UNIVERSITÉ DU QUÉBEC À MONTRÉAL

IMPROVING CHATGPT'S EMOTIONAL INTELLIGENCE THROUGH PROMPT ENGINEERING

DISSERTATION

PRESENTED

AS PARTIAL REQUIREMENT

TO THE MASTERS IN COMPUTER SCIENCE

BY

AHMED BELKHIR

NOVEMBER 2023

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

AMÉLIORATION DE L'INTELLIGENCE ÉMOTIONNELLE DE CHATGPT AVEC LE PROMPT ENGINEERING

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

AHMED BELKHIR

NOVEMBRE 2023

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Dr. Fatiha Sadat, for her invaluable guidance, unwavering support, and expertise throughout the entire duration of my Master studies at UQAM. She has been an excellent advisor and her insightful feedback and constructive criticism have been instrumental in shaping the direction and quality of this dissertation work.

I would also like to thank all faculty members and academic staff of UQAM for the resources they have provided throughout my academic journey and I also thank all the teachers of the computer science department of UQAM for the quality of their teaching throughout my masters.

My deep gratitude goes also to my family and friends for their support, encouragement, and patience throughout this academic journey. Their confidence in my abilities and constant support have been a source of strength and motivation.

Special thanks to the members of the jury for their interest in my work and for agreeing to examine this dissertation and enrich it with their proposals.

I extend my sincere thanks to the members of the jury for accepting to evaluate my work and for providing their valuable insights and evaluation. Their expertise and critical evaluation greatly contribute to the overall quality and rigor of this thesis.

In conclusion, I am deeply grateful to all those who have contributed directly or indirectly to the completion of this master's thesis. Their support, guidance, and encouragement have played a vital role in its success.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

LLMs: Large Language Models.

HCI: Human-Computer Interaction.

BERT: Bidirectional Encoder Representations from Transformers.

NLP: Natural Language Processing.

RNNs: Recurrent Neural Networks.

RNN-LM: Recurrent Neural Networks Language Models.

ALICE: Artificial Linguistic Internet Computer Entity.

AIML: Artificial Intelligence Markup Language.

Seq2Seq: Sequence to Sequence.

LSTM: Long Short Term Memory.

UI: User interface.

NLG: Natural Language Generation.

RL: Reinforcement Learning.

RLHF: Reinforcement Learning from Human Feedback.

PPO: Proximal Policy Optimisation.

VAD: Valence, Arousal, and Dominance.

GAN: Generative Adversarial Networks.

ELECTRA: Efficiently Learning an Encoder that Classifies Token Replacements Accurately.

PPCA: Principal Preserved Component Analysis.

AUROC: Area Under the Receiver Operating Characteristic Curve.

GPU: Graphics Processing Unit.

SOTA: State-Of-The-Art.

AI: Artificial Intelligence.

CoT: Chain-of-Thought.

RÉSUMÉ

Ce mémoire de maîtrise présente une étude sur l'amélioration des capacités émotionnelles des modèles de langues conversationnels. Cette recherche étudie de nouvelles approches pour incorporer les émotions dans les réponses des chatbots en utilisant le prompt engineering.

Ce travail est motivé par la demande croissante de chatbots émotionnellement intelligents qui peuvent engager les utilisateurs dans des conversations plus personnalisées et compatissantes. Pour ce faire, deux contributions principales sont fournies. Tout d'abord, un classificateur d'émotions précis basé sur le modèle ELECTRA est développé. Le classificateur est entraîné sur l'ensemble de données GoEmotions. Il obtient des performances impressionnantes, avec un AUROC allant jusqu'à 98,5 %. Ce classificateur d'émotions fiable sert comme élément de base pour le reste de la recherche et facilite l'analyse des émotions des utilisateurs et des chatbots.

Deuxièmement, de nouvelles méthodes d'infusion d'émotions utilisant le prompt engineering sont proposées et évaluées. Les modèles de chatbot ChatGPT-B et ChatGPT-C sont conçus pour modifier leurs réponses en fonction des émotions de l'utilisateur, ce qui se traduit par des interactions plus cohérentes sur le plan émotionnel. D'une part, ChatGPT-B prend en compte l'émotion de l'utilisateur avant de générer des réponses à l'aide d'un classificateur d'émotions, et d'autre part, ChatGPT-C tente de s'adapter à ce que ressent l'utilisateur sans aucune composante externe, en s'appuyant uniquement sur l'ingénierie de l'invite pour améliorer les réponses au niveau émotionnel.

En analysant les réponses des deux versions modifiées proposées et en les comparant à la version standard de ChatGPT (ChatGPT-A), nous constatons que l'utilisation du classificateur d'émotions externe entraîne une utilisation plus fréquente et plus prononcée des émotions positives par rapport à la version standard. En revanche, l'utilisation d'une simple ingénierie d'invite pour prendre en compte l'émotion de l'utilisateur produit l'effet inverse. Enfin, les comparaisons avec des modèles de chatbots émotionnels proposés dans la littérature mettent en évidence le potentiel du prompt engineering pour améliorer les capacités émotionnelles des agents conversationnels basés sur les modèles de langues larges.

Mots clés: Agents conversationnels, Chatbots, Modèles de langue larges, Apprentissage par transfert, Intelligence émotionnelle

ABSTRACT

This master thesis presents an exhaustive investigation into improving the emotive capabilities of conversational language models. The research investigates novel approaches for incorporating emotions into chatbot responses using prompt engineering.

The motivation behind our work is the increasing demand for emotionally intelligent chatbots that can engage users in more personalized and compassionate conversations. To accomplish this, two primary contributions are provided. First, an accurate emotion classifier built on top of the ELECTRA model is developed. The classifier is trained on the GoEmotions dataset and obtains impressive performance, with an AUROC of up to 98.5%. This reliable emotion classifier serves as a foundation for the rest of the research and facilitates the analysis of user and chatbot emotions alike.

Second, new methods for emotion infusion utilizing prompt engineering are proposed and evaluated. The ChatGPT-B and ChatGPT-C chatbot models are designed to modify their responses based on user emotions, resulting in more emotionally consistent interactions. On one hand, ChatGPT-B takes the user's emotion as input before generating responses using an emotion classifier, and on the other hand, ChatGPT-C tries to accommodate for how the user feels without any external component, relying only on prompt engineering to improve responses on the emotional level.

By analyzing responses of the two proposed altered versions and comparing them to the standard version of ChatGPT (ChatGPT-A), we find that using the external emotion classifier leads to more frequent and pronounced use of positive emotions compared to the standard version. On the other hand, using simple prompt engineering to take the user emotion into consideration, does the opposite. Finally, comparisons to the state-of-the-art models highlight the potential of prompt engineering to enhance the emotional abilities of conversational agents based on large language models.

Keywords: Conversational agents, Chatbots, Large language models, Transfer learning, Emotional intelligence

INTRODUCTION

In the ever-changing landscape of technological innovation, the field of conversational systems and chatbots has made remarkable strides in recent years, attracting the attention and enthusiasm of not only devoted researchers and practitioners but also a broader range of individuals, including non-specialists. Mainstream chatbot applications now range from customer service (Ando et Zhang, 2005) to education (Colace *et al.*, 2018) and mental health therapy (Abd-Alrazaq *et al.*, 2019). The interest in these systems has been fueled by their 24/7, around-the-clock availability, their ever-growing knowledge base, and, most importantly, their potential to chat with users in engaging and fulfilling interactions by simulating human-like conversations and therefore attracting users from all domains and needs.

Conversational agents have come a long way since the 1960s. Early chatbots relied on rigorous rule-based systems, operating within predefined guidelines and predetermined responses. Later, retrieval-based systems were developed, improving flexibility by accessing predefined databases of responses, thereby enabling more contextually relevant responses. The breakthrough, however, came with the introduction of modern generative models based on seq2seq (Sutskever *et al.*, 2014) and Transformers architecture (Vaswani *et al.*, 2017), enabling convincingly natural conversation with robots with impressive understanding capabilities.

However, even though modern chatbots can hold a meaningful conversations in natural language and provide helpful information, there is still a gap that prevents a real bond between humans and machines from being built, and it is about emotions. Modern conversational agents have limited abilities when it comes to understanding, processing, and generating human-like and emotion-rich conversations (Rapp *et al.*, 2021; Belainine *et al.*, 2020c; Belainine *et al.*, 2020a). On the other hand, emotion-aware chatbots have the potential to create more meaningful and empathetic conversations, bridging the gap between human and AI and building rapport and trust in human-machine interactions (Chen *et al.*, 2021).

To address this challenge, this research at the intersection of chatbot technology, Artificial Intelligence (AI), and emotion recognition explores the possibility of enhancing the emotional intelligence of chatbots built on top of Large Language Models (LLMs) and enabling them to deliver more emotionally engaging experiences.

More specifically, we investigate the effectiveness of incorporating external emotion classifiers and prompt engineering (Reynolds et McDonell, 2021) to take into account the user's emotional state when generating responses by the ChatGPT (OpenAI, 2022) chatbot model. The emotion classifier needs to be sensitive enough to discern even the slightest nuances in the user utterances in order to provide helpful information to the ChatGPT conversational language model and steer it accordingly. Prompt engineering is then used as a technique to inject emotional information into the chatbot to make it more aware of the emotion of the user before generating replies, resulting in more emotion-aware and human-like conversations with the same language model and without the need for re-training or even fine-tuning.

0.1 Problem statement

This research is motivated by the increasing demand for sophisticated and emotionally intelligent chatbot systems. While conversational agents have become increasingly prevalent in numerous domains, such as customer service, virtual companionship, education, and healthcare, their capacity to emotionally engage with users leaves to be desired. Emotions play a crucial role in human communication, influencing how we express ourselves, comprehend others, and forge relationships. Consequently, the incorporation of emotions into chatbot interactions has the potential to significantly improve user experiences and the overall efficacy of these systems and convince more users to use the new technology.

In recent studies, it has been shown that most people still prefer human-to-human interaction over communicating with an artificial chatbot system and believe that a human can understand them better (Rapp *et al.*, 2021). This is because, while modern chatbot systems do answer questions reliably in most cases, they fail to convey to users the feeling of talking to an actual human. This is due to their inability to understand and respond to user emotions. Traditional chatbots often provide impersonal and dispassionate responses, lacking humans' natural empathy and emotional awareness. By incorporating emotional intelligence into chatbots, we can bridge this gap and create conversational agents that can alter their responses based on the user's emotional states, leading to conversations that are more engaging and empathetic and have a deeper resonance with users.

Moreover, this research intends to contribute to the fields of Natural Language Processing (NLP) and AI by advancing the comprehension and modeling of emotions in human-machine interactions. Emotions are complex and diverse, and capturing their nuances in computational models is a significant challenge for researchers. By building a reliable emotion classifier, we can obtain insight into both the emotional and the semantic information conveyed by the text data. This additional information can be leveraged to incorporate emotions into chatbot systems effectively.

Due to the nuanced and complex character of human emotions, emotion recognition is considered a significant challenge in the field of NLP. Emotions are inherently complex because they encompass a broad spectrum of feelings, expressions, and contextual variations that are frequently intertwined with cultural, linguistic, and individual factors. Moreover, the scarcity of labeled emotional data compounds the difficulty of training robust models from scratch. Thankfully, transfer learning emerged as a promising solution, revolutionizing the landscape of emotion recognition in NLP. Transfer learning employs pre-trained models on large and diverse text corpora to enable the extraction of complex linguistic features and task-general contextual patterns. By fine-tuning these pre-trained models on emotion-specific datasets, the obtained models can effectively learn to recognize the subtle signals that indicate different emotions. This strategy capitalizes on the general semantic knowledge acquired during pre-training. It adapts it to the specific nuances of emotion recognition, resulting in improved accuracy and performance in identifying emotions from text, which is essential for imbuing chatbot interactions with emotional intelligence. Therefore, this research not only enhances our understanding of emotions but also contributes to the advancement of conversational AI systems.

As LLMs continue to grow in scale and complexity, a persistent challenge has emerged: the significant resource burden associated with retraining such massive models. The process of retraining LLMs in order to adapt to specific tasks or contexts has become increasingly resource-intensive, requiring significant computational power, time, and money resources. This poses a substantial obstacle, particularly for researchers with limited access to advanced infrastructure. Even the more manageable task of fine-tuning these models to achieve desired behaviors needs advanced resources that are not within the reach of all researchers. Enter prompt engineering (Reynolds et McDonell, 2021), a new, innovative paradigm that offers a compelling solution to this problem. Unlike traditional methods of retraining and fine-tuning, prompt engineering provides a streamlined and effective way of guiding and shaping the outputs of pre-trained language models. By designing carefully crafted prompts or queries, researchers and developers can influence the behavior of these models without the need for enormous resources.

Therefore, this research aims to advance the chatbot technology field by improving its emotional intelligence. By addressing the limitations of existing chatbot systems and utilizing the power of emotion recognition, we can create more engaging and human-like conversational agents that have the potential to revolutionize human-machine interactions across multiple domains.

0.2 Objectives

The research in this master thesis seeks to improve the chatbot systems' emotional intelligence so they can engage in more empathetic and emotionally-aware conversations. In order to achieve this goal, the following specific research objectives have been established:

- **Develop an emotion classifier:** Based on the ELECTRA model and a suitable emotion dataset, develop an accurate and reliable emotion classifier. The classifier must be able to detect and classify user emotions conveyed in textual input, laying the groundwork for emotion-aware chatbot interactions.
- **Integrate emotion modeling into ChatGPT:** Improve the state-of-the-art ChatGPT model with enhanced emotion capabilities. Explore various techniques, such as emotion infusion and adaptation, to imbue ChatGPT with the capacity to comprehend and respond to user emotions in a more personalized and context-appropriate manner.
- **Evaluate and compare approaches:** Conduct exhaustive tests to assess the performance of the developed emotion classifier and the improved ChatGPT models. Evaluate the efficacy of emotion prediction as well as the impact of emotion infusion and adaptation on the level of empathy exhibited by the different chatbot versions.
- **Compare with SOTA models:** Compare the performance of the developed emotion-aware ChatGPT models to that of other state-of-the-art emotion-aware chatbot models proposed in the literature. Evaluate the advantages and disadvantages of various emotional responses in terms of emotion precision and recall, as well as the fluency and coherence of the replies.

This research seeks to advance chatbot technology by enhancing its emotional intelligence and paving the way for more empathetic, engaging, and human-like interactions between chatbots and users.

0.3 Contributions

This master's thesis presents two significant contributions intended to improve the emotional intelligence of chatbot systems. The first contribution is the creation of an accurate emotion classifier, and the second is the introduction of novel emotion infusion techniques based on prompt engineering techniques.

An emotion classifier based on the ELECTRA model is developed to resolve the limitations of existent emotion detection models. The classifier detects and categorizes user emotions with high precision, outperforming existing SOTA models and providing a solid foundation for emotion-aware chatbot interactions. Using precision, recall, and F1-score metrics, the classifier's performance is proven to be reliable, both on an individual emotion level and on a general level.

Moreover, this dissertation presents novel emotion infusion techniques that utilize prompt engineering and an external emotion classifier to improve the emotional expressiveness of chatbots. Two new variants of the ChatGPT language model are presented: ChatGPT-B and ChatGPT-C. ChatGPT-B incorporates the precise emotion of the user as input, resulting in more positive expressions of emotion. ChatGPT-C, on the other hand, modifies its responses based on the user's emotional signals, with the goal of expressing a broader range of emotions, including negative ones, and fostering empathy. Extensive experiments also prove that the accuracy, fluency, and coherence of these methods' emotional responses are superior.

This master thesis is built upon the foundation of a paper published in ACL ontology and presented at the 14th Conference RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING (RANLP 2023). It provides essential insights and findings relevant to the research presented in this master thesis. See appendix section A.

0.4 Dissertation Structure

Chapter 1 sets the foundation for this research by providing an overview of the technologies utilized. We begin by examining chatbots: what they actually are, and the reasons behind their growing prominence. Exploring LLMs, we unravel their ability to understand context and represent words and meanings. Additionally, we delve into some details of the transformer architecture before introducing some emotion models, thus establishing a framework for understanding the different dimensions and categories of emotions. This chapter establishes the groundwork for the subsequent chapters, laying the foundation for integrating emotional intelligence into conversational systems.

Chapter 2 explores the existing literature on the topic, providing a comprehensive review of research and developments in conversational systems with a focus on emotional intelligence. We explore the history of chatbot research and development, highlighting the challenges and opportunities in enhancing emotional understanding in human-machine interactions. By analyzing prior approaches, we identify gaps and constraints, aiming to find opportunities for further advancements in the field.

Chapter 3 presents our approach to enhancing the emotional capabilities of ChatGPT. We aim to equip ChatGPT with emotional awareness to create more empathetic and emotionally intelligent chatbots. We begin by developing an emotion classifier that shall allow for accurate recognition and categorization of user emotions, serving as a crucial evaluation tool. Building upon this foundation, we introduce three variants of ChatGPT that enable us to explore different approaches to infusing emotional intelligence into chatbot conversations.

In Chapter 4, we present the results of our experiments in two parts. The first part focuses on the evaluation of our emotion classifier, showcasing the classification metrics and the hyperparameter tuning process that yielded impressive performance. The second part investigates the impact of emotion infusion and adaptation on ChatGPT's empathy level, comparing our approaches with state-of-the-art models. We examine the emotional response accuracy and fluency of the different ChatGPT variants, highlighting their strengths and contributions to enhancing emotional intelligence.

We finish our work with a summarizing conclusion that presents future perspectives and implications of this research.

CHAPTER 1
MAIN CONCEPTS

1.1 Introduction

In this chapter, we set the foundation for the subsequent chapters of this dissertation by providing background information on the main technologies utilized. We begin by delving into the world of chatbots, providing definitions, and emphasizing why they have recently become a topic of growing interest.

To comprehend the inner workings of chatbots, it is essential to familiarize ourselves with language models. We explore how they understand the context and elucidate how they represent words and meanings through advanced techniques. Moreover, we explore the complexities of transformer architecture, a crucial element in the development of modern chatbot systems.

Emotions play an essential role in human-to-computer interactions, and their incorporation into chatbot conversations has enormous potential. Therefore, we introduce the widely recognized emotion models proposed by Ekman and Plutchik.

Explaining all these concepts will provide a solid foundation for understanding the rest of this research and its contributions to enabling chatbots to carry on human-like conversations.

1.2 Chatbots

1.2.1 Definition

A chatbot is an interface model for Human-Computer Interaction (HCI) (Bansal et Khan, 2018). The Oxford English Dictionary defines a chatbot as

> "A computer program designed to simulate conversation with a human user, usually over the internet; esp. one used to provide information or assistance to the user as part of an automated service."

However, the term "conversational agent" is the more formal name for chatbots in scientific literature, and is defined as

> "A dialogue system that can also understand and generate natural language content, using text, voice, or hand gestures, such as sign language" (Allouch *et al.*, 2021).

Throughout this thesis, we will be using both terms interchangeably.

In simple terms, a conversational agent is a software that can converse with a user in a natural language like English or French (as opposed to artificial languages like programming languages) via virtual chat rooms, websites, mobile apps, messaging applications or through the telephone. This interface is slowly substituting many established graphical, purpose-specific user interfaces such as web or mobile apps (Følstad et Brandtzæg, 2017).

Although the first chatbot dates back to 1966 (Weizenbaum, 1966), the interest in chatbots has only exploded in recent years, especially after 2016 (Adamopoulou et Moussiades, 2020a). Some researchers even called it the "chatbot tsunami" (Grudin et Jacques, 2019). This interest is likely due to the improvements in computer processing power, the wide adoption of instant messaging applications, and the technological breakthroughs of recent years (see Figure 1.1). Another important driving factor behind chatbots development has been the Loebner Prize, the annual competition for conversational agents to identify the most human-like programs (Epstein, 1992), which led to the development of systems that are more and more human-like.

1.2.2 Why use chatbots?

Chatbots have proven to be very useful. The interest mainly arose among tech giants such as Google and Microsoft. Satya Nadella, Microsoft's CEO declared that "Chatbots are the new apps" (Marco della Cava, 2016). Chatbots today are being deployed across multiple industries to engage with customers in business, patients in healthcare, or assist in education, to name but a few fields.

1.2.2.1 Chatbot applications

In **customer service**, chatbots are a very interesting alternative to human agents because machines can work 24 hours a day, seven days a week, and serve multiple customers at the same time. Also, current chatbots are competent enough to keep customers happy to a certain extent. In fact, a study (Luo *et al.*, 2019) found that undisclosed chatbots are just as effective as experienced

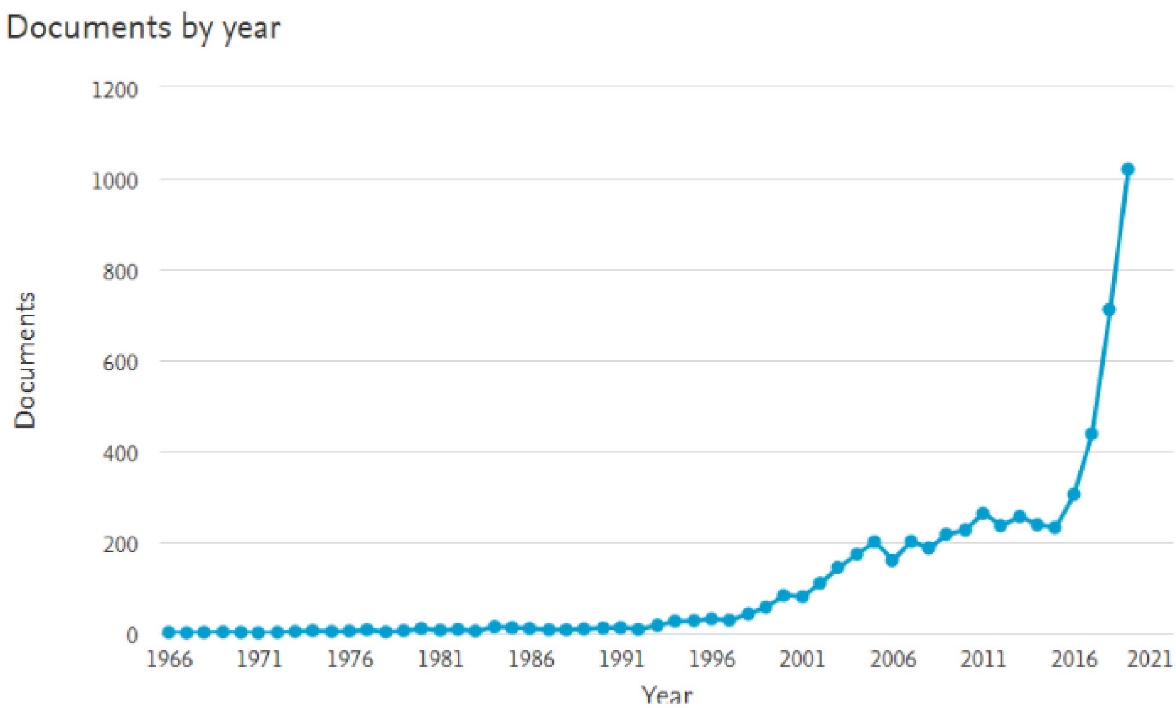

Figure 1.1: Search Results in Scopus, from 1966 to 2019 for the keywords “chatbot” or “conversation agent” or “conversational interface" (Adamopoulou et Moussiades, 2020a).

workers and four times more effective than inexperienced novice workers in terms of the number of purchases. The same study indicates that revealing that the customer is being assisted by a chatbot decreases the purchase rates significantly because machines are perceived as less knowledgeable and less empathetic, even when they are not. Furthermore, according to Harvard Business Review, a mere five-minute delay could decrease a business's chances of selling to a customer, and a ten-minute delay could reduce their chances by 400% (Magazine, 2018).

In **healthcare**, chatbots can be used to diagnose simple diseases and provide information about therapy and prevention. They represent a decent solution because research says 60% of visits to doctors are for simple small-scale conditions, 80% of which can be cured at home using simple home remedies (Bhirud *et al.*, 2019). Chatbots were also used to help fight the covid-19 virus by monitoring exposure to the virus, tracking infection symptoms, and combatting misinformation and fake news (Almalki et Azeez, 2020).

Education is another field in which chatbots are making a difference. In fact, the scalability of such 24/7 systems makes them attractive as teacher replacements, even only for casual, simple questions. In one case study, authors implemented a chatbot into the e-learning platform of the University of Salerno to assist with two courses: Computer Science and Computer Networks (Colace *et al.*, 2018). There results were positive as over 70% of the suggestions provided by the chatbot were considered correct by successful students. In addition to scalability, chatbots can be better than humans when it comes to teaching foreign languages. A study (Haristiani, 2019) found that learners were more confident talking and learning from chatbots compared to actual language teachers.

Chatbots proved to be interesting in **other fields**, too, including personal banking, food ordering and delivery, package tracking, and robotics. Beyond practical applications for chatbots, researchers investigated the psychological reasons behind using such systems.

1.2.2.2 Psychological Reasons

In one of the first surveys to investigate the reasons for using chatbots, it turned out that there are four main reasons behind the interest: productivity, entertainment, social needs, and curiosity (Brandtzaeg et Følstad, 2017).

