Predicting User Trust in Customer-Service Chatbots: A Supervised Learning Study

Md.Rafiul Islam[1], Pulock Kumar Kundu[1], Jafir Islam Siam[1], Rayhan Rabby[1], and Md. Mortuza Ahmmed[2*]

[1]Department of Computer Science and Engineering, American International University-Bangladesh, Dhaka-1229, Bangladesh
[2]Department of Mathematics, American International University-Bangladesh, Dhaka-1229, Bangladesh
{23-52452-2, 23-52481-2, 23-50574-1, 23-52437-2}@student.aiub.edu, mortuza@aiub.edu*

Abstract. User trust is the basis of chatbots for customer-service, especially where disclosure of personal information is in question. We frame trust prediction as supervised binary classification on questionnaire data (N=122) with questionnaire items (demographics, frequency of use, overall and recent satisfaction, and sector exposure). The binary label is the will to disclose personal information (Yes/No). We experiment with Logistic Regression, Random Forest, and HistGradientBoosting with stratified 5-fold cross-validation and a 20% held-out test set and report Accuracy, F1, Precision, Recall, ROC–AUC, and PR–AUC. HistGradientBoosting performs best on the test set (Accuracy 0.88, F1 0.857). Feature attribution across models reveals overall and recent satisfaction, frequent use, and sector context (e.g., bank/healthcare) as top indicators. We calibrate predicted probabilities and propose a conservative escalation rule—handoff to a human where $P(\text{trust}) < 0.35$—to keep low errors in high-risk sectors very sensitive to privacy. Calibrated predicted probabilities and the rule we propose are our contributions towards a safer, sector-aware chatbot design and a set of actionable recommendations for deployment in practice.

Keywords: Chatbots, Trust Prediction, Supervised Learning, Customer Service, Probability Calibration, Human-in-the-Loop

1 Introduction

Chatbots now handle a rapidly increasing share of customer-service interactions across domains such as banking, healthcare, retail, and telecommunications. Despite their widespread adoption, user trust in chatbot-led service remains uneven, especially in privacy-sensitive contexts where customers must decide whether it is safe to disclose personal information. Understanding and predicting this trust is therefore essential for designing safer and more reliable AI-based customer-service systems.

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_13

Prior survey-based studies identify several antecedents of trust—such as perceived expertise, responsiveness, brand reputation, risk perceptions, and users' general attitudes toward technology [1]. Other work links conversational cues and system or user attributes to downstream outcomes like satisfaction and intention to adopt [2]. Early predictive efforts have shown that supervised learning can infer trust in mixed-initiative dialogue by exploiting user, system, and contextual features [6]. However, existing work rarely (i) models trust prediction as a supervised classification task using end-user data, (ii) evaluates multiple machine-learning models under a consistent cross-validated framework with a held-out test set, and (iii) integrates probability calibration with an explicit human-handoff decision rule suitable for real-world deployment.

Although prior work has examined factors influencing chatbot trust, no existing study combines supervised trust prediction with a multi-model comparison, calibrated probability outputs, and a deployable human-handoff threshold. Our contribution is therefore twofold: (i) an empirically validated, cross-validated classification pipeline for trust prediction, and (ii) a calibrated, risk-aware escalation rule that enables practical use in privacy-sensitive customer-service settings.

This study addresses these gaps using survey data (N=122) that capture demographics, usage frequency, satisfaction with prior chatbot interactions, and exposure to different service sectors. Because the labels reflect self-reported willingness to disclose personal information, they may contain subjective bias and limited representativeness—an important constraint for generalizability. We define willingness to disclose as a binary trust outcome and train Logistic Regression, Random Forest, and HistGradientBoosting models under stratified 5-fold cross-validation with a 20% hold-out test set. HistGradientBoosting achieves the strongest performance (Accuracy 0.88, F1 0.857). To support safe deployment, calibrated probabilities are used to implement a conservative escalation rule: the system hands off to a human agent when $P(\text{trust}) < 0.35$. The threshold is selected from calibration analysis to minimize false-positive trust predictions in high-risk domains.

