Interaction Experience and Perceived Emotional Understanding of Generative AI: An Exploratory Study Using Explainable Machine Learning

Yiming Huang

University of the West of England, Bristol, BS16 1QY, UK

Keywords:

Generative Artificial Intelligence;
Perceived Emotional Understanding;
Interaction Experience;
Explainable Machine Learning;
Feature Ablation;
SHAP

ABSTRACT

As generative artificial intelligence (GenAI) is increasingly used for emotional expression and support, the relationship between interaction experience and users' evaluations of AI's emotional understanding remains to be examined. Based on questionnaire responses from 195 participants with prior experience of emotional disclosure to AI, this study applies logistic regression, random forests, and repeated nested cross-validation, together with feature-group ablation, permutation importance, and SHAP, to examine the contributions of different questionnaire features and the effects of semantically similar items on model results.The results show that the random forest model using interaction experience features alone achieved a ROC-AUC of 0.746, compared with 0.478 for background and usage behaviour features, indicating that interaction experience was the more informative source of classification information in the current sample. Perceived mechanical responses (Q18) and response immediacy (Q15) emerged as prominent interaction experience features. Removing Q18 reduced the random forest AUC to 0.672. Further removal of Q15 from the conservative model excluding Q18 and Q20 reduced the AUC from 0.666 to 0.590, indicating that response immediacy retained a notable classification contribution under the conservative feature setting.The findings indicate that, compared with user background and usage behaviour, interaction experience items contain more information relevant to classifying perceived AI emotional understanding in the current sample. Interpreting key features also requires consideration of overlapping questionnaire content and input feature combinations. Through explainable machine learning, this study identifies the main interaction experience features associated with users' subjective evaluations and provides empirical evidence for identifying important experience features in emotional interactions with GenAI.

CONTACT:

[1]Author: yiming2.huang@live.uwe.ac.uk

*Corresponding Author: yiming2.huang@live.uwe.ac.uk

Note: Authors designated with "†" are recognised as co-first authors.

DOI.10.64549/jaai-ii.v1i2.123

This work is licensed under the CC BY 4.0 HTTPS://CREATIVECOMMONS.ORG/LICENSES/BY/4.0/

1.Introduction

With the widespread adoption of generative artificial intelligence (GenAI) in everyday communication, some users have begun turning to AI to discuss personal concerns, express emotions, and seek support. Unlike conventional information retrieval tools, GenAI can sustain natural-language conversations, and its response speed, patience, and emotional expression may influence users' assessments of interaction quality. However, the relationship between these perceived interaction experiences and users' overall evaluations of AI's ability to understand complex human emotions remains insufficiently explored.

Previous research has shown that experiences of emotional interaction with AI are shaped by supportive responses, perceptions of the conversational partner, and individual preferences (Meng and Dai, 2021; Luo et al., 2025; Rubin et al., 2025; Wenger et al., 2026). Nevertheless, further investigation is needed into the extent to which user background, usage behaviour, and specific interaction experiences provide information relevant to perceived emotional understanding, and which experience-related features make the greatest contributions to classification models.

Explainable machine learning offers an approach to investigating these questions. Supervised classification models can assess how well different questionnaire variables distinguish between users' overall evaluations. Feature ablation and model interpretation methods can then be used to compare the contributions of different information sources and identify prominent interaction experience features. However, when predictor items and the outcome variable contain similar wording, models may rely on semantic consistency between questionnaire responses, potentially affecting the interpretation of feature importance. It is therefore necessary to examine not only classification performance but also the extent to which model results depend on particular input items.

Using questionnaire data from respondents with previous experience of emotional disclosure to AI, this study applies logistic regression, random forests, and repeated nested cross-validation, together with feature-group ablation, permutation importance, and SHAP analysis, to address three research questions:

RQ1: How much information relevant to classifying perceived AI emotional understanding is provided by user background, usage behaviour, and interaction experience?

RQ2: Which interaction experience features contribute most strongly to model classification, and to what extent are their importance and contribution directions consistent across data partitions?

RQ3: How do model performance and the classification contributions of key features change when items with semantic similarities to the outcome variable are excluded?

This study contributes by comparing the classification information provided by different questionnaire feature groups within a consistent machine learning evaluation framework, identifying prominent interaction experience features associated with perceived emotional understanding, and examining the effects of feature selection through item-removal experiments.

2.Literature review

2.1 Emotional Support and User Experience in Generative AI

The use of artificial intelligence in emotionally supportive communication has attracted growing attention in human–computer interaction research. Meng and Dai (2021) investigated supportive responses from chatbots and human conversational partners, finding that emotional support could influence participants' stress and concerns through perceived supportiveness. However, differences between human and chatbot interactions suggest that the effects of supportive communication also depend on the conversational partner and interaction context.

