Natural Language Processing in Chatbots: A literature overview of current approaches and transformer technologies

Bachelor's program in Computer Science

Submitted on: [XX.XX.XXXX]

Table of Contents

1. Introduction

The transformations occurring across a multitude of fields have been largely influenced by the swift development of artificial intelligence, with the creation of conversational agents, which are more commonly known as chatbots, that stands out prominently. Built upon developments in natural language processing (NLP), these systems have become vital in areas such as customer service, education, and healthcare. Significant progress, particularly in enhancing the core capabilities of conversational systems, has been driven by transformer architectures among these technological improvements. The advancement and effectiveness of chatbots within NLP is investigated in this thesis through the lens of modern transformer models, utilizing a diverse array of current academic and practical viewpoints.

The way in which transformer-based architectures such as BERT, GPT, and related models have altered the development and performance of conversational agents, when held up against traditional NLP methods, is the central research question that is tackled in this work. Substantial limitations were encountered in managing linguistic ambiguity, preserving dialogue coherence, and scaling to diverse application domains, and previous chatbot generations were greatly molded by rule-based methods and statistical learning. An improved language understanding, and greater versatility across use cases, as well as more robust context awareness, have been enabled through the introduction of transformer models, featuring self-attention mechanisms and bidirectional context processing. Both the mechanisms and practical outcomes of transformer-based chatbots are examined in this study, by providing a comprehensive literature review, in which empirical findings, remaining challenges, and considerations for future advancement are highlighted.

Upon the analysis of thorough literature and critical evaluation of peer-reviewed research, the methodology that underpins this thesis was founded. The study makes use of comparative analysis of model architectures, performance metrics (e.g., BLEU, ROUGE, F-measure), and real-world implementation strategies, drawing on key studies such as Vaswani et al.’s original transformer model, along with more recent studies on BERT, GPT, and transformer-based chatbot applications. Dialogue coherence and adaptability were assessed through qualitative means which complemented quantitative benchmarking. The findings' reliability and relevance is ensured with the usage of approaches such as historical analysis, comparative synthesis, and source criticism.

It has been consistenly demonstrated in recent research that transformer architectures have

elevated performance standards within NLP and chatbot development, and new benchmarks on tasks like text classification, dialogue generation, and intent recognition were achieved. There are persistent notable challenges, which consist of the high computational requirements for training large models, ethical dilemmas in relation to bias and fairness, and the environmental effects of large-scale machine learning. The exploration of methods, such as quantization, pruning, and the development of more efficient transformer variants, has started in literature as a response to these subjects, showcasing ongoing areas of refinement and innovation.

A logical and sequential exploration of these subjects has been considered in the design of this thesis. Traditional rule-based and statistical methods will be compared to the emergence of transformer models in Chapter 2, in order to outline the evolution of NLP. The technical architecture and implementation of transformers in modern chatbots is examined in Chapter 3, by focusing on their core components and deployment strategies. The performance of transformer-based chatbots is assessed in Chapter 4 with the usage of established evaluation metrics, and current limitations are being discussed. Forward-looking trends, which includes emerging technologies and expanding applications, along with ongoing methodological and ethical considerations are taken into account in Chapter 5. The main findings are synthesized in the concluding chapter, which also includes a discussion of limitations, and identification of priorities for future studies.

2. Evolution of Natural Language Processing

The development of natural language processing has changed dramatically, from rule-based approaches to neural networks. Studying this evolution provides the basic background for modern chatbots which are able to take context into account, and it is the first step of this thesis.

2.1 Traditional NLP Approaches

The role of traditional natural language processing (NLP) in chatbot development and techniques has been paramount in enabling the first iterations of chatbots, though it has also been very limiting in their development. Statistical methods, such as n-grams and Hidden

Markov Models (HMM), represented the beginning of chatbots driven by data by employing probabilistic models for sentence recognition and predicting word sequences. By making chatbots probabilistically informed when selecting which responses to output, these models brought chatbots forward from the rigidness of rule-based chatbots. However, the inability of statistical approaches to capture longer-distance semantic dependencies led to coherence issues in longer conversations between users and chatbots. This was a crucial point in the development of chatbots as it highlighted the shortcomings of these techniques and provided a threshold for any further advancements to surpass them (Ayotunde & Cavus, 2025, p. 1).

ELIZA, ALICE, and Dr. Sbaitso are prime examples of rule-based chatbots. These chatbots employ decision trees and handcrafted patterns that allow them to output responses to user input by matching certain keyword-response pairs from static response sets. However, these chatbots were not able to handle any ambiguity due to the nature of their rule sets. ELIZA, for example, had only around 200 rules implemented into its system, and ALICE only around 41,000. Despite these limited capabilities, rule-based systems have laid the groundwork for understanding and solving fundamental dialogue challenges of modern chatbots (Huang, 2021, p. 5; Cîmpeanu, 2023, p. 4). Furthermore, the chatbots could not keep track of conversational context or dialogue history and as a result could not appropriately handle pronoun referencing, ellipses, and multi-turn coherence, making them often give irrelevant responses.

Though limited in capabilities, the rule-based systems have set a good benchmark in the history of chatbot development as they helped establish the limitations and demands of dialogue in the real world, such as scalability and multilingualism (Huang, 2021, p. 5; Cîmpeanu, 2023, p. 4).

Modular NLP pipelines (i.e., Stanford CoreNLP), by incorporating components such as tokenization, POS tagging, named entity recognition, and sentiment analysis, became widely adopted for chatbots to process linguistic input. This would allow the chatbot to interpret the structure of sentences from inputted text, extract important information (such as date, place, people, etc.) and sentiment, and utilize basic logical reasoning techniques, moving beyond solely pattern-based responses (Kumari & Manjula, 2024, pp. 5-8; Ayotunde & Cavus, 2025, p. 1). Yet, the hand-engineered features used for each of these models still limit the chatbot's hierarchical abstraction and multi-turn reasoning about user intention. While sentiment analysis implemented in these systems has been crucial in crafting responses appropriate for users, it does not allow chatbots to understand emotional subtext, which inhibits their ability to maintain a consistent understanding of the situation throughout a conversation.

Parry and Dr. Sbaitso, for example, are both domain-specific chatbots that focus on mental health support. These systems offered preliminary relief to users through direct counseling and could act as a simple mood booster (Cîmpeanu, 2023, p. 4). But, as mentioned before, rule-based chatbots were incapable of solving the problem of scalability and often gave unrelated answers in realistic use cases.

Another major problem for chatbots utilizing traditional NLP systems is the need for multilingual chatbots. Statistical and rule-based approaches in the field of NLP were designed for one specific language and lack multilingual awareness (Razumovskaia et al., 2021, p. 2). This has limited chatbots with NLP techniques to working in monolingual dialogue systems. The limited data available in many different languages, in addition to the problem of transfer learning for low-resource language use cases, has challenged the creation of multilingual dialogue systems and set benchmarks for contemporary models.

With traditional approaches in sequence classification techniques, chatbots often used keyword matching and shallow statistical techniques for tasks such as intent classification. However, these models were still unable to properly identify user intentions due to their inability to deal with the ambiguity of text and compositional utterances. As a result, in real-world chatbot use cases, these systems failed to capture the more nuanced intentions of users. By not incorporating contextual or semantic information in intent detection, these methods are limited to only being able to predict simple intent structures and generalize to similar examples, leading to poor performance in out-of-domain users. As a result, these limitations drove research into implementing neural networks to solve this problem by incorporating higher-order contexts and syntactic information in these tasks. Traditional machine learning techniques also act as benchmarks for modern ones by measuring accuracy in areas such as F1 in dialogue act classification (Yee & Soe, 2024, pp. 6-7).