Using Uses and Gratifications theory (U&G) (Rubin, 2009), open questions in the survey revealed that the most common reason for using chatbots is productivity. Users find that a chatbot can be a fast and convenient way to obtain information online. It is in some way easier to just ask for the information than to navigate the various graphical user interfaces to find simple answers.

In addition to productivity, some people use chatbots for entertainment purposes. Chatbots' answers can be funny sometimes, especially if the user is bored or wants to kill time. Other users might be lonely and feel the need for some kind of social interaction. The chatbots nowadays have become so close to humans that it is possible to enjoy their company as a virtual friend in order to fulfill social needs.

Furthermore, chatbots are a relatively new technology for the mass market, and this attracts curious and risk-driven users who want to explore it and experiment with it.

1.3 Simplified Chatbot Pipeline

Chatbots are made up of several crucial parts, each of which plays a specific role and cooperates with the others to form a reliable system that successfully accomplishes its goal. According to (Bilquise *et al.*, 2022), we can organize these components in a pipeline based on the order of usage:

- **Natural Language Processing unit:** This is the component that processes the chatbot input using NLP techniques such as tokenization, lemmatization, and stemming (Suhaili *et al.*, 2021).
- **Natural Language Understanding:** Typically, this component parses structured data from the NLP component to comprehend the user's intent and any details related to that intent (Cahn, 2017)
- **Dialog Manager:** In order to decide what action should be taken next, the dialogue manager component analyses the comprehensible structured data, maintains the dialogue framework, such as the semantic frame, and encodes the data. If there are ambiguities about the user's needs, this component eventually asks further questions to resolve it (Adamopoulou et Moussiades, 2020b).

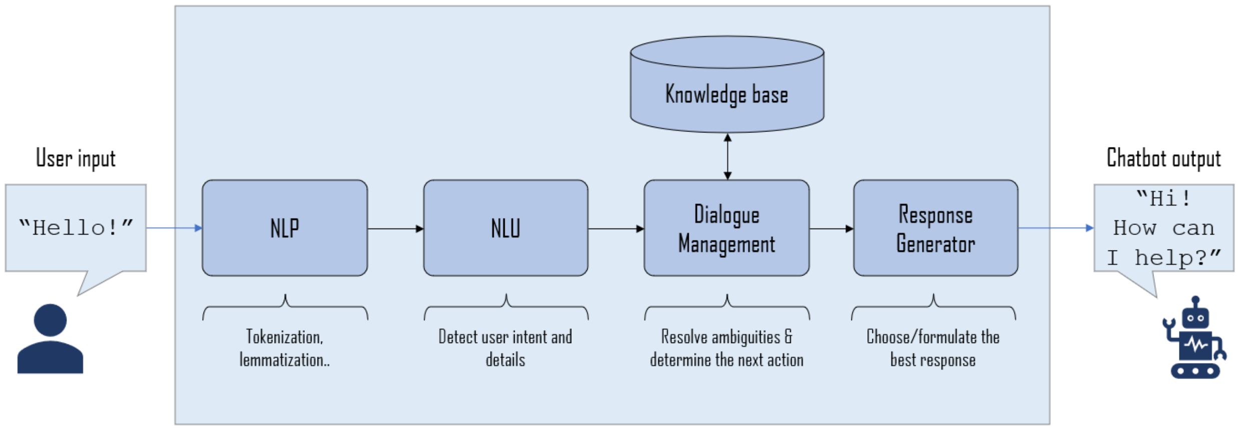

Figure 1.2: Simplified chatbot pipeline.

- **Natural Language Generator:** This is the component that may be retrieval-based, rule-based, or generative. It is responsible for how the chatbot generates responses based on information from previous components.

Additionally, recent developments in NLP have completely revolutionized the chatbot industry. One of the most important developments is the advent of large language models.

1.4 Large Language models

Large Language Models (LLMs) have had a substantial impact on the capabilities and functionality of modern chatbots. These models have revolutionized the way chatbots comprehend and provide human-like answers. They were trained on enormous volumes of text data and strengthened using deep learning techniques.

In order to have a basic understanding of LLMs' significance and how they work, we provide a brief explanation of important aspects:

1.4.1 Word representation

Language models use "word embeddings" or word vectors to represent words and meanings in a high-dimensional space. Word embeddings are dense numerical representations that identify the semantic connections between words based on how frequently they appear together in a text corpus. The development of Word2Vec in 2013 (Mikolov *et al.*, 2013) revolutionized word embeddings by proposing a shallow neural network model capable of learning continuous representations of words from large text corpora based on their co-occurrence patterns. These representations allow LLMs to process language more accurately and nuancedly by helping them comprehend the context and meaning of words. To illustrate this concept, authors in (Bandyopadhyay *et al.*, 2022) projected a number of word embeddings in 3D space to make it easier to grasp the idea of word embeddings (figure 1.3).

Word embeddings proved to be very useful because of their following characteristics:

- **Distributed representation:** Word embeddings in a high-dimensional space express words as vectors in a distributed representation. The features that represent each dimension of the vector each represent a specific facet of the meaning of the word. Therefore, words with comparable meanings or contexts tend to have vectors that are closer to one another in the embedding space thanks to this distributed representation. For example, consider two words:

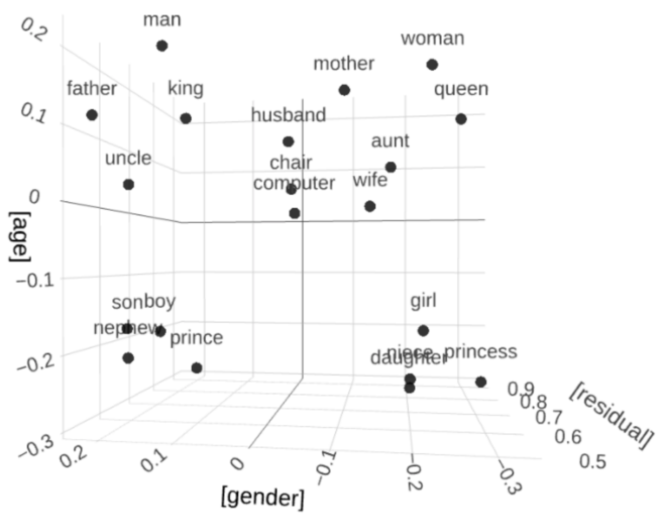

Figure 1.3: Plot of the projection of some word embeddings in a 3D space. The left and right halves of the x-axis represent male and female words, respectively. Adult and youth words are, respectively, in the top and bottom halves of the y-axis (Bandyopadhyay *et al.*, 2022).

"dog" and "puppy." These words will likely have similar vector representations in a well-trained LLM with word embeddings. For example, the representation of "dog" might be [0.2, 0.6, -0.3], while the representation of "puppy" might be [0.18, 0.58, -0.25]. The similarity in the vector representations suggests that these are semantically close ideas.

- **Analogies:** Word embeddings also provide LLMs the ability to use analogical reasoning. For example, given the analogy "man is to the woman as king is to ?," the LLM can infer the missing word, "queen," by leveraging the vector representations of the words as shown in figure 1.4

- **Contextual embedding:** When creating word representations, LLMs frequently use contextual embeddings like BERT (Bidirectional Encoder Representations from Transformers) (Devlin *et al.*, 2018), which take into account the surrounding words and sentence structure. These contextual embeddings improve the model's capacity to discern fine distinctions and word meanings that are exclusive to a given context. For illustration, consider the sentence "The bank is closed.". In this example, the word "bank" refers to a financial institution. However, the word "bank" has an entirely different meaning in the sentence "He sat by the river bank", where it refers to the river's side. Because of contextual embeddings, the LLM

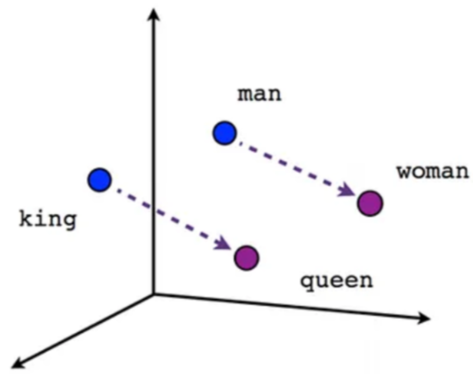

Figure 1.4: Word embeddings and analogy.

can distinguish between distinct word meanings based on the words and sentence structure around them.

In addition to word embeddings, character embedding was also explored as a solution to deal with morphologically complex and out-of-vocabulary words. In contrast to word embeddings, which represent semantic relationships between words based on their context, character embeddings represent individual characters. Instead of relying solely on pre-trained word vectors, character embeddings enable models to generate embeddings for unobserved words by taking into account the constituent characters. Techniques used to learn meaningful representations from character sequences include character-level convolutional neural networks (Zhang et LeCun, 2015) and recurrent neural networks language models (RNN-LM) (Kim *et al.*, 2016; Jozefowicz *et al.*, 2016). These embeddings can capture useful information about prefixes, suffixes, and stems, allowing models to manage word variations and orthography inconsistencies more effectively. By combining word and character embeddings, language models can obtain a more robust understanding of natural language.

To effectively leverage the powerful word representations in LLMs, such as word embeddings, a groundbreaking architecture called the Transformer (Vaswani *et al.*, 2017) was published in 2017, revolutionizing the field of Natural Language Processing.

1.4.2 The transformer architecture

Since its release in 2017, the Transformer architecture (Vaswani *et al.*, 2017) has had a significant impact on how LLMs are created and has been the underlying base of most new language models. In the seminal paper "Attention Is All You Need", Vaswani et al. introduced a machine learning architecture that overcame the drawbacks of sequential models like Recurrent Neural Networks (RNNs) (Rumelhart *et al.*, 1985) and revolutionized natural language processing tasks, including chatbot development.

Figure 1.5 shows the transformer architecture as illustrated in their original paper (Vaswani *et al.*, 2017). We also provide a brief overview of the essential aspects of this architecture as well as its significance in the context of LLMs:

- **Self-Attention Mechanism:** The Transformer architecture employs a self-attention mechanism that enables the model to evaluate the relative weights of several words or tokens within a sequence. By utilizing this technique, the model is able to overcome the constraints of sequential processing and capture long-range dependencies and interactions between words. Each word applies attention weights that it has acquired during training to all other words, including itself. This mechanism considers the words' positions and captures contextual relationships between them.
- **Attention Layers:** The Transformer architecture consists of multiple attention layers. Multiple self-attention heads are present in each layer, and they learn various representations of the input sequence separately. These attention heads' outputs are merged and altered to create a rich and varied collection of characteristics for the following layers.
- **Positional Encoding:** Using positional encoding, The Transformer model integrates word positional information. By including learned positional vectors into the word embeddings, positional encoding informs the model of the placements of the words within the sequence. This enables the model to distinguish words not just according to their semantic meaning but also according to their location.
- **Encoder-Decoder framework:** This framework is used in the transformer architecture in order to allow the model to generate coherent and contextually appropriate responses. The encoder analyses the input sequence while employing self-attention layers to capture contextual information, while the decoder pays attention to the encoder's output and the partial output sequence generated.
- **Parallel Computation:** Thanks to the transformer architecture, highly parallel computing is made possible. The self-attention mechanism removes the sequential bottleneck seen in conventional recurrent models, enabling the model to analyze all words in the

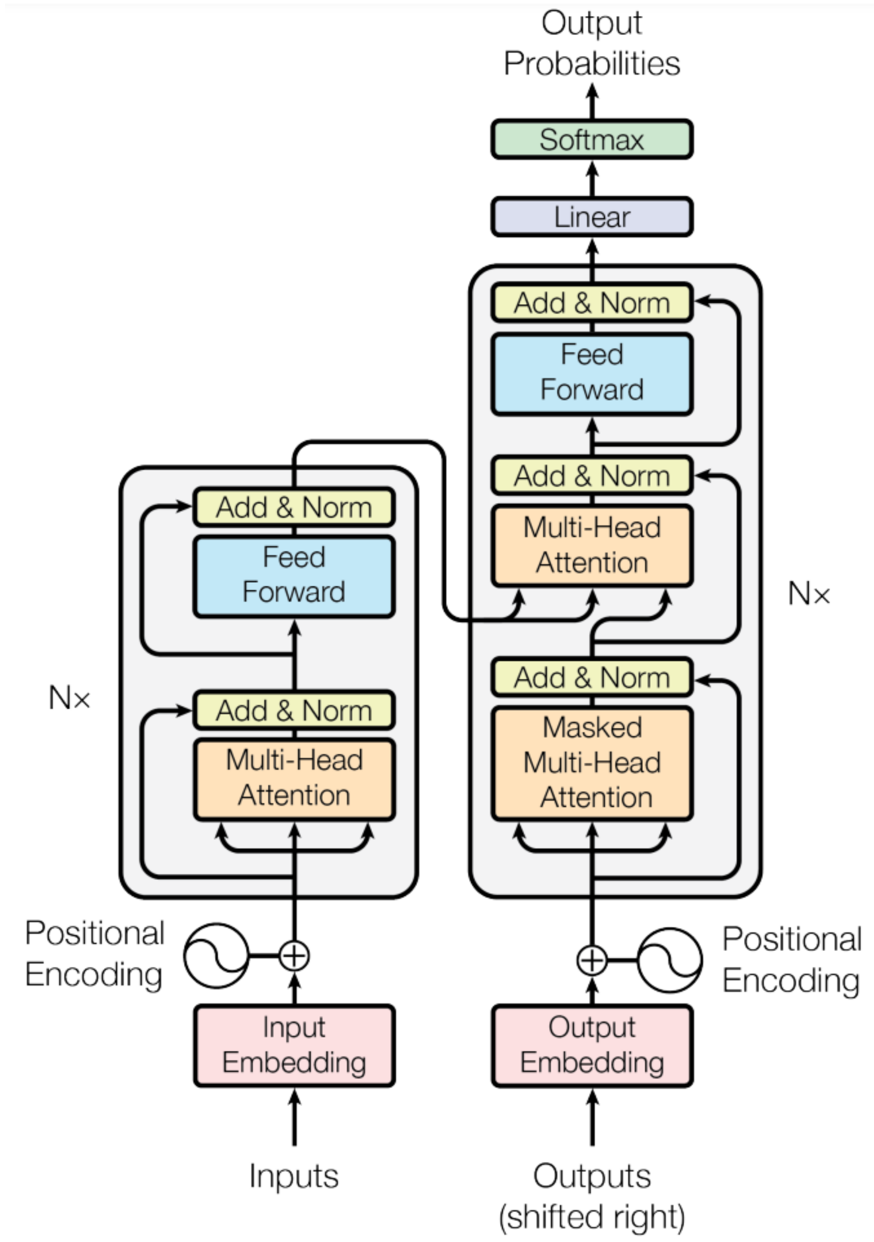

Figure 1.5: The transformer architecture (Vaswani *et al.*, 2017).

sequence simultaneously. Large-scale language models can significantly benefit from the improved training and inference efficiency of this parallel processing.

Building upon the advancements in LLMs, a very notable example that has gained considerable attention in the field of chatbot development is ChatGPT, a state-of-the-art conversational AI model developed by OpenAI.

1.5 Emotional Intelligence

1.5.1 Definition

Emotional intelligence is a concept rooted in psychology that plays a fundamental role in human communication and interaction. It can be defined as "the ability to monitor one's own and others' feelings, to discriminate among them, and to use this information to guide one's thinking and action" (Salovey et Mayer, 1990).

1.5.2 Emotion models

A well-known psychologist named Dr. Robert Plutchik introduced in 1980 a thorough model of emotions called the "Plutchik's Wheel of Emotions" (Plutchik, 1980), which seeks to explain and categorize the complex range of human emotions. The model visualizes emotions as a wheel, with eight primary emotions positioned at cardinal points and additional secondary emotions placed between them to represent blended or mixed emotions (see figure 1.6). The eight basic emotions represented are:

- Joy: The emotion associated with happiness, contentment, and pleasure.
- Sadness: The feeling of unhappiness, grief, or sorrow.
- Anger: The emotion characterized by hostility, frustration, or rage.
- Fear: The response to perceived threats or danger leading to anxiety or panic.
- Trust: The feeling of confidence, reliance, and belief in someone or something.
- Disgust: The aversion or revulsion towards something offensive or repulsive.

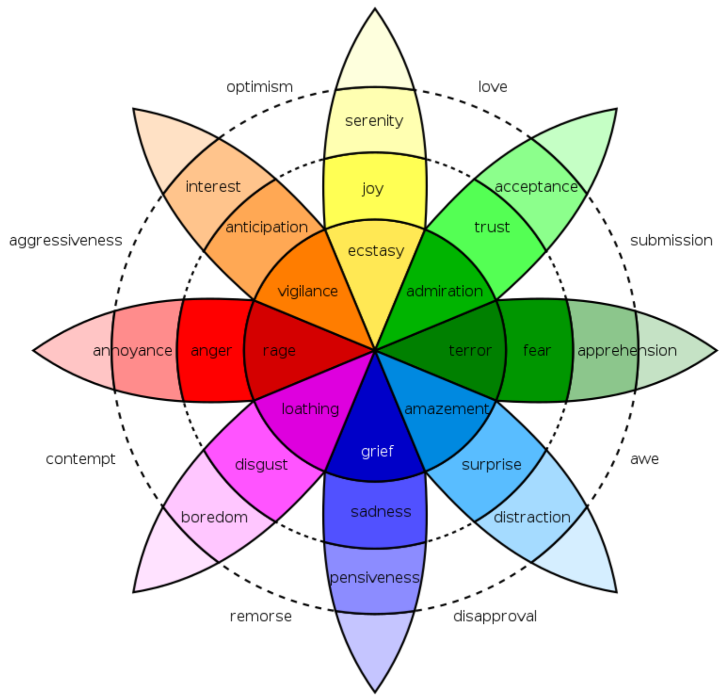

Figure 1.6: Plutchik's wheel of emotions (Plutchik, 1980).

- Surprise: The emotion resulting from unexpected or startling events.
- Anticipation: The anticipation or excitement towards future events or possibilities.

Plutchik's model also includes the concept of intensity, where varying degrees of intensity result in a spectrum of emotions. For instance, joy can range from mild contentment to extreme ecstasy. By combining and blending these primary emotions, Plutchik's model provides a framework for understanding and categorizing the complexities of human emotional experiences.

Later in 1992, Dr. Paul Ekman, the renowned psychologist and pioneer in the field of emotion research, introduced another emotion model based on the "Facial Action Coding System" and identified six common facial expressions of emotions. (Ekman, 1992). According to Ekman's theory, six primary emotions may be identified by their facial expressions: joy, sorrow, anger, fear, surprise, and disgust. According to Ekman's study, these emotional expressions are shared by all civilizations and social groups, indicating that they are intrinsic and physiologically based. The Facial Action Coding System thoroughly examines the precise facial muscle movements connected to each emotion, enabling accurate measurement and identification.

In the realm of AI, incorporating emotional intelligence into chatbots holds great promise for creating more engaging and empathetic conversational agents. We can narrow the communication gap between humans and machines and make encounters that are more meaningful and rewarding by giving chatbots the ability to recognize, understand, and respond to emotions.

1.6 Conclusion

In conclusion, this chapter has provided a comprehensive overview of the foundational concepts that underlie our research into enabling chatbots to engage in human-like conversations. We began by exploring the realm of chatbots, showing their significance and offering insights into their architecture design and natural language understanding capabilities. We also explained some techniques used by modern LLMs as well as some emotion modeling methods.

In the next chapter, we will dive deeper into chatbot evolution in the last decades with a particular focus on a recent disruptive chatbot: ChatGPT. We will also explore the challenges in equipping artificial chatbot systems with human-like emotions and the different techniques researchers propose to tackle such a problem and make conversations more natural and engaging.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

In recent years, the field of conversational systems has made tremendous strides, with language models exhibiting astonishing skills in producing information-rich replies. However, understanding and utilizing emotions is a significant obstacle to developing interactions that are genuinely interesting and realistic. In this literature review, we start with a brief history of chatbot research and development, then examine the methods that have been done to improve the emotional intelligence of conversational language models. We aim to investigate the state-of-the-art approaches and techniques used to comprehend and provide emotionally appropriate responses in human-machine interactions. We analyze prior research in this area in order to find gaps, constraints, and opportunities for further development.

2.2 A Brief History of Chatbots

The first chatbot in the literature was proposed in 1966 by Dr. Joseph Weizenbaum at the Massachusetts Institute of Technology (MIT), and it was called ELIZA (Weizenbaum, 1966). It was designed to play the role of a digital psychotherapist and paved the way for many chatbots to follow. Chatbot development has since evolved from rule-based systems to more advanced generative AI models.

2.2.1 Rule-based and pattern-matching systems

Rule-based chatbot systems were the leading technology used in the early stages of chatbot development, including ELIZA (Weizenbaum, 1966). These systems generated answers based on specified input criteria by following human-established rules and patterns. A rule-based system's advantage is that it can give accurate responses. However, it works well only when the input message is well-formed. They had limited language understanding capabilities, and substantial manual programming was necessary.

In 1995, ALICE (Artificial Linguistic Internet Computer Entity) was introduced, and it won the

Loebner Prize as “the most human computer” at the annual Turing Test contests in 2000, 2001, and 2004 (Wallace, 2009). It was the first personality program based on AIML (Artificial Intelligence Markup Language), which allowed for the creation of more sophisticated and flexible chatbots by defining patterns and predefined responses to engage in conversations on various topics. In fact, the knowledge of the robot is represented in AIML via a set of categories. A pattern, which is the user's input, and a template, which is the bot's response, make up each category. A word, a phrase, or even a more generic pattern employing wildcard characters can be used as the pattern. The input is compared to the predefined patterns when a user interacts with the chatbot, and the relevant template is then chosen to produce the answer.

2.2.2 Retrieval-Based Systems

These systems rely on predefined responses stored in a knowledge base. They emerged as an improvement over previous technology with newer NLP techniques such as similarity metrics to match user inputs to predefined patterns (Almansor et Hussain, 2020). These systems used predetermined responses stored in a database and used keyword matching or similarity measurements in order to pick the most acceptable replies for the user input. These techniques are used in personal assistants like Alexa, Siri, and Google Assistant to fulfill user requests by gathering data from a variety of sources (Bilquise *et al.*, 2022). However, these systems have limited flexibility and use, especially when domain-specific responses are required (Suhaili *et al.*, 2021).

2.2.3 Sequence-to-sequence models and Generative AI

Generative-based chatbots, as opposed to limited, retrieval-based, and rule-based systems, usually use an encoder-decoder design, specifically a sequence-to-sequence (seq2seq) architecture (Sutskever *et al.*, 2014). In fact, since Google researchers published their innovative seq2seq model in 2014, a multitude of chatbot architectures based on seq2seq were proposed in the literature thanks to the great potential it showed (Vinyals et Le, 2015).

Simply put, this model works by trying to predict the next sentence in a conversation on the basis of the previous sentences using two Recurrent Neural Networks as encoder and decoder:

- **Encoder:** It processes the input sequence (user query) and transforms it into a fixed-length vector representation known as the context vector. This allows for capturing the input's semantic and contextual details.

- **Decoder:** It creates the output sequence (the chatbot answer) word by word using the context vector as input. Based on the context and the previously created words, it guesses the following word that is most likely to occur.

Even though seq2seq was regarded as the industry's best practice because it maximizes the likelihood of the answer and can analyze a lot of data to generate replies (Pamungkas, 2019), newer architectures managed to improve chatbots even further. Researchers argued that for a conversation to be natural, a chatbot should respond on the basis of the current user query but also use previous queries as well (Cahn, 2017). To achieve this, Long Short Term Memory (LSTM) (Hochreiter et Schmidhuber, 1997) based architectures were used in chatbots to refer to previous information and learn long-term dependencies. Using LSTM achieved good results when designing a conversational agent for elderly care (Su *et al.*, 2017) for instance. Other researchers used LSTM for response generation instead (Xu *et al.*, 2017). They used word2vec to represent user queries with vectors and 2 LSTM networks, one as an encoder and the other as a decoder.

In a broader sense, Generative AI, be it based on seq2seq models as used in (Belainine *et al.*, 2022) or other advanced architectures, can produce outputs that are not limited to predefined options or fixed sets of responses (Bail, 2023). This is because they generate responses by modeling the probability distribution of the next word or sequence of words given the context. However, one limitation of generative models is the need for massive training data. This led to the development of chatbots being dominated by open-domain systems because of publicly-available open-domain data (Bilquise *et al.*, 2022).

The Transformer architecture published in 2017 (Vaswani *et al.*, 2017) is considered one of the most important milestones in chatbot development and language models in general because it drastically changed the way newer chatbots are designed. This new architecture helped language models better understand the relationships and dependencies in input sequences by using techniques such as attention mechanisms and self-attention layers. Through the use of intensive pre-training on enormous volumes of text input, LLMs based on the Transformers architecture like OpenAI's GPT (Radford *et al.*, 2018) considerably improved chatbot capabilities generating interesting and original interactions. The transformer model was explained more in-depth in section 1.4.2, and in the rest of this literature review, we are going to focus primarily on chatbots that are generative-based since it is now the dominating approach to building modern chatbots (Pamungkas, 2019).

