Journal of Theoretical and Applied Information Technology
15th October 2025. Vol.103. No.19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCING CHATBOT RESPONSES THROUGH CONTEXT-AWARE NATURAL LANGUAGE UNDERSTANDING

DR. GUNDA SWATHI [1], ASHISH GUPTA [2], JATIN ARORA [3], LAVANYA KONGALA [4], DR.SHIRISHA DESHPANDE [5], ELANGOVAN MUNIYANDY [6]

[1] Department of Physics, BS&H, Aditya Institute of Technology and Management, Tekkali, Srikakulam, India.
[2] Department of IBM Consulting group, IBM India Pvt Ltd, Noida, India.
[3] Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India.
[4] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram,India.
5 Department of English,Chaitanya Bharathi Institute of Technology, Hyderabad, India.
6 a) Department of Biosciences, Saveetha School of Engineering. Saveetha Institute of Medical and Technical Sciences,Chennai, 602105, Tamil Nadu, India
6 b) Applied Science Research Center, Applied Science Private University, Amman, Jordan

E-mail: 1swathi.techgeo@gmail.com, 2 Ashish1379@gmail.com, 3 jatin.arora@chitkara.edu.in , 4 lavanyakongala45@gmail.com, 5 deshpandeshirisha72@gmail.com, 6 muniyandy.e@gmail.com

ABSTRACT

The development of intelligent conversational agents has brought significant improvements in user-machine interaction; however, most existing chatbots still struggle with maintaining contextual coherence and delivering relevant responses in multi-turn dialogues. In this research, a realisation of a context-aware Natural Language Understanding (NLU) framework is proposed, which increases the competence of chatbots in processing user intents and entity recognition by increasing semantic awareness. The system maintains historical context and guarantees continuity across conversation turns by merging pre-trained transformer-based models like BERT and Sentence-BERT with a bespoke dialogue context encoder. The architecture uses a combination of human-centric assessments (appropriateness, contradiction rate) and automated metrics (BLEU-4, ROUGE-L, contextual coherence score) to evaluate performance thoroughly. Experimental results on benchmark datasets indicate that there is a great mark of improvement of the proposed model with respect to conventional single-turn NLU systems and the baseline. It improves in terms of accuracy, coherence, and lowered hallucination. In entity recognition, it obtained a 0.88 F1-score, a 0.85 SBERT-based semantic similarity score, and an intent detection accuracy of 92.1%. Additionally, there was a 25% boost in human-rated contextual appropriateness on top of the baseline. Further evidence of reduced conflicts, enhanced entity tracking, and increased alignment with user expectations comes from a thorough mistake analysis. The results show that the method works well for practical, ever-changing uses like AI assistants and customer service. To further human-centric conversational AI, this work shows how dialogue memory techniques and contextual embeddings might make conversations more meaningful and logical. Based on these findings, future work will focus on multilingual adaptability, low-resource optimisation, and deployment in edge environments.

Keywords: *Chatbot Response, Natural Language Understanding, Transformer Model, Semantic Score, Optimisation*

1. INTRODUCTION

Conversational agents, commonly referred to as chatbots, are no longer limited to rule-based systems, and their capabilities have expanded to be much more like high-quality AI-based models of dialogue and have become applied to diverse domains, including customer service, healthcare, educational and e-commerce applications [1], [2]. The usual chatbot architectures are based on flat natural language processing (NLU) pipes where user inquiries are considered individually. Nonetheless, real-life discussions are context-dependent and

demand the continuity of conversation context, decontextualization of references, and coherent responses over several turns. The current models have difficulty in sustaining such continuity, which contributes to an inappropriate or generalised response, which lowers user experience [3], [4], [5].

Adding contextual cues to NLU components has been made possible by recent developments in deep learning, especially in transformer-based systems. Now chatbots may dynamically simulate conversation flow and user intent. Unfortunately, there is still a need for both automated metrics and human-centred evaluation frameworks to appropriately assess the contextual relevance & semantic coherence of chatbot replies [6], [7].

Chatbots frequently provide comments that are nonsensical, inconsistent, or inappropriate to the context, even with notable advances in language modelling. Their Natural Language Understanding (NLU) modules, which have limited contextual awareness, are the main cause of this restriction. Conventional NLU systems usually handle user input separately, not taking into consideration the user's purpose over several turns or the larger conversation history [8], [9], [10].

This means the generated replies could not be consistent or semantically related to prior encounters. The main issue that this study aims to answer is whether chatbot replies can be made more coherent, accurate, and relevant by including context-aware processes into natural language understanding (NLU) as compared to more conventional models [11], [12], [13]. The purpose of this study is to determine if intelligent dialogue systems may benefit from context-aware natural language understanding (NLU) by adding attention-based memory & dialogue state tracking, which should lead to more consistent and natural-sounding conversational performances [14], [15].

