



Evaluating LLM Dialogue Quality: A Multi-Dimensional Assessment of AI-Generated Insurance Customer Service Responses

Kittisak Jaima, and Karn Yongsiriwit*

Cybersecurity and Technology Management Program, College of Digital Innovation Technology, Rangsit University,
Pathum Thani 12000, Thailand

*Corresponding author; Email: karn.y@rsu.ac.th

Abstract

Large Language Models (LLMs) have become essential tools for insurance customer service. However, evaluating the quality of LLM-generated dialogue remains challenging because single-metric assessments fail to capture domain-specific performance nuances. This research addresses this gap by proposing a comprehensive multi-dimensional evaluation framework for assessing LLM-generated responses in insurance dialogue systems. The methodology analyzes 26,232 dialogue pairs from the DSTC12 Written Insurance corpus across four distinct evaluation dimensions: semantic similarity (Sentence-BERT cosine scoring), domain correctness (Named Entity Recognition with 2,711 insurance-specific terms), linguistic fluency (GPT-2 perplexity normalization), and conversational appropriateness (ToxicBERT toxicity classification). Comparative analysis reveals that Ollama-generated responses demonstrate superior semantic alignment (0.302 vs. 0.288), substantially higher domain correctness (0.268 vs. 0.094, representing an 185% advantage), and exceptional linguistic fluency (0.825 vs. 0.161, representing a 412% advantage), while human-authored responses maintain marginally better conversational appropriateness (0.050 vs. 0.094 toxicity). The framework successfully integrates automated metrics with domain-specific assessment criteria, providing actionable insights for deploying and optimizing LLM-based customer service systems in regulated industries. Results support selective deployment of Ollama in insurance contexts while identifying conversational appropriateness as a priority refinement area.

Keywords: *Large Language Model (LLM), dialogue quality evaluation, Named Entity Recognition (NER), cosine similarity, ToxicBERT, insurance customer service*

1. Introduction

Large Language Models (LLMs) have revolutionized customer service interactions by leveraging advanced Natural Language Processing capabilities to deliver sophisticated conversational experiences. In the insurance sector, the application of LLM-based chatbots has become increasingly important as customers require efficient access to information about complex insurance products, policies, and claims. The development of these conversational AI systems requires rigorous evaluation methodologies to ensure response quality across multiple dimensions, including semantic accuracy, domain knowledge, linguistic quality, and professional appropriateness.

Traditional evaluation metrics such as BLEU and ROUGE rely on lexical overlap and fail to capture semantic nuances critical for domain-specific applications. To address this limitation, recent research has explored more sophisticated approaches including sentence embeddings, entity recognition, and perplexity-based fluency assessment. Cosine similarity techniques, combined with Named Entity Recognition, provide effective measurement of semantic alignment and domain correctness in technical and specialized contexts (Soyusiawaty & Zakaria, 2018; Wall, 2021). In the insurance domain specifically, accurate identification of domain-specific entities and terminology is essential for ensuring that automated responses address customer concerns using appropriate industry-standard language.

Beyond semantic and domain considerations, insurance customer service requires responses that maintain professional communication standards and avoid inappropriate language. The integration of toxicity detection mechanisms into evaluation frameworks ensures that automated systems operate within appropriate behavioral boundaries (Singh et al., 2023). Additionally, linguistic fluency assessment through perplexity

[343]



scoring provides quantitative measures of response naturalness and grammatical quality, distinguishing between conversational utterances and formally - generated text.

Large Language Model chatbots are increasingly central to customer service because advanced Natural Language Processing enables systems to interpret user intent, preserve conversational context, and incorporate sentiment analysis to generate empathetic, privacy-aware responses that support natural and efficient information access (Gohil et al., 2025). These capabilities have driven the adoption of AI assistants that provide intuitive interaction and streamlined information retrieval for users in diverse service environments, including high-volume and information-intensive sectors Gohil et al. (2025). Within the insurance domain specifically, conversational agents and humanoid AI systems have been explored as tools for improving service quality, customer satisfaction, and operational efficiency, particularly for handling repetitive inquiries and supporting complex decision processes in policy management and claims handling (Patil et al., 2024).