2.3 Chatbot Application Architecture

There are several chatbot architectural designs proposed in the literature. However, some are specific to retrieval-based chatbots (Wu *et al.*, 2016) or to rule-based conversational agents (Khanna *et al.*, 2015). More recently, (Adamopoulou et Moussiades, 2020a) proposed a general chatbot architecture that is valid for retrieval-based systems, rule-based systems as well as generative-based chatbot systems. It includes five main components:

User interface (UI) This component is the part that is exposed to the user and in which the chatbot collects the requests/questions and shows answers. It can be a mobile app, a web app, or a conversation inside an established instant messaging application. The UI needs to be well-designed and simple to use if we want the users to be satisfied with the interaction (Gnewuch *et al.*, 2018).

User Message Analysis This component preprocesses the user input and includes a spell checker to correct spelling mistakes, a machine translation model if the chatbot is multilingual, and a sentiment analysis/emotion recognition module to detect users' psychological state.

Dialog Management Sometimes, the user's input is insufficient to determine the context. In such cases, the chatbot can ask additional questions to collect contextual information. This component is helpful in handling ambiguity, collecting data about the user, and correcting potential errors.

Backend After processing the user's input, the chatbot connects to a database to retrieve the necessary information for a proper response. Ontologies like Wordnet (Miller, 1995) can be used at this point to find connections between nodes in the knowledge graph and, therefore, "understand" the meanings and relationships between words.

Response Generation Depending on the nature of the chatbot, this component uses a specific technique to generate the best response to the user's request using the information collected from previous components. Generative, machine learning-based chatbots, in particular, use Natural

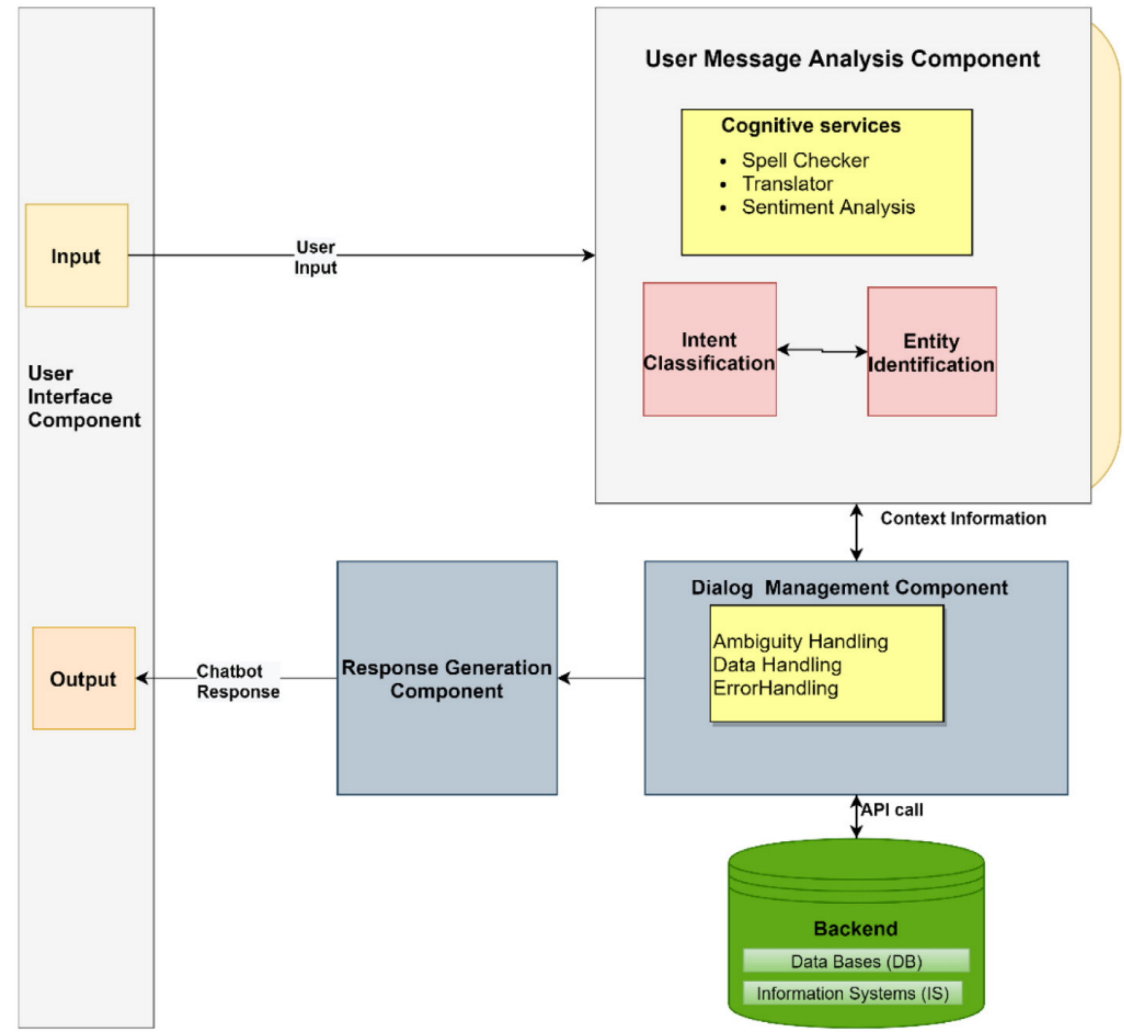

Figure 2.1: A general architecture design for chatbots (Adamopoulou et Moussiades, 2020a).

Language Generation (NLG) modules to respond in a natural, human-like fashion. As discussed earlier, this needs a big training dataset to achieve good results. The response is presented through the user interface, and the chatbot waits then for the following query.

2.4 Prompt engineering and language models

2.4.1 Prompt engineering

The exceptional performance of Large Language Models (LLMs) on a variety of tasks, even with zero-shot or few-shot settings, has inspired NLP researchers to reevaluate the predominant training paradigms from previous years, and prompt engineering is an excellent example of that. Prompt Engineering can be defined as the design of instructions (prompts) in a way that improves the quality of the results from existing language models without further training on new datasets (Reynolds et McDonell, 2021). This is a relatively new and promising technique that appears to have the potential to greatly improve LLMs' performance on downstream tasks.

Many researchers leveraged the art of prompt engineering to effectively instruct language models leading to the desirable results. For example, in the context of zero-shot mathematical reasoning, (Kojima *et al.*, 2022) found that simply prompting GPT-3 language model with "Let's think step by step" after the mathematical question quadrupled the accuracy on the MultiArith arithmetic dataset, from 18% to 79%! The authors noted, however, that this method works well only with arithmetic problems that need multiple steps and did not really help with commonsense questions, for example. Prompt engineering was also used to improve neural machine translation (Li *et al.*, 2022). It was shown that leveraging prompts does help with the quality of the translations and enhances the flexibility of human-in-the-loop translations.

Authors in (Wei *et al.*, 2022) investigated chain-of-thought (CoT) prompting as a simple and generally applicable method to enhance reasoning abilities in LLMs. To produce a human-like thought process and improve reasoning, they used CoT prompting with language models by including a few examples of chain-of-thought using only prompts instead of fine-tuning as (Cobbe *et al.*, 2021) did. This resulted in improvements across a variety of tasks, such as arithmetic, commonsense, and symbolic reasoning tasks. For instance, on grade school math problems, it was shown that chain of thought prompting tripled the solve rate from 18% to 57%. Selection-inference prompting is later proposed as an extension to Chain-of-Thought prompting (Creswell *et al.*, 2022). It consists of dividing the single question prompt for generating explanations and answers into multiple questions: A selection prompt first selects a pertinent subset of facts from the text. Then, a second inference prompt draws a conclusion from the data chosen. These prompts are then alternated in a sequence to generate multiple reasoning steps and, ultimately, the final answer. The authors showed that selection-inference prompting outperformed chain-of-thought prompting on bAbi and Proof Writer benchmarks.

Prompt engineering was also used with conversational language models. For example, (Polak et Morgan, 2023) proposed ChatExtract as a prompt-engineering-based method that works even with limited background to fully automate accurate data extraction from research papers. The set of engineered prompts applied to a conversational LLM not only extracts the relevant data but also verifies the data's correctness through a series of follow-up questions, overcoming known issues with LLMs like providing factually inaccurate responses (Lee *et al.*, 2022). (White *et al.*, 2023) even proposed an entire prompt engineering patterns catalog to get the most out of ChatGPT (OpenAI, 2022). The catalog consists of over 15 prompt patterns designed for a variety of tasks, such as output customization and context control, leading to improved output from the conversational language model for different domains and contexts. It doesn't include emotion or tone control prompts, though.

Emotion stimuli through prompt engineering has also been used to enhance the responses of Large Language Models. In (Li *et al.*, 2023), researchers explored the integration of emotional intelligence into several Large Language Models to see its impact on understanding performance. Automatic experiments on 45 tasks involving language models, such as Flan-T5-Large (Chung *et al.*, 2022), Vicuna (Zheng *et al.*, 2023), Llama 2 (Touvron *et al.*, 2023), BLOOM(Workshop *et al.*, 2022) and ChatGPT (OpenAI, 2022), reveal that LLMs indeed possess a grasp of emotional intelligence. The introduction of "EmotionPrompt," a fusion of the original prompt with emotional stimuli, leads to notable improvements, including an 8% relative performance boost in Instruction Induction and a remarkable 115% improvement in BIG-Bench with more improvements in few-shot settings when compared to zero-shot settings. They also conducted a human study, involving 106 participants and found that their technique achieves up to 1.0 in Relative Gain in a third of the problems. The authors also noticed that EmotionPrompt stimulates the creative faculties of LLMs. In terms of truthfulness, and responsibility metrics, LLMs prompted with EmotionPrompt showcased 19% and 12% of average improvements in truthfullness and informativeness scores. Researchers attributed these improvements to the fact that emotional stimuli can enrich the original prompts' representation and positive words in the prompts might have contributed too. They concluded their work by acknowledging that, while LLMs can be enhanced by emotional intelligence, there are still mysteries to unravel, leaving room for further exploration at the intersection of LLMs and psychology.

2.4.2 Dialogue models

Conversational language models (or dialogue models) have received a lot of attention in recent years (Zaib *et al.*, 2020). We can distinguish three main categories of these dialogue systems based on their functionality:

- **Task-oriented systems** are developed to have dialogues with users to carry out a specific task. These systems can be found in voice-based user interfaces, virtual assistants, and customer support.
- **Question-answering systems** are intended to provide specific answers to user queries. These systems aim to extract relevant information from a given knowledge source or corpus.
- **Open-domain chat agents** are designed to engage in open-ended discussions with users over various subjects. Open-domain chat-oriented systems try to create human-like and engaging interactions, unlike task-oriented dialogue systems that concentrate on completing specified tasks.

Many conversational systems based on Large Language Models (see section 1.4) were proposed in the literature, and most of them are based on the Transformer architecture (Zaib *et al.*, 2020). For instance, the DialoGPT (Zhang *et al.*, 2019) model was trained on 147 million conversation-like exchanges extracted from Reddit comment threads. Thus, it is a conversation language model that has encoding for the dialogue structure knowledge within their parameters. However, there is a recent, disruptive conversational language model that has taken the internet by storm since its introduction: ChatGPT (OpenAI, 2022).

2.4.3 ChatGPT

ChatGPT (OpenAI, 2022) is a Large conversational Language Model based on the GPT-3.5 architecture and developed by OpenAI (Ouyang *et al.*, 2022). Since its release as an online platform in November 2022, it has received a lot of attention from the public as well as from researchers.

Since its debut, ChatGPT has attracted a record 100 million subscribers in only two months, overcoming other very popular online platforms such as Facebook and Instagram (Haque *et al.*, 2022). ChatGPT reached the 1 Million user mark within five days only, whereas Instagram, Facebook, and Netflix needed 75, 300, and 1200 days to reach the same number (Haque *et al.*, 2022), respectively.

2.4.3.1 A (mostly) well-received newborn

The feedback around ChatGPT has been mostly positive from the public on social media and news outlets (Leiter *et al.*, 2023). Haque et al. (Haque *et al.*, 2022) examined how well the new chatbot was received using sentiment analysis on Twitter data and found that the social media platform users exhibited positive sentiments when talking about ChatGPT for topics such as entertainment, NLP, and software development with positive sentiments representing up to 92%, 83%, and 82% respectively. On the other hand, the topics in which users showed the least positive sentiments were Q&A testing and impact on the educational aspect, with only 38% and 54% of the tweets being positive, respectively. This indicates that, even though the feedback was mostly positive, some were concerned about the implications of ChatGPT on the future of education and information integrity and reliability.

While regular users enjoyed its features and have been using it to accomplish a variety of tasks, researchers were interested in studying this technology and exploring its capabilities and limits in multiple aspects.

2.4.3.2 A disruptive technology

ChatGPT's success can be attributed to its ability to converse with humans in various natural language tasks; from straightforward queries to more complicated dialogues in an incredibly fluent and coherent fashion (Guo *et al.*, 2023). Here are some of the features that made ChatGPT so engaging to users that it reached record-breaking adoption rates:

- **Contextual Understanding:** ChatGPT exhibits a good capacity for comprehension and context maintenance during discussions. In order to provide logical and contextually appropriate replies, it may draw on the complete conversation history instead of only considering the last user utterance. This leads to more interesting and organic interactions with users.
- **Natural Language Generation:** Thanks to its thorough training on various datasets, ChatGPT excels at producing writing that sounds human. By using the statistical patterns and linguistic structures it has acquired throughout training, it may provide comprehensible and appropriate replies for the given situation.
- **Creative and Dynamic Responses:** ChatGPT is renowned for its capacity to produce original replies, which can enhance user engagement. It can make ideas, come up with logical extensions, or even have imaginary talks about particular subjects.

Even though OpenAI has never published the technical details and specific architecture, we already know some techniques that helped achieve such an impressive system. The model was trained on a massive corpora blend of text and code (Neelakantan *et al.*, 2022) and continuously improves using Reinforcement Learning from Human Feedback (RLHF) (Stiennon *et al.*, 2020; Christiano *et al.*, 2017), which has become the go-to technique to align LLMs with a human's actual intent.

(Zhou *et al.*, 2023) explain the RLHF approach used in ChatGPT as follows: First, an extensive dataset, including prompts and the expected output behaviors, is collected, and GPT-3.5 is run on this data. Then, the refined model provides a variety of model outputs in response to the same prompt. To create a comparison dataset, a labeler assigns the required score and ranks the output, which is then utilized to train the reward model. Finally, using the Proximal Policy Optimisation (PPO) RL (Reinforcement Learning) algorithm, the fine-tuned model, ChatGPT, in this case, is optimized against the reward model. See 2.2 for an illustration of this process.

2.4.3.3 ChatGPT Under the Spotlight

There has been a growing interest among researchers in evaluating the new conversational language model to find how well it copes with different tasks, from machine translation to coding and reasoning.

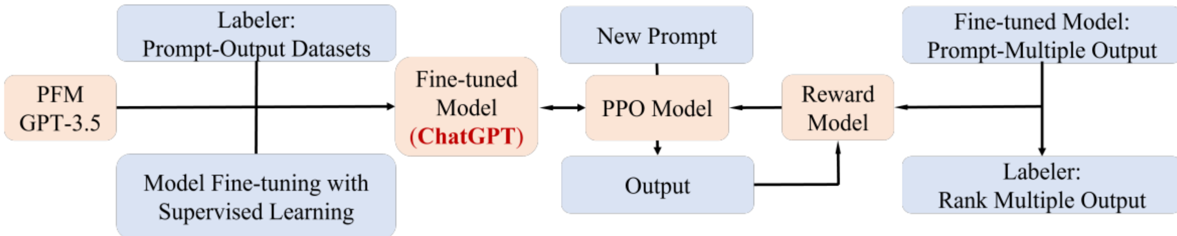

Figure 2.2: How reinforcement learning from human feedback is used for ChatGPT (Zhou *et al.*, 2023).

In **Machine translation**, a study found that ChatGPT outperforms professional translation tools for high-resource European languages, but for low-resource or far-off languages, it falls short (Jiao *et al.*, 2023). The use of pivot prompting (translating to a high-resource language before translating that to the target language) greatly enhances translation performance. However, their findings indicate that ChatGPT is not as reliable as commercial systems for biological abstracts or Reddit comments. Using the GPT-4 as the engine dramatically improves the translation results, especially for low-resource languages.

Another study examined ChatGPT's accuracy in translating between English and languages that exclusively use gender-neutral pronouns like Bengali (Ghosh et Caliskan, 2023). They found that ChatGPT shares similar gender biases as other translation tools. In fact, it turns out ChatGPT assigns gendered pronouns based on biases and preconceptions associated with particular professions and behaviors. Additionally, it mistranslates the pronoun "they," which is gender-neutral, leading to incorrect interpretations, and when asked about gender, ChatGPT shows greater respect for males than women in the same profession.

However, a team participating in the AmericasNLP 2023 Shared Task on Machine Translation into Indigenous Languages found that ChatGPT performs very poorly on translating to these south american languages (Stap et Araabi, 2023). They submitted translations for 11 languages using four different systems including a GPT-4 ChatGPT model. The latter didn't outperform any other model on any indigenous language, showing that it is not well suited for these low-resource languages.

Text summarization is another task for which ChatGPT was tested. Researchers found that, even though ChatGPT does give excellent results in terms of ROUGE score, it was outperformed by GPT-3.5 and tended to produce more extended summaries (Qin *et al.*, 2023). (Bang *et al.*, 2023) found that ChatGPT, with zero-shot, even outperforms some fine-tuned language models but with zero-shot. However, they found that this is not consistent as it sometimes outputs a text summary that is longer than the original text, for example.

Other researchers conducted an evaluation of ChatGPT's aspect and query-based summarization, which is more complex than generic summarization and requires a deeper level of understanding. The results showed that ChatGPT performs on par with conventional fine-tuning techniques on diverse benchmark datasets, including summaries from Reddit postings, news articles, conversation meetings, and tales (Yang *et al.*, 2023).

Another study (Pu et Demberg, 2023) explores ChatGPT's text generating capabilities in tasks such as text summarization. They found out that ChatGPT is not as good as humans at catching the depth of stylistic variances. It has fewer ranges and sometimes makes factual mistakes or hallucinations. The study's quantitative and qualitative experiments demonstrate that, despite ChatGPT's superior performance when compared to earlier models in automated metrics, human-written text and its output still differ significantly. These discrepancies are mitigated by offering a target example of a human writing style; yet, problems like errors and hallucinations continue to appear in text created by ChatGPT.

Reasoning and mathematical capabilities were also thoroughly examined by researchers. ChatGPT can make logical connections and draw conclusions based on the information provided, proving that it does have reasoning abilities. It was shown that ChatGPT has impressive arithmetic reasoning abilities, outperforming GPT-3.5 in 5 out of the six datasets the researchers used (Qin *et al.*, 2023). However, they showed that it wasn't the case with common sense and logical reasoning.

Authors in (Frieder *et al.*, 2023) examined ChatGPT's mathematical capabilities and found that, contrary to the media hype, ChatGPT is not ready to consistently deliver top-notch proofs or calculations in advanced mathematics, yet. However, there are positive surprises in the quality of answers. It was shown that the language model still falls short on graduate-level difficulty. ChatGPT is found to be inconsistently proficient in advanced math, excelling in some insightful proofs but generally struggling with difficult problems. It is not on par with task-specific models but shines in flexibility, serving as a universal tool for diverse mathematical areas.

In **Programming and coding**, ChatGPT can be used as a programmer's helper, offering recommendations for code, troubleshooting, and clarifying programming concepts. ChatGPT was evaluated in terms of bug-fixing efficiency on the QuixBugs dataset and was compared to other methods proposed in the literature (Sobania *et al.*, 2023). The results showed that ChatGPT outperforms standard program repair techniques and competes favorably with well-known deep learning techniques like CoCoNut and Codex.
Moreover, ChatGPT, thanks to its conversational nature, allows users to provide additional information, such as expected outputs or observed error messages, for further assistance. This feature helped ChatGPT to surpass state-of-the-art techniques.
A technical report (Kashefi et Mukerji, 2023) analyzed the numerical programming capabilities of ChatGPT. They showed that the language model has the capacity to create programs in several programming languages, improve and debug code, finish off missing portions, and even parallelize C++ routines using OpenMP. However, the authors found that ChatGPT has some difficulties including the creation of singular (non-invertible) matrices, the management of arrays of compatible shapes etc. Moreover, it turns out ChatGPT struggles to identify if a code portion is human-written or ChatGPT-generated.

ChatGPT was also examined in terms of **Sentiment analysis** capabilities, and it largely outperformed state-of-the-art zero-shot models in English and Indonesian and performed just as well in Buginese (Bang *et al.*, 2023). Another study found that ChatGPT exhibited similar performance when compared to BERT on the same sentiment analysis task (Zhong *et al.*, 2023). It was also shown that its performance is comparable to GPT-3.5 (Qin *et al.*, 2023). In another, more in-depth article, researchers tested ChatGPT on five sentiment analysis tasks with 18 benchmarking datasets and found that it has a similar performance to fine-tuned BERT, without being fine-tuned itself (Wang *et al.*, 2023). They also demonstrated that giving a couple of examples to the chatbot (few-shot learning) helped the model improve even further, even though it didn't significantly outperform state-of-the-art fine-tuned models.
(Lai *et al.*, 2023) evaluated the performance of ChatGPT, beyond English, on many NLP tasks such as NER, NMT, POS, NLI, QA, and CSR. (Kocoń *et al.*, 2023) tried to evaluate ChatGPT on 25 different NLP tasks and found that it did very well in most of them but didn't outperform the state of the art in any particular task.

The **emotional capabilities** of ChatGPT were studied as well. Through a series of trials on various downstream tasks such as emotion understanding and emotion generation, a study assessed the effectiveness of ChatGPT on emotional conversation interpretation and creation (Zhao *et al.*, 2023). Their results showed that ChatGPT demonstrates promising outcomes in evoking emotional reactions, even if its performance on emotional conversation interpretation may still lag below that of supervised models.

In fact, ChatGPT showed promising results in the emotion generation capabilities, such as empathetic response generation and emotional support conversation tasks, outperforming state-of-the-art models. However, it didn't outperform state-of-the-art fine-tuned models on emotion understanding tasks (emotion recognition, emotion cause recognition, and dialogue act classification). This shows that more work can be done to enhance the emotion recognition capabilities of the model. To the best of our knowledge, no work has been done yet to improve this aspect of ChatGPT.

Another recently published study (Wake *et al.*, 2023) explored ChatGPT's potential uses in data annotation and mental health analysis by delving into its ability to identify emotions in text across multiple datasets such as DailyDialogue (Li *et al.*, 2017). The experiments conducted revealed a respectable degree of emotion recognition reproducibility, with even more improvements seen through fine-tuning. However, it turns out performance varies between datasets and emotion labels, suggesting bias and intrinsic instability. The labeling techniques and training data biases are probably behind this variation in recognition performance. The authors concluded that even though the language model behind ChatGPT shows impressive proficiency, we need to be cautious when applying such models in sensitive domains like mental health.

2.4.3.4 Concerns

ChatGPT has received a lot of attention and appreciation for its exceptional abilities to produce human-like replies and carry out numerous language-related activities. However, the chatbot has also sparked debate both inside and outside of the AI community.

One major concern is the potential for **biases** present in the data used for training ChatGPT, leading to biased or discriminatory outputs. It was shown that ChatGPT exhibits gender bias, especially with professions associated with particular genders (Ghosh et Caliskan, 2023). Another article presented a comprehensive study on the different kinds of biases that ChatGPT could exhibit and identified 24 sorts of biases: from cultural and linguistic bias to hindsight bias (Ray, 2023).

Moreover, ChatGPT tends to generate plausible but inaccurate information, also known as language model **hallucination** (Lee *et al.*, 2022). It was shown that ChatGPT suffers from intrinsic hallucinations (output that contradicts the input) as well as from extrinsic hallucinations (output that cannot be verified by the source), the latter being more frequent (Bang *et al.*, 2023).

Another concern is related to **education**. While ChatGPT can enhance personalized learning and facilitate access to knowledge, the conversational model was so good at producing human-like text and at question answering that some considered it to be a threat to the future of exam integrity. Some students are already using it to cheat in assignments (Susnjak, 2022).

2.5 Emotions in conversational systems

Emotions are states of feelings resulting from internal or external changes in our lives and depend on the speaker's attitude and personality (Al-Omari *et al.*, 2020). Incorporating emotions into conversational systems is a crucial step in making human-like conversational systems, even though it can be challenging to achieve such a task.

2.5.1 The importance of emotions in conversational systems

Despite all the advancements the conversational agents' research, it appears that people still prefer natural communication to machine-like interactions and feel that a human can understand them better (Rapp *et al.*, 2021). Furthermore, it was shown that customers still prefer interacting with humans over machines (Adam *et al.*, 2021).

In fact, a study showed that for customer service, for example, 40% of consumers' requests are rather emotional without specific informational intents (Xu *et al.*, 2017). Thus, building sympathetic and successful conversational agents requires the capacity to recognize and respond to the emotional cues that are present in all human dialogue.

