# Error Analysis of BERT model for Chatbot using various Performance Measures

Bertilla Fernandes
School of Computer Sciences,
Kavayitri Bahinabai Chaudhari North Maharashtra University,
Jalgaon, Maharashtra, India.
Department of BSc Information Technology, Jai Hind College (Empowered Autonomous), Mumbai, Maharashtra, India

Snehalata B. Shirude
School of Computer Sciences,
Kavayitri Bahinabai Chaudhari North Maharashtra University,
Jalgaon, Maharashtra, India.
National Assessment and Accreditation Council, Bangalore

## ABSTRACT
Many opportunities for changing how information and computer systems engage with more naturally, accessible way are presented by Conversational Agents (CAs). There can be possibilities in which human expectations can be fallen short of and "failed" by these CAs. BERT, a Google creation, is a notable development in natural language processing (NLP) with impressive results on a variety of tasks including Chatbots. BERT models are designed to help understand the intricate contextual relationships between each word in a statement. The evaluation metrics of the Question Answering task, can assess the factuality of large language models (LLMs). In this study an explanation of the evaluation measures for error analysis with BERT transformer model for conversational agents is provided and also details of the strengths and limitations of using these evaluation measures for chatbots in response generation is given. The impact of six different types of conversational errors was systematically analyzed by us. Work is done on diverse variants of the BERT model and detailed analysis of the evaluation measures for error analysis on a python FAQ dataset which includes the question, answer and context is performed. It was analyzed that BERTSCORE supplements better with human decisions and brings forth better model selection performance compared to present metrics. Finally, the paper concludes with discussion on the strengths and limitations of the various metrics with error analysis for conversational agents.

## Keywords
Conversational-agent based error analysis, chatbots, education, error analysis, response generation, large language models.

## 1. INTRODUCTION
Our daily lives are increasingly integrated by Artificial Intelligence (AI) with the creation and exploration of intelligent software and hardware, called intelligent agents. A range of tasks can be done by intelligent agents, ranging from simple tasks to sophisticated operations. A classic example of an AI system is a chatbot. It is a computer program that responds to voice or text messages as though it were an intelligent entity, and has the ability to learn a variety of human languages by Natural Language Processing (NLP). In the realm of technology, a chatbot is defined as, "A computer program designed to mimic communication with human users, especially via the internet." Smart bots, interactive agents, digital assistants, or artificial conversational agents are also known as chatbots.

Specifically, the development of superior natural language responses for chatbots is considered a difficult and extensive task that is often dependent on superior training data and extensive domain knowledge. Consequently, it is crucial for experts with the required domain knowledge to be engaged in the process of creating chatbot responses. Natural Language Processing techniques are used by the chatbot to gradually enhance responses by examining user input and modifying algorithms accordingly.

The efficiency of generating responses by chatbots is highly dependent on NLP (Natural Language Processing) algorithms. These algorithms are driven by techniques from machine learning and artificial intelligence, which facilitate the improvement of Chatbot language generation and understanding. Knowledge bases and language models are also leveraged by chatbots, leading to more effective and natural language generation. Natural language processing (NLP) helps natural human language be understood by chatbots, and the intent behind queries or statements to be detected, while appropriate responses are generated. Fernandes and Shirude [19] in their work state that Natural language understanding (NLU) is widely used to extract identifying semantic information from users' query. The preciseness of chatbot responses is enhanced when language patterns and context are observed with the help of NLP algorithms. By understanding this process, places where typical errors could happen can be traced, and steps can be taken to prevent them. Accurate, natural, and engaging responses are ensured by a robust evaluation of a chatbot's knowledge base, language model, and neural network learning abilities. The robustness of chatbots is evaluated through essential components such as performance metrics. These indicators are crucial for evaluating a chatbot's effectiveness in practical applications, including user engagement rate, accuracy, response rapidness, and assessment of satisfaction. Mandlik et al. [13] in their work have used advanced natural language processing (NLP) techniques, utilizing the BERT algorithm for question understanding and the GPT algorithm for generating accurate and coherent responses.

With the progress in NLP technology, chatbots have become more sophisticated and are regarded as a vital tool for businesses that are seeking to enhance customer service while costs are being lowered. Repetitive tasks like answering FAQs can be efficiently handled by chatbots, allowing more complex issues to be addressed by human agents.