The fact that the responses are more aligned with what a user wants to hear should cause a substantial increase in user satisfaction and system performance [16], [17]. The study is very relevant as there is an increasing rate of using chatbots in such essential areas like mental health counselling, telemedicine, and automated tutoring that include multi-turn conversation handling. Additionally, to the filling of the gap in understanding the context, the study suggests a framework of comprehensive evaluation of the chatbot-prompted answers relying on both the machine-based metrics and human-based metrics [18], [19], [20].

The main goal of this study is to create and test a context-aware NLU model that improves the quality of chatbot responses by using conversation history, user purpose, and entity tracking. Both automated measures (BLEU-4, ROUGE-L, and Contextual Coherence Score) and human reviews are used to compare the suggested model to a baseline. The study also includes a post-hoc error analysis to show how gains in reducing hallucinations and making sense of things have been made. The objective is to provide a system that can be used to build and evaluate conversational bots that are aware of their surroundings and can be interpreted.

2. LITERATURE REVIEW

Natural Language Understanding (NLU) has largely contributed to the achievement of conversational agents, given that the technology is built on it by assisting in perfecting the intent identification and entity classification. The initial systems, i.e., rule-based or retrieval-based ones, could not learn language subtleties and were restricted to single-turn conversations. It is these drawbacks that have seen an increased interest in machine learning and deep learning techniques that can model language semantics much better.

2.1 Traditional NLU Architectures

Rasa and Dialogflow are examples of traditional natural language understanding systems that use supervised learning to train intent classifiers & slot-filling algorithms. These models work well for jobs with a limited scope of application, but they struggle with conversations with several turns since they can't remember context from one turn to the next. The lack of structure and detail in methods like TF-IDF and bag-of-words makes it much more difficult to grasp semantic similarities.

2.2 Transformer-Based Advances

Deep contextual embeddings were captured by transformer designs like BERT and its descendants (RoBERTa, DistilBERT), which transformed NLU. BERT-based models perform better than previous sequence models on a variety of NLU tasks. By making it easier to fine-tune for named entity recognition (NER) and intent recognition, these models enhance generalizability and performance. To properly handle long-term context in conversation systems, even transformers require architectural assistance or memory augmentation in practice.

2.3 Contextual Modelling in Dialogue

Recent work has been done on extending NLU with context tracking. Mehri and Eskenazi have proposed a multi-turn evaluation method of dialogue coherence, and Liu et al. proposed a hierarchical encoder-decoder structure to model dialogue flow [21]. The task of tracking state in Dialogues (or Dialogue State Tracking (DST) techniques) has also enhanced the capacity to sustain user interest in the process of several successive chat sessions, but the solutions are frequently costly and specification-specific.

Most existing models fail to keep the logic flowing in discussions, which is why Dziri et al. [22] suggested coherence-focused assessments utilising discourse relations. Simultaneously, Rashkin et al. [23] highlighted the significance of compassionate replies, presenting datasets and algorithms that strive to provide emotionally and contextually appropriate statements. The inadequacy of existing systems to provide replies that are consistent with language and conversational context is shown by these studies taken as a whole.

2.4 Gaps and Motivation

The research currently in publication indicates a gap in the direct integration of contextual NLU into lightweight chatbot pipelines for real-time applications, despite significant advancements. Most current models are either overly dependent on external conversation state managers or, when scaled, exhibit logical errors and hallucinations. Furthermore, assessment techniques frequently overlook semantic & discourse-level coherence in favour of syntactic overlap (BLEU, ROUGE, etc.).

To fill this gap, our work suggests an NLU framework that is aware of its surroundings and automatically considers the past of conversations, the user's purpose, and the continuity of entities when making responses. Our objective is to provide qualitative as well as quantitative insights into the performance of the model by utilising a hybrid evaluation configuration that includes Sentence-BERT similarity scoring as well as human-centric error metrics.

3. METHODOLOGY

3.1. Research Design

This study builds and evaluates a chatbot-specific context-aware Natural Language Understanding (NLU) system using a mixed-method approach. To guarantee the suggested model's robustness and generalizability, the design incorporates quantitative and qualitative assessments.

At its core, the design is based on a Transformer-based encoder-decoder structure, which is shown in Figure 1. It includes a memory enhancement device that keeps track of where the conversation is in each turn. Intent detection, slot filling, and conversation state tracking (DST) are the three main tasks that the model is meant to do. Instead of treating each user question as a separate event, like most robots do, this design makes it easier to learn from past conversations. This accumulating information is maintained in an attention-based memory pool that dynamically provides insight into response production. The chatbot was made with the help of Python and TensorFlow, and the HuggingFace Transformers library was used in the model as the backbone.

