# Hybrid Machine Learning Approach for Task-Oriented Dialog Systems

Ganesh Reddy Gunnam
The University of Texas at San Antonio
San Antonio, TX 78249 USA

Devasena Inupakutika
The University of Texas at San Antonio
San Antonio, TX 78249 USA

Rahul Mundlamuri
The University of Texas at San Antonio
San Antonio, TX 78249 USA

Sahak Kaghyan
The University of Texas at San Antonio
San Antonio, TX 78249 USA

David Akopian
The University of Texas at San Antonio
San Antonio, TX 78249 USA

## ABSTRACT
Nowadays, automated chatbots are commonly used since they easily provide essential information. While generic chatbots are essential for open-domain dialog, specific applications are better served with task-oriented dialog systems. These task-oriented dialog systems typically solve particular tasks in the application where the chatbot and user know what they are discussing (both sides know the scope and context of the conversation topic). The majority of these chatbots work based on keywords. Keyword extraction has been a well-established field in the natural language processing (NLP) domain for quite some time. It is crucial in various applications, such as information retrieval, search engine optimization, and content summarization. Recently, there has been a growing interest in the contextual recognition of keywords, which aims to identify keywords in a given text based on their contextual relevance. Additionally, integrating Large Language Models (LLMs) with intent prediction (IP) has opened new possibilities for interpreting and utilizing keywords in a more context-aware manner. In particular, one such LLM, BERT, a SQuAD dataset-based NLP model, has become a popular question-answer set. However, task-oriented systems still challenge specific questions, such as yes/no and synonym-based inquiries. Thus, a hybrid model involving LLMs and IP merits additional study. This paper explores the intersection of keyword extraction, LLMs, and Intent Prediction in the context of protocol-driven chatbots, particularly those designed for task-oriented applications, emphasizing their potential in addressing a niche application. Specifically, this paper presents a hybrid approach (TaskBERT) that addresses these challenges. The evaluation results demonstrate that TaskBERT outperforms Google Dialogflow and the performant keyword extraction tool KeyBERT.

## Keywords
artificial intelligence, natural language processing, closed domain chatbot, intent prediction

## 1. INTRODUCTION & BACKGROUND
Dialog machine-based chatbots are computational systems or artificial intelligence models designed to converse with humans or other machines through natural language. These chatbots are divided into open-domain and task-oriented dialog systems [1]. The open-domain interaction is more like talking to a friend, where the conversation can go in any direction. This free-flowing conversation has no defined objective, so the responses must adapt to whatever information the user asks [2]. The open-domain-based dialog systems were developed intensively [3], but it is more convenient to apply task-oriented chatbots when such a system is necessary for a specific conversation. Task-oriented dialog systems are intended to solve particular tasks in the application where the chatbot and user know what they are discussing. The majority of these chatbots work based on keywords. Keyword extraction has been a well-established field in natural language processing for quite some time. This approach is crucial in various applications, such as information retrieval, search engine optimization, and content summarization. Recently, there has been a growing interest in the contextual recognition of keywords, which aims to identify keywords in a given text based on their contextual relevance.

Keyword extraction tools are software applications or algorithms that automatically identify and extract significant keywords or key phrases from a given text or document [4]. These tools help summarize and categorize content and improve search engine optimization, information retrieval, content analysis, and various natural language processing tasks. A typical keyword extraction workflow (Figure 1) is as follows: The input text is preprocessed, including tasks like lowercasing, stemming, and removing stop words to clean the text, and is divided into words, phrases/tokens to identify units for analysis. Keywords are scored based on various metrics, such as term frequency-inverse document frequency (TF-IDF) [5], Text Rank [6], or other statistical measures. BERT embeddings may also be used in scoring. The extracted keywords are ranked based on their scores, and a threshold is often applied to select the top keywords. Additional steps like filtering out low-quality keywords or resolving synonyms can be part of the post-processing stage.

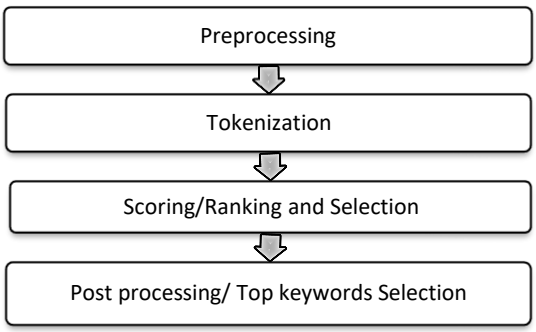


**Fig 1: keyword Extraction Workflow.**

Keyword extraction tools are widely used in task-oriented dialog systems. Since task-oriented dialog systems typically work on predefined keywords, these extraction tools sometimes fail to provide keywords for complex scenarios such as synonyms [7]- and yes/no-based questions [8].

To solve such complex scenarios, intent classification models can be used [9]. Intent classification and prediction models are designed to determine the user intention from the predefined intent categories. Traditionally, intent classification relied on keywords and grammar, but deep learning techniques like CNNs, LSTMs, and attention-based CNNs have recently become popular for intent recognition. These methods require substantial labeled data for high performance. At the same time, intent classification and slot filling are often separate. There are ongoing efforts to combine them. There's also a BERT-based model for joint intent and slot classification. However, commercial tools like Dialogflow [10], LUIS [11], and Amazon Lex [12] offer convenient solutions, but there are often use cases where more flexibility in customizing models is needed. Amazon Skills Kit is limited to the Alexa ecosystem for custom intents and training examples. Google Dialogflow intent matching flow uses a training phase for each intent [10]. Since task-oriented dialog systems will use custom datasets, sometimes these intent prediction models alone won't be sufficient to get an accurate response, especially for certain types of questions such as synonyms and yes/no. The advantage of BERT is that it was trained on Wikipedia and Book Corpus [9]. Fucheng et al. used a fine-tuned BERT model for keyword extraction to take advantage of transfer learning from the robust architecture of the pre-trained BERT model [13]. Thus, one can take advantage of BERT's training of Deep Bidirectional Transformers with intent prediction merged as a hybrid model for improving prediction accuracy.

