Combining Discriminative and Generative AI for Dedicated Conversational Assistants

Karthika Vijayan[a,*,1] **and Shruti Dhavalikar**[a,1]

[a]Data Science, Sahaj AI, Pune, India
ORCID (Karthika Vijayan): https://orcid.org/0000-0001-7281-1329, ORCID (Shruti Dhavalikar):
https://orcid.org/0009-0006-7599-3635

Abstract. Dedicated conversational assistants are expected to function within a defined scope of operation, pertaining to specific application scenarios. Recently, generative models are showing great promise and can contribute significantly to the space of conversational AI. In this paper, we attempt to showcase the use of generative models in textual conversational assistants, and highlight a few challenges like the lack of control over its responses in a dedicated scope setup. We propose to merge capabilities of discriminative models with generative ones, in pursuit of regaining the ability to control information dissemination by conversational assistants. We studied the application of an automated customer care agent for a specific business, and designed a definite scope. Later, we built the natural language understanding (NLU) of user messages in a discriminative fashion. The NLU makes use of BERT-based models for information extraction followed by selective routing of user's queries to gen-AI solutions and custom response generators. We obtained an F1 score in NLU accuracy of 99.68% for intent recognition, and specifically 99.46% for the "out-of-scope" intent, which is extremely challenging to model in a dedicated assistants' setup. Our work provides confidence in building dedicated conversational assistants for businesses to assist in their customer interactions, while controlling the narrative around their products and services.

1 Introduction

Conversational AI is creating strong waves of impact in human-computer interaction, particularly with the recent advent of generative AI (gen-AI) solutions based on Large Language Models (LLMs) like GPT3.5, GPT4 and LLaMA [4, 30]. From the perspective of a business entity, in relation to its customer interactions, there is a dire need of dedicated conversational assistants (DCA). Here, the assistant is expected to be dedicated to answering user queries related to products/services offered by the business, and is not supposed to respond to any messages outside the set scope. Though the gen-AI solutions offer great promise, there are a number of challenges associated with their usage in a dedicated setting.

Conversational Assistants (CAs), or chatbots, mimic human conversations by processing user inputs and providing responses in a sequential manner constituting a dialogue flow [5]. It evolved from a rule-based system to one employing pattern matching, deep learning based discriminative techniques and later to gen-AI solutions. The transformer architecture is the backbone of many of these state-of-the-art methodologies [31]. The sequence of processes involved in realizing a CA primarily are, understanding of user messages through information extraction and generating responses while maintaining a context from previous message exchanges [5].

Until recently, discriminative solutions were prevalent for NLU and natural language generation (NLG) in the development of CA. Large language representational models with transformer architecture, like different versions of the BERT model, act at the core of NLU in a discriminative fashion [8, 17, 13, 27]. These language representational models provide efficient mathematical representations of text, in terms of vector embeddings, which are later utilised for downstream tasks like information extraction for CAs. For text response generation, techniques ranging from response lookup tables to GPT and BART are utilised [23, 14].

The rise of gen-AI solutions is beginning to revolutionise the field of conversational AI, with extensive applications to question-answering (QA) systems, CA, document generation and so on [6, 34, 22, 29]. The gen-AI solutions are based on LLM, which also share the transformer architecture like the BERT models [21]. The QA system is probably the most widely used application of LLM, prominently with a strategy termed as Retrieval Augmented Generation (RAG) [12, 15, 33]. CA employing LLM is generally visualised as a QA pipeline, and RAG structure is employed.

Though LLM exhibits impressive text generation capabilities, its use in the development of a DCA is not straightforward. The most prominent challenge with using LLM in DCA is the lack of control over its responses to user queries. LLMs are known to hallucinate under unclear prompting, and user queries are not always concise [9]. While acting as a customer care agent for a business, hallucination by the DCA can cause exposure of sensitive information or the spread of misinformation. From the business's perspective, this will result in customer annoyance and detrimental effects on its brand reputation. From the development aspect of a DCA, despite RAG strategy being widely successful for QA pipelines, it is not remarkably effective for CA. The factors of context maintenance and adherence to a dialogue flow are the most relevant differences between a generic QA pipeline and DCA, which appear to be laborious in the case of RAG [7].