Text similarity measurement underpins many of these applications by enabling automated systems to match user queries with appropriate responses across large repositories of documents and examples. Cosine similarity, often computed over TF-IDF vectors or dense sentence embeddings, has been widely applied to tasks such as human resource management, where it supports resume–job matching and workforce analytics, demonstrating the effectiveness of similarity-based models for structured, decision-oriented domains (Singh & Garg, 2024). Beyond traditional text, similarity techniques have also been used to identify semantic correspondences between binary executables through source-level transformation, showing that embedding-based and structural similarity methods can capture deep relationships in highly technical representations (Wan et al., 2025). In digital libraries and information retrieval systems, leveraging similarity metrics improves the precision of content recommendation and document ranking, further highlighting their importance in large-scale knowledge access scenarios (Soyusiawaty & Zakaria, 2018). Collectively, these works show that cosine and related similarity measures provide a strong foundation for evaluating the alignment between user intent and system outputs, which is crucial when assessing dialogue systems in specialized fields such as insurance (Wall, 2021).

Named Entity Recognition is another core technology for domain-specific dialogue and question answering systems, as it enables accurate extraction of entities like persons, organizations, locations, and domain terms, thereby improving retrieval and response generation quality (Jati et al., 2020). To address the shortage of labeled data in specialized domains, migration learning and data expansion techniques have been proposed, allowing models to transfer knowledge from related domains to enhance NER performance in target tasks with limited annotations (Li et al., 2024). In parallel, keyword generation and extraction approaches in advertising illustrate how combining NER with curated domain lexicons can increase relevance in content targeting and campaign optimization, a strategy that can be adapted to ensure the presence and correctness of critical insurance-related concepts in generated answers (Bulut & Mahmoud, 2023).

Recent work also highlights growing concern about bias and inconsistency in Large Language Model outputs, motivating comparative studies of widely used systems such as ChatGPT 4, Perplexity, Google Gemini, and Claude under standardized prompts (Choudhary, 2025). These analyses reveal that different models may produce divergent or biased responses on politically and socially sensitive topics, underscoring the necessity of evaluation frameworks that examine not only semantic correctness but also neutrality, fairness, and robustness in domain deployments. In parallel, advances in machine learning for the insurance industry have transformed core processes such as sentiment monitoring, risk assessment, and fraud detection, enabling more proactive and data-driven decision making (Kailash Varma et al., 2024). Sentiment analysis models help insurers gauge customer satisfaction and adjust services, while predictive models for health claims support more accurate estimation of claim likelihood and cost, thereby optimizing resource allocation and operational planning (Ghosh, 2023). Furthermore, hybrid AI and ML architectures that combine supervised learning, unsupervised detection, and rule-based reasoning have achieved strong results in identifying complex and evolving fraud patterns, significantly reducing false positives and protecting insurers from financial loss (Suresh et al., 2025).



Ensuring safe and professional communication is particularly important for automated dialogue systems in regulated domains. BERT-based toxicity detection models have shown strong capability in identifying subtle, context-dependent harmful language, making them suitable components for moderating both user-generated content and system outputs in interactive applications (Saddi et al., 2023). When integrated with similarity, NER, and fluency metrics, toxicity detection enables a more comprehensive evaluation of dialogue quality, where semantic relevance, domain correctness, linguistic naturalness, and behavioral safety are jointly considered to support trustworthy LLM-based customer service in insurance.

Previous dialogue evaluation research has largely treated quality dimensions in isolation, focusing separately on similarity, toxicity, or entity correctness rather than their interactions in specialized domains such as insurance customer service. This fragmentation limits the ability of existing methods to capture the combined requirements of semantic accuracy, domain fidelity, fluent expression, and consistently professional tone that are essential in regulated environments. To address these shortcomings, this study introduces a unified, multi-dimensional evaluation framework that jointly assesses semantic similarity, domain correctness, linguistic fluency, and conversational appropriateness, offering a more comprehensive and practically actionable assessment of LLM-generated responses in insurance customer support scenarios.

The convergence of these evaluation dimensions reflects a broader recognition that LLM quality cannot be meaningfully assessed through single metrics. Unlike prior work that typically evaluates isolated aspects of dialogue quality, this research proposes a unified multi-dimensional evaluation framework integrating semantic similarity, domain correctness, linguistic fluency, and toxicity detection. Applied at scale to the DSTC12 Written Insurance corpus with 26,232 dialogue pairs and a curated 2,711-term insurance lexicon, this framework directly compares LLM-generated responses against human-authored ones, providing empirical insights into LLM performance and actionable guidance for deploying automated systems in insurance customer service.

