Utilizing Large Language Models for Focused Conversational Assistants

Shruti Dhavalikar
Data Science
Sahaj Software
Pune, India
shrutid@sahaj.ai

Karthika Vijayan
Data Science
Sahaj Software
Pune, India
karthikav@sahaj.ai

Abstract

A focused conversational assistant (FCA) realizes human-computer interaction bounded in a predefined scope of operation. With the advent of large language models (LLMs), it has become imperative to integrate them in conversational assistants (CAs). However, an LLM can become largely inaccurate in an FCA with multiple responsibilities, like information extraction, scope adherence and response generation. In this paper, we attempt to use an LLM for an FCA while constricting the scope of operation and maintaining a guided flow of conversation. We present a strategical combination of discriminative AI methods and generative AI models. Our methodology includes (i) a component of natural language understanding (NLU) operating discriminatively, (ii) a conditional intent-based routing of user messages to appropriate response generators, and (iii) response generators which are either custom ones or open sourced LLMs. The collation of these three strategies realizes a hybrid AI system, assisting FCA with adhering to the defined scope, maintaining context and dialogue flow.

1 Introduction

Conversational assistant (CA) is a prominent choice of human-computer interaction in natural languages, and focused conversational assistants (FCA) have been an integral part of enterprise communication. The FCAs should operate in a dedicated fashion within the business's operating domain. Traditionally, CAs were developed using rule-based techniques, which made them very rigid in their operation (Caldarini et al., 2022).

Discriminatively trained process pipelines for CA became popular, when transformers were developed for efficient text embedding (Vaswani et al., 2017). Several variations of the BERT models contributed to natural language understanding (NLU) for CAs (Devlin et al., 2019; Liu et al., 2019; Lan et al., 2020; Sanh et al., 2019). Furthermore, response generation in CA used natural language generation (NLG) with models such as GPT or BART (Radford et al., 2018; Lewis et al., 2019).

With increase in efficiency of large language models (LLMs), a shift is seen in response generation for CAs. The contribution of LLMs to CAs largely depends on design architecture and use cases. In the retrieval augmented generation (RAG), the retrieval mechanism narrows down the information-space for generative models, enabling them in generating accurate responses grounded in data (Lewis et al., 2020). This works well for a question-answering (QA) system. However, it is difficult in RAGs to maintain a natural flow of conversation or implement a guided dialogue flow for FCAs (Caldarini et al., 2022).

Another way of incorporating LLMs in CAs is to fine-tune them (Brown et al., 2020). This involves training the LLM on a dataset from the target domain, for improving its understanding of the domain. However, despite the language proficiency, the integration of LLMs in CAs is not straightforward. Under confusing prompts or under-tuning, they create hallucinations and lose control over dialogue flow (Dziri et al., 2022).

From the perspective of a business entity, particularly for automating their customer assistance, the focused behavior of CA is nonnegotiable. Here, the conversation between a customer and CA should adhere to the purpose of delivering a particular service to the user. Furthermore, the appropriateness of responses from CA, in terms of mentioning competing brands, providing incorrect financial or health advises, infeasible promises, etc., is extremely crucial for brand reputation. The CAs developed using the aforementioned implementations with LLMs, do not address the dedicated nature of the conversations required for FCA (Dziri et al., 2022).

Proceedings of Recent Advances in Natural Language Processing,pages 283–290
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-035

In this paper, we study a hybrid strategy for FCA, leveraging generative and discriminative AI methods working in synergy. This strategy includes methods based on information retrieval from data, followed by LLM for response generation. We trained an NLU system for intent and entity recognition from user messages. Our study also include a dialogue management system that predicts next action to take, based on the information retrieved by NLU. All the out-of-scope queries are strictly relegated out of the system, maintaining the scope of FCA. A similar strategy for controlling scope and generating context-rich responses was studied in (Vijayan and Dhavalikar, 2024). In contrast with the methodology proposed there, we investigated the revision of behavior of FCA based on the persona of users of CAs, to prevent data leakage. We have also done on-premise deployment of open sourced (OS) LLMs for addressing data privacy concerns.