With the development of GenAI, researchers have increasingly examined users' emotional disclosure experiences in everyday settings. In a mixed-methods study involving 270 adults who used ChatGPT for emotional or mental health support, Luo et al. (2025) identified emotional support, freedom of expression, and information quality as important reasons for users' positive experiences. At the same time, participants reported limitations such as insufficient personalisation, superficial emotional responses, and mechanical replies.

These findings suggest that emotional interaction with AI involves multiple dimensions of user experience. Response immediacy, patience, and freedom of expression may represent distinct aspects of interaction quality. Rather than examining only overall satisfaction, it is therefore useful to investigate how specific interaction experiences relate to users' broader evaluations of AI's emotional understanding.

2.2 Subjective Evaluations of Emotional Understanding and Related Factors

Users' assessments of AI's emotional understanding are not determined solely by the content of its responses. Rubin et al. (2025) found that participants could report higher perceived empathy when they believed a response came from a human, even when the responses were AI-generated. This suggests that beliefs about the source of a response can influence subjective evaluations.

Wenger et al. (2026) further reported that people may prefer to receive emotional support from humans even when they rate AI-generated empathetic responses more favourably on certain dimensions. These findings indicate that evaluations of response quality, perceived empathy, and preferences regarding conversational partners are related but distinct outcomes.

Accordingly, this study defines perceived emotional understanding as respondents' subjective evaluations of GenAI's ability to understand complex human emotions. It does not imply that AI possesses human empathy, nor does it equate favourable evaluations with improvements in mental health.

Although existing research highlights the importance of supportive content and perceptions of conversational partners, further analysis is needed to establish how different categories of user information distinguish between evaluations of emotional understanding.

2.3 Explainable Machine Learning and Questionnaire Measurement

Machine learning can identify statistical patterns linking multiple input variables to an outcome, making it suitable for analysing structured questionnaire data containing demographic information, usage behaviour, and interaction experience measures. Unlike analyses focusing on individual associations, classification models can evaluate the discriminative ability of feature sets on held-out observations. Feature-group ablation can then be used to compare the contributions of different information sources.

However, classification performance does not necessarily indicate that a model has identified an independent psychological mechanism. In questionnaire research, predictor variables and outcome measures may contain similar content. For example, an item assessing whether AI responses feel mechanical overlaps in wording with an outcome category describing AI as "sometimes mechanical". A model may exploit this consistency between responses, meaning that the feature's apparent importance may not entirely reflect distinct information about interaction experience.

Furthermore, experience-related items and overall evaluations completed by the same respondent may be affected by common method factors (Podsakoff et al., 2003). High classification performance or a prominent SHAP ranking alone is therefore insufficient to establish the robustness of feature interpretations.

To address these considerations, this study uses feature-group ablation to compare the classification ability of different information sources, alongside permutation importance, SHAP, and item-removal experiments to examine key interaction experience features under different input conditions.

3.Experiments and Results

3.1Data Source and Research Design

This study uses data from a Chinese online questionnaire entitled "AI Emotional Companionship Survey". The questionnaire was distributed through social media and voluntarily shared by participants. Respondents participated with informed consent. A total of 347 questionnaire records were collected.

Based on Q7, which asked whether respondents had disclosed emotions, concerns, or secrets to AI during the previous six months, 201 respondents with relevant experience were initially selected. Four records without an evaluable response concerning AI's understanding of complex emotions and two records containing inconsistent answers about AI usage were subsequently excluded. The final analytical sample consisted of 195 respondents, including 119 women and 76 men. Of these, 161 were aged 18–22, and 159 were undergraduate students.

The classification outcome was based on Q23: "How would you evaluate current AI systems in terms of their ability to understand complex human emotions?" Responses of "Excellent" (n = 19) and "Fairly good; able to understand most emotions" (n = 111) were combined into the positive evaluation category (n = 130), coded as 1. Responses of "Average; sometimes mechanical" (n = 62) and "Poor; completely unable to understand me" (n = 3) were combined into the other explicit evaluation category (n = 65), coded as 0.

Fourteen predictor items were divided into three feature groups:

Background information (B): Q1–Q5, covering gender, age, respondent status, daily internet use, and relationship status.

Usage behaviour (U): Q6–Q7, covering previous AI usage and emotional disclosure to AI during the preceding six months.

Interaction experience (E): Q14–Q20, covering freedom of expression, response immediacy, patience, objective analysis, mechanical responses and insufficient emotional resonance, privacy concerns, and generic responses.

Interaction experience items were measured using five-point agreement scales, with negatively worded items retaining their original coding direction. Q21 was excluded from the predictors because it directly addressed AI's ability to understand complex emotions and implicit meanings. A conservative feature set was also constructed by excluding Q18 and Q20, with additional item-removal experiments used to examine the contributions of key predictors.

