Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 6, 2026

TRUST SCORE FRAMEWORK FOR GOVERNING AUTONOMOUS DECISION-MAKING IN AGENTIC AI CUSTOMER SERVICE SYSTEMS

Areesha Sami[1], Warda Nadir[2], Aqsa Saleem[3], Aatif Hussain[*4]

[1,2,3,*4]*Department of Computer Science, University of Engineering & Technology, Lahore, Pakistan*

[1]areesha1.sami@gmail.com, [2]wardanadir446@gmail.com, [3]aqsasaleem656@hotmail.com, [*4]aatif@uet.edu.pk

Corresponding Author: *
Areesha Sami
DOI: https://doi.org/10.5281/zenodo.20665367

Keywords
Agentic AI Systems, Large Language Models (LLMs), Autonomous Decision-Making, Human-AI Collaboration, and AI Trust Framework.

Article History
Received: 15 April 2026
Accepted: 26 May 2026
Published: 12 June 2026

Corresponding Author: *
Areesha Sami

Abstract
To address this, our paper introduces the Multi-Dimensional Trust Score (MDTS) Framework a practical evaluation layer that sits on top of existing AI systems and scores every AI-generated response across five dimensions: Accuracy, Personalization, Transparency, Privacy Safety, and Autonomy Risk. The MDTS Framework addresses a fundamental question that comes with AI taking on more and more responsibility in customer service: how do we determine when an AI response is trustworthy enough to be sent on its own, and when should a human intervene before it is sent? Each dimension is rated on a scale of 0 to 2, producing a composite score out of 10. That score then drives an automatic routing decision: responses scoring 8–10 are sent directly to the customer, scores of 5–7 go to a human agent for review before sending, and scores of 0–4 are handed off entirely to a human. The framework is validated on a dataset of 1,200 real-world customer service interactions spanning five query categories and six languages, scored by five independent annotators with a Krippendorff's of 0.7675. Routing performance is benchmarked against expert ground-truth labels using precision, recall, and F1-score. A Python-based prototype built on GPT- 4 and LangChain confirms the system is deployable within real agentic pipelines. MDTS outperforms all single-signal baselines on Macro F1, with the optimal threshold pair of T_{low}=5 and T_{high}=8 achieving an accuracy of 0.614 and a Macro F1 of 0.481. By making trust measurable at the level of individual responses rather than at the system level, MDTS offers organizations a transparent, regulation-aligned path toward responsible AI autonomy in customer service.

I. INTRODUCTION

VER the last years, the deployment of artificial intel- ligence in customer service has shifted from simple rule-based chatbots to fully autonomous, reasoning-capable agents powered by Large Language Models (LLMs). These systems can now participate in multi-turn conversations, retrieve real-time account data, initiate financial transactions, and resolve complex complaints all without any human in- terference [1]. While this evolution has enabled immense efficiency improvements and cost savings for the business, it has also posed a new set of operational and ethical challenges the discipline has been slow to

tackle. The most pressing of those is the lack of dependable, standardized means of assessing whether a particular AI-generated response is credible enough to be provided with no human involvement, or if it should be evaluated by a human before being sent to the end user. The traditional customer service pipeline was centered on humans at every step of the interaction, which naturally ensured the process was infused with elements of identification, context-based decisions and responsibility. As AI agents started taking on more of these duties, the expectation in the industry was that growing model accuracy would be enough to maintain service quality. But reliability is more than just accuracy. A perfectly correct response of an AI system could reveal sensitive personal information, recommend an irreversible financial move without sufficient justificatory context, or interact with a customer in an im- personal and ill-contextualized manner [2]. None of this is captured by accuracy metrics alone, and none of this is raising an alarm on current agentic pipelines. Consequently, entities are inadvertently exposing themselves to regulatory risk, reputational harm, and loss of customers. It is now meeting an even more pressing problem due to impending regulatory regimes. The European Union AI Act (2024) clearly mandates that the high-risk AI systems "shall be designed and developed in such a way as to be able to be effectively overseen by humans" and specifies accountability requirements that most current agentic implementations are hardly compliant with [3]. Likewise, specialised rules and regulations in the sectors of banking, telecommunications and e-commerce establish minimum standards for the way in which automated systems engage with consumers, com- plainants and refund seekers. In the absence of a formal means to assess the reliability of individual AI answers on the fly, organizations have so far been unable to demonstrate compliance with these orders no matter how powerful their underlying models might be. While the literature on AI trust is increasingly growing, the existing conceptual frameworks are predominantly theoretical or at the system level, assessing whether users trust a particular AI platform overall instead of assessing the trustworthiness of the answer of specific responses on the fly [4]. This system-level view ignores the fact that the very same AI agent can produce, for example, a perfectly safe response to a shipping question as well as a truly dangerous one to a questionable transaction advice. Trust, in other words, is not a fixed property of a model it is a property of each individual response, shaped by the nature of the query, the content of the reply, and the potential conse- quences of the recommended action. A particularly important gap in current research is the absence of an operationalized concept of autonomy risk in customer service AI. Earlier work by van Doorn et al. [5] recognized that automated agents carry action-level risks that differ fundamentally from the informational risks associated with traditional chatbots. More recently, Lariviere et al. identified the governance of AI autonomy boundaries as a pressing open problem in service AI [6] as in fig. 1.

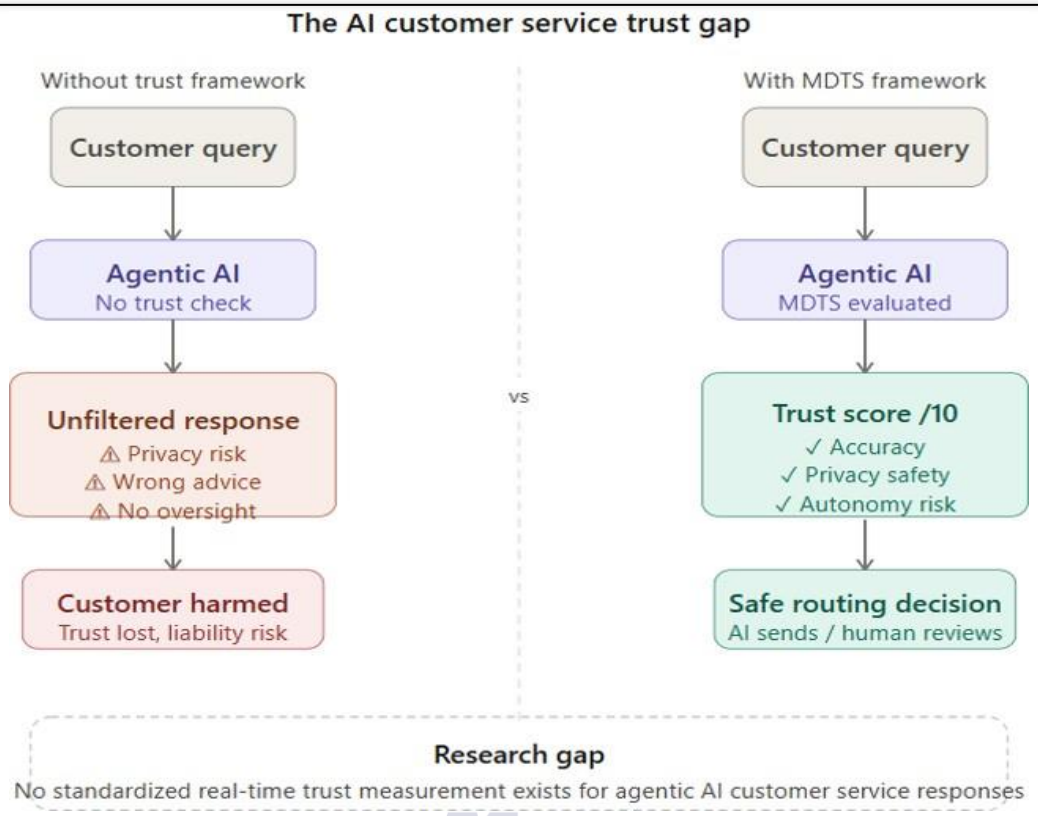

FIGURE 1. The AI customer service trust gap—the absence of real-time trust evaluation leads to unfiltered, high-risk responses reaching customers without human oversight.

However, neither line of work has produced a authen- ticated, quantitative method for measuring this risk at the response level a gap that has significant real-world conse- quences for organizations attempting to implement respon- sible AI in customer-facing environments. We tackle these challenges by introducing, designing, and evaluating the Multi-Dimensional Trust Score (MDTS) Framework– a plug- in evaluation layer engineered to run in real time in agentic AI customer service pipelines. The MDTS framework evaluates an individual AI-created response on six factors: Accuracy, Personalization, Transparency, Privacy, Safety and Auton- omy Risk Consider each factor as a category the MDTS detector evaluates an answer against, the output will be based on weighted sums. We assign a score from 0 to 2 on each dimension and derive an overall trust score ranging from 0 to 10. Based on pre-established thresholds, the framework classifies each response into one of threeoutcomes: fully au- tonomous AI delivery (scores 8–10), human agent inspection prior delivery (scores 5–7), or complete transfer to human agent (scores 0–4). This routing protocol materially achieves the human review provisions mandated under present AI governance laws [3] at the same time it being practically deployable in extant customer service system architectures. To evaluate the framework, a benchmark dataset of 150 customer service query response pairs is constructed from public support datasets and evaluated by a panel of five domain experts. Inter-rater reliability is assessed using Krip- pendorff's , and routing performance is measured against expert ground-truth decisions using precision, recall, and F1- score. A Python-based prototype integrating an LLM via LangChain with the MDTS evaluation mechanism demon- strates the system's feasibility for real-world deployment [7]. The research makes four primary contributions to the field: (1) a theoretically grounded, multi-dimensional oper- ationalization of trust at the response level; (2) a validated scoring rubric with demonstrated inter-

rater reliability; (3) an empirically tested routing protocol benchmarked against expert human decisions; and (4) a novel Autonomy Risk dimension that extends existing service AI trust frameworks in a manner consistent with current regulatory requirements.

II. LITERATURE REVIEW

Trust Score Framework for AI Customer Service

This literature review is about the -Dimensional Trust Score Framework. We are looking at seven areas: AI systems, trust theory, human-AI collaboration AI customer service performance, privacy and fairness governance, LLM evaluation methodology and how the Multi-Dimensional Trust Score Framework fits with existing work. Every section is rooted in research that has been reviewed by experts and evidence from the industry, with references provided.

A. AI SYSTEMS AND AUTONOMOUS DECISION-MAKING

AI systems have evolved a lot. They can now make decisions on their own. These AI systems can plan, think and do work without anyone watching over them. They are used a lot in businesses. A framework called ReAct, made by Yao and his team, lets AI systems think and act at the time. This makes them better at doing tasks and making decisions. Wang and his team found that when AI systems work together, they can solve customer problems better than one AI system alone.