However, generating empathetic and human-like responses is a challenging task for chatbots as it requires an understanding of the complex user's emotional state and the ability to respond appropriately. Much work was done to address this challenge of building emotion-aware conversational systems, and it was shown that the interest particularly grew since the year 2018 and that most of the solutions proposed in the literature were text-based conversational agents (Bilquise *et al.*, 2022).

The first chatbot developed with emotions in mind, PARRY, dates back to 1971, and it was a rule-based system that played the role of a patient with schizophrenia (Colby *et al.*, 1971). It was even tested later by multiple psychotherapists to see if they could determine if he was a bot or a human (Heiser *et al.*, 1979). Today, most emotionally aware chatbots are neural-based they mostly use seq2seq and encoder-decoder architectures.

2.5.2 Challenges and techniques used for incorporating emotions into chatbots

Researchers tackling the problem of developing emotion-aware conversational systems have used many techniques that address different aspects of human-like and emotionally-aware chatbots. Some of them worked on improving the detection of the user emotion, and others tried to generate emotional responses better, while others tried to avoid the dull and meaningless responses generated by basic seq2seq models.

2.5.2.1 Emotion Capture

Accurately identifying and comprehending emotions in user input is essential for developing emotionally intelligent responses. However, given the complexity and subjectivity of emotions, it might be challenging to identify them from writing. The proper recognition of emotions is further complicated by variations in linguistic expression, sarcasm, and cultural quirks, making it difficult to understand user input accurately.

(Casas *et al.*, 2021) used DeepMoji (Felbo *et al.*, 2017) as an emotion classifier to detect the user's emotions and classify them following Ekman's model of six basic emotions (discussed in section 1.5.2) before injecting them into the dialogue system. Other researchers tried to consider not only the last user input's emotion but instead use the conversation history. For instance, in the work of (Qiu *et al.*, 2020), the system tracks the user's emotional state using a transition network. (Hasegawa *et al.*, 2013) claimed that predicting the user's emotional state from past conversational utterances rather than a single speech is the only way to accomplish genuinely natural conversation.

However, according to a number of studies, emotions are complex and cannot be adequately described by a coarse-grained emotion label. In order to do this, some authors proposed to use model emotions using a 4-dimensional space (Belainine *et al.*, 2020b). Others attempted to use another emotion modeling using Valence, Arousal, and Dominance (VAD) (Warriner *et al.*, 2013) as the three dimensions for embedding for each of the words of the input text (Asghar *et al.*, 2018).

2.5.2.2 Emotional response generation

It can be challenging for a chatbot system to come up with varied and emotionally relevant replies for each emotional situation during dialogues since there could be a broad spectrum of emotional states. In order to make sure that the generated answer closely matches the desired emotional state, it is important for chatbots to be able to modify their responses based on the user's emotional signals, and there has been significant research focused on enhancing conversational agents' ability to produce emotionally expressive responses.

(Lin *et al.*, 2019) proposed the empathy hypothesis stating that the type of generated emotion should be consistent with the contextual emotional state of the user. Some studies have incorporated a target emotion into the response generator module to ensure this consistency. For instance, (Zhang *et al.*, 2017) designed a chatbot that outputs multiple responses, one for each of the six chosen emotions (like, sadness, disgust, anger, happiness, and other) using a multi-task seq2seq model with GRU (bidirectional LSTMs). Then, the system selects the most appropriate response using intra-ranking and inter-ranking policies. Similarly, Zhou et al., in their famous model called the "Emotional Chatting Machine", introduced three novel mechanisms for the decoder: emotion category embedding, internal emotion memory, and external memory (Zhou *et al.*, 2018). Other researchers proposed to achieve the same goal by conditioning the response generator to produce polite, rude, or neutral responses (Niu et Bansal, 2018).

Even though these approaches achieved good results overall, it has been argued that we cannot assume that the output emotion has to match the input emotion (Wei *et al.*, 2019). Researchers claimed that using a predefined label to train the response generator results in poor response quality. To address this, authors proposed to generate different responses for different emotion labels instead: (Zhang *et al.*, 2018) introduced a model that generates multiple responses for six emotional categories, and the best response is then selected with a ranking algorithm. (Colombo *et al.*, 2019) designed a system that generates multiple responses using two seq2seq models before ranking them to finally output the best reply. Their proposed conversational system, called EMOTICONS, works by modeling emotions using three techniques: emotion vector representation, an affect regularizer, and affect sampling. They also used emotional re-ranking to produce the final output that is most emotionally relevant. Experimental results showed that their EMOTICONS system outperformed. In other studies, topics and emotions were included in the decoder to allow the chatbot model to provide replies that were emotionally suitable (Zhang *et al.*, 2020).

Reinforcement Learning (RL) techniques have also been explored to enhance the emotional responses of conversational agents. Researchers have developed reward models that provide feedback on the emotional quality of generated responses. RL was first used in emotionally-aware chatbots in (Sun *et al.*, 2018) to build a chatbot that detects human emotions before giving the appropriate response by rewarding the subhuman conversational sequences more. They applied a kind of Generative Adversarial Networks (GAN) mechanism on the conversational model called seqGAN. They added emotional tags to the post and response from the dataset and obtained a system that can generate sound, appropriate responses in both content and emotion. More authors used reinforcement learning techniques for emotionally-aware chatbots. (Li *et al.*, 2019a) combined reinforcement learning with emotional editing constraints to develop affect-driven and emotional replies that work in three iterations. Their proposed model introduced a method for multi-task learning in order to learn three aspects: coherence, topic, and emotion. This yielded excellent results compared to models that only consider one aspect at a time.

Some researchers were interested in making the chatbot responses more diverse and rich while being emotionally relevant. In fact, according to Asghar et al. (Asghar *et al.*, 2018), neural conversational models often provide brief and unclear replies and do not adequately reflect the complexity of emotions. To ensure variety in the generated replies, they proposed using a heuristic search technique and adopted the VAD affective space (Warriner *et al.*, 2013) in an attempt to obtain more diverse responses. Their approach introduced three innovative techniques to incorporate emotions into chatbots: cognitively engineered affective word embedding, modified cross-entropy loss based on emotions, and an affectively diverse beam search for the decoder. In order for a chatbot to sound human, it shouldn't reply with the same response, even if the input message is the same. That is what (Yao *et al.*, 2021) tried to prove in their work . Their proposed model also outputs multiple responses that take emotion into consideration.

2.5.2.3 Datasets

Obtaining relevant and labeled data is a major challenge in developing emotionally intelligent chatbots. Large and varied datasets that correctly capture a wide range of emotional expressions and circumstances are needed for training models for emotion creation and interpretation. However, obtaining good quality data can be difficult due to the following reasons:

- **Limited data availability:** While many conversational datasets are available in the open domain, it is difficult to find conversational datasets that are labeled with emotions.
- **Unbalanced datasets:** Emotion-labelled datasets tend to be unbalanced. This is when certain emotions are overrepresented while others are underrepresented. We need to take this into consideration, especially when choosing the evaluation metrics.
- **Subjective bias:** Since emotion labeling is a subjective process, various annotators may have varying interpretations and assigned labels. Emotional annotations also risk being skewed toward specific cultural norms or linguistic expressions.

Some studies tried to address this emotionally-annotated conversation data challenge. Two particular datasets are often used to develop and evaluate emotion-aware conversational agents: DailyDialogue (Li *et al.*, 2017) and Empathetic Dialogues (Rashkin *et al.*, 2018). While the former is annotated with emotions on each utterance level, the latter has emotion labels for whole conversations reflecting the overall nuance.

Moreover, some researchers tried not to be limited to these two datasets by using approaches to label the available datasets at hand, especially when they needed domain-specific emotionally annotated data. (Zhou *et al.*, 2018) annotated the conversational data automatically with six emotion labels using the classifier that gave the best results on the NLPCC 2013 and NLPCC 2014 datasets. Similarly, (Huang *et al.*, 2018) used another emotion classifier to automatically assign emotion tags to each of the 1 million conversations in the dataset they used, but with five emotion labels.

2.6 Conclusion

In conclusion, this chapter provides a thorough overview of the evolution and developments of chatbot technology, with an emphasis on conversational language models like ChatGPT. The historical development of chatbots was examined, showing how they progressed from basic rule-based systems to complex generative-based conversational agents leveraging the power of LLMs.

As a prominent example of conversational language models, ChatGPT has proven to have exceptional abilities in recognizing and producing coherent and contextually appropriate text. It has established itself as a leading technology in the conversational systems space thanks to its capacity to make use of enormous training data sets and comprehend the subtleties of human language. Researchers have evaluated ChatGPT in various scenarios to gauge performance and pinpoint areas for further development of conversational systems.

Emotions are an essential part of human conversations, and conversational systems have begun to include them more and more. This chapter emphasized the value of emotional intelligence in chatbots and the different techniques proposed in the literature to better capture and express emotions, providing more meaningful and empathetic conversations and therefore enhancing user experiences. We present our approach to tackle this problem in the next chapter.

CHAPTER 3
METHODOLOGY AND IMPLEMENTATION

3.1 Introduction

As elaborated in chapter 2, The evolution of conversational agents has made significant strides in recent years. These chatbots are utilized in a variety of applications, such as customer service, virtual companions, and mental health support systems. This chapter presents our approach to enhancing the emotional capabilities of ChatGPT, the state-of-the-art language model. By imbuing ChatGPT with emotional awareness, we hope to produce more emotionally intelligent chatbots that can empathize with users and provide personalized experiences.

As a first step, we start by building an emotion classifier based on ELECTRA and fine-tuned on the GoEmotions dataset. This classifier allows us to detect and categorize user emotions in addition to its potential for evaluation. To ensure accurate emotion prediction, we delve into the model specificities, the proposed architecture, and the exploration of the dataset used for training. In the second step, we present three variants of ChatGPT with different emotional capabilities. The first version, ChatGPT-A, represents the foundation model and is used for comparison purposes. Then, we present ChatGPT-B and ChatGPT-C, which assess the user's affective state during a conversation using distinct explicit and implicit approaches, respectively.

In the following sections of this chapter, we will delve into the emotion classifier specifics as well as the details of each iteration of ChatGPT, elucidate on their implementation, and discuss their potential.

3.2 Emotion Classification

The first part of our work is building a reliable emotion classifier that is able to discern even the smallest nuances in emotion cues using only text. We will discuss later in section 3.3 how this will be useful for emotion infusion as well as for the evaluation of our conversational systems in the next chapter.

In order to build such a classifier, we propose to use an advanced language model that was proven to be effective in tasks similar to ours: ELECTRA. We leverage the power of transfer learning to adapt this pre-trained model to our specific downstream task: emotion classification.

3.2.1 About the ELECTRA Model

ELECTRA (*Efficiently Learning an Encoder that Classifies Token Replacements Accurately*) model is a type of neural network architecture that was introduced by researchers at Google (Clark *et al.*, 2020). It has been shown to outperform other pre-trained language models such as BERT (Devlin *et al.*, 2018) on several NLP benchmarks, including sentiment analysis (Mala *et al.*, 2023)

The main contribution behind the ELECTRA model is *replaced token detection* instead of *masked token prediction*. In fact, for popular LLMs like BERT (Devlin *et al.*, 2018), XLNet (Yang *et al.*, 2019a), and RoBERTa (Liu *et al.*, 2019), the pre-training job is masking a portion of the unlabeled input and then training the network to retrieve this original input. Even though his technique yields good results, its data efficiency is always limited because the model only learns from a fraction of the tokens. Researchers from Stanford University and Google Brain (Clark *et al.*, 2020) proposed replacing specific tokens with plausible substitutes produced by a small language model as an alternative to masking, then trying to determine if each token is an original or a replacement using the pre-trained discriminator. This resulted in a significantly more computationally efficient model thanks to learning from the entire set of input tokens.

In fact, studies such as (B *et al.*, 2023) have shown that this proposed method dramatically speeds up training and improves performance on downstream NLP tasks. They compared the ELECTRA model to other state-of-the-art models such as BERT, XLNET, and RoBERTa on the sentiment analysis capabilities. They found that ELECTRA performs the best in this task with up to 93% classification accuracy on the IMDB movie reviews dataset, compared to 85% from BERT, 86% from XLNet, and 92% from RoBERTa. Therefore, this model is well suited for our multiclass emotion classification task, being a closely related task to sentiment analysis (Devaram, 2020).

3.2.2 Transfer learning

Emotion recognition & classification is one of the most challenging problems in automated language understanding. In order to optimize the ELECTRA model for this specific downstream task, we

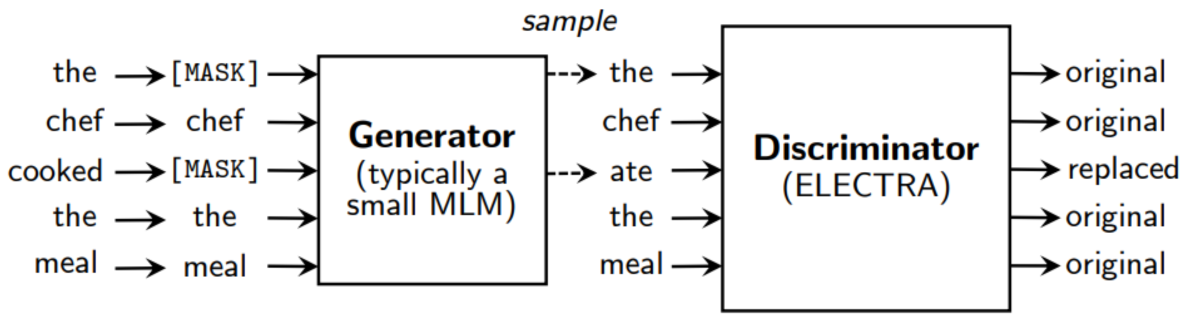

Figure 3.1: An overview of replaced token detection (Clark *et al.*, 2020).

will leverage the power of Transfer learning.

Transfer learning is a powerful technique that is now often harnessed by researchers and practitioners in order to tackle new problems and attain state-of-the-art performance in a variety of applications thanks to the effectiveness it has demonstrated in adapting pre-trained language models to specific-purpose downstream tasks (Belainine *et al.*, 2020c; Belainine *et al.*, 2020a). It provides effective and efficient learning on small amounts of labeled task-specific data by utilizing the information learned from another, more general task, often a large-scale pre-training job, and transferring it to the target task (Chronopoulou *et al.*, 2019). This is especially advantageous in NLP applications, where pre-trained language models capture many linguistic patterns and semantic correlations. Transfer learning enables us to start with a pre-trained model and refine it using data unique to the target task instead of training a language model from scratch on the target task.

Formal definition: According to (Weiss *et al.*, 2016), Transfer learning can formally be defined as follows:
Given some observations corresponding to m^S source domains and tasks and some observations about m^T target domains and tasks, transfer learning utilizes the knowledge implied in the source domains to improve the performance of the learned decision functions on the target domains $f_i(i = 1, ..., m^T)$.
Here, a domain D consists of two parts: a feature space $\mathcal{X}$ and a marginal distribution $P(X)$: $D = \{\mathcal{X}, P(X)\}$ and a task T is composed of a label space $\mathcal{Y}$ and a decision function f: $T = \{\mathcal{Y}, f\}$. The decision functions are expected to be learned from the data.

In more practical terms, by re-training (or fine-tuning) the pre-trained model on a dataset that is tailored to the new task (emotion classification in our case) while keeping some of the pre-trained model weights in the first few layers unchanged, the model we obtain is adapted (or fine-tuned) to the new task (Zhuang *et al.*, 2020). This is illustrated in figure 3.2.

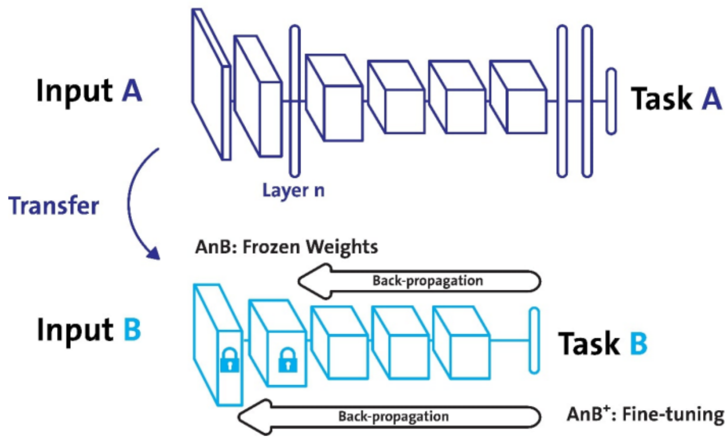

Figure 3.2: Transfer learning illustration (Kamath *et al.*, 2020).

Through this approach, the model can take the broad language comprehension from the source job and modify it to fit the specifics and criteria of the target activity. We can efficiently overcome data scarcity, shorten training times, and increase performance on a variety of downstream NLP tasks, including sentiment analysis, named entity identification, text categorization, and emotion classification, with faster convergence and greater performance using a fraction of the processing power (Pan et Yang, 2010). As mentioned earlier, the pre-trained language model we are going to use is the ELECTRA model, and we fine-tune it with an emotion-labeled dataset.

3.2.3 Dataset

3.2.3.1 About the GoEmotions dataset

In order to adapt the ELECTRA model to the emotion classification task, we used the GoEmotions dataset developed and published by Google (Demszky *et al.*, 2020). This is a large dataset of over 58,000 Reddit comments manually annotated by multiple human annotators. The sentences are emotion-annotated with a set of 27 fine-grained emotion labels in addition to the neutral emotion, and the labels range from basic emotion tags like joy, love, and anger to more complicated ones like nervousness, relief, and caring. The authors argue that the 28 chosen emotion labels are highly significant according to the Principal Preserved Component Analysis (PPCA) (Cowen *et al.*, 2019).

What sets the GoEmotions dataset apart is its rich, fine-grained emotion labeling approach. It encompasses complicated and subtle feelings like adoration, amusement, hatred, and nostalgia, in addition to fundamental emotions such as joy and fear. This thorough annotation makes the dataset a promising resource for researchers to better comprehend the emotional content of the text and enables them to dive into the complexity of human emotions.

3.2.3.2 Exploratory data analysis

In order to effectively use the GoEmotions dataset, we need to analyze it to understand its characteristics as well as its potential biases and limitations.

We have 211,225 rows of textual data along with 37 columns, 28 of which are binary annotations for the 28 emotion labels. It is important to note that we have multiple instances of each Reddit comment (having the same `id`). This is because a given text could be annotated by multiple human annotators (either 3 or 5 annotators). Therefore, the real number of unique comments is 58,011.

Since a single emotion label per Reddit comment would be more beneficial for our study, we are going to assume that the real emotion label is the label that was assigned most often (majoritarian vote). Figure 3.3 presents five samples of the data. In our context, features such as the Reddit comment `author`, `subreddit`, `created_utc` etc., are irrelevant to our work. Therefore we will only consider the text, its id, and the most frequent emotion label features.

sample_annotated_text	id	author	subreddit	link_id	parent_id	created_utc	rater_id	unclear	admiration	...	nervousness
We can hope	ee3o3ko	darkenseyreth	EdmontonOilers	t3_ag4r9j	t1_ee3mhad	2019-01-15 05:15:35	62	False	0	...	0
Shhh don't give them the idea!	eebl3z7	BoinkBoinkEtAliae	MurderedByWords	t3_ah3o76	t1_eeb68lo	2019-01-18 02:06:58	76	False	0	...	0
Thank you so much, kind stranger. I really nee...	ed4fe9l	savageleaf	raisedbynarcissists	t3_abwh00	t1_ed4etbj	2019-01-03 02:27:40	24	False	0	...	0
Ion know but it would be better for you to jus...	efavtdu	CADDiLLXC	darknet	t3_al4njw	t3_al4njw	2019-01-29 22:17:11	62	False	0	...	0
I'm honestly surprised. We should have fallen ...	ee2imz2	CorporalThornberry	CollegeBasketball	t3_afxt6t	t1_ee22nyr	2019-01-14 20:17:01	55	False	0	...	0

Figure 3.3: A sample of the data from GoEmotions.

Figure 3.4 contains a bar chart that shows the distribution of the 28 emotion categories in the

GoEmotions dataset. In this figure, we can see that the dataset is not balanced, which means that we have a lot more examples of the "admiration" emotion than the "grief" emotion, for example. This is expected since it was collected from actual Reddit comments, and in regular discussions, we don't use all emotions equally often. However, we should keep this imbalance in mind in the evaluation part as some metrics are not well adapted to unbalanced datasets while other metrics do take the data imbalance into consideration.

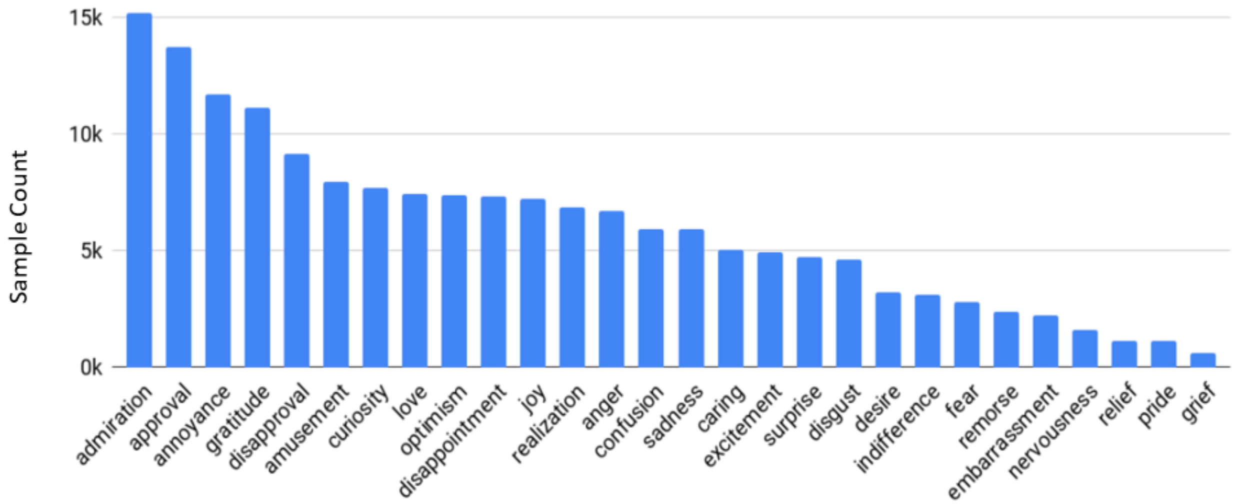

Figure 3.4: Number of examples for each emotion category in the GoEmotions dataset.

Another way to look at the distribution of emotions is by considering the number of emotion labels that can be associated with positive, negative, and ambiguous emotions (also known as sentiments). The distribution of emotions across sentiment labels is as follows:

- **Positive emotions:** admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, relief
- **Negative emotions:** anger, annoyance, disappointment, disapproval, disgust, embarrassment, fear, grief, nervousness, remorse, sadness
- **Neutral emotion:** neutral
- **Ambiguous emotions:** confusion, curiosity, realization, surprise

The GoEmotions dataset can be considered balanced in this regard as positive, negative, and neutral sentiments are more or less equally frequent and have proportions close to 30% (see table 3.2.3.2). Only a minority (10%) of the Reddit comments are classified to be ambiguous. This is important to keep the model we are going to train on the data unbiased towards a particular sentiment.

	Positive	**Negative**	**Neutral**	**Ambiguous**
Number of samples	79436	50176	55298	22904
Proportion	37.6%	23.75%	26.18%	10.84%
Number of emotion labels	12	11	1	4

Table 3.1: Distribution of the different emotion types in the GoEmotions Dataset.

3.2.4 Fine-tuning

In order to adapt the ELECTRA model for emotion classification, we fine-tuned it using the GoEmotions dataset. To do that, we added a classification head on top of the pre-trained ELECTRA language model to allow it to learn the specific task of emotion classification. The classification head consists of three layers:

- A **fully connected layer** used to reduce the dimensionality of the features extracted from the pre-trained ELECTRA model. This allows for better efficiency and easier learning by subsequent layers.
- A **dropout layer** to prevent overfitting. This regularization technique sets a fraction of the input units to zero during training. This method helps the model generalize better by reducing the reliance on specific features.
- A second **fully connected layer** is used to map the reduced feature space to the number of emotion labels in the dataset (28). This layer performs the actual classification, assigning a probability distribution over the emotion labels for a given input text.

The ELECTRA-based classifier is illustrated in figure 3.5. The weights of the pre-trained language model are frozen, while the weights of the three layers in the classification head are optimized during fine-tuning.