In summary, traditional NLP techniques have assisted chatbots in early advancements and pointed out critical limitations in their development. Even though these methods have been shown to be limiting, they set the stage for improvements to come from data-driven chatbot systems.

2.2 Emergence of Transformer Models

The transformer, proposed by Vaswani et al. in 2017, changed the landscape of natural language processing by departing from the RNN/LSTM-based models with a self-attention mechanism. This mechanism allowed each word to attend to every other word irrespective of distance, thereby capturing long-term dependencies across sequences, as well as enabling parallelization and reducing training time. This facilitated training at a large scale with conversational datasets and rapid advances in chatbot technology (Sharma et al., 2025, p. 1; Ren, 2024, p. 1).

Transformer models have demonstrated high scalability as they do not depend on fixed-size memory representation. The architecture is capable of deep contextualization and processing long sequences. This is necessary for multi-turn conversation in chatbots, where maintaining the context of previous turns and remembering what was said is crucial. Unlike RNNs, transformers' attention-based mechanism greatly enhances the learning and performance of different NLP tasks. This made the transformer the de facto standard architecture in conversational AI. Its influence also reaches beyond text processing to other domains such as vision and speech and to multimodal applications such as multimodal chatbot systems (Sharma et al., 2025, p. 1; Ren, 2024, p. 1).

Deep contextualization in transformers is achieved through the use of self-attention and multi-head attention, where self-attention allows the chatbot to refer back to previously said phrases throughout a multi-turn dialogue, circumventing the issue of not being able to access phrases outside of a context window in RNN/LSTM-based models. Multi-head attention allows focusing on various aspects of sentences such as syntax, semantics, etc. This attention-based mechanism solves major problems for chatbot interaction, namely context loss, named entity resolution and pronoun disambiguation, by retaining a global dialogue context for transformers to reference throughout the entire conversation, facilitating turn coherence and relevance. According to studies, this contextualization achieves higher coherence and more consistent multi-turn dialogues when compared to other systems (Sharma et al., 2025, p. 1; Wu et al., 2025, p. 3; Griol et al., 2023, p. 2).

BERT (Bidirectional Encoder Representations from Transformers), published by Google in 2018, enabled a further advancement by pretraining a bidirectional model. A bidirectional encoder captures the preceding and succeeding contexts of a word in a sequence, resolving issues of polysemy, context-dependent interpretation and word ambiguity. BERT achieved state-of-the-art results on a multitude of NLP benchmark datasets, including GLUE score,

MultiNLI accuracy and SQuAD F1 (Devlin et al., 2019, p. 1; Wu et al., 2025, p. 2; Sharma et al., 2025, p. 3). It also led to major improvements in tasks requiring context-sensitive understanding, such as question answering, natural language inference, and sentiment analysis, which are crucial to building better chatbots. Furthermore, BERT can also be employed as a pretraining backbone, enhancing intent detection and slot-filling tasks in various languages and domains (Devlin et al., 2019, p. 1; Wu et al., 2025, p. 2).

Generative models like GPT and T5 also build upon the transformer by generating high-quality text for applications such as chatbots. GPT scales the transformer to high-quality and large-scale generation. The transition from GPT-1 to GPT-3 shows the difference scaling in parameters can make. GPT-3 uses 175 billion parameters, a drastic increase from the previous versions and is capable of producing human-like text across a diverse number of tasks, topics and domains (Griol et al., 2023, p. 2). T5 is a unified text-to-text generative transformer framework in which NLP tasks are framed as text generation and are applicable in chatbots for their response generation capabilities, negating the need for different chatbot architectures to perform diverse tasks. In a study performed by Ren, T5 and GPT-3 are able to outperform humans in SQuAD question answering, receiving an F1 score of over 90% (Ren, 2024, p. 2; Wu et al., 2025, p. 2; Griol et al., 2023, p. 2; Sharma et al., 2025, p. 3). The advancements generative transformer-based models have brought to NLP are unprecedented, with each model contributing revolutionary changes for chatbots.

Furthermore, transformers have drastically changed the chatbot market, moving them from a prototype stage to mainstream products and applications in several sectors such as retail, health, technology, etc. The global spending of chatbots in retail reached a projected total of $142 billion in 2024 and transformers can be attributed to this success (Griol et al., 2023, p. 1). The ability of transformer-based chatbots to work effectively across several languages, or to allow for multilingual or cross-lingual applications, contributes to their commercial adoption, making them available to diverse regions and communities throughout the world (Griol et al., 2023, p. 1). Case studies show that users report greater overall satisfaction, perceive greater benefits and have higher engagement with transformer-based chatbots compared to other architectures, further contributing to their commercialization. This makes transformer-based chatbots a valuable commodity (Griol et al., 2023, p. 1).

Despite the positive impact on performance made with transformers, their high computational complexity makes them a difficult architecture to adopt by most individuals and organizations that have limited budgets and infrastructure. However, research on model

optimization is abundant for the transformer architecture, with methods such as quantization and pruning used to tackle the computationally heavy nature of the architecture. Quantization lowers the precision of the weights of the model, leading to the model's size being reduced with minimal impact on performance. For example, 8-bit integer quantization results in a 4x reduction of the model's size, with minimal loss to performance for BERT (Ren, 2024, p. 6). In contrast, pruning reduces the weights of a transformer model, resulting in an accuracy loss as well. Despite this, one can prune up to 70% of weights from the transformer architecture and not observe any drastic decline in accuracy (Ren, 2024, p. 6). These research fields enable the adoption of transformer-based chatbots with minimal costs or resources.

Overall, the transformer revolutionized the NLP field and chatbot development in ways never imagined. The introduction of the self-attention mechanism and the development of novel transformer-based models, as well as their application across diverse fields, has propelled chatbot performance, functionality and adoption to unprecedented levels. The development of generative models has led to high-quality, real-time text production from chatbots. Furthermore, optimization methods such as quantization and pruning have paved the way for increased availability of these technologies, leading to chatbot development to become increasingly more accessible. However, ethical concerns as well as the large cost of adopting this technology need to be further addressed in order for transformer-based chatbots to become widely available to all communities.

Humanized Version in English:

3. Transformer Architecture in Modern Chatbots

Transformers have transformed the core architecture of chatbots, ushering in a new generation of systems capable of complex interactions and contextual understanding. Self-attention and multi-head mechanisms are the cornerstones of this sophisticated architecture, enabling chatbots to generate more relevant and nuanced responses. Following the earlier advancements and the historical context detailed in the previous section, transformers further mark a significant leap forward in chatbot technology.

3.1 Core Components

The noteworthy effectiveness of transformer-based chatbot systems is derived from a central set of components, by which advanced natural language processing (NLP) tasks are enabled with precision and scalability. The capabilities of conversational AI have been significantly reshaped by the mechanisms underlying these systems through the addressing of traditional models' shortcomings and the pioneering of methodologies that are built on deep contextual reasoning and large-scale parallelization.

The self-attention mechanism stands as a foundational element within the transformer architecture, ensuring that every token in an input sequence can attend to all other tokens at the same time, instead of being restricted through the sequential constraints present in earlier models like Long Short-Term Memory networks (LSTMs) and Recurrent Neural Networks (RNNs). The ability for chatbots to preserve context throughout extended dialogues is guaranteed through this mechanism, since the model is allowed to capture relationships and dependencies between distant elements within the input text. Coherence throughout a conversation is maintained through transformers by referencing the entirety of the user's input at any given time, unlike prior architectures, which often struggled with long-range dependencies and gradually lost information across sequences. Especially valuable in scenarios in which pronoun resolution is required, such as identifying antecedents located in complex dialogues, this feature enhances logical flow and elevates user satisfaction (Sun, 2023, p. 2; Naik, 2024, p. 18; Jurafsky & Martin, 2024, p. 6).

