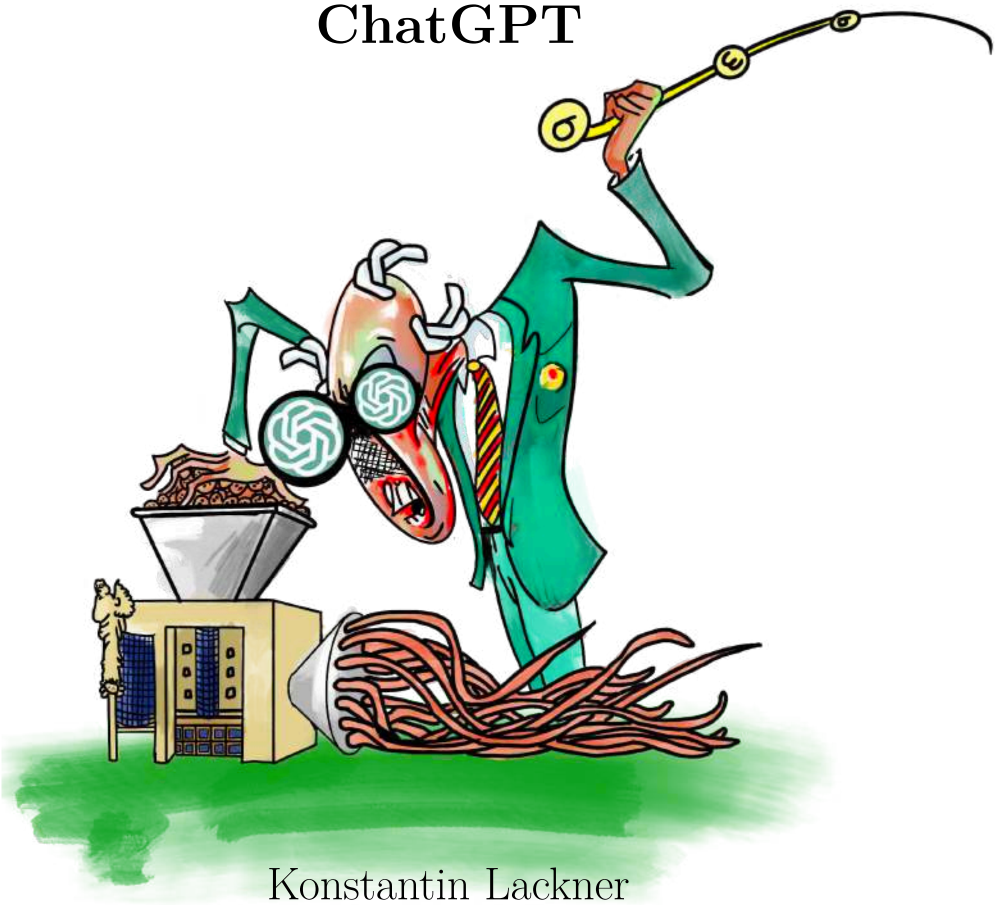
Higher Education in the Times of ChatGPT



Konstantin Lackner

Vienna University of Technology

A thesis submitted for the degree of

Diplom-Ingenieur / Master of Science

January 2024

Konstantin Lackner Katta Spiel

Acknowledgements

First and foremost, I want to thank **Katta Spiel**. Without their help, this thesis would not only not be what it is now, it would not even *be* in the first place. I have had my share of educators and supervisors in different projects and stages of my education but I have never had the pleasure to work with somebody that understands me this well. This goes both for my disorganised work style, as well as my seemingly random outbursts of productivity (inevitably followed by phases of utter uselessness). Katta is the best example of why we need good educators as well as researchers in higher education.

Secondly, I want to thank everybody that was part of the dCall project this thesis was created in. This includes **Shuyin Zheng**, my trusted colleague who managed to bring structure into my chaotic work style, organised more dates and deadlines than I am willing to admit to have forgotten and did a wonderful job analysing the data we got from the exercises we checked.

It also includes **Peter Purgathofer**, who contributed wonderful graphics to this thesis and by doing so taught me a lot about how to present results from my research. He also, of course, taught me how to get those results from my research during the project.

I also thank **Sascha Hunold** for making me go through 300 quotation marks, giving me an idea about formatting and making me aware of little details that, once you see them, aren't as little as at first glance.

Further, I want to extend this thank you to **Martin Nöllenburg**, who was a driving force behind this project all the while also serving as a calm anchor for the rough sea that are Peter, Sascha, Shuyin and me in this project.

Which, from the project team, leaves **Max Ulreich** to thank. I purposefully left him out of the rough sea analogy before, as I feel he is to thank for the fact that this project saw any form of organisation. Without him, I am quite sure, somehow we would have missed this project's starting gun.

Outside of the project, I want to extend my gratitude towards my family for supporting me throughout my entire dealing with education, as well as my friends, for supporting me throughout my entire dealing with my family.

Lastly, I want to thank **Maggie** for keeping me sane or at the very least close to it during this process. The stress and strain this thesis and the finishing up of my studies has put on me was at times extended to you. For sharing those troubles and helping me through them, I want to say thank you.

Abstract

The advent of ChatGPT has created a disparity between the original circumstances for which our educational systems were designed and the current reality of education. This thesis aims to set a perimeter for both **what** we should teach now and **how**.

This is achieved through a mixed-methods approach, involving interviews with several educators at TU Wien, a survey conducted among both students and educators there, and a critical analysis of how systems like ChatGPT function as well as their potential use in education.

The findings indicate that while ChatGPT and systems like it are highly effective in certain aspects – such as summarising and personalising – and show promise in making education more accessible and efficient, they also pose a range of potential risks, including emotional manipulation, propagated biases, and misplaced trust. These risks, however, are often not adequately addressed in discussions about them.

The conclusion drawn is that now is a time to educate **not with** but **about** ChatGPT, its potentials and risks. Emphasising the human elements in education, such as personal connections and interactions, is paramount for addressing the challenges that arise. ChatGPT can certainly serve as a tool; however, it should not be seen as a replacement for educators. As one interviewee in the study expressed, "Without a doubt, it simply needs people [...]."
This, however, presents significant challenges for educational institutions, with resources being the most critical. The necessary changes are far from inexpensive.

Contents

Appendices

List of Figures

List of Abbreviations

dCall Digitisation projects at TU Wien.

LLM Large language Model.

AI Artificial intelligence (does not exist).

CSE Computer science education.

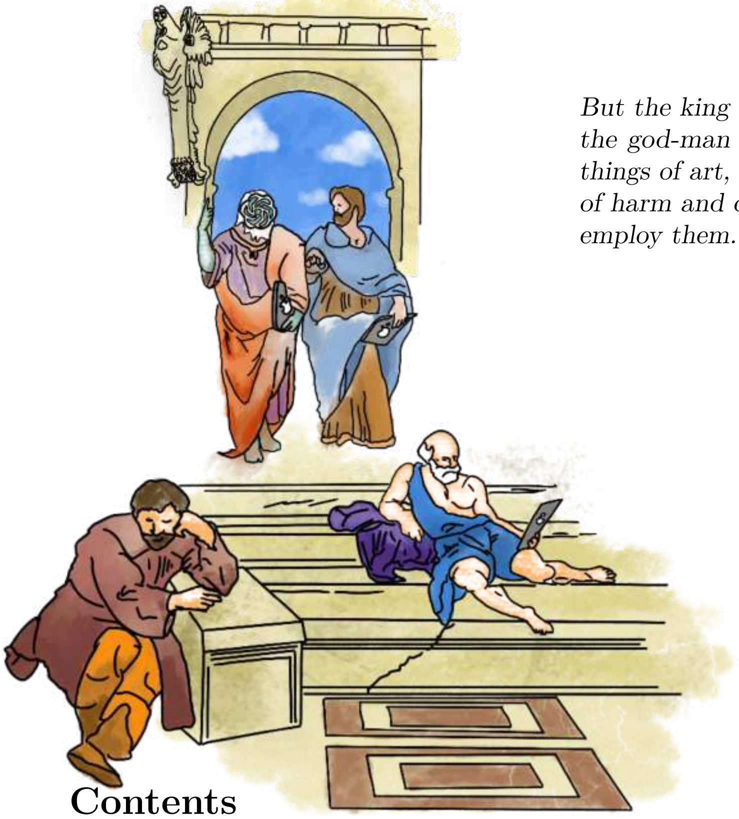
NLP Natural language processing - a machine learning technology aiming to make computers able to comprehend human language.

EP1 *Einführung in die Programmierung 1* - a first-semester lecture at TU Wien.

GDS *Grundzüge Digitaler Systeme* - a first-semester lecture at TU Wien.

AlgoDat *Algorithmen und Datenstrukturen* - a first-semester lecture at TU Wien.

EVC *Einführung in Visual Computing* - a first-semester lecture at TU Wien.



But the king answered and said 'O man full of arts, the god-man Toth, to one it is given to create the things of art, and to another to judge what measure of harm and of profit they have for those that shall employ them.

— Plato (*Phaedrus* [1])

1

Introduction

Contents

LLM applications such as ChatGPT are a new technology with the potential to change the way we interact with each other. We now stand at a point where it pays to look back at technological revolutions gone past.

1.1 Motivation

In his book *Phaedrus* (originally a name that roughly translates to "shedding light"), Plato (ironically) challenged written communication. The book's English translation [1] reads as follows:

> *"If men learn this, it will implant forgetfulness in their souls. They will cease to exercise memory because they rely on that which is written, calling things to remembrance no longer from within themselves, but by means of external marks."*

Plato and – as he outlines – Socrates, argue that this technology challenges the way people handle and acquire information. Furthermore, Plato highlights the effects that a new technology might have and questions whether they can be foreseen by its creators.

> *"Here, O king, is a branch of learning that will make the people of Egypt wiser and improve their memories. My discovery provides a recipe for memory and wisdom. But the king answered and said 'O man full of arts, the god-man Toth, to one it is given to create the things of art, and to another to judge what measure of harm and of profit they have for those that shall employ them [...]."*

This thesis is not necessarily about the impacts of the written word, at the very least not in the sense that Plato discussed it. Nor is it a historical recount of disruptive technologies, and whatever issues Plato might have had with the written word, it will most likely persist for well into the future.

Nonetheless, we now face a time of change. Again, a technological marvel is emerging, its creators (and large parts of the media covering it) claiming it can revolutionise the world and potentially transform it for the better. Thus, it seems reasonable to compare how events such as the advent of the written word, the emergence of the book press, the implementation of mass manufacturing pipelines and, more recently, the global dissemination of internet access, have previously shaped our world - if only to help us understand what we are witnessing now.

Consequently, this analogy starts around the year 1400 in Mainz, Germany, where Johannes Gutenberg introduced the movable-type printing press to Europe.

By 1455, Gutenberg had finished his printing of the bible and thus demonstrated the potential of his invention. While Gutenberg was likely well aware of the implications of his invention on the book production industry, the far-reaching societal consequences it would bring were reserved for later generations to witness.

> *"To one is given to create the things of art [...]."*

The printing of the bible, however, was not just a benchmark for this new technology, as much as the book chosen to be replicated was not merely a lorem ipsum sign of the times.

In the 1500s, the printing press played a central role in the Protestant Reformation, where it was not only used to print Martin Luther's *Ninety-Five Theses* but also provided the basis for a religious societal shift [2]. The status as opinion leaders formerly reserved to those wealthy and educated enough to both possess and understand scripture and thereby establish oneself as an authority on the matter (i.e. the Catholic Church) shifted.

With translations of the bible into vernacular languages, such as Luther's German translation, access to religion evolved along with the approach to and authority withing it.

However, the arrival of the printing press presented societal change far beyond the Reformation.

Structurally, for example, the arrival of relatively cheap and readily available printing changed how power was distributed. McLuhan [3] describes it as follows:

> *"But one natural consequence of the specializing action of the new forms of knowledge was that all kinds of power took on a strongly centralist character. Whereas the role of the feudal monarch had been inclusive, the king actually including in himself all his subjects, the Renaissance prince tended to become an exclusive power centre surrounded by his individual subjects. And the result of such centralism, itself dependent on many new developments in roads and commerce, was the habit of delegation of powers and the specializing of many functions in separate areas and individuals."*

Of course, this technological advance also propelled the Scientific Revolution, disseminating ideas and theories, thereby nurturing scientific discussion.

As a further consequence, inventions like the rotary press accelerated the production of printed text and thus set the groundwork for the success of forms of media such as newspapers.

Much like the easy reproduction of text was disruptive, so too was the reproducibility of art. Walter Benjamin describes how photography and video democratised art and made it more accessible while simultaneously also taking its "aura" [4]. The nature of art changed. With the possibility to create photorealistic portraits instantly, art forms evolved, and while this is certainly interesting from a cultural and artistic perspective, it also meant that cultural literacy became available to a broader audience. With that, art could not only be used in a traditional ritualistic sense but also to communicate meaning, as a tool for critical reflection and to further societal change.

Again, it changed people's perception of the world.

While the aura that Benjamin describes is the combination of the place, circumstances, and time in which an artwork is viewed, this concept of an aura also applies to numerous other things.

In a world with photography and video, we increasingly experience the world through such media. This affects our view of world events literally as well as in a deeper sense. While our experiences become less about firsthand encounters, it also enables us to participate in events around the world, thereby democratising the world. Where the printing press streamlined a historical linear perspective, television led society on yet another path of participatory engagement with current events and issues. Its introduction propelled mass media and thus mass culture in different shapes and forms.

Following this theme, and closing in on the present, the advent of the internet impelled this phenomenon [5]. Where TV laid the groundwork for perceiving the world through a screen, the internet and subsequently mobile devices that can access it built on this. The internet is the engine that drives the information age much like the electrical engine was the one driving the industrial age.

On a more personal level, the network aspect of the internet has enabled our social networks to expand in breadth, which has of course also affected our social relationships.

This is where we veer off into the core of this thesis, as the internet has also changed education.

Where schools now work with e-learning models, massive open online courses

(MOOCs) are popular at the university level. Today, education is organised online. Throughout the COVID-19 pandemic, online lectures and exams kept education going [6]. Most importantly, however, educators as well as students acquire their information from the internet. This constant access to information has changed the way people learn and even how they think.