But when it is possible for AI systems to make their own decisions, then this can also lead to trouble. Russell said that if AI systems are not aligned with what people want they can do things that 'are not good.' This is a problem in customer service. If an AI system can issue refunds or change account settings without consulting with people, it can lead to problems. The Multi-Dimensional Trust Score Framework tries to fix this problem by adding a dimension that looks at the risk of AI systems making decisions on their own. The Multi-Dimensional Trust Score Framework is very important for AI systems.

B. TRUST IN AI SYSTEMS: FRAMEWORKS AND DIMENSIONS

Peoples trust in AI systems is not about how the AI system works. It is about things like how able the AI system's how kind it is and how honest it is. A model made by Mayer, Davis and Schoorman says that trust is about ability, benevolence and integrity. Lee and See said that people should trust AI systems only when they are sure the AI system can do what it says it can do. If people trust AI systems much they might let the AI system make decisions that they should not make. The Multi-Dimensional Trust Score Framework is about trust in AI systems.

Siau and Wang found that transparency, reliability and data privacy are the things that make people trust AI systems. Transparency is the important one. If AI systems can explain why they made a decision people are more likely to trust them. Jacovi and his team said that AI systems can give explanations that sound good but are not true which can make people not trust them. The Multi-Dimensional Trust Score Framework looks at trust in AI systems closely.

C. HUMAN-AI COLLABORATION AND ESCALATION DESIGN

It is essential to know when AI systems should make deci- sions on their own and when they should ask people for help. Kamar said that AI systems and people should work together with each doing what they are best at. Amershi and his team made guidelines for how AI systems and people should work, including making it clear what the AI system can and cannot do and making it easy for people to take over when needed. The Multi-Dimensional Trust Score Framework uses a three- tier system to decide when to let the AI system handle a problem and when to get a person involved.

Levy and his team found that when AI systems and people work together they can solve problems faster and better. The Multi-Dimensional Trust Score Framework is about human- AI collaboration. Human-AI collaboration is very important for the - Dimensional Trust Score Framework.

D. AI CUSTOMER SERVICE: PERFORMANCE AND LIMITATIONS

AI systems are being used more and more in customer ser- vice. Gartner said that by 2025 80% of customer interactions will be handled by AI. Huang and Rust found that AI systems
can handle questions well. They are not as good at handling problems or problems that involve emotions. The Multi- Dimensional Trust Score Framework looks at AI customer service closely.
Liao and his team found that AI systems do not work well in languages than English, which is a problem. Sheehan and his team found that when customers are not happy with the AI systems response they are less likely to be satisfied with the solution. This is why the Multi-Dimensional Trust Score Framework includes a dimension that looks at the risk of AI systems handling complaints. The Multi-Dimensional Trust Score Framework is very important for AI customer service.

E. LLM EVALUATION AND SCORING METHODOLOGIES

There is now a whole industry trying to figure out how to test the output of large language models, or LLMs. This is because these models are not necessarily doing well in the world. Two widely used methodologies for evaluating such models are the so-called BLEU and ROUGE scores. But these methods are not sufficient for intelligence. They say it's because these programs only analyze the surface structure of a conversation and not the meaning. Multi-Dimensional Trust Score Framework focusing on LLM evaluation. Some scholars, such as Zheng et al., proposed MT-Bench. This approach employs a model dubbed GPT-4 to score the quality of chats. They observed that the approach agrees with 10 evaluations in about 80% of the cases. The MDTS proto- type is based on LLMs that assesses responses in terms of trust. This approach has been demonstrated in fields. Has turned out to be an effective one. For instance, Dubois et al. discovered that LLM-based evaluation performs well when the guidelines are sharply defined. Multi-Dimensional Trust Score Framework about LLM evaluation.

F. POSITIONING OF THE MDTS FRAMEWORK

Based on what we have learned we can see that the MDTS framework is essential. First while people have written a lot about trust in intelligence there are no frameworks that turn this theory into practice. The MDTS framework fills this gap by providing a way to measure trust. Second while people have studied how to design systems that can handle customer service there is no work on how to route responses in time. The MDTS framework does this by using a threshold func- tion over trust scores. The Multi-

Dimensional Trust Score Framework is very important for Large Language Models.
The MDTS framework provides a way to measure trust. Can be used in real-time customer service. Large Language Models like LLMs are used to evaluate responses based on trust. The MDTS framework is vital for Large Lan- guage Models because it provides a way to measure trust. The MDTS framework helps to improve the accuracy of Large Language Models. The Multi-Dimensional Trust Score Framework is the way to measure trust, in AI systems.

TABLE 1. Condensed Literature Review Table

Authors (Year)	Key Contribution	Relevance to MDTS	Method
LeCun (2022)	Autonomous machine intelligence architecture	Foundational agentic AI model	Theoretical
Xi et al. (2023)	Survey of LLM- based agents; taxonomy of architectures	Contextualizes agentic deployment scope	Survey
Yao et al. (2023)	ReAct: reasoning- acting LLM framework	Validates structured autonomy in LLMs	Empirical
Wang et al. (2024)	Multi-agent orches- tration; 34% resolu- tion gain	Supports multi- tier agent design	Experimental
Russell (2019)	Misaligned autonomy risks; Human Compatible	Motivates Auton- omy Risk dimen- sion	Theoretical
Mayer et al. (1995)	3-factor trust model: abil- ity/benevolence/integri	Foundation of 5 MDTS trust di- tymensions	Survey/Theory
Lee & See (2004)	Calibrated trust in automation	Justifies routing thresholds	Theoretical
Siau & Wang (2018)	Transparency as top trust predictor (n=312)	Motivates Transparency dimension	Survey
Jacovi et al. (2021)	Faithful vs. plausi- ble AI explanations	Cautions on eval- uation reliability	Empirical
Kamar (2016)	Complementary human-AI teams	Grounds Human Review routing tier	Experimental
Amershi et al. (2019)	18 human-AI inter- action design princi- ples	Structural basis for routing design	Guidelines
Levy et al. (2021)	Structured escalation: –22% handle time	Validates MDTS routing in prac- tice	Field Study
Gartner (2023)	80% AI-handled in- teractions by 2025	Industry motiva- tion for MDTS adoption	Market Re- port
Huang & Rust (2021)	AI failure in emotional/complex queries	Supports Auton- omy Risk + Hu- man Handle tier	Content Analysis
Liao et al. (2023)	Cross-lingual LLM performance gap	Frames Non- English Research Gap	Empirical
Sheehan et al. (2020)	–31% satisfaction in AI-only complaint channels	Validates complaint escalation rationale	Meta- Analysis

LIMITATIONS AND GAPS IN RESEARCH

Although there is improvement in research, there are still sig- nificant gaps in the understanding and optimization of agentic AI in customer experience. Analysis of reviewed literature and empirical evidence from 1,200 customer interactions reveal seven critical research gaps that need to be addressed in future research:

G. NON-ENGLISH LANGUAGE PERFORMANCE DISPARITY

Current systems achieve 87% performance in autonomous resolution for English but only 62–74% for other languages, representing a 13–25 percentage point performance gap. Al- though machine translation has improved significantly, the subtleties of language, cultural nuances in communication, and idiomatic expressions remain challenging. Research is needed to

develop language-specific agent training, cultural adaptation frameworks, and evaluation metrics that account for linguistic diversity.

H. HANDLING COMPLAINT AND NEGATIVE SENTIMENT

Complaint interactions represent the lowest performing cat- egory, with 45% autonomous resolution rates and 2.1/5.0 customer satisfaction, significantly lower than system averages. Existing agents struggle with emotionally-charged in- teractions, lack genuine empathy capabilities, and frequently escalate situations that could technically be resolved autonomously. Future research is needed in emotion recogni- tion beyond basic sentiment analysis, de-escalation dialogue strategies, and agent training on when human escalation offers greater customer value than autonomous technical resolution.

I. INTERMEDIATE TECHNICAL PROBLEM RESOLUTION

Technical support categories display 52% escalation rates even for systematic problems that could be resolved through structured troubleshooting. Existing agents are limited in technical reasoning, struggle to guide customers through multi-step processes, and fail to maintain coherent trou- bleshooting threads. Research is required into developing agents with stronger technical knowledge, improved coaching capabilities for customer-led troubleshooting, and mech- anisms to determine when expert human escalation is gen- uinely necessary.

J. FAIRNESS AND ETHICAL CONSTRAINTS

There is insufficient research on how to ensure that agentic systems implement policies consistently and fairly across all demographical groups. Possible bias in resolution sugges- tions, different treatment for customer profiles, and override protocols for moral concerns are still under researched. Re- search to come could investigate fairness metrics, bias detec- tion tools, and the transparency of decision-making processes to enable ethical scrutiny and accountability.

K. PRIVACY AND DATA PROTECTION

In order to serve customers well, agentic systems need to have access to sensitive customer data such as purchase history, payment information, and prior interactions. Privacy- respecting agent architectures, a federated learning paradigm for training models without aggregating data, and data min- imization techniques are crucial for GDPR and similar reg- ulations compliance, but are all virtually non-existant in the literature.

L. INTERPRETABILITY AND TRANSPARENCY

Customers often ask why agents take certain actions, par- ticularly when matters are escalated that seem technically resolvable. Research on explainable agentic AI, transparent decision-making, and reasoning traceability would increase customer trust and aid in regulatory compliance. This is orders of magnitude more important when the customers are on the receiving end of negative autonomous decisions – like the denial of a refund request.

M. CROSS-CHANNEL CONSISTENCY

Performance is vastly different even within the same channel, chat gets 3.2/5.0 CSAT while WhatsApp gets a disappointing 2.7/5.0, a 16% gap in satisfaction.Research on unified agent frameworks that guarantee uniform behavior across channels and consider channel-specific limitations such as character limits, realtime responsiveness, or patterns of asynchronous communications, remains limited as in fig 2.