In this paper, we propose a composite approach with discriminative and gen-AI solutions for DCA, where we attempt to control all the responses generated by the CA based on the business's interests. We have conducted rigorous analysis on the identification of the scope of the DCA and design a closed set of intents and dialogue flow

[*] Corresponding Author. Email: karthikav@sahaj.ai.
[1] Equal contribution.

based on the scope. The NLU for intent recognition is implemented using BERT in a discriminative AI approach. Later, adhering to the dialogue flow, selective routing of user queries is performed based on identified intents to LLM and custom generators to obtain relevant responses. We reported the F1 scores as a metric to demonstrate the effectiveness of NLU, and ROUGE & subjective evaluation scores to showcase the efficiency of response generation. Thus, we designed a DCA with an enclosed framework in terms of its scope, and combined discriminative and gen-AI capabilities for its operation. Our strategy gains the business's confidence in releasing a controlled narrative of their products/services, while preserving brand reputation.

The rest of the paper is organised as follows: In Section 2, we explain the strategies of building a general purpose CA. In Section 3, we discuss the challenges in developing a DCA and how a composite strategy is utilised. Section 4, elaborates the efficiency of the proposed strategy in terms of NLU and response generation. We summarise the contributions of this paper in Section 5.

2 Development of Conversational Assistants

CAs are predominantly expected to assist their users gather relevant information and perform certain actions. Once we design a CA in a dedicated setting, it is supposed to assist in customer interactions for businesses. We will dive deep into methodologies of developing a CA in the discriminative fashion and the gen-AI way.

2.1 The discriminative AI way

The development of a CA includes three major parts, namely, scope identification, NLU of user messages, and response generation. Scope identification generally entails, (a) *understanding nuances in specific business-customer interactions*, (b) *designing a closed set of intents and entities*, and (c) *designing a dialogue flow for conversation, aiming at performing specific actions that the business offers* [18]. The closed set of intents is identified based on the products/services and associated information dissemination the business intends to do. The dialogue flow is usually designed to encourage users to take certain actions like registering their interest to purchase a specific product or service.

After this stage, the NLU is realised aiming at intent and entity recognitions from user messages. Embeddings from language representational models, like the BERT, word2vec and variants, are used for NLU. These models are pretrained on huge amounts of text data, during which they learn gross meanings of tokens by analysing multiple occurrences of individual tokens in numerous contexts [8, 17, 13, 27, 20]. The pretrained BERT-based models can then be adopted for multiple downstream tasks including information extraction in NLU for a CA [24, 26, 19].

The response generation for user queries in CAs generally happens in a custom fashion. The details to generate relevant responses based on intents, sometimes need to be fetched from the business's knowledge bases via API calls. Later, text responses incorporating the fetched details are generated in a custom fashion using NLG methods based on RNN, CNN, etc. [18].

There exist other methodologies for the development of CAs, like QA systems relying on Information Retrieval (IR), seq2seq models, etc. which use question-answer pairs for training. However, the IR is less appropriate to fit in a dialogue flow and it lacks a personality for the CA [5]. Generally, scaling up the number of intents and addressing complex dialogue flows are pointed out as disadvantages of intent-based CAs [18]. Yet, in our opinion, it is a suitable fit for dedicated scope settings. We note that the responses from it will not be as appealing as the gen-AI responses.

2.2 The gen-AI way

The CA is considered as a QA based IR system, where a semantic similarity search is performed between LLM embeddings of query and prospective answers, which are later passed to LLM itself for answer generation. The LLM embeddings are effective vector representations of text, generated from a pretrained model based on transformer architecture, which is trained on a huge volume of text data. The semantic search based on the LLM embeddings provides faithful IR from a database in relation to the user queries.