The remainder of this paper is organized as follows. Section 2 briefly covers the state-of-the-art approaches to this topic and discusses the hybrid model approach, whereas Section 3 provides the methodology and experimental settings. In sections 4 and 5, the performance results and discussion are presented, and concluding remarks are provided in the following section.

# 2. STATE-OF-THE-ART

## 2.1 Keyword Extraction Tools

Keyword extraction tools [14] have been well-known and utilized for some time. These tools typically identify and extract essential words or phrases from a text, aiding content analysis and information retrieval. The output of a keyword extraction model generally is a list of keywords or phrases that represent the essential elements of the text. These keywords are not intent labels but rather important terms or concepts within the text. Keyword extraction models are used in various applications, such as content summarization, information retrieval, search engine optimization, and document categorization. They help in understanding the content of a document or text without categorizing it into predefined intents. Keyword extraction models can use unsupervised or semi-supervised learning techniques. They may not require explicit training data with labeled keywords but can use statistical or linguistic patterns to identify significant terms. These methods can vary widely, from simple rule-based approaches to more advanced algorithms like Term Frequency-Inverse Document Frequency (TF-IDF) [5], Text Rank [6], or graph-based techniques [15]. However, the evolving landscape of NLP has led to the development of more sophisticated keyword recognition techniques, often incorporating contextual understanding. This contextual recognition is particularly relevant to this proposed model, as it allows chatbots to understand better and respond more efficiently to user queries in a specific context.

One such keyword extraction tool that is widely used and performant is KeyBERT [16]. KeyBERT (Figure 3) is a minimal and easy-to-use keyword extraction technique that leverages BERT embeddings to create keywords and phrases most similar to a document. In this model, keyword extraction involves identifying sub-phrases within a document that closely resemble the document as a whole. Firstly, generate document embeddings from the user document using the BERT model. Subsequently, word embeddings are derived for N-gram words and phrases, forming a phrase-level representation. Lastly, cosine similarity is employed to pinpoint the words and phrases with the highest similarity to the document [16].

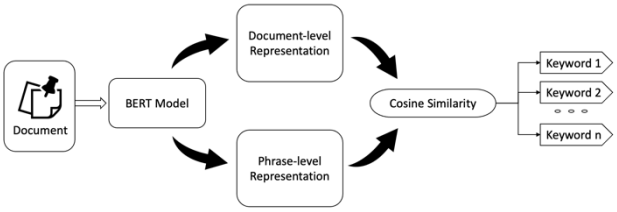


**Fig 2: KeyBERT Model Workflow [16].**

## 2.2 Large Language Models (LLMs)

Conversely, Large Language Models [17] have demonstrated remarkable capabilities in understanding and generating human-like text. The LLMs like BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), and T5 (Text-to-Text Transfer Transformer) [18] have continued to evolve. These models became popular in dialog systems with training in various topics and large datasets. The core benefits offered by the LLMs include versatility, support processing with multiple languages, language translation, transfer learning, and so on. However, they often fail to provide answers in the expected formats, especially when identifying and interpreting keywords for task-oriented dialog systems.

Recent advancements in cloud-based chatbot frameworks have seen the integration of Large Language Models (LLMs) using a hybrid approach. While the proposed method shares few similarities with these frameworks, several key differences set it apart.

The proposed method combines TaskBERT, a task-oriented language model, with Intent Prediction to enhance the understanding of user queries and generate contextually relevant responses. Unlike some cloud-based frameworks that rely solely on pre-trained LLMs, the proposed hybrid model leverages specialized task-oriented capabilities and intent prediction, offering a more tailored approach to dialogue management. Unlike some cloud-based frameworks that may offer limited customization options, the proposed method provides flexibility in model selection, training data, and fine-tuning strategies. This customization enables to adaptation of the chatbot's behavior to specific domains or applications, catering to diverse user needs and preferences.

While many cloud-based frameworks aim to facilitate general-purpose conversational interactions, the proposed method targets task-oriented dialogue scenarios. By incorporating domain-specific knowledge and intent prediction capabilities, the proposed approach guides users through structured tasks and facilitates efficient information exchange. The proposed method allows seamless integration with external tools and services, enabling access to additional functionalities such as language translation, sentiment analysis, or database querying. This integration extends the chatbot's capabilities beyond natural language understanding and generation, enhancing its utility in real-world applications.

The decision to integrate both TaskBERT and Intent Prediction in this hybrid method stems from the complementary nature of these components and their ability to address different aspects of the task-oriented dialogue challenge.