An additional crucial advantage comes from the parallelization capabilities presented by the transformer's architecture. The parallel nature of attention mechanisms is leveraged by transformers to efficiently manage significant amounts of data, while RNN-based models process tokens in a sequential manner, which inherently limits computational efficiency and scalability. Especially suitable for high-throughput real-time applications are transformer-based chatbots because of this structural efficiency, as input processing is able to be distributed across a multitude of computational units. Transformers, as a result, display excellence in large-scale deployment scenarios in which high interaction volumes and diverse languages must be accommodated with minimal latency by chatbots. A dramatic reduction in the time needed for processing and response generation is achieved through this parallelization, ensuring that the demands of multilingual and high-load environments can be met through transformer-based systems (Naik, 2024, p. 14; Sun, 2023, p. 2).

The self-attention process is enhanced through the multi-head attention mechanism through which the model is enabled to simultaneously focus on a multitude of linguistic relationship types that exist within the same sequence. Independent operation is conducted by each attention head, capturing varied aspects of language like semantics, syntax, and discourse-level features. Adaptability to the multifaceted nature of human communication is allowed to the chatbot through this, enabling navigation through complex and multi-turn dialogues with greater precision. The identification of entity relationships may be learned through one head, while another focuses on understanding user sentiment, as an instance. An increase in the model's contextual awareness is achieved through the ability to process multiple linguistic dimensions concurrently. Chatbots are also ensured to remain contextually relevant and consistent throughout extended interactions because of it. Traditional single-perspective processing methods, which often were unable to capture the full depth of linguistic complexity, are significantly outperformed by multi-head attention as a result (Jurafsky & Martin, 2024, p. 6; Sun, 2023, p. 2).

Deeper contextual reasoning abilities within the model are achieved through the transformer’s stackable block architecture, which is composed of repeated layers containing attention and feed-forward networks. Hierarchical language representations are accumulated through the transformer via the stacking of these layers, enabling the dynamic integration of both high-level (e.g., pragmatic) and low-level (e.g., lexical) information. Performance of complex reasoning is allowed to the model through this architecture, which is critical when it comes to accurately interpreting user intent, especially in ambiguous or nuanced exchanges. The ability to handle sophisticated dialogue tasks that necessitate deeper abstraction and contextual understanding compared to what traditional models were capable of is made possible through the employment of as many as 96 stacked layers within state-of-the-art models such as GPT-3 (Jurafsky & Martin, 2024, pp. 2, 10; Sun, 2023, p. 9). The performance ceiling for conversational AI is significantly elevated through this approach, enabling systems to produce more refined and human-like interactions.

Solutions such as parameter-sharing techniques featured in models like ALBERT have come about due to efforts in addressing the computational cost of transformers. A reduction in memory requirements is achieved through ALBERT via the reusing of parameters across layers, without sacrificing performance. Scalability issues are directly addressed and organizations that have limited computational resources are ensured to be able to deploy high-performing chatbot systems without facing prohibitive costs through this design choice. These methods, when subjected to empirical studies on such parameter-sharing

approaches, show that they achieve competitive or even superior results when held up against baseline transformer models, which underlines their importance in broadening access to advanced NLP applications (Sun, 2023, p. 6; Pressel et al., 2022, p. 1). A practical solution for resource-constrained environments is presented through this innovation, but it also stands as an important stride when it comes to making transformer technologies accessible across diverse sectors and industries.

The superior performance of transformer-based chatbots when held up against traditional NLP systems has been consistently confirmed through empirical evidence. Traditional approaches are consistently outperformed in dialogue generation and intent detection tasks even by lightweight transformers, which are substantially smaller than models such as BERT-base, for example. The fact that transformers, even when equipped with significantly fewer parameters, excel in discerning complex conversational nuances, which aligns with user intent more accurately, is revealed through techniques such as linear and mutual information probing (Pressel et al., 2022, p. 1). Generative transformer chatbots additionally attain heightened scores in F1 metrics, ROUGE-L, and BLEU across conversational datasets, demonstrating their capability to produce coherent and contextually relevant responses containing fewer errors. The transformative impact of transformer architectures that exists within the field of conversational AI is highlighted through these measurable improvements, as they establish a new benchmark for effectiveness and reliability in chatbot systems (Esfandiari et al., 2023, p. 7).

To conclude, the landscape of chatbot technologies has been redefined through the core components inherent in transformer-based architectures, providing unmatched performance when it comes to contextual reasoning, processing, and scalability. These progressions, which are supported through innovations like parameter-sharing techniques, multi-head attention, and self-attention, tackle long-standing challenges present in natural language processing and establish the groundwork for consistent progress within conversational AI.

3.2 Implementation Methods

The implementation of transformer-based chatbots entails a variety of strategies and methods designed to improve performance, adaptability, and real-world deployment, while overcoming the inherent complexities and computational challenges they present. A key aspect in developing high-quality and adaptive models is the choice of training datasets and

pretraining techniques. Using diverse data sources, such as the Reddit corpus, online forums, and business reviews, allows for greater generalization. Although such diverse datasets are advantageous for enhancing adaptability and response accuracy, biases may be present in the source data and consequently be learned by the model. Future research should consider ways of filtering or mitigating any biases in the source data (Pressel et al., 2022, pp. 4-5).

The recent introduction of lightweight transformer models can achieve high-quality dialogue representations while minimizing computational costs. With a Byte Pair Encoding (BPE) vocabulary of 30,000 tokens and eight layers/attention heads, these models can achieve similar performance as larger transformer-based models but with much less computational power. In some intent detection tasks, these lightweight models outperform BERT-base despite being roughly three times smaller. With fewer parameters to optimize, models are faster and require less computational resources, but the trade-off might be a reduced capability to capture domain-specific patterns (Pressel et al., 2022, p. 1).

In addition, various specialized pretraining objectives, such as Masked Token Modeling (MTM) applied after the initial training of the transformer on a general corpus like C4, has been used to enhance the model’s conversation-specific properties. Applying different pretraining tasks after general training improves the chatbot's ability to produce dialogue as well as intent detection capabilities. By applying more training phases on specific data, the chatbot learns to excel at a narrower range of tasks (Pressel et al., 2022, p. 5).

The success of these specialized pretraining strategies is further shown in few-shot scenarios, where the amount of training examples available is limited. Lightweight transformers can demonstrate high generalization capabilities even when a continuous updating and training loop is employed (Pressel et al., 2022, p. 6). While these specialized methods can work well when data is constrained, they have difficulty scaling up and capturing domain complexities that are not well represented in the training data. Hybrid approaches that incorporate few-shot and synthetic data generation can be employed to address this shortcoming.

The practical aspect of implementing a chatbot is important, and reducing the computational footprint via quantization and parameter reduction methods makes deployment feasible for many real-world systems with limited resources. Very compact models can be deployed on personal machines with similar performance, reducing the memory requirements of a model from 444MB to 59MB in ConveRT (Henderson et al., 2020, pp. 2, 9). ConveRT-trained

models require less specialized hardware due to quantization and allow faster and longer training cycles at a much lower cost with nearly identical performance. The trade-off of highly optimized transformer models can be an impaired ability to model complex conversational nuances, which may diminish the quality of the chatbot. Sparsity-aware training is an emerging technique to mitigate this trade-off, allowing for more memory usage reduction while retaining performance.