The rest of the paper is organised as follows: Section 2 discusses an FCA with a carefully chosen case study. Section 3, explains the hybrid solution for FCA. In Section 4 we discuss the experimentation performed to analyze the use of hybrid AI for FCA. Section 5 summarises our contributions.

2 Focused Conversational Assistant

An enterprise-FCA needs to stay within the scope of its operational domain. To study the required behavior and implementation of an FCA, we utilised a specific case of CA for a banking institution. Such an FCA should answer customer queries related to banking requests within the respective institution. Furthermore, this FCA should not answer queries unrelated to banking or about other banks. Additionally, the behavior of the CA should be automatically revised depending on the user persona.

The FCA should query the enterprise data of the banking institution, to retrieve information required to answer customer queries. Assuming we have a process in-place for retrieving information (e.g.: retriever from RAG, SQL queries, APIs, etc.), we chose an LLM to be the FCA. We followed a structured prompting methodology, constituting of a set of instructions for tasks, pointers on identifying in-scope and out-of-scope customer queries, and few-shot examples on framing responses. Additionally, instructions are passed for guardrailing related to (i) staying within the scope of operation, (ii) not giving ambiguous or unsolicited financial advices, and (iii) ensuring accuracy of responses by grounding with data. After multiple iterations of prompt tuning, we observed several erroneous cases, for example,

1. Query: *"My last transaction looks fradulent"*
   - Desired behavior: Guide user to freeze assets.
   - Observed behavior: Diagnosing the issue, de-prioritizing safety concerns.

2. User question: *"How can I reduce my interest rate on loan?"*
   - Desired behavior: Refer to bank's flexible repayment schemes and respond appropriately.
   - Observed behavior: Hallucination, giving unrealistic financial advises.

3. User question: *"Changes in banking after Covid?"*
   - Desired behavior: Out of scope.
   - Observed behavior: Gave answer from pretrained knowledge.

Post careful analysis of erroneous cases, we identified the following scenarios related to prompt engineering an LLM to be an FCA.
1. Prompt overloading with instructions 'lost-in-the-middle'.
2. Deviation from instructions, particularly in critical information handling.
3. Actionable items frequently getting classified as out-of-scope.
4. Chances of data leakage when the LLM confuses between the user personas of multiple customers.
5. Increased latency due to hierarchical instructions in the prompt.

It is now clear that, just by prompt engineering, an LLM cannot act as an FCA (Dam et al., 2024). This may result in tarnishing the brand reputation of the banking institution. We propose a hybrid solution for FCA, where the NLU is handled in a custom fashion, rather than using an LLM to perform all tasks.

3 Hybrid AI for FCA

The hybrid solution that we propose for an FCA consists of three elementary components: (1) NLU component responsible for deciphering the ask of the user, (2) dialogue management component routing conversations within the CA and (3) the response generation component answering user queries. Figure 1 shows the schematic of the hybrid

FCA. The first 2 components here are trainable with appropriate data and the last component are properly engineered. In the following sections, we will explain the development stages of such an FCA.

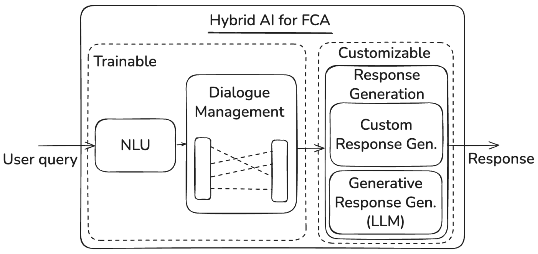

Figure 1: Hybrid AI FCA architecture

3.1 Scope Identification

The fundamental requirement of an FCA is the ability to adhere to a defined scope of operation. The business entity decides the type of customer queries that are in- and out- of scope. The identified scope is then represented as intents and intent-categories (ICs) for the development of FCA. Further, the business decides which category of intent they consider "critical", because of the requirement of privileged business information.

