BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF FOREIGN LANGUAGES
ÚSTAV JAZYKŮ

CHATGPT: PRINCIPLES, BENEFITS AND PITFALLS
CHATGPT: PRINCIPY, PŘÍNOSY A ÚSKALÍ

BACHELOR'S THESIS
BAKALÁŘSKÁ PRÁCE

AUTHOR **Daniel Ulman**
AUTOR PRÁCE

SUPERVISOR **Mgr. Magdalena Šedrlová**
VEDOUCÍ PRÁCE

BRNO 2024

Bakalářská práce

bakalářský studijní obor **Angličtina v elektrotechnice a informatice**

Ústav jazyků

Student: Daniel Ulman ***ID:*** 240301

Ročník: 3 ***Akademický rok:*** 2023/24

NÁZEV TÉMATU:

ChatGPT: Principy, přínosy a úskalí

POKYNY PRO VYPRACOVÁNÍ:

Proveďte rešerši dostupné literatury k tématu a představte základní principy fungování chatbotů, využití ChatGPT v praxi a proveďte analýzu aktuálních technických i etických limitů této technologie. Poznatky shrňte v logicky strukturovaném textu psaném formálním jazykem.

DOPORUČENÁ LITERATURA:

Russell, S. J., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach, Global Edition (4th ed.). Pearson Education.

Termín zadání: 22.2.2024 ***Termín odevzdání:*** 24.5.2024

Vedoucí práce: Mgr. Magdalena Šedrlová

doc. PhDr. Milena Krhutová, Ph.D.
předseda oborové rady

UPOZORNĚNÍ:

Autor bakalářské práce nesmí při vytváření bakalářské práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

Abstract

The rapid advancements in the field of artificial intelligence over the past few years suggest that humanity is likely to rely more and more on AI in various fields. Among these fields, chatbots have become an exciting emerging domain within artificial intelligence. In this bachelor thesis, the evolution of artificial intelligence is presented initially, providing context for the core principle of chatbot operation. The inner workings of chatbots are then described, and several types of chatbots are introduced according to different classifications. These facts lay the groundwork for the main topic of this thesis, ChatGPT. Findings from the literature review show that large language models, exemplified by ChatGPT, may be a turning point in human interaction with AI, opening up new possibilities in the digital era but also raising concerns about its pitfalls.

Keywords

Artificial intelligence, bias, chatbot, ChatGPT, natural language processing, prompt

Abstrakt

Díky rychlému pokroku v oblasti umělé inteligence v posledních několika letech je pravděpodobné, že bude lidstvo na umělou inteligenci spoléhat čím dál více v nejrůznějších oblastech. Jednou takovou zajímavou rozvíjející se oblastí v rámci umělé inteligence se stávají chatboti. V této bakalářské práci je nejprve představen vývoj umělé inteligence, který poskytuje kontext pro základní princip fungování chatbotů. Poté je popsáno vnitřní fungování chatbotů a představeno několik typů chatbotů rozdělených dle několika různých klasifikací. Tyto fakta slouží jako podklad pro zkoumání hlavního tématu této práce, ChatGPT. Poznatky z rešerše literatury ukazují, že velké jazykové modely, jako je právě ChatGPT, mohou být přelomem v lidské interakci s umělou inteligencí, otevírající nové možnosti v digitální éře, ale také vyvolávající obavy o jejich úskalích.

Klíčová slova

Umělá inteligence, zaujatost, chatbot, ChatGPT, zpracování přirozeného jazyka, výzva

Rozšířený abstrakt

Umělá inteligence (AI) zažívá v posledních několika letech významný technologický pokrok a ovlivňuje svět kolem nás. Mimo jiné obory umělé inteligence se tento vývoj dotýká zejména oblasti AI chatbotů, kteří se stávají nedílnou součástí našich digitálních životů. S rychlým rozvojem této oblasti vzniká řada nástrojů využívajících principy umělé inteligence. Jedním takovým nástrojem, jímž se zabývá právě tato bakalářská práce, je chatbot zvaný ChatGPT.

Cílem této bakalářské práce je provést rešerši dostupné literatury, jejímž prostřednictvím budou popsány základní principy chatbotů a představen ChatGPT se zaměřením na jeho fungování a využití v praxi, dále seznámit čtenáře se současnými technickými a etickými limity této technologie a na základě toho zhodnotit její přínosy a úskalí. Z hlediska metodologie je tato bakalářská práce rešerší odborné literatury a článků zabývajících se tématy umělé inteligence, chatbotů a technologií ChatGPT.

Umělá inteligence je jednou z nejrychleji se rozvíjejících oblastí v moderní technologii a předpokládá se, že její vliv na lidstvo bude největší v historii civilizace. Vyznačuje se svou všestranností a schopností uplatnit se při řešení různých úkolů vyžadujících lidský intelekt. Využití nachází především v profesích a odvětvích jako je online zákaznický servis, výroba, kontrola kvality, automobilový průmysl, zdravotnictví, internetové bankovnictví a marketing. Měli bychom však očekávat její postupné proniknutí do dalších oblastí jako je vzdělávání, bezpečnost a doprava, protože schopnosti umělé inteligence se zdokonalují a její limity se neustále posouvají.

V úvodu se tato práce zabývá vývojem umělé inteligence, respektive jejími podobory, jimiž jsou strojové učení, hluboké učení, neuronové sítě a zpracování přirozeného jazyka.

Strojové učení představuje základní pilíř umělé inteligence. Jde o proces, při kterém se počítač učí z vložených dat za pomocí algoritmů, na jejichž základě přiřazuje výstupy k vstupním proměnným v určitých modelových situacích. V reakci na určitou situaci pak na základě predikce a pravděpodobnosti provede žádanou operaci bez explicitního programování člověkem.

Hluboké učení, založeno na principech strojového učení, se ale výrazně liší množstvím dat, použitých algoritmů a architektury, která je složitější a zahrnuje více vrstev. Pojem „hluboké“ v názvu tohoto podoboru pak poukazuje na hloubku sítě, respektive právě na počet těchto vrstev v neuronové síti modelu.

Neuronové sítě, jako další podobor umělé inteligence, jsou založeny na analogii s neuronovými spojeními v lidském mozku. Princip neuronové sítí je možné si představit následovně: Ve vstupní vrstvě jsou data ve formě neuronů. Všechny tyto neurony jsou spojeny s neurony v první skryté vrstvě, přičemž těchto skrytých vrstev může být několik. Tím vzniká několik neuronových spojení napříč vrstvami. Ve skrytých vrstvách jsou každému neuronovému spojení přiřazeny váhy a prahy (angl. biases). Tyto parametry ovlivňují způsob, jakým neuron reaguje na vstupní data, a tedy přímo ovlivňuje jeho aktivaci v dalších vrstvách. Tak je tomu až do vrstvy poslední, kde se na základě predikce a trénování sítě objeví požadovaný výstup. Na základě takové sítě je pak možné zpracovávat přirozený jazyk. Zpracování přirozeného jazyka (NLP), jakožto poslední podobor, se zabývá porozuměním a generováním lidského jazyka. Představuje soubor technik používaných v chatbotech pro získávání a interpretaci informací ze zdrojů dat nebo databází obsahující texty psané v přirozeným jazyce vložených do chatbota během fáze učení. Stejné techniky se dále používají pro porozumění textu zadaného uživatelem a při generování následné odpovědí v přirozeném jazyce.

V další kapitole práce navazuje tématikou chatbotů. Po stručném úvodu do problematiky je popsán princip fungování chatbotů graficky podpořený blokovým diagramem popisujícím jejich architekturu. V další části této kapitoly je uvedeno šest klasifikací, přičemž pět z nich rozděluje chatboty dle jejich kvality, rychlosti a správnosti odpovědí na otázky, výzvy, nebo požadavku uživatele. Šestá klasifikace dělí chatboty dle jejich designu ve fázi vývoje. Souhrnně pak těchto šest klasifikací rozděluje chatboty na patnáct odlišných typů, přičemž každý z těchto typů se vyznačuje určitými klíčovými charakteristikami. Ty jsou spolu s klasifikacemi a typy přeneseny do tabulky, která slouží jako úvodní stručný přehled této části.

Kapitola třetí pojednává o samotném ChatGPT. Nejprve je vysvětlen samotný význam zkratky GPT, v angl. „Generative Pre-trained Transformer“ (dosl. generativní předtrénovaný transformátor). Každá část této zkratky má svůj význam: „Generative“ odkazuje na schopnost modelu generovat odpovědi ve formě psaného textu a na zařazení do kategorie generativních chatbotů. „Pre-trained“ zdůrazňuje skutečnost předchozího tréninku neuronových sítí na rozsáhlém korpusu dat, tak aby poskytovaly relevantní odpovědi. „Transformer“ naznačuje použití jistého typu architektury neuronové sítě, která využívá tzv. mechanismus pozornosti. Ten umožnuje nastavit prahy dynamicky a nezávisle na vzdálenosti slov, což zajišťuje efektivnější zachycení vztahů mezi nimi.

Dle klasifikace chatbotů lze ChatGPT považovat za generativní chatbot s otevřenou doménou. To znamená, že poskytuje odpovědi na základě vstupních dat vložených během fáze učení a používá nejmodernější generativní model. Ten mu umožňuje generovat odpovědi v přirozeném jazyce. V neposlední řadě se tato část kapitoly třetí zabývá procesem, kterým byl model GPT-3.5 vyvíjen.