Figure 1 presents the overall research workflow, from sample selection and predictor construction to model evaluation and interpretation. Following the identification of the final sample of 195 respondents, the study constructed the three feature groups and conducted cross-validation, feature ablation, and model interpretation analyses.

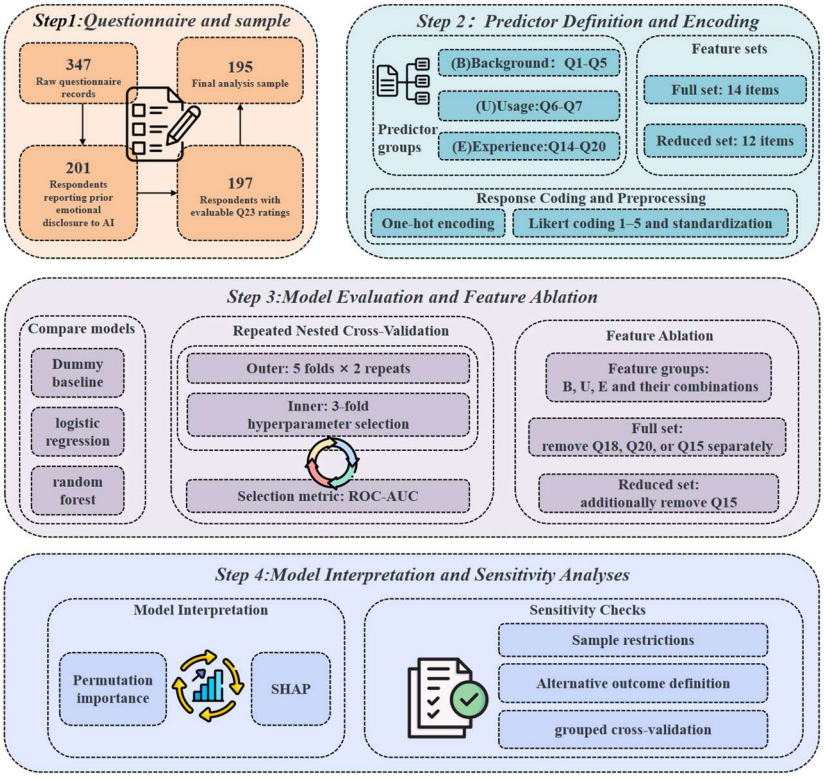


Figure1： Sample selection and explainable machine learning workflow

3.2Model Development and Evaluation

To assess whether questionnaire features could distinguish between respondents' evaluations of GenAI's emotional understanding, Q23 was treated as a binary classification outcome. Three models were compared: a Dummy Classifier, logistic regression (LR), and random forest (RF). The Dummy Classifier provided a baseline, LR served as a linear classification model, and RF was used to capture potential nonlinear relationships and feature interactions. RF was also selected for subsequent feature interpretation.

Background and usage behaviour variables (Q1–Q7) were one-hot encoded, while interaction experience variables (Q14–Q20) were standardised. All preprocessing was performed within the training folds. LR used balanced class weights and the liblinear solver, with the regularisation parameter C selected from 0.1, 1, and 10. RF consisted of 80 decision trees, used balanced class weights and a minimum leaf size of five, and considered maximum tree depths of three and five.

Because some interaction experience items overlapped in content with the outcome variable, two feature sets were defined: a full set containing 14 items and a conservative set containing 12 items. Q18 directly overlapped with the wording "mechanical" in Q23, while Q20, concerning generic responses, was considered potentially related in content to overall evaluations of emotional understanding. Both items were excluded from the conservative set, and their individual contributions were examined through additional item-removal experiments.

Model performance was evaluated using repeated nested cross-validation, consisting of five stratified outer folds repeated twice and three inner folds for hyperparameter selection based on ROC-AUC. Consistent data partitions were used across the main model comparisons. Performance was assessed using ROC-AUC, balanced accuracy (BA), positive-class F1 score, and Matthews correlation coefficient (MCC).

To address the research questions, feature-group ablation was first performed using B, U, E, and their combinations. Permutation importance and SHAP were then applied to identify key predictors in the random forest model. Finally, Q18, Q20, and Q15 were removed individually, and Q15 was additionally removed from

the conservative feature set to assess the effects of feature selection on model performance. Sensitivity analyses examined alternative sample restrictions, response-pattern grouping, and outcome definitions.