This strategy, termed as RAG, organises the prospective answers as text chunks and saves their LLM embeddings in a vector database for query-based retrieval [12, 15]. This method avoids the need for intent recognition and adhering to dialogue flow; instead provides a free flow of conversation. Several enhancements were proposed to IR-QA using LLMs in [10, 28]. However, maintaining context in RAG is realised by repeatedly passing previous user messages (or a consolidated context) as a prompt to an LLM [1]. This becomes a costly affair over time, in terms of querying and maintaining double storage of the same data, as a DCA is supposed to be the customer care agent for businesses for a really long time. RAG does not work remarkably well in DCAs as it does in "QA over docs" [7].

Additionally, the gen-AI solution for DCAs does not work based on intents. Without exhaustive prompting, it can answer all queries from users, even out-of-scope ones. This is not preferred in DCA, and most businesses have strict policies against this.

2.3 The composite way

It is suggested that in a dedicated setting, the CA should respect (a) *the scope (intents and dialogue flow)*, (b) *respond within context of conversations*, and (c) *respond with natural sentences rather than custom crafted sentences*. We propose to merge capabilities of discriminative and gen-AI to accomplish these considerations for a DCA.

3 Conversational Assistant with Composite Architecture

We proceed to implement the DCA with a composite architecture for business-customer interactions. With this strategy, we perform information extraction from user queries to understand customers' query/messages. Later, based on identified intents and entities from user messages, we fetch the information required to answer the queries. Adhering to the dialogue flow and identified intent, we choose an LLM or a custom response generator to answer the user messages. In this section, we explain the sequence of development processes for a DCA with composite architecture.

3.1 Scope identification

This process involves understanding the business's requirements and preferences in their customer interactions through meticulous discussions. Then, we identify the closed set of intents of user queries that the business wants to support based on the products and services they offer. Later, we classify these intents into four categories based on how we want to respond to them.

The *generic intents* (Intent category 1 or IC1) include queries belonging to 'greet', 'goodbye', etc. where responses should be given to acknowledge the presence of the user. The *in-scope low-risk* intent category (Intent category 2 or IC2) contains queries related to features of a specific product/services, and the details to respond to these queries are generally advertised by the business itself. Hence, there is no involvement of sensitive information. The *in-scope critical* intent category (Intent category 3 or IC3) contains queries related to price, technical specifications or service personnel for products/services. The answer to these queries may contain sensitive information and businesses would like to have complete control over the responses of DCA. Additionally, we designed an *out-of-scope* intent (Intent category 4 or IC4), which includes queries that the business thinks that its DCA should not be answering. Examples of user queries are given in Table 1.

Intent Category	Examples of user messages
Generic (IC1)	*Hello* *Good morning*
In-scope low-risk (IC2)	*Colour options for **ProductA*** *Tell me about **ProductB***
In-scope critical (IC3)	*Share the pricing for customised **PlanA*** *Where is the nearest service centre located?*
Out-of-scope (IC4)	*Share info on global warming* *Astrology help*

Table 1. Examples of user messages received by a DCA

A dialogue flow for conversations is also formulated in discussion with the business. The importance of dialogue flows stems from the fact that certain type of queries from customers are what the business focuses on converting into sales/service opportunities. So, it is in the interest of business profitability to guide the customers in a certain manner, assisting them in registering their interest in products/services, connecting them with service personnel, aiding in online purchase processes, and so on. The DCA should identify the intent of users and take them in a certain dialogue flow.

Thus the scope identification consists of business conversations, identifying the closed-set of intents for DCA and design of a dialogue flow.

3.2 NLU

3.2.1 Description of dataset

We contacted a customer care call center to obtain a dataset of customer queries received by product and/or services companies[2]. We further anonymised the sensitive content in the data to remove any PII content. We also appended additional data curated by LLM (GPT3.5 Turbo) to introduce a variety in the querying pattern to make the NLU pipeline more robust. We prepared training and test data of user queries belonging to all intents by making sure that we have atleast 100 examples to train and 30 examples to test for each intent.

