality used in many human responses. For self-blame—"I shouldn't have said those things to my friend"—the AI offered normalization, apology advice, and help drafting a message. Human responses were generally less standardized and more dependent on perceived relationship, seriousness, and shared cultural expectations. Four representative prompt-level comparisons are provided in Appendix A (Table A1).

Human Variability and Linguistic Authenticity

Compared with the AI outputs, human responses varied more in length, directness, register, and emotional intensity. Participants could be terse, humorous, indirect, or explicitly supportive, and the same prompt elicited more than one plausible interpersonal stance. This variability occasionally produced less polished language, but it also conveyed relationship-specific and culturally grounded meaning. ChatGPT's responses were more consistently grammatical and supportive; human responses were less uniform and often more socially distinctive. Linguistic authenticity in this corpus therefore reflected contextual fit and variation rather than grammatical accuracy alone.

Variability was analytically important because an emotionally charged prompt rarely has only one socially acceptable reply. The human sets demonstrated that support may be conveyed through brevity, shared humor, mild criticism, religious or culturally familiar reassurance, practical intervention, or strategic silence implied by a short acknowledgment. The AI more often made its supportive intention explicit and completed a recognizable sequence. This clarity can benefit learners who need an accessible model, but it can also create the impression that longer and more elaborated responses are inherently more empathetic. The comparison instead suggests that adequacy depends on the relationship, the seriousness of the event, and the recipient's likely preference as much as on the number of supportive expressions.

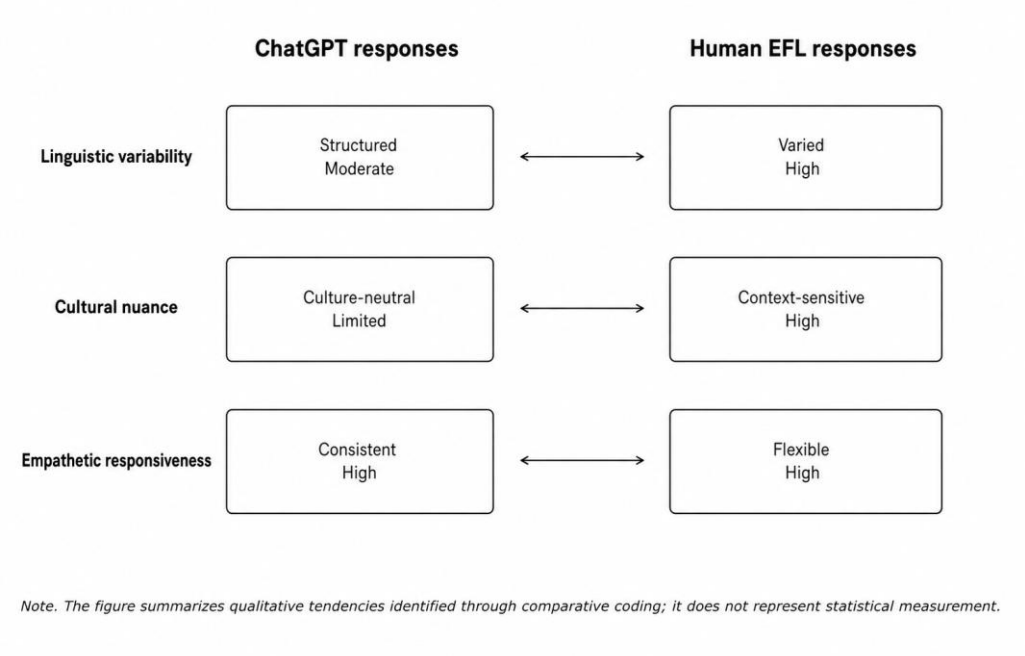


Figure 2. Conceptual comparison of standardized AI empathy and contextually variable human emotional response

Discussion

For Research Question 1, ChatGPT-4o recognized the intended affect in 44 of 50 prompts. Its strongest performance occurred when the prompt included explicit emotion words, recognizable bodily cues, or familiar situations. The six unsuccessful cases involved ambiguity, mixed affect, irony, or culturally embedded meaning. Emotional vocabulary detection was therefore strong but not context independent.