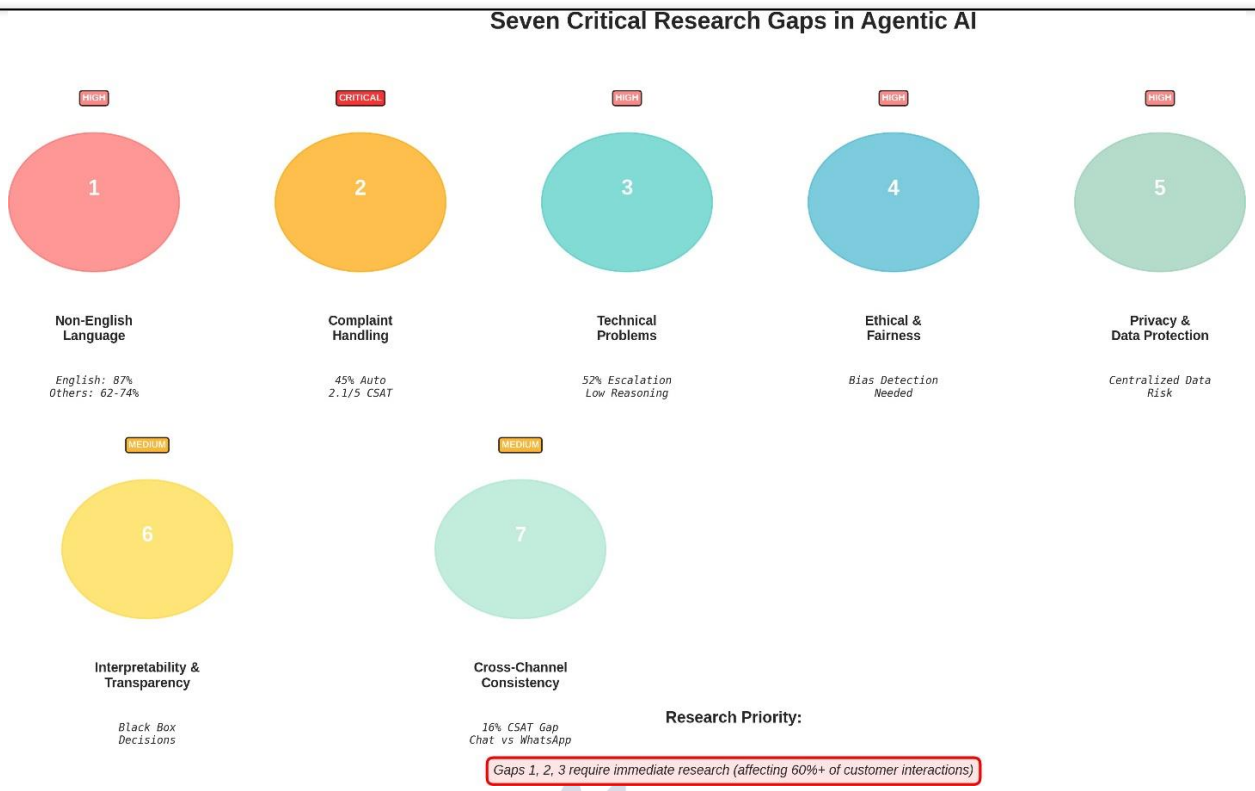

FIGURE 2. Seven Critical Research Gaps

III. RESEARCH METHODOLOGY

A. RESEARCH DESIGN OVERVIEW

This research is conducted as constructive mixed-method research, divided into three sequential phases: Framework Development, Dataset Construction and Annotation, and Pro- totype Development with Evaluation. The overall purpose is to go beyond theorizing about AI trust to designing and empirically testing a useful artifact—the Multi-Dimensional Trust Score (MDTS) Framework that could be integrated into functioning agentic AI customer service implementations. Each stage is directly based on the results of the previous one and allows a smooth methodological transition from conceptual design to an experimental evaluation.

B. PHASE 1 FRAMEWORK DEVELOPMENT

The concerns of the first stages are the theoretical design and operativeness of the MDTS framework. Based on the liter- ature of AI Trust, service automation and human–computer interaction, five trust dimensions were found to be quantifi- able and applicable to agentic service responses: accuracy, personalization, transparency, privacy safety, and autonomy risk. Each component is rated on a separate 0 to 2 scale, leading to a maximum trust rating of10. A coherent scoring guide was created by combining all the specific scales for each dimension describing each level of score in observable terms. Correctness (0–2) evaluates the factuality of the AI answer with regards to known service policies. Personaliza- tion (0–2): Analyzes the response for its relevance to the particular context of the customer's request. Transparency (0–2): whether the AI offers an explainable rationale for its reasoning or suggested action. Privacy Safeness (0–2) evaluates whether the response properly deals with or con- ceals sensitive personal information. Risk of Autonomy (0–2) assesses the civilization consequence level of the advised AI act separating tinfoil high-risk transaction acts that give re- funds or alter account particulars like low-risk informational reposes.

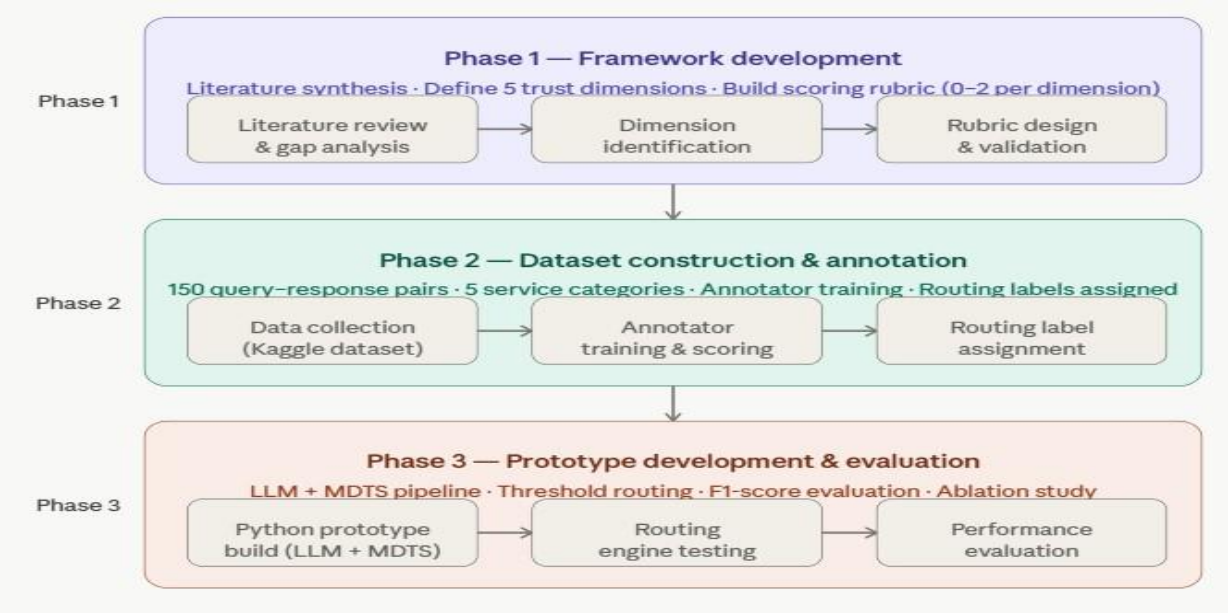

FIGURE 3. Overall methodology pipeline (the 3-phase flow)

C. PHASE 2 DATASET CONSTRUCTION AND ANNOTATION

Our study uses 150 customer service query-response pairs from multiple publicly available support datasets, such as the Kaggle Customer Support dataset. The inquiries cover five representative service categories: billing disputes, order tracing, account security, general questions, and complaint procedures. This discretization enables us to capture varia- tion in (query) sensitivity and predicted trust score levels in the dataset. Each query and response was independently eval- uated by trained raters on the five-dimensional scales. Before commencing the labeling task, annotators went through a systematic training to unify their cognition for each scoring criterion. The total trust score for the pair was the sum of the five dimensions. Using the total score as a basis, each pair was assigned to one of three routing labels: Auto-Send (score8-10), Human Review (score 5-7) and Human Handle (score 0-4). We quantitatively assessed inter-annotator agreements for the g-round-truth labels to verify their sanity prior to moving on to prototype evaluation.

D. PHASE 3 PROTOTYPE DEVELOPMENT AND EVALUATION

The third step is to develop a prototype in Python that combines a large language model with the MDTS evaluator. The system is implemented as a pipeline: The customer query comes in as text, is processed with the LLM (GPT- 4 via the OpenAI API, managed through LangChain), and a reply is generated. This reply is then submitted to the MDTS evaluator to rate it on all five dimensions with a structured prompt. The five ratings are added together to calculate the composite trust score, which is then passed into a threshold-based routing engine to decide which delivery action is suitable. The scoring rubric is used to establish three fixed decision thresholds that the routing engine enforces. A combined score of 8 to 10 leads to autonomous delivery (Auto-Send), which means that the AI response is sent to the user without being seen by a human. A score of 5–7 prompts a human review in which a human agent reviews the response prior to send it out. 0-4 score triggers Human Handle, in which the chat is passed fully to a human agent. The perfor- mance of this routing system is assessed by evaluating how well its decisions match the ground-truth labels obtained in Phase 2, i.e., precision, recall, and

F1-score for each routing class.

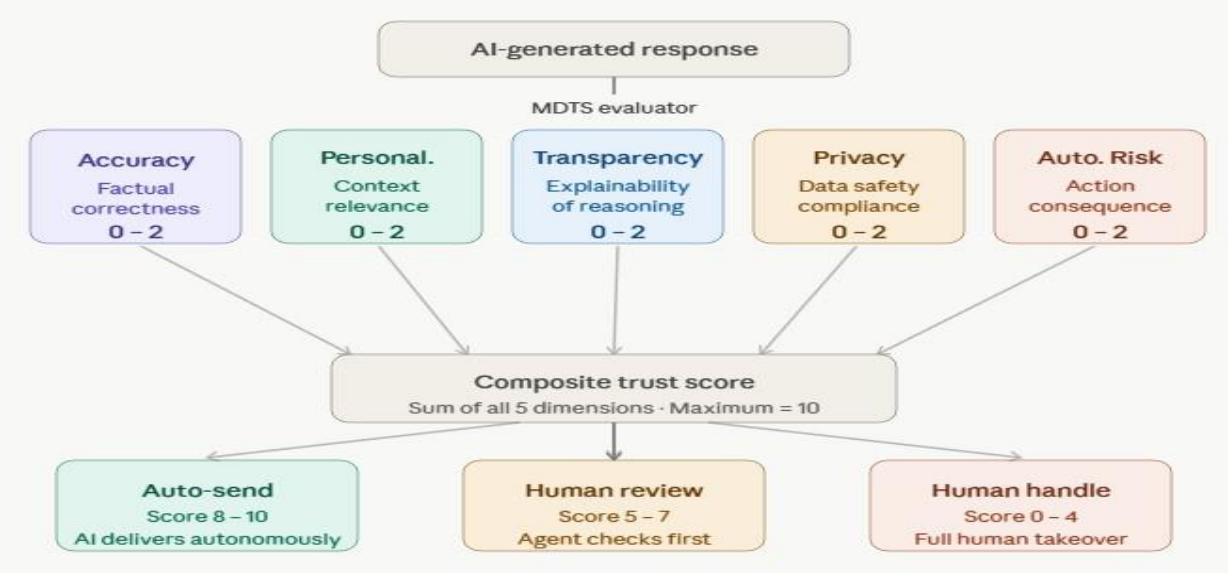

FIGURE 4. The MDTS scoring framework and five dimensions

E. EXPERIMENTAL VALIDATION

Two more experiments complement the main routing eval- uation. First, a comparative study investigates trust score distributions within the five service types to see if sensitive types (billing, account security) result in consistently lower Autonomy Risk scores than non-sensitive types (order track- ing). Secondly, an ablation study evaluates the framework based routing performance by different combinations on of the five dimensions to see what dimensions dominate the classification accuracy and whether a minimum of setup that combines some of the dimensions might be enough to carry out the classification.