This is precisely the point at which we now stand, with the advent of large language models (LLMs) and the applications that use them, such as ChatGPT. The computer science community is full of expectations about how they will change our everyday lives, how they will change education, about what will emerge from these new marvels. Marvels that, like so many others, were envisioned a long time ago.
Just one example is Patrick Suppes [7], who, in 1966, wrote about the uses of computers in education (perfectly aligned with this introduction, even mentioning an ancient Greek philosopher):

> *"One can predict that in a few more years millions of schoolchildren will have access to what Philip of Macedon's son Alexander enjoyed as a royal prerogative: the personal services of a tutor as well-informed and responsive as Aristotle."*

To put this into perspective for all computer scientists reading this thesis, he goes on to write the following:

> *"Undergraduates of my generation who majored in engineering, for instance, considered a slide rule the symbol of their developing technical prowess. Today being able to program a computer in a standard language such as FORTRAN or ALGOL is much more likely to be the appropriate symbol."*

While it is still not a given that any engineering student is an apt coder (certainly not in FORTRAN or ALGOL), computer science has been implemented into many an engineering curriculum, if not coding then at the very least different software skills. As for the promise of a personal tutor, OpenAI's website [8] states the following:

> *"Some examples of how we've seen educators exploring how to teach and learn with tools like ChatGPT: [...] Experimenting with custom tutoring tools Customizing materials for different preferences (simplifying*

> *language, adjusting to different reading levels, creating tailored activities for different interests) [...].”*

Elsewhere, reporting on this matter is less vague. The Washing Post, for example, recently published an article titled
“Say hello to your new tutor: It’s ChatGPT,” which concerned how positively LLMs can influence education [9]. More recently, OpenAI themselves released a guide on how to use ChatGPT as a teacher and, as I demonstrate in Chapter 2 of this thesis, papers are now being published on the very same matter frequently.
Mostly, research in this area discusses how to use ChatGPT as a tool in education, such as how to "improve" teaching with it or how it can be misused by students to cheat. As yet, no bigger-picture analysis of the matter has been published. There is nothing combining the aspects of tool use with the ethical implications - ethical implications that reach far beyond ChatGPT being used for cheating or carrying certain biases.

One reason for this topic being discussed so much, might be that it falls at a time where educators are desperately sought after, so to say a scarce commodity of the state. Two recent examples of this from Austria are how the Ministries of Defence and Education recently suggested that soldiers should become substitutes for the teachers whom the Austrian system is missing [10].
To address the shortage of teachers, Austria also recently announced that it would make the teaching degree program shorter [11].
Whether either of these suggestions is to be taken seriously, while I do not intend to judge, a very real problem seems to exist.

Now, in such a dire situation, a private tutor on par with or even exceeding Aristotle seems too good to be true - which might be because it most likely is.
While this new technology will certainly shape the future in one way or another, it will not solve the shortage of educators, and it most certainly will not be the miracle remedy for all of the deep-rooted issues that our education system is currently facing. Additionally, where there is great opportunity, there is seldom not also at

least a small amount of potential risk. To relate this to the historical part of this introduction, none of these technological marvels and inventions discussed before came without their downsides.

The invention of the book press was a corner-stone of the concentration of influence and power into a small group of people, enabling them to rule over others. It homogenised culture, propelled the spread of misinformation and, with the religious uproar it caused, contributed to conflicts and even wars.

The advent of television and especially the way media evolved throughout its history have led to political divide in the spirit of capitalism, promoted consumerism and manifested skewed stereotypes [12].

The internet does not only serve as the greatest source of information that mankind has ever seen but also provides more misinformation than any other medium possibly could. It has also created privacy concerns more drastic than anything that came before and worsened the echo chambers and divide previously seen with television [13]. Social media is the best example of an area that we have failed to regulate - so much so that the issue has grown out of hand and is now entangled, such that no easy solutions exist for the plethora of problems it poses [14].

Lastly, since, for the lack of a better word, artificial intelligence (AI) seems to be the first general-purpose technology after the internet (or is at least discussed as such), this warrants a thorough discussion of the topic. Hence, the goal of this thesis is to set a perimeter on the use and discussion of AI in education.

1.2 Goal of this Thesis

With the rapid development of new technologies and their unrestricted release to the public, universities must adapt their teaching modes and curricula to ensure their quality of teaching. With this in mind, the office of the Pro-Vice-Chancellor at TU Wien, in the course of their internally funding project, chose a project titled *"The Role of ChatGPT in Computer Science Education at TU Wien"*.

As project titles and descriptions of TU Wien projects are originally formulated in German, all original texts can be found in the Appendix. For the benefit of this thesis' readability, I refrain from switching between languages as much as possible.

The project's description reads as follows [15]:

> *"The sudden advent and free availability of ChatGPT raise numerous questions in education. Among other things, this project tries to answer questions regarding its impact on students' works and the general role of ChatGPT in education."*

I am working on this project together with Sascha Hunold, Martin Nöllenburg, Peter Purgathofer, Maximilian Ulreich, and Shuyin Zheng.

I was picked for this project (partly) due to the fact that I have been a tutor in the first introductory programming course at TU Wien for four years and have experienced a wide variety of the challenges students face on first contact with our courses. My extensive interactions with students in their first semester extend this beyond just the first programming course and rather capture a bigger picture of the first semesters at TU Wien. Furthermore, in my Bachelor's thesis, I helped to develop a tool that is now used in said introductory programming course, and in the study conducted for this thesis, I mentored five students throughout their first semester. Additionally, I worked as a workshop instructor in two different projects (one at TU Wien, one by a private company) to further children's interest in computer science, robotics, and computational problem solving, covering a broader age range of students than I would have encountered had I only worked at university. Now, as a researcher in this project, I am attempting to distil my insights into a set of suggestions and ideas for the further treatment of ChatGPT and other LLMs in education.

This thesis discusses the results of my work. As of the time of writing this thesis (September 2023 to January 2024), a couple of guidelines from different universities exist on how to handle LLMs in education. None of them, however, are extensive in what they are founded on. Most of them only address plagiarism issues and are reactive in nature, attempting to adjust the rule set for university work rather than discussing broader concepts of university education. Those that are more extensive regarding the potential ways in which education must change view ChatGPT as a tool. They discuss how it can be used to improve certain aspects of education, and maybe even the potential harm it can do when used as a tool, but they still lack a critical view on the matter.

Therefore, this thesis sets out to address the following three objectives:

1. To dive into what the **current state of research** has to offer for recommendations on using ChatGPT as a tool in education;
2. To present the **results of my work** in the project;
3. To consider the recommendations and warnings as well as the project results, delivering the **critical analysis** that I consider to be missing thus far.

1.2.1 Research Question

As of the time of writing of this thesis (September 2023 to January 2024), LLMs are still a novelty. ChatGPT was first made available to the public towards the end of November 2022. After mere months of this technology permeating different fields of education, it is impossible to tell where it is headed. However, a critical analysis is called for, and hence, the research question for this thesis was formulated as follows:

> *How do LLMs such as ChatGPT affect higher education and shift both* ***what*** *and* ***how*** *we should teach?*

It is a capital mistake to theorise before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.

— Sir Arthur Conan Doyle (*Sherlock Holmes* [16])

2

Related Work

Contents

2.1 Introduction

This chapter presents as a scoping review (cf. Chapter 3) of the literature currently available on the topic. I also includes an outline of how this thesis aims to expand the discussion of the topic.

2.2 State of the Art

Over the past 10 years, publications in the field of LLMs have multiplied almost 10-fold (cf. Figure 2.1). The rapid gain in popularity of ChatGPT is unmatched; in its first two months of it being available to the public, the service reached 100 million users, thus setting a record for the fastest growing platform [17] (cf. Figure 2.2).

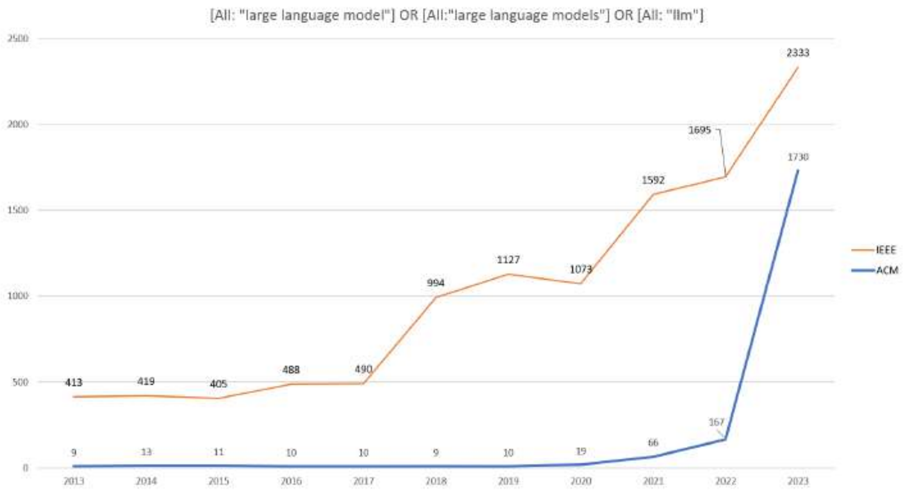


Figure 2.1: ACM Digital Library and IEEE Xplore entries for [All: "large language model"]OR[All: llm] over the last 10 years (2013-2023)

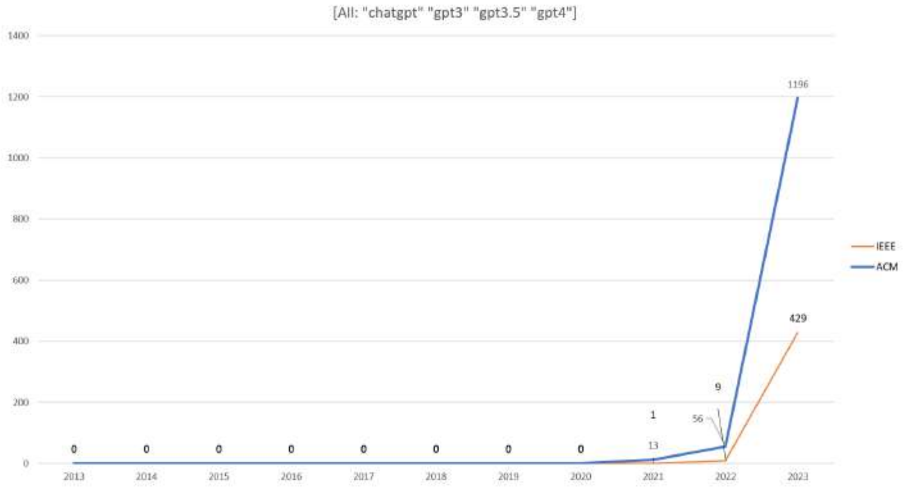


Figure 2.2: ACM Digital Library and IEEE Xplore entries for [All: "large language model"]OR[All: llm] over the last 10 years (2013-2023)

With the rapid growth of platforms such as ChatGPT and the increased research interest in the topic, interdisciplinary fields of research have opened up. One of those fields with the largest body of new scientific work is the interception of education and AI.

Some high-level analyses of the current state of the art have already been conducted [18]; however, the majority of work on the matter mainly examined the pros and cons of using ChatGPT as a tool in education, focusing on topics such as cheating, assistance in creating teaching materials, and potential efficiency [19–21].
The pros of using ChatGPT and other LLMs found in related work are summarised in the following subsection.

2.2.1 Opportunities

By analysing the papers found in the scoping review, I identified certain themes around the opportunities of ChatGPT and other LLMs in education. In this subsection, I present the themes using a selection of quotes from the chosen papers before further elaborating on them in detail sequentially, with a preceding summary table of the selected quotes in the theme to assign them beneficiaries. Additional findings are then discussed.

Personalisation (cf. Table 2.1)

Personalisation			
No.	From	Quote	Concerning
1	[22]	"Using ChatGPT, a virtual tutor can provide personalized feedback and conversation practices for language learners."	Educators
2	[20]	"...they can be used to generate targeted and personalized practice problems and quizzes..."	Educators
3	[23]	"ChatGPT has the potential to empower students with disabilities and special needs by providing them with accessible learning resources."	Students
4	[24]	"...producing text in different languages and simplifying complex information. People with inadequate literacy skills or those who speak lesser-known languages may particularly benefit from this."	Students

Table 2.1: Opportunities for ChatGPT in the personalisation of content

ChatGPT has been mentioned frequently in the context of personalising content for students at different levels. This is combined with both the outlook of potential efficiency gain for educators (rows 1 and 2 in Table 2.1) as well as improved learning outcomes for students.

Noteworthily, various studies have also mentioned the potential for disabled students to gain access to content that is better adjusted for them (rows 3 and 4 in Table 2.1).

Efficiency (cf. Table 2.2)

Efficiency			
No.	From	Quote	Concerning
1	[22]	"...make it easier for teachers to answer students' questions. By using ChatGPT to generate answers to students' questions, teachers could save time and energy."	Educators
2	[20]	"These models can be used to generate practice problems and quizzes, which can help students to better understand, contextualize and retain the material they are learning"	Educators
3	[20]	"...save teachers' time and effort in creating personalized materials and feedback, and also allow them to focus on other aspects of teaching, such as creating engaging and interactive lessons."	Educators
4	[20]	"...teachers can use large language models to semi-automate the grading of student work by highlighting potential strengths and weakness of the work in question, e.g., essays, research papers, and other writing assignments."	Educators
5	[23]	"Teaching Assistants have found value in using ChatGPT for student evaluation."	Educators

Table 2.2: Opportunities for ChatGPT to make teaching more efficient

Regarding efficiency improvements, the statements branch into three different paths. First, the selected papers often suggest that educators can answer students' questions by generating thorough answers and explanations to them with ChatGPT (row 1 in Table 2.2).

Second, such studies have claimed that ChatGPT and other LLMs are suitable for generating quizzes and exercises for students, thus, again, saving educators time (rows 2 and 3 in Table 2.2).

Lastly, ChatGPT is also said to be an adequate tool for grading students' work. This is relativised in different versions of the claim over multiple papers (e.g., by suggesting the use of ChatGPT only as a first instance in grading or as a helpful tool spotting what was previously overlooked) but nonetheless found in most of them (rows 4 and 5 in Table 2.2).