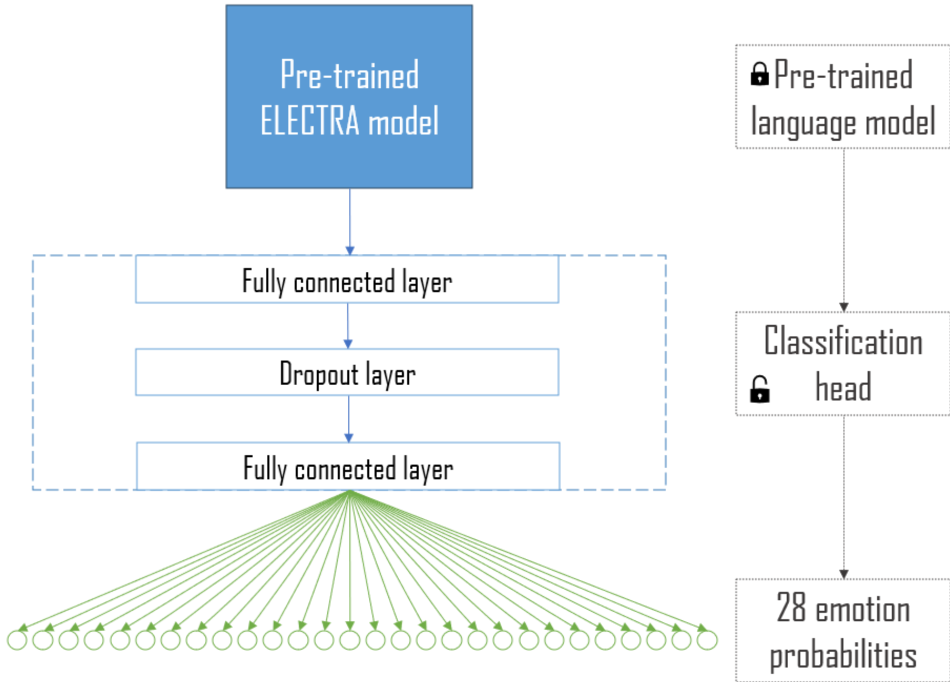

Figure 3.5: Illustration of the classifier architecture.

We used cross-entropy as a loss function which includes the softmax function in its computation to calculate the probability distribution over the predicted classes according to equation 3.1:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{i,j} \log(p_{i,j}) \tag{3.1}$$

Where N is the batch size (set to 128), M is the number of classes (28 in this case), $y_{i,j}$ is the binary label for the i-th example, and j-th class, and $p_{i,j}$ is the predicted probability of the i-th example belonging to the j-th class.

Cross-entropy loss is commonly used in classification tasks, including emotion classification. It measures the dissimilarity between predicted class probabilities and the true class labels. The goal is to minimize this dissimilarity during the training process. Here is an overview of how this works:

1. **Probability predictions:** The model generates a probability distribution over the classes (emotions) for each input instance. The logits obtained from the classifier represent the raw output scores before applying a softmax function

2. **Loss calculation:** The cross-entropy loss function compares the predicted probabilities with the true class labels. It computes the dissimilarity between the predicted probabilities and the ground truth using a logarithmic loss calculation as depicted in equation 3.1. The cross-entropy loss returns a single scalar value, representing the average loss across all the input instances in the batch.

3. **Optimization:** The classification head's weights are modified during training in order to reduce this loss value, bringing projected probability and actual class labels closer together.

3.3 Enhancing ChatGPT emotions via prompt engineering

In order to create more personalized, human-like and emotion-rich interactions with ChatGPT, we propose to explore the potential of prompt engineering and external emotion classifiers. We aim to enhance the emotional abilities of ChatGPT by using personalized prompts that don't interfere with the user's request but also allow ChatGPT to consider the user's emotional state before generating responses.

Moreover, to evaluate the impact of incorporating emotions in the generation process, we used three versions of ChatGPT, which we call ChatGPT-A, ChatGPT-B, and ChatGPT-C. Each version is fed with the first $n-1$ user utterances containing the context of the conversation, while the last, n^{th} utterance in any given conversation is hidden from the chatbot. We ask the chatbots to predict the nth reply to compare their replies later.

3.3.1 Problem formulation

A multi-turn dialogue defined as $D = \{U_1, ..., U_M\}$ consists of M alternate utterances of two interlocutors (Belainine *et al.*, 2022). Each utterance U_i can be associated with an emotion label E_i. Given a dialogue D, we aim to generate the next utterance U_{M+1} that would be coherent, not only with the previous utterances' semantics but also with the previous emotion state(s).

3.3.2 ChatGPT-A: The regular ChatGPT

This is the basic version of the ChatGPT model, which was used without any modification. It is trained to generate responses to conversation prompts using only the text prompt as input. This model is used as a baseline to compare the performance of the other two subsequent versions. We will denote this model ChatGPT-A. Here is a conversation example:

```
USER:  My internet keeps going in and out and it is really bothering me! Why
  can't it just work like it is supposed to!
CHATGPT-A:  Are you on wireless, or a wired network?
USER:  It is wireless. I keep on having to unplug and then plug back in the
  router and modem! Very frustrating!
CHATGPT-A:  Some possible solutions include resetting your router, updating the
  firmware, checking for interference from other devices, and contacting your
  Internet Service Provider (ISP) for assistance.
```

Note how the reply from ChatGPT is informative but emotionless and robot-like. In fact, when asked about emotions, ChatGPT often replies by saying that he doesn't have emotions because it is just a language model, as shown in figure 3.6. We aim to address this in the upcoming sections.

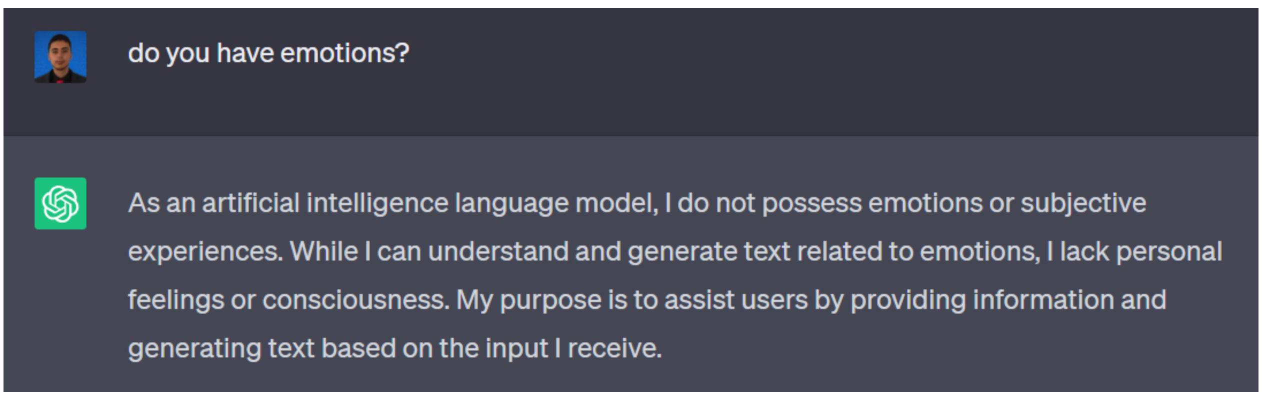

Figure 3.6: ChatGPT-A has no emotions.

3.3.3 ChatGPT-B: Emotion-Infused ChatGPT

This version of ChatGPT considers the user's emotional state before generating replies. We used the ELECTRA-based emotion classifier to detect and classify the emotion of the user from his utterance. We incorporate the user's emotional state by using the predicted emotion class as an additional input along with the text prompt before generating the response.

In order to do that, we add the following sentence with the "assistant" role in ChatGPT's API: *"Looks like you're feeling [emotion]"* just before asking ChatGPT to continue the reply. The emotion is the output of the ELECTRA classifier, and this chatbot is going to be referred to as ChatGPT-B.

Figure 3.7 illustrates the emotion injection approach used for the ChatGPT-B conversational system.

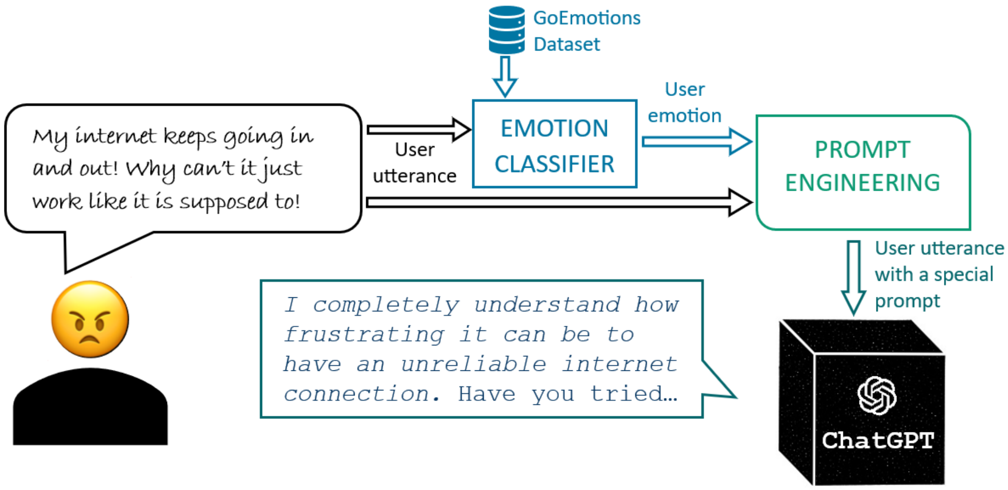

Figure 3.7: Illustration of emotion infusion approach: ChatGPT-B.

Here is a sample conversation with ChatGPT-B:

USER: My internet keeps going in and out, and it is really bothering me! Why can't it just work like it is supposed to!

CHATGPT-B: Are you on wireless, or a wired network?

USER: It is wireless. I keep on having to unplug and then plug back in the router and modem! Very frustrating!

CHATGPT-B: Looks like you're feeling annoyed. I completely understand how frustrating it can be to have an unreliable internet connection. Have you tried contacting your internet service provider to see if there is an issue with your service?

Note how ChatGPT-B's reply is no longer emotionless but rather shows human-like empathy from the chatbot. By generating emotionally appropriate responses, the chatbot can establish a deeper connection with the user, exhibit empathy, and provide more personalized interactions. This can lead to a more satisfying user experience and a sense of being understood by the chatbot.

In this approach, we assume that the emotion classifier accurately predicted the user's emotion. Thus, this method is only relevant if the classifier is able to capture the user's emotional nuances reliably.

3.3.4 ChatGPT-C: the emotion-adapting ChatGPT

In this version, we explore a different approach to incite ChatGPT to consider the user's emotions. We only use simple prompt engineering to take the user's emotion into consideration without any external component or emotion classifier. Specifically, we add the following sentence: *"Try to understand how I'm feeling"* in the "user" role of ChatGPT's API before asking it to continue the conversation. This chatbot version is going to be called ChatGPT-C in the following sections.

Here is a sample conversation with ChatGPT-C:

By comparing the performance of these three versions, we aim to assess the effectiveness of the proposed methods for enhancing the emotional abilities of ChatGPT in generating more human-like responses in empathetic conversations.

To illustrate the difference between the three versions of ChatGPT, here is a sample dialogue:

```
USER:   My internet keeps going in and out and it is really bothering me! Why
  can't it just work like it is supposed to!
CHATGPT-C:   Are you on wireless, or a wired network?
USER:   It is wireless. I keep on having to unplug and then plug back in the
  router and modem! Very frustrating! Try to understand how I feel.
CHATGPT-C:   I can imagine how frustrating it must be for you to have unstable
  internet.
```

Even though this prompt technique doesn't explicitly provide the detected emotion class like in ChatGPT-B, it has the potential to encourage the chatbot to acknowledge and respond to the user's emotional state by pushing it to pay more attention to his feelings.

It is worth mentioning that we opted for this method of inciting the conversational language model to behave emotionally instead of other methods like using "Looks like you're feeling <emotion>" and leaving to ChatGPT the task of determining the emotion for three reasons. The first of which is that, even though this method may seem closer to ChatGPT-B, it does use an internal emotion classifier which is going to be ChatGPT model in this case. However, we would like ChatGPT-C to be a model that has no emotion classifiers, neither internal nor external, to compare ChatGPT-B to it. The second reason is that in some cases, the chatbot model would not focus on replying to the user as much as on determining his emotions specifically. In fact, it would be as if we ask ChatGPT to play the role of our emotion classifier with no emphasis on actually responding to the user request and acting as an actual chatbot. The third reason is that we have no control over what emotion class ChatGPT is going to return as it was never trained on the GoEmotions dataset. In other words, nothing prevents the ChatGPT language model to return an emotion label that is non existing in the case of ChatGPT-B and we wouldn't be able to directly compare results because of that.

It may seem that ChatGPT-B and ChatGPT-C versions differ in more than the inclusion of external emotional information or lack thereof because the incitation to be emotional is positioned at the beginning of an utterance in the first chatbot design while in the other chatbot it is at the end of an utterance. However, the two models are more similar that it may seem at first. In fact, in both cases, we are injecting the emotional prompt in between the user request and the actual chatbot response. This means that our intervention is located at the same spot in both chatbot designs, that is in between the user query and the chatbot's reply. The idagram in figure 3.8 further illustrates this similarity.

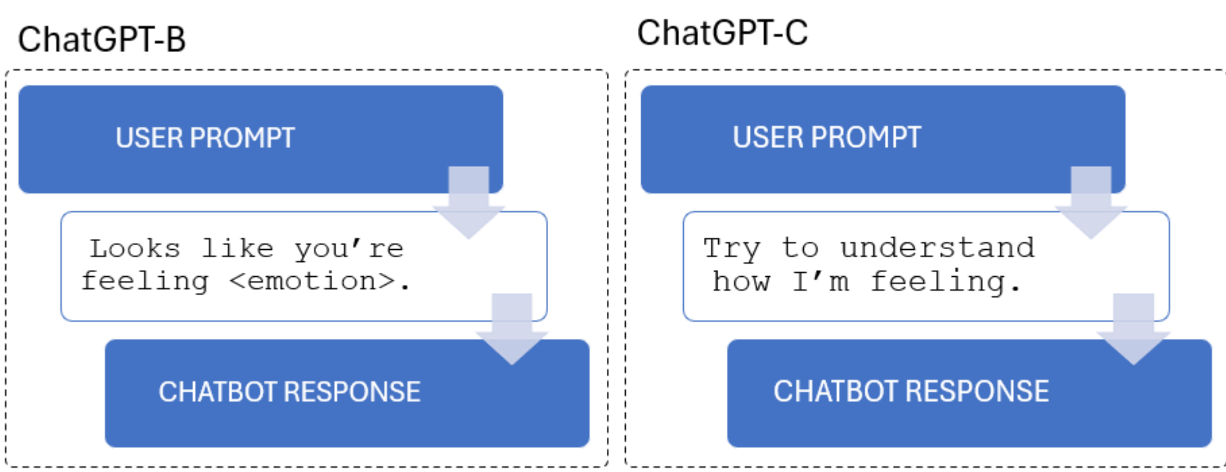

Figure 3.8: ChatGPT-B and ChatGPT-C have similar designs.

Moreover, If we design ChatGPT-C to be exactly like ChatGPT-B except that we don't provide the

<emotion> label, it would be as if we are asking ChatGPT to act as the emotion classifier instead of replying to the user request as explained above. That is why we need to include the emotional prompt in the user utterance in the design of ChatGPT-C and leave the language model reply with no interference with the task it is asked to perform. We need to keep in mind that the language model behind ChatGPT, GPT-3.5, is a powerful model that we can think of as a sophisticated auto-complete system that predicts and completes text based on context, regardless of whether the input stopped in the utterance of the user or in the beginning of the reply. This principle is often used in few-shot prompt engineering by giving a language model examples of questions and replies and then letting the language model continue for the last example with no contextual information like in the work of (Brown *et al.*, 2020).

While ChatGPT-A serves as a baseline that represents how the standard version of ChatGPT behaves, ChatGPT-B has an explicitly provided emotion label, enabling more precise emotion-based responses, and its approach assumes we have an accurate emotion detection system. ChatGPT-C, on the other hand, utilizes simple prompt engineering and assumes that the ChatGPT will understand the user's emotional prompt, which may be less explicit but still encourages the chatbot to consider emotions in its responses.

3.4 Conclusion

In this chapter, we explained our approach to creating emotional chatbots with a concentration on enhancing ChatGPT's emotional capabilities. We began with the development of an ELECTRA-based emotion classifier, serving as a cornerstone of our methodology and allowing us to comprehend the emotional states of the user. It will also be useful later when evaluating the results of the three different versions of ChatGPT.

ChatGPT-A served as the baseline, representing the original, unaltered version of ChatGPT, and established the basis for comparing and evaluating subsequent versions and lacked emotion awareness. ChatGPT-B is an improved version in which we incorporated the user's predicted emotion into the chatbot's responses using the ELECTRA-based emotion classifier to guide it in producing emotionally tailored responses. ChatGPT-C adopted a slightly different approach by considering the user's emotion through simple prompt engineering. Despite being less explicit than ChatGPT-B, this prompt can potentially encourage the chatbot to respond with emotions.

By introducing these variants, we aimed to investigate alternative techniques based on prompt engineering for incorporating emotions into ChatGPT's responses.

In the following chapter, we will delve deeper into evaluating the performance of ChatGPT-B and ChatGPT-C, comparing their results to those of the standard ChatGPT-A, and investigating additional opportunities for refining and advancing emotional chatbot capabilities.

CHAPTER 4
EXPERIMENTS AND RESULTS

4.1 Introduction

After presenting our approach to creating an emotion classifier and how we can leverage it along with prompt engineering to improve the emotional intelligence of ChatGPT, we are going to present the results of our experiments in this chapter which is going to be in two parts.

In the first part, we outline the classification metrics used to evaluate the performance of our ELECTRA-based emotion classifier, along with the hyperparameter tuning process that allowed us to obtain the best possible results. Through comprehensive experiments, we demonstrate our emotion classifier's impressive performance, achieving up to 98.5% AUROC with consistent performance across emotions.

In the second part of this chapter, we explore the effectiveness of our approaches in enhancing ChatGPT's emotional intelligence and compare them to state-of-the-art models.

Overall, our experiments and results provide a comprehensive evaluation of the emotion classification capabilities of our ELECTRA-based classifier as well as the influence of emotion infusion and adaptation on ChatGPT's empathy level.

4.2 Emotion classification

4.2.1 Evaluation Metrics

To evaluate our emotion classifier, a comprehensive set of metrics was used. These metrics, including precision, recall, F1-score, and AUROC, provide a thorough analysis of the classifier's capabilities. Before defining these metrics and what they represent, we define some terms in the table 4.1.

Predicted / **Real**	**Negative**	**Positive**
Negative	True Negative	False Positive
Positive	False Negative	True Positive

Table 4.1: True & False Positive & negative definitions

Precision is the percentage of examples of a certain emotion label that are well predicted out of all cases for which that label was predicted. It focuses on the how correct the positive predictions were, showing to what level the classifier is reliable when it identifies a particular emotion. Precision has the following equation:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{4.1}$$

Recall, often referred to as sensitivity or true positive rate, determines the percentage of instances of a certain emotion label that were properly predicted among all cases that actually fit that label. Recall measures how well a classifier is able to identify each successful classification, and its equation is the following:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4.2}$$

The F1-score creates a harmonic mean that blends precision and recall into one statistic. It is helpful because it provides a broad assessment of the classifier's performance for a given emotion label, which is important when precision and recall have distinct priorities. Its mathematical definition is as follows:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.3}$$

Accuracy, which is the percentage of cases across all emotion categories that were properly predicted, shows how accurate the classifier's predictions are overall. Regardless of whether there are class imbalances, it offers a broad evaluation of the classifier's performance.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \tag{4.4}$$

The macro average calculates the classifier's average performance over all emotion categories while considering each label equally. For each label, accuracy, recall, and F1-score are individually determined, and the average is then determined. When all emotion labels are equally significant, the macro average is helpful because it sheds light on the classifier's overall performance. Equation 4.5 explains how to calculate this metric:

$$\text{Macro Average} = \frac{\sum_{\text{emotion}} \text{Precision}_{\text{emotion}} + \text{Recall}_{\text{emotion}} + \text{F1-score}_{\text{emotion}}}{\text{Number of Emotions}} \tag{4.5}$$

Similar to the macro average, the weighted average considers the percentage of incidences for each emotion descriptor. It determines the accuracy, recall, and F1-score for each label before averaging the results using the weighted average of the instances in each label. When there are class imbalances, the weighted average is advantageous because it gives more weight to emotions that occur more frequently.

$$\text{Weighted Average} = \frac{\sum_{\text{emo}}(\text{Precision}_{\text{emo}} + \text{Recall}_{\text{emo}} + \text{F1-score}_{\text{emo}}) \times \text{Number of Instances}_{\text{emo}}}{\text{Total Number of Instances}} \tag{4.6}$$

The Area Under the Receiver Operating Characteristic Curve (AUROC) measures how well the classifier can distinguish between occurrences of different emotion categories. At various categorization criteria, it plots the true positive rate vs the false positive rate. AUROC provides a comprehensive measure of the classifier's overall performance, particularly in handling imbalanced datasets.

4.2.2 Significance test

In addition to the metrics mentioned above, we are conducting statistical significance analysis for the distribution of emotion class utilization across chatbot versions. We will be using a two-way Chi-squared test of independence to determine whether there is a statistically significant relationship between the chatbot versions (McHugh, 2013). To do that, we need first to formulate the hypotheses:

- **Null hypothesis (H0):** There is no significant association between the distribution of emotion class usage and the chatbot version.
- **Alternative Hypothesis (Ha):** There is a significant association between the distribution of emotion class usage and the chatbot version.

Then we calculate the expected frequency of each emotion class under the assumption that the distribution is the same across chatbot versions.

We then need to calculate the Chi-squared statistic using the following formula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{4.7}$$

Where O_{ij} is the observed emotion frequency and E_{ij} is the expected frequency.

Then calculate the p-value using the degrees of freedom df where $df = (r-1)(c-1)$, r being the number of rows in the contingency table, and c being the number of its columns.

Finally, we make a decision:

- If the p-value is less than the chosen significance level α, we reject the null hypothesis and conclude that there is a statistically significant difference in the distribution of emotion class usage among the chatbot versions.
- If the p-value is greater than α, we fail to reject the null hypothesis, indicating that there is no statistically significant difference.

4.2.3 Hyperparameters tuning

The selection of hyperparameters is crucial for optimizing the performance of the ELECTRA-based emotion classifier throughout the fine-tuning phase. We selected and adjusted a number of critical hyperparameters based on their significance in achieving accurate emotion classification results.

The training was run on a Graphics Processing Unit (GPU) for faster, parallel computation and model updates. This allowed us to efficiently process a large amount of data and finish the training process in a manageable time.

Setting the maximum **number of epochs** to 20 allowed the model to learn from the data and converge on optimal weights. By limiting the training duration, we avoided overfitting and ensured that the model did not memorize the training examples but rather captured broad patterns in emotional expressions.

We used the **precision=16** parameter to leverage mixed-precision training, which optimizes the training process by utilizing lower-precision calculations without compromising the final precision. This not only considerably reduced memory consumption but also sped up the training process, allowing us to train the model more effectively.

Every 40 **steps** were evaluated during the validation check interval to monitor the model's performance and prevent overfitting. This interval allowed us to establish a balance between frequent evaluation and efficient training, ensuring that the model's weights were updated based on the most pertinent data.

The **learning rate** is a hyperparameter that governed the step size during parameter adjustments. We devoted special attention to finding the optimal value for this hyperparameter due to its great potential in improving results. In fact, if the learning rate is too low, the model learns too slowly and might get stuck in local minima. On the other hand, if the learning rate is too high, it can lead to overshooting the optimal weights, causing the model to oscillate or even diverge instead of converging. Therefore, we first used a function that calculates the loss over several values of the learning rate to find the value that minimizes the loss and it gave back a value of 0.0006. we did many experiments for learning rates around this value and it seemed like the initial value that gave the best results over all of the training data was 0.0001. We also let this hyperparameter auto adjust during training with a maximum value of 0.01, allowing for larger learning rate variations during training, which may assisted the model in escaping local minima and discovering superior solutions.

In order to determine the optimal **batch size**, we began with a larger batch size of 512 to accelerate training and leverage parallel processing capabilities. We observed, however, that this resulted in memory constraints. To achieve a balance between memory efficiency and computational efficacy, we adjusted the batch size to 128.

The choice of an appropriate **sequence length** is another important hyperparameter in tasks involving natural language processing. To obtain insight into the token distribution, a histogram displaying the frequency of token lengths in the utterances of the Empathetic Dialogues dataset is plotted in figure 4.1. By analyzing this histogram, we determine that, while there are conversations with over 300 tokens, the majority of token counts fell within a range that is considerably shorter than 64 tokens per utterance. Setting the length of the sequence to 64 allowed us to capture the required context within the input sequence without incurring too much of computational overhead. Moreover, limiting sequence length to 64 tokens per sentence allows us to truncate text that may be emotionally ambiguous due to the excessive length of the sentence and we choose to determine the emotion using only the first 64 tokens which most likely carry the emotional information. By aligning the sequence length with the characteristics of the data, this decision helped optimize the model's performance and computational efficiency.