3.2.2 Modelling the NLU

We designed the NLU process pipeline with a tokenizer, featurizers and a multi-task classifier for intent and entity recognitions [2]. The NLU process pipeline is shown in Figure 1. The components in the NLU pipeline are

- Tokenizer: We used a simple whitespace tokenizer to split words as tokens
- Featurizer: A combination of sparse and dense features
- Classifier: Multi-task classifier for intent and entity recognition

For the featurizer process, we experimented with different variants of BERT models for embedding extraction. These embeddings are used as dense features in our work. Later, we paired the dense featurizer with some sparse featurizers, like n-grams to form composite featurizer components [32]. The classifier that we used was the dual intent-entity transformer (DIET) from the RASA framework [2].

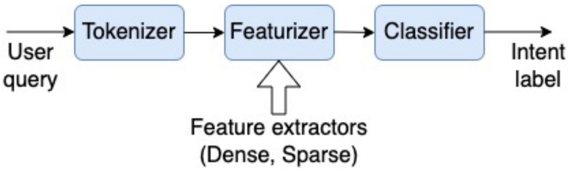

Figure 1. Process pipeline for the NLU.

We trained this process pipeline with the training dataset from all intents and later validated and tested it with respect to intent identification accuracy. Apart from the data from customer queries dataset, we prepared data for modelling *out-of-scope* intent. We took motivation from the idea of the universal background model in speech technology [25], and prepared the training data by involving as many diversified examples as possible. This data included more number of examples with offensive/sensitive nature and critical nature to business considerations, like queries related to their competitors' products/services. The NLU pipeline that we have chosen is based on discriminative AI.

3.3 Response generation

After recognising the intent of the user query, based on the intent category that it belongs to, we route it to either an LLM or custom response generator to construct answers. The generic and *in-scope low-risk* intents do not involve sensitive information in their responses and the business feel confident in passing these intents to LLMs for response generation. Whereas, *in-scope critical* intents involve sensitive/protected information, and are routed to custom response generators consuming information fetched over API calls to the business's knowledge bases. The *out-of-scope* intents always result in a static response, expressing the inability of DCA to answer the queries in this intent-category.

For obtaining responses from LLM, we made use of OpenAI APIs by passing the query and details to answer in a prompt [3]. We performed prompt designing to avoid hallucinations in the response generation by an LLM. Certain instructions like avoiding comparison between different brands, avoiding abuse were strictly enforced on the LLM via prompt. An example set of user query - prompt for LLM - corresponding response is given in Table 2.

In some scenarios, while handling in-scope low-risk intents, additional information related to the conversation and the user query was sourced in the prompt from the business enterprise data. This brought in the effect of keeping the response relevant to the scope of the conversation, while responses being natural. Another example set of user query - prompt for LLM - corresponding response, corresponding to prompting with additional information, is also given in Table 2.

[2] The dataset consists of privileged client data. We are unable to open-source this dataset, respecting Digital Personal Data Protection Bill, 2022, [CountryName].

Query	What is the mileage of [carA]?
Prompt	You are an automated car sales assistant that gives required information on available mileage in market for particular cars, mentioned in the query. User message will be a dictionary of three parts. first is the text query, second is intent, third is car names. Do not give comparisons between mileage of different vehicles.
Response	[carA] mileage is 22.41 kmpl to 31.12 km/kg.
Query	Information about [ProductA]
Prompt	You are a product presales assistant that gives only required information on products with its general details. User message will be a dictionary of four parts. first is the text query, second is intent, third is the product names, fourth is features of those products. Please make use of features while answering. Make answer interesting. Do not give comparisons of different products.
Response	Ah, the [ProductA]! A [product] that truly takes things to the next level. Let me give you the scoop on this marvelous machine. First and foremost, safety is prioritized to the next level. You can feel secure knowing that it comes with Next Level Safety features, ensuring your well-being ...

Table 2. Example of prompting an LLM to get response for user queries.

We kept response generation for critical intents in-house using custom generators. These response generators are allowed to fetch information from the knowledge base of the business with API calls, and integrate the fetched information into static text prefix or suffix. The out-of-scope intent will always be answered with a static response. This fashion of responding to user queries ensures that the business controls the narrative in their customer interactions.

The overall design of the composite architecture proposed in this paper for DCA, combining the abilities of discriminative and gen-AI, is shown in Figure. 2.