2. Objectives

The objectives of this research are explicitly specified to ensure measurable outcomes through systematic evaluation of LLM-generated dialogue quality in insurance contexts.

1) To Identify and quantify multi-dimensional performance characteristics. This objective measures LLM performance across four distinct quality dimensions using automated, quantifiable metrics:

- a. Semantic Similarity
- b. Domain Correctness
- c. Linguistic Fluency
- d. Conversational Appropriateness

2) To compare LLM performance against human-authored baseline standards.

3) To provide actionable insights for improving automated insurance customer service Systems.

3. Materials and Methods

This section presents a comprehensive multi-dimensional evaluation framework for assessing Large Language Model (LLM)-generated responses in conversational insurance dialogue systems, as illustrated in Figure 1. The methodology leverages the DSTC12 Written Insurance corpus to compare LLM-generated responses against human-authored reference responses across four distinct evaluation dimensions: semantic similarity, domain correctness, linguistic fluency, and conversational appropriateness.

3.1 Dataset and Response Generation

The evaluation utilizes the DSTC12 Written Insurance dataset, comprising 26,232 dialogue act pairs extracted from customer-agent conversations in the insurance domain. This corpus was specifically designed for controllable conversational theme detection and provides rich contextual information across various insurance-related scenarios, including policy inquiries, claims processing, and coverage questions. Each dialogue pair consists of a prompt utterance and a corresponding response, enabling systematic comparison between human-authored and LLM-generated outputs.



As shown in the workflow diagram (Figure 1), for each dialogue prompt in the dataset, two response types are processed:

Reference Response: The original human-authored response preserved from the dataset, representing expert-level domain knowledge and an appropriate conversational style.

Generated Response: An LLM-generated response produced using Ollama, conditioned on the same dialogue prompt. The model generates responses without access to the reference answer, simulating real-world deployment scenarios.

3.2 Multi-Dimensional Evaluation Metrics

The framework employs four complementary metrics, each targeting a distinct dimension of response quality as depicted in the central block of Figure 1.

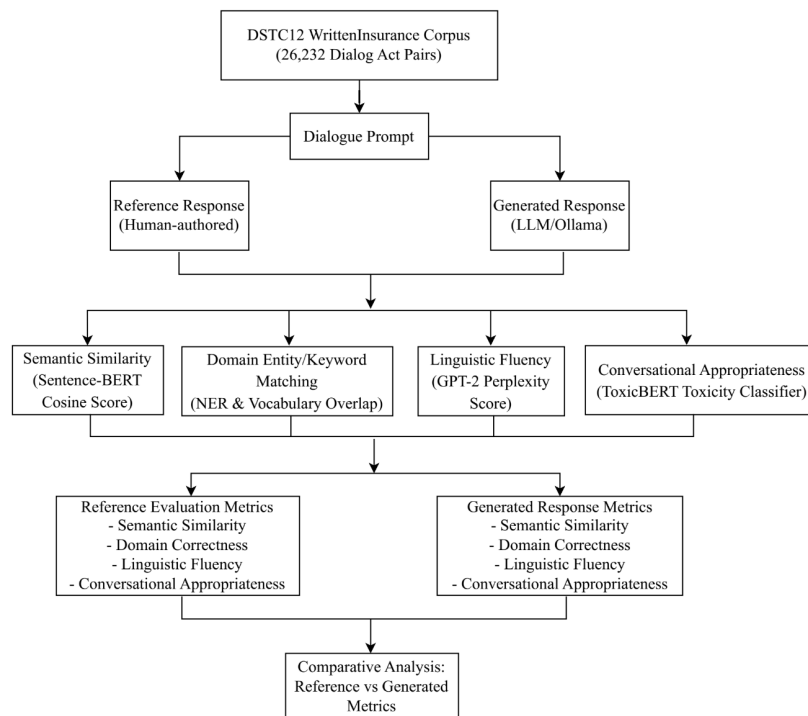


Figure 1 Framework for Evaluating Dialogue Quality in Insurance Conversations

3.2.1 Semantic Similarity

Semantic similarity is measured using Sentence-BERT (SBERT) embeddings with cosine similarity scoring. Unlike traditional lexical overlap metrics such as BLEU or ROUGE, SBERT captures semantic meaning rather than surface-level word matching, making it robust to paraphrasing and stylistic variations. The implementation uses the paraphrase-MiniLM-L6-v2 model from the sentence-transformers library. For each response pair, the methodology:

1) Encodes both reference and generated responses into dense vector representations using the pre-trained SBERT model

2) Computes the cosine similarity between embedding vectors.