TaskBERT, a pre-trained language model fine-tuned on task-oriented dialogue data, excels in capturing nuanced contextual information and semantic representations relevant to the dialogue task. Intent Prediction specializes in classifying user intents based on input utterances. By training on labeled intent data, Intent Prediction can accurately identify the underlying purpose or goal behind user queries, enabling the chatbot to generate appropriate responses tailored to the user's needs. TaskBERT and Intent Prediction allows us to effectively leverage both components' strengths. While TaskBERT provides rich contextual embeddings that capture the semantic nuances of user queries, Intent Prediction complements this by focusing on intent classification, ensuring that the chatbot's responses align with the user's goals.

Simultaneously using TaskBERT and Intent Prediction enhances the robustness and generalization capabilities of the chatbot. TaskBERT's contextual understanding enables the chatbot to handle complex dialogue scenarios and adapt to diverse domains. At the same time, Intent Prediction ensures accurate intent classification, even in the presence of noise or ambiguity in user queries. By integrating TaskBERT and Intent Prediction, the proposed hybrid method achieves a synergistic effect, resulting in improved performance compared to using either component individually. The combined approach enables the chatbot to understand user intents more accurately, generate contextually relevant responses, and provide a seamless conversational experience in task-oriented dialogue systems.

# 3. METHODOLOGY

Under commercial tools, the functionality of Google Dialogflow is similar to the "keyword extraction tool + intent prediction" as a complete chatbot framework. Dialogflow [10] categorizes end-user inputs into the relevant intent. An intent is the verb or action part of the conversation with the user. In each agent, a user can define multiple intents, with the combined intents controlling the flow of a conversation. It can be specified how to respond to the user for each intent based on the use case with training phrases, which are the set of sample utterances/similar words. Google Dialogflow uses these phrases to train built-in machine learning algorithms for intent classification. This work utilizes four intents: "Male", "Female", "Yes", and "No". Once the intent is matched with the user input, the action field triggers the logic in the workflow, and the corresponding response will be displayed to the user [10]. Figure 3. shows Google Dialogflow intent matching architecture.

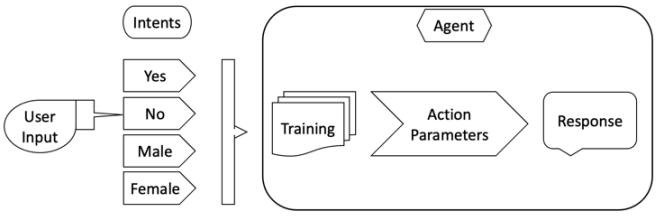


**Fig 3: Google Dialogflow Intent Matching Workflow [10].**

Since task-oriented dialog systems are particular towards applications and have limited knowledge, the chatbot and user should know what they expect from each other. Instead of enhancing keyword extraction tools for task-oriented dialog systems, this paper approaches a different method to add an extra layer to LLMs to be usable as a keyword extraction tool. Combining LLMs with intent prediction models [19] is exciting. This hybrid approach can effectively convert LLMs into keyword extraction tools. The model can accurately identify keywords within the context by predicting the user's intent.

The IP model will determine the underlying intention or purpose behind a given input text or utterance and classify input text into predefined categories or intents. The intent prediction can be used to understand the user's input request, such as weather information, a question about a product, or a complaint. This model typically provides a single intent label, representing the most likely intention behind the input text. These labels are predefined based on the expected actions or responses the system can take and are usually trained in supervised learning. The main difference between the IP model in the hybrid approach and modern techniques is the way the model is trained to solve certain types of questions, such as synonym-based, yes/no type questions, etc. The complete details will be discussed in the experimental setup section.

In the proposed niche application, these protocol-driven or task-oriented chatbots operate within a closed domain, often following specific protocols or guidelines. The hybrid and traditional keyword extraction techniques compete for relevance in this scenario. However, this competition may present an exciting opportunity for a new problem formulation. Leveraging combined LLM and IP for keyword recognition in the context of protocol-driven chatbots could be valuable and potentially lead to innovative solutions. Fine-tuning of LLM models has been a hot topic in recent years. The proposed approach serves as a form of fine-tuning model with a distinct purpose. Instead of fine-tuning LLM models, the current approach adds an extra interpretation layer to accommodate protocol-driven chatbots' minimal vocabulary and specific expectations. This customization ensures that the chatbot can understand and respond to user queries in a manner consistent with its predefined protocol, enhancing its overall performance and usability.

The BERT-based model combines intent classification and slot filling into a single token classification task [20]. The text also touches on commercial intent recognition tools like Google's Dialogflow, Microsoft's LUIS, and Amazon Lex, allowing users to create custom intents and upload example utterances with limited customization options [20]. The Amazon Skills Kit, which is limited to Amazon's Alexa ecosystem, is also mentioned, allowing the creation of custom intents and providing various training examples [21].

The SQuAD [22] dataset is a publicly available research dataset published by Stanford University for questions and answers from Wikipedia and is manually composed and annotated by an external crowdsource, where the answer is an executive text span in a Wikipedia paragraph [23]. These SQuAD dataset-based NLP models perform well in open-domain and task-oriented chatbots except for certain types of questions, such as yes-or-no and synonym-based. The hybrid approach proposed in this paper is discussed further in Section III, Figure 4.

The Performance Challenges in Task-Oriented Dialog Machines are as follows.

- On the Reasoning Ability: The products related to a dialog system that can be seen on the market give people a feeling of not being smart. The main reason is that the existing dialog system cannot reason like human beings, which is the most critical factor restricting the development of the dialog system to a higher level of intelligence [3].
- On yes/no type Questions: In their dialog systems, a few questions do not include a simple linguistic expression corresponding to ‘‘yes’’, so it is not easy to recognize its intention [1].
- On synonym-based Questions: The existing ML models in these dialog systems could answer questions based on context, but the dialog systems require a specific answer to reply to the user, for example, a synonym of a particular word, such as female for woman. Google Dialogflow [10] has this feature since it is a commercial tool, and this can be utilized in the proposed task-oriented dialog systems.