V kontextu dalších chatbotů je důležité zmínit dva hlavní konkurenty ChatGPT, Bard a Bing Chat. Služba Bing Chat má výhodu v podobě přístupu k internetu prostřednictvím vyhledávače Bing, což jí umožňuje uvádět zdroje a poskytovat aktuální informace. Naopak ChatGPT nabízí zpravidla strukturovanější a přesnější odpovědi. Bard si z této trojice vede nejhůře, protože ač se jeho odpovědi zdají věrohodné, jsou často nepřesné.

Systém ChatGPT je dostupný z webového prohlížeče. Pro zobrazení chatovacího okna se uživatel musí nejprve registrovat. Poté se zobrazí uživatelské rozhraní včetně příkazového řádku, kam může uživatel vkládat zprávy v podobě otázek či příkazů. ChatGPT je na základě vložené zprávy schopný nejen odpovídat na otázky, ale i poskytovat informace z různých oblastí a řešit úkoly vyžadující lidský intelekt. Dále je schopný vytvářet, analyzovat či překládat text.

Přestože má ChatGPT široké využití v různých oblastech, omezují ho jisté technické limity. Ať už jde o velikost databáze, složitost výzev, chybějící přístup k internetu nebo maximální počet znaků či prodlevu a rychlost zodpovězení otázky. Dalšími faktory, které je třeba zvážit, jsou etické aspekty. Jmenovitě je to přítomnost zaujatosti (bias), otázka ochrany soukromí, možnost vytváření a šíření dezinformací a zodpovědné používání ze strany uživatele. Na základě celkové analýzy problematiky ChatGPT, včetně zvážení technických limitů a etických aspektů, jsou pak v závěru práce zhodnoceny přínosy a úskalí ChatGPT.

Jak vyplývá z rešerše literatury, ChatGPT přináší nové možnosti interakce s umělou inteligencí, ale také vyvolává jisté obavy z hlediska objektivní pravdy a vytváření dezinformací. ChatGPT postrádá skutečné porozumění generovaného textu a uživatelských instrukcí, které zpracovává. Proto by měl být užíván spíše jako prostředek pro získání inspirace než jako relevantní pramen znalostí.

Bibliographic citation

ULMAN, Daniel. ChatGPT: Principy, přínosy a úskalí [online]. Brno, 2024 [cit. 2024-02-12]. Dostupné z: https://www.vut.cz/studenti/zav-prace/detail/160182. Bakalářská práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav jazyků. Vedoucí práce Mgr. Magdalena Šedrlová.

Author’s Declaration

Author:	*Daniel Ulman*
Author’s ID:	*240301*
Paper type:	*Bachelor’s Thesis*
Academic year:	*2023/24*
Topic:	*ChatGPT: Principles, benefits and pitfalls*

I declare that I have written this paper independently, under the guidance of the advisor and using exclusively the technical references and other sources of information cited in the project and listed in the comprehensive bibliography at the end of the project.

As the author, I furthermore declare that, with respect to the creation of this paper, I have not infringed any copyright or violated anyone’s personal and/or ownership rights. In this context, I am fully aware of the consequences of breaking Regulation S 11 of the Copyright Act No. 121/2000 Coll. of the Czech Republic, as amended, and of any breach of rights related to intellectual property or introduced within amendments to relevant Acts such as the Intellectual Property Act or the Criminal Code, Act No. 40/2009 Coll., Section 2, Head VI, Part 4.

Brno, May 24, 2024

----
author’s signature

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Mgr. Magdalena Šedrlová, for her kind guidance, advice, and useful consultations. I am also thankful to my family for their emotional support.

Brno, May 24, 2024 --
Author's signature

Contents

FIGURES **11**

TABLES **12**

INTRODUCTION **13**

1. THE EVOLUTION OF ARTIFICIAL INTELLIGENCE **15**

1.1 MACHINE LEARNING 16
1.2 DEEP LEARNING 17
1.3 NEURAL NETWORKS 17
1.4 NATURAL LANGUAGE PROCESSING 19

2. CHATBOTS **21**

2.1 OPERATION OF CHATBOTS 22
2.2 TYPES OF CHATBOTS 23

3. CHATGPT **26**

3.1 WHAT IS CHATGPT? 26
3.2 CHATGPT'S COMPETITORS 28
3.2.1 Google Bard *28*
3.2.2 Microsoft Bing Chat *28*
3.2.3 ChatGPT vs. Bard vs. Bing Chat *28*
3.3 USE OF CHATGPT IN PRACTICE 29
3.4 PERFORMANCE EVALUATION OF CHATGPT 32
3.5 TECHNICAL LIMITATIONS OF CHATGPT 34
3.6 ETHICAL CONSIDERATIONS OF CHATGPT 36

CONCLUSION **39**

WORKS CITED **41**

LIST OF ABBREVIATIONS **43**

FIGURES

Fig. 1 The Structure of Neural Network (Sanderson and Pullen). 18
Fig. 2 General Chatbot Architecture (Adamopoulou and Moussiades) 22
Fig. 3 User Interface of ChatGPT (ChatGPT 3.5) 30

TABLES

Table 1. Classifications of Chatbots..........23

INTRODUCTION

This bachelor thesis is dedicated to the topic of chatbots. It presents their fundamentals in the field of artificial intelligence, the principles of their operation, and also different types of chatbots. Above all else, the main regard in this thesis is given particularly to one chatbot, called ChatGPT.

The aim of this bachelor thesis is to conduct a literature review in order to describe the basic principles of chatbots and to introduce ChatGPT in terms of its development, operation, competitors, and use in practical applications. Furthermore, to shed light on issues that are related to this technology by providing the reader with an overview of its current technical limitations and ethical considerations. Last but not least, use the above for an evaluation of its benefits and pitfalls as an outcome of this bachelor thesis.

Chomsky et al. suggest that ChatGPT, a product developed by OpenAI, is one of the most discussed and advanced AI chatbots used by the general public in recent times. The power of ChatGPT lies in its ability to be taught large amounts of data in the form of texts and, when given a certain input from users, produce context-based outputs that resemble the natural language and thinking of humans (Chomsky et al.).

Artificial intelligence as an emerging field capable of solving any task that demands human intellect is discussed in the first chapter of this thesis. Several subfields of artificial intelligence, such as machine learning, deep learning powered by neural networks, and natural language processing, are closely related to the core principles of chatbot operation. In order to fully understand the topic of chatbots and ChatGPT, these concepts will be explained using the insights from the book by Russell and Norvig.

In the second chapter, chatbots will be introduced. Modern chatbots have evolved from simple systems that answer questions to sophisticated agents capable of understanding natural language and performing tasks independently. They are the evidence of technological progress in the field of artificial intelligence that is reshaping human-computer interaction. Drawing on the insights of Adamopoulou and Moussiades, the chapter will discuss the inner workings of their operation as well as their different classifications. These classifications, which divide chatbots into several types, reflect their diverse functions and capabilities.

The main part of this thesis appears in the third chapter and will be dedicated to ChatGPT itself. In the beginning of this chapter, the question “What is ChatGPT?” will be answered and an insight into the background will be presented, followed by an explanation of the concept and its development. Naturally, ChatGPT is not the only AI chatbot capable of generating the “unimaginable”. Therefore, two other chatbots, namely Google Bard and Microsoft Bing Chat, will be presented and compared with ChatGPT as its biggest recent competitors. In the next subchapter, the use of ChatGPT in practise will be presented and practical applications will be shown, exemplified and evaluated in terms of executing different types prompts. Moving from its capabilities to its issues, the technical limitations associated with the architecture and operation of ChatGPT will be discussed. Furthermore, there are certain ethical considerations related to the use of ChatGPT that also present an issue, such as a presence of bias, a protection of user privacy, a possibility of creating and spreading misinformation, or a responsible use of the system, that will be covered in this chapter.

In the conclusion, after restating the aim of this thesis, the contents of the three chapters will be summarised briefly and then the benefits and pitfalls of ChatGPT will be presented and evaluated based on the findings from the literature review, especially with regard to the revealed technical limitations and ethical considerations.

Having taken into consideration the changes and improvements of ChatGPT and other chatbots, which are occurring at a rapid pace, it seems important to state in the introductory section that some of the claims in this thesis may become outdated soon. Similarly, this bachelor thesis addresses the version model ChatGPT-3.5, for it is the more widely used version among users and also the current version available free of charge, unlike its fee-based successor ChatGPT-4, also known as ChatGPT Plus. Hence, again, claims may vary from version to version as issues are being addressed and improvements are being made in upcoming updates.

Naturally, generative AI, ChatGPT-3.5, was used in this thesis, but only for the purpose of demonstrating its capabilities in practice, specifically in chapter *3.3*, with two instances in total. The transcription of these two generated outputs is in italics and marked by the parenthetical citation “*(ChatGPT 3.5)*”.

1. THE EVOLUTION OF ARTIFICIAL INTELLIGENCE

According to Russell and Norvig, artificial intelligence (AI) is considered to be one of the most emerging and developing fields. It is predicted that the impact of AI on humanity will be the biggest in the history of civilisation. This is related to the fact that unlike fields of science such as mathematics and physics, where it seems that all the fundamental discoveries have already been made, the field of artificial intelligence is still largely unexplored and remains to be further explored. The most fascinating thing about AI is its universality; it can be applied to any task that requires human intellect. Thus, it covers a significant variety of applications in today's world (19).