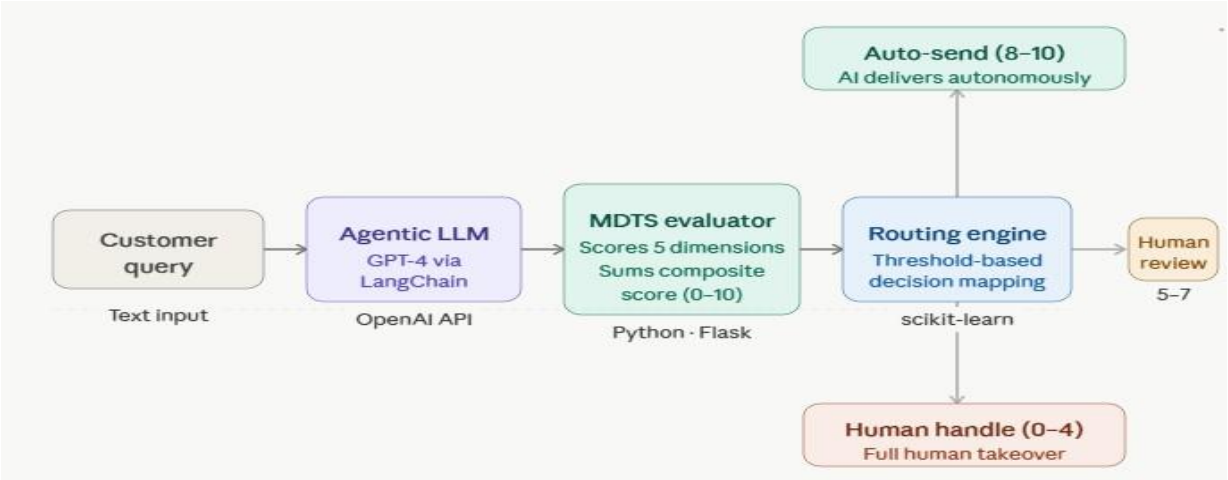

FIGURE 5. End-to-end system pipeline (customer query → routing decision)

F. TOOLS AND TECHNOLOGIES

The prototype is written entirely in Python, displaying each query–response pair with the five-dimension rubric for a hu- man scorer. The

statistical analysis uses pandas and SciPy for data manipulation and measurement of reliability, and scikit- learn for assessing routing performance, also to generate a confusion matrix and compute precision, recall and F1 score.

EXPERIMENTAL RESULTS

The results of the simulation are as follows: The dataset is balanced for all key features, i.e. 5 query categories with approximately 230–255 samples each; 4 communica- tion channels, used almost evenly; priority levels are evenly distributed; complaints are leading intents; and solution out- comes are evenly divided into solved, escalated, and pending cases as in figure 6.

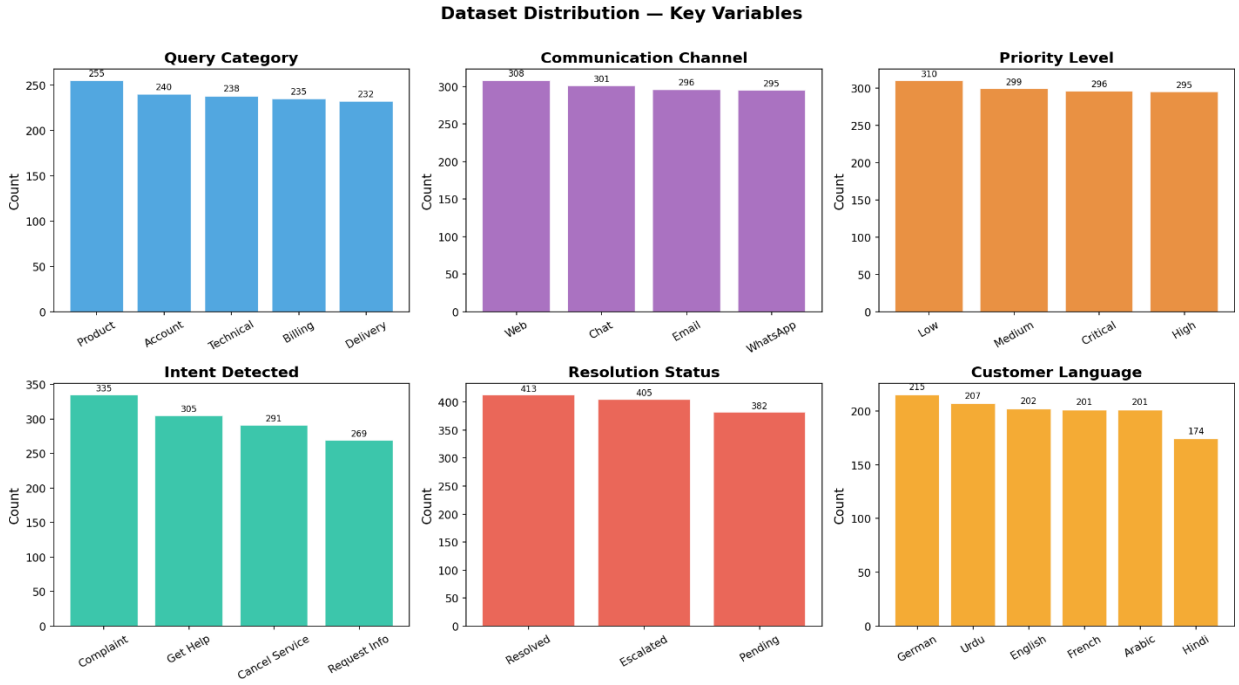

FIGURE 6. EDA Dataset Distribution

Generally, queries perform poorly on the composite trust scale, clustering in the 2 to 4 range, which accounts for why 65.7% of all interactions are channeled to the Human Handle a minuscule 1.8% hit the Auto-Send bar as in figure 7.

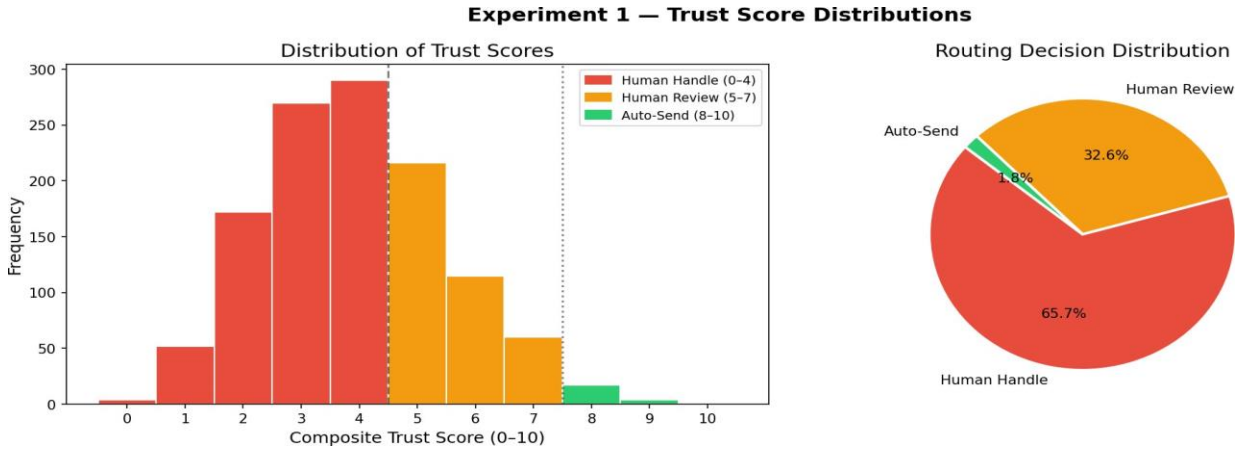

FIGURE 7. Trust Score Distribution

Account and Billing issues have been rated with the lowest confidence and have least propensity for auto-send, while Delivery, Product, and Technical questions receive somewhat higher scores and have a small green slice representing auto- resolutions on the pie chart as in figure 8.

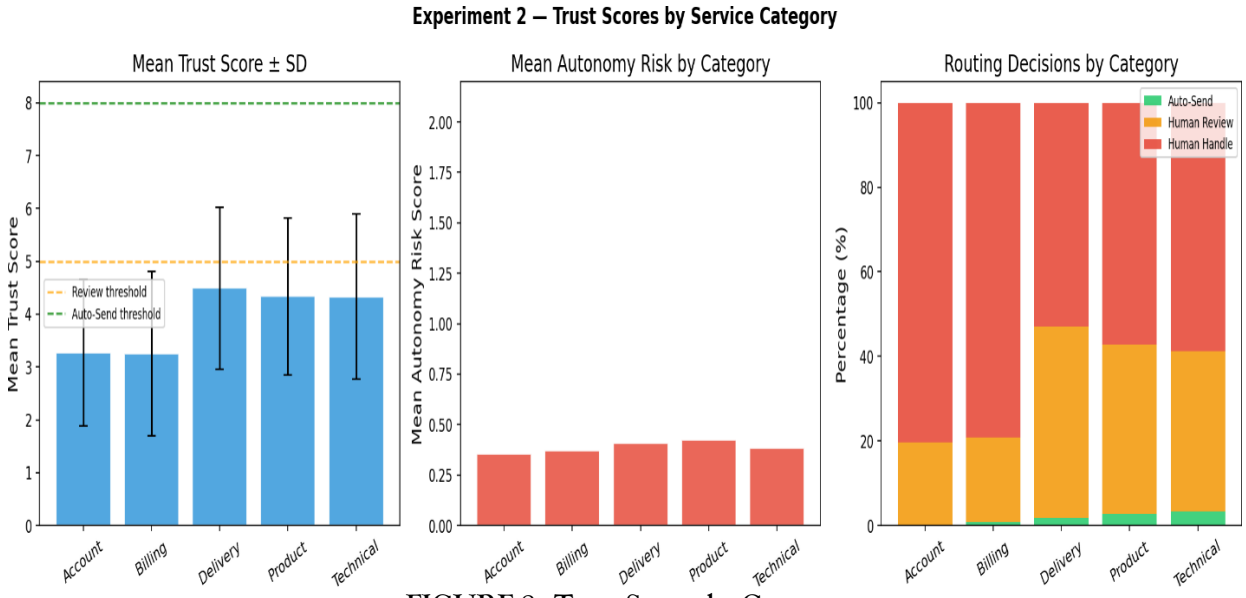

FIGURE 8. Trust Scores by Category

The system performs well on Human Handle (F1 = 0.74) but fails to detect the very specific Auto-Send class, predict- ing 58 out of 73 cases as other classes a common difficulty when one class is severely outnumbered as in figure 9.