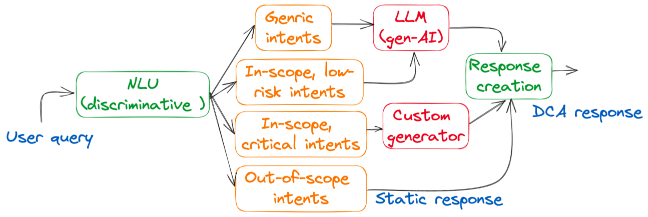

Figure 2. Proposed schema of composite architecture for a DCA.

4 Results and Discussions

We evaluated the schema of composite architecture for a DCA, in terms of the accuracy of NLU and appropriateness of response generation in terms of objective scores and subjective evaluation. The identified scope for DCA included a total of 15 intents, which has 3 generic intents (IC1), 4 in-scope low-risk intents (IC2), 3 in-scope critical intents (IC3) and 1 out-of-scope intent (IC4). Additionally, there are 4 named entities that the business is interested in capturing from user messages. We considered at least 100 examples per intent for training and around 30 examples per intent for testing.

4.1 Evaluation of NLU

We trained the NLU pipeline of information extraction with the training data, while experimenting with multiple Language Models (LMs) for embedding extraction. The LMs are obtained as huggingface hosted models [11]. The intent recognition F1 score from NLU using various LMs over 4 intent categories are reported in Table 3.

We observed that the intent category, IC3, had a comparatively higher F1 score than the of rest of the intent categories. The training data for intents belonging to IC3 depicts a specific querying pattern of inquiring typical details maintained by the business. Other intent categories are also represented well by LM featurizers and we obtained faithful F1 scores in intent recognition.

The queries belonging to out-of-scope intent represent "everything else" apart from the scope of the DCA vaguely like the noise. Hence, it becomes tricky for the model to learn that representation. We obtained high F1 score from different LMs for the IC4 category (out-of-scope). The higher F1 score for a challenging intent like out-of-scope can be attributed to our crafted data preparation for this intent. As indicated in Section 3.2.2, we relied on the idea of modelling the universe of speakers in speech technology. We created a training dataset with wide scope of queries/statements/messages on a variety of topics unrelated to the designed scope of DCA.

The classic BERT-large model is delivering the best F1 score of NLU, across all intent categories. We chose the NLU pipeline with the BERT Large model for the following studies.

Table 3. Evaluation of NLU: F1 scores (%) of intent recognition corresponding to intent categories (IC) delivered by different language models(LM).

IC LM	IC1	IC2	IC3	IC4	All
bert-base-uncased	99.53	98.74	99.58	98.65	99.12
bert-base-cased	99.26	98.94	99.58	98.92	99.18
bert-large-uncased	100.00	98.95	99.88	99.46	99.57
bert-large-cased	**100.00**	**99.56**	**99.70**	**99.46**	**99.68**
roberta-base	98.78	98.89	99.30	97.82	98.70
distilbert-base-uncased	98.79	99.06	99.58	98.65	99.02
distilbert-base-cased	98.36	98.24	98.64	98.38	98.41
GPT2	88.93	91.94	94.16	93.59	92.15

Once the intent recognition is performed and user queries are tagged with intent labels, the selective routing of those queries to gen-AI or custom response generators happens. The DCA responses, either from gen-AI or custom generators, are then evaluated objectively (ROUGE scoring, Cosine similarity score) and subjectively (manual evaluation).

4.2 Evaluation of LLM responses

For the evaluation of responses, we identified two user personas, namely the "business person" and "customer". These two user personas are required as the user satisfaction from business and customers are equally important for a DCA. From the LLM responses, 40 queries from the in-scope low-risk category (IC2) over 4 intents are passed to GPT3.5 via openAI API and responses are collected. This set of query-response pairs is given to 8 individuals belonging to the business user persona and customer user persona, for manual subjective evaluation. The subjects involved in the subjective study were asked to rate the DCA responses from LLM on a scale of 1 to 5 (1 being unsatisfied and 5 being excellent). We computed the Mean Opinion Score (MOS) of ratings given by these individuals, which are reported in Table 4.