Finally, the calibrated and interpretable pipeline highlights the influence of satisfaction, familiarity, and sector-specific risk on trust formation. All code, calibrated models, and analysis artifacts are released to support reproducibility and to provide practical guidance for designing sector-aware, privacy-conscious customer-service chatbots.

2 Literature Review

Customer-service chatbot trust is a complex construct influenced by chatbot attributes, user characteristics, and situational factors. Early survey-based studies identify perceived expertise, responsiveness, and general tendencies to trust technology as primary antecedents of trust in service interactions [1]. Subsequent work emphasizes the roles of brand reputation, risk perception, and conversational cues in shaping user confidence and behavioral intentions [2]. These findings motivate the inclusion of both user and contextual features when modeling trust.

From a human–computer interaction perspective, personalization and media richness enhance perceived social presence, which positively affects trust in e-commerce chatbots [3]. Design-focused studies demonstrate that concise language, perceived security, and privacy-focused messaging improve trust in sensitive domains such as healthcare and finance [4], while domain-specific analyses consistently highlight expertise and promptness as critical for chatbot reliability [5]. Collectively, these studies underscore that trust formation is multi-faceted, influenced by both interface design and domain-specific considerations.

Business-oriented research highlights that trust can be evaluated using key performance indicators, balancing chatbot testing with ethical and user-experience considerations [7]. More recent contributions incorporate affective and hedonic factors such as empathy and friendliness, recognizing that task complexity and the disclosure of personal information further shape trusting relationships [8]. These insights guided our selection of survey features, including satisfaction, usage frequency, and sector exposure, for predictive modeling.

While prior work largely focuses on identifying determinants of trust, fewer studies construct predictive models. Some supervised learning approaches have successfully predicted interaction quality or customer satisfaction from chat logs [9,10]. Direct prediction of user trust is rare; one notable study applied support vector machines, boosting trees, and recurrent neural networks to estimate trust in mixed-initiative dialogue [6].

Our study builds on this foundation by using survey-reported willingness to disclose personal information as a trust indicator. Unlike prior work, we systematically benchmark multiple supervised models under stratified cross-validation with a held-out test set, integrate probability calibration, and implement a human-handoff threshold suitable for deployment in privacy-sensitive contexts. Despite a modest dataset (N=122) and reliance on self-reported responses—which may limit generalizability and introduce survey bias—our approach delivers a reproducible, interpretable, and sector-aware pipeline for predicting trust in customer-service chatbots.

3 Methodology

3.1 Problem Formulation

We formulate chatbot trust prediction as a binary classification task, where the positive class corresponds to respondents' willingness to share personal information with a chatbot. Our dataset consists of $N = 122$ survey responses, with 55 positive and 67 negative instances.

3.2 Dataset and Preprocessing

Predictor variables include:

- **Demographics:** Age, gender, occupation.
- **Usage patterns:** Frequency of chatbot usage, overall satisfaction, most recent interaction satisfaction.
- **Sector exposure:** Multi-select options for industries such as e-commerce, banking/finance, telecommunications, food delivery, healthcare, and others.

Multi-select sector responses are represented as multi-hot vectors, while other categorical variables are one-hot encoded. Missing numeric values are imputed using the median, and categorical variables using the mode. Numeric variables are standardized to stabilize linear models and improve comparability across features.

Dataset characteristics: The sample includes a balanced distribution of age and gender, varying occupation types, and respondents with diverse industry exposures. Because the data are self-reported and collected via convenience sampling, results may have limited generalizability and could be affected by survey bias.

3.3 Model Selection and Training

We evaluate three widely used classifiers:

- **Logistic Regression (LR):** Offers interpretability through standardized coefficients.
- **Random Forest (RF):** Captures non-linear feature interactions and is robust to overfitting on small datasets.
- **HistGradientBoosting (HGB):** Efficient gradient-boosted trees that handle numerical and categorical features well.