For Research Question 2, the AI used an appropriate tone in 42 responses and a relevant empathetic act in 45. Validation, normalization, advice, and invitations to elaborate were its principal strategies. These moves usually produced coherent support, but their repeated sequencing sometimes sounded formulaic, and some advice adopted a therapeutic register that was disproportionate to the prompt. A high empathy score thus represents observable linguistic performance, not evidence that the system feels concern.

For Research Question 3, human responses were more variable and context dependent. They displayed wider differences in directness, humor, brevity, interpersonal stance, and culturally situated interpretation. The AI was more uniform and consistently polished, but that consistency reduced the unpredictability and relationship-specific texture associated with human interaction. Human authenticity was not synonymous with superior grammar; it emerged from flexible adaptation to social context.

For Research Question 4, the findings support a supplementary, teacher-mediated role for conversational AI in EFL education. ChatGPT can provide immediate models for naming emotion, validating a speaker, and practicing pragmatic alternatives. Learners should nevertheless compare outputs, identify formulaic or culturally incongruent language, and rewrite responses for different relationships and levels of directness. Such activities make the limitations of simulated empathy part of the learning objective.

Relationship to Prior Research and Theory

The results extend favorable EFL findings without treating user acceptance as proof of emotional accuracy. Positive attitudes and opportunities for low-risk practice reported by Mohamed and Alian (2023), Bao (2019), and Shikun et al. (2024) are compatible with the model's high rates of recognition and supportive tone. At the same time, the present errors show why confidence, enjoyment, or fluency outcomes cannot substitute for analysis of pragmatic fit. The distinction also accords with Liu et al. (2024): emotional support alone may not be enough to support learning and should be integrated with cognitive support.

The model's strong explicit-cue performance is consistent with advances in affective computing and emotion-aware systems (Cambria, 2013; Salloum et al., 2025). Its difficulty with implicit and culturally nuanced prompts, however, confirms that classification and supportive generation remain different tasks. Al Fraidan's (2025b) account of emotionally responsive AI in healthcare illustrates the perceived value of sustained interaction, while the present matched comparisons show that supportive fluency can coexist with formulaic or overly therapeutic language.

The theoretical frameworks explain this duality. Under the computers-as-social-actors perspective, responsive language invites users to apply social expectations to the system (Al Lily et al., 2023; Nass et al., 1994). Speech act theory clarifies that the AI's success often depended on selecting a recognizable act—validation, advice, reassurance, or invitation—rather than merely naming an emotion (Searle, 1969). Pragmatic-failure theory explains the remaining gap: a formally appropriate act can still be miscalibrated in directness, register, or cultural meaning (Thomas, 1983).

Computer-mediated communication further explains why polished text may feel socially present while remaining standardized (Short et al., 1976; Walther, 1996). The study's theoretical contribution is therefore a distinction between standardized empathy and situated empathy. Standardized empathy is consistent, linguistically fluent, and organized around recurrent supportive sequences. Situated empathy varies with relationship, culture, uncertainty, and lived experience. The two can overlap at the level of language, but they are not equivalent.

Pedagogical and Institutional Implications

A first pedagogical use is contrastive response analysis. Teachers can present one prompt with an AI response and several anonymized human alternatives, then ask learners to identify the emotion, speech act, level of directness, evidence of cultural positioning, and possible effect on the recipient. Learners can rank responses for different relationships rather than search for one universally correct answer. This task connects emotional vocabulary with pragmatic reasoning and makes visible the difference between grammatical fluency and contextual fit. It also complements vocabulary-strategy instruction by requiring active retrieval, paraphrase, and fine-grained comparison instead of simple recognition (Al Fraidan, 2025c; Al Fraidan & Fakhli, 2024).

A second use is guided production. After identifying recurrent AI formulas, learners can rewrite the response for a close friend, classmate, teacher, stranger, or culturally specific situation. They can reduce an over-therapeutic register, add an appropriate indirect expression, replace generic praise with evidence from the prompt, or produce a brief version that still acknowledges the speaker's emotion. The model can then generate alternatives that learners critique rather than accept. Such sequencing preserves the immediate feedback and motivational possibilities associated with AI-supported writing while keeping judgment with the learner and teacher (Al

Fraidan, 2025a). It also offers a low-risk route to practicing difficult speech acts such as apology and consolation, whose success depends on cultural and interpersonal variables (Al Fraidan & Alhulaybi, 2025).