The primary research problem addressed in this study is the development of an effective task-oriented dialogue system that can accurately understand user intents and provide contextually relevant responses in real-time interactions. This problem encompasses challenges related to natural language understanding, context modeling, and response generation in dynamic conversational environments.

1. To propose a hybrid framework**:** To introduce a hybrid framework that integrates LLMs with intent prediction mechanisms to improve the accuracy and effectiveness of task-oriented dialogue systems.
2. To evaluate the performance**:** To empirically evaluate the performance of the proposed framework on benchmark datasets and compare it with existing approaches, such as Google Dialogflow and KeyBERT.
3. To demonstrate applicability**:** To demonstrate the applicability of the proposed approach across various domains and dialogue scenarios, showcasing its versatility and effectiveness in real-world settings.

## 3.1 BERT MODEL WITH SQUAD DATASET

The trained model is deployed in the cloud and exposed as an application programming interface (API) endpoint. This API will take questions, evaluate them based on the context, and give the appropriate answer. A customized script was used to send API requests to the natural-language understanding (NLU) service. A pre-trained language model based on the Bidirectional encoder representation for transformers (BERT) was uploaded as an Amazon Web Services (AWS) S3 object.

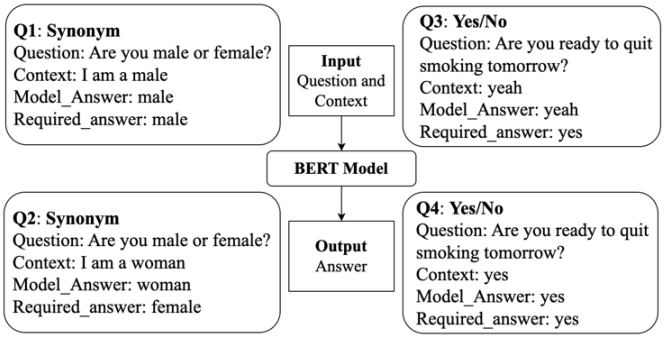


**Fig 4: BERT model Sample Output.**

This BERT model can accept questions and context as input data and provide answers as output, as shown in Figure 8. The BERT model is fine-tuned with the custom task-oriented related question-answer-based dataset. This model works well except for specific questions, such as synonym-based and yes/no-type questions. Figure 8, Q2 shows a synonym-based question. In Q2, for the question “Are you male or female?” the expected required answer to be “female” or “male,” but this model gave “woman”. In Figure 8, Q3 and Q4 are yes/no type questions except “yes” or “no.” In Q3, the model expected “yes” but gave “yeah.”

## 3.2 NLU MODEL WITH INTENT PREDICTION

In this setup, the NLU service is integrated with chatbot infrastructure deployed in the cloud. An intent prediction-based dataset was used, and this model takes the context as input and the answer as output, as shown in Figure 9. In both Q1 and Q2, the required response is matched with the model output.

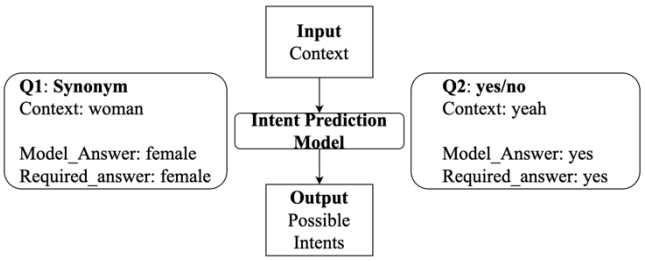


**Fig 5: An intent-prediction sample output.**

## 3.3 HYBRID MODEL

The SQuAD model is combined with the Intent Prediction model in this approach. The output of the fine-tuned BERT model is fed as input to the intent prediction (IP) model. With this approach, better accuracy can be achieved by addressing synonym-based questions and yes/no-type questions. Figure 10, Q2 showed a synonym-based question processing approach that takes a specific question, “Are you male or female?” and context, “I am a woman” and gives “woman” as a BERT model output, which is fed to the Intent Prediction model and gets “female” as a final output which is expected outcome. Similarly, Q3 and Q4 also provide a result that is the same as the desired output.

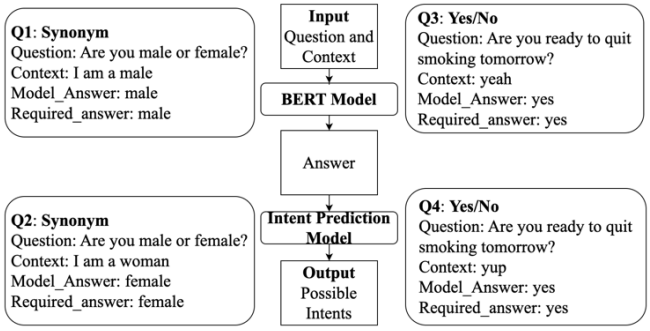


**Fig 6: Hybrid Model Sample Output.**

The performance of each approach is evaluated on a range of accuracy metrics. Statistical analysis and significance testing were conducted to validate the observed differences.