3.3 Overall Classification Performance

To account for possible content overlap between interaction experience items and the outcome measure, model performance was compared under full and conservative feature settings. The full feature set comprised 14 items from background information (Q1 – Q5), usage behaviour (Q6 – Q7), and interaction experience (Q14 – Q20). The conservative set excluded Q18 and Q20. Q18 directly shared the term "mechanical" with an answer option in Q23, while Q20, which concerned generic responses, was considered potentially related to respondents' overall evaluations of emotional understanding. This comparison was intended to examine the extent to which classification performance depended on these items, while subsequent item-removal experiments assessed their individual contributions.

The results of repeated nested cross-validation are summarised in Table 1.

Table 1：Classification performance under full and conservative feature settings

Feature Set	Model	AUC (SD)	BA	MCC	F1+
Full	Dummy	0.500 (0.000)	0.500	0.000	0.800
Full	LR	0.695 (0.085)	0.637	0.262	0.697
Full	RF	0.730 (0.086)	0.667	0.325	0.743
Conservative	Dummy	0.500 (0.000)	0.500	0.000	0.800
Conservative	LR	0.512 (0.095)	0.527	0.052	0.593
Conservative	RF	0.666 (0.097)	0.644	0.289	0.751

Under the full feature setting, RF achieved a mean ROC-AUC of 0.730, compared with 0.695 for LR and 0.500 for the Dummy Classifier. RF also achieved higher balanced accuracy (0.667) and MCC (0.325) than LR (0.637 and 0.262, respectively), indicating relatively better overall classification performance in the present dataset.

After removing Q18 and Q20, the RF AUC declined from 0.730 to 0.666, while the LR AUC fell from 0.695 to 0.512. Both models showed reduced performance, although RF retained some discriminative ability. This suggests that Q18 and Q20 jointly provided information relevant to the outcome and motivates further examination of their individual contributions.

Beyond the mean performance metrics, Figure 2a presents the distribution of ROC-AUC values across the ten outer test folds. Although RF achieved a relatively high mean AUC, performance varied across folds, indicating some sensitivity to data partitioning. Figure 2b compares RF performance under alternative sample

restrictions, data partitioning procedures, and outcome definitions. These sensitivity analyses are discussed further in Section 3.7.

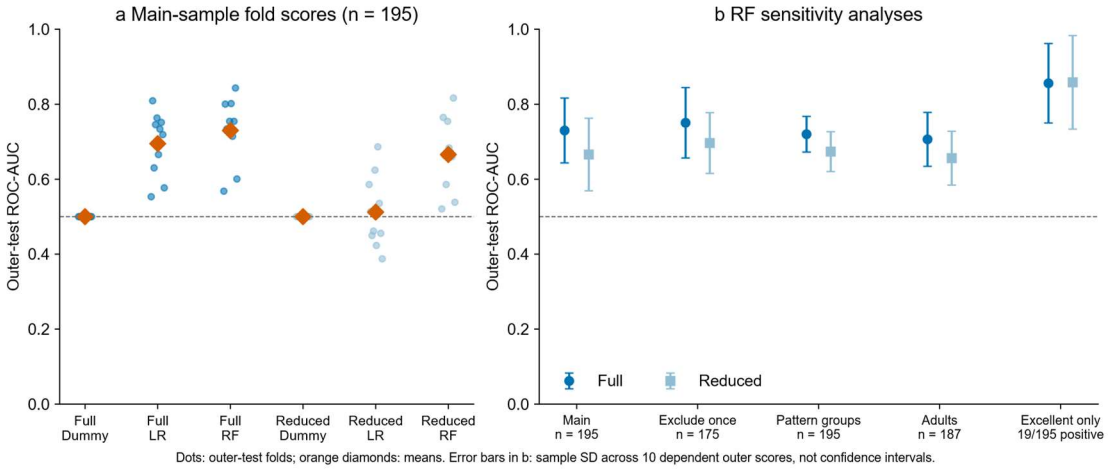


Figure2:Out-of-fold model performance and random forest sensitivity analyses

(a) Out-of-fold ROC-AUC values and mean performance for different models under full and conservative feature settings.

(b) Mean ROC-AUC and standard deviations for random forests under alternative sample and outcome definitions. Error bars indicate standard deviations across outer test folds rather than confidence intervals for population-level performance.

3.4 Classification Ability of Different Feature Groups

To compare the discriminative information provided by different categories of questionnaire variables, separate models were developed using background information (B), usage behaviour (U), interaction experience (E), and their combinations. The results are presented in Table 2. This experiment examined the relative classification ability of each information source and whether combining additional feature groups improved performance.