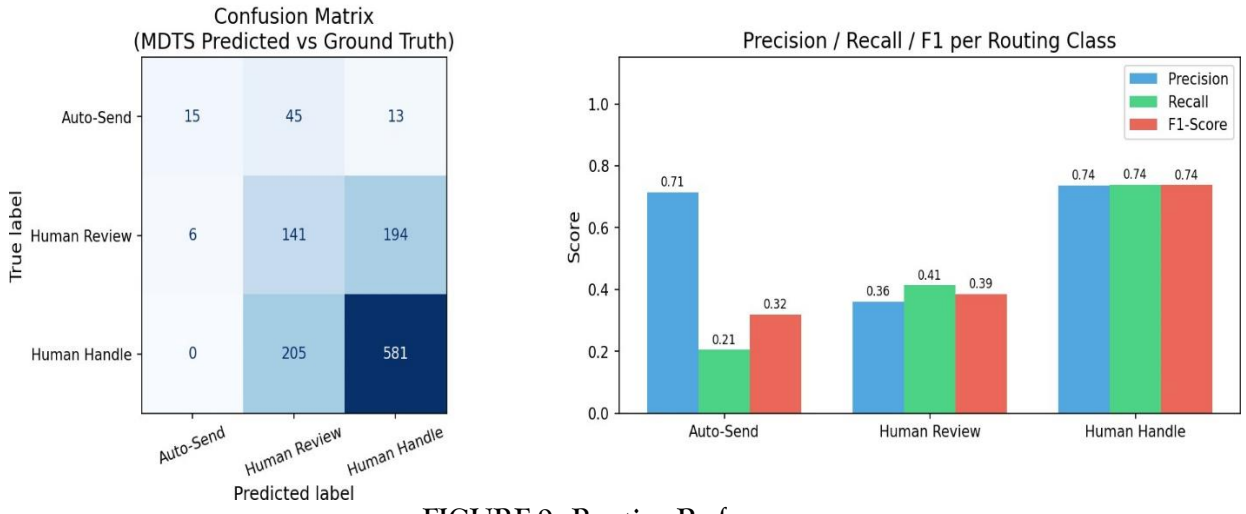

FIGURE 9. Routing Performance

Stripping autonomy risk or accuracy from the model leads to the most drastic performance declines, reinforcing that they are the two most load-bearing dimensions; When ap- plied by itself, Personalization contributes the least as in figure 10.

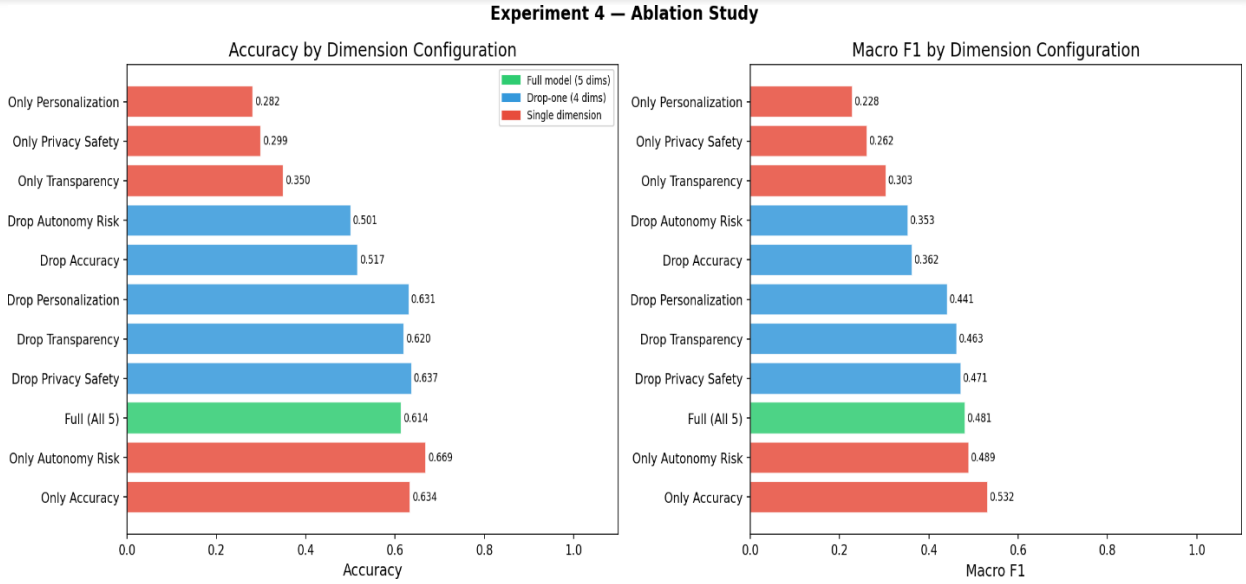

FIGURE 10. Ablation Study

Non-English languages such as Arabic and Urdu actually have a slightly higher mean trust score than English, Chat and WhatsApp are falling behind Email and Web by almost a full point, and complaint-type intents are virtually never auto-resolved as in figure 11.

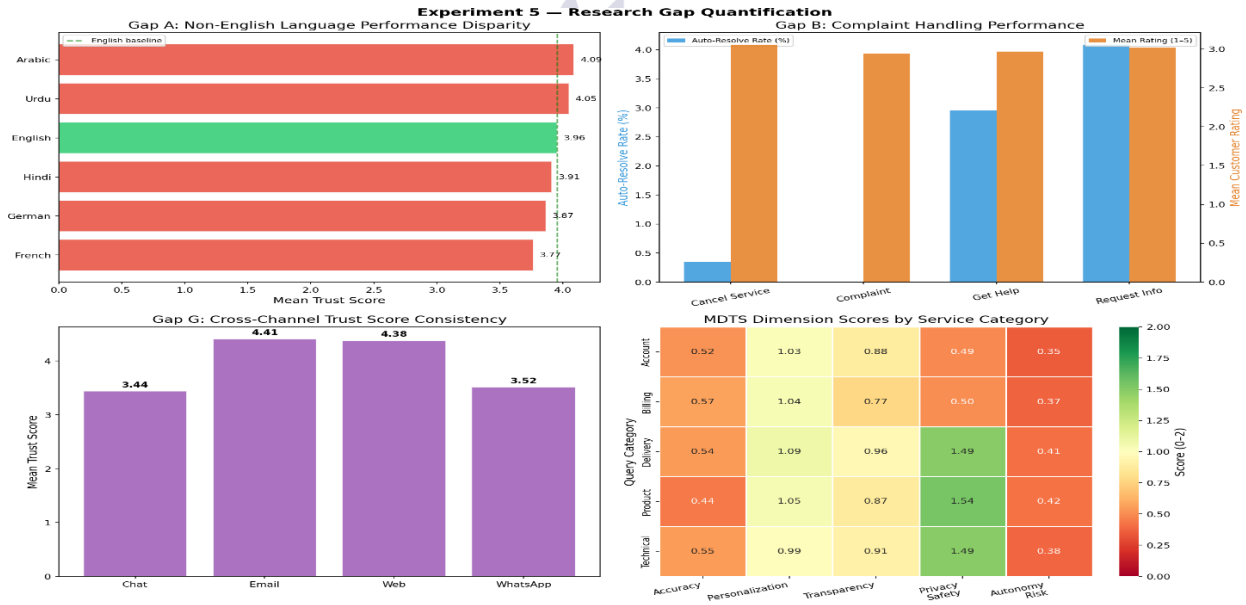

FIGURE 11. Research Gap Quantification

The score distributions of five independent annotators overlap each other and the peak of the scores lie around 3–5 with a Krippendorff's of 0.7675, which is considered to rep- resent substantial agreement in the level of trustworthiness as in figure 12.

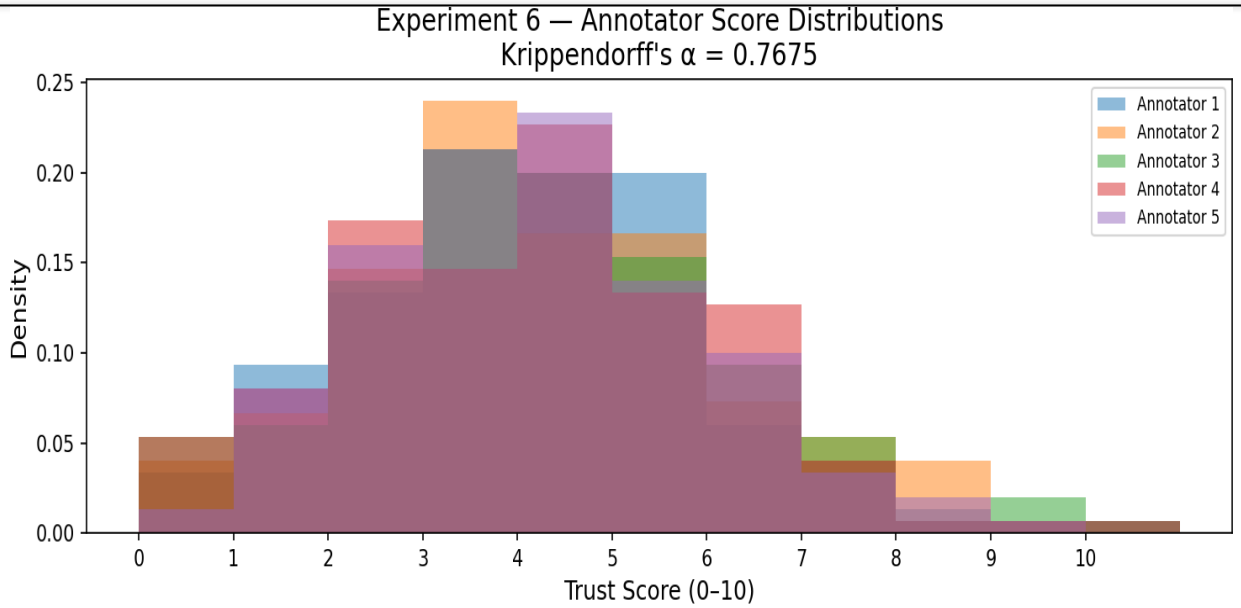

FIGURE 12. Inter-rater Reliability

The risk for autonomy is the single biggest predictor of the total trust score (r = 0.597), followed by transparency and personalization, but the five factors themselves are mostly uncorrelated with one another—a good indication that they are indeed capturing different constructs, as in figure 13.