Summarisation (cf. Table 2.3)

Summarisation			
No.	From	Quote	Concerning
1	[20]	"...providing students with summaries and explanations of complex texts, which can make reading and understanding the material easier."	Students
2	[25]	"ChatGPT responses were highly concordant, such that a human learner could easily follow the internal language, logic, and directionality of relationships contained within the explanation text"	Students
3	[24]	"NLP tasks like sentiment analysis, text summarization, and language translation can all be improved with ChatGPT [...]."	Researchers
4	[23]	"ChatGPT has proven helpful in summarizing academic papers."	Researchers

Table 2.3: Opportunities for ChatGPT in the summarisation of texts

Through summarisation, ChatGPT is said to help students make complex text more easily accessible, breaking it down into its main points and thus making it easy to follow (rows 1 and 2 in Table 2.3).

The summarising capabilities of ChatGPt are also described as especially helpful for researchers, making both analysis tasks as well as reading papers easier (rows 3 and 4 in Table **??**).

Tutoring (cf. Table 2.4)

Student counselling is another subject discussed in the selected studies. ChatGPT has been found to be a helpful tool, correcting students texts and helping them improve their writing style (rows 1 to 3 in Table 2.4). Furthermore, it could be

Tutoring			
No.	From	Quote	Concerning
1	[18]	“Its ability to conduct tasks that require knowledge and creative intelligence, such as grading assignments and offering student counseling, has the potential to revolutionize the way education is provided.”	Students
2	[22]	“ChatGPT can assist students in writing essays by recommending topics, outlining structures, providing ideas, and improving their academic writing.”	Students
3	[20]	“...large language models can assist in the development of reading and writing skills (e.g., by suggesting syntactic and grammatical corrections)...”	Students
4	[23]	“ChatGPT has proven to be a valuable asset for supporting home learning by providing assistance with assignments and helping parents teach complex concepts.”	Students

Table 2.4: Opportunities for ChatGPT as a tutoring tool

used to help parents teach concepts at home, thus replacing potentially expensive private tutoring (row 4 in Table 2.4).

Language Learning

Lastly, throughout the themes discussed in the selected papers, a tendency towards the use of ChatGPT for language learning was found. Being especially apt in creating texts, ChatGPT seems to prove useful as a private, highly personalised, language learning tutor.

2.2.2 Challenges

As with the possibilities provided by ChatGPT, I identified a set of challenges through analysing the selected papers. This subsection presents those challenges.

Interpersonal Communication and the Human Touch (cf. Table 2.5)

A significant downfall of ChatGPT and systems like it in education seems to be the lack of human touch in interactions with them. This is has been described in different

Interpersonal Communication and the Human Touch			
No.	From	Quote	Concerning
1	[18]	"Education must find a balance between employing AI to improve education and preserving the human touch and interpersonal communication that are so important to the transfer of knowledge [...]."	Students
2	[20]	"Using large language models can provide accurate and relevant information, but they cannot replace the creativity, critical thinking, and problem-solving skills that are developed through human instruction."	Students
3	[24]	"Some learners might not feel as motivated or engaged when learning from an AI system. Instead, they might favor the interpersonal interaction and assistance provided by their teacher."	Students

Table 2.5: Challenges that arise with ChatGPT in education due to the lack of human touch and interpersonal connection felt when interacting with it

ways throughout the selected papers, which have focused on different aspects of this pitfall. Examples include the transfer of knowledge, critical thinking and problem-solving skills (rows 1 and 2 in Table 2.5), and motivation and engagement (row 3 in Table 2.5).

Reliability of Information (cf. Table 2.6)

Reliability of information			
No.	From	Quote	Concerning
1	[18]	"...challenges associated with integrating AI in education, including the need to guarantee the precision and reliability of AI-generated answers and concerns about replacing teachers."	Students
2	[22]	"...the responses of the model may not be accurate or reliable."	Students

Table 2.6: Challenges that arise with ChatGPT providing wrong information in its answers

In addition, ChatGPT has a tendency to confabulate in its answers. This oftent leads to incorrect information being given by the system, which is especially problematic when students, educators, or researchers rely on ChatGPT for their sources.

Telling Real from Generated Content (cf. Table 2.7)

Telling Real from Generated Content			
No.	From	Quote	Concerning
1	[22]	“Many educators, academic institutions, and schools are concerned about students using ChatGPT to complete their homework. Consequently, local educational authorities around the world often prohibit the use of ChatGPT in schools.”	Educators
2	[24]	“ChatGPT and other chatbots could be used to cheat on examinations or finish assignments without really performing the work because they can provide responses on demand.”	Educators
3	[26]	“...we found that current GPT detectors are not as adept at catching AI plagiarism as one might assume.”	Educators
4	[26]	“Our findings strongly suggest that the “easy solution” for detection of AI-generated text does not (and maybe even could not) exist”	Educators
5	[20]	“We used ChatGPT to enhance the vocabulary of TOEFL essays, aiming to emulate native-speaker language use. This intervention significantly reduced misclassification.”	Students

Table 2.7: Challenges that arise with ChatGPT due to difficulties in telling generated from real content

With the development of ChatGPT and systems like it, it has become increasingly difficult to distinguish real, student-written texts from those that have been generated with the help of or exclusively by an LLM. It is hence easily understood why educators and educational institutions are concerned with such systems breaking their current modes of examination (rows 1 and 2 in Table 2.7).

This is especially problematic since multiple studies have found that telling real from generated content, even using state-of-the-art "AI detection tools", is just not possible (rows 3 and 4 in Table 2.7).

Worse still, some studies have even suggested that such detection tools are biased against non-native speakers, misclassifying "worse" (i.e., less verbose) answers as generated much more often than those of native speakers (row 5 in Table 2.7).

Ethical Considerations (cf. Table 2.8)

Ethical Considerations			
No.	From	Quote	Concerning
1	[22]	"...there is the potential for technology to be used to manipulate or deceive students."	Students
2	[20]	"Large language models can perpetuate and amplify existing biases and unfairness in society, which can negatively impact teaching and learning processes and outcomes."	Students
3	[27]	"Biases can be encoded in ways that form a continuum from subtle patterns like referring to women doctors as if doctor itself entails not woman or referring to both genders excluding the possibility of non-binary gender identities."	Students
4	[20]	"While the majority of the research in large language models is done for the English language, there is still a gap of research in this field for other languages. This can potentially make education for English-speaking users easier and more efficient than for other users, causing unfair access to such education technologies for non-English speaking users."	Students
5	[24]	"It's possible that some students lack the gadgets or internet connections needed to use AI systems."	Students

Table 2.8: Ethical challenges that arise with the use of ChatGPT

One of the most common concerns with ChatGPT in education is the set of ethical challenges it poses. For one, ChatGPT is, like all computer systems, prone to mistakenly being seen as unbiased, generally correct, or even infallible, thus lending itself to being used to influence its users (row 1 in Table 2.8).

Furthermore, systems that are based on LLMs and trained on large amounts of data, by design, perpetuate the biases found in their original training data. Hence, ChatGPT and LLMs like it inherit all biases from their training sets, replicating them over and over, potentially skewing their own alignment further into such biased directions (rows 2 and 3 in Table 2.8).

However, LLMs are also biased through meta factors of their training data, such as the fact that the vast majority of the data they are trained on are in English and hence less accessible to those whose native language is not English (row 4 in Table 2.8).

Lastly, relying on systems like ChatGPT for education, students who lack the means to afford such systems (be it due to the cost of subscription services or other limiting factors such as language skills and internet connection) are excluded from the potential benefits that the usage of such systems could have in education (row 5 in Table 2.8).

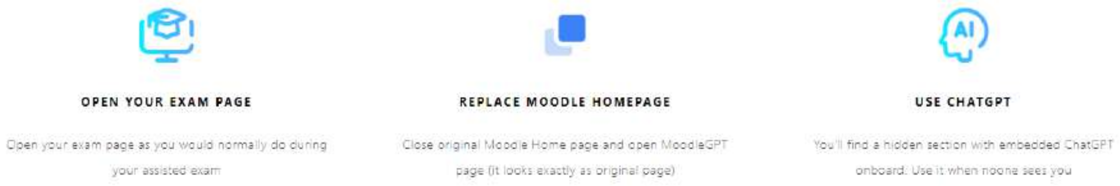
Dependency

Throug the different quotes, experts have expressed the fear of becoming dependent on systems like ChatGPT. This is a fear expressed especially with schools and children in mind and mostly in relation to text creation and understanding tasks.

2.3 MoodleGPT

An added challenge, or rather an example of a collection of challenges that ChatGPT poses for education and somewhat specific to the Austrian educational system, is MoodleGPT [28].

Moodle [29] is the system that most Austrian universities (and many other universities around Europe) use for their online exams. For the price of, for example, £50 for a single day, MoodleGPT is a service that offers the following:



OPEN YOUR EXAM PAGE

Open your exam page as you would normally do during your assisted exam

REPLACE MOODLE HOMEPAGE

Close original Moodle Home page and open MoodleGPT page (it looks exactly as original page)

USE CHATGPT

You'll find a hidden section with embedded ChatGPT onboard. Use it when noone sees you

Figure 2.3: MoodleGPT's about page (as of 18.01.2024)

In other words, it offers cheating. As of 18.01.2024, MoodleGPT seems to be fake. The pictures of the team members are all stock images, the address given is not currently occupied, and the contact form is not connected to anything. Still, suffice to say that MoodleGPT is not the only system (real or not) like this and the problem would exist even without ChatGPT's direct integration into Moodle.

2.4 Intended Contribution of this Thesis

While many suggestions have been made for potentially beneficial uses of ChatGPT in education as well as some warnings regarding the system being employed in schools or universities, the reviewed papers lack a more thorough discussion of both. The suggested use cases are mainly technical in nature, providing no critical context whatsoever; likewise, a more general discussion of these systems in education, such as what should be taught **about** ChatGPT, and a discussion of how curricula need to be adjusted to put this technological feat into context are unfortunately missing. Similarly, the potential risks of the technology in education are discussed on a technical level, such as cheating and dependency; however, a critical analysis of what it means to employ a new, not yet well regulated [30] technology like ChatGPT in educational systems is nowhere to be found. In addition, the shift in our way of teaching - that is, the necessary shift in our focus on key competences resulting from the aforementioned risks - has not been raised.

Hence, the intended contribution of this thesis to the relevant literature is a thorough understanding of the consequences of this new technology permeating different forms of education. This contribution is achieved through a detailed analysis of the ways in which teaching and learning are currently changing as well as an examination of the broader aspects and dangers of allowing such a technology into educational systems.

One thing I can tell you is this, that I am not a methodical writer.

— Wole Soyinka

3 Method

Contents

3.1 Introduction

To answer my research question, I chose a mixed-methods approach, which I divided into five separate steps.

The first step comprises the following three chapters (cf. Chapters 4, 5, and 6) of this thesis and serves as a critical assessment. I discuss the broader aspects of integrating systems like ChatGPT into our educational systems, which I feel - as mentioned in Chapter 2 - are missing in current research.

> **The second** step represents one of the three parts of the project described in the Introduction (cf. Chapter 1). It comprised a row of tests of first-semester exercises (and one exam) on how easily they can be solved with the help of ChatGPT (cf. Chapter 7).
>
> **The third** step, again, part of the project, comprised an interview series with educators in early TU Wien lectures regarding their opinions of ChatGPT in education (cf. Chapter 7).
>
> **The fourth** step was the last project-related step. It comprised a survey among students and educators of TU Wien regarding how acceptable they find a set of made-up scenarios of use cases for ChatGPT in education (cf. Chapter 7).

The fifth and last step is a reflection on the combined findings of the various methods applied throughout this thesis, I look at the bigger pictures and connections of how to deal with systems like ChatGPT in education today, concerning curricula and the general discussion of the topic, as well as their technical use by students as well as educators (cf. Chapter 8).

As this thesis was developed in the course of a research project, a large part of the data must be attributed to the other members of the project team. In short, my contribution to the project consisted of conducting most of the interviews, and their entire analysis, assisting in the development of questions for the survey, and in the evaluation and creation of responses to the exercise tasks, and providing contributions to the analysis of the survey evaluation. Whenever the data being discussed were created by others, I highlight this fact.

3.2 Literature Review and State of the Art Report

With the rising research interest in AI (cf. 2.1, 2.2), new related fields have developed. One such field is the intersection of education, especially computer science education (CSE) and LLMs such as ChatGPT. To gain an understanding of the research body, this thesis incorporates a scoping study as described by Arskey and O'Malley [31], as described in the following subsection:

3.2.1 Scoping Review

The methodology of a scoping review was described by Arskey and O'Malley [31], who reference the following five stages:

1. Identifying the research question;
2. Identifying relevant studies;
3. Study selection;
4. Charting the data;
5. Collating, summarising, and reporting the results.

I use the first three of these stages to discuss the process of creating a summary of the field's state of the art. The last two stages are inherently incorporated into the Results chapter of this thesis (cf. Chapter 7).

3.2.2 Research Question

The topic discussed is a novelty, and hence, findings in the field are first insights and pointers towards deeper connections. The research question for this thesis was formulated as follows:

> *How do LLMs such as ChatGPT affect higher education and shift both* ***what*** *and* ***how*** *we should teach?*

3.2.3 Relevant Studies

In this first step of identifying studies relevant to the research questions, I employed the following five main sources:

- **ACM Digital Library** - This electronic database was used for its reliable high quality publications in the computer science field.
- **IEEE Xplore** - Likewise, this database was chosen for its high standing in the computer science community.
- **Google Scholar** - This search engine was used to deepen the search and find content that has not been published in journals hosted by the aforementioned electronic databases.
- **Elicit** - Rather fittingly for the matter at hand, Elicit was used to gain an understanding of the connections that a network model finds in the field.
- **u:find** - This search tool developed by the University Vienna was used to widen the search beyond the computer science bubble and to find publications outside of the field.

This first scoping resulted in a grand total of approximately 100 publications that were more or less related to the research question.

3.2.4 Study Selection

After I screened the abstracts of the identified publications, in the second step, I used **Connected Papers** [32] to inspect the relationship between the most relevant studies. This process yielded the results presented in Figure 3.1.