## 4. EXPERIMENTAL SETUP

In this section, the experimental setup of the proposed hybrid model is discussed. Firstly, NLP API is implemented for the chatbot application and deployed in the AWS cloud. In this NLP model, the SQuAD and BERT models are utilized by fine-tuning custom task-oriented datasets. This fine-tuning step is optional since it was initially used to solve specific questions, such as synonym-based and yes/no-type questions. However, this is not ideal for the proposed task-oriented dialog systems. Thus, there arose the need for the implementation of an intent prediction model with keywords/synonyms generated from open-source websites such as Thesaurus [24-25], which is one of the most well-known online thesauruses that provides synonyms/intents and antonyms for words and is a feasible resource for finding similar words. The second tool used was ChatGPT [26], which is related to the keyword processing associated with the proposed chatbot system. In this work, a continuous bag of words (CBOW) model is used. Finally, these two models are combined to solve the above question types. The experimental setup is shown in Figure 4.

Testing the task-oriented dialog systems is still challenging because they work on a specific topic. This work was intended to evaluate the model based on internal, unseen testing data. Wanting et al. used 12 questions to assess their task-oriented dialog system [27]. In this work, the focus was on two questions and, as a result of input processing through the recognition model, four answers (2 pairs of answers) were expected. While more questions can be added for assessing the dialog system, the chosen two questions are representative of a broad set of yes/ no and synonym-based examples. Furthermore, any further questions added for evaluation would result in similar responses as covered by the two questions in this paper.

The target questions were:

Q1: Are you Male or Female?
A1: Male/Female

Q2: Are you ready to quit smoking tomorrow?
A2: Yes/No

Each question is repeated 10 times in the custom test dataset, totaling 20 questions with unseen context. First, this dataset was tested with a BERT model trained on the SQuAD dataset, followed by intent prediction alone. Finally, these two models are combined, the output of the BERT model is provided to the input of the intent prediction model, and this experiment was repeated 100 times each. The following subsections cover the sample examples for datasets utilized for evaluation in this paper.

### 4.1 SQUAD SAMPLE DATASET

The SQuAD [28] dataset contains the columns of title, context, question, and answers. In the answer section, the dataset provides the answer and span of the response from the context. The sample dataset record is shown below.

{
"**id**": "5733be284776f41900661182",
"**title**": "University_of_Notre_Dame",
"**context**": "Architecturally, the school has a Catholic character. Atop the …...",
"**question**": "To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?",
"**answers**": {
"**text**": [Saint Bernadette Soubirous"],
"**Answer start**": [515]
}
}

### 4.2 INTENT PREDICTION SAMPLE DATASET

The intent prediction model dataset contains tags, patterns, responses, and context. The patterns section will train the model with different combinations of the answer and match to the tag. The sample dataset record is shown below.

{
"**tag**": "yes",
"**patterns**": ["yes", "absolutely", "yeah", "sure"],
"**responses**": ["yes"],
"**context**": "Did you go through the past 24 hours without smoking cigarettes?"
}

Finally, the test dataset is evaluated with the KeyBERT model. KeyBERT is an approach for extracting keywords from text documents by leveraging BERT embeddings to identify the most significant terms within the content. This method operates unsupervised and involves three sequential steps: Candidate Keywords or Key phrases, BERT Embedding, and Similarity measurement [29].

## 5. RESULTS

A statistical analysis was conducted using appropriate tests such as t-tests or ANOVA to assess the significance of observed differences in performance metrics. This analysis helped quantify the level of improvement offered by TaskBERT over existing approaches and establish its statistical significance.

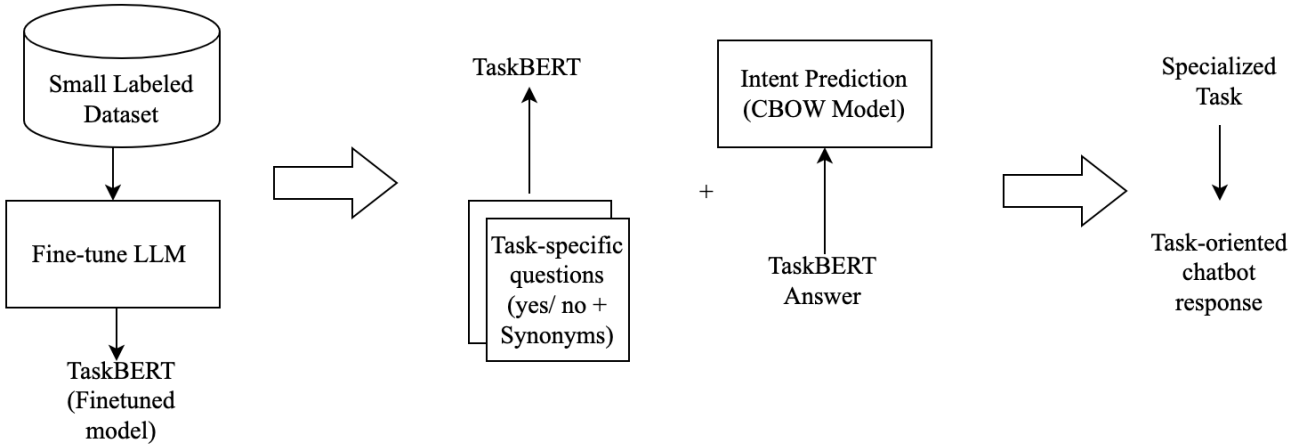


**Fig 7: Hybrid Model Experimental Setup.**

## 5.1 HYBRID MODEL

This section presents the results of the experimental setup discussed above. This testing involved verification and validation of a set of user queries covering the scope of the chatbot. A user query list of 20 test cases will be tested by comparing its actual response tag list to the expected response tag list, and using two questions in this paper, as mentioned before.