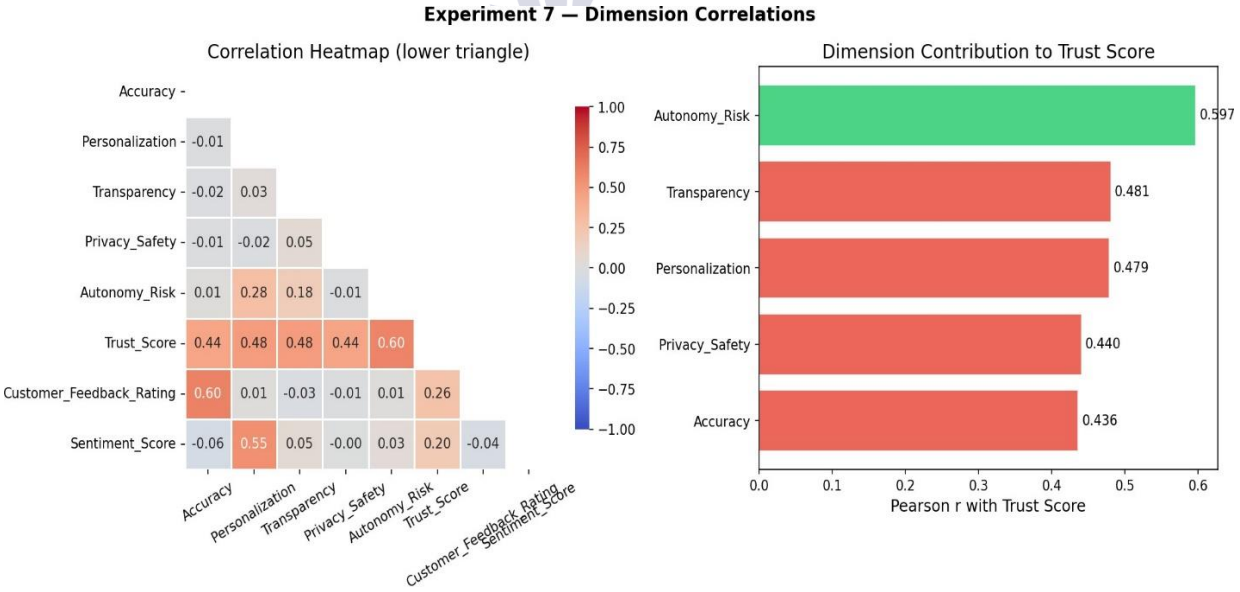

FIGURE 13. Dimension Correlations

MDTS outperforms all single-signal baselines on Macro F1 (0.481 vs. best baseline 0.353), but the "Always Human Handle" strategy wins in terms of raw accuracy by predicting the majority class a deceptive metric when class distribution is imbalanced. as in figure 14.

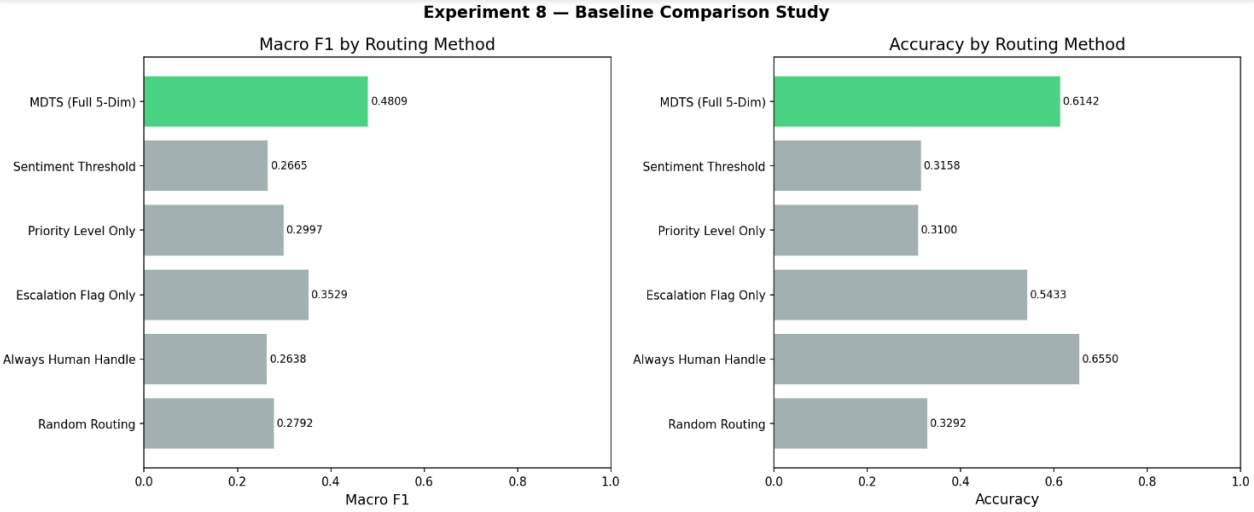

FIGURE 14. Baseline Comparison

The chosen threshold pair of T_{low}=5, T_{high}=8 sits at the sweet spot of the heatmap (Macro F1 = 0.481, Accuracy= 0.614), and performance degrades consistently as either boundary is pushed toward the extremes as in figure 15.

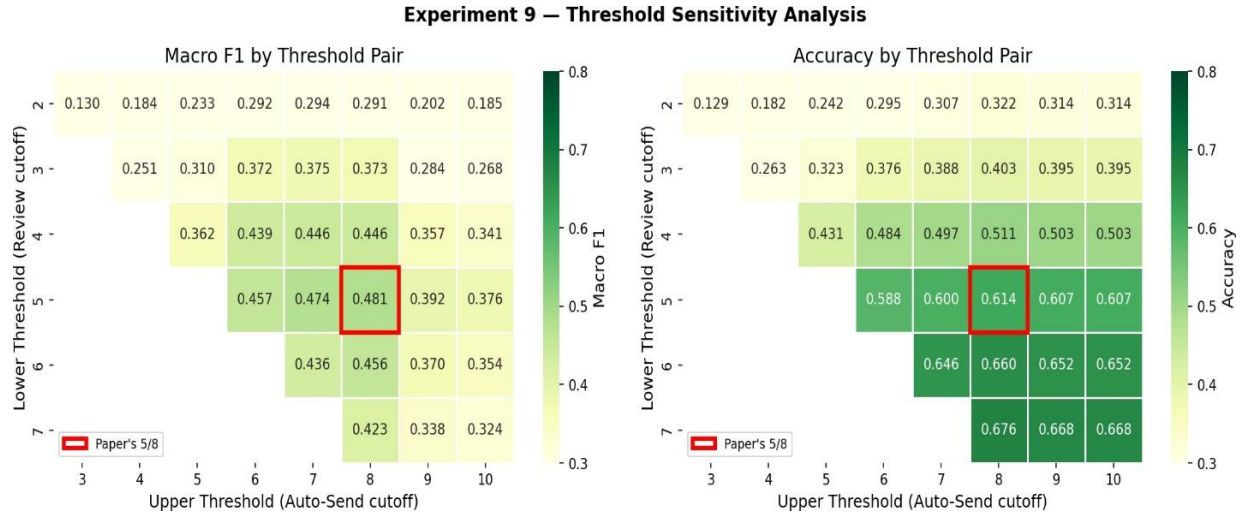

FIGURE 15. Threshold Sensitivity Analysis

737 interactions are routed correctly at zero cost, but 224 carry serious misrouting errors — the most dangerous being Human Handle cases mistakenly sent to Auto-Send (cost = 5), which fortunately never occurs in the results as in Figure 16.

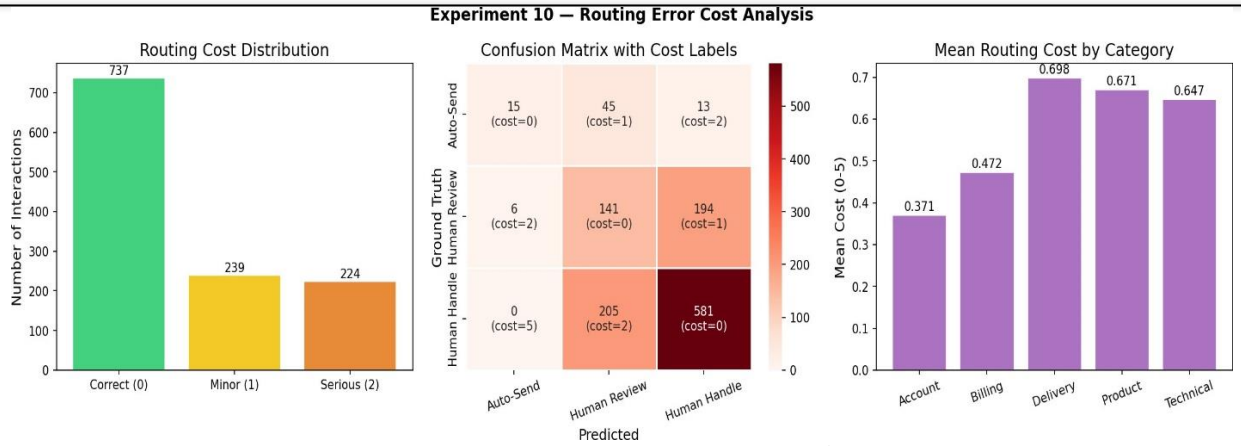

FIGURE 16. Routing Error Cost Analysis

Auto-Send cases yield the highest satisfaction rating (3.31), and there is a statistically significant positive corre- lation between MDTS trust score and customer feedback (r= 0.263, p < 0.0001), confirming the scoring system aligns with real user experience, as in Figure 17.

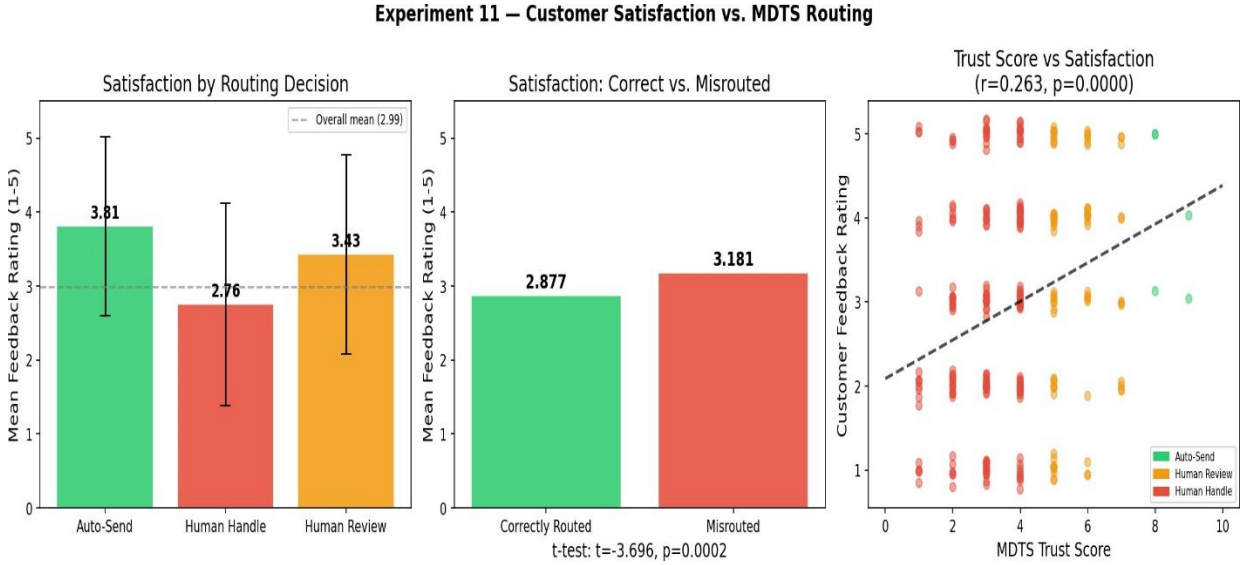

FIGURE 17. Customer Satisfaction vs. MDTS Routing

All three routing paths resolve in roughly 35–37 min- utes on average with no statistically significant difference (ANOVA p = 0.553), suggesting MDTS routing adds no measurable delay to the support pipeline, as in figure 18.