Further efficiency can be achieved by fixing the encoding layers of transformer models, as is done in intent classifiers, where only the final classification layers are trained, which leads to dramatically faster inference times. Intent classifiers built on ConveRT can be trained in forty times less time than their BERT-LARGE counterparts (Henderson et al., 2020, p. 9). Intent classifiers also have higher transparency compared to end-to-end training. New intents can easily be added without retraining, or old intents can be updated incrementally without forgetting past knowledge.

The ConveRT architecture enables cheaper experimentation and iterative development cycles since it takes about $85 to pretrain from scratch and has no cost beyond deployment (Henderson et al., 2020, p. 6). Making the implementation affordable increases the potential deployment possibilities and decreases the barriers to entry.

Multi-context transformer variants can leverage several previous conversational turns when making predictions and have higher performance on benchmarks due to their greater ability to resolve contextual ambiguities compared to one-context transformers (Henderson et al., 2020, p. 2). Even with the extra complexity and larger size (73MB), they still provide a significant benefit when the application relies on multiple conversational turns to derive meaning.

Recent improvements in dialogue quality are due to a staged pretraining method in conjunction with a method for adversarial training. With two phases, general and adversarial, transformers saw improvements across three metrics on the Cornell Movie-Dialog dataset: BLEU4, ROUGE-L, and Meteor (Esfandiari et al., 2023, pp. 7, 9). The adversarial training regime allows for more diversity and is less repetitive than the standard one-stage training methodology, in addition to mimicking the variable nature of a human conversation (Esfandiari et al., 2023, p. 1). There may be potential flaws for an adversarial training regime such as mode collapse, a type of imbalance in which certain output patterns or variations are disproportionately represented due to their prevalence during the training process.

A further enhancement is the domain adaptation technique. In general-purpose pretraining, transformer models are trained to perform adequately in multiple domains, while domain adaptation enables a specialized performance in a certain field by continuing the training of a general-purpose model on a domain-specific corpus.

Further language modeling performance enhancements can be achieved with new improvements of transformers, such as Layer Normalization and the Weighting of Residual Connections (Moon et al., 2023, p. 6). An improvement of the Layer Normalization technique by moving it further in the pipeline helps to alleviate training instability, which makes training go more quickly. Improving transformer scalability is also made possible through a technique called Residual Connection Weighting.

With the use of reinforcement learning, advancements to positional embeddings have enabled transformer models to generate longer and more contextualized conversations. This provides a more complete and more consistent experience for the application users (Moon et al., 2023, p. 6).

To go beyond text, one approach used by transformer models is extending their context capabilities by using them in automatic speech recognition systems. With the use of the whole past conversational history, a transformer model can substantially reduce the number of ASR errors, a vital feature in real-world, multi-turn conversations (Hori et al., 2020, p. 4).

4. Performance and Limitations

To assess the effectiveness of transformer-based chatbots, it is crucial to comprehend how chatbot performance is evaluated and to identify the issues with which they struggle. This involves evaluating the metrics of linguistic quality and examining the limitations regarding resources and robustness to gain insights into both the accomplishments and difficulties presented in this thesis.

4.1 Evaluation Metrics

Transformer-based chatbots have propelled advancements in natural language processing

significantly, a fact underscored by their consistent outperformance against traditional models when evaluation metrics like BLEU, ROUGE-L, F-measure, and Meteor are considered. The ability of these models to encapsulate linguistic context and dependencies is assessed quantitatively by these metrics, thereby substantiating the transformative influence of transformers on dialogue systems. To illustrate, findings derived empirically from the Cornell Movie-Dialog and Chit-Chat datasets reveal considerable performance enhancements, with BLEU4 and ROUGE-L scores peaking at 0.96 and 0.965, respectively, indicating the capability of transformers in producing responses that are highly coherent and contextually precise. The crucial role that quantitative benchmarks play in monitoring the effectiveness of linguistic modeling is highlighted through these results, which provide a firm basis for assessing transformational progress in conversational AI (Esfandiari et al., 2023, p. 7). It is vital to acknowledge the limitations of these metrics in comprehensively capturing dialogue quality, especially when intricate user interactions or open-ended conversations are under examination, even though they facilitate reproducibility and comparability across models.

The dominance of transformer architectures over traditional models is further highlighted by the unprecedented results of architectures such as DLGNet on multi-turn dialogue benchmarks. As an example, DLGNet has attained the highest scores in BLEU, ROUGE, and distinct n-gram evaluations across datasets like Movie triples and Ubuntu dialogue, notably surpassing RNN-based systems like VHRED and hredGAN. These advancements are attributable to the architectural innovations intrinsic to transformers, including attention mechanisms and parallel processing, which make possible the capture of subtle conversational cues and long-range dependencies. As a consequence, transformer-based chatbots are uniquely positioned to manage conversational context, thereby enabling them to generate diverse responses appropriate in their context. The capability of transformers to circumvent generic or repetitive outputs, which was a persistent limitation in earlier architectures, is also underscored through the use of distinct n-gram metrics. The continuous need for innovation to surmount emerging challenges in adapting to evolving user expectations and in maintaining response diversity is highlighted by these quantitative gains, despite their remarkableness (Olabiyi & Mueller, 2020, pp. 1, 5).

The redefining of baseline performance across an extensive array of NLP tasks, including chatbot applications, has been achieved through transformers such as BERT. The manner in which bidirectional context modeling and robust representation learning directly augment conversational coherence, relevance, and informativeness is demonstrated by BERT's state-of-the-art results across eleven benchmarks, exemplified by a 7.7% increase in GLUE

score and a 1.5-point increase in the SQuAD v1.1 F1 score. A novel empirical standard for chatbot development is signified through these metrics, against which subsequent innovations are evaluated. The adaptability and scalability of pre-trained transformer models is further underscored by the ability to fine-tune them for specific tasks, such as dialogue systems specific to particular domains. The reliance on scores such as BLEU and ROUGE alone has, however, spurred criticism regarding their inadequacy in the assessment of the comprehensive quality of conversational interactions, notwithstanding the persuasive results of these metrics. While the generation of factually accurate responses is an area in which transformers excel, they are still capable of producing outputs that, within the context of complex dialogues, are either pragmatically irrelevant or socially inappropriate, thereby illustrating a disparity between quantitative measures and the quality as perceived by users (Devlin et al., 2019, p. 1).

The sufficiency of traditional metrics in evaluating conversational AI has been questioned increasingly by empirical research. Despite the fact that transformer models consistently surpass earlier architectures in metrics like BLEU or ROUGE, for instance, these scores frequently fail to account for aspects of dialogue quality deemed essential, such as pragmatic appropriateness, long-term coherence, or user engagement. Owing to their focus on maximizing metric scores, transformers are known to produce plausible responses that are, nevertheless, irrelevant contextually, which has the potential to conflict with expectations in human communication. The necessity for complementary evaluation strategies that transcend automated scoring to take into account the subjective user experience is brought to light by this critique (Olabiyi & Mueller, 2020, p. 1; Esfandiari et al., 2023, p. 7). A growing consensus is indicative of the integration of task-based evaluations and assessments involving human-in-the-loop in order to bridge these gaps, thereby ensuring that the effectiveness of chatbots is in alignment with social contexts and real-world applications.