Table2:Classification performance across feature-group combinations

Feature group	Number of questions	LR AUC (SD)	RF AUC (SD)
B	5	0.464 (0.074)	0.527 (0.062)
U	2	0.502 (0.070)	0.482 (0.086)
E	7	0.720 (0.073)	0.746 (0.102)
B + U	7	0.438 (0.070)	0.478 (0.070)
B + E	12	0.684 (0.083)	0.738 (0.087)
U + E	9	0.712 (0.074)	0.747 (0.095)

Feature group	Number of questions	LR AUC (SD)	RF AUC (SD)
B + U + E	14	0.695 (0.085)	0.730 (0.086)

For RF, the AUC was 0.527 when using background information alone, 0.482 when using usage behaviour alone, and 0.478 when combining background and usage behaviour. All three values were close to the no-discrimination benchmark of 0.5.

In contrast, using interaction experience features alone produced an RF AUC of 0.746 (SD = 0.102), compared with 0.478 for the combined background and usage behaviour features. The mean matched-fold AUC difference was 0.269, with the same direction observed across all ten outer test folds. LR showed a similar pattern, achieving an AUC of 0.720 using interaction experience features alone, compared with 0.438 using background and usage behaviour.

Further comparisons showed RF AUC values of 0.738 for B+E, 0.747 for U+E, and 0.730 for the full B+U+E feature set. The E-only model achieved an AUC of 0.746, comparable to U+E and higher than the full model. Thus, including additional background and usage behaviour features did not produce a clear improvement in classification performance.

These findings indicate that interaction experience was the main source of information associated with perceived emotional understanding in the current questionnaire. Compared with demographic characteristics and previous AI usage, respondents' subjective assessments of AI response styles and interaction quality were more informative for distinguishing between overall evaluations.

This finding is consistent with previous research emphasising the importance of response quality and interaction experience (Meng and Dai, 2021; Luo et al., 2025). However, interaction experience items and the outcome variable are both subjective evaluations, and their similarity in content may contribute to classification performance. Further analysis of individual items is therefore necessary.

Overall, Table 2 shows that interaction experience features achieved higher classification performance than background and usage behaviour features under both RF and LR. For RF, the AUC values were 0.746 and 0.478, respectively, with a mean matched-fold difference of 0.269. Adding background and usage behaviour to the experience features produced no clear additional benefit. In the present dataset, the primary discriminative information was therefore concentrated in respondents' subjective assessments of their interactions with AI.

3.5 Interpretation of Key Interaction Experience Features

Having established that interaction experience features contained more discriminative information, permutation importance and SHAP were used to examine the individual predictors on which the RF model relied. Figure 3 presents item-level SHAP contributions for the full RF model (Figure 3a) and permutation importance (Figure 3b). The former measures the magnitude of each feature's contribution to predicted probabilities, while the latter measures changes in classification performance when an individual item is randomly permuted.

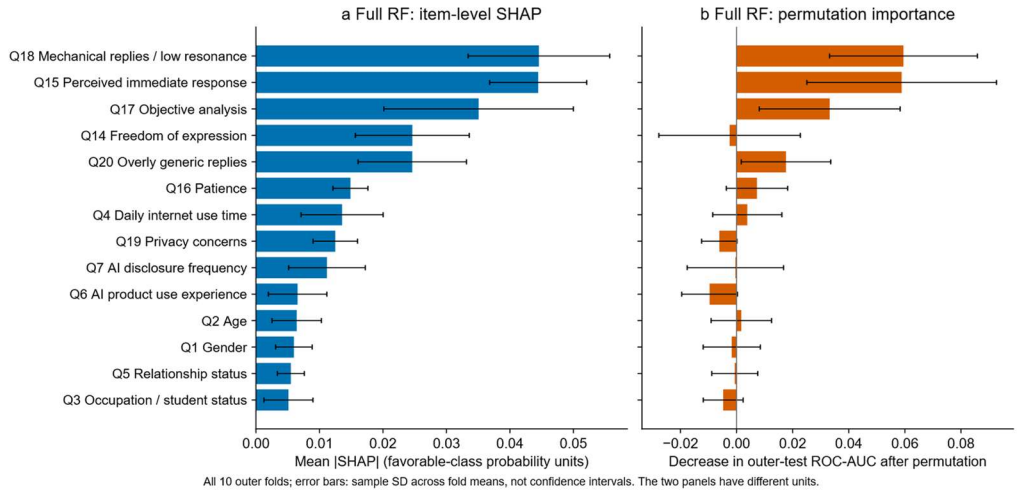


Figure3:SHAP contributions and permutation importance for the full random forest model

As shown in Figure 3, Q18 (mechanical responses and insufficient emotional resonance) and Q15 (perceived response immediacy) ranked among the most important features under both interpretation methods. Permuting Q18 and Q15 reduced the model's test AUC by an average of 0.0595 and 0.0589, respectively, representing the largest decreases among the questionnaire items. This indicates that the model relied strongly on information from these two interaction experience features when distinguishing respondents' evaluations of emotional understanding. Q17 (objective analysis) also contributed to classification, although its importance was lower.