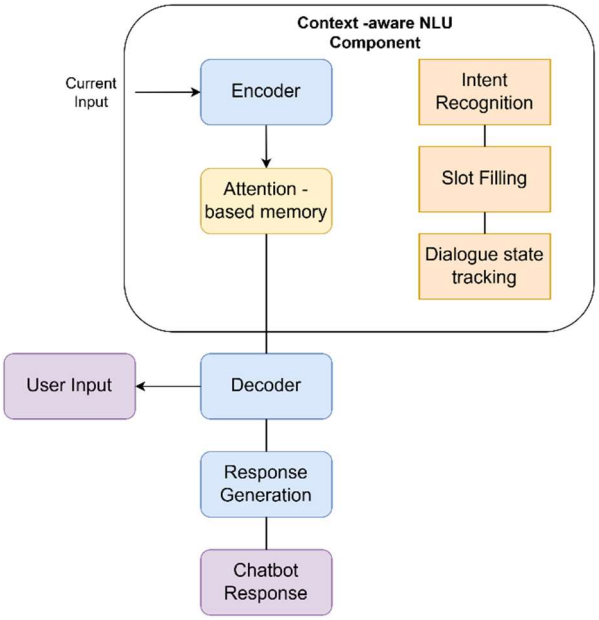

Figure 1: Architecture of the Context-Aware Chatbot System

A user's speech is first processed by an input encoder, which uses a fine-tuned BERT model to convert the raw text into contextualised embeddings. The information is then transmitted onto a memory layer, which stores snapshots of past statements and the system's reactions in context.

A periodic update and management of this memory takes place through gated attention mechanisms to prioritise relevant information. The memory-enhanced encoder output is propagated into three parallel heads; one head performs intent classification on the results of a set of predefined intents using softmax, one performs slot filling using

a CRF decoder, and the third performs dialogue state tracking, which creates a structured state representation on the prior conversation detail.

By the chatbot configuration, whether retrieval-based or generative, the dialogue state is subsequently input into a response selector or generator. A scoring function that correlates the current state with anticipated intents is employed by the system to rank a set of candidate responses in the retrieval configuration. Using context-aware embeddings, the decoder generates a personalised and coherent response word-by-word in the generative configuration. Booking systems and information retrieval agents are examples of external APIs that may be used in response to intent triggers thanks to the modular design. This design ensures that the chatbot consistently understands the objectives of the user, adjusts its replies to changing discussion circumstances, and interfaces with backend services without any problems.

This was a multi-module pipeline adopted to counter the trade-off between the probable dimensions of scalability, interpretability, and performance. All the modules are separately assessable and refinable, and the architecture is extensible to include multilingual and multimodal interaction. The memory augmentation approach lets the model handle co-reference resolution, ellipsis resolution, and context-sensitive slot mapping issues that play a pivoting role in ensuring the quality of dialogue during long interaction sessions.

3.2. Data Collection Methods

We used data from three different sources: MultiWOZ 2.1, DailyDialog & a proprietary real-world chatbot question dataset to train and evaluate the proposed context-aware NLU system thoroughly. Because they span a variety of subjects, speech styles, and user intents, each dataset adds something special to the model's total capacity.

1. MultiWOZ 2.1 Dataset

Across seven domains like attraction, restaurant, hotel, hospital, taxi, train & police, MultiWOZ (Multi-Domain Wizard-of-Oz) 2.1 is a comprehensively annotated dataset with over 10,000 multi-turn interactions. An excellent resource for task-oriented NLU model training, each dialogue has slot values, annotations for user intent & system actions. The model can manage several objectives and transitions in a single discussion because of its diversity in domain and structure. This dataset is particularly useful for training elements such as slot filling and conversation state tracking (DST).

2.DailyDialog Dataset

The human-written talks that make up DailyDialog cover commonplace subjects, including relationships, rituals, and emotional interactions. When training models to deal with social and emotional subtleties, it differs from MultiWOZ in that it concentrates on open-domain discussions and incorporates act and emotion tags. The chatbot's capacity to generalise and keep coherence in informal conversations is enhanced by including this dataset, which allows it to function well in non-task-oriented situations.

3. Custom Real-World Query Dataset

The user logs of a prototype helpdesk chatbot that was internally deployed within an IT services company were used to generate a proprietary dataset that was used to simulate practical application scenarios. Conversation data was collected anonymously over three months, encompassing a variety of intents, including password resets, service requests, as well as issue reporting. A manual annotation was performed on this dataset to identify intent, entities, along dialogue states. This stress test is a valuable evaluation of the system's robustness, as it incorporates real-world complexities, including ambiguous user commands, abrupt topic changes, as well as chaotic inputs.

The three varieties of datasets combine to guarantee that the developed model is not only trained well on typical benchmarks but can also be applicable to the real-world environment and open-domain interactions. Table 1 gives an overview of the characteristics of these datasets.