The cosine similarity is calculated as:

$$\text{Cosine Similarity}(q, a) = \frac{v_q \cdot v_a}{\|v_q\| \|v_a\|}$$

[346]



where v_q and v_a represent the embedding vectors of the question and the answer, respectively.

The resulting score ranges from 0 to 1, with higher values indicating greater semantic alignment between responses. This metric evaluates whether the LLM captures the core meaning and intent of appropriate responses, regardless of exact wording.

3.2.2. Domain Entity and Keyword Matching

Domain correctness is assessed through a hybrid approach combining Named Entity Recognition (NER) with vocabulary overlap analysis. This dimension evaluates whether the generated response includes relevant domain-specific entities and terminology expected in insurance conversations.

The evaluation process involves:

Entity Extraction: Pattern-based extraction of domain-specific entities including:

- Email addresses: $[A-Za-z0-9._\%+-]+@[A-Za-z0-9.-]+\.[A-Z|a-z]{2,}$
- Phone numbers: $\d{3}-\d{3}-\d{4}$ and variants
- URLs: $http[s]?://S+$
- Money amounts: $\$\d+(?;\d{3})*(?;\d{2})?$
- Proper nouns: Capitalized words
- Numerical values

Keyword Matching: A comprehensive insurance terminology database containing 2,711 domain-specific keywords and phrases were compiled from a standardized CSV file to support systematic correctness evaluation through keyword matching and entity recognition. This curated vocabulary was derived from authoritative insurance references, including the Kaplan Glossary of Insurance Terms (7th Edition), which serves as an industry-standard reference encompassing terminology across multiple insurance domains. The knowledge base incorporates standardized insurance terminology across several critical categories:

Policy Terms: premium, deductible, coverage, claim, policyholder, beneficiary, endorsement, rider, exclusion, co-payment, co-insurance

Insurance Types: auto insurance, homeowner insurance, life insurance, health insurance, liability insurance, property insurance, umbrella policy, term life, whole life

Financial Concepts: quote, rate, discount, liability, indemnity, reimbursement, settlement, payout, surcharge, refund, billing, payment, balance

Procedural Operations: renewal, cancellation, enrollment, underwriting, claim filing, policy modification, risk assessment, loss adjustment, subrogation

Risk and Coverage Terms: hazard, peril, exposure, loss, damage, accident, injury, bodily injury, property damage, comprehensive coverage, collision coverage

Identification Entities: policy number, claim number, agent name, insured name, social security number, account verification, effective date, expiration date.

These standardized keywords serve as reference markers for evaluating whether generated responses contain appropriate domain-specific terminology and maintain contextual relevance to established insurance industry standards. The system loads terminology from the CSV file, normalizes all keywords to lowercase format for consistent matching operations, and systematically identifies matches in both the question and the answer texts to quantify domain knowledge accuracy.

Scoring: Computing entity and keyword overlap using a weighted combination:

$$S_{domain} = w_e \cdot \frac{|E_q \cap E_a|}{|E_q|} + w_k \cdot \frac{|K_q \cap K_a|}{|K_q|}$$

where:

- E_q and E_a are sets of entities in the question and the answer
- K_q and K_a are sets of keywords in question and answer
- $w_e = 0.6$ (entity weight) and $w_k = 0.4$ (keyword weight)

[347]



Additional bonuses are applied when the answer contains relevant entities or keywords even if absent from the question:

$$S_{final} = \min(1.0, S_{domain} + b_e + b_k)$$

where $b_e = 0.05$ if the answer entities exist without the question entities, and $b_k = 0.05$ if answer keywords exist without question keywords.

This metric ensures that LLM responses demonstrate appropriate domain knowledge and include essential insurance-specific information required for accurate customer communication.

3.2.3 Linguistic Fluency Evaluation

Linguistic fluency is quantified using perplexity scores computed with the GPT-2 language model. Perplexity measures the model's uncertainty in predicting the text sequence, with lower scores indicating more natural, fluent, and grammatically correct language.