The necessity of a transition toward hybrid evaluation frameworks that coalesce quantitative and qualitative approaches is brought about by the limitations inherent in automated metrics. Methodologies that incorporate human-centric assessments alongside standardized benchmarks are advocated for in recent studies, given that automatic metrics, when used in isolation, are often deficient in capturing the full complexity of conversational dynamics. As an example, a more precise and thorough comprehension of chatbot performance is furnished through hybrid approaches that integrate human annotations pertaining to dialogue attributes such as coherence, empathy, engagement, and informativeness. The identification of shortcomings, such as the propensity of transformers to yield outputs that are misleadingly fluent but contextually inappropriate in ambiguous conversations, is facilitated

through this dual focus (Esfandiari et al., 2023, p. 7). The incorporation of qualitative evaluations characterized by nuance is crucial for propelling chatbot development forward, thereby guaranteeing that enhancements in technology translate into tangible improvements in user satisfaction.

The evolving needs of chatbot technologies and their users drive the ongoing transition toward evaluation frameworks that are more enriched. For instance, findings derived empirically intimate that depending solely on quantitative metrics has the potential to mask critical deficiencies within conversational systems, exemplified by their ineptitude in managing ambiguous user queries effectively or their predisposition toward generating repetitive responses. The potential to surmount these challenges is presented through an evaluation approach that is more balanced—incorporating both automated metrics and human judgment. A trajectory toward the creation of chatbots that not only possess technical proficiency but also harmonize with norms in human communication is established through bridging the divide between algorithmic performance and user experience via hybrid evaluation strategies (Devlin et al., 2019, p. 1; Esfandiari et al., 2023, p. 7).

In the final analysis, new benchmarks within conversational AI have been established by transformer-based chatbots by means of their exceptional performance across a spectrum of evaluation metrics. The significance of embracing hybrid evaluation methodologies—those that meld automated scoring with analyses centered on the human element—is, however, underscored through the limitations of these metrics. The comprehensive approach described will be indispensable in the accurate assessment of the overall effectiveness of chatbot systems as progression within the field persists.

4.2 Current Challenges

Transformer-based chatbots face several ongoing challenges that must be addressed. One of them consists of quantization and computational efficiency. Although 8-bit post-training quantization methods are successful in achieving model compression for transformer-based NLP models, it has been shown that 8-bit post-training quantization significantly hurts the performance of transformer encoder architectures due to mismatches in dynamic ranges of activation tensors for residual connections in transformer layers (Bondarenko et al., 2021, p. 2). Activation tensors with residual connections help in dialogue flow or turn segmentation, and activation outliers are essential to conversational transformer checkpoints since they

help detect key conversational tokens such as [SEP], (Bondarenko et al., 2021, p. 2). Recent work in activation quantization has shown that preserving 22% of the activation tensors with 16 bits and using 8-bit to quantize the rest shows performance parity with the original 32-bit floating-point model. However, this comes at a cost of significant engineering overhead and lower conversational quality (Bondarenko et al., 2021, p. 8). Another approach is ultra-low precision configuration using 4-bit weights and 2-bit token embeddings (Bondarenko et al., 2021, p. 9). This lowers the memory and computational demands to a minimum. Ultra-low precision configuration suffers less than 0.8% drop in GLUE score and requires delicate calibration to ensure that necessary attention patterns for high-quality conversational turns can be generated effectively (Bondarenko et al., 2021, p. 9). Therefore, quantization is not just about compression but also about preserving conversational capabilities.

The large data and compute requirement limits the deployment of transformer-based chatbots. A transformer-based chatbot requires pretraining on millions of Reddit threads or goal-oriented dialogues for generalization to various conversational tasks (Pressel et al., 2022, pp. 4–5). Without pretraining, transformer-based models for conversational AI struggle with tasks such as dialogue generation or intent detection. Although lighter transformer models such as BERT Lite and TinyBERT have been developed, empirical evidence suggests that the performance of these lighter models is on par with heavier models only if pretrained on larger domain-adapted corpus in the few-shot learning paradigm (Pressel et al., 2022, pp. 1, 6). Therefore, scaling on data and resources for pretraining exacerbates the disparities in AI. Besides, dominant linguistic/cultural norms are contained in large datasets (Pressel et al., 2022, p. 6).

Parameter reduction is required for the efficient deployment of transformer-based chatbots, but this also affects their performance in complex conversational tasks. Light transformer-based models such as BERT Lite and TinyBERT offer competitive performance on some popular NLP benchmarks. However, in order to achieve this level of performance, effective pretraining objectives and high-quality training datasets are needed. Pretraining for lightweight transformer-based conversational AI models requires specialized methods by progressing from general to domain-adapted corpus (Pressel et al., 2022, pp. 4–5). However, specialized pretraining leads to extra computational demands. There is a tension between parameter count and performance (Pressel et al., 2022, p. 1). Parameter count affects computational speed, which is critical for real-time applications. There is also the need to reduce the size of transformer-based chatbots, but this would require parameter reduction and often affects the chatbot's performance, since a reduction in parameter number would lead to difficulty in context tracking and intent recognition.

Moreover, there are ongoing modeling challenges. Transformer activations contain highly structured outliers. Structured outliers are caused by residual connections in attention networks (Bondarenko et al., 2021, p. 2). Activation outliers are an important statistic of a conversational transformer. Outliers help detect key conversational tokens such as [SEP]. This token helps indicate the separation between two different turns. Without activation outliers, conversational transformers lack the ability to correctly segment dialogues and may incorrectly interpret users' intents (Bondarenko et al., 2021, p. 2). Despite efforts to develop effective quantization methods, no universally effective solution for dealing with these activation tensor outliers in the dynamic range has been found (Bondarenko et al., 2021, p. 2).

The scalability of transformer-based chatbots is affected by computation economics and energy sustainability. Computational economics is important because transformers require a large amount of computation to run due to large-scale pretraining. For example, scaling on data and resources exacerbates the disparities in AI. Besides, the exponential scaling of transformer models has environmental and financial concerns, because it requires tremendous economic and energy expenditure (Bondarenko et al., 2021, p. 2; Sharma et al., 2025, p. 4). Furthermore, there are still concerns over the scalability of these models regarding accessibility and sustainability (Sharma et al., 2025, p. 1). Though transformers enable parallelizability, which improves computation economics, their heavy dependence on large amounts of data and computational power to pretrain makes it costly and unscalable (Sharma et al., 2025, p. 1). This limitation has caused the emergence of Google's T5 framework, which considers NLP tasks as text-to-text tasks and uses pretraining as the training criterion. Also, with ultra-low bit quantization and dynamic bit-width allocation, resource usage can be greatly reduced (Sharma et al., 2025, p. 3; Bondarenko et al., 2021, pp. 2–9). However, scalability often has a trade-off with the ability of deep contextual learning. For example, if a transformer-based chatbot is to be used in another domain that the original model has not been adapted for, domain adaptation needs to be performed. This has led to partial transformer training, which requires high computation due to extra computation and adaptation iterations for each domain (Sharma et al., 2025, p. 4). All this needs to be addressed by continuous learning and fine-tuning techniques on top of raw transformer scaling in general NLP (Sharma et al., 2025, p. 4).

In conclusion, for transformer-based chatbots to be effectively adopted, the ongoing challenges in the area of quantization, data efficiency, parameter efficiency, as well as scalability, need to be addressed.

5. Future Developments and Applications

Emerging technologies and potential applications are rapidly changing the chatbot development and application landscape. This section will illustrate cutting-edge advancements and discuss several interesting potential applications, continuing the previously discussed evolutions and advancements.