The rest of the paper is structured as follows. In Sect. 2, we briefly outline the related works explaining the Chatbot Technology with BERT model and literature studies on Error

analysis with Chatbots. In Sect. 3, in methodology, discussion on the Response generation system for Conversational agents using BERT model used in the study is provided, while in Sect. 4, experiments and results section, analysis of errors relevant to chatbot technology are described which are encountered in via experiments in point of view of the dataset, error distribution by type of errors. Next, in Sect. 5, discussion on the strengths and limitations of metrics for errors analysis with conversational agents is explained. Finally, Sect. 6 reports conclusions and highlights directions for further research.

# 2. RELATED WORK

## 2.1 Chatbot Technology with BERT

A long way has been come by chatbot technology with the launch of powerful language models like BERT. A transformer language model known as BERT (Bidirectional Encoder Representations from Transformers) is used in transfer learning to learn unsupervised from a corpus on unlabeled data; common model weaknesses like overfitting and underfitting can be overcome by BERT due to the size of the dataset. Because of its substantial pre-training on large textual datasets, understanding of word meanings within their contextual context is enabled by BERT's bidirectional architecture. When a smaller set of labeled data is used to fine-tune BERT, several natural language tasks, such as sentiment analysis and question answering can be performed. ABINAYA et al. [12] in their work state that for text-based emotion analysis, Bidirectional Encoder Representations from Transformers (BERT) is employed to assess user queries and pinpoint emotional cues. BERT is a useful model for linguistically challenging tasks such as intent recognition because of its architecture, which is essentially the same across tasks.

## 2.2 Literature review for Error Analysis with Chatbots

The literature on error analysis with chatbots highlights the growing importance of improving conversational AI systems through advanced evaluation and correction mechanisms. The role of error correction in enhancing chatbot intelligence was emphasized by Izadi and Forouzanfar [1] through the examination of issues such as factual inaccuracies and misinterpretations. Several machine learning approaches, including supervised learning, reinforcement learning, meta-learning, and data-driven feedback systems, were explored in their study, while the importance of human oversight in chatbot training was also stressed. Similarly, intent detection in task-oriented conversational agents was focused on by Jbene et al. [2] through the comparison of Recurrent Neural Networks (RNNs) and Transformer-based models. Chatbot performance was evaluated using metrics such as accuracy, precision, recall, and F1-score, and the need for more advanced methods capable of handling complex and diverse datasets was highlighted. It is demonstrated by these studies that improving chatbot accuracy and adaptability remains a key research focus in conversational AI systems. Another significant area of research in chatbot error analysis is concerned with hallucination detection and human-centered evaluation methods. Semantic entropy–based uncertainty estimation techniques were introduced by Farquhar et al. [3] to detect hallucinations in large language models (LLMs). Uncertainty at the semantic level was measured by their approach rather than being relied on only by word sequences, enabling better identification of confabulated or factually incorrect responses across different datasets and tasks. In addition, Giorgi et al. [4] proposed psychologically grounded metrics for evaluating dialogue systems from a human-centered perspective. Factors such as empathy, emotional entropy, agreeableness, and language matching were examined in their research, and comparisons were made with automatic evaluation metrics like BARTScore and BLEURT. It was shown by the findings that unique conversational qualities not reflected in conventional automated evaluation methods were captured by psychological metrics, thereby improving the prediction of human judgments in open-domain dialogue systems.

Several researchers have also contributed to the development of automatic evaluation metrics for text generation and question-answering systems. A comprehensive survey of question-answering systems was provided by Farea et al. [5] through the analysis of benchmark datasets and evaluation techniques such as AES, MTES, BERTScore, and other automatic scoring methods. Yuan et al. [6] proposed BARTScore, which evaluates generated text using pre-trained sequence-to-sequence models and assesses aspects such as fluency, informativeness, and factuality. Their results demonstrated strong performance across multiple datasets and evaluation conditions. BERTScore was introduced by Zhang et al. [7], which is an automatic evaluation metric based on contextual embeddings from BERT and was shown to have stronger alignment with human judgments compared to traditional metrics. This area was further expanded by Deriu et al. [8] through the surveying of evaluation methods for dialogue systems, including correctness metrics, trained metrics, and fine-grained evaluation approaches. The growing reliance on advanced automated evaluation methods for measuring chatbot response quality and conversational effectiveness is underlined by these studies.