**Table 1: Accuracy of Hybrid Model with Custom Test Dataset.**

| Intent Count | BERT Model | Intent Prediction Model | Hybrid Model |
|---|---|---|---|
| 10 | 19.03 | 60 | 72.4 |
| 20 | 19.03 | 65 | 85.25 |
| 30 | 19.03 | 60 | 85.5 |
| **40** | **19.03** | **65** | **92.4** |
| 50 | 19.03 | 65 | 89.3 |

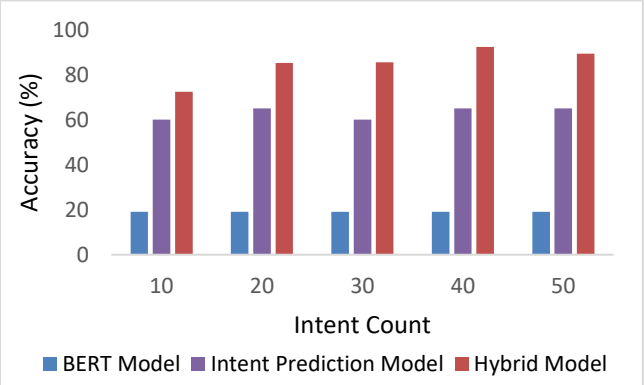


**Fig 8: Accuracy of Hybrid model and comparison with BERT and Intent Prediction model.**

Firstly, the top 10 synonyms/intents were generated for the four keywords used in this experiment, namely "Yes", "No", "Male", and "Female". The custom intent prediction model was trained with these keywords and tested with a test dataset in the BERT, intent prediction, and hybrid models. The experiments resulted in 19 % accuracy in the BERT model, 60% in the Intent model, and 72.4% in the Hybrid model. With the increase in intent count to 20, the accuracy increased to 65% in the Intent model and 85.25% in the Hybrid model. With the top 30 intents, 60% in the Intent model and 85.5% in the Hybrid model are achieved. Later, with the top 40 intents, the accuracy was 65% with the Intent model and 92.4% with the Hybrid model. Lastly, the experiments were performed with the top 50 intents. The accuracy was observed to be 65% with the Intent model and 89% with the Hybrid model, as shown in Table 1. Overall, the top 40 intents resulted in better accuracy, Additionally, with 50 intents, the accuracy decreased because of the correlation between those keywords, as shown in Table 1 and Figure 5.

## 5.2 COMPARISON WITH THE KEYBERT MODEL

With the same test dataset, the tests were performed with the keyBERT and 12 different pre-trained models [30] that achieved 10% accuracy for all the models. In this testing, the intent count doesn't affect the KeyBERT model accuracy because there is no possibility to fine-tune this model with similar words [16], as shown in Table 2 and Figure 6.

**Table 2: Accuracy of comparison of KeyBERT model and Hybrid Model.**

| Intent Count | KeyBERT Model | Hybrid Model |
|---|---|---|
| 10 | 10 | 72.4 |
| 20 | 10 | 85.25 |
| 30 | 10 | 85.5 |
| **40** | **10** | **92.4** |
| 50 | 10 | 89.3 |

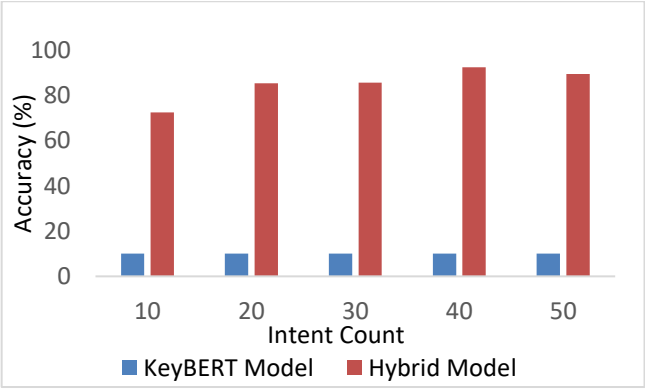


**Fig 9: Accuracy Comparison of KeyBERT model and Hybrid model.**

## 5.3 COMPARISON WITH THE DIALOGFLOW

**Table 3: Accuracy of comparison of Google Dialogflow and Hybrid Model.**

| Intent Count | Dialogflow | Hybrid Model |
|---|---|---|
| 10 | 70 | 72.4 |
| 20 | 80 | 85.25 |
| 30 | 80 | 85.5 |
| **40** | **80** | **92.4** |
| 50 | 80 | 89.3 |

As mentioned, the models were trained with ten intents and achieved 70% accuracy, approximately the same as the Hybrid model. The intent count was increased to 20, increasing accuracy to 80%. Further, the increase of intents to 30, 40, and 50 did not exhibit any improvement in accuracy. This shows at 20 intents, the Dialogflow gets saturated, whereas the hybrid model got saturated at 40 intents and achieved 92.4% accuracy, as shown in Table 3 and Figure 7.