SHAP analysis further clarified the relationships between these features and model predictions. For Q18, higher agreement was associated with a negative mean SHAP contribution (−0.0579), whereas lower agreement was associated with a positive contribution (0.0849). Thus, when respondents perceived AI replies as mechanical or lacking emotional resonance, the model tended to assign a lower probability to the positive evaluation category.

Q15 showed the opposite pattern. Respondents who more strongly recognised the immediacy of AI responses had a positive mean SHAP contribution (0.0318), whereas lower agreement corresponded to a negative contribution (−0.0517). In the current model, more favourable perceptions of response immediacy were therefore associated with higher predicted probabilities of positive emotional understanding evaluations.

The two interpretation methods produced broadly consistent rankings of the leading features. Across the ten outer test folds, the mean pairwise Spearman correlation between SHAP feature rankings was 0.799, indicating some consistency across data partitions. Overall, the model relied not only on whether respondents perceived AI responses as mechanical but also on their evaluations of response immediacy.

However, Q18 directly overlaps with the wording "sometimes mechanical" in Q23. Its high importance may therefore partly reflect similarities in questionnaire content. The following section uses item-removal experiments to examine the classification contributions of Q18 and Q15 under different feature settings.

3.6 Semantic Overlap and Item-Removal Experiments

The preceding analysis identified Q18 and Q15 as prominent interaction experience features. However, Q18 directly overlaps in wording with the outcome item Q23, raising the possibility that its apparent importance partly reflects semantic consistency between questionnaire responses. To investigate this issue, Q18, Q20, and Q15 were removed individually, and the models were retrained to compare classification performance. Q15 was

also removed from the conservative feature set, which already excluded Q18 and Q20, to examine whether its contribution changed under different input conditions. The results are presented in Table 3.

Table3:Item-removal experiments and further ablation of the conservative feature set

Condition	LR AUC (SD)	ΔAUC	RF AUC (SD)	ΔAUC
Full set: 14 items	0.695 (0.085)	+0.000	0.730 (0.086)	+0.000
Remove Q18	0.605 (0.099)	-0.090	0.672 (0.089)	-0.058
Remove Q20	0.706 (0.079)	+0.011	0.699 (0.080)	-0.030
Remove Q15	0.662 (0.099)	-0.033	0.695 (0.088)	-0.035
Remove Q18 and Q20	0.512 (0.095)	-0.183	0.666 (0.097)	-0.064
Conservative set: additionally remove Q15	0.472 (0.123)	-0.040	0.590 (0.105)	-0.076

The result shows that ΔAUC values are calculated relative to the full feature set, except for the final row, which is compared with the conservative feature set.

Removing Q18 alone reduced AUC by 0.058 for RF and 0.090 for LR, indicating that Q18 contained substantial discriminative information. However, its contribution may partly reflect overlapping wording with Q23. By comparison, removing Q20 reduced RF AUC by 0.030 but increased LR AUC by 0.011, indicating that its contribution differed between the two models.

The contribution of Q15 varied with the feature set. Removing Q15 from the full feature set reduced RF AUC by 0.035. When Q15 was further removed from the conservative set excluding Q18 and Q20, RF AUC decreased by 0.076. This difference suggests that Q15 may share partially overlapping or substitutable classification information with other experience-related items, although its precise contribution depends on the feature combination.

Taken together, SHAP, permutation importance, and item-removal results identify Q18 and Q15 as important features in the current RF classification task. While the contribution of Q18 may partly reflect semantic overlap with the outcome measure, Q15 retained a substantial classification contribution after Q18 and Q20 were excluded. Perceived response immediacy therefore remained an important source of information for distinguishing users' evaluations of AI emotional understanding.

3.7 Sensitivity and Robustness Analyses

To assess whether model performance depended on particular sample restrictions, data partitions, or outcome definitions, four sensitivity analyses were conducted. These involved excluding respondents who had disclosed emotions to AI only once, grouping identical response patterns during cross-validation, restricting the sample to adults, and redefining the positive evaluation category. RF classification performance under the full and conservative feature settings is summarised in Table 4.

Table4:Sensitivity analyses under alternative samples and outcome definitions

Analysis condition	Total / positive cases	Full RF AUC (SD)	Conservative RF AUC (SD)
Main analysis	195 / 130	0.730 (0.086)	0.666 (0.097)
Excluding one-time disclosure respondents	175 / 116	0.750 (0.094)	0.696 (0.081)
Grouped by complete response patterns	195 / 130	0.720 (0.048)	0.674 (0.053)
Adults only	187 / 123	0.706 (0.072)	0.656 (0.072)
"Excellent" as the only positive category	195 / 19	0.856 (0.106)	0.858 (0.125)

After excluding respondents who had disclosed emotions to AI only once, the AUC of the full RF model changed from 0.730 to 0.750. Grouping identical response patterns during cross-validation produced an AUC of 0.720, while restricting the sample to adults resulted in an AUC of 0.706. These values were relatively close to the main analysis, indicating that classification performance showed some consistency across the alternative sample and partitioning conditions. The conservative RF model exhibited a similar pattern.