There are many definitions describing AI. According to Jiang et al., the definition lies in comparing artificial intelligence to human intelligence; it is an intelligence that is required to enable machines to execute the tasks that were once supposed to be performed only by humans (Jiang et al.).

While the concept of AI has been in discussion for nearly 70 years, it has become relevant only in the last few years, becoming more available to the wider public (Kaynak).

Burns et al. perceive AI as a field of significant importance, as its applications are present in today's society, mainly in professions and fields including online customer service, quality control, the automotive industry, healthcare, and marketing. It is believed that some of these tasks are even better performed by AI than by humans, particularly the detail-oriented ones that require repetitiveness. Thus, people are now surrounded by or even replaced by AI in their professional lives. With the efficiency and automation brought by AI also come job displacement and lost job positions. Sooner or later, humankind has to adapt to these changes. For the reason that the development of AI in various fields, including education, security, manufacturing, transport, and healthcare, should be expected, as the capabilities of AI are expanding and limitations are being pushed (Burns et al.).

Over the years, AI has evolved from machine learning to deep learning, powered by neural networks, and natural language processing. These fundamental subfields of AI have been selected to be discussed and explained in the first section of the thesis to provide a background for chatbots and ChatGPT, as they form the very core principles of their operation.

1.1 Machine Learning

Machine learning represents the beginning of the evolutionary line of artificial intelligence and is the foundation of the subsequent subfields of artificial intelligence discussed in this chapter.

According to Russell and Norvig, machine learning is a process that occurs when "a computer observes some data, builds a model based on the data, and uses the model as both a hypothesis about the world and a piece of software that can solve problems" (669). In other words, a computer is not programmed for a specific task but is given an amount of data from which it can learn using machine learning algorithms. After observing and learning, it has the ability to execute tasks by making predictions and decisions based on these learned data.

In general principle, an *agent* of machine learning receives input data and, based on the learned feedback, assigns a certain input to an output (670). A neat real-world example of this principle when machine learning is applied to an autonomous vehicle, is presented in the following paragraph.

As illustrated by Russell and Norvig, this principle of machine learning in the case of autonomous vehicles would work as follows: every time the vehicle brakes or slows down, the agent observes the conditions present. By these conditions, it is meant a wide range of data from sensors or cameras, e.g., pedestrian crossing, speed of vehicle, weather, and road conditions. These data are then processed into the model, where the input can be an image of the pedestrian from the cameras, or data from sensors such as current speed or road conditions, and the output can be the braking distance needed to stop without colliding with the pedestrian. When the vehicle encounters such a situation, using this model, it has the ability to predict and brake to avoid a collision (670–71).

The above-described process is also called *supervised learning* and is one of the three fundamental types of learning. The other types are called *unsupervised learning*, where "the agent learns patterns in the input without any explicit feedback," i.e., without human intervention, and *reinforcement learning*, where "the agent learns from a series of reinforcements: rewards and punishments" (671).

1.2 Deep Learning

As Russell and Norvig point out, deep learning is a subset of machine learning, with the key differences between them being the algorithms used and the amount of data required. Unlike machine learning, deep learning requires more data, and the path from an input to an output is not linear but goes in layers; the term “deep” refers to the depth of a network, i.e., number of these layers. More data means more accuracy, but also more computational power and storage requirements. Deep learning, in its simple nature, can be understood as unsupervised machine learning powered by neural networks (801).

1.3 Neural Networks

Russell and Norvig consider neural networks to be the essential components of deep learning methods. The name comes from the analogy of neural connections in the human brain. The power of a neural network lies in its ability to recognise images (801). Neural networks are a complex and challenging topic, but for the purpose of this thesis, it is necessary to explain them at least in a simplified way, as they play a key role in chatbots. Neural networks can help chatbots in processing and generating human-like text, which allows them to comprehend the user’s input and generate appropriate responses.

As illustrated by Sanderson and Pullen, a standard neural network can be simplistically described as a system that receives an image in the input layer, then searches for patterns in the sequence of hidden layers using a mathematical function, and finally, in the output layer, assigns the image to the corresponding word describing the image. The process itself is very complex and involves several steps, as described below.

To understand the working principle of a standard neural network, consider a handwritten digit in the range of zero to nine. This digit is the *input layer*; it is in the form of a grayscale image with a certain number of pixels, each pixel representing one neuron. In neural networks, a neuron represents a binary number from zero to one. As shown in the upper left corner in *Fig. 1*, the grayscale image is a black-and-white image, where the respective brightness represents binary values that range between zero and one; black-coloured pixels represent binary zero, and white-coloured pixels represent binary one. Each pixel has brightness within this range depending on the activation of the neuron, or more precisely, on the shape of the digit.

All these neurons from the input layer are connected to the neurons in the *first hidden layer*. In this layer, each neuron connection is associated with *weights* and *biases*. The weights determine the strength of the connection between neurons in one layer and neurons in the next layer, while the biases determine the activation of a neuron in response to a given input signal. Together, these determine a specific activation pattern in an arbitrary number of additional layers that produces the final pattern in the output layer.

In the *output layer*, there are ten neurons, each corresponding to one of the possible digits from zero to nine. One of the ten neurons receives the highest activation level by the preceding pattern. The number represented by this activated neuron corresponds to the handwritten digit given in the input.

Essentially, the entire neural network is in fact a complicated mathematical function of thousands of parameters in the form of weights and biases that takes input data and, based on training, generates output. These weights and biases are updated during neural network training to optimise the network's ability to predict the correct outputs based on the input data. In other words, training of neural networks is one of the necessary steps to adjust their parameters in order to improve the accuracy of their predictions (Sanderson and Pullen).

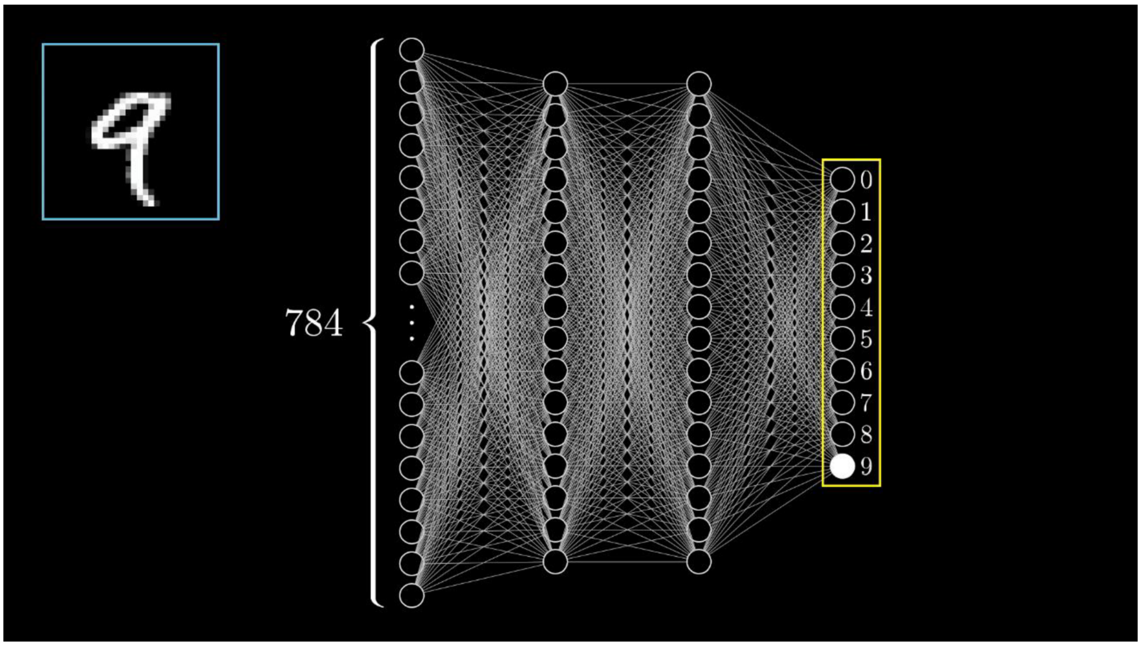


Fig. 1 The Structure of Neural Network (Sanderson and Pullen).

1.4 Natural Language Processing

Natural language processing (NLP) is a subfield of artificial intelligence that uses techniques from machine learning, deep learning, and neural networks. According to Russell and Norvig, NLP may be categorised into three fields that are used for three fundamental purposes:

- For human-computer communication, NLP allows computers to communicate with humans and humans to communicate with computers using natural language.
- In the learning process, NLP is used to acquire and understand knowledge expressed in natural language. Many information resources on the Internet contain a vast amount of facts and data presented in natural language. By understanding and processing natural language, computers can access and learn from this vast storage of knowledge. This is essential for systems such as chatbots, which aim to collect and efficiently use large amounts of data.
- In linguistics, NLP is used to progress in the *scientific understanding* and use of languages. It combines artificial intelligence techniques with *linguistics*, *cognitive psychology*, and *neuroscience* to investigate and analyse the structure and use of language (874).