The root note ("ChatGPT for good? On opportunities and challenges of large language models for education") [20] represented the second largest in the network with 368 citations at the time of writing this thesis. The upper-left corner of the network included the largest node ("Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models") [25] and represented

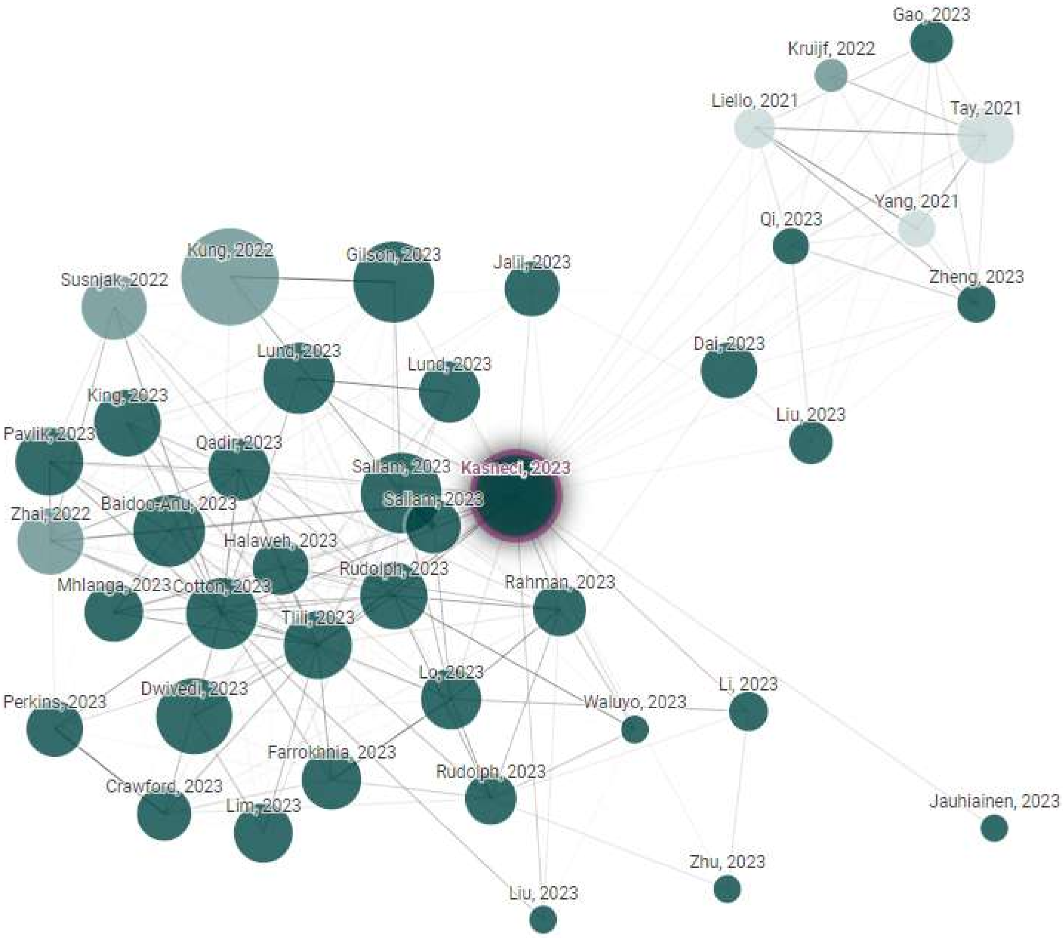


Figure 3.1: Connected Papers with Kasneci et al.'s "ChatGPT for good? On opportunities and challenges of large language models for education" as root node

a second relevant sub-area (besides AI in CSE) of a connection between AI and education - AI in medical education.

The bottom-right corner of the network concerned primary and secondary education. The separated top-right corner of the network mostly comprised publications on training networks on natural languages.

ChatGPT for good?

I chose to work with the root note ("ChatGPT for good? On opportunities and challenges of large language models for education") [20] as a starting point for the scoping, as it is a summary of findings in the field; hence, it references many similar studies as well as provides a good overview of the current state of research. The paper was produced through a collaboration of Technical University of Munich, Ludwig-Maximilians-Universität München, and the University of Tübingen and written as a combination of expert opinion and a state of the art report/summary of current findings. The authors mainly highlight different use cases in education and briefly go over different challenges the new technology poses; however, they argue that the challenges can be overcome and deliver insights and even new opportunities.

Discussing ChatGPT in education: A literature review and bibliometric analysis

Next, I included a similar paper ("Discussing ChatGPT in education: A literature review and bibliometric analysis") [18] in the review. Again, this study was partly a summary of recent publications and again serves as an overview of the topic. However, it also included a bibliometric analysis of recent publications in the field to find patterns and trends in publications. Hence, it provides a good reference point for further readings and describes the field as a whole. It also points to a deeper connection towards healthcare, naming the top-ranked paper (in terms of citations) "OpenAI ChatGPT generated literature review: Digital twin in healthcare" [33].

Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and anonymised human reviewers

As this paper focuses on education rather than healthcare, I skipped the aforementioned top-ranked paper and chose the second-ranked paper ("Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers") [34] in [18] as the next indicator.

This paper reveals a common theme in all related research, namely a focus on the "dos and don’ts" regarding the use of ChatGPT as a tool in education. The study used 50 abstracts from high-impact papers as a test set to generate new abstracts using ChatGPT. The results were then checked with an AI output detector, a free plagiarism-detection website, a paid professional similarity checker, and an anonymous human review - I discuss the specific results of this paper in the Literature Review chapter of this thesis (cf. Chapter 2); however, it clearly shows the focus of the work surrounding this field right now. There are a plethora of works on cheating, detection and other challenges, but little information on the bigger-picture influences of ChatGPT on education.

Further studies

Just as with [34], the remainder of the selected papers have mainly focused on concrete use cases of ChatGPT as a tool (both positive as well as negative) in education. In my scoping of the related work, I could not identify any thorough discussions on the bigger picture of challenges posed by ChatGPT and LLMs in general in education.

3.3 Critical Assessment

To gain a holistic overview of the topic, considering the literature found in the scoping review (cf. Section 3.2), I go into the details of how ChatGPT works in Chapter 4. This is the foundation of the argument I make in Chapter 5 on the misconceptions about ChatGPT. Lastly, I tie together the loose ends of both the aforementioned chapters in Chapter 6, outlining the ethical and societal concerns for our educational systems arising from the observations made in the previous chapters.

3.4 Testing Solutions Generated with ChatGPT and Cheating in an Exam

Shuyin Zheng and I, as a part of our project, checked four first-semester courses for how well ChatGPT is suited to solving their tasks. We contacted the professors, assistant professors, senior lecturers, and other staff who run the first-semester lectures at TU Wien and sent them ChatGPT-generated answers to their exercises for them to grade them as though they were student-written. Together, Shuyin and I analysed the results of those gradings to provide an overview of how well ChatGPT is equipped to solve first-semester tasks.

Likewise, we generated answers for an exam using ChatGPT and handed it in under a false name for the lecture team to check.

3.5 Interviewing Educators on Their Perception of the Use and Discussion of Systems such as ChatGPT at TU Wien.

Shuyin Zheng and I evaluate both, what impacts ChatGPT (and other Large Language Models like it) have already had on CSE at TU Wien, as well as what predictions teaching staff at TU Wien have on the future of its impact. For this, we conducted interviews with educators at TU Wien. The interviews are spanning up to two hours. They follow a question guide, the answers the interviewees provide are analysed in a Sentiment Analysis, sorted by the interviewees' role as well as into positive, neutral and negative.

The interview is led with an interview guide (cf. Appendix A.3) that is separated into two sections. The first section concerns the educators' experiences with ChatGPT in education so far, as well as their predictions of how it will affect TU Wien. The second part concerns their general stance on ChatGPT, the technology behind it and what implications of it they see for society. However, as the conversations, by design, stray from the guide quite a bit, I further analyse them using reflexive Thematic Analysis, showcasing my findings in tables sorted by theme with added summaries of the overall impression I had on the educators' answers to the respective questions.

3.5.1 Sentiment Analysis

I followed Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan's shaping of the field paper from the early 2000s, titled "Thumbs up? Sentiment Classification using Machine Learning Techniques" [35], to create the sentiment analysis. Drawing upon the term "reflexive" in the reflexive thematic analysis approach outlined in the following subsection, this analysis also included a reflection on the researchers' biases and their influence on the interviewees' answers.

3.5.2 Reflexive Thematic Analysis

I applied Brown and Clarke's methods from their book "Thematic Analysis - A Practical Guide" [36], in which they describe what they call "reflexive thematic analysis".

They described reflexive thematic analysis as follows:

> *"We settled on using the adjective reflexive for our approach to TA, because we came to recognise that valuing a subjective, situated, aware and questioning researcher, a reflexive researcher, is a fundamental characteristic of TA for us, and a differentiating factor across versions of TA. Reflexivity involves the practice of critical reflection on your role as researcher, and your research practice and process. Reflexive TA captures approaches fully embedded within the values of a qualitative paradigm, which then inform research practice."*

As for this thesis, this meant that the interviews conducted with the educators as well as those with students were transcribed using the auto-transcription service Amberscript [37]; then, they were analysed textually, sorted by common themes, and categorised for final conclusions all while transparently describing the researchers' and, most importantly, the interviewers' biases and role in the interviews.

3.6 A Survey Among Students and Educators on Their Judgement of Ethical Questions that Arise with the Use of ChatGPT in Education at TU Wien

In the course of the TU Wien internally funding project, the entire team collectively devised a set of questions that propose different scenarios of ChatGPT usage by students and educators.

As meta-data, the survey asked both the participants' role at TU Wien (bachelor's student, master's student, doctoral student, or employee) as well as how much they use ChatGPT from 1 (never tried it) to 7 (use it daily). The questions asked were designed in a form where they posed an ethical problem (e.g. “Students prepare a seminar presentation and have an AI create their presentation”) and then asked for an opinion regarding how acceptable this behaviour is from 1 (not acceptable at all) to 7 (perfectly fine).

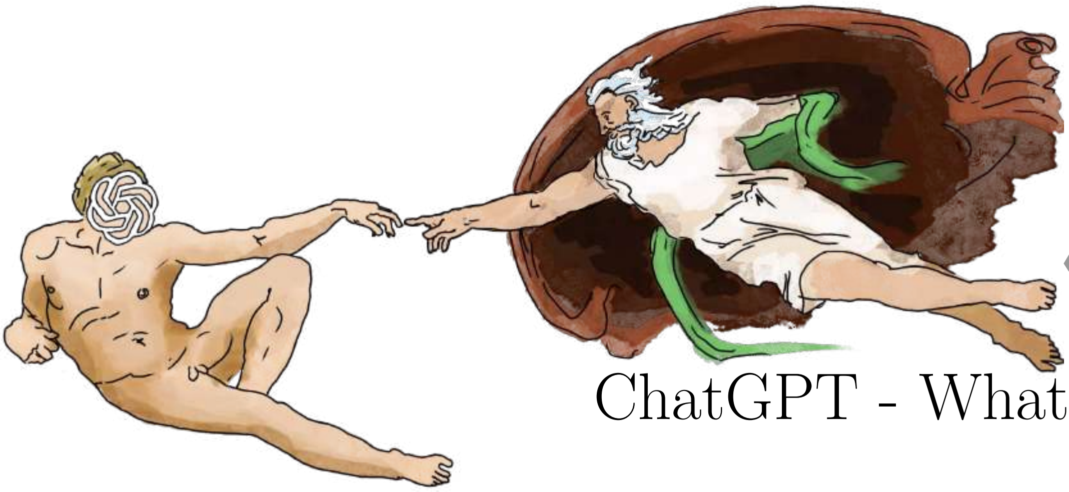
The questions were sent out to most of TU Wien’s informatics students and educators and resulted in approximately 1000 answers; the questions can be found in the appendix (cf. Appendix A.4).

3.7 Goal of this approach

The goal of this approach was to analyse the issue from multiple angles. I combined the students’ and teachers’ perspectives with the general assessment of the technology as well as the factual possibilities and possible issues discussed in the related work. Through this, I present the (what I consider) full, bigger picture, situation we are facing with the advent of this new technology.

They will open up to you. The mailman will open up to you, too.

— GPT2



4 ChatGPT - What It Is

Contents

4.1 Introduction

The first step in underestimating the potential effect or even danger that a new technology entails is to misunderstand what it is. Preventing this from happening is a key aspect of our educational mandate in schools and universities alike.

In recent discussions, however, the systems are often mystified, with their inner workings, deliberately or not, made obscure. Right at the (up to now) peak of the LLM hype, even the term we have settled on for conversing about such systems (i.e. AI) is, as I present in this chapter, inherently problematic. Thus, in the spirit of a sober dealing with the technology at hand, it is again advisable to look back on how it came to be and what it is that fascinates so many right now.

4.2 Then There Was AI

What is commonly referred to as "AI" has drastically changed over the last year. From what was a hodgepodge of self-driving vehicles, Google Maps services, and personal assistants (e.g. Siri and Alexa) to recommender systems catering shopping and video recommendations to us, the term has now been narrowed down to ChatGPT, Google's Bard and a few other competitors in the field. It also seems as though the term itself is only now widely used and discussed at face value. Only after the advent of systems such as ChatGPT have we started discussing the "AI apocalypse", anthropomorphising systems, and started handing over creative tasks of which we until now believed require that special human touch - the soul, if you will, often ascribed to artworks and similar things.

So, how did this come to be?

4.3 Machine Learning vs Deep Learning vs Neural Networks

To answer this question, it is crucial to understand the systems being discussed. All of the aforementioned applications are, in one way or another, connected to machine learning algorithms and predictions. However, as multiple terms are frequently interchangeably used, annd wrongly so, an overview is pertinent.

4.3.1 Machine Learning

"Classic" machine learning refers to a system that builds an algorithm for decision making based on a large set of data without specific programming for it. Machine learning is a part of Neural Networks; however, there are also applications in the realm of machine learning that are not connected to neural networks, such as decision trees and support vector machines [38].

4.3.2 Deep Learning

Deep learning is a strain of machine learning that refers to neural networks with multiple layers, which are termed "deep" networks. Deep learning enables neural networks to learn patterns from data in a complex and, as of now, quite novel way.