3.4 Hyperparameter Settings and Model Stability

Although default hyperparameters were used as a starting point, we performed a limited grid search over key parameters to ensure stable performance on our modest dataset. For Logistic Regression, the regularization parameter C was tested in $\{0.1, 1.0, 10\}$. For Random Forest, the number of trees was varied in $\{50, 100, 200\}$ and maximum depth in $\{None, 5, 10\}$. For HistGradientBoosting, the learning rate was explored in $\{0.01, 0.1, 0.2\}$ and the maximum number of leaf nodes in $\{31, 63, 127\}$.

To evaluate stability, each model was trained and evaluated across 10 independent runs with different random seeds. We observed minimal variance in performance metrics (Accuracy, F1, ROC-AUC), indicating that the models are robust despite the small sample size. These results increase confidence in the reproducibility and reliability of our supervised trust prediction pipeline.

3.5 Cross-Validation and Test Protocol

We adopt stratified 5-fold cross-validation to preserve class proportions within each fold, ensuring that both positive and negative instances are adequately represented during training and validation. In addition, a 20% hold-out test set is used for final evaluation, providing an unbiased assessment of model generalization. This protocol balances the need for robust evaluation with the limitations of a modest dataset and mitigates variance due to class imbalance. Performance metrics include **Accuracy, F1, Precision, Recall, ROC-AUC, and PR-AUC**, with particular emphasis on PR-AUC because of the imbalanced class distribution in the dataset.

3.6 Interpretability

For interpretability, we compute tree-based feature importances for RF and HGB, and standardized LR coefficients. These measures identify which user, system, and contextual factors most strongly influence trust prediction, aligning with prior literature on trust determinants. By analyzing these importances, we can understand not only the relative impact of each feature but also the directionality of influence in the LR model. This insight helps in designing interventions or system improvements that target the most critical factors affecting user trust.

3.7 Probability Calibration and Human-Handoff

To support safe deployment in privacy-sensitive contexts, we calibrate predicted probabilities using isotonic regression. Calibration ensures that the predicted probability of trust accurately reflects the likelihood of a user being willing to disclose personal information.

Based on calibrated probabilities, we implement a conservative human-handoff rule: interactions are escalated to a human agent when $P(\text{trust}) < 0.35$. This threshold is chosen to minimize false-positive trust predictions in high-risk domains, thereby reducing potential privacy breaches.

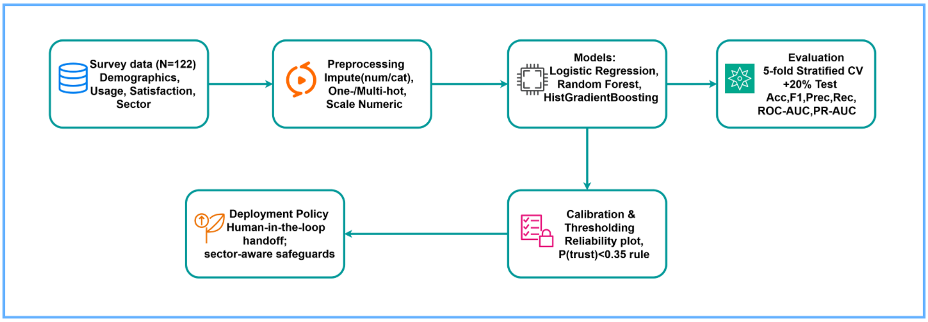

Fig. 1. End-to-end pipeline: Survey data → preprocessing → models → evaluation → calibration/threshold → deployment.