Foundational research on evaluation metrics has also significantly influenced modern chatbot assessment methods. BLEU was introduced by Papineni et al. [10], as a language-independent automatic evaluation metric for machine translation that demonstrated a strong correlation with human judgments using n-gram matching techniques. Later, Lin [9] proposed the ROUGE evaluation package, which became widely adopted for summarization and text evaluation tasks through measures such as ROUGE-N, ROUGE-L, and ROUGE-S. More recent healthcare-oriented studies have applied chatbot evaluation methods to domain-specific applications. The validity and consistency of chatbot responses to prosthodontic patient FAQs were assessed by Gheisarifar et al. [18] using GPT-3.5, GPT-4, Gemini, and Bing, with varying levels of reliability among the systems being found. The ability of AI chatbots to generate single best answer questions for medical education was examined by Abouzeid et al. [20], and technical flaws and inconsistencies across platforms were identified. It is indicated by these studies that while significant advancements have been made in chatbots, extensive research and refinement continue to be required for challenges related to factual accuracy, reliability, and contextual understanding.

In summary of the literature review for error analysis with chatbots response generation, it is seen that numerous metrics are used for testing the quantitative as well as the qualitative assessment of the conversational agent. In this research study exploration of the BERT based variant models for error analysis is done, Section 4 gives a detailed explanation of the different kinds of errors with the practical experimental results. As per literature studies, in comparison of the research study to the non-BERT based models for evaluation of responses from conversational agents, it is seen that automatic evaluation conversational system specific metrics like BLEURT, generation-based evaluation metric like BARTScore, also

traditional metrics like BLEU, Response length (word count) , Flesch reading ease are used along with transformer models like BART, ChatGPT , GPT-3.5, ChatGPT-4 , Gemini, Generative AI and Llama3. Furthermore , it is observed in literature studies that comparative analysis of errors is evaluated using statistical specific metrics like t-test, standard deviation, chi-square test and Anova. In this study, Conversational agent based specific metrics for response generation of quantitative assessment like Accuracy and ROUGE (1/2/L F1) were used, and qualitative assessment metrics like BERTScore (Precision/Recall/F1) which approximates semantic quality were analyzed. These metrics were found suitable for study and hence error analysis with qualitative as well as quantitative assessment is done on the same.

# 3. METHODOLOGY

In the study for response generation with conversational agents, implementation of the BERT model and its variant models is done. Fig 1 gives the general framework used for response generation with the implementation of the BERT model and its variants.

## 3.1 Phases of the Response Generation System for Chatbots

In the Fig. 1. The flow of response generation is described by six phases which is explained as follows:

### *3.1.1 Input Layer*

The question from the user is received by the input layer, and the related text passage or FAQ entry is retrieved as context.

### *3.1.2 Preprocessing Layer*

Sentence splitting and tokenization are performed by the pre-processing layer.

### *3.1.3 Computation Layer*

In the computation layer, the computation is carried out as per the BERT model applied to predict and respond appropriately.

### *3.1.4 Model*

The relevant BERT-based model is applied. Evaluation in this study is done on the BERT model variants of BERT, RoBERTa, DistilBERT and SBERT.

### *3.1.5 Response Generation*

Next, the response text is generated in the response generation phase.

### *3.1.6 Evaluation Metrics Layer*

The model is evaluated using evaluation metrics specific to conversational agents for response generation. Specialized chatbot metrics is used for evaluation like Accuracy, ROUGE-1 F1(Unigram Overlap), ROUGE-2 F1(Bigram Overlap), ROUGE-L F1(Longest Common Subsequence), BERTScore Precision, BERTScore Recall and BERTScore F1.

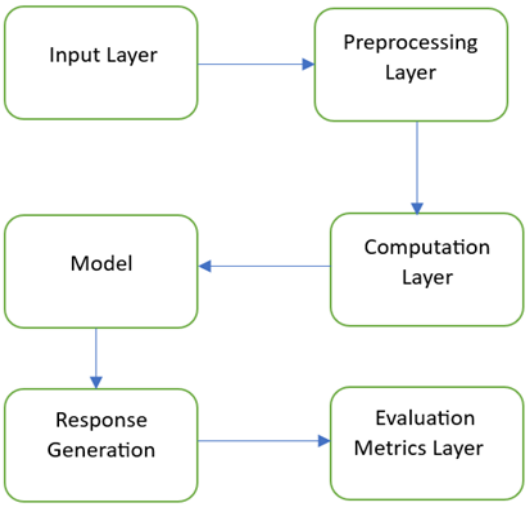


**Fig 1: Response generation system for Conversational agents using BERT model.**

# 4. EXPERIMENTS AND RESULTS

In the research study for response generation using BERT model with conversational agents, detailed experiments with variant of BERT models like BERT (bert-large-uncased-whole-word-masking-finetuned-squad), RoBERTa (deepset/roberta-base-squad2), DistilBERT (distilbert- base-cased-distilled-squad) and SBERT(sentence-transformers/all-mpnet-base-v2) using the diverse performance measures for error analysis of BERT model with conversational system is done. The following section gives a detailed explanation of the same.