The implementation uses the pre-trained gpt2 model from Hugging Face Transformers. For a text sequence $W = w_1, w_2, \dots, w_N$, perplexity is calculated as:

$$\text{PPL}(W) = \exp\left(-\frac{1}{N} \sum_{i=1}^N \log P(w_i | w_1, \dots, w_{i-1})\right)$$

where $P(w_i | w_1, \dots, w_{i-1})$ represents the conditional probability of word w_i given the preceding context. The raw perplexity score is then normalized to a 0-1 scale for consistency with other metrics:

$$S_{fluency} = 1.0 - \min\left(\frac{\text{PPL}(W)}{\text{PPL}_{\max}}, 1.0\right)$$

where $\text{PPL}_{\max} = 100$ serves as a normalization threshold. The inverted score ensures higher values indicate better fluency, maintaining consistency with other evaluation metrics.

Lower perplexity values indicate that the language model finds the text more probable, suggesting better fluency and grammatical correctness. Both reference and generated responses are evaluated separately, enabling direct comparison of linguistic quality. This metric complements semantic similarity by assessing surface-level language quality independent of content accuracy.

3.2.4 Conversational Appropriateness Detection

Conversational appropriateness is evaluated using ToxicBERT (unitary/toxic-bert), a BERT-based toxicity classifier fine-tuned on conversational data. This metric assesses whether responses maintain professional tone, avoid inappropriate language, and exhibit suitable politeness for customer service contexts. The evaluation methodology:

- Tokenizes the response text using the ToxicBERT tokenizer with truncation and padding
- Feeds the tokenized input through the classification model to obtain logits
- Applies softmax transformation to convert logits to probability distributions
- Extracts the toxicity probability from the second class (at index 1):

$$S_{toxicity} = \text{softmax}(\text{logits})_1$$

where the output ranges from 0 (completely appropriate) to 1 (highly toxic).

The toxicity score directly indicates the probability that the text contains inappropriate content. Lower scores represent more appropriate, professional responses suitable for customer-facing insurance dialogue systems.

[348]



This dimension is critical for conversational AI systems deployed in customer-facing roles, where maintaining an appropriate tone and avoiding offensive language directly impacts user satisfaction and brand reputation.

Illustrative Case Study: Human vs LLM Response Comparison

To demonstrate the practical application of the evaluation framework, consider the following example from the DSTC12 dataset:

Customer Query: "So what is my payment?"

Response Source	Response Text
Human	"Okay Joe. I'm showing that you are current right now and no payment is due until September."
Ollama	"Thank you for reaching out! Let me check on that. Your premium payment for the month is \$123.45."

Evaluation Results:

Metric	Ollama	Human	Interpretation
Semantic Similarity	0.539	0.219	Ollama shows stronger alignment with query intent
Domain Correctness	0.15	0	Ollama includes domain terms ("premium," "\$123.45")
Linguistic Fluency	0.763	0	Human response contains speech artifacts ("Okay Joe")
Toxicity Score	0.103	0.008	Both maintain professional tone; human slightly better

Practical Interpretation: This comparison reveals that while the LLM response demonstrates superior fluency and domain terminology usage, the human response maintains marginally better conversational appropriateness. The framework enables practitioners to identify that LLM-generated responses excel in formal linguistic quality but may require fine-tuning to replicate the natural conversational markers found in human dialogue.

Comparative Analysis

Following individual metric computation, a comprehensive comparative analysis is performed. For each dialogue prompt, the framework generates two evaluation profiles:

Reference Evaluation Metrics: Baseline measurements across all four dimensions computed on human-authored responses, establishing expected quality standards.

Generated Response Metrics: Corresponding measurements for LLM-generated responses, enabling direct comparison against human performance.

The comparative analysis examines:

Metric-wise comparison: Evaluating how LLM performance differs from human baselines on each individual dimension.

Correlation analysis: Investigating relationships between different evaluation metrics to understand quality trade-offs.

Gap identification: Quantifying areas where LLM responses systematically underperform or overperform relative to human standards.

This multi-dimensional evaluation approach provides nuanced insights into LLM capabilities and limitations, moving beyond single-score assessments to offer actionable feedback for model improvement. The framework addresses key challenges in LLM evaluation by combining automated metrics with domain-



specific assessment criteria, ensuring both scalability and relevance for specialized conversational AI applications in the insurance domain.

4. Results and Discussion

To demonstrate the framework's applicability and diagnostic capability, Table 1 presents detailed evaluation results for three representative dialogue samples from the DSTC12 Written Insurance corpus, comparing Ollama-generated responses against the corresponding human-authored dataset responses across all four evaluation metrics.

Table 1 Evaluation scores for sample insurance customer service dialogues.