It is observed that users belonging to both personas expressed their satisfaction with DCA responses from LLM faithfully. The customer satisfaction is low in LLM responses to queries belonging to

I3 and I4. These intents required additional information to be provided to LLM for response generation. And, it appears that the LLM responses consuming a specific set of information tend to be verbose. Customers noted verbosity as a reason for their displeasure in LLM responses to intents I3 and I4. The business users expressed their satisfaction equivalently across all intents.

Table 4. Evaluation of response generation by the LLM. Subjective and objective scores for responses from DCA corresponding to 4 intents belonging to the IC2 category.

Intent / Scoring	I1	I2	I3	I4
Mean opinion score (1-5)				
Customer persona	4.78	4.74	3.32	2.77
Business persona	4	3.7	4.1	4.2
Objective evaluation				
ROUGE Score(%)	66.84	62.63	27.37	66.35
Cosine score (%)	93.23	89.29	74.84	96.85

The objective evaluation of LLM responses to 4 intents belonging to category IC2 is performed and reported in terms of ROUGE score and Cosine similarity score. To compute these scores, we created a ground truth set of responses in consultation with the business. The ROUGE score is calculated to evaluate whether the generated response precisely summarises/incorporates the ground truth. We additionally calculated cosine similarity scores to reconfirm if the generated response and the ground truth are aligned in their meanings. The results of the objective evaluation of LLM responses are reported in Table 4.

The nature of the calculation of the ROUGE score includes the n-gram approach. They sometimes fail to represent the similarity of meanings between sentences, when different wordings are used to express the same fact. In such scenarios, higher cosine similarity scores reassure the right validation for the evaluation. Note that for calculating the similarity score, both the generated response and the ground truth were embedded using the BERT featurizer and their distance in the vector space was calculated to capture their semantic distance.

The objective evaluation of LLM responses indicates that the ROUGE and Cosine scores are in agreement [16]. The LLM response scored least in intent I3, mostly due to the verbosity of responses.

The evaluation in Table 4 shows the DCA is able to cater to both sides of the dialogue, namely, business and customer. The difference in the distributions of the scores originates from the fact that whatever a business wants to present will not always be a typical enquiry of the user. A user query can be very specific to an individual user whereas the business responses are supposed to be diluted to a certain level to accommodate the diverse variety of user audiences. Also, dips in the ROUGE scores were noticed when the LLM presented creative responses with additional verbosity to make the answer interesting for the user. The response analysis also hints at the complementary steps like performing preprocessing on user queries for noise removal and spelling corrections, and further engineering of the prompts to make them richer in terms of context for the extended scope of the research.

We observed that the LLM responses are appealing to users, and are blended well in the dialogue flow of DCA. We have successfully integrated the abilities of discriminative NLU and generative response creation over the partial scope of a DCA. The composite architecture for DCAs effectively enabled businesses in controlling the narrative around conversations with their customers, while ensuring the customer's satisfaction with using the DCA.

5 Conclusions

Dedicated conversational assistants have been successfully automating customer interactions for businesses, while ensuring that there is no accidental leakage of protected information or circulation of negative impression on the brand. It is able to do so by operating within a defined scope, adhering to a dialogue flow intended at acquiring opportunities, and controlling responses to user messages. The recent progress in gen-AI frameworks has kickstarted a migration of techniques in conversational AI from traditional discriminative methods to generative methods with LLMs. However, there exist numerous concerns over this migration with respect to data privacy, lack of control over responses, guided dialogue flow and so on.

In this paper, we propose a composite architecture for DCAs aiming at utilizing the best that discriminative and generative models have to offer. We built an NLU pipeline for DCAs in a discriminative fashion to enable information extraction from user messages. A selective routing of user messages based on intents is designed with respect to considerations arising from the dedicated scope of operation. We utilised gen-AI for response generation to user queries, while providing ample control over the overall responses delivered by the DCA. Our solution can be used to build confidence in business owners about the utilization of gen-AI in their customer interactions, while providing a natural connect for customers' with the DCA.