## 4.1 Dataset

The dataset which used in the study is a custom dataset which is prepared from the official website of Python documentation. The link is https://docs.python.org/3/faq/. The dataset can be included in a csv file format or as a Custom dataset which includes list of dictionaries containing questions, contexts, and ground-truth answers. The dataset is included in a csv file format which includes fields for questions, con-texts, and ground-truth answers and 445 total questions defined.

## 4.2 Error Analysis due to the Dataset

The following are the errors and inaccuracies analyzed due to constraints of the dataset for response generation.

### *4.2.1 Span Prediction Error*

Incorrect start and end tokens for unseen contexts are predicted.

### *4.2.2 Dataset Imbalance Errors*

If one category of questions is dominated (e.g., “What is…” vs. “How to…”), biased attention weights are learned by the model. When trained on biased data, certain attention patterns are dominated, reducing the ability of the model to generalize

across question types.

## 4.3 Evaluation measures for Chatbot

In the study of response generation with BERT model for conversational agents the metrics of Accuracy, ROUGE and its types - ROUGE-1 F1, ROUGE-2 F1, ROUGE-L F1, BERTScore types - BERTScore Precision, BERTScore Recall and BERTScore F1 are used. Khalid and Lee [11] in their work state that singular measures of improvement are not enough to capture the variability of performance exhibited by the evaluation metrics in judging complex conversation behaviors like contradictions.

Table 1. briefly describes the metrics used.

**Table 1. Metrics used in the study for response generation with Chatbot**

| Metric Name | Description |
|---|---|
| Accuracy | Measures how often the chatbot provides a correct response to a user's query. It Ranges from 0 to 1 (or 0% to 100%) |
| ROUGE(Recall-Oriented Understudy for Gisting Evaluation) | The quantity of word overlap between the generated response and the ground truth is determined by ROUGE. |
| ROUGE-1 F1 | Calculates the overlap of individual words (unigrams) between the chatbot's response and the reference. |
| ROUGE-2 F1 | Calculates the overlap of two consecutive words (bigrams) between the chatbot's response and the reference. |
| ROUGE-L F1 | Emphasizes the longest common subsequence (LCS) between the between the chatbot's response and the reference text. |
| BERTScore | BERTScore is a metric used to assess the quality of text generation by contrasting predicted answers with reference answers using contextual embeddings. It is computed using Precision, Recall, and F1-score based on cosine similarity in the embedding space. |
| BERTScore Precision | Calculates the number of words from the chatbot's response which are semantically similar to the words in the reference text. |
| BERTScore Recall | Measures how much of the reference text is semantically discovered by the chatbot's response. |
| BERTScore F1 | It is the harmonic mean of BERTScore Precision and Recall, balancing how well the chatbot's responses both match the reference and cover the reference's semantic content. |

**Table 2. Analysis of the BERT variant models used in the study**

| Metric Name | BERT (bert-large-uncased-whole-word-masking-finetuned-squad) | RoBERTa (deepset/roberta-base-squad2) | DistilBERT (distilbert-base-cased-distilled-squad) | SBERT (sentence-transformers/all-mpnet-base-v2) |
|---|---|---|---|---|
| Accuracy | 0.2584 | 0.2854 | 0.2270 | **1.0000** |
| ROUGE-1 F1 (Unigram Overlap) | 0.0281 | 0.0275 | 0.0298 | **1.0000** |
| ROUGE-2 F1 (Bigram Overlap) | 0.0058 | 0.0055 | 0.0069 | **1.0000** |
| ROUGE-L F1 (Longest Common Subsequence) | 0.0270 | 0.0267 | 0.0284 | **1.0000** |
| BERTScore Precision | 0.8303 | 0.8297 | 0.8331 | **1.0000** |
| BERTScore Recall | 0.7139 | 0.7137 | 0.7155 | **1.0000** |
| BERTScore F1 | 0.7674 | 0.7671 | 0.7695 | **1.0000** |