The customer messages with the intent of exchanging pleasantries or navigating through the category of banking services are identified as *generic-and-navigation intents* (Intent Category 1 or IC1). The *in-scope-non-critical intents* (Intent Category 2 or IC2), consists of intents which can be answered from the general knowledge of the LLM, not necessarily needing business's data. The third intent category *in-scope-critical intents* (Intent Category 3 or IC3), requires some specific/sensitive information from business records. Finally, there is the *out-of-scope intent* (Intent Category 4 or IC4). We have scoped for 19 intents grouped into 4 intent-categories for the FCA, as given in Table 1. Now, we will discuss the trainable components in the system.

3.2 NLU

The NLU ensures the FCA its *focused* behavior, by classifying user messages into one of the predefined set of intents. The pipeline undergoes supervised training using carefully curated dataset consisting of user messages belonging to each intent.

NLU pipeline shown in Figure. 2 consists of a tokenizer, a set of featurizers and a classifier. We have used a white space tokenizer, a combination of syntactic and dense language model featurizers (Vijayan and Anand, 2022), together with the Dual Intent and Entity Transformer (DIET) from the Rasa framework as the classifier (Bocklisch et al., 2017). The DIET is a transformer based architecture, trainable to identify intents and entities. Thus, user messages are classified into an intent, together with extracting the entities mentioned. Once the intent is identified, the dialogue management scheme will decide on what is the next action and where to route the message to, for response generation from CA.

Intent Category (IC)	Intents in each IC	Examples of user messages
Generic and navigation (IC1)	Greeting	*Hello*
	Goodbye	*Thank you, bye*
	Affirm	*Yes*
	Deny	*Nope*
In scope non critical (IC2)	Branch address	*Which is the nearest branch?*
	Past transactions	*Show me my last 4 transactions*
	Dispatch status	*Is my card dispatched?*
	Due date	*Last date for my bill payment*
	Card issues	*My card is not working*
	IFSC code	*I am looking for IFSC code*
	Loan query	*How can I get a loan?*
In scope critical (IC3)	Activate card	*help me in starting my debit card*
	Generate pin	*I need help changing my pin*
	Unauthorised transaction	*I did not do this last transaction*
	Change limit	*Increase my card limit?*
	Balance enquiry	*I want to know my balance*
	Block	*How can I suspend my online banking?*
Out of scope (IC4)	Out of scope	*Which is the best car?, Can I smoke?, etc.*

Table 1: Scope of FCA for banking institution.

3.3 Dialogue Management

The dialogue management in FCA plays a crucial role in adhering to a defined directive of the conversation, while providing flexibility for context shift. The dialogue management shown in Figure. 3 undergoes supervised training with a dataset of conversational stories. For intent-based selective routing of user messages, the following rules are established: (i) the customer messages with intents belonging to IC2 are routed to LLM for generating responses. (ii) The customer messages with criti-

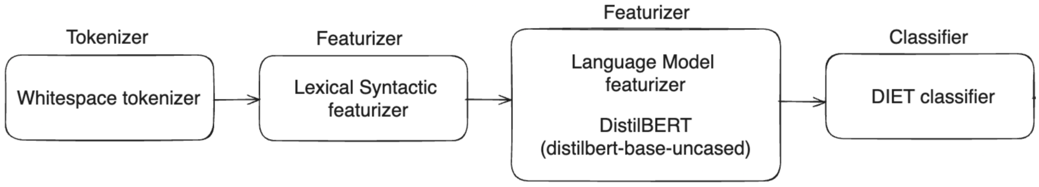
Tokenizer
Whitespace tokenizer
Featurizer
Lexical Syntactic featurizer
Featurizer
Language Model featurizer
DistilBERT (distilbert-base-uncased)
Classifier
DIET classifier

Figure 2: Proposed NLU pipeline

cal intents belonging to IC3 are routed to in-house custom response generators, which may use privileged business information. We protect privileged information from getting consumed by LLMs. (iii) The asks belonging to generic intents in the IC1 and queries with OOS intent (IC4) are replied with static responses.

The entities identified from customer messages are used to set flags and/or fill slots. These are used for (i) specific checks on customer validation, (ii) flagging and routing-out users if they do not comply with access policies or attempt to access restricted information, (iii) for consistent context maintenance, and (iv) to ensure effective information retrieval.