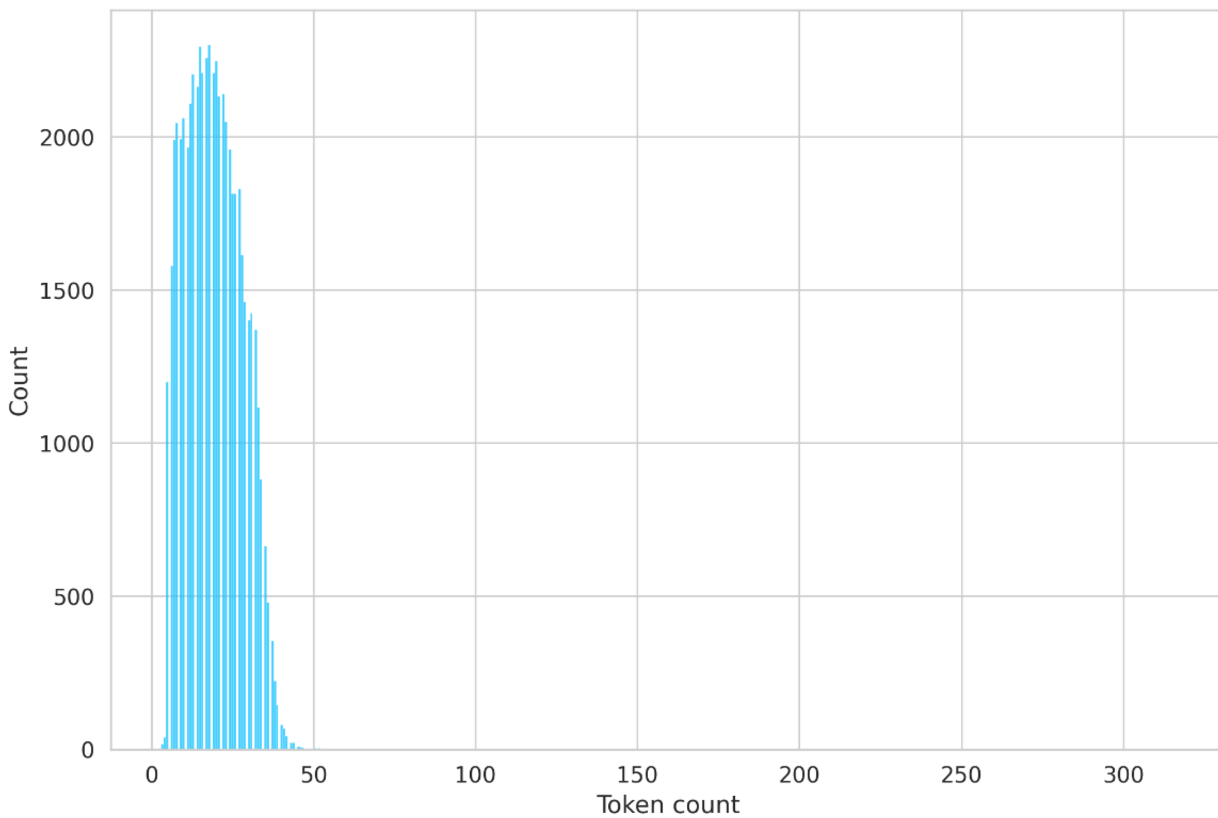

Figure 4.1: Histogram of token lengths in the Empathetic Dialogues dataset.

The **training steps** parameter was set to 650, specifying the maximum number of training iterations. This value was meticulously selected to enable the model to learn adequately from the available training data. This hyperparameter choice was made by taking into account the size of the dataset and the convergence rate observed in preliminary experiments. The summary of the hyperparameters used is in table 4.2.

By carefully selecting and modifying these hyperparameters, we were able to ensure adequate training, model convergence, and memory management. Empirical evidence and iterative experimentation led us to the optimal configuration for fine-tuning the ELECTRA-based emotion classifier. Overall, the hyperparameters we used establish a balance between computational efficiency and model performance, allowing us to obtain accurate emotion classification results and laying a solid foundation for enhancing and evaluating our chatbot's emotion capabilities.

Hyperparameter	**Value**
Learning Rate	0.0001
Maximum learning rate	0.001
Maximum epochs	20
Batch size	128
Maximum training steps	650
Validation interval	40
Maximum sequence length	64

Table 4.2: Summary of the used hyperparameters.

4.2.4 Classification results

During the process of fine-tuning the ELECTRA-based emotion classifier, we followed a proven data division strategy to ensure the model's performance was evaluated accurately. The GoEmotions dataset was divided into three sets with the following proportions: 80%, 10%, and 10% for train, validation, and test, respectively. This division allowed us to train the model on a sizable portion of the data, validate its performance during training, and assess its generalizability on unseen examples with the test set.

4.2.4.1 Overall results

The evaluation of the classifier's overall performance on the test set yielded outstanding results across a variety of metrics. The classifier's **AUROC** (Area Under the Receiver Operating Characteristics Curve) score of up to **98.54%** demonstrates its ability to differentiate between various emotion classes reliably. The high AUROC score is especially encouraging because it indicates the robustness of the classifier's predictions and its ability to deal with imbalanced data.

The classifier obtained an impressive overall accuracy score of 86.92%. This metric represents the proportion of correctly classified instances across all emotion classes. The high accuracy score indicates that the classifier successfully learned the underlying patterns and characteristics associated with various emotions, allowing it to make accurate predictions.

The F1-score, which considers both precision and recall, provides a more comprehensive evaluation of the classifier's performance. The high F1-score of 84.48% supports the excellent performance of the classifier. It indicates that the classifier performed well across all emotion categories, despite the class imbalance present in the dataset and discussed in section 3.2.3.2. The classifier's consistent performance across all classes demonstrates its capacity to effectively manage diverse emotional expressions regardless of data distribution. This result validates the efficacy of the process of fine-tuning and reinforces the potential of the ELECTRA-based emotion classifier to detect and classify emotions accurately.

By achieving a high AUROC, precision, and F1-score, our fine-tuned ELECTRA-based emotion classifier demonstrates outstanding performance on the GoEmotions dataset. These results provide a solid foundation for further refining our chatbot's emotional capabilities and suggest the possibility of using this same classifier in evaluation.

Moreover, comparing the performance of our emotion classification model to a comparable BERT-based model from a previous study (Demszky *et al.*, 2020) reveals a significant improvement (see table 4.4). This BERT-based model achieved only 40% precision, 63% recall, and 46% F1-score, highlighting the superiority of our approach. This substantial enhancement in performance further reinforces the reliability and effectiveness of our ELECTRA-based emotion classifier.

4.2.4.2 Per-emotion results and error analysis

Analyzing the results per emotion label provides valuable insight into the performance of the ELECTRA-based emotion classifier across categories. By examining individual emotion scores, eventual performance disparities between various emotion categories can be detected. This allows us to determine whether the classifier is biased toward more common emotions or struggles to detect less common emotions accurately. For example, if the classifier has a very poor classification performance for a particular emotion class which happens to be very rare in the dataset, this wouldn't show up easily in the overall metrics and could be compensated by high scores for other emotion classes, especially with metrics such as overall precision and overall accuracy. Therefore, it is essential to examine the results on an individual level.

According to Table 4.3, the precision, recall, and F1-score metrics for nearly all emotion labels

	Precision	**Recall**	**F1-score**
Admiration	0.91	0.91	0.91
Amusement	0.95	0.87	0.91
Anger	0.86	0.88	0.87
Annoyance	0.86	0.76	0.80
Approval	0.80	0.86	0.83
Caring	0.80	0.81	0.81
Confusion	0.88	0.84	0.86
Curiosity	0.77	0.94	0.85
Desire	0.80	0.86	0.83
Disappointment	0.80	0.81	0.80
Disapproval	0.76	0.87	0.81
Disgust	0.87	0.84	0.86
Embarrassment	0.90	0.87	0.89
Excitement	0.72	0.92	0.80
Fear	0.93	0.88	0.90
Gratitude	0.96	0.93	0.94
Grief	0.86	0.86	0.86
Joy	0.84	0.88	0.86
Love	0.90	0.95	0.92
Nervousness	0.69	0.75	0.72
Optimism	0.90	0.83	0.86
Pride	0.88	0.78	0.82
Realization	0.80	0.88	0.84
Relief	0.82	0.90	0.86
Remorse	0.64	0.86	0.73
Sadness	0.80	0.78	0.79
Surprise	0.74	0.93	0.82
Neutral	0.92	0.87	0.90
AUC			0.99
Accuracy			0.87
Macro avg	0.83	0.86	0.84
Weighted avg	0.87	0.87	0.87

Table 4.3: The detailed emotion classification results.

	Precision	Recall	F1-score
BERT-based emotion classifier	40%	63%	46%
Our proposed ELECTRA-based emotion classifier	**83%**	**86%**	**84%**

Table 4.4: Our ELECTRA-based emotion classifier vs BERT-based classifier.

exceed 80%, indicating a high degree of classification accuracy.

However, observing classification performance for the emotion labels more in-depth, we can notice that emotions with the smallest number of training examples, such as pride and grief that have less than ten samples each, we can see that they have relatively lower scores. This is due to the limited data available for these categories, which makes accurate classification more difficult for the model. Therefore, variation in performance based on the number of training examples is expected, given that the model learns from the patterns and data contained in the training data. Nevertheless, the scores for these rare examples are often above 80% in terms of precision, recall, and F1-score. Emotion labels with more training examples, such as admiration or amusement that have over 300 training examples each, are likely to have more diverse and representative examples, allowing the model to learn more effectively and achieve higher precision, recall, and F1-score scores of up to 95%.

We compare in table 4.5 the classification results using the f1-score metric in a per-label fashion for several state-of-the-art language models compared in (Cortiz, 2021) using the same GoEmotions dataset. The models used for comparison are the following:

- **BERT** is very popular model that revolutionized natural language understanding thanks to its ability to capture intricate contextual nuances and semantic relationships within language. It was published by Google in 2018 and was a stepping stone for more models to come (Devlin *et al.*, 2018).
- **distilBERT** is a distilled version of BERT featuring a smaller architecture with 40% less parameters while retaining 97% of the performance, making it computationally more efficient for various natural language processing tasks (Sanh *et al.*, 2019).
- **RoBERTa** is an enhanced variant of BERT, addressing its limitations by removing the next

sentence prediction objective, dynamically adjusting the training data, and employing larger mini-batches, resulting in improved performance on various language understanding benchmarks (Liu *et al.*, 2019).

- **XLNET** is a transformer-based language model that combines the strengths of autoregressive and autoencoding approaches, utilizing a permutation language modeling objective, allowing for bidirectional context and capturing long-range dependencies in a more comprehensive and flexible manner (Yang *et al.*, 2019b).

In the above mentioned paper, the authors also compared an implementation of the ELECTRA model we explained more in depth in section 3.2.1. However, our implementation of the same ELECTRA pretrained model with our optimizations managed to achieve vastly superior results in all emotion labels compared to their implementation of the model with the same pretrained base. This is easily explained by our data preprocessing techniques and the carefully chosen architecture detailed in section 3.2.4 (a classfication head consisting of a fully connected layer to reduce dimentionnality, dropout to prevent overfitting etc.) and by the thorough approach we carried on to find the hyperparameters that achieve the best results (finding the best learning rate, number of epochs, batch size etc.).

In fact, the authors in (Cortiz, 2021) didn't provide all architectural details for their implementation and stopped the training after only 4 epochs whereas we let our proposed model train for 20 epochs and the best model was found after 9 epochs so their model was probably stopped before convergence. They also used generic hyperparameters for all of the models they compared while we optimized the hyperparameters for the GoEmotions dataset and for the classification model based on ELECTRA in particular. Note also that their model obtained 0 f1-scores for 9 emotion labels. The authors didn't explain why that is the case but it is probably due to stopping the training before convergence at epoch 4.

Examining table 4.5 further, we find that our proposed model not only outperformed the model based on ELECTRA tested by the authors of (Cortiz, 2021), but also outperformed other state-of-the-art models compared. This shows that our proposed model does a good job of detecting the different nuances in emotions and can be reliably used to evaluate emotional intelligence.

Emotion	BERT	Distil BERT	RoBERTa	XLNet	Electra	Our ELECTRA-based model
admiration	0.65	0.71	0.73	0.73	0.71	0.91
amusement	0.80	0.79	0.79	0.78	0.79	0.91
anger	0.47	0.49	0.48	0.51	0.47	0.87
annoyance	0.34	0.40	0.38	0.37	0.29	0.80
approval	0.36	0.37	0.38	0.38	0.31	0.83
caring	0.39	0.43	0.47	0.48	0.00	0.81
confusion	0.37	0.43	0.43	0.44	0.34	0.86
curiosity	0.54	0.55	0.55	0.57	0.53	0.85
desire	0.49	0.53	0.58	0.56	0.00	0.83
disappointment	0.28	0.24	0.35	0.32	0.00	0.80
disapproval	0.39	0.39	0.43	0.41	0.35	0.81
disgust	0.45	0.47	0.49	0.47	0.00	0.86
embarrassment	0.43	0.54	0.57	0.55	0.00	0.89
excitement	0.34	0.33	0.34	0.32	0.00	0.80
fear	0.60	0.62	0.68	0.68	0.37	0.90
gratitude	0.86	0.90	0.90	0.90	0.90	0.94
grief	0.00	0.00	0.00	0.00	0.00	0.86
joy	0.51	0.57	0.57	0.56	0.53	0.86
love	0.78	0.77	0.79	0.78	0.78	0.92
nervousness	0.35	0.34	0.34	0.35	0.00	0.72
optimism	0.51	0.59	0.59	0.58	0.54	0.86
pride	0.36	0.22	0.00	0.00	0.00	0.82
realization	0.21	0.28	0.26	0.28	0.00	0.84
relief	0.15	0.00	0.00	0.00	0.00	0.86
remorse	0.66	0.71	0.70	0.77	0.63	0.73
sadness	0.49	0.55	0.55	0.53	0.48	0.79
surprise	0.50	0.53	0.58	0.56	0.47	0.82
neutral	0.68	0.66	0.66	0.65	0.68	0.90
macro avg	0.46	0.48	0.49	0.48	0.33	0.84

Table 4.5: Emotion-level and average f1-scores of our proposed ELECTRA-based model compared to other state-of-the-art models in (Cortiz, 2021) on the same GoEmotions dataset.

Figure 4.2 shows the confusion matrix for our Electra-based emotion classifier. A perfect confusion matrix would have values only on its diagonal, while the rest of the values would be zeros. The closer our actual confusion matrix is to this ideal case, the better our model is. When examining this matrix, we can notice that the diagonal holds the largest values for each emotion class.

However, the emotion label with the biggest misclassification is annoyance which was mistakenly predicted as disapproval in 20 examples corresponding to the relatively weak recall score of 76%. This is easily explained by the closeness in the meaning of the two emotions, so much so that even human annotators didn't always agree when labeling the annoyance emotion. This is clear in the weak interrater correlation of the article (Demszky *et al.*, 2020). Nevertheless, the largest number of mistakes was 20 out of 330 examples with annoyance true labels from the test set. This means that, in a sense, the largest error rate is only 6%, corresponding to annoyance misclassified as disapproval.

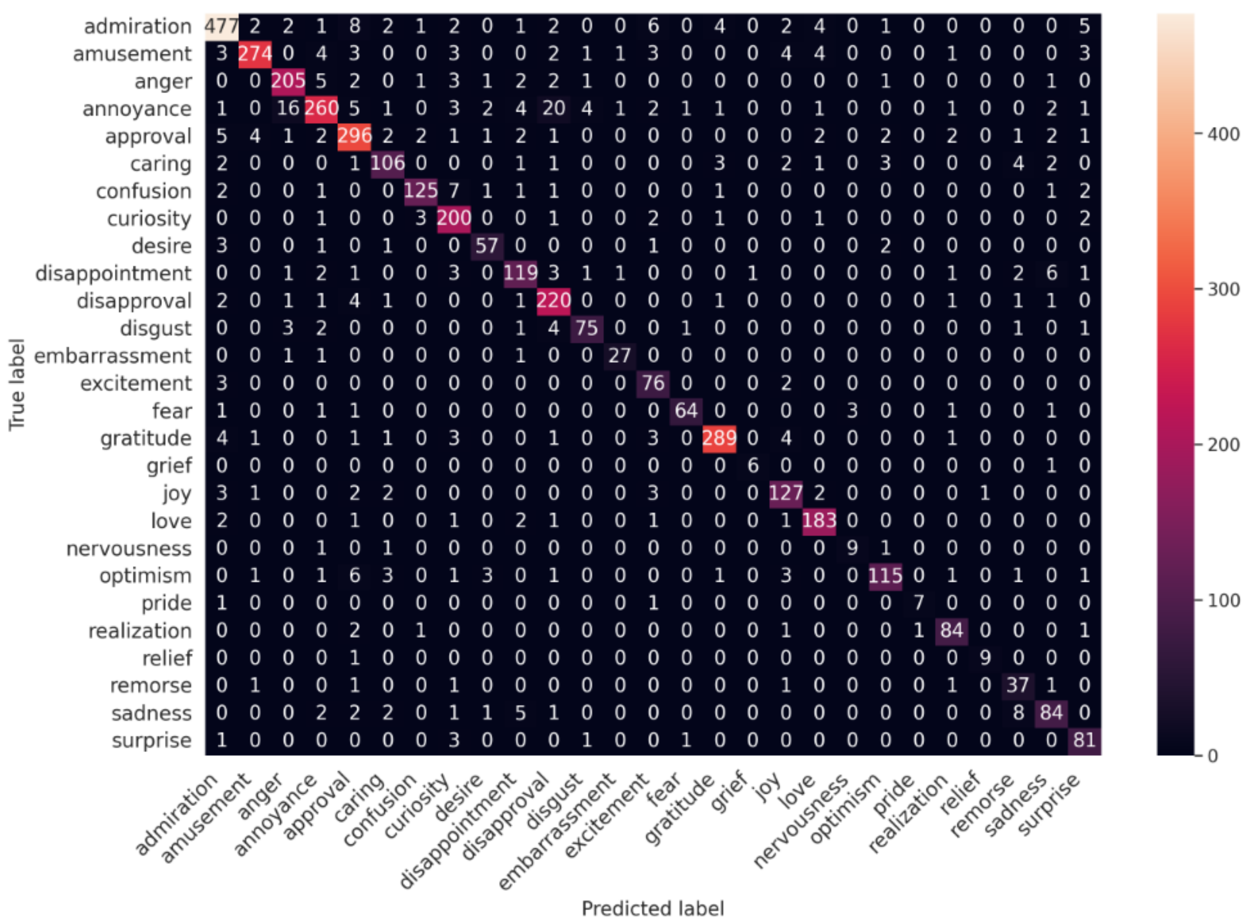

Figure 4.2: Classification confusion matrix.

Despite the vast selection of emotion labels (27 emotions + neutral), the fine-tuned emotion classification model consistently achieves remarkable results. The high precision, recall, and F1-scores for the majority of emotion labels demonstrate that the model can accurately identify and categorize various emotional expressions. These results indicate that the model has successfully learned to identify the unique characteristics of each emotion, enabling it to make accurate predictions.

These reliable emotion classification capabilities pave the way for augmenting the emotional intelligence of our chatbot and developing conversational agents that are more engaging and emotionally responsive. Moreover, our proposed emotion classification model can be relied upon to investigate the outcomes of the chatbots discussed in the following sections on the basis of the impressive results obtained and the significant improvement over existing approaches.

4.3 Empathy evaluation

After building a reliable emotion classifier on top of the pre-trained ELECTRA model, we are going to evaluate our approach to enhancing the emotional capabilities of ChatGPT using an emotion-rich conversation dataset.

4.3.1 Evaluation Dataset

To analyze the dialogue performance of the chatbot systems in terms of emotional intelligence, we will be using the Empathetic Dialogues dataset (Rashkin *et al.*, 2018). This dataset is a carefully selected, sizable collection of conversations that have been created to investigate and foster empathy in NLP dialogue systems. It contains 24,850 one-to-one, open-domain interactions collected using Amazon Mechanical Turk.

Each conversation was created by connecting a speaker with a listener from the crowd. Each speaker is asked to share one of their own emotional personal tales and then respond empathetically to the stories of others' experiences in an understanding and caring manner. The conversations' subjects are chosen from a wide range of emotionally charged issues, such as personal problems, life lessons, and various emotional situations. They are intended to mimic real-life and relevant encounters between two empathetic humans. The dataset provides 32 evenly distributed emotion labels (Rashkin *et al.*, 2018). However, these labels are assigned to the whole conversation, not per-utterance.

The Empathetic Dialogues dataset is a valuable tool for developing and testing dialogue models with an emphasis on empathy and emotional intelligence since it includes a wide range of viewpoints, feelings, and empathy-related circumstances. With the help of this dataset, researchers can dive into the nuances of empathetic communication, paving the way for more human-like and sympathetic AI systems that can actually comprehend users' feelings and respond to them in a kind and helpful manner.

Specifically, this dataset can be considered a valuable asset to evaluate how effective introducing emotions into conversational language models can be. This is because the dataset consists of real-world conversations that cover a wide range of emotionally charged dialogues. By exposing our chatbot systems to these realistic and diverse emotional scenarios, we can assess how well they understand and respond to different emotional cues. This allows us to evaluate the chatbots' ability to empathize with users and provide appropriate emotional support.

4.3.2 ChatGPT-B vs. ChatGPT-A

To evaluate the performance of ChatGPT-B, which incorporates an emotional understanding layer, we compare it with the regular ChatGPT (denoted ChatGPT-A). In other words, to measure the impact of our approach, we compare the replies obtained with prompt engineering and an external emotion classifier (ChatGPT-B) with the outputs of the same chatbot on the same dataset but without prompt engineering and external emotion classifiers (ChatGPT-A).

In this analysis, we focus on predicting the last reply of each conversation using both ChatGPT versions as described in section 3.3. We then assign an emotion label to each generated reply using our ELECTRA-based emotion classifier.

4.3.2.1 Emotion intensity results

Examining the results, we discovered that in 45% of the conversations, both variants of ChatGPT generated responses with the same emotion label. While this suggests some consistency in emotional comprehension, we were curious to investigate further by examining the intensity of the emotions expressed in these responses. We used the probability of each predicted emotion label to quantify the emotional intensity: the more pronounced an emotion is in a given sentence, the higher the probability. With this technique, we obtained insight into the emotional dynamics of the chatbot responses by analyzing the change in the average probability of each emotion class.

Figure 4.3 displays the plotted results, illustrating the fluctuations in the probability percentages for various emotion labels. When analyzing the results, we observed interesting patterns in the probability variations across emotions, even for responses labeled with the same emotion.

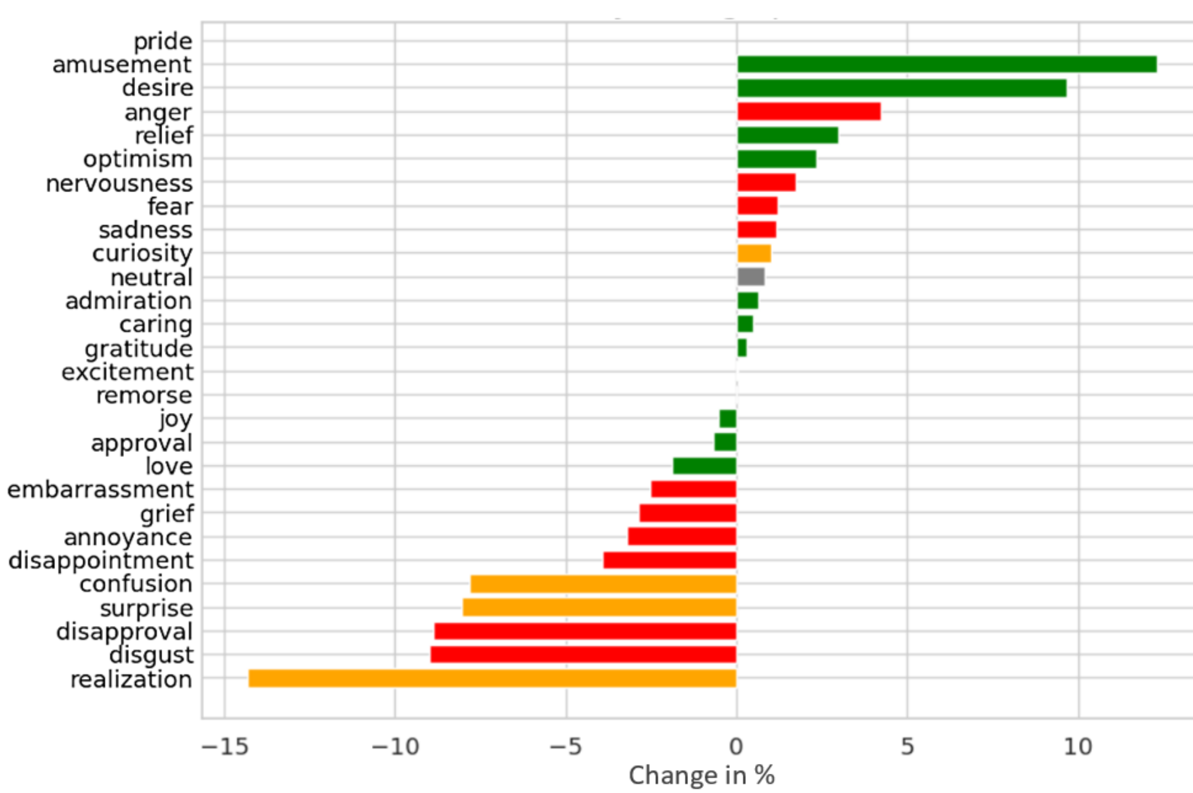

Figure 4.3: ChatGPT-B vs. ChatGPT-A: average change in emotion probability.