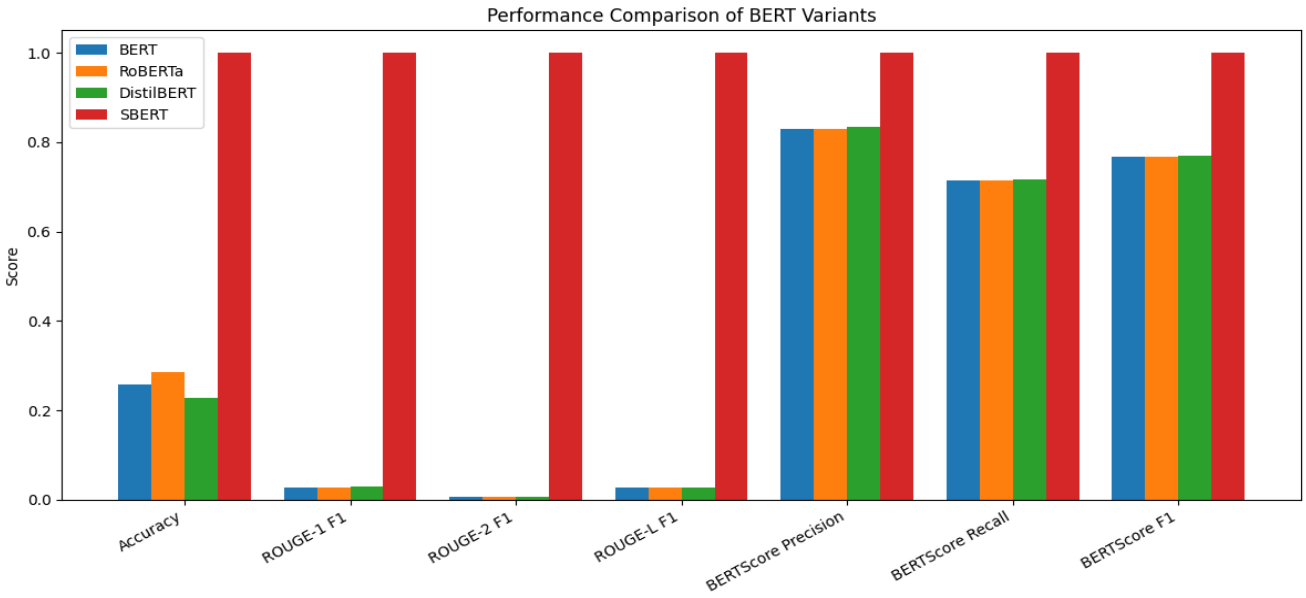


**Fig 2: Comparative analysis of BERT variants.**

In Table 2. Summarization of the quantitative results obtained via the numerous metrics on the BERT variant models is presented. It is noticeable that SBERT model outperformed the other BERT variant models across all metrics of BERTScore (Precision / Recall / F1), ROUGE (ROUGE-1 / ROUGE-2 / ROUGE-L F1) and Accuracy indicating an exact match between predicted and ground-truth answers, likely due to its retrieval-based semantic matching approach. The results conclude that SBERT is effective for FAQ-style question answering.

Fig 2 graphically represents comparative results in terms of the performance of four variants of BERT across the Accuracy, ROUGE, and BERTScore. For all evaluation metrics, SBERT obtained the highest score (1.0) which proves it performs superior among BERT, RoBERTa, and DistilBERT.

The ROUGE metrics (ROUGE-1, ROUGE-2, and ROUGE-L) show relatively low values for BERT, RoBERTa, and DistilBERT, indicating limited lexical overlap with the reference texts. Among these three models, DistilBERT slightly outperforms the others on ROUGE-based measures, indicating better text generation or summarization quality.

## 4.4 The problem of Response Generation in Conversational Agents and Errors encountered

In artificial intelligence and natural language processing, conversational response generation is a technique used to generate responses that appear human in a conversational context. Training of Large Language models is involved to understand and generate meaningful responses to user inputs, including queries or requests. Large datasets of conversational data are used to train these models, allowing for patterns to be learned and contextually appropriate responses to be generated. In this research of the BERT model for conversational agents, the responses are generated but there are responses which are incorrectly generated.

The following section provides a detailed explanation of the Chatbot Errors in BERT-Based Conversational Systems encountered in the research study.

## 4.5 Error Distribution by Type of Error

### *4.5.1 Context Misinterpretation*

This type of error refers to the Semantic Understanding Errors. The semantic relationship between the question and the context is failed to be captured by the model. Table 3 . Presents an example from experimental results for Context Misinterpretation.