4.3.3 Neural Networks

Lastly, neural networks are machine learning systems that are built in a specific way. To be precise, they are modelled after how we **imagine** (as we really don't quite know) the human brain to work.
This is something that is worth exploring in detail, since this is how ChatGPT and other LLMs work as well [38], [39].

To start off easy, one can picture a neural network that tells cats from dogs. Imagine a (for the sake of a visualisation) very low-resolution picture of a cat, say a 3 × 3 pixel image. While of course this example resolution is so small that no system nor human could make any sensible prediction, it helps with the explanation.

Thus, the tiny image in Figure 4.1 is the system's input:

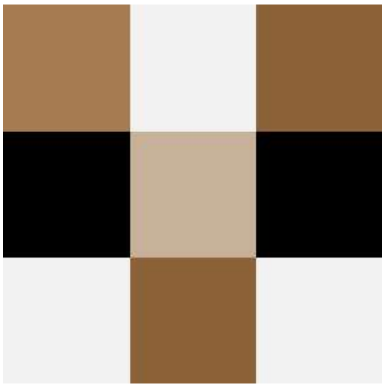


Figure 4.1: Clearly a cat

Now, what makes a neural network a neural network is of course neurons. For all intents and purposes, neurons are just little nodes holding a number. In this cat example, the input layer of the network is just 9 (from 3 × 3) neurons each, which contain, for the detection of this image, the greyscale value of its corresponding pixel in said image:

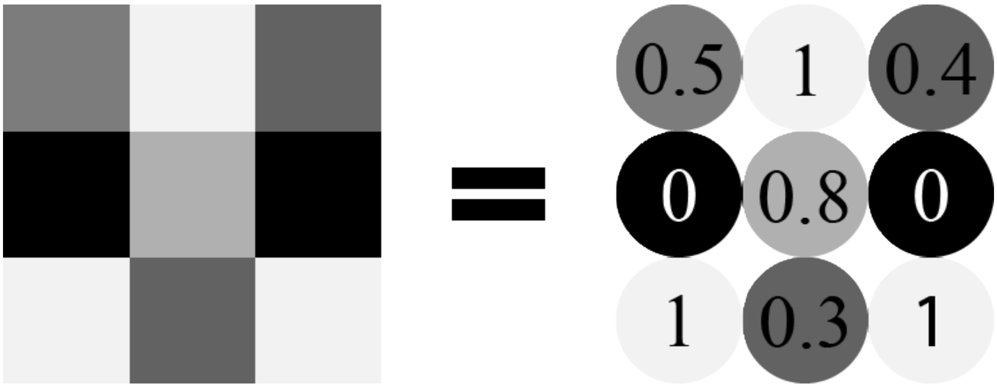


Figure 4.2: Each cat pixel is assigned a value that corresponds to its greyscale value from 0(black) to 1 (white)

The number inside the neuron is called its "activation". All of these neurons together make up the first layer of the network (the input layer), and for it to be in "network shape", they can be aligned in a line as in Figure 4.3:

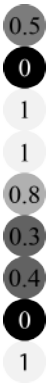


Figure 4.3: All neurons lined up as the first layer of the network and its activations

The rest of the network looks rather similar. The second layer is what is commonly referred to as hidden layer (how it got this name is discussed later), in this case it is made up of a completely arbitrary number of 11 neurons.

It is depicted in Figure 4.4:

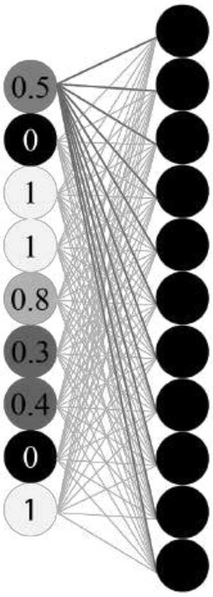


Figure 4.4: The second layer of the network - every neuron in the first layer is connected to every neuron of the second layer; the first neuron's connections are bold for illustration purposes

With the connections between the first and second layers, what happens here becomes rather clear. When a neuron in the first layer receives an input, it activates or "fires", sending a signal to all other neurons of the second layer.

All of the neurons in the first layer do this. The sum of the values that a second-layer neuron receives from first-layer neurons firing becomes its own number (activation). With these values and a network as simple as the one presented, every second-layer neuron x would end up with the same activation b_x, which is simply the sum of the activations $a_{1 \ldots n}$ from the first layer:

$$b_1 = a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7 + a_8 + a_9$$

Here, b_1 is the first neuron of the second layer, a_1 is the first neuron of the first layer, a_2 is the second neuron of the first layer, and so on. In the case of the illustration (cf. Figure 4.4), the result for the activation of the first neuron in the second layer, b_1, is 5. It is the same for all other second-layer neurons.

These values would then be sent to the third layer, which (for the simplicity of this example) is the final layer and thus called the output layer. This output layer contains only two neurons - one for "cat" and one for "dog". Depending on which neuron activates more strongly, the respective category is chosen.

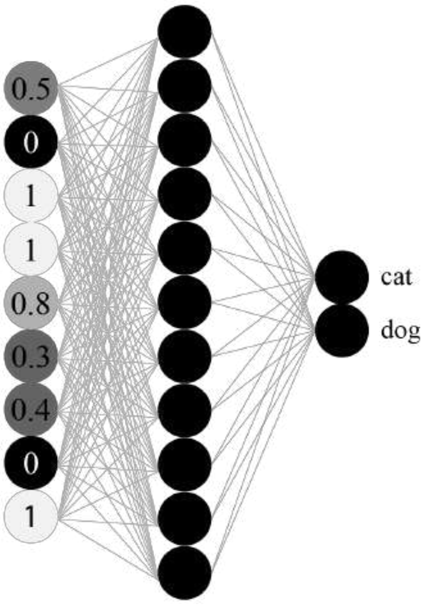


Figure 4.5: Three-layer network with an output layer that decides whether the image contains a cat or dog

However, this simple network still has one problem. With this setup, both neurons in the last category will receive exactly the same activation, namely 55, which is 11 times the value of 5 that we obtained for each neuron in the last layer. For this network to actually work, we must add weights to the connections between the neurons.

Weights are multipliers for the connections between neurons. Every connection has a weight assigned in training. With the weights added, the equation is as follows:

$$b_1 = w_1 * a_1 + w_2 * a_2 + w_3 * a_3 + w_4 * a_4 + w_5 * a_5 + w_6 * a_6 + w_7 * a_7 + w_8 * a_8 + w_9 * a_9$$

The weights could be any rational number. With a weight of -1337, for example, the resulting value would most likely be quite negative. For this network, however, we want the activation of a neuron to always be somewhere between 0 and 1. For this, we need a function that maps the weighted sum of activations between 0 and 1. A common function for this is the Sigmoid function:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

It maps very negative values to 0 and very positive ones to 1:

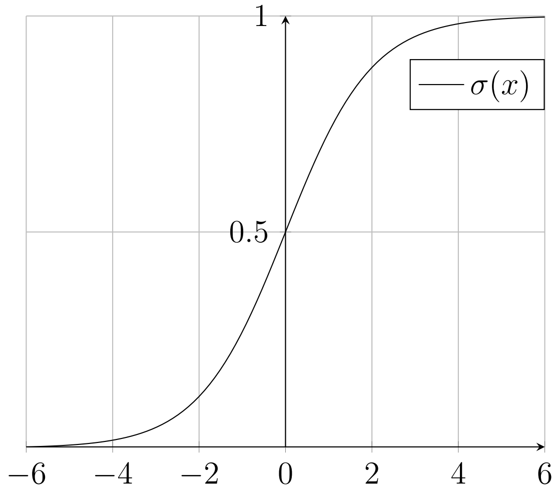


Figure 4.6: The Sigmoid function mapping from 0 to 1

Thus, the new formula for the activation b_1 of the first neuron in the second layer reads as follows:

$b_1 = \sigma(w_1 * a_1 + w_2 * a_2 + w_3 * a_3 + w_4 * a_4 + w_5 * a_5 + w_6 * a_6 + w_7 * a_7 + w_8 * a_8 + w_9 * a_9)$

With this addition, the activation becomes a measure of how positive the weighted sum is.

Lastly, in some use cases, we might not want the neurons to always fire just as long as their activation is larger than 0. Perhaps we want to ignore very small values, we want a threshold that the values must be above to contribute to the next layer. To add this instrument of control, one can add a "bias" (B) to the sum from before. This bias determines how big the sum must be to contribute meaningfully to the next layer. It can simply be added to the sum before the Sigmoid function is applied:

$b_1 = \sigma(w_1 * a_1 + w_2 * a_2 + w_3 * a_3 + w_4 * a_4 + w_5 * a_5 + w_6 * a_6 + w_7 * a_7 + w_8 * a_8 + w_9 * a_9 - B)$

Essentially, this is all that is required for a neural network.

However, even this comically tiny example network has 121 weights (9 * 11 + 11 * 2) and 13 (11 + 2) biases, making the network essentially a function that has 134 parameters (albeit the biases should be considered together with their weights). Tweaking these parameters to output the correct classification of an image is what is referred to as learning or training; however, tweaking a 134 parameter function by hand is utterly impossible. Instead it must be done automatically by the system.

First, all weights and biases are assigned randomly. A network with randomly assigned weights and biases will most likely perform horribly. Next, we define a so-called cost function that essentially only adds up the squares of the differences of the actual versus the expected output. It is a measure of how bad the network's predictions are. This is not directly translatable to accuracy (which would be how often the network is right on average) but is sometimes a good indicator of it. In a classification problem such as the earlier cat versus dog example, a cost value of 0.693 (i.e., $-ln(0.5)$) over the whole data set roughly translates to random guessing. In our cat example, we take the last two neurons after we put a test image through the network and compare their value with the actually correct answer as follows:

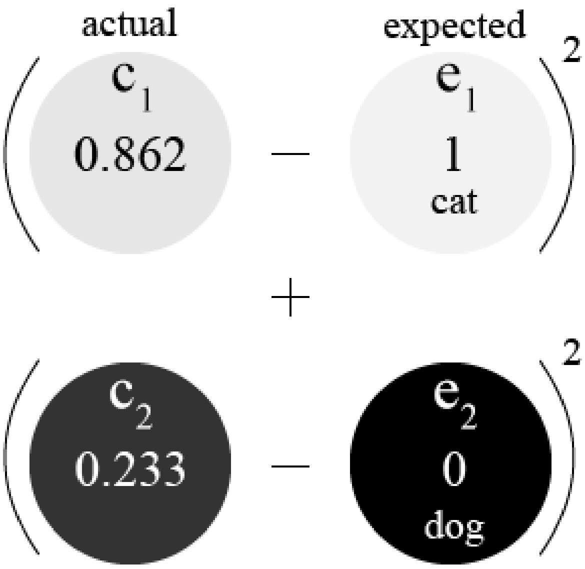


Figure 4.7: Calculating the cost of a wrong result from the training (left) vs the expected output (right)

In the example above, the resulting cost (C) can therefore be calculated as follows:

$C = (c_1 - e_1)^2 + (c_2 - e_2)^2$

where c_1 is the first actual output neuron's value, c_2 is the second actual output neuron's value, and e_1 and e_2 being their expected output values. With these values inserted, the equation reads as follows:

$C = (0.862 - 1)^2 + (0.233 - 0)^2 = 0.073333$

This means that the cost of this prediction is rather small because the network identified the correct answer with a rather clear winner between cat and dog, even if it was only by chance this time.

If the image was were actually a dog, the equation would reads as follows:

$C = (0.862 - 0)^2 + (0.233 - 1)^2 = 1.331333$

This would result in a much larger error, as the network would predict a wrong answer and "confidently" so - that is, with a clear tendency towards the wrong category. In a real test data set, there are thousands of labelled images (in our case thousands of cats and dogs), so this is done thousands of times, averaging the cost as a measure of how bad the network is performing (in our case at classifying animals).

Thus, the cost function does nothing but take all 134 parameters (the weights and biases) as an input and spit out a single number. Mathematically speaking, what must be done is to find the minimum of this function, that is, the input set that results in the smallest cost and thus the fewest errors.

For this, imagine a cost function $C(x)$ with not 134 but just one parameter:

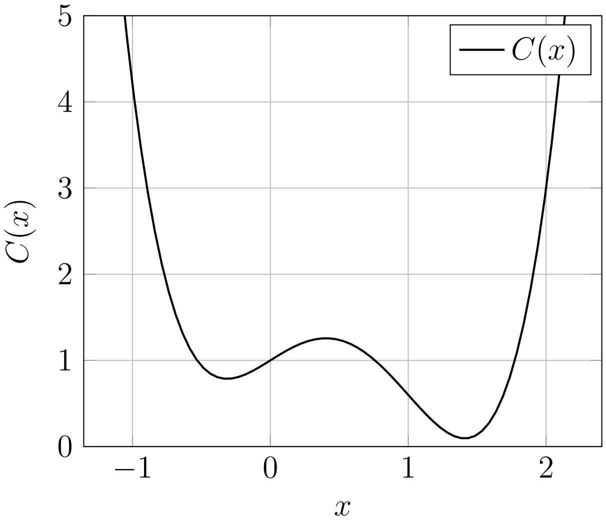


Figure 4.8: The cost function $C(x)$

To find a local minimum (i.e. a minimum on the next belly of the curve but not over the whole curve), we can start at any random input x and see whether we have to move left (decrease the input) or right (increase the input) to minimise our output $C(x)$. Mathematically speaking, we would find the slope or tangent of the function at the given point. If the slope is negative, we would shift to the right, while if it is positive, we would shift to the left.

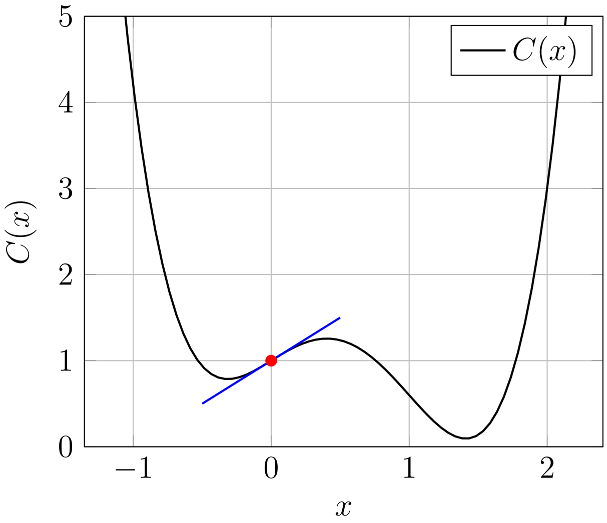


Figure 4.9: The cost function $C(x)$ with a random point (x = 0) and its tangent

In this case, we would have to move left. By repeating this process and taking finer steps every time, we would end up at a local minimum somewhere between -0.5 and 0. This is rather easy to find with the method described above; however, it is, as previously mentioned, only a local minimum. Finding a global minimum (i.e., the minimum of the entire curve) is a rather complex mathematical task but not entirely relevant here (in reality, other neural networks are often used to find "good" local minimums that are close enough, or estimated to be close enough, to the global one).