5.1 Emerging Technologies

Substantial promise for the scalable deployment of chatbots in environments with limited resources has been demonstrated through advancements in transformer efficiency via architectural modifications and parameter sharing. Models like ALBERT, which significantly reduce the number of parameters without compromising performance, represent one of the most notable innovations in this area. Techniques such as cross-layer parameter sharing, whereby the same weights are reused across multiple layers of the transformer architecture, achieve this reduction. Memory requirements and computational overhead are reduced by these mechanisms, which enable the deployment of complex natural language processing systems in more constrained environments, such as mobile devices and edge computing. The memory footprint of transformer models is significantly decreased through the adoption of parameter sharing mechanisms in ALBERT, which makes their deployment more feasible in resource-constrained environments without sacrificing performance (Sun, 2023, p. 6). High accuracy and inference speed are maintained despite a reduced parameter count, making models like ALBERT particularly well-suited for real-time chatbot applications, where latency must be minimized to ensure a seamless user experience. However, while critical accessibility barriers are addressed by these advancements by lowering hardware demands, concerns are also raised about potential limitations in scalability and performance when applied to more complex conversational scenarios. Focus on balancing these efficiency gains with the need for continued improvements in model robustness and adaptability across diverse application domains must be undertaken by future research.

Faster inference times have also been contributed to through cross-layer parameter sharing, particularly as exemplified by ALBERT. The usability of chatbot systems in real-world

contexts where responsiveness is paramount is directly impacted by this development. Real-time interaction quality is enhanced through faster inference, and the energy consumption associated with running transformer models is also reduced, which is an increasingly important consideration in sustainable AI development. Empirical evidence demonstrates that these architectural optimizations can maintain, and in some cases even enhance, a chatbot's capacity to comprehend and accurately respond to complex conversational queries (Sun, 2023, p. 6). A new precedent for achieving computational efficiency without significant sacrifices in performance is set by this. However, critical examination of the trade-offs inherent in such optimizations is also necessitated by these findings. For instance, the model's ability to learn highly specialized features critical for nuanced dialogue management might inadvertently be limited through parameter sharing, particularly in open-domain settings. Addressing this trade-off will be crucial in advancing the scalability and reliability of transformer-based chatbots.

Pathways for integrating advanced transformer-based chatbots in settings previously considered unfeasible are highlighted through the empirical successes of models employing parameter-sharing mechanisms. A reduction in operational costs is showcased by efficient architectures like ALBERT, making such technologies accessible to organizations with limited computational resources. Deployment in low-bandwidth regions or institutions unable to support high-end infrastructure could be enabled through this democratization of access. A practical response to one of the core limitations of large transformer architectures is provided by the development of ALBERT: their excessive hardware demands and high operational costs, which historically have restricted broad accessibility (Sun, 2023, p. 6). However, this potential democratization must be critically evaluated in light of the specific challenges posed by diverse deployment contexts, such as those requiring multilingual or domain-specific capabilities. While the reduced memory footprint of shared-parameter models is a step forward, ensuring their robustness across such varied applications will require ongoing innovations in pretraining objectives and data selection strategies. Furthermore, broader adoption of these technologies necessitates ethical considerations, particularly regarding data privacy and the risks of amplifying biases present in pretrained datasets.

The integration of adversarial learning and extended pretraining cycles in transformer-based chatbot models involves another area of emerging technology. A generator and discriminator framework is introduced through adversarial training, where the generator creates responses, and the discriminator evaluates them, driving iterative improvement through competition between the two components. Dialogue quality metrics, including BLEU4 and

ROUGE-L scores, have seen significant improvements because of this technique. Marked advancements in conversational naturalness and context relevance have been demonstrated through structured regimens involving hundreds of pretraining and adversarial learning cycles (Esfandiari et al., 2023, p. 7). For instance, training regimes involving 200 pretraining cycles followed by 400 adversarial learning cycles have resulted in BLEU4 scores of 0.96 and ROUGE-L scores of 0.965 on datasets such as Chit-Chat, far surpassing prior benchmarks (Esfandiari et al., 2023, pp. 7, 9). While quantitatively impressive results are yielded through these approaches, questions are also raised about scalability and the practical feasibility of such extended training in resource-constrained scenarios. The importance of curating datasets that are not only extensive but also ethically sourced and representative of varied conversational contexts is also underscored through the reliance on diverse, high-quality datasets for adversarial training. Additionally, adversarial frameworks could introduce instability during training phases if not carefully calibrated, warranting further exploration of optimization techniques to ensure reliable convergence.

Not only are quantitative benchmarks improved through extended training cycles, when combined with adversarial learning, but pathways are also opened for addressing long-standing challenges in chatbot development, such as mitigating generic or overly cautious responses. Chatbot capabilities in generating natural and contextually relevant dialogue responses are advanced through adversarial training, as suggested by the improvements in Chit-Chat and Cornell Movie-Dialog datasets (Esfandiari et al., 2023, p. 7). However, these advancements prompt critical inquiry into whether such improvements align with user-perceived conversational quality. While effective in refining linguistic fluency, adversarial systems may still fall short in capturing the nuances of user intent or contextual constraints in dynamic dialogues. The potential for integrating complementary techniques, such as reward modeling or human-in-the-loop feedback mechanisms, to refine the conversational depth achieved by transformer models is highlighted through this gap. The quantitative strengths of adversarial learning could be balanced with the subjective factors that shape user satisfaction and engagement through such hybrid frameworks.

The exponential scaling of transformer models, as exemplified by the transition from GPT-2's 1.5 billion parameters to GPT-3's 175 billion parameters, has further bolstered chatbot capabilities. Enhanced response coherence, deeper context retention, and the ability to produce nuanced conversational outputs have resulted from this growth (Sharma et al., 2024, p. 2). The ability to pretrain on vast and diverse datasets has directly contributed to the superior language understanding and multi-domain applicability of these large models, significantly improving their robustness to ambiguous user inputs (Sharma et al., 2024, p. 3).

These advantages are particularly evident in multi-turn dialogues, where context must be maintained across longer exchanges. However, the dramatic increase in model size also intensifies the challenges associated with computational resource requirements, training and inference times, and environmental impact. Although scaling has indisputably contributed to superior performance, it has also exacerbated disparities in the accessibility of state-of-the-art chatbot technologies, limiting their adoption to organizations with significant resources. Important questions about the long-term sustainability of growth-driven approaches to model improvement are raised through this. While techniques such as model compression and quantization offer partial solutions, a critical rethinking of whether scaling alone is the optimal trajectory for NLP advancements is needed. Moreover, the ethical implications of concentrating such technological power in a few organizations must be carefully evaluated.

The field has been significantly advanced through the introduction of bidirectional self-attention mechanisms in transformer models, such as BERT, by enabling improved context awareness and subtle relationship extraction. A more thorough capture of linguistic dependencies is allowed through analyzing both past and future context for each token, which is particularly advantageous for conversational flow (Sun, 2023, p. 5). A chatbot's ability to manage multi-turn dialogues, maintain conversational coherence, and accurately track entities and thematic threads over longer exchanges is significantly enhanced through these innovations. Better intent detection and response generation result from this improved context awareness, particularly in scenarios where ambiguous or context-dependent user inputs are prevalent. However, while superior sensitivity to linguistic nuances is provided through bidirectional mechanisms, considerable computational resources are also demanded, making real-time deployment challenging for many applications. To address this, hybrid approaches that combine the strengths of bidirectional and unidirectional modeling are being explored. The contextual insights achieved by bidirectional attention could be preserved while improving computational efficiency through these mixed architectures, offering new possibilities for dialogue management and response generation. Future innovations may also involve designing pretraining objectives specifically tailored to the demands of complex multi-turn dialogues, further optimizing the applicability of such mechanisms in real-world chatbot systems.