**Table 3. Example from results indicating Context Misinterpretation**

| Question | Predicted | Ground Truth | Analysis |
|---|---|---|---|
| "What is the Python Software Foundation ?" | "Python Software Foundatio n" | "The Python Software Foundation is an independent non-profit organization that holds the copyright on Python versions 2.1 and newer. The PSF's mission is to advance open source technology..." | "What is the Python Software Foundation?" |

### *4.5.2 Span Prediction Error*

This refers to the Boundary Misalignment error. An answer span is predicted by the model whose start or end boundaries do not fully capture the correct text fragment. Table 4. Presents an example from experimental results for Span Prediction Error.

**Table 4. Example from results indicating Span Prediction Error**

| Question | Predicted | Ground Truth | Analysis |
|---|---|---|---|
| "How do I program using threads?" | "threads" | "Be sure to use the threading module and not the _thread module. The threading module builds convenient abstractions on top of the low-level primitives provided by the _thread module." | Only a keyword fragment ("threads") was predicted by the model, not the full explanatory span. Boundary truncation is shown, a classical BERT limitation where good alignment of start and end logits does not occur. |

### *4.5.3 Generative Limitation Errors*

It can be Unnatural or Fragmented Responses in which Fluent natural sentences cannot be synthesized by BERT, which is encoder-only. Else it can be Lack of Paraphrasing where answers cannot be reformulated in a conversational tone. Table 5. Presents an example from experimental results for Unnatural or Fragmented Responses.

**Table 5. Example from results indicating Unnatural or Fragmented Responses**

| Question | Predicted | Ground Truth | Analysis |
|---|---|---|---|
| "How do I delete a file?" | "deleting a file" | "Use os.remove(file name) or os.unlink(filen ame)..." | A noun phrase fragment is identified as the prediction, not a complete, fluent response. A lack of generative capability is indicated — rephrasing or elaboration cannot be performed by BERT. |

Table 6. Presents an example from experimental results for Lack of Paraphrasing.

**Table 6. Example from results indicating Lack of Paraphrasing**

| Question | Predicted | Ground Truth | Analysis |
|---|---|---|---|
| "How do I send mail from a Python script?" | "send mail from a Python script" | "Use the standard library module smtplib. Here's a simple interactive mail sender that uses it…" | The question structure is copied by the model instead of a conversational or reworded answer being generated. A lack of paraphrasing ability is reflected — responses cannot be naturally reformulated by BERT. |

### *4.5.4 Inappropriate or Irrelevant Response*

The contextually irrelevant response or the response that does not address the actual question is provided. Table 7. Presents an example from experimental results for Inappropriate or Irrelevant Response.

**Table 7. Example from results indicating Inappropriate or Irrelevant Response**

| Question | Predicted | Ground Truth | Analysis |
|---|---|---|---|
| "How do I access the serial (RS232) port?" | "access the serial (RS232) port" | "For Win32, OSX, Linux, BSD, Jython, IronPython: https://pypi.or g/project/pyse rial/ …" | Part of the input is repeated by the model instead of relevant technical instructions being provided. This occurs when excessive focus is placed on the question tokens by the attention weights, rather than on the context passage — an attention collapse problem. |

### *4.5.5 Repetitive Response Error*

The same generic output pattern is produced by the model across multiple, different inputs. Table 8. Presents an example from experimental results for Repetitive Response Error.

**Table 8. Example from results indicating Repetitive Response Error**

| Question | Predicted | Analysis |
|---|---|---|
| Across multiple questions such as:<br><br>"How do I copy a file?"<br><br>"How do I delete a | All return short, fragmented keyword-based responses like:<br><br>"Copying a file"<br><br>"deleting a file" | The question's key noun phrase is being repeated as the "answer" by the model, indicating pattern overfitting or template bias from fine-tuning data. |

| | | |
|---|---|---|
| file?"<br><br>"How do I send mail from a Python script?" | "send mail from a Python script" | |

### *4.5.6 Hallucination Errors*

Fabricated data is used by an AI model to fill in knowledge gaps. Table 9. Presents an example from experimental results for Hallucination Errors.

**Table 9. Example from results indicating Hallucination Errors**

| Question | Predicted | Ground Truth | Analysis |
|---|---|---|---|
| "What is Python?" | "Python is a programming language and it has many features." | A very long official definition detailing object-oriented, interpreted nature, and PSF references. | While the output sounds plausible, factual details from the actual context (e.g., licensing, paradigms, PSF) are omitted. This is qualified as a mild hallucination — a hallucinated summary rather than an extraction. |

## 5. DISCUSSION

In this section discussion on the strengths and weakness of each of the metrics used in the study for evaluation and error analysis for a python FAQ chatbot using conversational agents with the BERT model is done. Table 10. states the strengths and weakness of each of the metrics used in the study.