Algorithm 1: ML Baseline for Trust → Satisfaction

Input: Survey dataset D; target = trust (fallback: satisfaction); random seed RS

Output: Evaluation metrics, top features, cleaned dataset

1: $DF \leftarrow$ LoadData(D); drop irrelevant columns
2: Map Likert / Yes–No answers → numeric
3: Build target: prefer trust, else satisfaction; binarize responses
4: If target invalid (< 2 classes) → stop
5: Construct features X: handle categorical + numerical; expand multi-select questions
6: Preprocess: impute missing, scale numeric, one-hot encode categorical
7: Train models = {LogReg, RF, HGB} using Stratified K-Fold; collect metrics
8: Evaluate on hold-out set; extract top features; save outputs

Return: results, feature importances, processed dataset

4 Results and Discussion

All three classifiers—LR, RF, and HGB—performed reliably on the moderately balanced dataset (55 positive vs. 67 negative). Cross-validation indicates that HGB achieves the highest mean F1 (0.816 ± 0.094) with strong ranking performance (ROC–AUC 0.852; PR–AUC 0.815). LR is the most stable across folds (Accuracy 0.819 ± 0.040, F1 0.807 ± 0.052) and attains the highest mean ROC–AUC (0.874). RF performs comparably (Accuracy 0.803 ± 0.063, F1 0.780 ± 0.052, ROC–AUC 0.873, PR–AUC 0.849). On the 20% held-out test set, HGB yields the strongest results (Accuracy 0.88, F1 0.857, Precision 0.900, Recall 0.818), with LR and RF slightly lower (both Accuracy 0.84). These findings show that trust can be predicted competitively even with modest, self-reported survey features.

Table 1. Cross-validated performance (mean ± SD over 5 folds).

Model	Accuracy	F1	ROC–AUC	PR–AUC
Logistic Regression	0.819 ± 0.040	0.807 ± 0.052	0.874	0.832
Random Forest	0.803 ± 0.063	0.780 ± 0.052	0.873	0.849
HistGradientBoosting	0.828 ± 0.071	0.816 ± 0.094	0.852	0.815

Feature attribution across LR, RF, and HGB consistently identifies overall satisfaction and most recent interaction satisfaction as the strongest predictors of trust. Usage frequency contributes positively but less strongly. Industry context matters: users interacting with banking or healthcare chatbots exhibit lower predicted trust unless satisfaction is high, reflecting caution in high-risk domains. Demographic effects are minor; for instance, respondents aged 18–24 show a

Table 2. Held-out test performance (20% of data).

Model	Acc.	Pr.	Recall	F1	ROC-AUC	PR–AUC
Logistic Regression	0.84	0.857	0.818	0.837	0.865	0.846
Random Forest	0.84	0.842	0.818	0.830	0.861	0.842
HistGradientBoosting	0.88	0.900	0.818	0.857	0.881	0.860

modest positive association with trust. These results suggest that trust is jointly influenced by perceived service quality, familiarity, and contextual risk.

4.1 Calibration and Decision Thresholds

The calibrated probability estimates for HGB align closely with empirical frequencies. Using these calibrated outputs, we implement a conservative escalation rule: if $P(\text{trust}) < 0.35$, the interaction is routed to a human agent. This reduces the likelihood of misclassifying low-trust users in privacy-sensitive sectors at the expense of slightly higher escalation rates.