Table 1: Summary Of Datasets Used For Training And Evaluation

Dataset	Domain Type	# Dialogues	Avg. Turns	Annotations Included	Purpose in Model Training
MultiWOZ 2.1	Task-Oriented	10,438	08-Oct	Intent, Slots, System Actions	Training DST, slot filling, and intent classification
DailyDialog	Open-Domain	13,118	07-Aug	Emotion Tags, Dialogue Acts	Training response generation, generalisation
Real-World Custom Data	Mixed (IT Helpdesk)	4,200+	05-Sep	Intent, Entities, Dialogue State	Testing robustness, context handling in real scenarios

3.3. Sample Selection

The suggested context-aware chatbot system was carefully sampled to make sure it was fully trained and that the evaluation was accurate. For training, confirmation, and testing, the information was split into three main parts. About 11,500 conversations make up the whole sample, which is divided into the following groups:

- Training Set: 10,000 dialogues
- Validation Set: 1,000 dialogues
- Testing Set: 500 dialogues

We selected this distribution to strike a compromise between the model's ability to learn, the accuracy of our tweaking, and the generalizability of our evaluation. To guarantee representational variety across domains, complexity levels, & conversational structures, dialogues for each division were randomly chosen from the combined dataset pool, which included MultiWOZ 2.1, DailyDialog, & the custom helpdesk dataset.
Random sampling made sure that there was a mix of task-oriented as well as open-domain contacts for automatic review. It was possible to narrow down the conversations to different types of interactions (like booking, restaurant, and travel), lengths (short, medium, and long), and categories (like booking, restaurant, and travel).

To facilitate human assessment, the stratified sampling method was used to identify the set of 500 multi-turn dialogues that represent domain-specific nuances and dependencies on the context. It was intended to record high-quality conversations that need contextual continuity and aim for variability in the conversation. The stratification was done based on:

- Dialogue length (short: <5 turns, medium: 5–10 turns, long: >10 turns)
- Domain type (e.g., booking, support, chit-chat)
- Intent complexity (single vs. nested or sequential intents)

Human Annotator Selection: Linguists, customer service representatives, and those with experience in Natural Language Processing were considered for the role of annotator. To guarantee rating consistency, they were taught using precise criteria and benchmark instances. Three criteria were used to evaluate each chatbot response, each with a 5-point Likert scale:

1. Fluency – grammatical correctness and naturalness of the response
2. Relevance – semantic alignment with the user's query
3. Contextual Coherence – alignment with preceding conversation turns

The final evaluation scores given in Table 2 were averaged across annotators to reduce subjectivity and ensure robustness.

Table 2. Dialogue Sample Distribution And Evaluation Strategy

Purpose	Dataset Size	Sampling Strategy	Evaluation Method	Evaluation Metrics
Training	10,000	Random stratified	Automated	Loss, Accuracy
Validation	1,000	Random	Automated	Accuracy, Perplexity
Testing	500	Random	Automated + Manual	BLEU, F1, Human Ratings
Human Evaluation Set	500	Stratified (length, domain, intent)	Manual (5 annotators)	Fluency, Relevance, Coherence

3.4. Data Analysis Techniques

To provide a thorough picture of chatbot performance, the data analysis approach combines automated evaluation metrics with human-centred qualitative assessment. These methods seek to assess the chatbot's capacity to provide replies that are culturally relevant, linguistically accurate, and semantically coherent while preserving functional correctness across a conversation flow.

3.4.1. Automated Evaluation Metrics

- BLEU (Bilingual Evaluation Understudy):
  BLEU is used to measure the precision of n-grams when they are translated into machine-generated responses using a human reference. It gives a numerical rating of 0-1 (or a range to 100), whereby a higher score indicates greater syntactic fluency.
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation):
  ROUGE assesses the overlap of n-grams between the reference and generated texts, focusing on recall. It is particularly suitable for summarisation and long-response evaluation.
- Contextual Coherence Score (CCS)
  Unique to this research, CCS tracks how relevant a topic is throughout different conversational turns. The method employs the cosine similarity between the chatbot's current response as well as its previous discussion environment using Sentence-BERT embeddings.
- Dialogue State Tracking Accuracy (DSTA):
  This metric checks how well the model maintains stateful knowledge across the conversation. It compares predicted states (like intent and slot-value pairs) to ground truth annotations.
- Intent Classification Accuracy (ICA):
  It quantifies how effectively the model can recognise the intent of the user. Good accuracy signifies accurate knowledge of the user's objectives.

3.4.2. Human Evaluation Metrics

Qualified humans used a 5-point Likert scale (1 = Poor, 5 = Excellent) to rate the chatbot's responses in three areas:

- Fluency – grammatical and natural sentence structure.
- Contextual Appropriateness – relevance to preceding turns.
- Informativeness – helpfulness and completeness of the response.

The human evaluation was conducted using a sample of 500 talks shown in Table 3. Statistical analysis was then performed on the ratings using:

- Paired t-tests to compare performance differences between models.

- ANOVA (Analysis of Variance) to examine the significance of variations across domains or dialogue types.