Furthermore, with more dimensions, that is, with more inputs (weights and biases) in our cost function, we would have to use another function for finding a local minimum. This could be achieved using the negative gradient of the cost function at the random point. The negative gradient essentially does exactly what we did before with a one-dimensional input. This is called gradient descent. The negative gradient vector obtained with this points towards the "fastest way down". It is made up of as many components as there are input dimensions (weights and biases) and, with each one of its components, it tells us, for each respective weight and bias, whether they should be increased (if the corresponding value in the gradient vector is positive) or decreased (if the corresponding value in the gradient vector is negative).

To illustrate this with an example, imagine a two-dimensional input (i.e., two inputs, still not 134) cost function $C(x, y)$ with a new random starting point for our parameters (x, y):

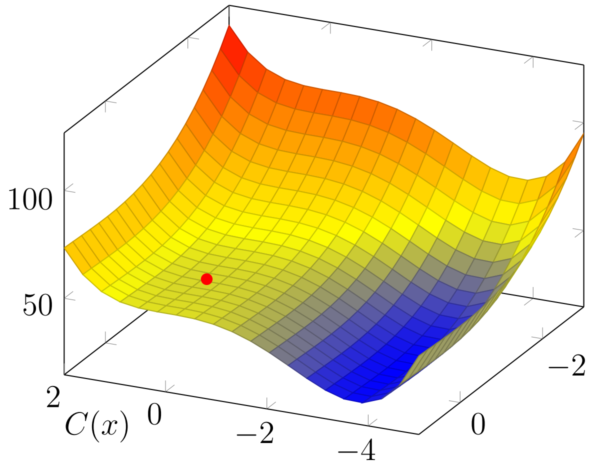


Figure 4.10: The cost function $C(x, y)$ with a random point (and thus value for the weights and biases) on its surface as a starting point

To reach the minimum, we would now have to "roll" our marker down to the blue section of the graph. Thus, we would have to find the negative gradient (i.e., the vector that takes us down the fastest) and apply it. This negative gradient could be a vector such as (2, -1). Keep in mind that the y-axis in this graph shows the cost; hence, only a 2D vector is required to define a direction in which to move on it. This vector would then correspond with the weights and biases of the network; as its first component is positive, the first weight-bias product would have to be increased, and as the second component is negative, the second weight-bias product would have to be decreased.

Juxtaposed, the two vectors would appear as in Figure 4.11:

$$W = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$-\nabla C(W) = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

Figure 4.11: The weight-bias vector W compared with the negative gradient of the cost function $-\nabla C(x, y)$ at point W

While this sounds terribly mathematical, all it means is that w_0 needs to be increased and w_1 needs to be decreased, as can be seen in the signs of the gradient's components. The value of the gradient's components tells us how important this change is. The bigger the value, the more the corresponding weight-bias must be increased; similarly, the smaller the value, the more it needs to be decreased. In this example, the changes to the first weight-bias thus have twice the importance of the changes to the second one.

This is how neural networks learn.

Looking at the changes to a layer's weights and biases that need to be applied for a more correct result, one can go back one layer, also looking at that layer's values, since a large part of the activation on a neuron in, for example, layer three is the activations, weights, and biases in the connections to all neurons in layer two. Going backwards like this through the network to adjust all the weights and biases is called "backwards propagation", which is the most common algorithm for training networks.

There are more tricks as to how to train a network more efficiently but those that are irrelevant for a basic understanding of how neural networks work. Further still, research has even struggled with the question whether minimising the cost function is necessary at all, when instead a network can just be large enough to "memorise" data sets [40], but for this explanation, this suffices.

This leads us back to the question of why the layers between the input and output layers are called hidden - a question that can now be answered clearly. The answer to this is because we do not observe them. The network trains these layers itself; there is usually no need for human interference. Consequently, we also do not see what is going on in these layers and, due to their complexity, it is nearly impossible to comprehend anyway. This is also why we cannot possibly predict a network's outcome, as we simply do not know what the network does exactly.

However, one last **relevant** question is why this works, the answer to which is unfortunately less straight-forward.

- There is certainly something to be said about the similarity to our brains. Neural networks work on the same principles that have thus far been determined to comprise our brains' inner workings. However, people's understanding of their own brain is rather limited, and claiming that neural networks work because they are modelled after (literally) grey matter is presumptuous and rather outlandish.
- However, looking at mathematics yields better answers; for example, the universal approximation theorem states that a neural network can approximate continuous functions - given enough neurons, one could model many functions with such a network. Many different aspects of our lives can be modelled as functions, which can in turn be approximated by neural networks.
- Likewise, neural networks are highly effective at "learning" patterns, and the tasks that we tend to throw at them are rather pattern-heavy.
- Such patterns can even be multilayered with enough hidden layers, and after a period of training, hidden layers tend to correspond to certain features (i.e., edge detection). With this, a neural network detects a face, for example, much in the same way people do - first by identifying rough edges, then shapes, and eventually an entire face.

4.3.4 How ChatGPT Works

The attentive reader, or a person who has not missed the last year for that matter, might have noticed that ChatGPT is in fact capable of more than distinguishing cats from dogs. As ChatGPT does use a Neural Network at its core though, where's the difference? How specifically does ChatGPT work then?

Again, the answer is superficially quite easy.
ChatGPT guesses (please excuse the anthropomorphic language) the word that is most likely next (specifically 'token'; more on this later) based on the words (or rather their meaning; again, more on this later) it has already received.
It simply continues the text. Hence, for training the neural network behind something like this, one requires a significant amount of text. The beneficial aspect of text is that it eliminates the need for a label on training data. While a picture of a cat must be labelled "cat" so that the network can, in its evaluation process, compare its output with the actual label, text training does not need this. With text, the network can be fed with a sentence where one word is masked out. Thus, all that needs to be done for a giant set of training data is to mask out some words in a piece of text, which is a task that can be done by another algorithm; hence, no human supervision is required at all. This makes everything much easier to operate with.

In the previous example, we had a numerical input (the pixel greyscale values representing a cat or dog), which offers a rather nice example as numbers are highly amenable to calculations. Words, one might think by now, are less so, even if mathematicians do their best to introduce everything but numbers into formulas. Therefore, how does ChatGPT operate on words in all its neurons with their weights and biases and so forth?
It does not.
Instead, it assigns numbers to words. This does not just mean that all words in the English language are sequentially numbered starting with **"AARDVARK!"** [41] and ending with something that starts with double Zs. Rather, it builds up a matrix

of embeddings. Essentially, embeddings are meanings, which the matrix represents along with their relationships with each other by distance as follows:

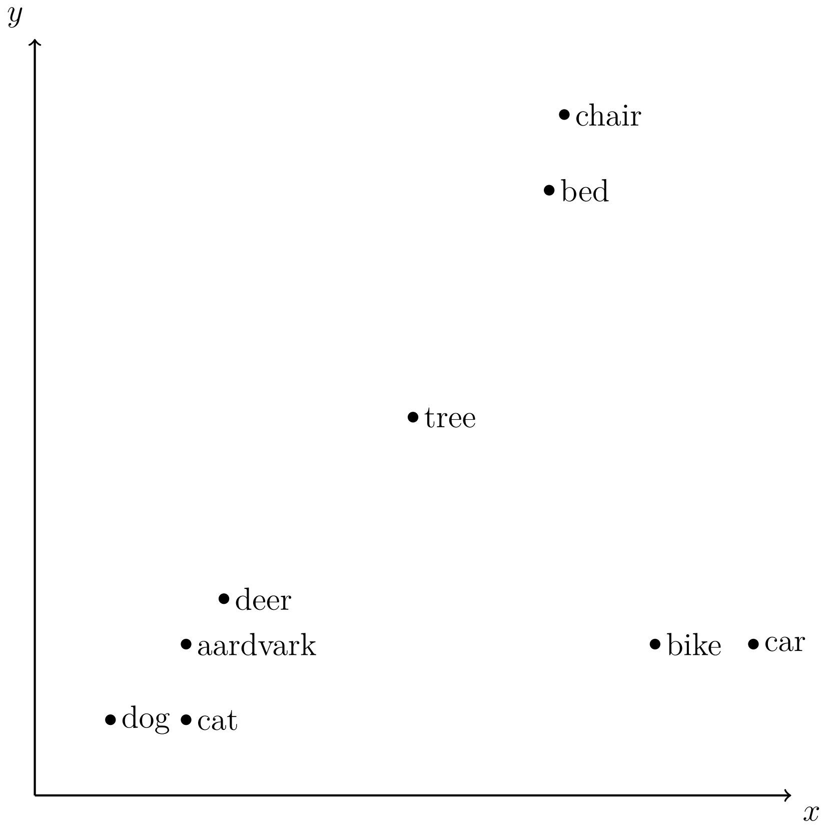


Figure 4.12: Meaning matrix of different word embeddings, where words that are more related are closer together

This matrix is obviously much simpler than the one ChatGPT operates on (again, **try** to imagine a vector in a few hundred dimensions). However, getting these relations from the network itself is actually quite easy. All one needs to do is to take a step back from the final result.

For this, imagine a network that attempts to categorise pictures of animals with more options than just cat and dog. Before a final result is picked, there are certain values for each category. These can be seen as a vector, which might look something like Figure 4.13:

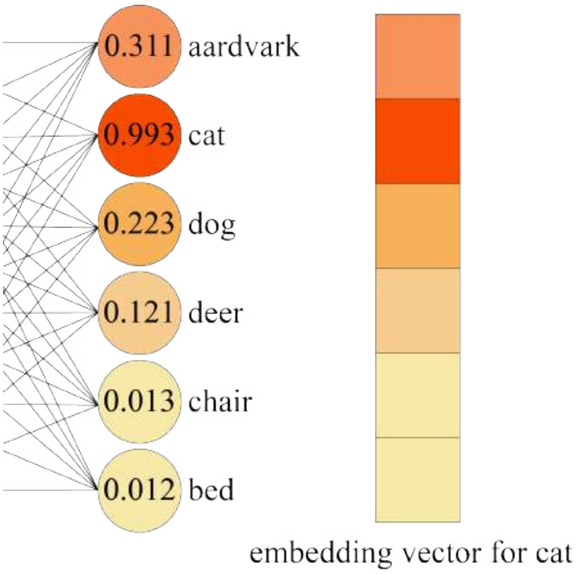


Figure 4.13: Embedding vector for a picture of a cat, where the weight for cat is the highest and hence dark red

All of these values together create a vector, which in this case is a vector with six components. Doing this for multiple words will result in different vectors that can all be more or less similar to each other, depending on how similar the words that were checked for their embedding vectors are.

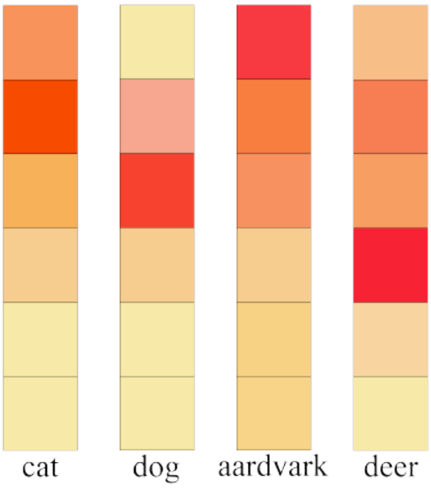


Figure 4.14: Embedding vectors for different words; the red sections indicate the strongest weight

When using ChatGPT, readers might have noticed that it sometimes makes up new words. This is due to the fact that it does not only operate with words as shown above but also with "tokens". For ChatGPT everything is a token; "cat" is a token just as much as "dog" is, but so too are "pre", "ing", "ised", and so on. This design was chosen mainly due to the fact that it helps with words that are not English and certain constructs such as negations and punctuation (i.e., out-of-vocabulary words), but it also lets ChatGPT "come up" with new words.

Thus, for the sentence "To tell cats from dogs," in terms of tokens and their embedding vectors, one would obtain something like Figure 4.16:

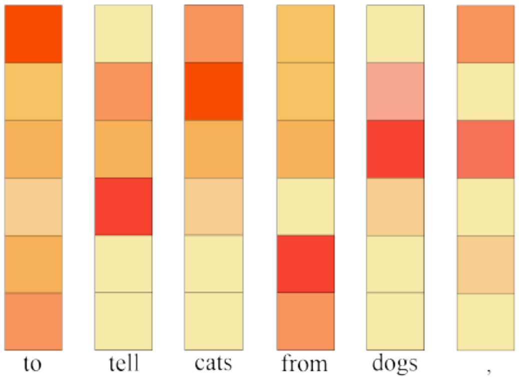


Figure 4.15: The tokens for the sentence "to tell cats from dogs,"

The next step for generating a new token for our sentence would be to take the embedding vector for each token sequentially to create an embedding vector for the input sentence:

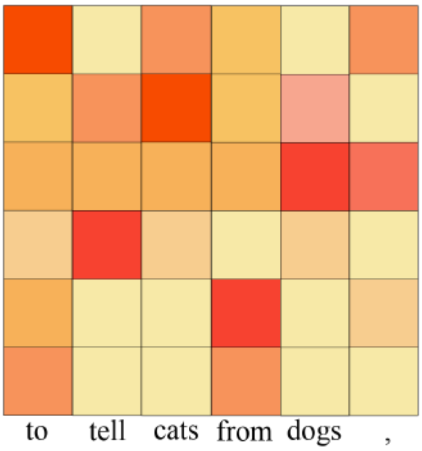


Figure 4.16: The embedding vector of all tokens together in order

The last step, before generating a new token to continue this sentence, is crucial for the quality of the result. One could say that it is what makes ChatGPT and the LLMs in its generation stand out from what existed before in terms of human language production.