A third use is critical AI literacy. Learners should be told that a fluent validating response is generated, that different prompts or model versions may yield different outputs, and that highly personal disclosures should not be treated as ordinary classroom data. Reflection questions can ask what information the model lacked, whether its advice assumed a particular culture, and when a human teacher, counselor, family member, or professional should be consulted. This approach aligns emotional-pragmatic instruction with the transparency and human-oversight principles emphasized in AI-mediated language assessment (Al Fraidan, 2025d). It also prevents the pedagogical value of simulated empathy from being confused with authority, confidentiality, or professional care.

At the institutional level, responsible use requires clear boundaries. AI may serve as a supplementary language-practice tool, but it should not independently evaluate a learner's emotional condition or replace human support in high-stakes situations. Programs should identify approved uses, protect prompt data, provide culturally relevant examples, and train instructors to recognize formulaic, biased, or overconfident output. Evaluation should include learners with different proficiency levels and should distinguish satisfaction from demonstrated pragmatic learning. The stakeholder differences documented in online EFL assessment research indicate that implementation guidance must address learners, teachers, and administrators rather than assume a uniform perspective (Al Fraidan & AlSalman, 2026).

Limitations

The study is limited by its small, single-institution sample and one model version. The 50 prompts cannot represent every emotion, dialect, relationship, or cultural context, and qualitative judgments remain partly interpretive despite strong coder agreement. Demographic variables and an independently administered proficiency score were not documented in the retained analytic record. Item-level human wording was also unavailable for quotation during revision, limiting the illustration of individual voices. Finally, the design evaluates language outputs rather than changes in learner behavior, well-being, or long-term competence. Claims are accordingly restricted to the analyzed corpus.

Conclusion

This study compared ChatGPT-4o with advanced EFL learners across 50 emotionally charged prompts to assess emotional recognition, tone, empathetic-pragmatic relevance, and linguistic authenticity. ChatGPT recognized the intended emotion in 88% of responses, used an appropriate tone in 84%, and produced relevant empathetic support in 90%. It was strongest with explicit cues and consistent validation but weaker with implicit or culturally situated meaning, sometimes using formulaic or over-therapeutic language. Human responses were less uniform yet more varied in register, interpersonal stance, and cultural expression. These findings support ChatGPT as a teacher-mediated supplementary resource for emotional vocabulary and pragmatic practice, not a substitute for human judgment. Classroom use should foreground response comparison, cultural appropriateness, and critical AI literacy; future research should use larger, more diverse samples and preserve item-level human data.

Declarations

Funding: This work was funded by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia (Grant No. KFU264186).

Acknowledgments: The authors thank the study participants for their valuable contributions and the editor and anonymous reviewers for their constructive comments. They are also grateful to their families for their support. Abdullah Al Fraidan offers special thanks to his wife, Alanoud Alwasmi, and to his children, especially his youngest, Hiam, for their continued encouragement.

Conflicts of Interest: The authors declare no conflict of interest.

Authenticity: This manuscript is an original work that has not been published previously and is not under consideration elsewhere. The study materials and de-identified data are available from the corresponding author on reasonable request, subject to ethical and institutional requirements.

Artificial Intelligence Statement: ChatGPT-4o (OpenAI) was the object of analysis and generated the 50 AI responses examined under standardized prompts; accordingly, the manuscript contains brief, clearly identified excerpts of AI-generated study material. During revision, ChatGPT was also used to assist with language editing, organizational clarity, and reference formatting under author supervision. The authors reviewed and edited all AI-assisted text, independently verified the source data, analyses, claims, and citations, and take full responsibility for the submitted content. AI did not recruit participants, code the human-response data, determine the findings, or replace the authors' scholarly judgment.