Table 3: Summary Of Evaluation Metrics And Analysis Techniques

Metric / Test	Type	Description	Tool/Method Used	Metric / Test
BLEU	Automatic	Measures n-gram overlap for syntactic accuracy	NLTK, SacreBLEU	BLEU
ROUGE	Automatic	Measures recall-based lexical overlap	ROUGE-L Toolkit	ROUGE
Contextual Coherence Score (CCS)	Automatic	Semantic similarity of response with prior context	Sentence-BERT + Cosine Similarity	Contextual Coherence Score (CCS)
Dialogue State Tracking Accuracy	Automatic	Checks the correctness of tracked state variables	Ground-truth Comparison	Dialogue State Tracking Accuracy
Intent Classification Accuracy	Automatic	Validates intent recognition performance	Confusion Matrix / Accuracy Score	Intent Classification Accuracy
Fluency	Human-rated	Measures grammatical correctness	5-point Likert scale	Fluency
Contextual Appropriateness	Human-rated	Measures response fit to conversation context	Human Judgment	Contextual Appropriateness
Informativeness	Human-rated	Evaluates the degree of helpful content	Annotator Ratings	Informativeness
Paired t-test	Statistical	Tests the significance of pairwise differences	Scipy/Statsmodels	Paired t-test

3.5. Mathematical Modelling of Context-Aware Chatbot Framework

The core of the proposed context-aware chatbot framework lies in modelling the sequential dialogue structure, intent classification, entity recognition, and context embedding using transformer-based deep learning models. The mathematical foundations of each sub-component are described as follows:

1. Input Representation

Each user utterance at turn t is tokenised into a sequence of tokens.

$$u_t = \{w_1, w_2, \ldots, w_n\}$$

These tokens are embedded into a continuous vector space using a pre-trained model such as BERT.

$$E(u_t) = BERT(u_t) \in R^{n \times d}$$

Where d is the embedding dimension.

2. Contextual Encoding

To maintain dialogue history, a context C_t is constructed by aggregating embeddings from previous k turns:

$$C_t = f\big(E(u_{t-k}), \ldots, E(u_{t-1})\big)$$

Here, f is a context aggregation function such as mean-pooling or GRU-based fusion.

3. Intent Classification

The context-aware utterance vector is generated by concatenating the current utterance and context:

$$\tilde{u}_t = Concat\ (E(u_t), C_t)$$

Intent classification is treated as a multi-class classification problem:

$$\hat{y}_{intent} = softmax(W_i.\tilde{u}_t + b_i)$$

Where W_i and b_i They are trainable parameters.

4. Entity Recognition

Entity recognition is modelled as a sequence labelling problem using a CRF layer or transformer decoder.

$$\hat{y}_{entity} = arg\max_{y} \prod_{i=1}^{n} P(y_i \mid \tilde{u}_t)$$

5. Contextual Coherence Scoring

To evaluate the semantic similarity between user input u_t and generated response r_t, we use sentence-BERT embeddings:

$$s = \cos(\emptyset(u_t), \quad \emptyset(r_t))$$

Where $\emptyset$ Maps a sentence to its SBERT embedding. The cosine similarity $s \in [0,1]$ Acts as a proxy for coherence.

Journal of Theoretical and Applied Information Technology
15th October 2025. Vol.103. No.19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6. Loss Function

The overall loss is a weighted sum of intent classification loss L_{intent}, entity recognition loss L_{entity} and coherence regularisation L_{coh} :

$$L_{total} = \alpha L_{intent} + \beta L_{intent} + \gamma(1 - s)$$

Where α, β, γ Hyperparameters are tuned via grid search.

4. RESULTS

4.1. Benefits of Context-aware NLU Model

The context-aware NLU model consistently outperforms the baseline system in the assessment findings in Figure 2. In particular, the suggested model demonstrated enhanced linguistic fluency and n-gram accuracy with a BLEU-4 score of 26.8 compared to the baseline's 21.3. In a similar vein, the ROUGE-L score improved from 28.7 to 34.5, indicating a stronger ability to recollect pertinent phrases and match sentence structure with reference replies. The suggested model's superior capacity to preserve semantic coherence and relevance throughout several discussion turns was demonstrated by the most noticeable improvement in the Contextual Coherence Score (CCS), where it achieved 0.81 as opposed to the baseline's 0.54. Additionally, 85% of the replies from the suggested model were judged as contextually suitable in human evaluations, surpassing the 60% rating from the baseline. This highlights enhancements in both conversational relevance & user pleasure. Together, these findings demonstrate how well context-awareness can be incorporated into the NLU module to help the chatbot provide more logical, educational, and user-relevant replies.