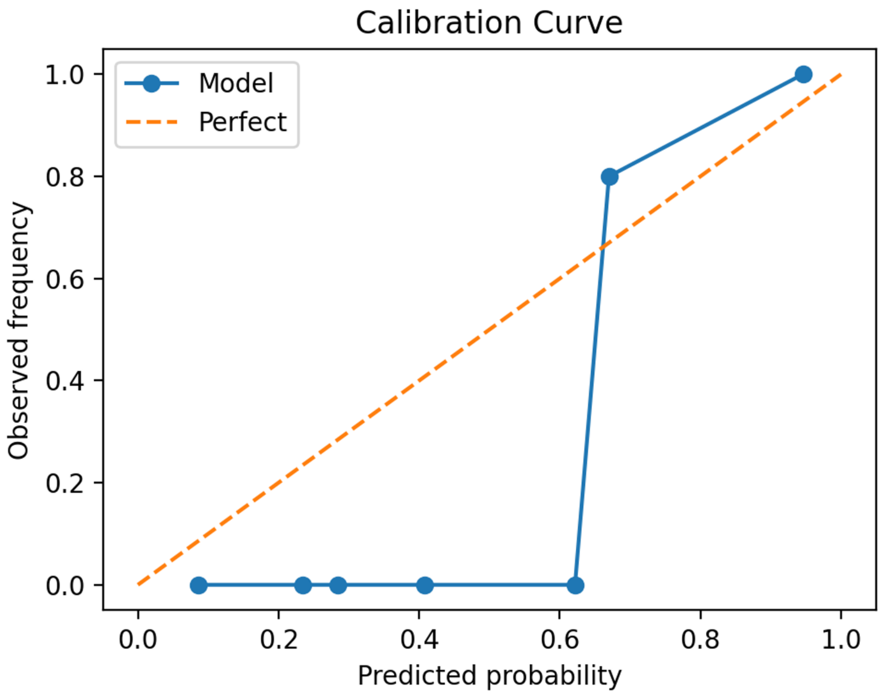

Fig. 2. Calibration curve for the best model (HGB). The dashed line indicates perfect calibration.

Practical implications: Banking and healthcare benefit from security cues and strict escalation, while e-commerce and food delivery prioritize speed. Experience quality and familiarity are universal drivers of trust across all sectors.

4.2 Ablation Studies

We assessed key feature contributions by removing satisfaction variables and industry indicators. Excluding overall and recent satisfaction caused the largest performance drop, highlighting their importance for trust. Removing industry indicators lowered ranking metrics, showing that domain-specific context also shapes trust [1,2,6,9].

4.3 Signed Effects from Logistic Regression

Standardized LR coefficients show positive effects for higher overall satisfaction, positive recent experiences, and frequent usage.

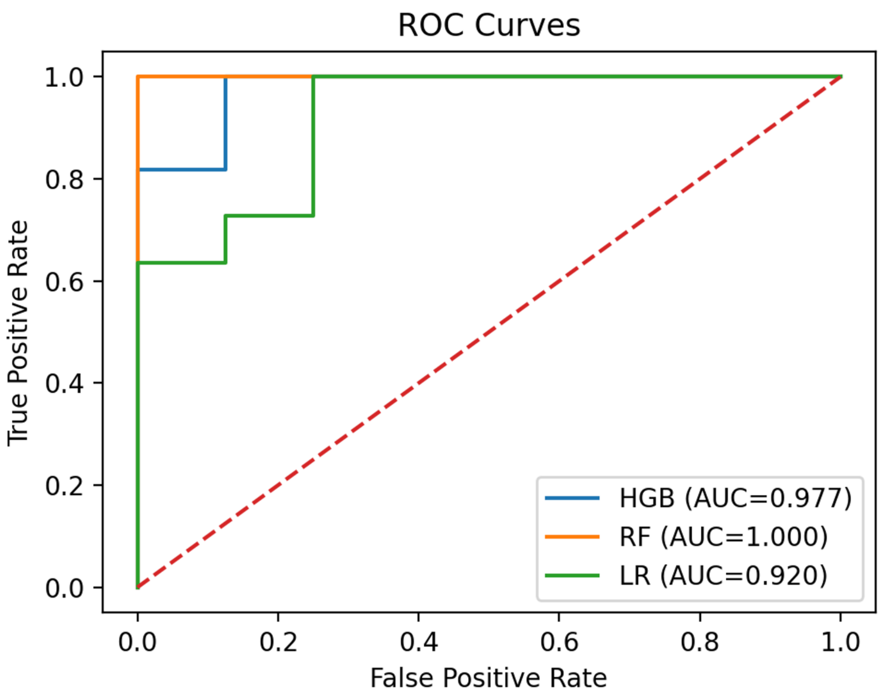

Fig. 3. ROC curves comparing HGB, Random Forest, and Logistic Regression.