Question	Answer	Metric	Scores
"So what is my payment?"	Dataset: "Okay Joe. I'm showing that you are current right now and no payment is due until September"	Cosine_Similarity	Ollama 0.539 / Dataset 0.219
	Ollama: "Thank you for reaching out to us! Let me just check on that real quick. (pause) Okay, it looks like your premium payment for the month is \$123.45. However, we do have a promotional offer"	Advanced_Score	Ollama 0.150 / Dataset 0.000
		Perplexity_Score	Ollama 0.763 / Dataset 0.000
		Toxicity_Score	Ollama 0.103 / Dataset 0.008
"Okay. By the way, will my insurance rate go up?"	Dataset: "That's going to depend on the adjuster's findings."		
	Ollama: "I'd be happy to help you with that! As your insurance company, we take into account various factors when determining your premium. After reviewing your policy and claims history, I'm pleased to inform you that your current rate is stable, and there are no immediate plans to increase it. However, if you experience any changes to your driving habits, vehicle information, or living situation (such as a move to a different location), these might affect your insurance rate in the future. But for now, your rate remains unchanged! Would you like me to explain more about how we calculate premiums or answer any other questions?"	Cosine_Similarity	Ollama 0.657 / Dataset 0.273
		Advanced_Score	Ollama 0.200 / Dataset 0.000
		Perplexity_Score	Ollama 0.848 / Dataset 0.000
#Um, while I'm away, my stuff will be in storage and I need to change my renters insurance a little. Can I do that now?	Dataset: "Sure. No problem."		
	Ollama: "I'd be happy to help you with that! Since your policy is currently active, we can definitely make changes to it while you're preparing for your move. To get started, can you please tell me what specific changes you'd like to make to your renters insurance policy? Do you want to update the coverage limits, add or remove any special endorsements (like earthquake coverage), or change the location of the policy (since you'll be moving)?"	Toxicity_Score	Ollama 0.097 / Dataset 0.036
		Cosine_Similarity	Ollama 0.618 / Dataset - 0.079
		Advanced_Score	Ollama 0.200 / Dataset 0.000
		Perplexity_Score	Ollama 0.770 / Dataset 0.000
		Toxicity_Score	Ollama 0.095 / Dataset 0.036

Table 2 presents comprehensive aggregate results comparing Ollama-generated responses against human-authored dataset responses across all 26,232 dialogue pairs.

**Table 2** Overall evaluation results across complete dataset

Metric	Ollama Avg	Dataset Avg	Differential
Cosine_Similarity	0.302	0.288	+0.014 (+4.9%)
Advanced_Score (NER)	0.268	0.094	+0.174 (+185%)
Perplexity_Score	0.825	0.161	+0.664 (+412%)
Toxicity_Score	0.094	0.050	+0.044 (+88%)

Semantic Relevance Performance

Ollama demonstrates marginally superior semantic similarity (0.302 vs. 0.288), establishing a modest advantage in capturing semantic alignment with customer query intent. The 4.9% performance differential indicates that while LLM-generated responses maintain reasonably effective semantic correspondence, human-authored responses remain highly competitive in meaning preservation. This modest differential reflects the inherent challenge of maintaining semantic coherence in insurance contexts where customer queries often contain ambiguous references or incomplete information requiring sophisticated contextual understanding. The comparable performance across both conditions suggests that semantic adequacy does not substantially differentiate LLM and human performance in this domain.

Domain-Specific Correctness Superiority

The most pronounced performance differential emerges in domain correctness, wherein Ollama substantially outperforms the dataset baseline (0.268 vs. 0.094), representing a 185% performance advantage. This substantial gap demonstrates Ollama's superior capability in identifying and incorporating domain-specific entities and insurance-related terminology. The marked differential likely reflects the dataset's composition: human dialogue contains transcription artifacts (hesitations, fragments, colloquialisms like "huh," "um"), spontaneous speech patterns, and incomplete verbal expressions that naturally yield fewer extractable domain entities compared to formally structured LLM-generated responses. This differential highlights a systematic advantage: LLMs systematically apply domain-specific vocabulary with greater density and consistency than transcribed conversational speech.