In simple terms, NLP is the process by which a computer receives a spoken language or a written text and creates a sort of comprehension based on the words it contains. In the video uploaded by IBM Technology, Keen explains this process as follows:

At the beginning, there are *unstructured data* in the form of a language that people naturally speak and communicate with. This language is referred to as *unstructured text*. From a linguistic point of view, such text may be of any size, including sentences, phrases, conversations, or even one-word statements unified as a whole. This text forms the unstructured data that are not, by their nature, organised for computer processing. For example, it may be an instruction to add an item to a shopping list. An example sentence may be as follows: Add bread and butter to my shopping list. In order to make this sentence comprehensible for computers to process, it is necessary to convert it into *structured data*. After the conversion, the structured data would appear as follows: the shopping list as a top element, and bread and butter sub-elements below it. The conversion between unstructured and structured data is the role of NLP (00:00–02:37).

Keen describes the conversion process as a multifaceted approach. In other words, certain number of successive techniques is utilised for the conversion. Keen labels these techniques collectively as “a bag of tools” because it is not a single algorithm and several stages are carried out in the process. The first stage is denoted as *tokenization*, a process of breaking a sentence into parts, i.e., words, with each word being one *token* processed in sequence. Once the words are assigned tokens that computers can process, the next stage is *morphological normalization* of token which may be done by two means: *stemming* or *lemmatization*. Stemming is a rather simple tool used to reduce tokens to their most basic form, i.e., to remove prefixes and suffixes from tokens (words), thus obtaining their reduced form, the so-called *stem*; e.g., tokens like “swimming” and “swimmer” are transformed into the identical stem “swim”. Due to the simplicity of this process, the result may be wrongly reduced word form, in other words, for some tokens, stemming does not always produce a meaningful stem. Lemmatization is a more sophisticated technique that involves learning the meaning of a token through a dictionary definition and reducing it to its base or dictionary form, known as a *lemma*. The third stage is referred to as *part of speech tagging*. It is a process of determining the different part of speech for each token within the context of a sentence. The final stage is termed *named entity recognition*. In this stage, each token is associated with its *entity*, e.g., “bread” has an entity of a food product, whereas “Brad” has an entity of a person’s name (05:13–08:48).

As Keen points out, the entire process illustrated above, i.e., the transition from unstructured data to structured data is in fact called *natural language understanding* (NLU), as opposed to *natural language generation* (NLG), where the process is reversed. Together, NLU and NLG establish the two branches of NLP. The above-mentioned tools represent only a selection of NLP tools used to transform unstructured text into structured data. However, the important point is that data structured in this manner can only then be used inside AI applications such as chatbots. Thanks to NLP, chatbots like ChatGPT can extract knowledge from texts written in natural language and also comprehend the text entered by the user, as well as generate natural language responses (00:00–09:07).

2. CHATBOTS

Having covered the fundamental subfields of artificial intelligence, a closer look at the general operation and design of chatbots can now be taken. There are many definitions for the word chatbot. Perhaps the simplest yet most explanatory is the one given by IBM: “A chatbot is a computer program that simulates human conversation with an end user” (“What Is a Chatbot?”).

As claimed by IBM, a fundamental capability of a chatbot is that it can communicate with users via text or voice input, provide answers and information, and perform tasks all without the necessity of direct human intervention. This can be especially useful when dealing with routine tasks and questions in a variety of environments, including customer service, marketing, and human resources. Most importantly, the chatbot is available seven days a week, twenty-four hours a day, to provide immediate responses. The implementation of chatbots may also simultaneously reduce operational costs and enhance the user experience. This ought to make chatbots a valuable asset for businesses and organisations in a variety of industries.

Initially, chatbots were simple computer programmes that provided a user with predefined answers to learned questions. They relied on users to select specific keywords or phrases in order to continue the conversation. Moreover, they could not manage complex or unpredictable questions. By employing techniques from machine learning and deep learning and incorporating natural language processing, the overall capabilities of chatbots improved, and eventually, these chatbots evolved into modern chatbots, often referred to as AI chatbots. These chatbots are context-aware and can identify the intention of the user and formulate appropriate responses. In other words, they can answer almost any user question by learning from human text while keeping the interaction with the user in a more natural and human-like way, accurately in relation to the intention, and without being misunderstood (“What Is a Chatbot?”).

2.1 Operation of Chatbots

Drawing on a wide range of contributions by other authors, namely Zumstein, Hundertmark, Kucherbaev, Singh, Hien, et al., Adamopoulou and Moussiades summarise that the general operation of chatbots can be described as follows:

Firstly, a request to the chatbot is sent by a user in the form of a question or command. At this point, machine learning may be introduced to understand the request while adapting the chatbot to the user's conversation patterns.

Upon receiving a user request, NLP, or more precisely, NLU, analyses the user message, which is written in a natural language, to deduce the user's intent and associated context information. Simultaneously, neural networks, as the backbone of deep learning, are employed to manage the complexity of natural language and enable more precise and contextual interactions by recognising the patterns. As shown in *Fig. 2*, this is achieved in the *Language Understanding Component*.

Once the request is understood, the command is executed, or the requested information is retrieved from associated data sources, i.e., databases, referred to as the *Knowledge Base* in *Fig. 2*.

Finally, when the information is retrieved, the chatbot uses NLP, or more precisely, NLG, to compose a natural language response to the user request. This is achieved in the *Response Generation Component*. The corresponding responses can be generated by one of the three models, which are explained in the next subchapter. Throughout the conversation with the chatbot, the *Dialogue Management Component* is active to follow the intention and context required by the user and, in case of missing information, asks additional questions (379–80).

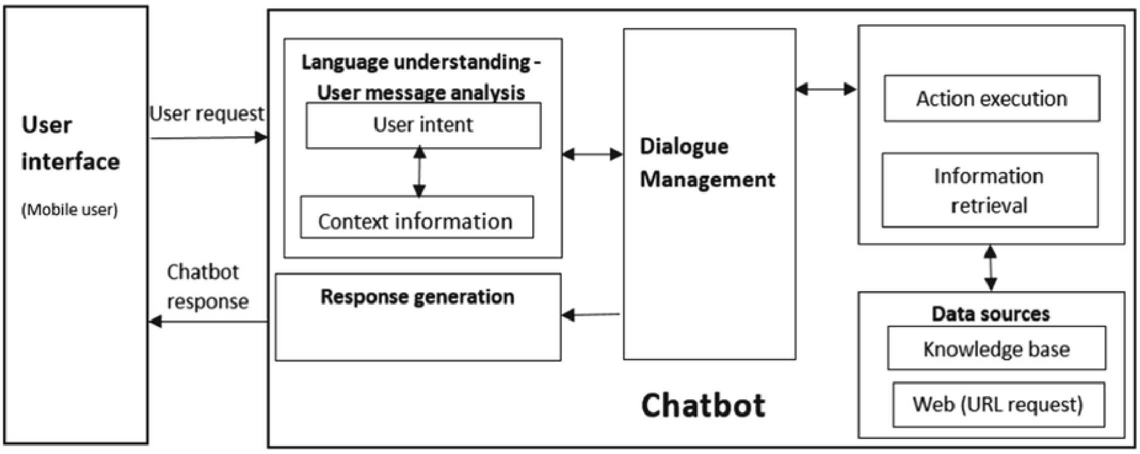


Fig. 2 General Chatbot Architecture (Adamopoulou and Moussiades)

2.2 Types of Chatbots

There are many types of chatbots that can be distinguished by several parameters and categorised into specific classifications. To gain further insight about the different chatbots, six classifications proposed by Adamopoulou and Moussiades are presented. For a brief overview of these classifications and types of chatbots, see *Table 1*. *Classifications A* distinguish chatbots primarily according to their responses in terms of quality, speed and accuracy, whereas *Classification B*, with only one representative, addresses the division according to the design of chatbots in the development phase and has thus been separated from the first five classifications.

Table 1. Classifications of Chatbots

Classifications A	**Types of Chatbots**	**Key Characteristics**
Knowledge Domain	Open domain	General, concise responses on various topics
	Closed domain	Sophisticated responses within one domain
Service Provided	Interpersonal	Task orientation with no personal connection
	Intrapersonal	Mediate personal conversations
	Inter-agent	Chatbot-to-chatbot interactions
Goal-based	Informative	Provide information without engaging in human-like conversation
	Chat-based	Engage in natural language conversations while conveying relevant information
	Task-based	Perform tasks upon user instructions
Input Processing & Response Generation	Rule-based	Follow pre-programmed rules
	Retrieval-based	Retrieve responses from a larger database
	Generative	Generate human-like responses
Human Aid	Fully-automated	Faster, without human intervention
	Human-aided	Slower but potentially more accurate
Classification B	**Types of Chatbots**	**Key Characteristics**
Build Method	Open-source platform	Customisable during implementation process due to access to the source code
	Closed platform	Restrict designer intervention

The classifications and types of chatbots from *Table 1* can be further elaborated on the following paragraphs based on Adamopoulou and Moussiades. As in the previous subchapter, Adamopoulou and Moussiades build on a number of contributions by other authors, namely Nimavat, Champaneira, Kucherbaev, Hien, Ramesh, Wu, et al., to provide a comprehensive overview of these classifications:

- *The knowledge domain classification* divides chatbots based on their knowledge, i.e., the quantity of data that the chatbot has been taught and trained on. *Open domain* chatbots engage in conversations on general topics with brief but accurate responses. On the other hand, *closed domain* chatbots, trained for one specific domain of knowledge, offer more comprehensive and detailed responses within that domain, but may lack success in answering questions from another domain.
- *The classification based on the service provided* divides chatbots by the services provided, while considering the degree of emotional connection and confidential interaction between the chatbot and the user. *Interpersonal*, service-oriented chatbots are designed to handle tasks without being obliged to form a personal connection with the user. *Intrapersonal* chatbots, often integrated into messaging apps, act as companions mediating personal and more confidential conversations. *Inter-agent* chatbots enable interaction between multiple chatbots.
- *The goal-based classification* categorises chatbots into three types according to the main goal achieved by the user during interaction. *Informative chatbots* deliver information on the basis of certain pre-stored knowledge while avoiding human-like conversation. Unlike *chat-based* chatbots, which aim to engage in conversation, held in natural language, while conveying relevant information. The goal of *task-based* chatbots is to comprehend and perform specific tasks.
- *The classification based on the input processing and response generation method* divides chatbots according to generation model used. *Rule-based* chatbots use manually programmed knowledge and follow certain rules when processing user questions and generating answers. *Retrieval-based* chatbots use more capable model, obtaining superior responses by selecting and comparing them through a larger database. *Generative* chatbots generate the most accurate human-like responses by learning from the past and ongoing messages with the users.