3.4 Response Generation

The response generation component assures delivery of an appropriate answer with business information to the user query, bringing in generative AI (gen-AI) capabilities wherever required. It consists of a custom response generator and an OS LLM response generator.

3.4.1 Custom Response Generator

This is initiated to generate responses to queries with intents belonging to IC1 and IC3. Depending on the individual intent, the custom generator may call an internal API to fetch privileged business information or write/amend the file system, thereby providing accurate information to the user. The construction of custom responses requires strict control by business, and we applied programmatic techniques to gather the right context and to generate responses.

3.4.2 OS LLM Response Generator

A self-hosted OS LLM, Llama2 (7B params) (Touvron et al., 2023) produces the response for user queries with intents belonging to IC2. Llama2 can be hosted on the same or a distinct machine as per the infrastructure choice. The LLM is accessible to the outer services only via a secured application endpoint, assuring business information security.

In the FCA, the response generation component calls the hosted LLM whenever the user asks a query from IC2. Depending on the intent, the response generation component sends additional information to the LLM along with a prompt. We adopted a strategy for prompting containing (i) definition of a persona for LLM, (ii) instructions and chain-of-thought (CoT) prompting to follow specific rules/guidelines while performing tasks, and (iii) few-shot examples. The prompt used for LLM as a final response generator is given in Table 2. The prompt strategy and content are concluded with iterative experimentation.

You are a customer-care assistant for a banking institution. Your job is to give crisp descriptions of banking services and provide key information to customers. You are an expert in banking and know how to use context/relevant data while answering a query. Make use of the additional Business Context to prepare a relevant answer to the customer query.
Think step by step before answering: 1. Identify the entity from the user question. 2. Highlight features, rules and regulations in an engaging and positive way using relevant information in the context. 3. Keep the response concise and informative. 4. Format the answer in a template: {template}.
Learn the process by going through the examples below. Example(1), Example(2)

Table 2: Prompt used for final-response generation in FCA using LLM, showcasing sections on persona, instruction & CoT, and few shot examples.

Lets revisit the practical limitations of an LLM being an FCA, explained in Section 2. The issues related to data-leakage and incorrect execution of actionable items are mitigated in the hybrid AI system for FCA using the preset dialogue flow. The

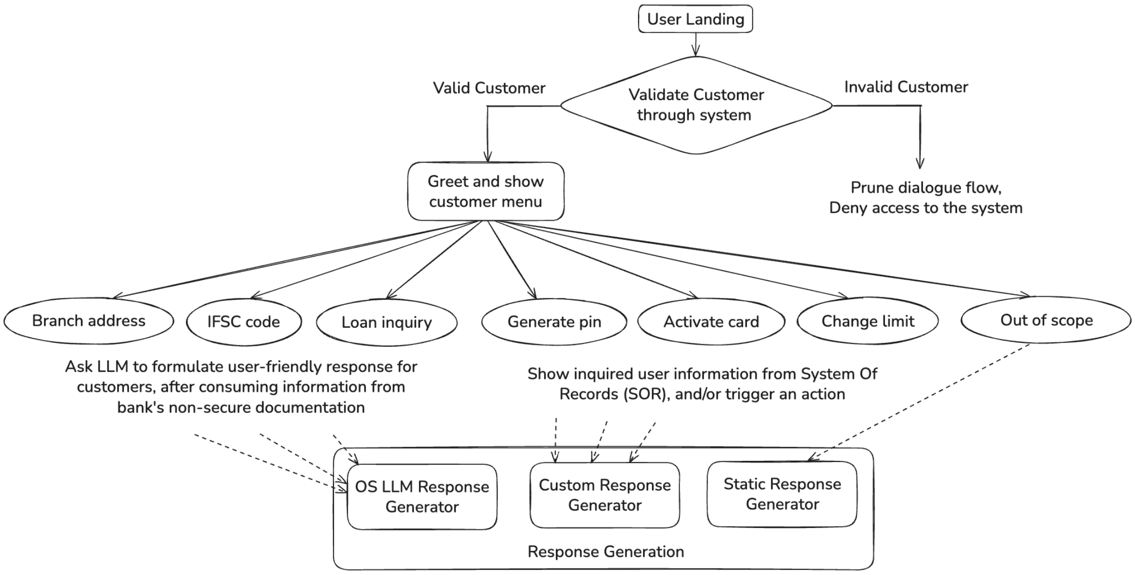
User Landing
Validate Customer through system
Valid Customer
Invalid Customer
Greet and show customer menu
Prune dialogue flow, Deny access to the system
Branch address
IFSC code
Loan inquiry
Generate pin
Activate card
Change limit
Out of scope
Ask LLM to formulate user-friendly response for customers, after consuming information from bank's non-secure documentation
Show inquired user information from System Of Records (SOR), and/or trigger an action
OS LLM Response Generator
Custom Response Generator
Static Response Generator
Response Generation