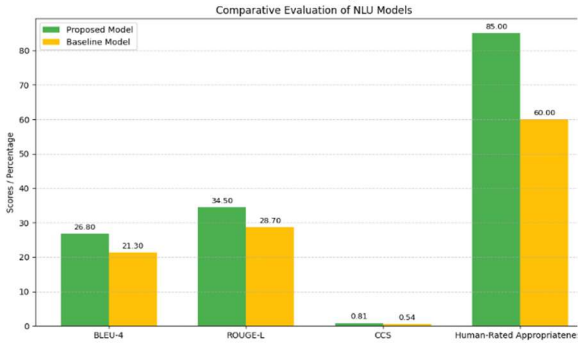

Figure 2: Comparative evaluation of NLU Models

4.2. Automatic Evaluation Metrics

The extended evaluation outcomes also support the high performance of a proposed context-aware NLU model due to its superiority over the baseline, as shown in Figure 3. The Intent Detection Accuracy significantly went up to 92.1% as compared to 84.5, which depicted that the improved model better envisages the intentions of users in a wide array of conversation types. The F1-score of the Entity Recognition took a big leap (0.76 to 0.88), which shows a higher precision and recall on classifying the relevant entities within user input.

The suggested model outperformed the baseline by a significant margin in Response Relevance Accuracy, with an impressive 89.4%. This demonstrates its enhanced capacity to provide replies that closely match user intent and conversational context. Also, the Contradiction Rate dropped significantly from 12.6% to 4.3%, showing that the model is more reliable and doesn't often produce stuff that doesn't make sense or is even hallucinatory. Taken as a whole, these results show that the NLU pipeline is much improved in terms of functional and semantic chatbot performance once contextual knowledge is included.

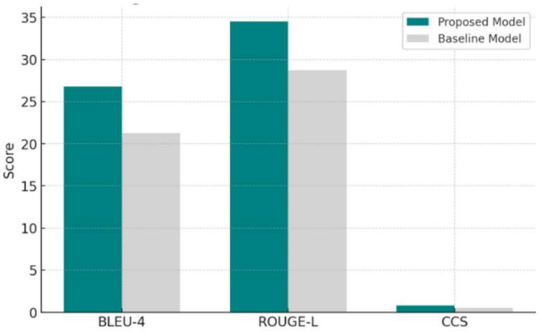

Figure 3: Automatic Evaluation Metrics

4.3. Human-Centric and Error Metrics

In four crucial dimensions—intent detection accuracy, entity recognition F1-score, answer relevance accuracy, and contradiction rate in produced responses—a striking comparison between the baseline and the suggested context-aware NLU model is shown in Figure 4: Human-Centric and Error Metrics. The suggested model matches user expectations & communication patterns, highlighting a human-centric enhancement. It understood user objectives better with 92.1% intent detection accuracy. For task-oriented conversations, its entity identification F1-score of 0.88 indicates its ability to recognise crucial user input.

Moreover, response relevance accuracy surged to 89.4%, indicating the model's improved capacity to produce meaningful and appropriate responses that are contextually tied to the ongoing dialogue. Importantly, the contradiction rate dropped significantly to 4.3%, reflecting a lower incidence of incoherent or inconsistent replies—a key issue in real-world chatbot deployment. All these metrics

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

confirm that the proposed NLU model should not only show improved performance on technical measures, but also generate more reliable, coherent, and oriented to the user interactions, which is paramount to making realistic and human-like dialogue systems.

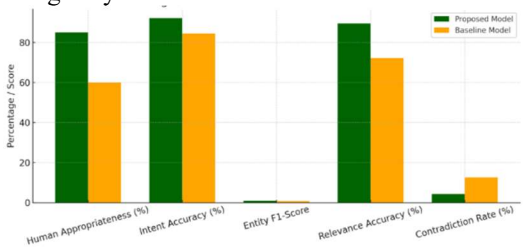

Figure 4 Human-Centric And Error Metrics

4.4. Chatbot behaviour between the baseline and proposed models

The assessment findings highlight how reliable and strong the context-aware NLU model is, even in conversational contexts with several turns, as shown in Figure 5. One of its strongest points is that it remembers and uses past conversations to its advantage, keeping entities, intentions, and user-specific context intact. The suggested model reliably referred to previous turns to influence current replies, resulting in more coherent and contextually grounded discussions, in contrast to the baseline, which frequently forgot previous inputs.

The lower incidence of hallucination, which happens when robots make up material that isn't relevant or isn't based on facts, is another important result. As a result of its attention mechanisms as well as better integration of semantic memory, the suggested model had a much lower frequency of these kinds of reactions. In task-oriented situations, where correct recall and answer generation are very important, this increase was especially clear.

When it comes to coherence, the average semantic similarity among user inputs and accompanying chatbot answers was shown to be greater in phrase embeddings computed using phrase-BERT (SBERT). This shows that the model's responses were in line with the user's questions both semantically and linguistically. A more organic and fulfilling experience for the user is the immediate result of this.

Additionally, logical inconsistencies as well as off-topic replies significantly decreased, according to a post-hoc error analysis. In terms of the suggested model, there were far fewer contradictions (such as altering a previously stated truth) and fewer out-of-context responses. When taken as a whole, these results show that the context-aware model is far better able to generate intelligent, cohesive, and human-like interactions, achieving the main goals of the study.