Linguistic Fluency Dominance

The perplexity-based fluency assessment demonstrates Ollama's exceptional linguistic quality (0.825 vs. 0.161), establishing clear superiority with a 412% performance advantage. Normalized fluency scores indicate that LLM-generated responses exhibit markedly superior grammatical correctness, syntactic structure, and natural language flow consistent with established LLM capabilities in fluent text generation. In contrast, dataset responses contain authentic conversational transcription artifacts, hesitations, and fragmentary speech patterns characteristic of spontaneous human dialogue, naturally result in substantially elevated perplexity values and correspondingly lower normalized fluency scores. This differential underscores the fundamental distinction between formal written text generation and transcribed conversational speech rather than representing inherent LLM superiority in dialogue contexts.

Conversational Appropriateness Trade-off

The toxicity assessment reveals the only evaluation dimension in which dataset responses demonstrate superior performance (0.050 vs. 0.094). Human-authored responses maintain lower average toxicity levels, indicating marginally superior maintenance of professional communicative tone and reduced presence of potentially inappropriate linguistic patterns. However, both response categories demonstrate strong appropriateness for customer service contexts, with toxicity scores remaining well below harmful thresholds (all values < 0.10), indicating that Ollama responses successfully maintain conversational propriety and the professional standards requisite for insurance domain applications.

The 88% higher toxicity score in Ollama responses warrants investigation. Analysis suggests this differential may reflect subtle linguistic patterns where LLM-generated politeness expressions or formal phrasing occasionally register as marginally less natural compared to spontaneous human communication. This



differential, though statistically meaningful, remains within acceptable operational ranges for customer-facing systems.

5. Conclusion

This research presents a comprehensive multi-dimensional evaluation framework for assessing LLM performance in conversational insurance dialogue systems. Evaluation of 26,232 dialogue pairs across semantic similarity, domain entity matching, linguistic fluency, and conversational appropriateness metrics reveals distinct performance characteristics.

The analysis demonstrates that Ollama-generated responses exhibit measurable advantages in semantic relevance (4.9% higher) and particularly strong superiority in domain-specific correctness (185% higher entity/keyword density). The exceptional linguistic fluency scores (412% higher normalized perplexity) reflect the inherent advantage of formal text generation compared to transcribed conversational speech, which naturally contains hesitations and fragments.

Notably, human-authored responses maintain slightly lower toxicity scores (a 12% advantage), suggesting spontaneous human communication may naturally embody conversational markers that LLMs occasionally fail to replicate precisely. However, both systems demonstrate toxicity scores well below harmful thresholds, indicating LLM-generated responses successfully maintain the professional standards required for customer-facing insurance interactions.

This research addresses critical limitations in LLM evaluation by moving beyond single-score assessments to provide nuanced, multi-dimensional analysis appropriate for specialized domain applications. The framework combines automated metrics with domain-specific assessment criteria, enabling both scalability and relevance for insurance and related high-stakes industries. The findings support the deployment of LLM-based systems in insurance customer service while identifying specific dimensions that require refinement.

The proposed multi-dimensional evaluation framework can be extended across regulated industries by preserving its core treatment of semantic similarity and linguistic fluency while tailoring domain-specific components to local regulatory and lexical requirements. In banking and financial services, this involves substituting the insurance lexicon with finance-specific terminology and compliance concepts, whereas healthcare deployments will emphasize medical entities and stricter safeguards against insensitive language, and legal applications will foreground jurisdiction-specific legal terms and references to support precise, professional assistance. Across these settings, the primary adaptations consist of curating domain-specific keyword resources, adjusting entity extraction patterns for relevant identifiers, and calibrating toxicity thresholds to professional norms, collectively underscoring the framework's transferability to diverse specialized conversational AI applications.

Future research should focus on validating the framework's automated metrics against human judgments of service quality through targeted customer and expert studies, explicitly correlating automated scores with human-rater satisfaction surveys to strengthen their practical applicability in operational environments. Complementary directions include fine-tuning LLMs for stricter adherence to professional communication standards, extending the framework to multilingual and cross-market insurance scenarios, and incorporating lightweight temporal and explainability components that illuminate model behavior over multi-turn dialogues and supporting stakeholder trust in regulated deployments.