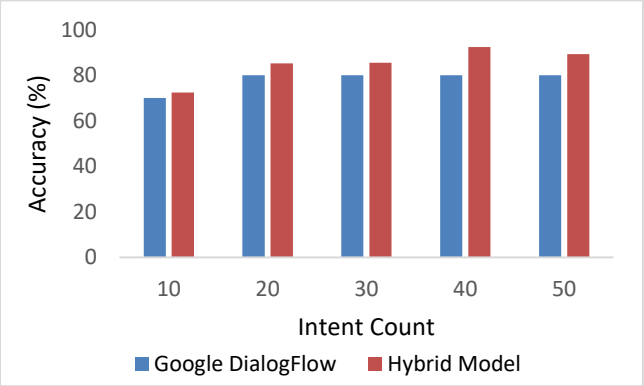


**Fig 10: Accuracy Comparison of Google Dialogflow model and Hybrid model**

## 6. DISCUSSION

A statistical analysis was conducted by comparing two different models to assess the significance of observed differences in performance metrics. This analysis helped quantify the level of improvement offered by TaskBERT over existing approaches and establish its statistical significance. The results demonstrate that TaskBERT outperforms both Google Dialogflow and KeyBERT across all evaluated metrics. The hybrid nature of TaskBERT, combining the strengths of BERT-based models with intent prediction, enables more accurate and contextually relevant responses in task-oriented dialogue scenarios. The observed differences in performance metrics between TaskBERT and existing approaches are statistically significant, indicating the superior effectiveness of the proposed model. By incorporating contextual information and fine-tuning task-specific datasets, TaskBERT achieves higher accuracy and better generalization compared to off-the-shelf solutions like Google Dialogflow. This work chose to compare with the proposed method with Dialogflow primarily because of its widespread use of intents, which aligns closely with the implementation of the hybrid model. Dialogflow's intent-based approach provides a suitable basis for evaluating the effectiveness of the proposed method in handling task-oriented dialogues, as the proposed model also focuses on intent prediction and keyword extraction to understand user input and generate appropriate responses.

While it's true that there are many other chatbot frameworks available, such as AWS Lex and Azure Bot Service, Dialogflow's emphasis on intents makes it particularly relevant to the proposed study. Intents serve as a fundamental building block for Dialogflow and the hybrid model, allowing for the classification of user queries and extracting relevant information to drive conversational interactions. Future research intends to expand the comparison to include other chatbot frameworks like AWS and Azure, thereby providing a more comprehensive assessment of the proposed method across diverse platforms. By doing so, the paper aims to evaluate the scalability and adaptability of the Hybrid approach in different environments and address the broader landscape of conversational AI solutions. Future work could explore additional optimization techniques, such as ensemble learning or active learning, to further enhance the performance of TaskBERT in diverse dialogue domains.

## 7. CONCLUSION

This paper proposed a hybrid NLP model for chatbot applications. Since most chatbot applications are task-oriented and designed for specific tasks, the SQuAD dataset-based BERT model is one the most potential question-answer-based chatbot models. However, this BERT model did not address a few question types. The results section addressed yes/no type-based and synonym-based questions with the hybrid model approach. The evaluation results demonstrate that TaskBERT outperforms Google Dialogflow by 82% and KeyBERT by 12%, as shown in Table 3 and Table 2, respectively.

The synergy between keyword extraction, LLMs, and intent prediction in the context of protocol-driven chatbots offers a promising avenue for innovation. Recognizing the unique demands of this niche application and tailoring LLM+IP approaches to suit these demands can create more effective and context-aware chatbot solutions. As the fine-tuning of LLM models continues to evolve, the Hybrid approach enhances the capabilities of protocol-driven chatbots, making them more efficient and user-friendly. This area of research holds great potential and should be explored further to unlock new possibilities in chatbot development, emphasizing the essential role of keyword extraction and interpretation. While traditional keyword extraction tools have their merits, combining Large Language Models with intent prediction aligns with the growing interest in fine-tuning LLM models for specific applications, potentially leading to more contextually relevant chatbot interactions. Further research and experimentation are warranted to explore this approach's potential fully.

The hybrid model works best when people use specific keywords to interact with the dialog systems, especially in task-oriented conversations. However, it might not perform as well in scenarios where the system needs to understand more about user expectations beyond matching keywords. The introduction and evaluation of the hybrid model on a range of question types, including synonym-based and yes/no questions provided valuable insights into the model's performance under specific circumstances. While these examples gave a starting point to show the feasibility of the hybrid model in improving the prediction performance of task-oriented dialog systems, it's crucial to do a more thorough analysis with a more extensive and varied set of questions to understand the strengths and weaknesses of the model entirely.

Future research should explore enhancements to the model to accommodate a wider variety of question types and structures

beyond the focus on keyword-based interactions. Doing a more comprehensive evaluation with a wider variety of questions will help us understand how well the model performs in different scenarios. Furthermore, we need to consider model scalability and efficiency with larger sets of data and more complex conversations.

## 8. REFERENCES

[1] M. Nakano and K. Komatani, "A framework for building closed-domain chat dialogue systems," *Knowledge Based Systems*, vol. 204, p. 106212, Sep. 2020, doi: 10.1016/j.knosys.2020.106212.

[2] P. H. Saurav, A. R. Limon, R. Amin, and M. S. Rahman, "Multi-Layer Open-Domain Bangla Conversational Chatbot with a Hybrid approach," *2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM)*, Jun. 2023, doi: 10.1109/ncim59001.2023.10212816.

[3] F. Cui, Q. Cui, and Y. Song, "A Survey on Learning-Based Approaches for Modeling and Classification of Human–Machine Dialog Systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 4, pp. 1418–1432, Apr. 2021, doi: 10.1109/tnnls.2020.2985588.