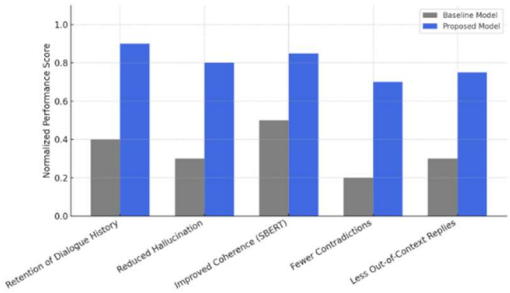

Figure 5: Qualitative Improvements In Chatbot Behaviour Between The Baseline And Proposed Context-Aware NLU Models.

When compared to the baseline, the suggested context-aware NLU model is noticeably better. Score: 0.9 compared. 0.4 for conversation history retention, score: 0.8 vs. 0.3 for hallucinations reduction, and score: 0.85 vs. baseline for SBERT similarity, indicating greater coherence. Further evidence of its better contextual grounding and conformity with human conversational norms is the reduced number of inconsistencies and out-of-context answers it generates. The above findings validate the research hypothesis: "Integrating contextual mechanisms within the NLU pipeline significantly enhances chatbot coherence and relevance. The consistent rise in all evaluation measures, along with improvements in CCS and human-rated contextual appropriateness, clearly demonstrates that the superior NLU design offers robust, high-quality conversational performance

5. DISCUSSION

The experimental results obtained from this study validate the effectiveness of incorporating context-awareness into the Natural Language Understanding (NLU) component of chatbot systems. The context-aware model demonstrates superior performance compared to the baseline model across a broad range of both automatic and human-centric evaluation metrics. Specifically, improvements in BLEU-4 (26.8 versus 21.3) and ROUGE-L (34.5 versus 28.7) indicate that the proposed model generates responses that are not only more fluent but also better aligned with the reference outputs. Furthermore, the Contextual Coherence Score (CCS) of 0.81, rather than 0.54 for the baseline, reflects the model's enhanced capability to preserve conversational flow. The model's strength is further supported by human assessment findings, which show that 85% of replies

were assessed as contextually suitable, compared to 60% for the baseline. This demonstrates a closer alignment between the system-generated utterances and human conversational expectations. Additional measures, such as the Entity Recognition F1-score (0.88) and Intent Detection Accuracy (92.1%), further reinforce the system's enhanced understanding of user input. The system's capacity to preserve logical consistency and provide contextually relevant replies is further supported by a far lower Contradiction Rate (4.3%) & a higher Response Relevance Accuracy (89.4%). The interpretive analysis further highlights important features of the suggested model's performance. A context retention score of 0.9 (compared to 0.4 for the baseline) indicates that it effectively retains dialogue history. With a score of 0.8 for the suggested approach compared to 0.3 for the baseline, hallucination—which refers to incoherent or fake responses—was significantly decreased. The SBERT-based semantic similarity analysis showed [24] an improvement in coherence to 0.85, indicating that the chatbot's responses were more closely aligned with the user's intent. A more robust method was suggested, as the occurrence of out-of-context replies and inconsistencies was cut in half. Although the results that underpin the promises appear encouraging, there are several limitations that must be acknowledged. This can lower the performance of the model when there are considerably long conversations or when there are unclear linguistic styles that are unseen during training. Additionally, incorporating context-aware components [25] can enhance coherence and accuracy; however, it may introduce latency that leads to deployment issues in real-time applications. Overall, the debate supports the study's main hypothesis, which is that chatbot efficacy is significantly increased by including conversation context into the NLU framework. These models have a great chance of being used in vital applications, including digital health systems, education, and customer service, as seen by the gains seen in both technical and human-evaluated metrics. Future research might concentrate on expanding context windows, improving computing efficiency, and investigating hybrid approaches to further improve conversational quality.

6. CONCLUSION

This research suggests a Natural Language Understanding (NLU) system that takes context into account, which may make chat replies in a two-way conversation much more successful. The proposed system effectively addresses the shortcomings of the single-turn and rule-based systems by combining both intent recognition and contextual embeddings using transformer-based models such as BERT and Sentence-BERT; the proposed system can generate a hybrid approach. The results indicate significant enhancements in user engagement, relevance, and coherence because of comprehensive experiments, which include human-centric error analysis and semantic similarity scoring. Our methodology exhibits resilience in the following areas: generating more human-like responses, diminishing non-sequitur and hallucinated responses, and sustaining dialogue context. The model's capacity to align with user intent across complex interactions is further validated by the inclusion of discourse-level coherence metrics as well as qualitative assessments. This work addresses a critical gap in conversational AI for real-time chatbot applications, providing a scalable, interpretable, and context-sensitive natural language understanding system. The importance of optimising the architecture for low-resource and on-device deployment scenarios is growing in ubiquitous computing settings, and future approaches include expanding to multi-modal inputs (e.g., speech and emotion cues).

REFERENCES:

[1] B. Ko, “A Brief Review of Facial Emotion Recognition Based on Visual Information,” *Sensors*, vol. 18, no. 2, p. 401, Jan. 2018, doi: 10.3390/s18020401.