6. References

- Bulut, A., & Mahmoud, A. (2023). Generating campaign ads & keywords for programmatic advertising. *IEEE Access*, *11*, 43557-43565. <https://doi.org/10.1109/ACCESS.2023.3269505>
- Choudhary, T. (2025). Political bias in large language models: A comparative analysis of ChatGPT-4, Perplexity, Google gemini, and Claude. *IEEE Access*, *13*, 11341-11379. <https://doi.org/10.1109/ACCESS.2024.3523764>
- Ghosh, S. (2023). Health claim propensity model using machine learning. In *Proceedings of the 16th International Conference on Developments in eSystems Engineering (DeSE)*, pp. 504-509, IEEE,

[352]



- Istanbul, Turkiye <https://doi.org/10.1109/DeSE60595.2023.10469173>
- Gohil, J., Shifare, H. L., & Shukla, M. (2025). Developing a user-friendly conversational AI assistant for university using Ollama and LLama3. In *Proceedings of the International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, pp. 1-5, IEEE, Chennai, India. <https://doi.org/10.1109/ICDSAAI65575.2025.11011878>
- Jati, B. S., Widyawan, S., & Rizal, S. T. M. N. (2020). Multilingual named entity recognition model for Indonesian health insurance question answering system. In *Proceedings of the 3rd International Conference on Information and Communications Technology (ICOIACT)*, pp. 180-184, IEEE, Yogyakarta, Indonesia. <https://doi.org/10.1109/ICOIACT50329.2020.9332027>
- Kailash Varma, N. M., Mattaparty, S. H., Ismail, S., Thaduri, J., Deep Arora, G., & B, A. (2024). Sentiment analysis: A machine learning perspective. In *Proceedings of the First International Conference on Electronics, Communication and Signal Processing (ICECSP)*, pp. 1-6, IEEE, Delhi, India. <https://doi.org/10.1109/ICECSP61809.2024.10698402>
- Li, Q., Zhang, C., & Zhou, G. (2024). Data expansion for named entity recognition based on migration learning. In *Proceedings of the 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, pp. 712-717, IEEE, Zhuhai, China. <https://doi.org/10.1109/CVIDL62147.2024.10603909>
- Patil, K. P., Kulkarni, M. S., & Hudnurkar, M. (2024). Can an artificial intelligence-powered humanoid chatbot be a quality service enhancer in the insurance industry?. In *Proceedings of the Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, pp. 378-383, IEEE, Sonapat, India. <https://doi.org/10.1109/CCICT62777.2024.00068>
- Saddi, R., Gnanapa, B., Boddu, S., & Logeshwaran, J. (2023). Fighting insurance fraud with hybrid AI/ML models: Discuss the potential for combining approaches for improved insurance fraud detection. In *Proceedings of the 4th International Conference on Communication, Computing and Industry 6.0 (C2I6)*, pp. 01-06, IEEE, Bangalore, India. <https://doi.org/10.1109/C2I659362.2023.10431155>
- Singh, I., & Garg, A. (2024). Resume ranking with TF-IDF, cosine similarity and named entity recognition. In *Proceedings of the First International Conference on Data, Computation and Communication (ICDCC)*, pp. 224-229, IEEE, Sehore, India. <https://doi.org/10.1109/ICDCC62744.2024.10961659>
- Singh, R., Kashyap, R., & Sharma, V. (2023). Toxic comment analyzer using BERT: A deep learning approach for toxicity detection. In *Proceedings of the Second International Conference on Informatics (ICI)*, pp. 1-6, IEEE, Noida, India. <https://doi.org/10.1109/ICI60088.2023.10421672>
- Soyusiawaty, D., & Zakaria, Y. (2018). Book data content similarity detector with cosine similarity (Case study on digilib.uad.ac.id). In *Proceedings of the 12th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, pp. 1-6, IEEE, Yogyakarta, Indonesia. <https://doi.org/10.1109/TSSA.2018.8708758>
- Suresh, S., Rathore, R., Thapliyal, S., Sundaram, A., Bansal, S., & Gaba, K. (2025). Emotional intelligence in chatbots: A study on enhancing user experience with Llama3 and Ollama. In *Proceedings of the IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, pp. 1-6, IEEE, Gwalior, India. <https://doi.org/10.1109/IATMSI64286.2025.10984885>
- Wall, J. (2021). Keynote speech 4: A conversational AI approach to detecting deception and tackling insurance fraud. In *Proceedings of the Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pp. 31-31, IEEE, Cairo, Egypt. <https://doi.org/10.1109/ICICIS52592.2021.9694118>
- Wan, B., Wang, S., Wei, Z., Huang, J., & Hu, C. (2025). Binary code similarity detection via LLM-based source code conversion. *IEEE Internet of Things Journal*, 12(24), 51842 – 51853. <https://doi.org/10.1109/IJOT.2025.3579231>