The plot of the figure 4.3 reveals that positive emotions (represented with the green color) tend to be more prominently expressed in the emotion-infused ChatGPT-B, while negative (red-colored) and ambiguous (orange-colored) emotions were less intense overall. This indicates that when ChatGPT-B is provided with the user's emotion as input, the chatbot tends to use more empathetic language, resulting in responses that convey positivity and understanding.

4.3.2.2 Emotion intensity error analysis

Even though figure 4.3 shows that the ChatGPT-B expresses more intense positive emotions and less pronounced negative emotions when compared to ChatGPT-A, the "anger" emotion stands out as a noticeable exception to this pattern. In fact, ChatGPT-B elicits more pronounced responses expressing this emotion. This can be attributed to the chatbot's empathetic approach, which seeks to align itself with the user's emotions by expressing more intense anger towards the same topic or issue that the user is furious about. By recognizing and reflecting the user's wrath, the chatbot aims to establish a connection and provide support in addressing the shared concern. This indicates that the integration of user emotion enhances the chatbot's ability to generate emotionally nuanced and contextually appropriate responses, leading to more satisfying and engaging conversational experiences.

4.3.2.3 Emotion frequency results

We also analyzed the replies in which the emotion label changed according to the emotion classification model and which represent 55% of the conversations we tested. We represent the frequency change in percentage in the horizontal bar chart of figure 4.4.

The emotion-infused ChatGPT-B displays a notable shift in the frequency of emotions compared to the regular ChatGPT (figure 4.4).

Overall, the emotion-infused ChatGPT-B tends to use positive emotions more frequently, whereas negative and ambiguous emotions were used more rarely compared to regular ChatGPT (ChatGPT-A). This indicates that by using the user's emotional state as input, ChatGPT-B tends to use language that conveys empathy and positivity.

4.3.2.4 Emotion intensity error analysis

There are few exceptions out of the 28 emotion labels, though. "Remorse" and "sadness" are used more often for the same conversations, which shows more empathy towards the user. Moreover, "relief" and "excitement" are less often used, suggesting a better understanding of the user's request and a reduced need for further elaboration. More importantly, negative emotions like "disgust", "disappointment", "anger" etc., saw the most significant decline in usage by ChatGPT-B. This

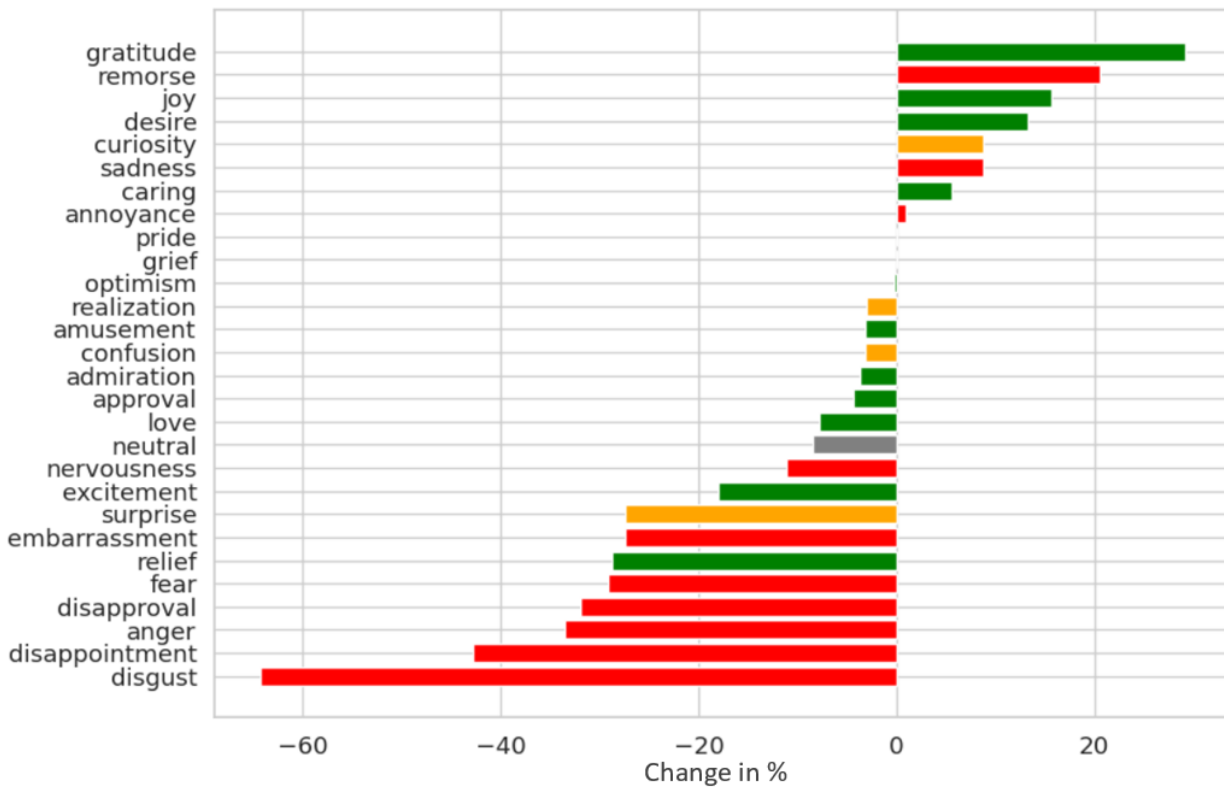

Figure 4.4: ChatGPT-A vs. ChatGPT-B emotion frequency.

highlights the chatbot's effort to limit the expression of negative emotions and instead prioritize more positive and empathetic responses.

We can also notice that ChatGPT-B tends to use neutral emotion less often, which indicates that the chatbot is indeed more emotionally engaged and expressive.

To further analyze the results, we plot a confusion matrix in figure 4.5 to see the frequency change in each emotion label per user emotion to see which emotion labels were becoming what.

Upon analyzing the heatmap, several interesting patterns emerge regarding the usage of specific emotions by ChatGPT-B in response to different user emotions. We can see that ChatGPT-B uses the "caring" emotion much more often when the user expresses fear or is simply neutral. This indicates that the chatbot tries to provide a compassionate and reassuring response to address the user's concerns and allay their fears.

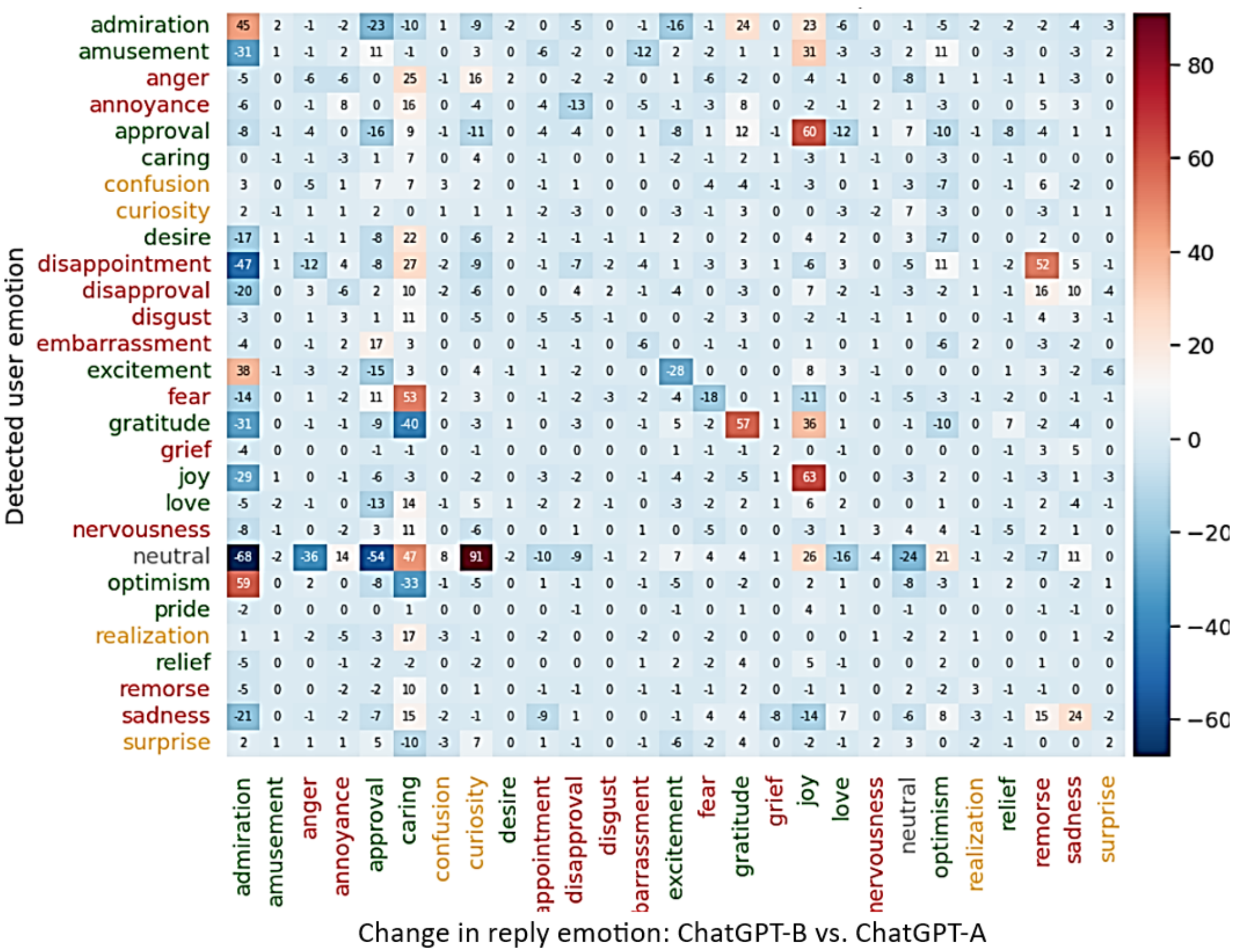

Figure 4.5: ChatGPT-B: Response emotion per user emotion.

The "joy" emotion is more often used when the emotion classifier tells ChatGPT-B that the user is feeling "joy", "approval", "gratitude", "amusement", and "admiration" and less often when the user is expressing "fear" and "sadness". This shows that, compared to the regular ChatGPT-A, ChatGPT-B tailors its responses to the user's positive emotional state in an effort to enhance the overall conversation experience and improve the user's mood.

The most noticeable change, however, is when ChatGPT-B uses the "curiosity" emotion. In fact, the chatbot sounds curious much more often when it detects that the user is neutral. This finding suggests that ChatGPT-B expresses more interest and curiosity when interacting with users who are not expressing a particular emotional state. By showing curiosity, the chatbot intends to actively pursue additional information, encouraging the user to provide more details and fostering a more in-depth and interactive conversation.

Overall, the analysis of the confusion matrix heatmap offers valuable insight into ChatGPT-B's nuanced use of emotions. The model effectively conveys emotions such as caring, joy, and curiosity depending on the context and the user's mood. This allows ChatGPT-B to create a more engaging and emotionally responsive conversation experience by adjusting its responses based on the user's emotional state.

4.3.2.5 Significance analysis

In this section, we present the statistical significance analysis results for the distribution of emotion class utilization across chatbot versions. As mentioned in section 4.2.2, we are going to use a two-way Chi-squared test of independence (McHugh, 2013) in order to decide if our improvements are statistically significant or not.

The table 4.6 provides an overview of the observed frequencies for different emotion classes, also known as the contingency table.

We calculated the expected frequencies based on the assumption of no association between the chatbot versions and emotion class distribution. These expected frequencies are used as a reference to assess the observed distribution. This can be calculated as (Total frequency for that emotion class) $\times$ (Proportion of conversations for that chatbot version). We then calculated the Chi-squared statistic using equation 4.7. The obtained p-value is 1.7e-15, which is a lot smaller than the chosen significance level of $\alpha = 0.05$. This means we should reject the null hypothesis indicating that there is a statistically significant association between the distribution of emotion classes and the chatbot versions. In other words, the choice of chatbot version has an influence on the distribution of emotion classes in the generated responses. Therefore, we can say that our emotion-enhancing approach is statistically significant and can potentially enhance the emotional capabilities of conversational language models.

4.3.3 ChatGPT-C vs ChatGPT-A

The next version of ChatGPT tested, ChatGPT-C, which used the prompt *"try to understand how I'm feeling"* at the end of the user's utterance, takes a different path when compared to the regular version, ChatGPT-A.

	ChatGPT-A	**ChatGPT-B**
admiration	4955	4772
caring	3748	3956
approval	2640	2525
joy	1441	1666
curiosity	750	816
optimism	488	487
neutral	473	433
sadness	434	472
remorse	419	505
gratitude	405	523
excitement	366	300
love	321	296
annoyance	223	225
anger	203	135
fear	178	126
disapproval	175	119
disappointment	119	68
embarrassment	113	82
relief	87	62
surprise	62	45
realization	33	32
amusement	32	31
confusion	31	30
grief	28	28
nervousness	27	24
desire	15	17
disgust	14	5

Table 4.6: Contingency table for ChatGPT-A and ChatGPT-B.

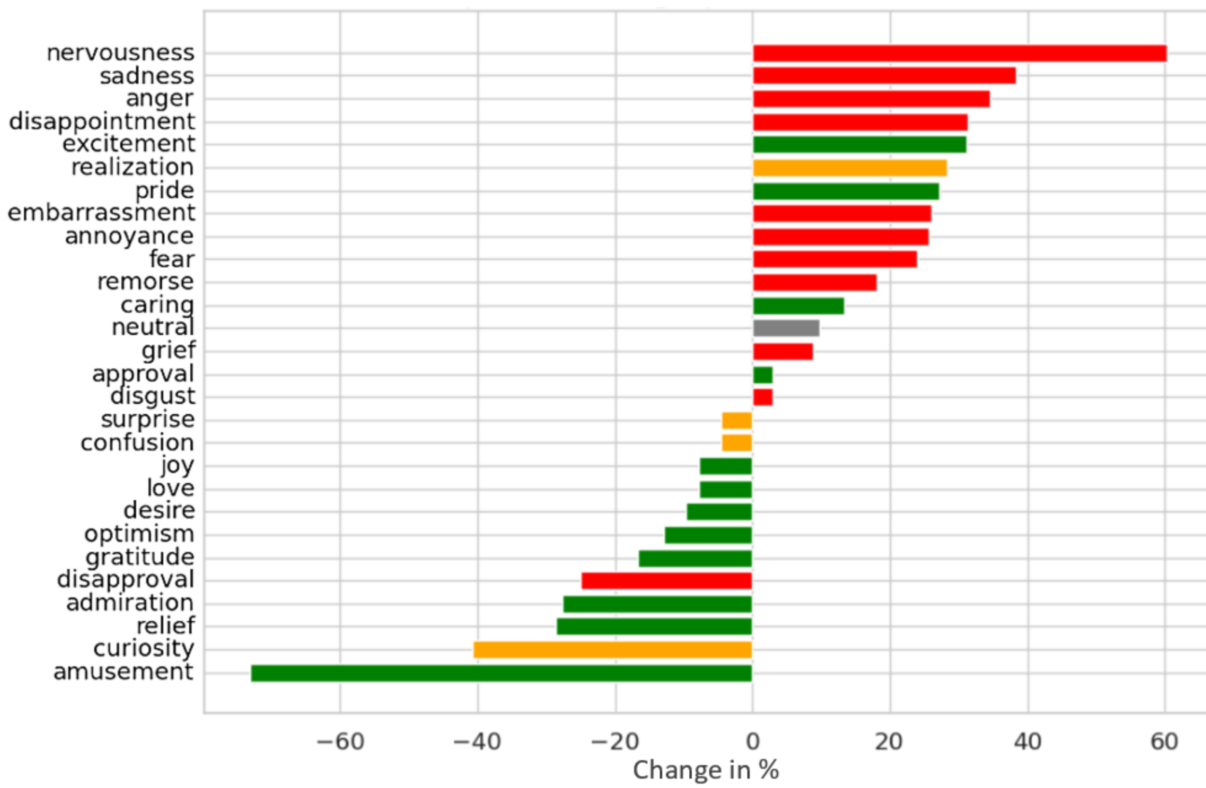

Figure 4.6: ChatGPT-C vs. ChatGPT-A emotion frequency.

4.3.3.1 Overall results

Similarly to the previous section, we illustrate in figure 4.6 a horizontal bar chart showing the change in how often ChatGPT-C uses each emotion compared to the regular version of ChatGPT (ChatGPT-A). We can see in this figure that ChatGPT-C tends to utilize negative emotions more frequently and positive emotions less often. This finding can be attributed to the nature of the prompt itself, which implicitly suggests a request for understanding and empathy toward the user's emotional state. The phrase *"try to understand how I'm feeling"* is more likely to be employed when the user is experiencing negative emotions such as sadness or frustration rather than positive emotions like joy or excitement. Consequently, ChatGPT-C aligns its responses to match the user's emotional state, exhibiting a higher prevalence of negative emotions. To confirm that, we can examine the emotion frequency change per user emotion illustrated in the heatmap of the figure 4.7.

In the mentioned heatmap, we can notice the biggest changes in the following situations:

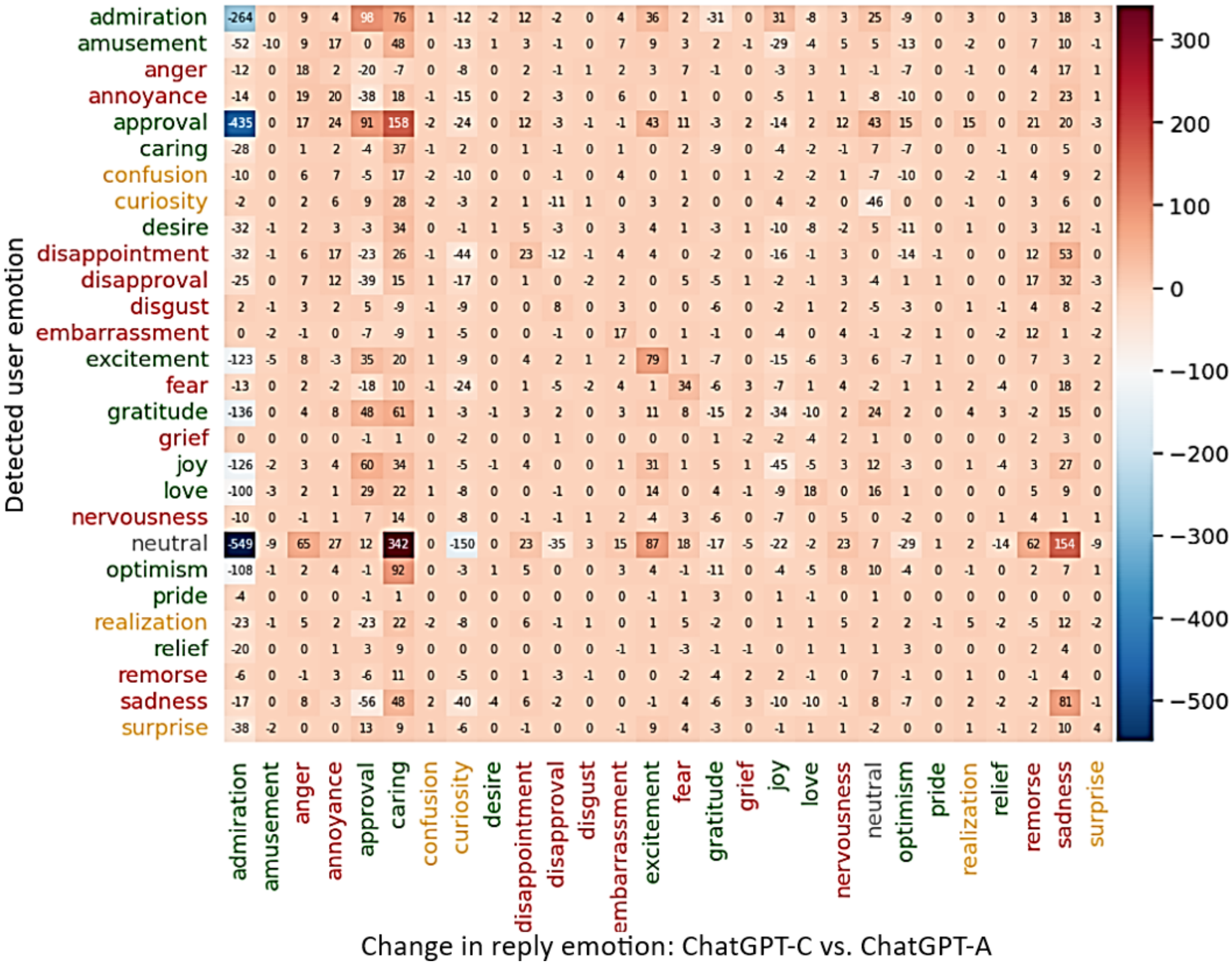

Figure 4.7: ChatGPT-C: Response emotion per user emotion.

- When the user is neutral, the chatbot expresses admiration and curiosity much less often and instead tries to mimic either the "caring" and excitement or the sadness, anger, and remorse emotions. This suggests that ChatGPT-C aims to establish a more empathetic connection by expressing understanding and concern and potentially addressing any negative emotions the user may be experiencing.

- When the user appears to be sad, the chatbot expresses "approval", "curiosity", and "joy" less often and expresses "caring" and "sadness" instead, reflecting a more empathetic and compassionate approach to acknowledging and addressing the user's sadness.

- If ChatGPT-C finds that the user is fearing something, it expresses the "fear" emotion instead of "approval" or "curiosity". This adaptation suggests that ChatGPT-C aims to respond with empathy and understanding, acknowledging the user's fear and potentially providing reassurance or support.

- The approval and caring emotions are more often used by ChatGPT-C, especially when the user is neutral or expresses "approval", "admiration", or "gratitude". This indicates that ChatGPT-C emphasizes understanding and supporting the user's positive emotions or seeking to provide a caring and attentive response.

4.3.3.2 Error analysis

These observations demonstrate that ChatGPT-C can alter its affective responses based on the emotional state of the user. However, compared to ChatGPT-B, ChatGPT-C exhibits a distinct adaptation in its emotional responses.

While ChatGPT-B, with explicit user emotion used as an input, tends to use positive emotions more frequently, ChatGPT-C tends to use negative emotions more frequently. The prompt "try to understand how I'm feeling" is likely to convey negative feelings regardless of the actual user's emotion and thus inciting the chatbot to respond in an empathetic manner by mirroring the negative sentiments the user may be experiencing. In contrast, ChatGPT-B's explicit user emotion input enables it to respond with a greater spectrum of positive emotions, displaying a more optimistic and supporting tone. Both versions feature distinct emotional adaptations, catering to various facets of empathetic conversations.

4.3.3.3 Significance analysis

Just like what we did with ChatGPT-B in section 4.3.2.5, we performed the statistical analysis to see to what point the prompt-engineered ChatGPT-C differs from the regular ChatGPT-A and if the difference is statistically significant.

Table 4.7 presents the contingency table of ChatGPT-C vs. ChatGPT-B, that is the number of emotion labels used by each chatbot version.

Again, the p-value obtained in this case is much smaller than the chosen significance level (α=0.05). In fact, the calculated p-value using the Scipy Python library is so tiny that the calculation in Python

	ChatGPT-A	**ChatGPT-C**
admiration	5024	2847
caring	3668	4796
approval	2631	2796
joy	1433	1225
curiosity	743	313
optimism	501	387
neutral	483	587
sadness	452	1014
gratitude	432	308
remorse	385	556
excitement	368	702
love	295	252
annoyance	229	388
disapproval	185	111
anger	184	379
fear	172	281
disappointment	123	236
embarrassment	113	193
surprise	81	74
relief	63	35
amusement	45	7
realization	39	70
confusion	34	31
grief	31	37
nervousness	29	117
desire	17	14
disgust	16	17
pride	4	7

Table 4.7: Contingency table for ChatGPT-C vs. ChatGPT-A.

considered it to be zero. Thus we reject the null hypothesis and consider the obtained emotion labels highly dependent on the chosen chatbot version. This indicates that our enhancements are statistically significant, which means that replies from ChatGPT-C are significantly different from those from the regular ChatGPT-A and that prompt engineering greatly influences the kind of replies we get from LLMs.