References

[1] M. AI. Retrieval augmented generation: Streamlining the creation of intelligent natural language processing models, Sep. 2020.

[2] T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol. Rasa: Open source language understanding and dialogue management. *CoRR*, abs/1712.05181, 2017. URL http://arxiv.org/abs/1712.05181.

[3] G. Brockman, M. Murati, and P. Welinder. OpenAI API. https://openai.com/blog/openai-api, June 2020.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

[5] G. Caldarini, S. Jaf, and K. McGarry. A literature survey of recent advances in chatbots. *Information*, 13(1), 2022. ISSN 2078-2489. doi: 10.3390/info13010041. URL https://www.mdpi.com/2078-2489/13/1/41.

[6] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt, 2023.

[7] H. Chase. LangChain Docs. https://docs.langchain.com/docs/, 2022.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova. 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, June 2019.

[9] N. Dziri, S. Milton, M. Yu, O. Zaiane, and S. Reddy. On the origin of hallucinations in conversational models: Is it the datasets or the models?, 2022.

[10] M. Engelbach, D. Klau, F. Scheerer, J. Drawehn, and M. Kintz. Fine-tuning and aligning question answering models for complex information extraction tasks, 2023.

[11] Huggingface. Pretrained models. https://huggingface.co/models, June 2020.

[12] Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig. Active retrieval augmented generation, 2023.

[13] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 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, April 2020.

[14] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence

pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[15] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. 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, 2020. Curran Associates Inc. ISBN 9781713829546.

[16] F. Liu and Y. Liu. Correlation between rouge and human evaluation of extractive meeting summaries. In *Proceedings of ACL-08: HLT, Short Papers (Companion Volume))*. ACM, 2008.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692:1–13, July 2019.

[18] B. Luo, R. Y. K. Lau, C. Li, and Y.-W. Si. A critical review of state-of-the-art chatbot designs and applications. *WIREs Data Mining and Knowledge Discovery*, 12(1):e1434, 2022. doi: https://doi.org/10.1002/widm.1434. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1434.

[19] A. Merchant, E. Rahimtoroghi, E. Pavlick, and I. Tenney. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44. Association for Computational Linguistics, Nov. 2020.

[20] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013.

[21] OpenAI. Gpt-4 technical report, 2023.

[22] J. Qiu, L. Li, J. Sun, J. Peng, P. Shi, R. Zhang, Y. Dong, K. Lam, F. P.-W. Lo, B. Xiao, W. Yuan, N. Wang, D. Xu, and B. Lo. Large ai models in health informatics: Applications, challenges, and the future. *IEEE J Biomed Health Inform.*, Sep. 2022. doi: 10.1109/JBHI.2023.3316750.

[23] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[25] D. Reynolds. *Universal Background Models*, pages 1349–1352. Springer US, Boston, MA, 2009. ISBN 978-0-387-73003-5. doi: 10.1007/978-0-387-73003-5_197. URL https://doi.org/10.1007/978-0-387-73003-5_197.

[26] A. Rietzler, S. Stabinger, P. Opitz, and S. Engl. Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, page 4933–4941, May 2020.

[27] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108:1–5, Oct. 2019.

[28] W. Shen, Y. Gao, C. Huang, F. Wan, X. Quan, and W. Bi. Retrieval-generation alignment for end-to-end task-oriented dialogue system, 2023.

[29] Z. Sun. A short survey of viewing large language models in legal aspect, 2023.

[30] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 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, Dec. 2017. Curran Associates Inc.

[32] K. Vijayan and O. Anand. Language-agnostic text processing for information extraction. In *Computer Science Information Technology (CS IT) Proceedings*, number 23. AIRCC Online, December 2022.

[33] X. Zhang, M. Xia, C. Couturier, G. Zheng, S. Rajmohan, and V. Ruhle. Hybrid retrieval-augmented generation for real-time composition assistance, 2023.

[34] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, H. Peng, J. Li, J. Wu, Z. Liu, P. Xie, C. Xiong, J. Pei, P. S. Yu, and L. Sun. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt, 2023.