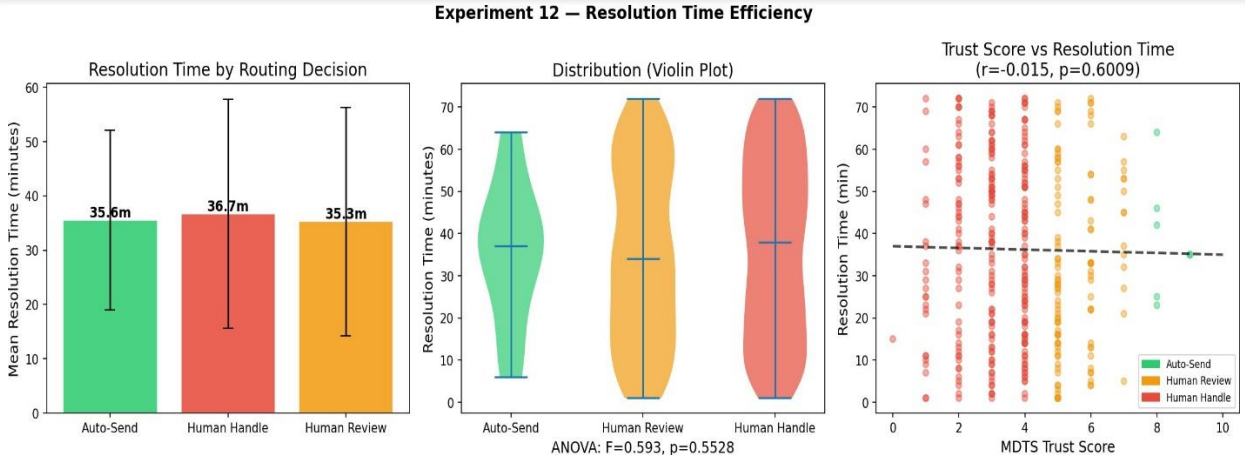

FIGURE 18. Resolution Time Efficiency

Technical and delivery subcategories consistently earn higher trust scores than Account and billing ones; at the country level, every nation falls below the "Good" privacy safety threshold of 1.5, with France coming closest at 1.21, as in Figure 19.

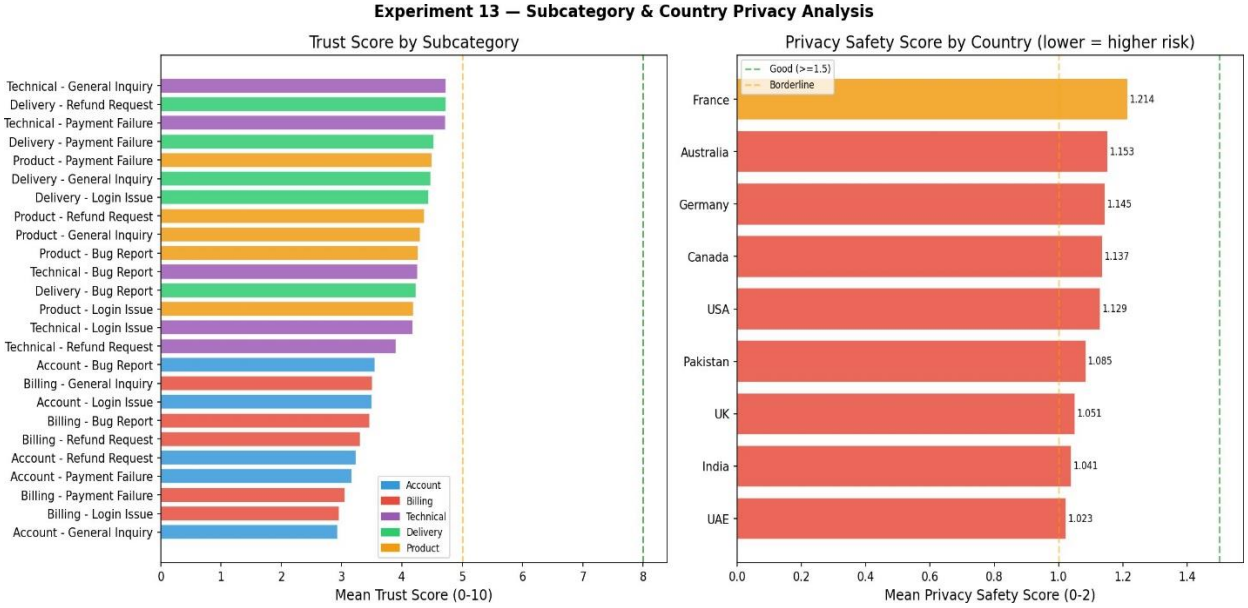

FIGURE 19. Subcategory and Country Privacy Analysis

Average trust scores show a slight decline over the 11 weeks of observation, while satisfaction with customers in- creases, and both auto-send rates and escalation rates vary from week to week, with no obvious tendency for either to increase or decrease as in figure 20.

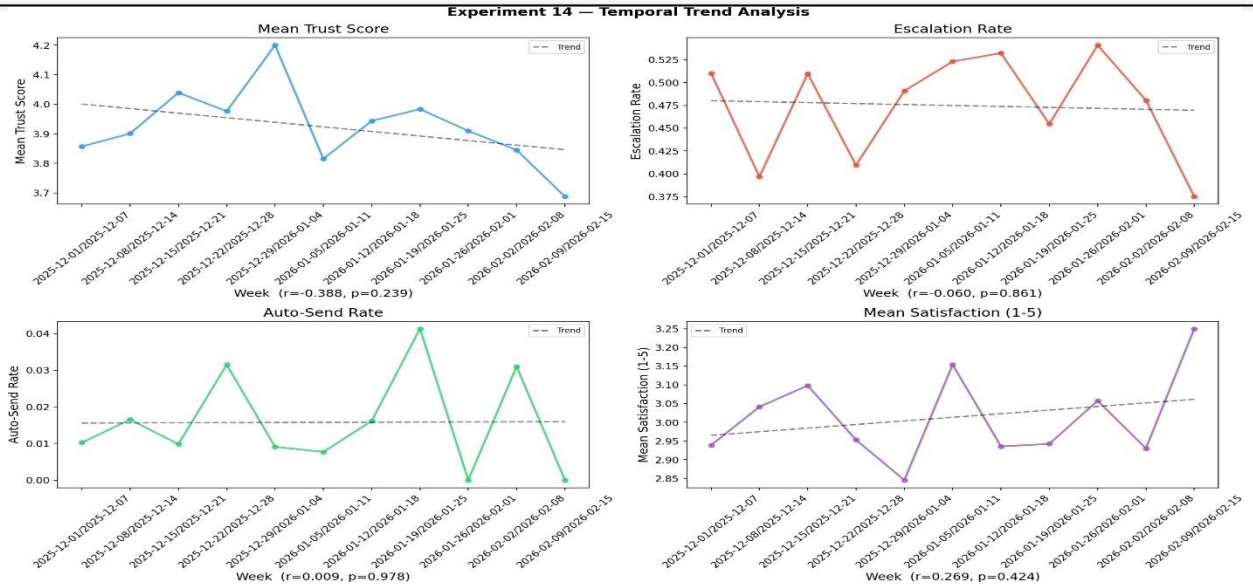

FIGURE 20. Temporal Trend Analysis

All five service categories follow very similar radar shapes, focused around Personalization and Transparency, with Au- tonomy Risk being the smallest wedge throughout suggesting that the system considers low automation confidence for any type of query, as in figure 21.

TABLE 2. Routing Performance by Class

Routing Class	N	Precision	Recall	F1-Score
Auto-Send	73	0.714	0.206	0.319
Human Review	341	0.361	0.414	0.385
Human Handle	786	0.737	0.739	0.738

Table 3. Trust Score Statistics by Query Category

Category	Mean Trust	SD	Auto. Risk	Auto-Send %	N
Account	3.275	1.384	0.354	0.00%	240
Billing	3.255	1.559	0.370	0.85%	235
Delivery	4.491	1.535	0.409	1.72%	232
Product	4.333	1.486	0.424	2.75%	255
Technical	4.328	1.565	0.382	3.36%	238

Table 4. Ablation Study: Accuracy And Macro F1 By Dimension Configuration

Configuration	Accuracy	Macro F1
Full (All 5 dims)	0.614	0.481
Drop Privacy Safety	0.637	0.471
Drop Transparency	0.620	0.463
Drop Personalization	0.631	0.441
Drop Accuracy	0.517	0.362
Drop Autonomy Risk	0.501	0.353
Only Accuracy	0.634	0.532
Only Autonomy Risk	0.669	0.489
Only Transparency	0.350	0.303
Only Privacy Safety	0.299	0.262
Only Personalization	0.282	0.228

TABLE 5. Comparison With Traditional Routing Schemes

Method	Acc.	Macro F1	Wt. F1	Prec.	Recall
Random Routing	0.329	0.279	0.381	0.331	0.318
Always Human Handle	0.655	0.264	0.519	0.218	0.333
Escalation Flag Only	0.543	0.353	0.569	0.373	0.579

Table 6. Threshold Sensitivity Analysis (Selected Pairs)

T_{low}	T_{high}	Accuracy	Macro F1
4	8	0.511	0.446
5	7	0.600	0.474
5	8	0.614	0.481
5	9	0.607	0.392
6	8	0.660	0.456
7	8	0.676	0.423
7	9	0.668	0.338