**Table 10. Strengths and Weakness of the Metrics used in Study**

| Name | Description | Strengths | Weaknesses |
|---|---|---|---|
| Accuracy (Exact Match / Relaxed-match) | Generally used for extractive QA and checks for the correctness of the response. | Rajpurkar et al. [14] in their work state that it is Simple, interpretable, and easy to compute — good for extractive QA tasks where an exact span is required. When answers are short and deterministic (e.g., factual slots), Exact Match (EM) is meaningful and can be directly interpreted. | Liu et al. [15] in their work state that It has low correlation to human judgment for free-form replies. Open-ended dialogue and long multi-sentence answers are not suitable for it (as partial correctness is rarely captured). |
| ROUGE (ROUGE-1 / ROUGE-2 / ROUGE-L F1) | It can be used for Content-overlap checks (e.g., comparing a generated answer against a factual reference) and combined with embedding-based metrics for semantic assessment. | Lin [9] in his work state that Sentence-level sequence similarity can be better reflected by ROUGE-L (LCS) than by simple n-gram counts. It works well when most of the important content is captured by the reference (recall-focused). | Novikova et al. [16] in their work have explained that Semantic equivalence (paraphrases, synonyms) is missed by ROUGE, which is lexical. For dialogue/QA, good paraphrased answers can be misrepresented by it. |
| BERTScore (Precision / Recall / F1) | It is best Used as a primary automatic metric for semantic adequacy in generative QA and dialog, always complemented by factuality checks or human evaluation for high-stakes domains. | BERTScore *Precision* helps detect irrelevant verbosity as it indicates how much of the generated text is supported by the reference.<br><br>BERTScore *Recall is* useful when references are comprehensive as it indicates how much of the reference is covered.<br><br>BERTScore *F1* balances the two i.e. | The task determines interpretation: terse but correct interpretations might be produced with high precision and low recall, while verbose interpretations might result from high recall and low precision. Human judgment is still required to decide which tradeoff is better. Hanna and Bojar [17] present a Fine-Grained Analysis of BERTScore and state that BERTScore fails to assign low scores when a bad candidate sentence has |

| | | | |
|---|---|---|---|
| | | precision and recall. It is useful for comparing models across trade-offs. Zhang et al. [7] in their work state that BERTSCORE exhibits significantly higher performance compared to the other metrics in context of Machine Translation, Image Captioning and speed. | high lexical overlap with the reference in terms of content words. |

In summary, of the metrics used in the study for conversational agents for response generation, it can be concluded that BERTScore based metrics for Precision, Recall and F1 score gave good scores to test the responses in context of irrelevancy, similarity and semantic understandings of the responses generated. ROUGE based metrics check the content overlaps but need improvement in results for the study as per the scores. Accuracy is good to prove the correctness of the responses but as per the study it needs improvement in the metric scores for error analysis. In general, a combined evaluation for error analysis for response generation of conversational agent based on BERTScore (Precision / Recall / F1), ROUGE (ROUGE-1 / ROUGE-2 / ROUGE-L F1) and Accuracy can be used to test the overall performance.

# 6. CONCLUSION

In concluding observations, as per the study in context of conversational agents with the BERT model for error analysis, it can be stated that no single metric is considered sufficient for conversational agents. It is recommended by surveys and comparative studies that lexical (ROUGE), contextual (BERTScore), and human judgments of relevance, fluency, and appropriateness be combined. It has been shown by the extensive experiments that better correlation is achieved by BERTSCORE than by common metrics, and that it is effective for model selection. Nevertheless, no BERTSCORE configuration is proven to perform noticeably better than any other. As there are some responses which are not generated appropriately and as per the metrics evaluation for error analysis, the enhancement of decision making of the Chatbot is needed. Considering future research, the studies will be leading to consider an appropriate approach using large language model for Conversational systems or potentially an hybrid approach to enhance the decision making of Chatbots using Intelligent Conversational Agents for response generation.

# 7. REFERENCES

[1] Saadat Izadi and Mohamad Forouzanfar,” Error Correction and Adaptation in Conversational AI: A Review of Techniques and Applications in Chatbots”. AI 2024, 5, 803–841. https://doi.org/10.3390/ai5020041 (2024).