Overall, RF retained some discriminative ability under different sample restrictions and data partitioning procedures, suggesting that the main classification results were not entirely dependent on a single analytical setting. Nevertheless, these analyses were conducted using the same questionnaire dataset, and performance on independent samples remains to be established.

4.Conclusion

Using questionnaire data from 195 respondents with previous experience of emotional disclosure to AI, this study applied explainable machine learning to examine the relationships between user background, usage behaviour, interaction experience, and perceived emotional understanding of GenAI.

The results indicate that interaction experience was the primary source of discriminative information in the current questionnaire. Using interaction experience features alone, RF achieved an AUC of 0.746, compared with 0.478 for the combined background and usage behaviour features. Adding background and usage variables did not produce a clear improvement in classification performance.

Model interpretation further identified perceived mechanical responses (Q18) and perceived response immediacy (Q15) as prominent interaction experience features. Because Q18 overlapped in wording with the overall evaluation measure, item-removal experiments were conducted to examine the contributions of key predictors. Removing Q18 reduced model performance, while Q15 retained a substantial classification contribution after Q18 and Q20 were excluded. These findings indicate that the model relied not only on questionnaire items with similar evaluative content but also on other interaction experience features, particularly perceived response immediacy.

By combining feature-group ablation with SHAP, permutation importance, and item-removal experiments, this study provides a more detailed examination of the classification contributions of interaction experience features. The findings offer empirical evidence concerning users' subjective evaluations of emotional understanding in GenAI interactions and highlight the need to consider measurement overlap and dependencies between features when interpreting explainable machine learning results based on questionnaires.

Given the cross-sectional questionnaire design, reliance on subjective evaluations, and absence of independent external validation, the findings do not establish causal relationships between interaction experience and perceived emotional understanding. Future research could incorporate independent samples, more clearly differentiated measures of emotional experience, and behavioural indicators from actual interactions to examine the consistency of key experience features across user groups and application contexts.

Acknowledgments

The author would like to thank Xinyi Cai for her assistance with questionnaire data collection and Yu Zou for reviewing the manuscript and providing valuable feedback.

Disclosure Statement

Participation in the questionnaire survey was voluntary, and informed consent was obtained from respondents before data collection. The analysis was conducted using questionnaire responses for research purposes.

References

Cawley, G.C. and Talbot, N.L.C. (2010) 'On over-fitting in model selection and subsequent selection bias in performance evaluation', Journal of Machine Learning Research, 11, pp. 2079 – 2107. Available at: https://www.jmlr.org/papers/v11/cawley10a.html.

Lundberg, S.M. and Lee, S.-I. (2017) 'A unified approach to interpreting model predictions', Advances in Neural Information Processing Systems, 30. Available at: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.-I. (2020) 'From local explanations to global understanding with explainable AI for trees', Nature Machine Intelligence, 2, pp. 56 – 67. https://doi.org/10.1038/s42256-019-0138-9.

Luo, X., Wang, Z., Tilley, J.L., Balarajan, S., Bassey, U.-A. and Cheang, C.I. (2025) 'Seeking emotional and mental health support from generative AI: Mixed-methods study of ChatGPT user experiences', JMIR Mental Health, 12, e77951. https://doi.org/10.2196/77951.

Meng, J. and Dai, Y. (2021) 'Emotional support from AI chatbots: Should a supportive partner self-disclose or not?', Journal of Computer-Mediated Communication, 26(4), pp. 207 – 222. https://doi.org/10.1093/jcmc/zmab005.

Pedregosa, F., Varoquaux, G., Gramfort, A. et al. (2011) 'Scikit-learn: Machine learning in Python', Journal of Machine Learning Research, 12, pp. 2825 – 2830. Available at: https://jmlr.org/papers/v12/pedregosa11a.html.

Podsakoff, P.M., MacKenzie, S.B., Lee, J.-Y. and Podsakoff, N.P. (2003) 'Common method biases in behavioral research: A critical review of the literature and recommended remedies', Journal of Applied Psychology, 88(5), pp. 879 – 903. https://doi.org/10.1037/0021-9010.88.5.879.

Rubin, M., Li, J.Z., Zimmerman, F., Ong, D.C., Goldenberg, A. and Perry, A. (2025) 'Comparing the value of perceived human versus AI-generated empathy', Nature Human Behaviour, 9(11), pp. 2345 – 2359. https://doi.org/10.1038/s41562-025-02247-w.