[4] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic Keyword Extraction from Individual Documents," in *Text mining: applications and theory*, Wiley online library, 2010. doi: 10.1002/9780470689646.ch1.

[5] S. Qaiser and R. Ali, "Text mining: Use of TF-IDF to examine the relevance of words to documents," *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/ijca2018917395.

[6] S. Pan, Z. Liu, and J. Dai, "An improved TextRank keywords extraction algorithm," *Proceedings of the ACM Turing Celebration Conference - China*, May 2019, doi: 10.1145/3321408.3326659.

[7] A. Andy, M. Robert, and M. F. Chouikha, "Exploiting synonyms to improve question and answering systems," *International Journal of Computer Applications*, vol. 108, no. 18, pp. 24–27, Dec. 2014, doi: 10.5120/19012-0523.

[8] M. Grice and M. Savino, "Information structure and questions: evidence from task-oriented dialogues in a variety of Italian," in *Regional variation in intonation*, P. Gilles and J. Peters, Eds. 2004. [Online]. Available: https://www.cs.columbia.edu/~julia/papers/grice&savino04.pdf

[9] G. N, V. G, and T. A. Vinetia, "Intent Classification using BERT for Chatbot application pertaining to Customer Oriented Services," *International Conference on Combinatorial and Optimization, ICCAP*, Dec. 2021, doi: 10.4108/eai.7-12-2021.2314563.

[10] N. Sabharwal and A. Agrawal, "Introduction to Google Dialogflow," in *Apress eBooks*, 2020, pp. 13–54. doi: 10.1007/978-1-4842-5741-8_2.

[11] "Microsoft Luis," *Language Understanding (LUIS)*. https://www.luis.ai/ (accessed Dec. 01, 2023).

[12] "Amazon," *Amazon Lex*. https://aws.amazon.com/lex (accessed Dec. 01, 2023).

[13] F. You, S. Zhao, and J. Chen, "A topic information fusion and semantic relevance for text summarization," *IEEE Access*, vol. 8, pp. 178946–178953, Jan. 2020, doi: 10.1109/access.2020.2999665.

[14] N. Firoozeh, A. Nazarenko, F. Alizon, and B. Daille, "Keyword extraction: Issues and methods," *Natural Language Engineering*, vol. 26, no. 3, pp. 259–291, Nov. 2019, doi: 10.1017/s1351324919000457.

[15] S. Beliga, A. Meštrović, and S. Martinčić-Ipšić, "An Overview of Graph-Based Keyword Extraction Methods and Approaches," *DOAJ (DOAJ: Directory of Open Access Journals)*, Jul. 2015, [Online]. Available: https://doaj.org/article/c60517233bf44eae8807eaba0a2ebf59

[16] M. Grootendorst, "KeyBERT: Minimal keyword extraction with BERT," *Zenodo*, 2010, doi: 10.5281/zenodo.4461265.

[17] I. Alberts *et al.*, "Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be?," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 50, no. 6, pp. 1549–1552, Mar. 2023, doi: 10.1007/s00259-023-06172-w.

[18] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, Jan. 2020, [Online]. Available: https://jmlr.org/papers/volume21/20-074/20-074.pdf

[19] A. Bhargava, A. Çelikyılmaz, D. Hakkani-Tür, and R. Sarikaya, "Easy contextual intent prediction and slot detection," *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, doi: 10.1109/icassp.2013.6639291.

[20] Q. Chen, Z. Zhu, and W. Wang, "BERT for joint intent classification and slot filling," *arXiv (Cornell University)*, Feb. 2019, [Online]. Available: https://arxiv.org/pdf/1902.10909.pdf

[21] M. Huggins, S. Alghowinem, S. Jeong, P. Colón-Hernández, C. Breazeal, and H. W. Park, "Practical Guidelines for Intent Recognition," *ACM/IEEE International Conference on Human-Robot Interaction*, Mar. 2021, doi: 10.1145/3434073.3444671.

[22] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," *arXiv (Cornell University)*, Jun. 2016, doi: 10.48550/arxiv.1606.05250.

[23] Y. Xiao, "A Transformer-based Attention Flow Model for Intelligent Question and Answering Chatbot," *International Conference on Computer Research and Development (ICCRD)*, Jan. 2022, doi: 10.1109/iccrd54409.2022.9730454.

[24] "Thesaurus.com - The world's favorite online thesaurus!," *Thesaurus.com*, Dec. 01, 2023. https://www.thesaurus.com/ (accessed Dec. 01, 2023).

[25] "Synonym Finder," *WordHippo*. https://synonym.wordhippo.com/ (accessed Dec. 01, 2023).

[26] "ChatGPT." https://chat.openai.com (accessed Dec. 01, 2023).

[27] W. Cai, Y. Jin, and L. Chen, "Task-Oriented user evaluation on Critiquing-Based recommendation chatbots," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 3, pp. 354–366, Jun. 2022, doi: 10.1109/thms.2021.3131674.

[28] "squad · Datasets at Hugging Face," Apr. 06, 2001. https://huggingface.co/datasets/squad (accessed Dec. 01, 2023).

[29] M. Q. Khan *et al.*, "Impact analysis of keyword extraction using contextual word embedding," *PeerJ*, vol. 8, p. e967, May 2022, doi: 10.7717/peerj-cs.967.

[30] "Pretrained Models Sentence Transformers documentation."https://www.sbert.net/docs/pretrained_models.html (accessed Dec. 01, 2023).