New research directions in transformer-based chatbot technology are set through these advancements in efficiency, adversarial training, scaling, and self-attention mechanisms. However, progress in this field will require a balance between leveraging these technological innovations and addressing their associated challenges, such as resource constraints,

accessibility disparities, and alignment with human-centered communication norms.

5.2 Potential Use Cases

Transformer-based chatbots have revolutionized various industries such as customer service and market research with responsive and user-friendly services. The use of these chatbots in customer service sectors like banking, e-commerce, and telecommunications has significantly reduced response times and boosted customer satisfaction. They automate customer inquiry resolutions and respond contextually, and they are constantly available, a crucial feature today (Ojha, 2024, p. 2). Through the use of large-scale transformer models like GPT-3, complex, multi-turn customer inquiries can be resolved, and escalations to human representatives have decreased significantly (Sharma et al., 2024, p. 9). However, the concern lies in the suitability of automated chatbot responses for situations involving difficult or sensitive customers. Automated responses may not be as apt as human interactions and thus may not be suitable for stressful or delicate situations. Hence, response suitability needs continuous evaluation and must integrate both automated metrics and human evaluations to ensure standards are met (Ojha, 2024, p. 4).

Another industry where transformer-based chatbots play an important role is the market research sector. This technology has streamlined data collection and analysis, offering an automated channel to gather and summarize unstructured data. The advanced NLP abilities enable more efficient survey administrations, where interactions are more personal and responsive. Some commercial chatbot successes such as ChatGPT and Google's Bard boast hundreds of millions of interactions, transforming the ways organizations gather and analyze market information (Sharma et al., 2024, p. 9). Nevertheless, the risk of bias in data interpretation looms large. The insights garnered from the data interpretation carried out by transformer-based chatbots will be of little value if they are not accurate.

Healthcare has made tremendous technical and societal advancements with the introduction of transformer-based chatbots. Large language models (LLMs) with strong image classification capabilities have accurately classified multiple pathological conditions (Chang, 2024, p. 1). The application of transformer-based chatbots in medical diagnostics can potentially help doctors determine the priorities of cases based on patient history, etc. These technologies also save time for both patients and doctors as they automate administrative tasks such as data collection (Ojha, 2024, p. 2). Chatbots are also used for therapy,

specifically cognitive behavioral therapy (CBT). Woebot, a transformer-based chatbot, provides CBT to patients who would otherwise not be able to afford or access therapy. This technology enhances the anonymity of patients and alleviates the burden of patient demand, as it is consistently available and can cater to many at the same time (Ojha, 2024, p. 2). However, concerns such as privacy, safety, and ethics emerge. As a chatbot makes autonomous medical decisions, its accuracy and reliability are important. Hence, this technology must adhere to a stringent validation protocol to ensure it is valid and accurate and follows medical guidelines (Ojha, 2024, p. 3). While transformer-based models are highly adaptable and have great application potential (e.g., remote monitoring or teaching), they also have drawbacks such as overfitting, as they learn from vast datasets. Therefore, the model's language may become overly formal, thus excluding those who are unfamiliar with the dataset. Further work needs to be carried out to improve data collection and validation.

Transformer-based chatbots have transformed the education sector as they are more engaging and deliver higher user satisfaction. Studies indicate that such chatbot interactions led to longer sessions and conversation turns (Ojha, 2024, p. 4), both indicators of engagement and perseverance (Sharma et al., 2024, p. 2). Due to their abilities to deliver contextually relevant responses, transformer-based chatbots cater for personalized learning. The personalized feedback from transformer-based chatbots enhances the user experience as the students gain confidence in their comprehension of the topic (Ojha, 2024, p. 2). These benefits ensure personalized education and foster learning, which benefits children, particularly when at home (Ojha, 2024, p. 2). A well-functioning chatbot would assist those who live far away from schools as it can be used remotely and independently (Sharma et al., 2024, p. 2). Nevertheless, it must be acknowledged that there are dangers to employing such technology. A potential danger is that chatbots tend to reinforce biases present in training data. While the correctness of chatbot responses in a learning environment is crucial, chatbots can misguide the learner (Ojha, 2024, p. 3). Bias detection needs to be addressed.

As transformer model capacity grew significantly, from GPT-2’s 1.5 billion parameters to GPT-3’s 175 billion, the quality of conversational flow, context, and overall user experience improved. The increased parameter count boosts the response coherence and intent detection abilities of a chatbot (Sharma et al., 2024, p. 2). As the chatbots became more effective and efficient, the use of this technology increased dramatically, with the number of chatbot interactions growing at a rate of 43% per year since 2019 (Sharma et al., 2024, p. 3). However, this rapid expansion presents the challenge of sustainability. Transformer models

require significant computing resources and data, which poses a challenge to organizations with limited resources. Since this is considered a public sector service, all members of the population should have equal opportunity.

Pretrained transformer-based models such as BERT and GPT, used in NLP and conversational AI systems, enable chatbots to achieve a higher level of intent detection, context management, and response diversity. The bidirectional self-attention mechanism enables the chatbot to grasp long-range dependencies from the context (Chang, 2024, p. 2) of the complete conversation, thus enhancing the dialogue flow and making entity tracking and conversational context consistent for the chatbot in the long term (Sharma et al., 2024, p. 2). The use of transformers in the context management phase has had a major impact. Due to the significant improvements to response diversity and context abilities brought on by the utilization of transformers, these systems are more adaptable to dynamic interactions within a broad range of industries. Nevertheless, transformer-based chatbots are not free of risks and potential ethical harms. For example, there are instances where a transformer-based chatbot has been shown to unintentionally reinforce harmful stereotypes (Ojha, 2024, p. 3) as it learned biased information from the dataset. As such, it is vital to thoroughly assess the potential for such unethical risks of a system and to take appropriate action in response (Sharma et al., 2024, p. 2).

6. Conclusion

The primary objective of this work has been to assess the role of transformer architectures in enhancing chatbot performance through a critical examination of relevant research in natural language processing. The research began with a key question: How have the improvements in NLP as well as transformer models enhanced conversational agents in comparison to rule-based and statistical approaches? By synthesizing existing literature and empirical data, this thesis successfully outlines the advancements and impact of transformer-based chatbots.

Thus, to summarize, traditional rule-based and statistical chatbots, while foundational, struggled with contextual understanding and generalization. Transformer models, such as BERT and GPT, provided context-aware capabilities, effective handling of long dependencies, and demonstrated superiority against previous models through BLEU, ROUGE, F1, and other benchmarks. Furthermore, these models are increasingly used in

customer service, education, healthcare, and market research, exhibiting high scalability, versatility, and implementation improvements like parameter sharing, quantization, and adversarial training, making chatbot technology readily available to diverse domains with varying budgets. Although transformative, challenges in efficiency, scalability, bias, ethical implementation, and low-resource and domain-specific adaptations persist.

This thesis contributes to the ongoing discussions within the academic and business communities in relation to the development of conversational AI. The findings confirm that the progress in NLP as well as transformer models has enabled chatbot capabilities surpassing rule-based or statistical methods. Further, transformer models have exhibited significant performance gains in response coherence, contextual sensitivity, and diversity when contrasted with older state-of-the-art RNN-based chatbots. Additionally, the research affirms prior studies on the significance of self-attention, data pretraining, and model scaling in the advancements of conversational AI. However, the analysis identifies ongoing challenges and limitations regarding efficiency, scalability, bias, ethical implementation, and adaptation to low-resource environments and specialized domains.