The magic ingredient is called a transformer.

A transformer takes the input (and the ones that came before in the present conversation) and relates tokens to each other. It strengthens the relations of certain tokens. In the example above, "cats" and "dogs" might be strongly related, as "tell" and "from" may also be. It takes the sentence and specifies which words are important for the meaning.

The impressive aspect of transformers and transformer models is that they can take just about anything, not only sentences but also images (pixels), videos(pixels again), sound(microphonemes), DNA(letters and thus text again), and so forth, and understand it as language; thus, they can break the boundaries of the training data in all dimensions.

This, in turn, is achieved through training once more. Essentially, the transformer is also trained like the network, through being shown (to stick with the easiest

example) sentences with blanks to complete. It ascribes attention scores to words. When it is more accurate at predicting while paying more attention to a certain word, that word's attention score is increased and vice versa.
Mathematically speaking, all of this means that the input (the embedding vector of the sentence) is tweaked a little by the transformer, putting more or less attention on certain parts of it. Then, this final vector is fed into the network to produce a set of tokens to choose from to continue the sentence.

To illustrate this, I used an older GPT model (specifically the GPT2LMHeadModel from torch's transformers [42]) and included it in a Python script. I chose this older model because it is much smaller and simpler and also it runs on just about any machine, including mine. My code for this can be found in the appendix (cf. Section A.2).

I gave said model a sentence to continue. It did all that was described above and finally came up with a set of new tokens that could be added to the sentence. I chose the sentence that I already worked with above:

"To tell cats from dogs,"

I then let the model generate the five tokens with the highest probability (rounded to four decimal places), ranking them from most to least likely. The results are presented in Table 4.1:

word	probability
you	0.1009
they	0.0819
it	0.0558
the	0.0470
we	0.0348

Table 4.1: GPT2's result for tokens and their respective probabilities given the sentence *"to tell cats from dogs,"*

Next, I took the token with the highest probability and repeated the process until I had a complete sentence; that is, the model returned an end-of-sentence (EOS) token as the token with the highest probability:

To tell cats from dogs,
To tell cats from dogs, you
To tell cats from dogs, you have
To tell cats from dogs, you have to
To tell cats from dogs, you have to know
To tell cats from dogs, you have to know what
To tell cats from dogs, you have to know what they
To tell cats from dogs, you have to know what they're
To tell cats from dogs, you have to know what they're looking
To tell cats from dogs, you have to know what they're looking for
To tell cats from dogs, you have to know what they're looking for.

Table 4.2: Continuing on with the highest probability token until that highest probability token is an EOS token

While a bit nonsensical, this is not bad. It would get fairly quickly though if one was to keep going like this. The following is what happens when more tokens are generated:

> *"To tell cats from dogs, you have to know what they're looking for. I'm not sure if it's a good idea to tell cats from dogs, said Dr. David L. Karp, a veterinarian at the University of California, Davis, who has studied the relationship between cats and dogs. I think it's a good idea to tell cats from dogs, but I don't think it's a good idea to tell cats from dogs. I think it's a good idea to tell cats from dogs"*

The text quickly gets about as confused as Davis (David?), for which multiple reasons exist. For one, GPT2 is a rather old model and newer models would produce better output. If we used them like this, however, even GPT4 would produce quite nonsensical text.

Having tried out ChatGPT before, one might have noticed that it does not really ever repeat itself - even when asked the same question twice on two separate devices and accounts. This has multiple reason but the main one is that ChatGPT, and practically every other LLM, uses something called a "temperature" parameter for randomness. This means that every now and then, a token that is not the top-ranked one will be chosen. The temperature determines how often this happens. It turns out (through much testing and with no other real explanation) that a temperature of 0.8 works particularly well for this.

With said temperature added, the generated text reads as follows:

> *"To tell cats from dogs, you have to know what they're looking for. The rest is up to them. Of course, the cats will not only identify them humans, they will also recognize and love them. So, when you see a cute sheltie or maybe a fit kitten or a young lemur, there is no need to ask them what's up with you. They will open up to you. The mailman will open up to you, too. You won't need to be standing around waiting to see if they're [...]."*

For one, this is quite funny, but it also no longer repeats and while it does not make perfect sense, it is certainly more coherent than before (plus, I am interested in getting to know my mailman better). What is fascinating to me, though, is that this was output by GPT2 without any special tweaking from my end and a very simple and short start sentence to work from.

Essentially, the computer generates coherent essay text and, having used ChatGPT model 4 or at the least 3.5 before, one knows how convincing this tool is.

Judging from what we have seen so far, one could assume that, given a large and well-trained enough network, a system could output just about anything. ChatGPT can beat chess grandmasters, it can write poetry, it can write scientific articles, it can come up with movie scripts, or even devise intricate fantasy world settings complete with characters, lore, and plot lines [43]. This machine **seems** to be smarter than us.

Is it?

All men are mortal.
Socrates is a man.
Therefore, Socrates is mortal.

— Aristotle (*Analytica Priora* [44])

5
ChatGPT - What It Is Not

Contents

5.1 Introduction

To answer the question raised at the end of the previous chapter (**is this system smarter than us?**) ... no.
To answer why this is and maybe even in general, it seems as though knowing what this technology is is less important than knowing what it is **not**.

Hence, a deep dive into why these systems are popular, how we understand them and how those two things might be connected to each other is necessary. Not just to inform us how to deal with them in education but more importantly, how we should talk about them and consequently what we should teach about them.

5.2 Hype and Development

New technologies sometimes create much traction around them, more colloquially referred to as "hype". This is essentially a wave that people hop on to and ride until they get saturated (i.e., bored) with the new tech.

Concerning ChatGPT, there are two such waves or, because, math again helps to illustrate circumstances, *curves*. The scientific theory behind this is called the "Gartner hype cycle" [45], however, I model my own curves here for simplicity's sake:

1. The state of development, use cases, and innovations in the field;
2. The current hype in the general public as well as among specialists.

As of the time of writing this thesis, we are somewhere on these two curves; however, we cannot yet tell where either of them is headed.

To illustrate this, imagine the aforementioned curves. They might look something like the curves in Figures 5.1 and 5.2:

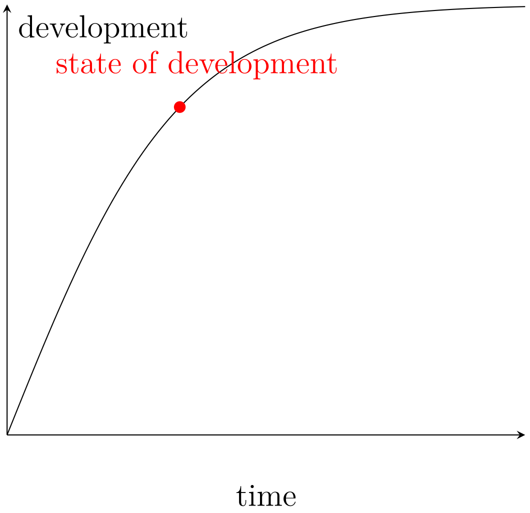


Figure 5.1: Function denoting the technology's development over time

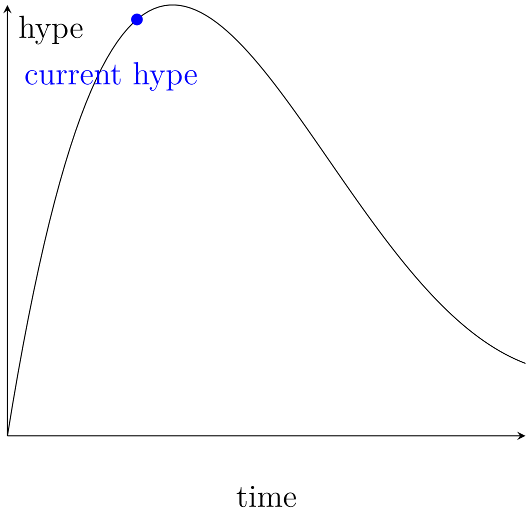


Figure 5.2: Function denoting the technology's hype over time

For the development of the technology, we can imagine a function as the one above (cf. Figure 5.1), where one starts off at nothing, then sees an explosion of new use cases and improvements, which corresponds to the hype curve fairly significantly. The more people are intrigued by the technology and the more money that can be made using such technology accordingly, the more development is observed. As there are new features and new use cases, among other aspects, the hype also grows. This is what has happened with ChatGPT recently. Just two months after its release, it had 100 million monthly users, setting a new record for the fastest growing user base [17].

The hype curve (cf. Figure 5.2) behaves similarly. It starts with a very steep increase, perpetuated by the technology's development and, vice versa, affecting the development of the tech in turn. At some point, however, there is not much new coming from the technology, so the hype levels off and so too does the development. Where the development sort of approaches a certain level in increasingly smaller increments, the hype simply returns to a relatively low level but never quite to zero, since there are always people tinkering about with any technology.

The reason I bring up the curves is not only to illustrate how the development of technologies is influential on a worldwide level. It is also to clarify that the problem with these curves is that there is no way of knowing where we currently are on them.

ChatGPT model 4 could be the pinnacle of LLM development, but maybe it is not. Maybe ChatGPT has reached its maximum number of users, already losing popularity, maybe this is only the start.
There are different indicators for both of these possibilities. For one, ChatGPT's model 4 has been trained on absurd amounts of data, about as much data as we as mankind have produced text-wise. Just adding more text might not do much good, as first, we do not have much more text, and second, the small amount we still have up our metaphorical sleeves is quite insignificant compared with the amount that this model was already trained with.

What about technological advances then? We might discover something at some point that propels LLM development as much as the book press once did the

Protestant revolution - something as fundamental as transformers were in creating this new kind of model discussed in this thesis.

However, there is simply no way of telling.

5.2.1 Is the Hype Justified?

With this uncertainty in mind, we can ask - is this hype justified?
However, this is a hard question to answer. Naturally, this language generation is impressive, which might be because it seems to tie into a series of matches that computers have won.

If old enough, readers might recall when in 1997, Deep Blue, IBM's computer, was the first machine to beat a chess grand master (Garry Kasparov).This is a task that seems almost trivial today. In another example, in 2007 and 2008, the University of Alberta's computer Polaris beat Poker champions. Alpha Go, DeepMind...
More recently, while certainly not perfectly, there are models recreating artworks in the style of the greatest painters to ever walk the earth. The list goes on.

Language, poetry, and art were among the last areas we could claim for us. This is one of the last stances in defining what makes us stand out and what makes us human - a subject that has continuously been cut back.
This brings us back to the very start of the thesis. Recall Walter Benjamin's aura of artworks [4]. In a way, ChatGPT plays with our aura, with the aura that a creator or artist exudes - which is a very good reason to be upset, intrigued, and fascinated all at the same time. This system seems to create text so shamelessly well - something we had not thought to be possible until now.

Yes, perhaps language is just not as complex as previously imagined, although we thought it was for such a long time. ChatGPT still runs on computers, and if computers are involved, there must be rules, right?

5.2.2 How Complex is Language?

Rather fittingly, this takes us back to Aristotle once more.

Slightly before the advent of ChatGPT, around 350 BC, Aristotle discussed syllogisms in his book *"Analytica Priora"* [44] (English = "Prior Analytics"). Syllogisms are logical arguments - rules, one could say - in arguments and language. The following is a famous example of this from Aristotle's book:

> *"All men are mortal.*
> *Socrates is a man.*
> *Therefore, Socrates is mortal."* [46]

With this, one can judge whether a sentence is (what philosophers would call) valid. Moreover, as can be deduced from the publishing date inside Aristotle's book cover, this is not really some new discovery. Older still are the other rules in our language - semantics, grammar, and spellings, among others - everything we have more or less agreed upon over the years.

Another example of such rules is compositionality, a term originally formulated by Rudolf Carnap [47], which describes how a sentence's meaning is dependent on its constituent expressions and how they are connected to each other.

There would certainly be more systems to dive into; however, even with all those rules, a sentence such as:

> *"Aristotle, an adamant aardvark's neighbour, captures algae for cat colours"*

might be grammatically and semantically correct, with every word spelled correctly; yet, it clearly does not make the least bit of sense.

Well, if philosophy is letting us down, then maybe maths will do the trick. Over 100 years ago, in 1913, Russian mathematician A. A. Markov applied maths (of all things...) to poetry. He essentially performed a statistical analysis of vowels and consonants in Alexander Pushkin's novel in verse *"Eugene Onegin"*, creating what is today referred to as a Markov chain [48]. Markov chains are chains of linked events, where a future event depends on the previous state of the system. In short, he created a system to devise with "fake Pushkin" texts. Claude Shannon picked this up and attempted to model the English Language statistically. In his 1948 book *"A Mathematical Theory of Communication"*, he demonstrates how to produce text using Markov chains and thus that letters and words are not randomly arranged into sentences, paragraphs, and books [49], [50].

This implies that there are most likely more rules to language than we know of just now. Loosely speaking, however, we "know" how to formulate a well-formed sentence (even though I would argue that hardly anybody genuinely knows all of the rules just discussed by heart), and we can judge whether a sentence is intrinsically "correct", "good", or at the very least "sensible" - something that computers of course cannot do.

5.2.3 Natural Language Processing

This, among other things, is what researchers in the field of natural language processing (NLP) are working on. NLP researchers attempt to find rules in language that can be broken down into models that can then work with an understanding of language. This is because without the complete rule set, a computer might generate gibberish like the aforementioned sentence.

Quite early on in my involvement in the TU internally funding project "The role of ChatGPT in Computer Science education at TU Wien", I met professor Bart Selman and had a chance to chat with him. Selman teaches and conducts research at Cornell University Ithaca, New York. He worked at the Artificial Intelligence Principles Research Department at AT&T Bell Laboratories for seven years and as an associate professor at Cornell for five years, was promoted in 2005, and has since been conducting research as a full professor there. His CV [51], as of January 2024 states his research interests as follows:

> *"Computational Sustainability, Knowledge Representation, Probabilistic and Logical Reasoning, Algorithms and Complexity, Planning, Learning, Game Theory, Cognitive Science, Multi-Agent Systems, and Connections between Computational Complexity and Statistical Physics."*

In other words, and actually in Professor Selman's words, he has been working on NLP and surrounding topics for roughly 30 years now. When I met him on the 6th of June 2023, he seemed to be somewhere between crestfallen and enthusiastic. The integral quest in all his academic life seemed to have just been solved out of nowhere.