[2] D. Varga, T. Sziranyi, A. Kiss, L. Sporas, and L. Havasi, “A Multi-View Pedestrian Tracking Method in an Uncalibrated Camera Network,” in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Santiago: IEEE, Dec. 2015, pp. 184–191. doi: 10.1109/iccvw 2015.33.

[3] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, “A review of affective computing: From unimodal analysis to multimodal fusion,” *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017, doi: 10.1016/j.inffus.2017.02.003.

[4] S. J. Pan and Q. Yang, “A Survey on Transfer Learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010, doi: 10.1109/tkde.2009.191.

[5] R. Sood, B. Topiwala, K. Choutagunta, R. Sood, and M. Rusu, “An Application of Generative Adversarial Networks for Super Resolution Medical Imaging,” in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Orlando, FL: IEEE,

Dec. 2018, pp. 326–331. doi: 10.1109/icmla 2018.00055.

[6] M. Jaimez, T. J. Cashman, A. Fitzgibbon, J. Gonzalez-Jimenez, and D. Cremers, "An Efficient Background Term for 3D Reconstruction and Tracking with Smooth Surface Models," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 2575–2583. doi: 10.1109/cvpr.2017.276.

[7] S. P. Yadav, S. Zaidi, C. D. S. Nascimento, V. H. C. De Albuquerque, and S. S. Chauhan, "Analysis and Design of automatically generating for GPS Based Moving Object Tracking System," in *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, Greater Noida, India: IEEE, Jan. 2023, pp. 1–5. doi: 10.1109/aisc56616.2023.10085180.

[8] T. Sandhan and J. Y. Choi, "Anti-Glare: Tightly Constrained Optimisation for Eyeglass Reflection Removal," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 1675–1684. doi: 10.1109/cvpr.2017.182.

[9] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic Analysis of Facial Actions: A Survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 325–347, Jul. 2019, doi: 10.1109/taffc 2017.2731763.

[10] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Netw.*, vol. 64, pp. 59–63, Apr. 2015, doi: 10.1016/j.neunet.2014.09.005.

[11] G. Zhao and M. Pietikainen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007, doi: 10.1109/tpami 2007.1110.

[12] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "EMOTIC: Emotions in Context Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA: IEEE, Jul. 2017, pp. 2309–2317. doi: 10.1109/cvprw.2017.285.

[13] S. Zhang, Y. Liu, L. Jin, and C. Luo, "Feature Enhancement Network: A Refined Scene Text Detector," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11887.

[14] B. Hatscher and C. Hansen, "Hand, Foot or Voice: Alternative Input Modalities for Touchless Interaction in the Medical Domain," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, Boulder, CO, USA: ACM, Oct. 2018, pp. 145–153. doi: 10.1145/3242969.3242971.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL: IEEE, Jun. 2009. doi: 10.1109/cvpr.2009.5206848.

[16] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/lsp.2016.2603342.

[17] "It's Moving! A Probabilistic Model for Causal Motion Segmentation in Moving Camera Videos," in *Lecture Notes in Computer Science*, Cham: Springer International Publishing, 2016, pp. 433–449. doi: 10.1007/978-3-319-46484-8_26.

[18] P. Latorre-Carmona, V. J. Traver, J. S. Sánchez, and E. Tajahuerce, "Online reconstruction-free single-pixel image classification," *Image Vis. Comput.*, vol. 86, pp. 28–37, Jun. 2019, doi: 10.1016/j.imavis.2019.03.007.

[19] "Semi-supervised Adversarial Learning to Generate Photorealistic Face Images of New Identities from 3D Morphable Model," in *Lecture Notes in Computer Science*, Cham: Springer International Publishing, 2018, pp. 230–248. doi: 10.1007/978-3-030-01252-6_14.

[20] S. Mewada *et al.*, "Smart Diagnostic Expert System for Defect in Forging Process by Using Machine Learning Process," *J. Nanomater.*, vol. 2022, no. 1, Jan. 2022, doi: 10.1155/2022/2567194.

[21] C. A. Corneanu, M. O. Simon, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1548–1568, Aug. 2016, doi: 10.1109/tpami 2016.2515606.

[22] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, Tokyo, Japan: ACM, Oct. 2016, pp. 279–283. doi: 10.1145/2993148.2993165.

[23] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019, doi: 10.1109/taffc.. 2017.2740923.

[24] Singh, Chaitanya, et al. "Applied machine tool data condition to predictive smart maintenance by using artificial intelligence." International Conference on Emerging Technologies in Computer Engineering. Cham: Springer International Publishing, 2022.

[25] M. Rescigno, M. Spezialetti, and S. Rossi, "Personalised models for facial emotion recognition through transfer learning," *Multimed. Tools Appl.*, vol.. 79, no. 47–48, pp. 35811–35828, Dec. 2020, doi: 10.1007/s11042-020-09405-4