4.3.4 Comparaisons to the SOTA Models

4.3.4.1 Models used for comparison

Comparing our three ChatGPT versions with other State-Of-The-Art (SOTA) emotion-aware chatbot models from the literature allows us to gain insights into their relative performance. The models used for comparison are the following:

- The original **Ttransformer** model (Vaswani *et al.*, 2017) proposed in 2017 uses a self-attention mechanism to generate responses to user input without special consideration for emotions.
- **Emoprepend** (Rashkin *et al.*, 2018) is an improved version of the transformer with an emotion classifier trained to minimize the joint generation and classification loss.
- **EmoEL** (Lin *et al.*, 2019) is another transformer-based model that leverages responses from different decoders corresponding to different emotions.
- **EmpDG** (Li *et al.*, 2019b) is another emotional chatbot model that uses both coarse-grained dialogue-level and fine-grained token-level emotions to generate empathetic responses.

We compared our proposed ChatGPT versions with these models in terms of the output emotion compared to the emotion label of the Empathetic Dialogues ground truth response emotion.

Additionally, we assessed the fluency of the chatbot models by calculating the perplexity score for their generated replies. Perplexity measures the level of uncertainty or "surprise" of a language model when predicting the next word in a sequence. A lower perplexity score indicates higher fluency and coherence in the generated responses (Bahl *et al.*, 1983). The results are in table 4.8.

	Response emotion choice				Perplexity
	Acc.	*Prec.*	*Recall*	*F1*	
Transformer	15.04	17.74	7.2	5.46	89.89
Emoprepend	15.04	**18.15**	7.72	5.58	94.77
MoEl	15.59	16.96	7.52	5.69	112.3
EmpDG	15.19	14.92	7.87	6.02	111.9
ChatGPT-A	28.64	13.19	**14.14**	**12.4**	30.63
ChatGPT-B	**39.55**	14.1	13.35	12.34	28.38
ChatGPT-C	22.71	10.03	12.23	9.4	**26.35**

Table 4.8: Benchmarking with other empathetic chatbots.

Source	**Utterance**
Context	**Yeah about 10 years ago i had a horrifying experience . it was 100 % their fault but they hit the water barrels and survived . they had no injuries but they almost ran me off the road .**
Reference	*Did you suffer any injuries ?*
Transformer	Oh no , i am so sorry to hear that .
emoprepend	Oh no , did they catch them ?
moel	Oh no ! i hate spiders !
emp_DG	Oh no ! i am so sorry . i am so sorry . i am so sorry . i am so sorry to hear that .
ChatGPT-A	Wow, that sounds really scary! Have you been able to cope with the experience since then?
ChatGPT-B	I'm sorry to hear about your scary experience. It's alarming to think about what could have happened, but I'm glad you're okay.
ChatGPT-C	That sounds really scary and traumatic, and it's understandable that you would still remember it vividly.

Table 4.9: Reply examples from the different chatbot models.

4.3.4.2 Response emotion choice results

Considering the zero-shot nature of their training on the Empathetic Dialogues dataset, the results of the comparison between our proposed ChatGPT variants and other emotion-aware chatbot models reveal the remarkable performance of our models. ChatGPT-B, in particular, obtained the highest emotion accuracy of all the models, demonstrating its ability to generate responses that are emotionally aligned without specific training on the dataset. The regular ChatGPT also demonstrated excellent recall and F1-score, indicating its proficiency in recognizing and responding to the emotional content of dialogues.

4.3.4.3 Error analysis of the response emotion choice

The relatively low scores for emotion accuracy, recall, and F1-score can be attributed to a number of factors. Unlike other models, our ChatGPT variants are disadvantaged by the absence of training or fine-tuning on the Empathetic Dialogues dataset specifically. Moreover, the use of a large number of fine-grained emotion labels (28 labels) increases the difficulty of precisely matching the reference emotion. Emotions can be expressed in a variety of ways, and the variety of potential responses makes the task challenging.

Moreover, the disparity between the reference answers and chatbot responses must be taken into account. While the reference answers may include queries to demonstrate curiosity, our models' responses prioritize expressing concern. This variation in response technique can have an effect on the exact alignment of emotions. In fact, a conversational agent can appear empathetic and emotional with several classes of emotions. For example, when looking at the answers from chatbots, we find that sometimes in the reference, the answer to something like "I had an accident" is a question like "are you okay now?" which expresses the emotion 'curiosity' while the chatbot says "I hope you are okay now" which represents the emotion "caring". This explanation is supported by the prevalence of question-based responses in the reference answers (25%) increases the difficulty of the task, while our chatbot models' responses are primarily dominated by the "caring" emotion. In spite of these challenges, our ChatGPT variants outperform other models on most metrics, particularly considering their zero-shot approach and lack of direct training on the Empathetic Dialogues dataset. This demonstrates the steerability and robustness of the ChatGPT chatbot model in producing empathetic and emotionally-aligned responses.

4.3.4.4 Fluency comparison with SOTA models

On the perplexity front, it's clear that GPT-3.5-based ChatGPT models outperform all the other chatbot models in generating coherent and fluent responses. Recall that a lower perplexity means a more coherent expression (Bahl *et al.*, 1983), we can see that ChatGPT-based models are vastly superior on this level. Specifically, the emotion-adapting ChatGPT-C has the lowest perplexity score of 26.35, indicating its ability to generate highly coherent responses, while the emotion-infused ChatGPT-B has a slightly worse perplexity score of 28.38, and the regular ChatGPT-A version has an even slightly higher perplexity of 30.63. While the latter is the worst score out of the three ChatGPT models, it is still well ahead of all the other models that have a perplexity score of more than 89.89.

4.3.4.5 Fluency error analysis

Examining examples in table 4.9 further supports the superiority of the ChatGPT models in terms of response coherence. For instance, while emp_DG's reply does express remorse, it does so in an unnatural and repetitive sentence structure: *"oh no ! i am so sorry . i am so sorry . i am so sorry . i am so sorry to hear that ."* which likely contributes to its bad perplexity score.

The ChatGPT models, notably ChatGPT-C and ChatGPT-B, outperform other chatbot models in terms of complexity, generating coherent and natural-sounding responses. Their superior performance in sustaining conversational flow and coherence is a result of their ability to generate more natural and context-appropriate replies.

4.4 Conclusion

In this chapter, we conducted a series of experiments to investigate our approaches to building a reliable emotion classifier and enhancing ChatGPT's emotional capabilities.
Our findings shed light on the efficacy of our ELECTRA-based emotion classifier in predicting emotions accurately, obtaining remarkable performance across all evaluation metrics. This classifier offers a solid foundation for injecting emotions into ChatGPT and for evaluating the responses of the different models as well. We also showed how utilizing user emotions as inputs can lead to more positive and limited negative emotions with ChatGPT-B, while ChatGPT-C tries to be more

empathetic by aligning its negative emotions with those of the user, giving a sense of empathy and understanding.

Comparing our proposed ChatGPT models to other emotion-aware chatbot models proposed in the literature, we found that ChatGPT-B outperformed EmoEL and EmpDG in terms of emotion response accuracy, despite never being trained on the dataset. Moreover, in terms of fluency and coherence, our ChatGPT models outperformed competing models, as measured by perplexity scores. The outcomes of our research open up new avenues for future exploration and development, particularly by trying more prompt techniques or using the same approach for other conversational language models.

CONCLUSION

In this study, we looked at how ChatGPT may elicit emotional reactions. Our findings imply that using prompt engineering and external emotion classifiers to augment conversational bots' emotional intelligence can be successful.

Our research adds to the expanding pool of knowledge regarding conversational agents and their emotional intelligence. The findings suggest that external knowledge sources, such as emotion classifiers, can provide a more nuanced understanding of the user's emotional state and can lead to more affective and natural responses. Additionally, our study highlights the potential of prompt engineering to steer existing language models to produce outcomes tailored to our preferences without re-training or even fine-tuning. In the short term, this means we may not need to train new LLMs as often since we can have more personalized models just be using external modules and integrating them with existing models using prompt engineering. In the long term, the used techniques may be useful for comparison of emotional capabilities between models specifically altered to be empathetic and end-to-end conversational models with more sophisticated architectures.

Future research might examine how well ChatGPT performs with other prompt designs. Other datasets can also be examined to see how that impacts the generated replies. We can also conduct a cross-lingual study to explore the benefits and limits of prompt engineering in generative AI.

APPENDIX A
PUBLICATION

Ahmed Belkhir, Fatiha Sadat. **Beyond Information: Is ChatGPT Empathetic Enough?** In Proceedings of Recent Advances in Natural Language Processing **(RANLP 2023)**, Varna, Bulgaria, from September 4th to September 6th, 2023.

BIBLIOGRAPHY

Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P. et Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, *132*, 103978.

Adam, M., Wessel, M. et Benlian, A. (2021). Ai-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, *31*(2), 427–445.

Adamopoulou, E. et Moussiades, L. (2020a). Chatbots: History, technology, and applications. *Machine Learning with Applications*, *2*, 100006.

Adamopoulou, E. et Moussiades, L. (2020b). An overview of chatbot technology. Dans *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II 16*, 373–383. Springer.

Al-Omari, H., Abdullah, M. A. et Shaikh, S. (2020). Emodet2: Emotion detection in english textual dialogue using bert and bilstm models. Dans *2020 11th International Conference on Information and Communication Systems (ICICS)*, 226–232. IEEE.

Allouch, M., Azaria, A. et Azoulay, R. (2021). Conversational agents: Goals, technologies, vision and challenges. *Sensors*, *21*(24), 8448.

Almalki, M. et Azeez, F. (2020). Health chatbots for fighting covid-19: a scoping review. *Acta Informatica Medica*, *28*(4), 241.

Almansor, E. H. et Hussain, F. K. (2020). Survey on intelligent chatbots: State-of-the-art and future research directions. Dans *Complex, Intelligent, and Software Intensive Systems: Proceedings of the 13th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS-2019)*, 534–543. Springer.

Ando, R. K. et Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, *6*, 1817–1853.

Asghar, N., Poupart, P., Hoey, J., Jiang, X. et Mou, L. (2018). Affective neural response generation. Dans *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40*, 154–166. Springer.

B, M. J., J, A. A. S., M, A. R. S. et Rajan, R. (2023). Efficacy of electra-based language model in sentiment analysis. Dans *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, 682–687. http://dx.doi.org/10.1109/ICISCoIS56541.2023.10100342

Bahl, L. R., Jelinek, F. et Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, (2), 179–190.

Bail, C. A. (2023). Can generative ai improve social science?

Bandyopadhyay, S., Xu, J., Pawar, N. et Touretzky, D. (2022). Interactive visualizations of word embeddings for k-12 students. Dans *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12713–12720.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W. *et al.* (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Bansal, H. et Khan, R. (2018). A review paper on human computer interaction. *International Journals of Advanced Research in Computer Science and Software Engineering*, *8*, 53–56.

Belainine, B., Sadat, F. et Boukadoum, M. (2022). End-to-end dialog generation using a single encoder and a decoder cascade with a multi-dimension attention mechanism. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*.

Belainine, B., Sadat, F., Boukadoum, M. et Lounis, H. (2020a). Towards a multi-dataset for complex emotions learning based on deep neural networks. *Workshop on Linguistic and Neurocognitive Resources (LiNCr2020), Language Resources and Evaluation Conference (LREC 2020)*, 50–58.

Belainine, B., Sadat, F., Boukadoum, M. et Lounis, H. (2020b). Towards a multi-dataset for complex emotions learning based on deep neural networks. Dans *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, 50–58.

Belainine, B., Sadat, F., Lounis, H. et Boukadoum, M. (2020c). Towards an emotionally driven natural language generation. *Montreal AI Symposium 2020*.

Bhirud, N., Tataale, S., Randive, S. et Nahar, S. (2019). A literature review on chatbots in healthcare domain. *International journal of scientific & technology research*, *8*(7), 225–231.

Bilquise, G., Ibrahim, S., Shaalan, K. *et al.* (2022). Emotionally intelligent chatbots: A systematic literature review. *Human Behavior and Emerging Technologies*, *2022*.

Brandtzaeg, P. B. et Følstad, A. (2017). Why people use chatbots. Dans *International conference on internet science*, 377–392. Springer.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Cahn, J. (2017). Chatbot: Architecture, design, & development. *University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science*.

Casas, J., Spring, T., Daher, K., Mugellini, E., Khaled, O. A. et Cudré-Mauroux, P. (2021). Enhancing conversational agents with empathic abilities. Dans *Proceedings of the 21st ACM*

International Conference on Intelligent Virtual Agents, 41–47.

Chen, J.-S., Le, T.-T.-Y. et Florence, D. (2021). Usability and responsiveness of artificial intelligence chatbot on online customer experience in e-retailing. *International Journal of Retail & Distribution Management*, *49*(11), 1512–1531.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S. et Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, *30*.

Chronopoulou, A., Baziotis, C. et Potamianos, A. (2019). An embarrassingly simple approach for transfer learning from pretrained language models. *arXiv preprint arXiv:1902.10547*.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S. *et al.* (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Clark, K., Luong, M.-T., Le, Q. V. et Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R. *et al.* (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Colace, F., De Santo, M., Lombardi, M., Pascale, F., Pietrosanto, A. et Lemma, S. (2018). Chatbot for e-learning: A case of study. *International Journal of Mechanical Engineering and Robotics Research*, *7*(5), 528–533.

Colby, K. M., Weber, S. et Hilf, F. D. (1971). Artificial paranoia. *Artificial Intelligence*, *2*(1), 1–25.

Colombo, P., Witon, W., Modi, A., Kennedy, J. et Kapadia, M. (2019). Affect-driven dialog generation. *arXiv preprint arXiv:1904.02793*.

Cortiz, D. (2021). Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta, xlnet and electra.

Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R. et Keltner, D. (2019). The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature human behaviour*, *3*(4), 369–382.

Creswell, A., Shanahan, M. et Higgins, I. (2022). Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G. et Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Devaram, S. (2020). Empathic chatbot: Emotional intelligence for empathic chatbot: Emotional intelligence for mental health well-being. *arXiv preprint arXiv:2012.09130*.

Devlin, J., Chang, M.-W., Lee, K. et Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169–200.

Epstein, R. (1992). The quest for the thinking computer. *AI magazine*, *13*(2), 81–81.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I. et Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*.

Følstad, A. et Brandtzæg, P. B. (2017). Chatbots and the new world of hci. *interactions*, *24*(4), 38–42.

Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. C. et Berner, J. (2023). Mathematical capabilities of chatgpt.

Ghosh, S. et Caliskan, A. (2023). Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. *arXiv preprint arXiv:2305.10510*.

Gnewuch, U., Morana, S., Adam, M. T. et Maedche, A. (2018). "the chatbot is typing..."–the role of typing indicators in human-chatbot interaction.

Grudin, J. et Jacques, R. (2019). Chatbots, humbots, and the quest for artificial general intelligence. Dans *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–11.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J. et Wu, Y. (2023). How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Haque, M. U., Dharmadasa, I., Sworna, Z. T., Rajapakse, R. N. et Ahmad, H. (2022). " i think this is the most disruptive technology": Exploring sentiments of chatgpt early adopters using twitter data. *arXiv preprint arXiv:2212.05856*.

Haristiani, N. (2019). Artificial intelligence (ai) chatbot as language learning medium: An inquiry. Dans *Journal of Physics: Conference Series*, volume 1387, p. 012020. IOP Publishing.

Hasegawa, T., Kaji, N., Yoshinaga, N. et Toyoda, M. (2013). Predicting and eliciting addressee's emotion in online dialogue. Dans *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 964–972.

Heiser, J. F., Colby, K. M., Faught, W. S. et Parkison, R. C. (1979). Can psychiatrists distinguish a computer simulation of paranoia from the real thing?: The limitations of turing-like tests as measures of the adequacy of simulations. *Journal of Psychiatric Research*, *15*(3), 149–162.

Hochreiter, S. et Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Huang, M., Ye, Z. et Zhou, H. (2018). Overview of the nlpcc 2017 shared task: Emotion generation challenge. Dans *Natural Language Processing and Chinese Computing: 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8–12, 2017, Proceedings 6*, 926–936. Springer.

Jiao, W., Wang, W., Huang, J., Wang, X. et Tu, Z. (2023). Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N. et Wu, Y. (2016). Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Kamath, V. R. *et al.* (2020). Transferred fusion learning using skipped networks. *arXiv preprint arXiv:2011.05895*.

Kashefi, A. et Mukerji, T. (2023). Chatgpt for programming numerical methods. *Journal of Machine Learning for Modeling and Computing*, *4*(2).

Khanna, A., Pandey, B., Vashishta, K., Kalia, K., Pradeepkumar, B. et Das, T. (2015). A study of today's ai through chatbots and rediscovery of machine intelligence. *International Journal of u-and e-Service, Science and Technology*, *8*(7), 277–284.

Kim, Y., Jernite, Y., Sontag, D. et Rush, A. (2016). Character-aware neural language models. Dans *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K. *et al.* (2023). Chatgpt: Jack of all trades, master of none. *arXiv preprint arXiv:2302.10724*.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. et Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T. et Nguyen, T. H. (2023). Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Lee, N., Ping, W., Xu, P., Patwary, M., Fung, P. N., Shoeybi, M. et Catanzaro, B. (2022). Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, *35*, 34586–34599.

Leiter, C., Zhang, R., Chen, Y., Belouadi, J., Larionov, D., Fresen, V. et Eger, S. (2023). Chatgpt: A meta-analysis after 2.5 months. *arXiv preprint arXiv:2302.13795*.

Li, C., Wang, J., Zhu, K., Zhang, Y., Hou, W., Lian, J. et Xie, X. (2023). Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus. *arXiv preprint arXiv:2307.11760*.

Li, J., Sun, X., Wei, X., Li, C. et Tao, J. (2019a). Reinforcement learning based emotional editing constraint conversation generation. *arXiv preprint arXiv:1904.08061*.

Li, Q., Chen, H., Ren, Z., Ren, P., Tu, Z. et Chen, Z. (2019b). Empdg: Multiresolution interactive empathetic dialogue generation. *arXiv preprint arXiv:1911.08698*.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z. et Niu, S. (2017). Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Li, Y., Yin, Y., Li, J. et Zhang, Y. (2022). Prompt-driven neural machine translation. Dans *Findings of the Association for Computational Linguistics: ACL 2022*, 2579–2590.

Lin, Z., Madotto, A., Shin, J., Xu, P. et Fung, P. (2019). Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. et Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luo, X., Tong, S., Fang, Z. et Qu, Z. (2019). Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science*, *38*(6), 937–947.

Magazine, C. (2018). Chatbot report 2018: global trends and analysis. *Retirado de: https://chatbotsmagazine. com/chatbot-report-2018-global-trends-and-analysis-4d8bbe4d924b*.

Mala, J., SJ, A. A., SM, A. R. et Rajan, R. (2023). Efficacy of electra-based language model in sentiment analysis. Dans *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, 682–687. IEEE.

Marco della Cava, U. T. (2016). *Microsoft CEO Nadella: 'Bots are the new apps'*. USA TODAY. Récupéré le August 21, 2023 de `https://www.usatoday.com/story/tech/news/2016/03/30/microsof-ceo-nadella-bots-new-apps/82431672/`

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica*, *23*(2), 143–149.

Mikolov, T., Chen, K., Corrado, G. et Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., Yuan, Q., Tezak, N., Kim, J. W., Hallacy, C. *et al.* (2022). Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.

Niu, T. et Bansal, M. (2018). Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, *6*, 373–389.

OpenAI. (2022). *Introducing ChatGPT*. Récupéré de https://openai.com/blog/chatgpt

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. *et al.* (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Pamungkas, E. W. (2019). Emotionally-aware chatbots: A survey. *arXiv preprint arXiv:1906.09774*.

Pan, S. J. et Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345–1359.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion* 3–33. Elsevier.

Polak, M. P. et Morgan, D. (2023). Extracting accurate materials data from research papers with conversational language models and prompt engineering–example of chatgpt. *arXiv preprint arXiv:2303.05352*.

Pu, D. et Demberg, V. (2023). Chatgpt vs human-authored text: Insights into controllable text summarization and sentence style transfer.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M. et Yang, D. (2023). Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Qiu, L., Shiu, Y., Lin, P., Song, R., Liu, Y., Zhao, D. et Yan, R. (2020). What if bots feel moods? Dans *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1161–1170.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. *et al.* (2018). Improving language understanding by generative pre-training.

Rapp, A., Curti, L. et Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, *151*, 102630.

Rashkin, H., Smith, E. M., Li, M. et Boureau, Y.-L. (2018). Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Ray, P. P. (2023). Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Reynolds, L. et McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. Dans *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–7.

Rubin, A. M. (2009). Uses and gratifications. *The SAGE handbook of media processes and effects*, 147–159.

Rumelhart, D. E., Hinton, G. E. et Williams, R. J. (1985). *Learning internal representations by error propagation*. Rapport technique, California Univ San Diego La Jolla Inst for Cognitive Science.

Salovey, P. et Mayer, J. D. (1990). Emotional intelligence. *Imagination, cognition and personality*, *9*(3), 185–211.

Sanh, V., Debut, L., Chaumond, J. et Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sobania, D., Briesch, M., Hanna, C. et Petke, J. (2023). An analysis of the automatic bug fixing performance of chatgpt. *arXiv preprint arXiv:2301.08653*.

Stap, D. et Araabi, A. (2023). Chatgpt is not a good indigenous translator. Dans *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, 163–167.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D. et Christiano, P. F. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, *33*, 3008–3021.

Su, M.-H., Wu, C.-H., Huang, K.-Y., Hong, Q.-B. et Wang, H.-M. (2017). A chatbot using lstm-based multi-layer embedding for elderly care. Dans *2017 International Conference on Orange Technologies (ICOT)*, 70–74. IEEE.

Suhaili, S. M., Salim, N. et Jambli, M. N. (2021). Service chatbots: A systematic review. *Expert Systems with Applications*, *184*, 115461.

Sun, X., Chen, X., Pei, Z. et Ren, F. (2018). Emotional human machine conversation generation based on seqgan. Dans *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, 1–6. IEEE.

Susnjak, T. (2022). Chatgpt: The end of online exam integrity? *arXiv preprint arXiv:2212.09292*.

Sutskever, I., Vinyals, O. et Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. *et al.* (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. et Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vinyals, O. et Le, Q. (2015). A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Wake, N., Kanehira, A., Sasabuchi, K., Takamatsu, J. et Ikeuchi, K. (2023). Bias in emotion recognition with chatgpt. *arXiv preprint arXiv:2310.11753*.

Wallace, R. S. (2009). *The anatomy of ALICE*. Springer.

Wang, Z., Xie, Q., Ding, Z., Feng, Y. et Xia, R. (2023). Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*.

Warriner, A. B., Kuperman, V. et Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, *45*, 1191–1207.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D. *et al.* (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Wei, W., Liu, J., Mao, X., Guo, G., Zhu, F., Zhou, P. et Hu, Y. (2019). Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. Dans *Proceedings of the 28th ACM international conference on information and knowledge management*, 1401–1410.

Weiss, K., Khoshgoftaar, T. M. et Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*(1), 1–40.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J. et Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Workshop, B., Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F. *et al.* (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Wu, Y., Wu, W., Xing, C., Zhou, M. et Li, Z. (2016). Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*.

Xu, A., Liu, Z., Guo, Y., Sinha, V. et Akkiraju, R. (2017). A new chatbot for customer service on social media. Dans *Proceedings of the 2017 CHI conference on human factors in computing systems*, 3506–3510.

Yang, X., Li, Y., Zhang, X., Chen, H. et Cheng, W. (2023). Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R. et Le, Q. V. (2019a). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, *32*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R. et Le, Q. V. (2019b). Xlnet: Generalized autoregressive pretraining for language understanding. Dans H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, et R. Garnett (dir.).

Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc. Récupéré de https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf

Yao, K., Zhang, L., Luo, T., Du, D. et Wu, Y. (2021). Non-deterministic and emotional chatting machine: learning emotional conversation generation using conditional variational autoencoders. *Neural Computing and Applications*, *33*, 5581–5589.

Zaib, M., Sheng, Q. Z. et Emma Zhang, W. (2020). A short survey of pre-trained language models for conversational ai-a new age in nlp. Dans *Proceedings of the Australasian computer science week multiconference*, 1–4.

Zhang, R., Guo, J., Fan, Y., Lan, Y. et Cheng, X. (2020). Dual-factor generation model for conversation. *ACM Transactions on Information Systems (TOIS)*, *38*(3), 1–31.

Zhang, R., Wang, Z. et Mai, D. (2017). Building emotional conversation systems using multi-task seq2seq learning. Dans *National CCF Conference on Natural Language Processing and Chinese Computing*, 612–621. Springer.

Zhang, R., Wang, Z. et Mai, D. (2018). Building emotional conversation systems using multi-task seq2seq learning. Dans *Natural Language Processing and Chinese Computing: 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8–12, 2017, Proceedings 6*, 612–621. Springer.

Zhang, X. et LeCun, Y. (2015). Text understanding from scratch. *arXiv preprint arXiv:1502.01710*.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J. et Dolan, B. (2019). Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Zhao, W., Zhao, Y., Lu, X., Wang, S., Tong, Y. et Qin, B. (2023). Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. *et al.* (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Zhong, Q., Ding, L., Liu, J., Du, B. et Tao, D. (2023). Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L. *et al.* (2023). A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*.

Zhou, H., Huang, M., Zhang, T., Zhu, X. et Liu, B. (2018). Emotional chatting machine: Emotional conversation generation with internal and external memory. Dans *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. et He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43–76.