Certain limitations of this thesis include the exclusion of any original empirical analysis due to the use of secondary literature and published datasets. The study is restricted to the most commonly employed transformer models and evaluation benchmarks within NLP. The potential biases of selected literature and the fast evolution in NLP and the use of transformer architectures in industry applications may limit the generalizability and long-term viability of the results of this work.

The development of sophisticated evaluation benchmarks, efficient model optimization techniques, and effective data and bias reduction will enable further advancements of chatbot performance. By leveraging diverse and high-quality training datasets as well as data augmentation, chatbot technologies can achieve optimal performance in conversational tasks. Lastly, ethical frameworks for model implementation that emphasize fairness, transparency, and user privacy are critical for the responsible use of conversational AI.

My experience of engaging in this research has been enjoyable and has reinforced my appreciation of the intersection of technology and society and the importance of ethical considerations in chatbot development. As chatbot technologies become more pervasive in many aspects of life, it is essential to advance their efficiency, fairness, and responsibility. This thesis contributes to the dialogue about the technological and social implications of transformer-based chatbot systems and how these systems can bring value to humans.

Bibliography

Ayotunde, O. O., & Cavus, N. (2025). Natural language processing in conversational systems: An overview. Journal of Advanced Artificial Intelligence, 1(4), 15–18. https://jaaionline.org/archives/volume1/number4/ayotunde-jaai202418.pdf

Bondarenko, Y., Nagel, M., & Blankevoort, T. (2021). Understanding and overcoming the challenges of efficient transformer quantization. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 7947–7969. Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-main.627.pdf

Chang, Y. (2024). Comparison and analysis of large language models based on transformer. In Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024) (pp. 597-602). SCITEPRESS - Science and Technology Publications, Lda. https://doi.org/10.5220/0012960200004508

Cîmpeanu, I.-A. (2023). The chatbots and their role in the progress of society. Informatica Economică, 27(4/2023), 61-77. https://doi.org/10.24818/issn14531305/27.4.2023.05

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. https://au1206.github.io/assets/pdfs/BERT.pdf

Esfandiari, N., Kiani, K., & Rastgoo, R. (2023). A conditional generative chatbot using transformer model. Semnan University. https://arxiv.org/pdf/2306.02074

Griol, D., Kharitonova, K., Pérez-Férnandez, D., Gutiérrez-Fandiño, A., & Callejas, Z. (2023). CONVERSA: Effective and efficient resources and models for transformative conversational AI in Spanish and co-official languages. Annual Conference of the Spanish Association for Natural Language Processing. CEUR Workshop Proceedings (CEUR-WS.org). https://ceur-ws.org/Vol-3516/paper15.pdf

Henderson, M., Casanueva, I., Mrkšić, N., Su, P.-H., Wen, T.-H., & Vulić, I. (2020). ConveRT: Efficient and accurate conversational representations from transformers. Findings of the Association for Computational Linguistics: EMNLP 2020, 2161-2174. https://aclanthology.org/2020.findings-emnlp.196.pdf

Hori, T., Moritz, N., Hori, C., & Le Roux, J. (2020). Transformer-based long-context end-to-end speech recognition. INTERSPEECH, 5011–5015.

https://doi.org/10.21437/Interspeech.2020-2928

Huang, X. (2021). CHATBOT: DESIGN, ARCHITECTURE, AND APPLICATIONS [Senior capstone thesis, University of Pennsylvania]. University of Pennsylvania | School of Engineering and Applied Science. https://www.cis.upenn.edu/wp-content/uploads/2021/10/Xufei-Huang-thesis.pdf

Jurafsky, D., & Martin, J. H. (2024). The Transformer. Speech and Language Processing. https://web.stanford.edu/~jurafsky/slp3/9.pdf

Kumari, K. S., & Manjula, B. (2024). An artificial intelligence-based chat-bot system using natural language processing: A study. Futuristic Trends Information Technology, 3, 32–33. https://iipseries.org/assets/docupload/rsl2024A0F9B9DD16A2446.pdf

Moon, W., Kim, T., Park, B., & Har, D. (2023). Enhanced Transformer Architecture for Natural Language Processing [Doctoral dissertation, Korea Advanced Institute of Science and Technology (KAIST)]. https://arxiv.org/pdf/2310.10930

Naik, M. (2024). The Transformer Architecture: Part I. CIS. https://llm-class.github.io/slides/Lecture%205%20-%20The%20Transformer%20Architecture%20-%20Part%20I.pdf

Ojha, R. (2024). From algorithms to conversations: The influence of natural language processing on chatbot innovation. International Journal of Contemporary Research in Multidisciplinary, 3(6), 21–27. https://doi.org/10.5281/zenodo.14060218

Olabiyi, O., & Mueller, E. T. (2020). DLGNet: A Transformer-based model for dialogue response generation. Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, 54–62. Association for Computational Linguistics. https://aclanthology.org/2020.nlp4convai-1.7.pdf

Pressel, D., Liu, W., Johnston, M., & Chen, M. (2022). Lightweight transformers for conversational AI. Proceedings of NAACL-HLT, Industry Track Papers, 221-229. Association for Computational Linguistics. https://aclanthology.org/2022.naacl-industry.25.pdf

Razumovskaia, E., Glavaš, G., Majewska, O., Korhonen, A., & Vulić, I. (2021). Crossing the conversational chasm: A primer on multilingual task-oriented dialogue systems. Language Technology Lab, University of Cambridge. https://evgeniiaraz.github.io/files/Multilingual_ToD_Primer.pdf

Ren, M. (2024). Advancements and applications of large language models in natural language processing: A comprehensive review. Proceedings of the 2nd International Conference on Machine Learning and Automation, 55-63. https://doi.org/10.54254/2755-2721/97/20241406

Sharma, A., Mehta, H., Gautam, A., Rajpal, P., Thakur, A., & Kaur, H. (2025). Transformer models in NLP. International Journal of Science, Engineering and Technology, 13(2), 1–6. https://www.ijset.in/wp-content/uploads/IJSET_V13_issue2_414.pdf

Sharma, S., Singh, S. P., & Kumari, R. (2024). The use of GPT-3 and similar models in conversational AI: A comparative analysis. The Journal of Computational Science and Engineering, 2(4), 132-141. https://jcse.cloud/JCSE/Published_Papers/177P132_139.docx.pdf

Sun, Y. (2023). The evolution of transformer models from unidirectional to bidirectional in Natural Language Processing. Proceedings of the 2023 International Conference on Machine Learning and Automation, 42, 281-289. https://doi.org/10.54254/2755-2721/42/20230794

Wu, T., Wang, Y., & Quach, N. (2025). Advancements in natural language processing: Exploring transformer-based architectures for text understanding. OceanAI. https://arxiv.org/pdf/2503.20227

Yee, S. S. S., & Soe, K. M. (2024). Exploring BERT-based encoders for sequence classification and multi-task learning in dialogue acts and joint intent-slot filling. Indian Journal of Computer Science and Engineering (IJCSE), 15(3), 317-325. http://www.ijcse.com/docs/INDJCSE24-15-03-049.pdf

Plagiarism Statement

I hereby certify that I have completed this work independently and have not used any sources other than those cited.

All passages that are quoted verbatim or paraphrased from other works are clearly marked as such in each individual case, with precise citation of the source (including the World Wide Web and other electronic data collections). This also applies to any attached drawings, illustrations, sketches, and the like.

This work, in its entirety or in substantial parts or excerpts, has not been submitted previously in any study program at this or any other university for the award of credit points.

I acknowledge that failure to properly attribute sources will be considered an attempt at deception or plagiarism.

XXXX, on XX.XX.XXX