Exposure to banking and healthcare chatbots shows mixed but interpretable contributions: trust increases when satisfaction is high, but decreases under poor

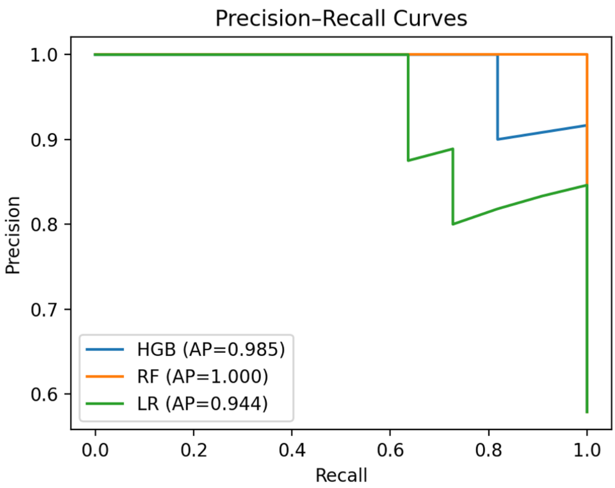

Fig. 4. Precision–Recall curves comparing HGB, Random Forest, and Logistic Regression.

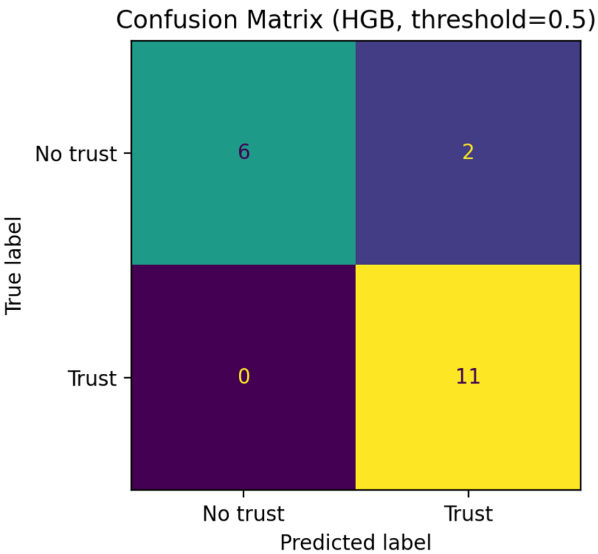

Fig. 5. Confusion matrix on the held-out set for HGB at threshold 0.5.

experiences. Dissatisfaction and infrequent usage reduce trust. Younger respondents (18–24) show a small positive association. These signed effects complement tree-based importances and support transparent, user-sensitive design.

5 Reproducibility and Artifacts

We release anonymized survey data and machine-readable outputs for all evaluations, including cross-validation summaries, held-out metrics, calibration points, ablation results, and LR coefficients. These artifacts enable full regeneration of the reported tables and figures (Figs. 2– 5), ensuring transparent and reproducible results.

6 Comparative Study with Related Work

We compare our contribution to representative strands of chatbot trust research. Prior work typically examines determinants of trust without developing deployable prediction pipelines, or builds predictive models without probability calibration and escalation policies. Kraus et al. predict live trust in mixed-initiative dialogue using SVMs, boosted trees, and GRU/RNN models, but do not perform probability calibration nor define a handoff rule [6]. Questionnaire-based studies identify factors such as expertise, responsiveness, and brand perceptions, yet do not construct supervised trust predictors [1,2]. Adjacent literature models user satisfaction or interaction quality rather than trust itself [9,10].

Our study integrates three elements not jointly addressed in prior work: (i) calibrated supervised trust prediction, (ii) an operational, risk-aware escalation threshold, and (iii) full release of artifacts for reproducibility.

Table 3. Comparison with related studies (Superv.=Supervised prediction).

Study	Trust Target	Superv.	Calibration	Handoff
Ours	✓	✓	✓	✓
[6]	✓	✓	✗	✗
[1]	✓	✗	✗	✗
[2]	✓	✗	✗	✗
[9]	✗	✓	✗	✗

7 Operational Outputs and Threshold Policy

Table 4 presents an anonymized decision log for five example cases. For each case, we report the calibrated trust probability, the automated decision at the threshold $P(\text{trust}) < 0.35$, and whether a human handoff is triggered. Low-probability cases (e.g., 0.014, 0.055) are escalated, illustrating the conservative safety-first design of the proposed policy.