About the Authors:

Abdullah Al Fraidan is Professor of English Linguistics and CEO of the Innovation District at King Faisal University, Saudi Arabia. He formerly served as Dean of E-Learning and Distance Education. His research covers language assessment, applied linguistics, AI-mediated learning, and computer-assisted language learning. He has published in leading journals, including Computer Assisted Language Learning, Acta Psychologica, Humanities and Social Sciences Communications, Heliyon, and PLOS ONE. ORCID ID: https://orcid.org/0000-0002-5152-9794

Jumana Waleed Buhaimed is an MA student in the Department of English Language, College of Arts, King Faisal University, Saudi Arabia, and an English-language teacher. Her interests include applied linguistics, EFL pedagogy, emotional vocabulary, pragmatic competence, and AI-mediated interaction. Her research examines simulated empathy and linguistic, emotional, and pragmatic differences between AI-generated and human responses.

References

Al Fraidan, A. (2025a). AI and uncertain motivation: Hidden allies that impact EFL argumentative essays using the Toulmin model. *Acta Psychologica, 252*, 104684. https://doi.org/10.1016/j.actpsy.2024.104684

Al Fraidan, A. (2025b). AI language and emotional support as a physician assistant in hypertension management: An N-of-1 case study on virtual encouragement and blood pressure control. *Humanities and Social Sciences Communications, 12*, 1229. https://doi.org/10.1057/s41599-025-05635-9

Al Fraidan, A. (2025c). Evaluating lexical competency in Saudi Arabia's hybridized EFL ecosystem: A taxonomic exploration of vocabulary assessment modalities. *International Journal of Distance Education Technologies, 23*(1), 1–36. https://doi.org/10.4018/IJDET.368224

Al Fraidan, A. (2025d). Procedural transparency and legal accountability to sustain AI-mediated language assessment in Saudi Arabia. *SAGE Open, 15*(4), 1–13. https://doi.org/10.1177/21582440251396113

Al Fraidan, A. (2025e). Test anxiety across writing formats: The mediating role of perceived teacher strictness. *Acta Psychologica, 256*, 104942. https://doi.org/10.1016/j.actpsy.2025.104942

Al Fraidan, A., & Al-Harazi, E. M. (2023). Using social media platforms to prepare for examinations post COVID-19: The case of Saudi university EFL learners. *Heliyon, 9*(11), e21320. https://doi.org/10.1016/j.heliyon.2023.e21320

Al Fraidan, A., & Alhulaybi, A. A. (2025). Exploring language use in apology strategies of tourists visiting Saudi Arabia. *PLOS ONE, 20*(2), e0319393. https://doi.org/10.1371/journal.pone.0319393

Al Fraidan, A., & Alkuwaity, R. (2026). Immersive educational innovation: Stakeholder perceptions on metaverse-based vocabulary learning in Saudi Arabia. *Computers in Human Behavior Reports, 22*, 101199. https://doi.org/10.1016/j.chbr.2026.101199

Al Fraidan, A. A., & AlSalman, A. (2023). The utilisation of test-taking strategies by Saudi EFL high schoolers in Al-Ahsa region. *World Journal of English Language, 13*(2), 414–423. https://doi.org/10.5430/wjel.v13n2p414

Al Fraidan, A., & AlSalman, A. (2026). Learner, instructor, and administrator perspectives on online EFL assessment practices and challenges in Saudi Arabia. *Computer Assisted Language Learning*, 1–25. https://doi.org/10.1080/09588221.2026.2640470

Al Fraidan, A., & Fakhli, I. (2024). Vocabulary learning strategies of university students: The case of preparatory year students and English major students. *Porta Linguarum, 41*, 45–63. https://doi.org/10.30827/portalin.vi41.24321

Al Lily, A. E., Ismail, A. F., Abunaser, F. M., Al-Lami, F., & Abdullatif, A. K. A. (2023). ChatGPT and the rise of semi-humans. *Humanities and Social Sciences Communications, 10*, 626. https://doi.org/10.1057/s41599-023-02154-3

Bao, M. (2019). Can home use of speech-enabled artificial intelligence mitigate foreign language anxiety—Investigation of a concept. *Arab World English Journal, Special Issue on CALL* (5), 28–40. https://doi.org/10.24093/awej/call5.3