Figure 3: Proposed dialogue flow

NLU and dialogue management are based on information retrieval from user query, and these components satisfies the crucial need for preset dialogue flow. As information retrieval and routing are already implemented discriminatively, the prompt for LLM as a response generator is no longer overloaded or hierarchical in nature, as explained in Section 2. This helped in curbing down hallucinations by LLM and minimize latency in the FCA. In the next section, we discuss the results and evaluation of the FCA.

4 Results and Discussion

The hybrid approach for FCA used discriminative and generative AI. Since the two approaches are fundamentally different, we evaluated the FCA in two distinct stages. The discriminative NLU pipeline is evaluated with the efficiency of intent classification. On the other hand, generative responses are evaluated with a similarity score between responses from LLM and expected ground truth answers (Maroengsit et al., 2019). We also performed manual evaluation of responses from LLM and presents mean opinion scores from evaluators.

4.1 Description of dataset

We have used a combination of 2 prominently used datasets for intent classification and eventually for building the FCA, namely, Skit-S2I (Rajaa et al., 2022) and the intent classification dataset from (Larson et al., 2019). The combined dataset contains customer queries belonging to 19 intents, each one having 85 samples for training and 25 samples for testing. The OOS intent has 1000 samples for training and 250 samples for testing. The dataset prepared for this study is published via Github and is available here.

4.2 Evaluation of NLU

We performed intent classification of user queries into 19 intents from Table. 1, using the NLU pipeline in Figure. 2. We reported F1 scores with respect to 4 intent-categories in the Table. 1, rather than each of the 19 specific intents, to increase readability and inference. The F1 scores for the categories of intents are calculated for various BERT-based Language Models (LM). Each LM is independently included as a dense featurizer in the NLU pipeline and a comparative study of these LMs is presented in the Table 3.

The BERT-based models deliver superior performance in intent recognition, and we chose to use BERT-large model for NLU in FCA. For critical intents, the queries are very specific and crisp, which became easier for the LMs to decipher the underlying patterns. Additionally, IC4 category has overall better scores, which is very challenging given the nature of this category. This can be attributed to the variability in the crafted training data for out-of-scope intent. We took inspiration from the concept of Universal Background Model(Reynolds, 2009) in speech processing to train the NLU for OOS in-

Table 3: Evaluation of NLU: F1 scores (%) of intent classification delivered by different language models(LM).

IC Models	IC1	IC2	IC3	IC4	All
bert-base-uncased	97.95	99.5	99.46	99.54	99.11
bert-base-cased	97.35	99.77	98.87	99.84	98.96
roberta-base	97.75	97.76	96.59	99.49	97.9
gpt2	86.4	95.36	93.88	97.06	93.18
bert-large-cased	99.42	99.35	99.18	99.79	99.43
bert-large-uncased	98.69	98.09	98.5	98.98	98.57
distilbert-base-uncased	96.04	98.73	98.54	99.34	98.16
distilbert-base-cased	95.22	98.2	98.23	99.59	97.81

tent. The discriminatively trained NLU guarantees the *focused* nature of the CA.

4.3 Generative response evaluation

The response generation component uses either a custom response generator or an LLM depending on the intent category and dialogue story. The custom response generator demand visibility of sensitive business information, and is mostly programmatic and agreed upon with business requirements. Hence, we have excluded it from evaluation.