[2] Mourad Jbene, Abdellah Chehri, Rachid Saadane, Smail Tigani, Gwanggil Jeon,” Intent detection for task-oriented conversational agents:A comparative study of recurrent neural networks and Transformer models”,Expert Systems.2025;42:e13712. https://doi.org/10.1111/exsy.13712 (2024).

[3] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, Yarin Gal,”Detecting hallucinations in large language models using semantic entropy”, Nature. Vol 630. https://doi.org/10.1038/s41586-024-07421-0 (2024).

[4] Salvatore Giorgi, Shreya Havaldar, Farhan Ahmed, Zuhaib Akhtar, Shalaka Vaidya, Gary Pan, Lyle H. Ungar, H. Andrew Schwartz, João Sedoc,”HUMAN-CENTERED METRICS FOR DIALOG SYSTEM EVALUATION”,DOI:10.48550/arXiv.2305.14757 (2023).

[5] Amer Farea, Zhen Yang, Kien Duong, Nadeesha Perera, Frank Emmert-Streib,” Evaluation of Question Answering Systems Complexity of judging a natural language”. ACM Computing Surveys, Volume 58, Issue 1 Article No.: 1, Pages 1 – 43. https://doi.org/10.1145/3744663 (2021).

[6] Weizhe Yuan, Graham Neubig, Pengfei Liu,” BARTSCORE: Evaluating Generated Text as Text Generation”,35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia. (2021).

[7] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi,” BERTSCORE: EVALUATING TEXT GENERATION WITH BERT”, ICLR 2020 (2020).

[8] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre and Mark Cieliebak,”Survey on evaluation methods for dialogue systems”, Artificial Intelligence Review (2021) 54:755–810. https://doi.org/10.1007/s10462-020-09866-x . Springer (2021).

[9] Chin-Yew Lin,” ROUGE: A Package for Automatic Evaluation of Summaries”, Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics. (2004).

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu,” BLEU: a Method for Automatic Evaluation of Machine Translation”, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318. (2002).

[11] Baber Khalid and Sungjin Lee,”Explaining Dialogue Evaluation Metrics using Adversarial Behavioral Analysis”, Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5871- 5883 July 10-15, 2022 ©2022 Association for Computational Linguistics (2022)

[12] S. ABINAYA, K. S. ASHWIN, A. SHERLY ALPHONSE,” Enhanced Emotion-Aware Conversational

Agent: Analyzing User Behavioral Status for Tailored Reponses in Chatbot Interactions", VOLUME 13, IEEE Access (2025)

[13] Dipak Mandlik, Roshan Chaudhary, Mayur Kotkar, Rushikesh Zende, Dr. R. S. Bhosale," AI-Powered College Enquiry Chatbot Using NLP with BERT and GPT", IJIRMPS, ISSN: 2349-7300 March - April 2025 Volume 13 Issue 2 (2025)

[14] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang," SQuAD: 100,000+ Questions for Machine Comprehension of Text", Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2016)

[15] Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, Joelle Pineau,"How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation", EMNLP (2016)

[16] Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, Verena Rieser," Why We Need New Evaluation Metrics for NLG", Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2231-2242, Copenhagen, Denmark, September 7-11, (2017)

[17] Michael Hanna, Ondrej Bojar ,"A Fine-Grained Analysis of BERTScore", *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics. (2021)

[18] Maryam Gheisarifar, Marwa Shembesh, Merve Koseoglu, Qiao Fang, Fatemeh Solmaz Afshari, Judy Chia-Chun Yuan, Cortino Sukotjo," Evaluating the validity and consistency of artificial intelligence chatbots in responding to patients' frequently asked questions in prosthodontics", THE JOURNAL OF PROSTHETIC DENTISTRY, Volume 134 Issue 1,(2025)

[19] Bertilla Fernandes, Snehalata B. Shirude,"Intent Classification and Response Generation of Conversational Agents: A Literature Review", In: Bansal, J.C., Saha, S., Coello, C.A.C., Rathore, H. (eds) Advances in Data-driven Computing and Intelligent Systems. ADCIS 2024. Lecture Notes in Networks and Systems, vol 1304. Springer, Singapore. https://doi.org/10.1007/978-981-96-3652-5_3,(2025)

[20] Enjy Abouzeid, Rita Wassef, Ayesha Jawwad, Patricia Harris," Chatbots' Role in Generating Single Best Answer Questions for Undergraduate Medical Student Assessment: Comparative Analysis", JMIR Med Educ 2025;11:e69521; doi: Abouzeid et al 10.2196/69521(2025)