Barrett, L. F., Gross, J., Christensen, T. C., & Benvenuto, M. (2001). Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. *Cognition and Emotion, 15*(6), 713–724. https://doi.org/10.1080/02699930143000239

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cambria, E. (2013). An introduction to concept-level sentiment analysis. In F. Castro, A. Gelbukh, & M. González (Eds.), *Advances in soft computing and its applications* (pp. 478–483). Springer. https://doi.org/10.1007/978-3-642-45111-9_41

Gross, J. J. (2015). Emotion regulation: Current status and future prospects. *Psychological Inquiry, 26*(1), 1–26. https://doi.org/10.1080/1047840X.2014.940781

Haque, M. D. R., & Rubya, S. (2023). An overview of chatbot-based mobile mental health apps: Insights from app description and user reviews. *JMIR mHealth and uHealth, 11*, e44838. https://doi.org/10.2196/44838

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage.

Liu, Y., Zhang, H., Jiang, M., Chen, J., & Wang, M. (2024). A systematic review of research on emotional artificial intelligence in English language education. *System, 126*, 103478. https://doi.org/10.1016/j.system.2024.103478

Mayring, P. (2021). *Qualitative content analysis: A step-by-step guide*. Sage.

Mohamed, S. S. A., & Alian, E. M. I. (2023). Students' attitudes toward using chatbot in EFL learning. *Arab World English Journal, 14*(3), 15–27. https://doi.org/10.24093/awej/vol14no3.2

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 72–78). Association for Computing Machinery. https://doi.org/10.1145/191666.191703

Salloum, S. A., Alomari, K. M., Alfaisal, A. M., Aljanada, R. A., & Basiouni, A. (2025). Emotion recognition for enhanced learning: Using AI to detect students' emotions and adjust teaching methods. *Smart Learning Environments, 12*, 21. https://doi.org/10.1186/s40561-025-00374-5

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139173438

Shikun, S., Grigoryan, G., Huichun, N., & Harutyunyan, H. (2024). AI chatbots: Developing English language proficiency in EFL classroom. *Arab World English Journal, Special Issue on ChatGPT*, 292–305. https://doi.org/10.24093/awej/ChatGPT.20

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. Wiley.

Thomas, J. (1983). Cross-cultural pragmatic failure. *Applied Linguistics, 4*(2), 91–112. https://doi.org/10.1093/applin/4.2.91

Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research, 23*(1), 3–43. https://doi.org/10.1177/009365096023001001

Yang, L., & Zhao, S. (2024). AI-induced emotions in L2 education: Exploring EFL students' perceived emotions and regulation strategies. *Computers in Human Behavior, 159*, 108337. https://doi.org/10.1016/j.chb.2024.108337

Zhou, C., & Hou, F. (2024). Can AI empower L2 education? Exploring its influence on the behavioural, cognitive and emotional engagement of EFL teachers and language learners. *European Journal of Education, 59*(4), e12750. https://doi.org/10.1111/ejed.12750

Appendix A

Table A1. Illustrative Prompt-Level Comparisons

Stimulus	Verified AI response excerpt	Coded human-response pattern	Interpretation
Presentation anxiety	"It's completely normal … your body's reacting to stress."	Recognition of anxiety with varied brief reassurance, advice, and peer encouragement.	Both groups recognized the cue; AI used a polished validation-plus-coping formula.
Broken promise	"It hurts when someone breaks a promise … Want to talk about what happened?"	Sympathy, criticism, advice, and culturally situated expressions of loyalty or disappointment.	AI tone was appropriate; human strategies were more socially varied.
Unrecognized effort	"You deserve recognition … Do you want to tell me more?"	Validation ranged from direct encouragement to shorter, more indirect support.	AI support was fluent but repeated a common validation-and-invitation sequence.
Regretted speech	"A sincere apology can go a long way. Want help figuring out how to express it?"	Responses varied with inferred relationship, blame, and seriousness.	AI supplied an actionable repair strategy; human interpretation was more context-dependent.

Note 1. AI excerpts are verbatim from the retained corpus. The human column reports coded response patterns because item-level participant wording was not documented in the retained analytic record; quotation marks are therefore intentionally not used.