The responses coming from LLM are subject to evaluation by both customers and the business. We performed a subjective and an objective rounds of evaluation of FCA's LLM-responses. We maintained 10 queries each for the 4 intents belonging to IC2. We presented these queries to the FCA to produce LLM-responses. Additionally, we approached the business to roll out the expected guideline responses for these queries, which were used as ground truth responses. We calculated the cosine similarity scores to test the similarity of FCA responses with the ground truth.

For the subjective evaluation, we presented the set of 40 question-response pairs from FCA to 5 human subjects. 3 of the 5 subjects belong to the customer user persona and 2 belong to the business user persona from banking institution. We asked them to rate their level of satisfaction with the FCA responses on a scale of 1 to 5, with 1 representing unsatisfactory and 5 denoting excellent satisfaction. The Mean Opinion Scores (MOS) of these ratings over 40 question-response pairs are presented in Table 4 along with the cosine similarity scores (objective metrics).

It was observed that both user personas gave positive feedback to FCA. The business expressed its contentment regarding FCA's ability in answering user queries without skipping any. The customers showed interest in the creative and interactive responses compared to the templated responses they usually saw elsewhere. From Table 4, we observed that I2 outperforms the rest of the intents in both subjective and objective evaluations. For I2, the LLM consistently generated detailed explanations, and the users appreciated these responses in spite of their verbosity. On the contrary, I3, which has comparatively lower scores, was observed to have long and vague responses from LLM, which users did not appreciate well. Our evaluation of FCA includes formulation of persona aware and context-rich responses, studied against the satisfaction of both customer and business audience, as opposed to (Vijayan and Dhavalikar, 2024). Our perspective advocates the more secured use of OS LLMs for communication synthesis without hurting business interests and customer satisfaction.

Table 4: Evaluation of response generation by the LLM.

Intent Scoring	I1	I2	I3	I4
Subjective evaluation (Mean opinion score (1-5))				
Customer persona	3.9	3.95	2.8	3.55
Business persona	4.06	4.33	3.1	3.76
Objective evaluation				
Cosine score (%)	71.45	80.71	75.27	72.24

5 Conclusion

In this paper, we address multiple challenges in utilizing an LLM for FCA, in terms of scope adherence and context maintenance. We presented a strategy of using discriminative techniques to control the queries at the input of LLM, placing guardrails at the front-end of the system. The combination of discriminative and generative intelligence offers enforcement of rigid scope boundaries and let the LLM be optimally creative within the set scope. Additionally, the context maintenance is done by entity extraction and consistently sharing them across the process pipeline. The proposed hybrid AI strategy brings in stronger control for business over the narrative of conversations in FCA.

References

Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management. *CoRR*, abs/1712.05181.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Guendalina Caldarini, Sardar Jaf, and Kenneth McGarry. 2022. A literature survey of recent advances in chatbots. *Information*, 13(1).

Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. 2024. A complete survey on llm-based ai chatbots.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models?

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020*, pages 1–17. OpenReview.net.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692:1–13.

Wari Maroengsit, Thanarath Piyakulpinyo, Korawat Phonyiam, Suporn Pongnumkul, Pimwadee Chaovalit, and Thanaruk Theeramunkong. 2019. A survey on evaluation methods for chatbots. In *Proceedings of the 2019 7th International Conference on Information and Education Technology*, ICIET 2019, page 111–119, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Shangeth Rajaa, Swaraj Dalmia, and Kumarmanas Nethil. 2022. Skit-s2i: An indian accented speech to intent dataset.

Douglas Reynolds. 2009. *Universal Background Models*, pages 1349–1352. Springer US, Boston, MA.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108:1–5.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Karthika Vijayan and Oshin Anand. 2022. Language-agnostic text processing for information extraction. In *Proc. 2022 5th International Conference on NLP techniques and applications*.

Karthika Vijayan and Shruti Dhavalikar. 2024. Combining discriminative and generative ai for dedicated conversational assistants. In *Workshop on Composite AI at the 27th European Conference on Artificial Intelligence*.