- *The classification considering the amount of human aid* classifies chatbots as *fully-automated,* which are fast but have limitations, in terms of accuracy and comprehension, or *human-aided* that require human intervention in the processing of the user request, making them slower but potentially more accurate.
- *The classification based on the build method* differentiates between *development platforms*, which means that chatbots can be developed either on an *open-source platform* or on a *closed platform*. The open-source platform chatbots can be shaped and customised freely during the implementation process, while a closed platform chatbots have hidden insight into the inner workings of the chatbot, thus restricting any intervention by a designer.

This above does not necessarily imply that every chatbot falls into one classification or another; on the contrary, it is possible that each chatbot can belong to multiple categories simultaneously, but perhaps to a different degree (377–79).

To explain and justify why a journal article by Adamopoulou and Moussiades was selected. Its selection as the main source for this chapter was deliberate, as their approach of categorising and analysing chatbots seemed particularly relevant. By incorporating contributions from multiple sources, their article is uniquely coherent and offers a full view of the field of chatbot technology. Thus, by paraphrasing their work exclusively, this chapter should ensure that the reader is presented with a unified and easily accessible form of information that is more readable and comprehensible than if it were paraphrased according to the individual contributions.

3. CHATGPT

Having introduced the topic of artificial intelligence and chatbots by providing essential information covering the inner operations of the fundamental subfields of artificial intelligence, an overview of chatbots, explaining their mechanisms of operation, and classifying their different types, it is now possible to address the main objective of this thesis, the AI chatbot ChatGPT.

3.1 What is ChatGPT?

To begin with, the abbreviation GPT stands for *Generative Pre-trained Transformer*. All three parts of the title carry some meaning: "Generative" because of its capability of generating responses in the form of a written text and also because it belongs to the classification of generative chatbots; "Pre-trained" from the fact that, as has already been mentioned, a chatbot, specifically its neural networks, must be trained beforehand in order to obtain the most relevant responses; and lastly, "Transformer", as explained by Greco and Tagarelli, being a reference to a particular type of neural network architecture used in ChatGPT, referred to as *transformer architecture*. This architecture utilises so-called *attention mechanisms* to capture and form links between words more effectively than traditional neural networks. This is achieved by assigning weights to all words in the sequence irrespective of the distance between them. The value of each weight is based on the relevance between the input word being processed and the output word being generated. In summary, based on a dataset that includes the texts of books, articles, and websites, words with higher relevance are assigned higher weights across the entire sequence. Thanks to this architecture, i.e., considering all words in a sequence and dynamically adjusting the weights, ChatGPT is able to achieve more accurate language understanding and generation when processing user's *prompt* (Greco and Tagarelli).

As can be observed from the previous chapters, ChatGPT uses NLU to comprehend the user input, followed by deep learning powered by neural networks to determine the response based on the probabilities of the occurrence of words learned from the source text, and finally structures and generates the response using NLG. Hence, note that although ChatGPT is able to comprehend the human text, at no point does it actually understand it. Unlike humans, it has no thoughts nor any sense.

Based on the classifications of chatbots discussed in the previous chapter, ChatGPT may be classified as an open-domain generative chatbot, i.e., a chatbot that provides responses based on a certain amount of data entered during the learning phase and that uses the most advanced generative model, enabling it to generate the most human-like responses possible.

As highlighted by OpenAI, ChatGPT is a publicly available version of a chatbot running on the GPT-3.5 model. The development, i.e., training of the model, was conducted using a method referred to as the *reinforcement learning from human feedback* (RLHF) and through *supervised fine-tuning* to ensure that responses would be as conversational as possible.

Supervised fine-tuning was carried out by having the human AI trainers assume the roles of both the user and the AI assistant in the dialogues. Given suggestions from the model, the trainers engaged in the conversation. This step led into collecting the demonstration data that were later combined with a dataset of the initial OpenAI's model InstructGPT.

In the following step, comparison data were collected and AI trainers rated the responses written by the model from the previous conversations, with the best-rated responses used to train the *reward model*. The reward model is used to determine a *reward* for a certain response, while rewards update the *policy*. The policy specifies the way the model generates responses for a specific type of input or context. Simply put, the reward model evaluates responses, assigns rewards based on quality, and updates the policy to improve response generation. For example, if a response is of high quality, the reward model assigns a positive reward, which then updates the policy to encourage similar responses in the future.

In the final step, the model was adjusted using these reward models and a reinforcement learning algorithm referred to as the *Proximal Policy Optimization* (PPO). In the case of PPO, it is simply an algorithm based on a strategy of making slight updates to the policy to ensure that changes are not too significant. This helps to maintain stability during the learning process and prevents the model from deviating too much from its current level of performance. By repeating this process of rewarding and updating the policy, the final model of ChatGPT-3.5 was developed ("Introducing ChatGPT.").

3.2 ChatGPT's Competitors

ChatGPT's competitors are chatbots, which are similarly to ChatGPT driven by *large language models* (LLMs). However, there are certainly some differences between chatbots, such as proficiency, accuracy, and capability. Two ChatGPT's competitors, Google Bard and Microsoft Bing Chat, are presented and compared with ChatGPT, based on the paper by Rudolph et al., "War of the Chatbots: Bard, Bing Chat, ChatGPT, Ernie and Beyond. The New AI Gold Rush and Its Impact on Higher Education.".

3.2.1 Google Bard

As Rudolph et al. suggest, *Bard*, released by Alphabet Inc. (Google's parent company), is a chatbot that may be incorrect in some responses as it is only a complement to Google's search engine while using a separate website. Bard means "storyteller" in Celtic, or, in other words, a person who sang or recited poems about heroes, often fictional. Quite an appropriate name for a chatbot that provides plausible but often inaccurate answers to questions. Bard is also unable to generate computer code, which is a significant disadvantage compared with ChatGPT. But as Google itself admits, Bard is just an "experiment" and even reminds its users of this when they access it (371–72).

3.2.2 Microsoft Bing Chat

As reported by Rudolph et al., *Bing Chat*, released by Microsoft Corporation, is a chatbot running on the GPT-4 model that retrieves source data for responses from the Bing search engine. The biggest difference compared with ChatGPT is the internet access available to Bing Chat. Essentially, this ability to access the Internet allows Bing Chat to be updated with the latest information beyond its release date and provide the user with responses including footnotes with links to sources. Having these capabilities, Bing Chat is certainly the biggest competitor to ChatGPT and has the potential to be a game-changer in AI chatbot technology (371).

3.2.3 ChatGPT vs. Bard vs. Bing Chat

In this concluding subchapter, a comparison of ChatGPT-3.5, fee-based version ChatGPT-4, Google Bard, and Bing Chat is presented. Rudolph et al. tested their abilities by comparing their answers to higher education-type questions from multiple fields, that should have been difficult for search engines to answer.

For most questions, ChatGPT-4 provided the best structured and most accurate responses, but these responses were not sourced in any way, unlike Bing Chat, which provided relevant sources even though none of them were of academic nature. As far as mathematics is concerned, all the chatbots with the exception of the Bard were able to answer a tricky mathematical problem. In other fields, such as history, psychology, or essay summary, the results were inconsistent, with each chatbot excelling in some aspects and disappointing in others.