Varoquaux, G. (2018) 'Cross-validation failure: Small sample sizes lead to large error bars', NeuroImage, 180, pp. 68 – 77. https://doi.org/10.1016/j.neuroimage.2017.06.061.

Wenger, J.D., Cameron, C.D. and Inzlicht, M. (2026) 'People choose to receive human empathy despite rating AI empathy higher', Communications Psychology, 4, article 19. https://doi.org/10.1038/s44271-025-00387-3.

Appendix

Appendix A. Questionnaire Variables and Coding

This appendix summarises the questionnaire variables included in the analysis. The predictors were divided into background information (B), usage behaviour (U), and interaction experience (E). The outcome variable was derived from Q23.

Table A1. Questionnaire variables used in the analysis

Item	Variable description	Feature group	Encoding
Q1	Gender	Background (B)	Categorical
Q2	Age	Background (B)	Categorical
Q3	Respondent status	Background (B)	Categorical
Q4	Daily internet usage	Background (B)	Categorical
Q5	Relationship status	Background (B)	Categorical
Q6	Previous AI usage	Usage (U)	Categorical
Q7	Emotional disclosure to AI	Usage (U)	Categorical
Q14	Freedom of emotional expression	Experience (E)	Likert scale (1–5)
Q15	Perceived response immediacy	Experience (E)	Likert scale (1–5)
Q16	Perceived patience	Experience (E)	Likert scale (1–5)
Q17	Perceived objectivity of analysis	Experience (E)	Likert scale (1–5)

Item	Variable description	Feature group	Encoding
Q18	Perceived mechanical responses and insufficient emotional resonance	Experience (E)	Likert scale (1–5)
Q19	Privacy concerns	Experience (E)	Likert scale (1–5)
Q20	Perceived generic responses	Experience (E)	Likert scale (1–5)
Q23	Evaluation of AI's ability to understand complex human emotions	Outcome	Binary

Categorical predictors were one-hot encoded, while the interaction experience items retained their original five-point response direction and were standardised within the training folds.

Table A2. Binary coding of the outcome variable (Q23)

Original response	Number of respondents	Binary label
Excellent	19	1
Fairly good; understands most emotions	111	1
Average; sometimes mechanical	62	0
Poor; does not understand me at all	3	0
Total	195	

The positive evaluation category (1) comprised 130 respondents, while the other explicit evaluation category (0) comprised 65 respondents.

Q21 was excluded from the predictors because its content directly concerned AI's understanding of complex emotions and implicit meanings.

The full feature set comprised 14 items (Q1–Q7 and Q14–Q20). The conservative feature set excluded Q18 and Q20, leaving 12 items. Q18 was excluded because its wording concerning mechanical responses directly overlapped with an answer option in Q23. Q20 was additionally excluded as a more conservative measure because its content concerning generic responses might overlap with overall evaluations of emotional understanding.

Appendix B. Model Configuration and Validation

B.1 Model Configuration

Table B1. Model and validation settings

Component	Configuration
Main analytical sample	195 respondents
Classification task	Binary classification based on Q23
Baseline model	Dummy Classifier
Linear model	Logistic Regression (LR)
Tree-based model	Random Forest (RF)
Categorical preprocessing	One-hot encoding
Interaction experience preprocessing	Standardisation
LR class weights	Balanced
LR solver	liblinear
LR regularisation parameter (C)	0.1, 1, 10
LR maximum iterations	3,000
RF number of trees	80
RF class weights	Balanced
RF minimum samples per leaf	5
RF maximum depth	3, 5
Outer cross-validation	5 stratified folds × 2 repeats
Inner cross-validation	3 folds
Hyperparameter selection metric	ROC-AUC
Classification threshold	0.5
Evaluation metrics	ROC-AUC, balanced accuracy, positive-class F1, MCC

All preprocessing and hyperparameter selection were performed within the corresponding training folds. Outer test folds were reserved for model evaluation. The same outer data partitions were used when comparing the main model and feature-ablation conditions.

B.2 Feature-Group Ablation

Seven feature-group configurations were evaluated:

B, U, E, B+U, B+E, U+E, and B+U+E.

Each feature configuration was evaluated using the same repeated nested cross-validation framework to compare the classification information provided by different questionnaire variable groups.

B.3 Item-Removal Experiments

The following configurations were evaluated:

1.Full feature set (14 items).
2.Full feature set excluding Q18.
3.Full feature set excluding Q20.
4.Full feature set excluding Q15.
5.Conservative feature set excluding Q18 and Q20.
6.Conservative feature set additionally excluding Q15.

Models were retrained and hyperparameters were selected separately for each configuration.