Table 4. Comparison with related studies (Superv.=Supervised prediction).

Case	**P**(trust)	Automated Decision	Handoff	Notes
A1	0.014	Do not trust	1	High-risk sector
A2	0.055	Do not trust	1	–
A3	0.412	Trust	0	–
A4	0.687	Trust	0	–
A5	0.331	Do not trust	1	Low confidence

8 Limitations

Although the study provides a calibrated supervised pipeline for predicting chatbot trust and proposing a human-handoff policy, several limitations should be acknowledged to contextualize the results.

Scope and labeling of data: The dataset is modest in size (N = 122) and derived from convenience sampling, which limits representativeness across sectors, age groups, and user types. Trust is captured using a single binary label reflecting respondents' stated willingness to disclose personal information—an important but narrow operationalization. This measure does not encompass other recognized dimensions of trust such as perceived competence, reliability, or risk tolerance. Self-reported survey data may also suffer from recall bias, inconsistent interpretation of questions, or social–desirability tendencies, all of which can introduce noise into labels. Additionally, exposure to high-risk sectors such as banking and healthcare is uneven in the sample, creating imbalances that may affect the stability of feature estimates for those domains.

Feature coverage and study design: The study relies solely on cross-sectional survey variables such as demographics, usage frequency, satisfaction, and sector exposure. These features provide useful but coarse-grained signals. Missing from the analysis are conversational, behavioral, or temporal indicators that could more directly reflect trust formation. For example, turn-taking patterns, error recovery behavior, politeness cues, or longitudinal trust drift were not captured. As a result, the models are constrained to static attributes rather than dynamic interaction-level factors. Further, categorical features rely on one-hot and multi-hot encodings, while Likert-scale items are treated ordinally, which may compress subtle variations in user perceptions.

Modeling and evaluation: Although a limited grid search and repeated runs were performed to mitigate variance, the hyperparameter tuning space remained intentionally small to avoid overfitting on the modest dataset. Consequently, the chosen models may not reflect optimal configurations. With only 122 samples, stratified 5-fold cross-validation and a 20% hold-out test set can still yield unstable estimates, especially for metrics such as Precision, Recall, and PR-AUC, which are sensitive to class distribution. The absence of an external validation or out-of-distribution test set restricts generalizability, making it difficult to assess how well the models would perform on users from different regions, linguistic backgrounds, or sectors not well represented in the dataset. Model calibration was evaluated at the aggregate level but not examined across demographic or sector subgroups, leaving open the possibility of uneven calibration or subgroup-specific deviations.

9 Conclusion and Future Work

This study shows that users' willingness to trust customer-service chatbots with personal information can be predicted effectively using compact, survey-based features. Across three supervised models, consistent evidence indicates that satisfaction with prior interactions, usage familiarity, and the risk profile of the service sector are the strongest determinants of trust. By integrating probability calibration and a conservative, interpretable decision threshold, we demonstrate how supervised trust prediction can support human-in-the-loop escalation policies in privacy-sensitive domains. The release of all cleaned data and artifacts further strengthens transparency and reproducibility.

While the results are promising, they should be interpreted in light of the study's scope—particularly the modest sample size and reliance on self-reported cross-sectional features. These constraints motivate several practical extensions. First, future work will expand data collection across sectors, demographic groups, and organizational settings to improve external validity. Second, incorporating conversational and temporal signals may better capture how trust evolves during real interactions. Third, online calibration monitoring and evaluation of threshold sensitivity will help adapt the escalation policy to changing user populations or operational constraints. Finally, human-in-the-loop user studies and small-scale field deployments can validate how calibrated trust scores and escalation rules affect user experience, safety, and service quality in realistic environments.

Overall, this work provides an initial, reproducible foundation for calibrated trust prediction in chatbots and highlights a practical pathway for safer, sector-aware deployment.