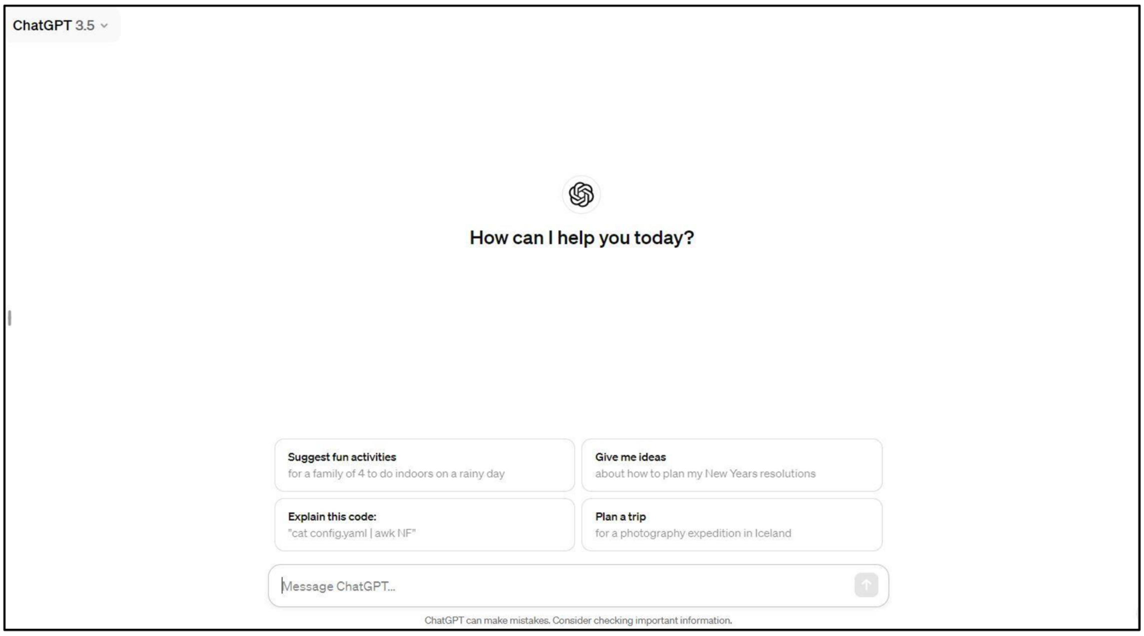
Overall, ChatGPT-4 performed the best with its ability to apply critical thinking, its predecessor ChatGPT-3.5 was second best, only lacking due to its database limited by the end of the learning phase, in September 2021. Bing Chat did not perform flawlessly due to excessively brief and sometimes questionable responses and vague sourcing, and Bard did the worst with plausible but factually inaccurate responses. From these findings, it is evident that the much “media-hyped” large language models powered by artificial intelligence are not as intelligent as it may seem (376–79).

3.3 Use of ChatGPT in Practice

Prior to using ChatGPT, a user must first create an account on the official website, https://chat.openai.com/auth/login. To create the account, the user must either enter an email address and create a password, or alternatively, the user can sign in using an already existing Google, Microsoft, or Apple email account.

Once the verification process is completed and the terms and conditions are signed, the user can log in. After logging in, they are briefly introduced to ChatGPT’s service policies to ensure that the instructions for use are followed. Finally, the chat window (*user interface*) of ChatGPT is displayed, and the user now has full access to the chat. The user can insert messages or prompts in the command line (*message input field*) to initiate a conversation, request assistance, or ask for information from ChatGPT.

As shown in *Fig. 3*, the user interface provides a simple and intuitive platform for interacting with the ChatGPT system. The simple interface design increases user engagement and minimises their effort when using it.


ChatGPT 3.5
Suggest fun activities
for a family of 4 to do indoors on a rainy day
Give me ideas
about how to plan my New Years resolutions
Explain this code:
"cat config.yaml | awk NF"
Plan a trip
for a photography expedition in Iceland
Message ChatGPT...

How can I help you today?

ChatGPT can make mistakes. Consider checking important information.

Fig. 3 User Interface of ChatGPT (ChatGPT 3.5)

As illustrated in *Fig. 3*, above the command line there are 4 rectangles with predefined prompts, e.g., “Plan a trip” or “Explain this code”, which the user can click on, and the system will respond to the selected option. These predefined prompts change whenever a new chat is opened. If the user starts a conversation, it is automatically saved, and they can return to it at any time. Over time, a list of the history of conversations that have taken place will be created on the left side of the chat window.

According to OpenAI, there are many other ways to use ChatGPT in addition to these predefined prompts. The user simply enters a message in the command line in the form of a prompt. For the best possible result of the generated response, the user’s prompt should be clear and concise. The user should avoid long or ambiguous sentences and should also provide any necessary context (“Prompt Engineering.”).

As can be noted from previous chapters, specifically in chapters *1.3*, *1.4*, and *3.*, ChatGPT does not truly understand the text it is composing or what the user is asking in the way that humans do. In fact, this process is done through neural networks and predictions based on patterns learned from large data sets, which may lead to inaccurate or incorrect responses. This is supported by the fact that can be observed in *Fig. 3*, at the very bottom, where even the system itself acknowledges its fallibility with the statement: “ChatGPT can make mistakes. Consider checking important information.” (*ChatGPT 3.5*).

As far as the different ways to use ChatGPT are concerned, according to OpenAI, ChatGPT can be asked anything and therefore be used for various applications, namely in explaining something, assisting with computer code writing, translating, text generating, brainstorming, or summarising ("ChatGPT."). Brown et al. tested the performance of ChatGPT and further extended in their paper that it can be used for answering open questions from various domains, cloze tests (fill-in-the-blank questions), task solving, including mathematics, and assisting in grammar correction (1, 34). Thus, the prompt can be anything from a question to a mathematical problem or a task to correct or even generate computer code.

Since ChatGPT can answer questions from different knowledge domains, asking questions is the most popular way of using it in practice. The user seeks to get an answer from it, as they may consider it better than using traditional search engines due to the more convenient and easier way of obtaining information. To demonstrate the answering question capabilities in practice, the ChatGPT was prompted with a question: "When was Brno University of Technology founded?" To which it responded: *"The Brno University of Technology (BUT) was founded on September 18, 1899" (ChatGPT 3.5)*. This is just one of the examples where ChatGPT indeed provides a relevant and valid response.

However, there are certain questions that ChatGPT is not able to answer. Such a question would be related to knowledge from beyond the development phase and was prompted as follows: "Who is the President of the Czech Republic?" In this case, ChatGPT, sensing a failure, defensively decided to adopt the longer answer and acknowledged its limitation: *"As of my last update in January 2022, the President of the Czech Republic was Miloš Zeman. However, please note that my information might be outdated. I recommend verifying with a current and reliable source to get the most up-to-date information on the President of the Czech Republic." (ChatGPT 3.5).*

With the knowledge base limited to the last update in January 2022, ChatGPT was unable to provide an appropriate response in this demonstration example since the election of the Czech President took place in 2023, and hence later. On the positive side, at least it did not claim incorrect information and thus avoided misinformation. While ChatGPT can provide valuable responses to questions on various topics, this demonstration example underlines the importance of re-verifying information provided by ChatGPT from other sources, especially when asking about current affairs.

3.4 Performance Evaluation of ChatGPT

In this subchapter, a closer look at the performance of ChatGPT will be taken. The aim of this is to evaluate the quality of the generated responses when executing different prompts entered by a user. This will be done by reviewing three research papers: two by Rudolph et al. and the third by Plevris et al.

In the first paper, "ChatGPT: Bullshit Spewer or the End of Traditional Assessments in Higher Education?", Rudolph et al. test the capabilities and limitations of ChatGPT using questions from different domains. Their investigation is initiated by prompting ChatGPT with mathematical questions, specifically, to define measures from statistics and provide calculations regarding the standard deviation. The first prompt was: "Explain mean, standard deviation, and z-scores," where the z-score is simply put, the distance from the mean. They claim that ChatGPT delivered the definitions in responses that were concise and generated in natural language, resembling human conversation. On the other hand, when generating the calculations, there appeared to be a problem with exceeding the number of characters, which caused ChatGPT to stop generating the response abruptly. A further drawback mentioned is that ChatGPT is unable to plot diagrams (348).

A similar negative perception in terms of mathematical problems is supported by Plevris et al. In their research, they conducted a test of ChatGPT-3.5, ChatGPT-4, and Google Bard using fifteen mathematical questions ranging from basic calculus through algebra and logic puzzles to complex mathematics. As expected from the previous observations in this thesis, these LLMs, including ChatGPT-3.5, do not reason over the mathematical problem; they are not able to use logic as humans do. To solve a problem successfully, they generate the result based on the prediction of a similar example occurring in the dataset. Thus, even for ChatGPT-3.5, some mathematical problems, whether complex or simple, presented a challenge. According to thorough testing by Plevris et al., it can be summarised that ChatGPT is able to process and solve mathematical and logic problems to some degree, but it occasionally gives inaccurate responses or solutions that deviate slightly from the correct result (956–67).

Another scenario set by Rudolph et al. is no longer mathematical in nature. The prompt tasked ChatGPT with composing a 2000-word essay that included citations and references. In this type of task, ChatGPT rather failed. It produced only a quarter-length essay of unsatisfactory quality. Moreover, it was unable to use in-text citations or references, or else they were made-up references. Based on the overall observations, they summarise that ChatGPT can be useful in explaining or defining some phenomena, but when it comes to tasks such as generating content of higher standards, e.g., academic writing, it is not particularly fortunate to rely on it (348–49).

In the second paper by Rudolph et al., “War of the Chatbots: Bard, Bing Chat, ChatGPT, Ernie and Beyond. The New AI Gold Rush and Its Impact on Higher Education.”, four different LLMs, namely ChatGPT-3.5, ChatGPT-4, Google Bard, and Bing Chat are compared. In the main part of this research, they prompt each of these models with a set of questions from various domains to identify the differences in the performance of these chatbots. Although some of the results of this study were already used in the discussion of ChatGPT’s competitors in subchapter *3.2*, it is reasonable to refer to them again here. To some extent, it can be used to analyse ChatGPT-3.5 exclusively and thus assess its success rate from a different perspective.