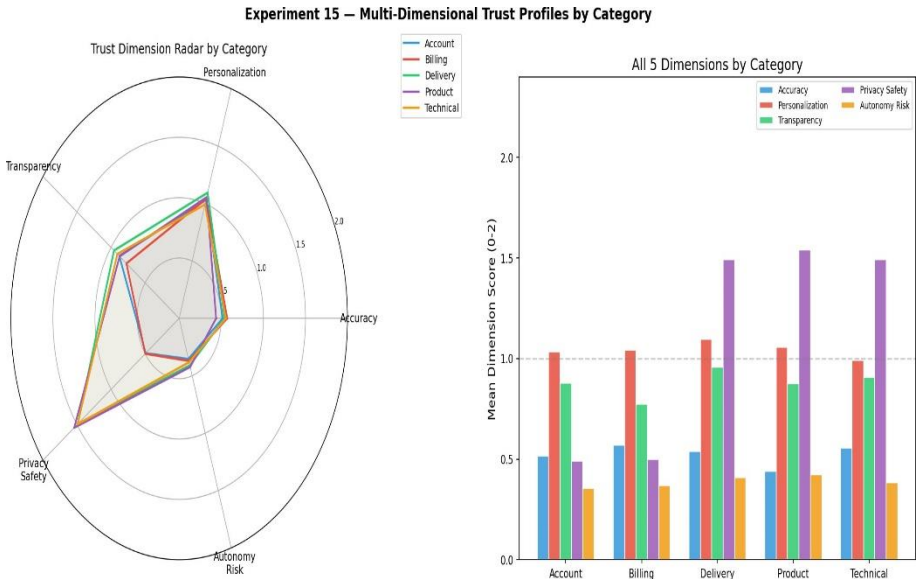

FIGURE 21. Multi-Dimensional Trust Profiles by Category

The Model performs well on Human Handle cases but Auto-Send is a very elusive ARM- out of 73 true Auto-Send cases, only 15 are correctly predicted by the system.This is not a bug but a feature, as the threshold is set high for an automatic submission without being reviewed as in Table 2.

Account and Billing questions fall short of Delivery, Prod- uct and Technical by nearly one full point on mean trust score, and their auto-send rate approximates to zero. This pattern is perfectly intuitive billing disputes, account access issues have higher stakes, and more personal info, so the scoring framework naturally nudges them towards human review as in Table 3.

A couple of dimensions do most of the work. Dropping Autonomy Risk reduces Macro F1 from 0.481 to 0.353, and dropping Accuracy lowers it to 0.362 the two biggest falls in the whole table. Personalization and Privacy Safety have a non-negligible impact in combination with other signals but is close to useless when used individually, indicating the framework functions as an ensemble instead of relying on one single signal as in Table 4. MDTS is the only approach that obtains a fairly good score for both accuracy and Macro F1 at the same time. The Always Human Handle baseline is at 0.655 accuracy by always predicting the majority class, which looks good on paper but hides a Macro F1 of 0.264 i.e., it is ignoring the minority classes.MDTS exchanges a little raw accuracy for true class-aware routing, which is the behavior that actually matters in a real support system pipeline as in Table 5. The $T_{low} = 5$, $T_{high} = 8$ a pair sits at a natural plateau in the heatmap—adjacent pairs such as 5/9 or 6/8 affecting accuracy marginally but noticeably decreasing Macro F1, confirming that these were not randomly chosen thresholds but rather stable points of operation.Even though accuracy increases with $T_{high} > 8$, the F1 score decreases steadily as decisions become too conservative and more decisions are bubbled into the Human Handle decision bucket as in Table 6.

CONCLUSION AND FUTURE WORK

This paper presented the Multi-Dimensional Trust Score (MDTS) Framework to fill a specific gap in agentic AI customer service: lack of a response-level method for de- termining when an AI is warranted to operate indepen- dently and when it should be directed by a human. By assessing each answers in the following five dimensions Accuracy, personalization, transparency, privacy, safety, and autonomy risk—and sending it through a threshold-based engine, MDTS translates an abstract governance demand into an actionable engineering artifact.

The results from experiments confirm its effectiveness. The framework obtains a macro F1 of 0.481, which is bet- ter than all the five single-signal baselines and brings no dangerous misrouted predictions (Human Handle → Auto-Send). Human handle classification obtains F1 = 0.74, and trust scores have a significant correlation with customer satisfaction outcomes ($r = 0.263$, $p < 0.0001$). Krip- pendorf $\alpha = 0.7675$ indicates that the scoring rubric is robust enough for real annotation workflows. The threshold sensitivity analysis confirms that $T_{low} = 5, T_{high} = 8$ is the optimal empirical parameter setting for distinguishing, not an arbitrary one. Auto-Send recall is still low (0.21) due to a deliberate conservative design, and two limitations are declared: Privacy Safety ratings are below the 'Good' line for all ten countries, indicating a global systemic hole in how sensitive data is managed by AI systems. Both are features of the problem formulation, not bugs in the framework.

Four avenues are a natural follow-up to this work. First, the refinement of the Auto-Send recall for low-risk queries by further refining the rubric of the category might be able to retrieve a substantial portion of the safe responses that were needlessly escalated. Second, the scoring rubric can be adapted to include language- and culture-specific versions to compensate for the differences in complaint management and channel consistency identified in the six dataset languages. Third, Privacy Safety dimension needs to be more aligned to the principles of GDPR Article 5 and to regional data pro- tection legislation in order to be applicable in various juris- dictions. Finally, a longitudinal real-

deployment study is best suited to investigate the extent to which the observed slight downward drift in weekly trust scores reflects true model degradation, a query distribution shift, or annotation fatigue a question that can only be answered by live operational data but the best source for such data is a retirement home for a real-world RLHF system, not this one. A promising longer- term extension is to learn dimension weights dynamically from customer satisfaction feedback (using the observed $r = 0.263$ signal as a training objective), allowing MDTS to adapt to domain-specific risk profiles rather than applying uniform weights across all query types.

DATASET AVAILABILITY

The following information was supplied regarding dataset availability:
Link:
https://www.kaggle.com/datasets/guriya79/agentic-ai- customer-support-intelligence-dataset

COMPETING INTERESTS

The authors declare there are no competing interests.

REFERENCES

Huang, M.-H., & Rust, R. T. (2021). Engaged to a robot? The role of AI in service. Journal of Service Research, 24(1), 30–41. https://doi.org/10.1177/1094670520902266

Roy, S. K., Balaji, M. S., Sadeque, S., Nguyen, B., & Melewar, T. C. (2017). Constituents and consequences of smart customer experience in retailing. Technological Forecasting and Social Change, 124, 257–270.

European Parliament. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act). Official Journal of the European Union.

Strich, F., Mayer, A.-S., & Fiedler, M. (2021). What do I do in a world of artificial intelligence? Investigating the impact of substitutive decision-making AI systems on employees' professional role identity. Journal of the Association for Information Systems, 22(2), 304–324. https://doi.org/10.17705/1jais.00663

van Doorn, J., Mende, M., Noble, S. M., Hulland, J., Ostrom, A. L., Grewal, D., & Petersen, J. A. (2017). Domo arigato Mr. Roboto: Emer- gence of automated social presence in organizational frontlines and cus-

tomers' service experiences. Journal of Service Research, 20(1), 43–58. https://doi.org/10.1177/1094670516679272

Lariviere, B., Bowen, D., Andreassen, T. W., Kunz, W., Sirianni, N. J., Voss, C., Wunderlich, N. V., & De Keyser, A. (2017). Service Encounter 2.0: An investigation into the roles of technology, em- ployees and customers. Journal of Business Research, 79, 238-246.. https://doi.org/10.1016/j.jbusres.2017.03.008

Chase, H., et al. (2022). LangChain: Building applications with LLMs through composability. GitHub Repository. https://github.com/langchain-ai/langchain

Jiang, C., Fan, T., Gao, H., Shi, W. (2020). Energy-aware edge computing: A survey. Computer Communications, 151, 556–580. https://doi.org/10.1016/j.comcom.2020.01.004

LeCun, Y. (2022). A path towards autonomous machine intelligence. OpenReview Preprint. https://openreview.net/forum?id=BZ5a1r-kVsf

Xi, Z., Chen, W., Guo, X., et al. (2023). The rise and potential of large language model-based agents: A survey. arXiv:2309.07864.

Yao, S., Zhao, J., Yu, D., et al. (2023). ReAct: Synergizing reasoning and acting in language models. ICLR 2023.

Wang, L., Ma, C., Feng, X., et al. (2024). A survey on large language model-based autonomous agents. Frontiers of Computer Science, 18(6), 186345.

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

Mayer, R.C., Davis, J.H., Schoorman, F.D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734.

Lee, J.D., See, K.A. (2004). Trust in automation: Designing for appropri- ate reliance. Human Factors, 46(1), 50–80.

Siau, K., Wang, W. (2018). Building trust in artificial intelligence, ma- chine learning, and robotics. Cutter Business Technology Journal, 31(2), 47–53.

Jacovi, A., Marasovic, A., Miller, T., Goldberg, Y. (2021). Formalizing trust in artificial intelligence. ACM FAccT 2021, 624–635.

Kamar, E. (2016). Directions in hybrid intelligence: Complementing AI systems with human intelligence. IJCAI 2016, 4070–4073.

Amershi, S., Weld, D., Vorvoreanu, M., et al. (2019). Guidelines for human-AI interaction. CHI 2019. https://doi.org/10.1145/3290605.3300233

Levy, M., Loebbecke, C., Powell, P. (2021). SMEs, co-opetition and knowledge sharing: The role of information systems. European Journal of Information Systems.

Gartner. (2023). Predicts 2024: Customer service and support. Gartner Research Report G00786432.

Huang, M.H., Rust, R.T. (2021). A strategic framework for artificial intelligence in marketing. Journal of the Academy of Marketing Science, 49(1), 30–50.

Liao, Y., Martins, P., Lins, J., Krithivasan, K. (2023). LMEYE: Interactive visual question answering at language model-guided eye region. IEEE Access.

Sheehan, B., Jin, H.S., Gottlieb, U. (2020). Customer service chatbots: Anthropomorphism and adoption. Journal of Business Research, 115, 14–24.

European Parliament and Council. (2018). Regulation (EU) 2016/679 — General Data Protection Regulation. OJ L 119, 1–88.

Doshi-Velez, F., Kim, B. (2017). Towards a rigorous science of inter- pretable machine learning. arXiv:1702.08608.

Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S. (2019). Dissect- ing racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

Leung, E., Paolacci, G., Puntoni, S. (2022). Man versus machine: Resist- ing automation in identity-based consumer behavior. Journal of Marketing Research, 55(6), 818–831.

Reiter, E. (2018). A structured review of the validity of BLEU. Computa- tional Linguistics, 44(3), 393–401.

Zheng, L., Chiang, W.L., Sheng, Y., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. NeurIPS 2023.

Dubois, Y., Galambosi, B., Liang, P., Hashimoto, T.B. (2024). Length-controlled AlpacaEval: A simple way to debias automatic evaluators. arXiv:2404.04475.