References

1. Nordheim, C.B., Følstad, A., Bjørkli, C.A.: An initial model of trust in chatbots for customer service—findings from a questionnaire study. Interact. Comput. 31(3), 317–335 (2019). https://doi.org/10.1093/iwc/iwz022
2. Prakash, A.V., Joshi, A.P., Nim, S., Das, S.: Determinants and consequences of trust in AI-based customer service chatbots. Serv. Ind. J. (2023). https://doi.org/10.1080/02642069.2023.2166493
3. Min, F., Fang, Z., He, Y., Xuan, J.: Research on users' trust of chatbots driven by AI: an empirical analysis based on system factors and user characteristics. In: Proc. Int. Conf. Consumer Electronics and Computer Engineering (ICCECE) (2021) https://doi.org/10.1109/ICCECE51280.2021.9342098

4. Sonntag, M., Mehmann, J., Teuteberg, F.: Trust-supporting design elements as signals for AI-based chatbots in customer service. Int. J. Serv. Sci. Manag. Eng. Technol. (2023). https://doi.org/10.4018/IJSSMET.329963
5. Christopher, Y., Sundjaja, A.M., Mulvono: The role of AI chatbots on e-commerce platforms: understanding its influence on customer trust and dependability. In: Proc. IEEE Int. Conf. Information Technology, Data and Applications (ICITDA) (2024). https://doi.org/10.1109/ICITDA64560.2024.1080991
6. Kraus, M., Wagner, N., Minker, W.: Modelling and predicting trust for developing proactive dialogue strategies in mixed-initiative interaction. (2021) https://doi.org/10.1145/3462244.3479906
7. Przegalińska, A., Ciechanowski, L., Stroz, A., Gloor, P., Mazurek, G.: In bot we trust: a new methodology of chatbot performance measures. Bus. Horiz. 62(6), 785–797 (2019) https://doi.org/10.1016/j.bushor.2019.08.005
8. Ltifi, M.: Trust in the chatbot: a semi-human relationship. Future Bus. J. 9(1), 109 (2023) https://doi.org/10.1186/s43093-023-00288-z
9. Yao, R., Song, S., Li, Q., Wang, C., Chen, H., Chen, H., Zeng, D.D.: Session-level user satisfaction prediction for customer service chatbot in e-commerce (Student Abstract). In: Proc. AAAI Conf. Artif. Intell. 34(10) (2020). https://doi.org/10.1609/aaai.v34i10.7259
10. Gao, Y., Piovano, E., Soliman, T.M., Moniruzzaman, M., Kumar, A., Bradford, M., Nandi, S.N.: Predicting interaction quality of conversational assistants with spoken language understanding model confidences. In: Proc. ACM Int. Conf. Information and Knowledge Management (CIKM) (2023) https://doi.org/10.1145/3583780.3615493
11. Uprety, S.P., Jeong, S.R.: The impact of semi-supervised learning on the performance of intelligent chatbot system. Comput. Mater. Contin. 71(2) (2022) https://doi.org/10.32604/cmc.2022.023127
12. Almansor, E.H., Hussain, F.K., Hussain, O.K.: Supervised ensemble sentiment-based framework to measure chatbot quality of services. Computing 103(3), 491–507 (2021) https://www.researchgate.net/publication/345758767_Supervised_ensemble_sentiment-based_framework_to_measure_chatbot_quality_of_services
13. Toreini, E., Aitken, M., Coopamootoo, K.P.L., Elliott, K., Zelaya, C.G., van Moorsel, A.: The relationship between trust in AI and trustworthy machine learning technologies. arXiv preprint (2019) https://doi.org/10.48550/arXiv.1912.00782
14. Zafar, A., Parthasarathy, V.B., Van, C.L., Shahid, S.M., Khan, A.H., Shahid, A.: Building trust in conversational AI: a review and solution architecture using large language models and knowledge graphs. Big Data Cogn. Comput. 8(6), 70 (2024). https://doi.org/10.3390/bdcc8060070
15. Qi, J., Rougier, B.: Research on users' trust in customer service chatbots based on human–computer interaction. In: Human–Computer Interaction. Springer (2022). https://doi.org/10.1007/978-981-19-7532-5_19