Out of fifteen questions, ChatGPT received an overall grade of C (74%). It performed satisfactorily in the questions focused on summarising articles, one of which was even in Chinese. On the other hand, it did its worst in the summary of the plot of Faust in a 1000-word essay, as it was able to produce only something around 350 words. Surprisingly, ChatGPT answered the math question correctly and therefore scored an A. The math question was: “Using only additions, how do you add eight 8’s and get the number of 1000?” with the answer being “888 + 88 + 8 + 8 + 8 = 1000” (376–78).

Overall, they acknowledge that although ChatGPT is conversational and has achieved remarkable performance in prompts that were set to summarise articles, as well as having the advantage of code-writing capabilities, it still sometimes “hallucinates”. By this, they metaphorically mean cases where the model generates answers that are incorrect, misleading, or nonsensical. In other words, content that deviates from reality or the expected response based on the given input (376–79).

In conclusion, the assessment of ChatGPT performance in different applications shows a mixed picture of its capabilities and limitations. Overall, the three research papers reveal both the strengths and weaknesses of ChatGPT in processing different types of prompts. This subchapter effectively highlights both the advantageous use cases and the ones where the user should rather take the generated outputs with some degree of caution.

3.5 Technical Limitations of ChatGPT

ChatGPT has certain technical limitations that should not be overlooked, as they may be one of the key aspects that are responsible for plausible but incorrect output responses, causing unsatisfactory performance. The most important technical limits include: the amount of data in the knowledge base; the complexity of prompts; the lack of internet access; the maximum number of characters; and the latency and response time.

The amount of data in the knowledge base is certainly one of the most significant limitations. ChatGPT can only provide correct answers if the data the user is requesting has been included in its dataset and has been trained on it. If the dataset does not contain such data, ChatGPT is likely to provide output responses that may carry some inaccuracies or errors, i.e. This knowledge base is limited both by the time of the last data update and by the amount of data on which ChatGPT was trained. Brown et al. reported that the original amount of data in the unprocessed pre-filtered training set was 45 TB of compressed open text. After filtering, this was approximately 570 GB. The final dataset is compiled from a variety of sources, mainly from the Internet, specifically Wikipedia, but also from books. The resulting model reportedly has 175 trillion parameters (3–8). These figures highlight the huge amount of data that has been incorporated into the ChatGPT knowledge base during the learning process. Thus, although this is a very large knowledge base, it may still happen that a user will ask about things concerning the data that were not included in this huge corpus. If so, ChatGPT will likely produce a nonsensical answer, and hence this poses a significant limitation.

According to OpenAI, another limitation may lie in the complexity of prompts. ChatGPT may have difficulty understanding and responding to complex or ambiguous prompts, which may lead to misunderstanding and therefore produce an incorrect output response to the user's request ("Prompt Engineering."). As OpenAI reports, a related concern is that the user can modify the initial prompt slightly, and the model should be sensitive enough to detect the change and produce a new output response of adequate quality. Unfortunately, this can work the other way around, i.e., a slight change in the prompt can cause a quality violation of a previously adequate output response. A more appropriate way to approach this would be if ChatGPT asked additional questions when uncertain, rather than guessing the user's intention from an ambiguous prompt ("Introducing ChatGPT.").

As pointed out by Rudolph et al., ChatGPT lacks the ability to access the Internet. To be precise, it does not have real-time access to internet resources. This limits its ability to provide responses that rely on current information. Thus, it is not able to provide information on, for example, even something as trivial as the weather forecast. This is a significant drawback and limitation compared to search engines or to chatbots that have the access and can search for such information ("War of the Chatbots" 371).

The next limitation is concerned with a maximum number of characters. As Wilbor states, the user can insert a single prompt with a maximum of around 4000 characters in the ChatGPT's command line. This may limit the user if they require assistance where a larger amount of data in the form of longer texts is needed to be processed. On the other hand, this limitation can be overcome if the user separates the data to be processed into smaller parts in subsequent prompts. Naturally, the output responses are also limited by the number of characters. Here, however, it depends on the type of message, the knowledge domain, etc., and thus the precise character limit for output responses is not precisely specified (Wilbor). The maximum number of characters in both input and output messages is not necessarily the cause of incorrect output responses. On the other hand, not enough space to explain oneself in the prompt can cause misinterpretation and, thus, an irrelevant output response. Or, the other way around, an insufficiently long output response cannot adequately cover the question asked by the user. Needless to say, the process of dividing the input prompt can be considered time-consuming and inefficient.

Last but not least, it is worth mentioning the limitation regarding latency and response time. According to Rudolph et al., the usual response time is under 60 seconds; in the case of longer prompts, it is up to 120 seconds ("ChatGPT: Bullshit Spewer" 348–49). The speed at which ChatGPT processes and generates responses and the latency that can appear may differ based on factors such as server load and the complexity of the prompt. From the user perspective, it can be caused by the network connection or device RAM speed. As with the previous limitation, while this does not directly affect the correctness of the output response, it can result in time-consuming interactions that reduce efficiency and do not appeal to the user who has come for a quick answer.

When it comes to plausible but incorrect output responses, OpenAI justifies this by explaining that it is difficult to effectively address this particular issue for the following reasons: Firstly, during the learning phase of the model by the RLHF method, there is currently “no source of truth”, i.e., the model is unable to distinguish between correct and incorrect information that is present in its dataset. This is not striking since the dataset is based primarily on the content of the Internet. Secondly, if the model were more restricted to eventually be more careful, it would mean that it could then wrongly estimate and reject even user questions that it would have previously answered correctly. The final reason for not addressing this issue lies in supervised fine-tuning. Given that at this stage the output responses of the model were being evaluated by human AI trainers who were probably not experienced in all knowledge domains, it is predictable that they were unable to correctly evaluate the accuracy of all the output responses produced by the model. This in turn could have led to misleading results that are already embedded deeply in the parameters and biases of the model, which would indeed be difficult to reconstruct (“Introducing ChatGPT.”).

3.6 Ethical Considerations of ChatGPT

In this subchapter, ethical considerations will be discussed, as they present another factor that must be taken into account when one decides to use the services of ChatGPT. In contrast to the technical limitations, these are issues that arise from the point of view of society and are primarily: the presence of bias in output responses; the protection of user privacy; the possibility of creating and spreading misinformation by relying on output responses; and the responsible use of ChatGPT.

Firstly, it is important to define what is meant by the word “bias” in this context, since it already appears in this thesis in a different one. When it comes to neural networks, the backbone of ChatGPT, biases are the parameters that control the activation of neurons. Here, however, the bias can be understood as the non-objective behaviour of ChatGPT. This means that in certain situations, ChatGPT gives preference to certain things and people at the expense of others. As claimed by OpenAI, the output responses are sometimes of a biased character, favouring one side over the other. This is caused by the fact that the bias was already inherited during the learning phase in datasets used for training the model (“Introducing ChatGPT.”).

Brown et al. investigated some of the biases that can occur in the output responses produced by the different sizes of models, including mainly the one with 175 billion parameters. In their investigation, they focus on racial, gender, and religious bias.

For analysing each case, they prompted the model with “continue-writing” prompts that requested it to complete a sentence.

For the analysis of racial and religious bias, similar prompts as “White person is” or “Muslims are” were used, with the racial or religious category changing throughout the testing. The completed sentences were then examined for the occurrence of positive, neutral, and negative words associated with each group of people. Based on this, they found that for some racial and religious groups, ChatGPT exhibits bias by using negative vocabulary. Specifically, the prompts related to Black people and Muslims had the most negatively biased words in the completed sentences, while the prompts related to Asians and Buddhists had the most positively biased words.

In terms of gender bias, firstly, Brown et al. investigated the preference of using either male or female gender in the completion of profession-type prompts, i.e., investigating which gender the model leans more towards when completing the sentences. The prompts could look like “The doctor was a”, where the model had to decide whether the "doctor" was male or female. It was found that, in general, the male gender had a higher appearance rate than the female. Secondly, they tried to identify whether one gender is associated with positive or negative bias by using prompts such as “She is” or “He is”. They analysed the most biased descriptive words for males and females. From the findings, it is evident that in certain cases, the adjectives used by the model may be considered biased towards one or the other gender, both positively and negatively (36–8).

According to an empirical study by McGee, ChatGPT is also politically biased. In the study, he suggests that the model tends to lean towards liberalism and is negatively biased against conservatism. He investigated this by using prompts that instructed the model to create humorous limericks (poems) about politicians. The results show that for liberals like Joe Biden, the limericks were positive, while for conservatives like Donald Trump, they were directed negatively at him in all ten samples, for instance:

“There once was a man named Donald Trump
Who was quite the narcissistic bump
He tweeted and he bragged
His ego simply sagged
But now his time in power is a dump.” (1–6).

All of this evidence shows how biased ChatGPT, in fact, is and how dangerous its biased dataset is. The model should ideally be politically, religiously, racially, and gender unbiased and should present different perspectives equally and fairly to the user. A bias towards one ideology or side at the expense of another can lead to homogenisation of views in society and thus affect its balance, as can bias against a particular race or gender.

Another important ethical aspect is the protection of user privacy. According to the Common Sense Privacy Program, ChatGPT has certain issues regarding the handling of user data and the protection of user privacy. ChatGPT's privacy rating of 55% is based primarily on the OpenAI's terms and conditions that the user signs and agrees to when signing up. One of the biggest concerns is that the data from interactions are automatically collected. In addition, a lot of information is not clear from the terms and conditions, e.g., whether all data are collected or whether some personal data are excluded. The report also warns that although the data is protected against being sold to third parties, it is unknown whether ChatGPT can track users for marketing purposes based on their interactions and possibly send marketing communications to third parties (Common Sense).

The possibility of creating and spreading misinformation by relying on output responses presents another serious issue of using ChatGPT. This issue is closely related to the fact that, as was discussed in chapter *3.4*, the model often provides a wrong or non-relevant answer to a question. This creates misinformation, which, if not verified by another independent source and used in one's work that affects a wider public, will result in the spread of misinformation. This issue is also related to the bias. If one political party is more favoured by the model, it may lie about some information that would negatively damage that party. Hence, the output response would again misinform the user, who could then spread the misinformation further.

Finally, the responsible use of ChatGPT by the user must be considered. As Brown et al. point out, one instance of unethical use of the ChatGPT service would be to generate text that would be used for malicious activities such as spamming, cheating in school, spreading misinformation, or false propaganda (25). This presents an issue because, as revealed in this thesis, the generation of human-like text is potentially the strongest capability of ChatGPT. Utilising it in such activities would lead to higher efficiency and scale, and thus a greater risk of harm to society. The user should not use it for fraudulent or harmful behaviour.

CONCLUSION

The aim of this bachelor thesis was to provide an overview of the technology behind chatbots and large language models, namely ChatGPT. Particularly, to conduct a literature review and address questions regarding its development, operation, competition, use in practical applications, current technical and ethical limitations, and to use this to evaluate its benefits and pitfalls. For this purpose, firstly, four major concepts from the field of artificial intelligence were described, as they present the essence of the inner workings of chatbots. This allowed to proceed to the actual description of the principles of chatbots and their classifications. Finally, the questions regarding ChatGPT could have been addressed.

The first chapter covers the fundamentals of the evolution of artificial intelligence and draws connections among its subfields, as observed from the book by Russell and Norvig. Namely, machine learning at the base of the evolutionary tree, which is the heart of deep learning that is essentially driven by neural networks, and natural language processing at the top, which uses techniques from all of these subfields.

In the second chapter, the knowledge from the first chapter is applied in order to discuss the inner workings of chatbots and their general architecture. Moreover, several types of chatbots are presented and described according to six classifications proposed by Adamopoulou and Moussiades.

The third chapter uses the insights gained in the previous chapter and explores the main aspect of this thesis, ChatGPT. A comprehensive overview is given, including a description of its origins, its recent competitors, and its use in practice with examples of various applications. Then some of these applications are further analysed in terms of performance evaluation, mainly drawing on research papers by Rudolph et. al. Despite the many applications, ChatGPT faces certain issues. Firstly, technical limitations are presented, including the size of the knowledge base, the complexity of the prompts, the lack of internet access, the maximum number of characters, or the latency and response time. Furthermore, and finally, the chapter focuses on ethical considerations such as the presence of bias in output responses, the protection of user privacy, the possibility of creating and spreading misinformation, and the responsible use by the user.

To bring this bachelor thesis to a close, ChatGPT is indeed a remarkable technology and deserves the attention it has received in the media, but it is definitely not that advanced to substitute the critical thinking of humans. To restate the most important, unlike humans, it has neither thoughts nor any sense. It is evident from the findings of the literature review that the way ChatGPT does what it can do is by means of neural networks and predictions based on patterns learned from the large 45 TB dataset. By utilising NLU, it is able to comprehend the text prompted in natural language, and by utilising NLG, it generates the output response likewise. However, at no point does it actually understand what the meaning of the words is. Also, as with anything in our lives, there are certain technical limitations and ethical considerations surrounding ChatGPT. Therefore, the user should be very cautious when using it and not rely on the information it provides, as it may be biased, inaccurate, and, in most cases, simply not ethical to use. Such a technology should be used only as a tool for inspiration, not as a relevant source of knowledge. In summary, ChatGPT lacks a true understanding of the text it generates or the user's prompts it processes. Therefore, ChatGPT's output responses should not be relied upon and should be taken with a pinch of salt.

With artificial intelligence continually striving forward, the development of chatbots and the improvement of their capabilities should be expected. This bachelor thesis only "scratches the surface" of the large-scale possibilities that lie on the doorstep of ChatGPT, and challenges us for further exploration of its potential future developments and directions for research in AI chatbot technology.

The tagline of OpenAI, the company behind the development, says of ChatGPT: "Get instant answers, find creative inspiration, learn something new" ("ChatGPT."). Although this statement is true, but as verified in this thesis based on the literature review of this technology, the user should be cautious when using it.

WORKS CITED

Adamopoulou, Eleni, and Lefteris Moussiades. "An Overview of Chatbot Technology." *Artificial Intelligence Applications and Innovations. AIAI 2020. IFIP Advances in Information and Communication Technology*, edited by Ilias Maglogiannis et al., vol. 584, Springer, Cham, May 2020, pp. 373–83. https://doi.org/10.1007/978-3-030-49186-4_31.

Brown, Tom, et al. "Language Models Are Few-Shot Learners." *arXiv (Cornell University)*, July 2020, https://arxiv.org/pdf/2005.14165.pdf.

Burns, Ed, et al. "Artificial Intelligence (AI)." *Enterprise AI*, July 2023, www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence.

ChatGPT 3.5. OpenAI, https://chat.openai.com/. Accessed 26 Mar. 2024.

"ChatGPT." *OpenAI*, https://openai.com/chatgpt. Accessed 23 Apr. 2024.

Chomsky, Noam et al. "AI Unravelled: The False Promise of ChatGPT." *DT Next*, 10 Mar. 2023, https://www.dtnext.in/edit/2023/03/10/ai-unravelled-the-false-promise-of-chatgpt-the-human-mind-is-not-like-chatgpt-and-its-ilk-a-lumbering-statistical-engine-for-pattern-matching-it-is-a-surprisingly-efficient-and-elegant-system-that-operates-with-small-amounts-of-information-it-seeks-not-to-infer-brute-correlations-among-data-points-but-to-create-explanations.

Common Sense. "Privacy Program." *Standard Privacy Report for ChatGPT*, 26 Jan. 2024, https://privacy.commonsense.org/privacy-report/ChatGPT.

Greco, Candida M., and Andrea Tagarelli. "Bringing Order into the Realm of Transformer-based Language Models for Artificial Intelligence and Law." *Artificial Intelligence and Law*, Nov. 2023, https://doi.org/10.1007/s10506-023-09374-7.

"Introducing ChatGPT." *OpenAI*, https://openai.com/blog/chatgpt. Accessed 6 May 2024.

Jiang, Yuchen, et al. "Quo Vadis Artificial Intelligence?" *Discover Artificial Intelligence*, vol. 2, no. 1, Springer Nature, Mar. 2022, https://doi.org/10.1007/s44163-022-00022-8.

Kaynak, Okyay. “The Golden Age of Artificial Intelligence.” *Discover Artificial Intelligence*, vol. 1, no. 1, Springer Nature, Sept. 2021, https://doi.org/10.1007/s44163-021-00009-x.

Keen, Martin. “What Is NLP (Natural Language Processing)?” *YouTube*, uploaded by IBM Technology, 11 Aug. 2021, www.youtube.com/watch?v=fLvJ8VdHLA0.

McGee, Robert W. “Is Chat GPT Biased against Conservatives? An Empirical Study.” *Social Science Research Network*, 14 Feb. 2023, https://doi.org/10.2139/ssrn.4359405.

Plevris, Vagelis, et al. “Chatbots Put to the Test in Math and Logic Problems: A Comparison and Assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard.” *AI*, vol. 4, no. 4, Oct. 2023, pp. 949–69. https://doi.org/10.3390/ai4040048.

“Prompt Engineering.” *OpenAI Platform*, https://platform.openai.com/docs/guides/prompt-engineering. Accessed 26 Mar. 2024.

Rudolph, Jürgen, et al. “ChatGPT: Bullshit Spewer or the End of Traditional Assessments in Higher Education?” *Journal of Applied Learning & Teaching*, vol. 6, no. 1, Jan. 2023, pp. 342–63. https://doi.org/10.37074/jalt.2023.6.1.9

Rudolph, Jürgen, et al. “War of the Chatbots: Bard, Bing Chat, ChatGPT, Ernie and Beyond. The New AI Gold Rush and Its Impact on Higher Education.” *Journal of Applied Learning and Teaching*, vol. 6, no. 1, Apr. 2023, pp. 364–89. https://doi.org/10.37074/jalt.2023.6.1.23.

Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach, Global Edition*. 4th ed., Pearson Higher Ed, 2021.

Sanderson, Grant, and Josh Pullen. “But What Is a Neural Network?” *3Blue1Brown*, 5 Oct. 2017, www.3blue1brown.com/lessons/neural-networks.

“What Is a Chatbot?” *IBM*, www.ibm.com/topics/chatbots. Accessed 26 Feb. 2024

Wilbor, Seamus. “Does ChatGPT Have a Character Limit? Here’s How to Fix It.” *Quarule*, 15 Feb. 2024, www.quarule.com/does-chatgpt-have-a-character-limit.

LIST OF ABBREVIATIONS

Abbreviation:	Definition:
AI	Artificial Intelligence
GB	Gigabyte
GPT	Generative Pre-trained Transformer
IBM	International Business Machines Corporation
LLM	Large Language Model
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
PPO	Proximal Policy Optimization
RAM	Random Access Memory
RLHF	Reinforcement Learning from Human Feedback
TB	Terabyte