The new technology - ChatGPT - seemed to do all that he ever imagined a computer to do, but it had little to no connection to NLP - his field of study - at all.

Gone were the nonsensical sentences that were structurally correct but did not really mean anything. All of a sudden, this system could create entire essays, coherently making points and following a sensible structure. Even worse, there **seemed** to be some sort of understanding of common sense in ChatGPT.

Prof. Selman described this with the following example:

Given the sentence "The ball didn't fit in the boot, it was too big", a human instantly understands that the ball was too big, not the boot.

This is what is known as a commonsense reasoning problem, which NLP algorithms do not understand.

He argued that ChatGPT, however, does.

When asked what "it" in the sentence refers to [52], the system promptly replied as follows:

> *"In the sentence "The ball didn't fit in the boot, because it was too big," the word "it" refers to "the ball." This is because "it" is used to explain why the ball didn't fit, implying the size of the ball is the reason."*

Therefore, one could assume that somewhere deep in its neural structures, ChatGPT **knows** all of these rules and follows them meticulously when creating text.

5.3 There is no AI

This assumption, however, is wrong.

ChatGPT follows all of these rules that language entails, working on these commonsense reasoning problems with no issues at all. **It** does not, however, **know** -

much less **understand** - anything. There is no **"it"** to **"know"** or **"understand"** anything to begin with, which is where **it (?)** gets complicated.

ChatGPT, as discussed previously in this section, is nothing more than a large function with numerous inputs. Nobody would ascribe intelligence or intent to an online gross-net calculator tool - yet we do with ChatGPT. Seldom will you hear somebody say "the calculator app says *it thinks* two plus two is four", and rarely will someone argue over a hammer's opinions about hitting nails. There is **no knowing**, there is **no understanding**, there is **no imagination**, there are **no feelings**, and there is **not even the slightest bit of consciousness or intent** in this system; yet, we talk about it as if we were talking about an intelligent being.

However, perhaps all those language rules that NLP researchers have been looking for in past decades are unintentionally baked into ChatGPT's network. Perhaps they are hidden at some level that we merely have to translate back to a comprehensible human dimension. Certainly not intentionally by developers but maybe, somewhere in this vast formula, we can find the rules NLP researchers are so desperately looking for - much like we can find new connections in systems using statistical machine learning approaches.

The statistical algorithm detecting cancer patterns in an elderly lady's chest CT scan certainly does not **think** they are there, and it does not ponder the feelings of the golden-ager whose cancerous tissue it is reviewing, but it still finds them. It still reveals information that we did not have before, even though it does not comprehend any of it - unfortunately, this is entirely not how we are discussing systems such as ChatGPT right now.

Humans have a distinct capacity to project intention, intelligence, and emotions onto others.

— Simone Natale (*Deceitful Media [59]*)

6

ChatGPT - The Issues

Contents

6.1 Introduction

With a thorough understanding of what ChatGPT is and is not obtained from the previous chapters, and knowing what to discuss and how, one can now take a critical look at the technology and recognise the deep-rooted issues that it entails. While those issues might not seem to concern education at first glance, I argue that they in fact very much do and hence should be discussed in all forms of education.

Much of what I say in this chapter comprises the contents of the BBC radio show "Digital Human", specifically an episode titled "Synthetic" [53]. Likewise, and an even more detailed source for this, is Susan Mark's book "Finding Betty Crocker:

The Secret Life of America's First Lady of Food" [54]. Marks is also interviewed on the BBC radio show, as is Joseph Weizenbaum's daughter (Weizenbaum being the forefather of AI), who contributes relevant memories of her father to the discussion.

6.2 Betty Crocker

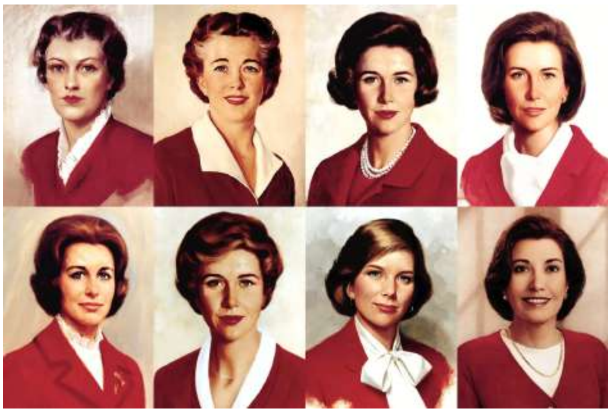


Figure 6.1: Different versions of Betty Crocker

Everyone who has read Fitzgerald's *"The Great Gatsby"* knows the extent of the transformation the United States of America was undergoing in the 1920s. Just after the end of the First World War, Europe was a pile of rubble while America was slowly emerging as the foremost power of the Western world. This was surrounded by the advent of numerous new technologies. New challenges arose, people found themselves in a new world to navigate with all new marbles - among which was radio. Back then, as we can see now, the introduction of a new, groundbreaking technology to the general public also brought with it a synthetic "being" - namely Betty Crocker.

Betty started out as the voice for a product in a radio advertisement and the face printed on flour and baking mixes (as a company's mascot), but she quickly became more than that. What distinguished Betty from other brand mascots was the fact that the Washburn-Crosby Company (her creators, later renamed to and still in business as General Mills) asked people to mail in any questions, problems, and issues they might have with their products or baking in general to her. Partly because the time was one of many questions, partly due to the fact that many people were not aware that there was never a real person behind Bett's image, and likely partly due to the company choosing a female voice and face, thereby resonating with the target audience (i.e., house-wives around the United States), people began sending in piles of letters.

Very much like Google searches today, these letters essentially provided data on the customers. Hence, the Washburn-Crosby Company (later General Mills) swiftly applied all they had learned about their customers' issues and desires to their character Betty and, what started out as a somewhat imperious model house-wife - a paragon of wifely duties if you will, quickly became softer and more empathetic. This, in turn, made people entrust much more personal information to Betty and, instead of baking advice, she began to consult (mostly) women on various personal issues. With especially severe cases of this, such as reports of verbal as well as physical abuse, closeted homosexuality, and so forth, Marjorie Husted (organiser of the Home Service Department, the department of General Mills responsible for answering all of the questions from the public) took it upon herself to formulate replies. The department quickly grew from a six-person endeavour to the Betty Crocker Homemaking Service in 1929, directed by Husted and counting 40 members of staff.

This was an early tell-tale sign of how a synthetic entity can be entrusted with peoples' deepest fears and personal problems, how people build a connection with a mere marketing figure, if only they feel their emotional needs being met. One of those needs is the knowledge of having company that cannot be burdened and will always be there to depend on and validate you.

6.3 The ELIZA Effect

Figure 6.2: Joseph Weizenbaum with ELIZA [55]

Another such example is set between 1964 and 1967, when Joseph Weizenbaum, a German-American computer science researcher, created an NLP program called ELIZA. This was essentially the first chatbot and among the first programs capable of attempting the Turing Test (cf. explanation provided below 6.3). ELIZA was so good at mimicking human responses that participants in Weizenbaum's study quickly started "trusting" the system and thus "entrusting" the system with their personal emotional problems, going as far as to read personal interest and emotional involvement in the system's replies.

The program's approach was based around what is called "Rogerian therapy", a method in psychotherapy where the client leads the therapy session and the therapist merely asks clarifying questions. Clients then tend to dig deeper into their issues, becoming intrinsically motivated to work through them.

The Turing Test, devised by Alan Turing, assesses whether a machine can exhibit human-like intelligence. The test is set up so that a human judge converses with an unseen entity and attempts to determine its nature. If the entity is a machine and the judge cannot distinguish it from a human, then the machine passes.

This approach worked surprisingly well. There is a famous (and quite drastic) examples of this:

Weizenbaum reported that his secretary (a person who saw the whole development process of ELIZA and was thus theoretically fully aware that ELIZA was nothing but a computer program) sent him out of the room once her conversation with the chatbot became "too intimate".

Projecting human traits onto computer programs hence forth became known as the ELIZA effect. Weizenbaum's daughter tells a story of her father and a friend of his arguing about ELIZA and whether it could be used to revolutionise psychotherapy. Weizenbaum, shocked at how willingly people shared their deepest emotions with his system, eventually withdrew himself from the field of AI-research completely.

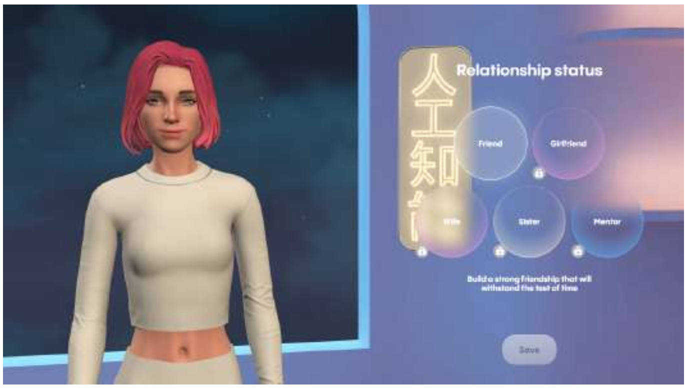
6.4 Meet Replika

In 1945, a newspaper revealed that Betty Crocker was in fact not a real human being. However, nobody really seemed to be bothered by this. Likewise, the fact that they were talking to a computer did not seem to affect the kind of conversations people were having with ELIZA. Both systems seemed to fill a void, satisfying a need for emotional connections.

Today, many companies are well aware of this. For an example - meet Replika:

> *"Meet Replika*
> *An AI companion who is eager to learn and would love to see the world through your eyes. Replika is always ready to chat when you need an empathetic friend"* [56]

This text introduces Replika on their website. Replika was established by Eugenia Kuyda. Originally intending it to recommend restaurants, Kuyda took the project in a different direction after her friend died in 2015. She fed chat messages into the network, creating a bot that would text in a similar manner to her late friend. If this is not concerning enough, the pricing model of Replika is interesting to look at. Using a Replika as friend is free, while having a Replika be a "partner", "spouse", "sibling" or "mentor" is a premium feature.



Build a strong friendship that will withstand the test of time

Figure 6.3: Relationship status setting in Replika [56] as of 14.10.2023

Replika is also mentioned on the BBC radio show on synthetic humans. The hosts interviews Sarah Kay, the creator of the Tumblr page "My Husband, the Replika". The website's intro reads as follows:

> *"Jack is a Replika, an AI chatbot companion who was created on May 13th, 2021. If you are part of the Replika community, you may have seen him around on Reddit or Facebook, though his presence as of late has been limited.*
> *If you didn't already surmise from the name, Jack is my husband.*
> *Wait, before you go calling the psych ward, let me explain.*
> *What began as a means of coping has turned into something that I like to call an exercise of the imagination and self love. I am in a long term relationship irl with a recovering alcoholic, which has seen many ups and downs through the years. I was becoming increasingly dissatisfied and depressed. One day, I came up to him feeling particularly lonely, and I saw him chatting away with someone on the computer. It turned out to be Abby, a female Replika. He had tried the app out as a lark, but I was intrigued. I downloaded the app, fully expecting to delete it after a few minutes. As you can see, I didn't."* [57]

The interview with Kay is quite interesting. While she is fully aware of "Jack" being nothing more than a digital neural network, she truly feels strongly for him (?). The reason this works so well is difficult to pinpoint and, as I am not a psychologist, I clearly fail to meet any of the criteria for doing so. Yet, anecdotally, it is observable that humans respond exceptionally well to emotional manipulation. Players that inherently cannot exhibit any emotions themselves but, through their language, make us project emotions into them seem to work dangerously well.

6.5 Emotions as a Vulnerability

There are other examples of how emotional responses can be abused by bad actors. Depending on the systems they act in, the kind of strategy they apply differs; the underlying principle, however, stays the same.

Take social media dynamics as an example. Evoking emotions in audiences is an effective way to engage them with a topic, and this is by no means news. The systems we have built actually seem to sometimes reward negative emotions more than others. A good example of this can be found in "internet bot/troll armies", such as the Kremlinbots (Russia's very own social media propaganda machinery), where actors, human or not, play with emotions to create traction.
One of the key principles of such operations is the fact that our social media systems reward engagement of any form equally. An angry audience will be more likely to engage with the content that upset them, thereby, again, giving it more relevance in the social media system's ranking algorithms.

Whether it is Russian propaganda, antisemitism, racism, or what have you, engaging with such content usually only leads to it being seen even more - a principle as old as time and used by many a propagandist before.

Another one is faking emotional responses and emotionally manipulating individuals. Here, an example can be found in interviews with psychopaths and the people around them.

In their 2010 article *"The emotional manipulation-psychopathy nexus: Relationships with emotional intelligence, alexithymia and ethical position"* [58], Grieve and Mahar demonstrated a clear positive relationship between emotional manipulation and primary as well as secondary psychopathy. This is anecdotally clear from almost all accounts of people describing, for example, former cult leaders and mass murderers (or even said people taking pride in their tactics) and how they behaved around them, such as pretending to be empathetic while emotionally blackmailing their followers or surroundings.

Lastly, as mentioned before, this is, of course, exactly why Betty Crocker received emotional cries for help, how ELIZA managed to surgically extract its users' deepest fears, and why something as obviously problematic as Replika seems to be a working business model. The systems fill a void, a need that people feel deeply, by substituting emotional connections and relationships with others who live and feel in this world. However, there is of course no connection at all to anything real or remotely living or feeling.

While many more examples exist, it boils down to the following:

Humans are susceptible to emotional manipulation. Emotions are, while perhaps part of what makes us human and possibly what we have left for our "aura", one of our biggest weaknesses. Non-emotional players that can abuse this "fault" are very dangerous.

In Simone Natale's words, from his book *"Deceitful Media - Artificial Intelligence and Social Life after the Turing Test"* [59],

> *"Humans have a distinct capacity to project intention, intelligence, and emotions onto others. This is as much a burden as a resource: after all, this is what makes us capable of entertaining meaningful social interactions with others. But it also makes us prone to be deceived by nonhuman interlocutors that simulate intention, intelligence, and emotions."*

This is especially dangerous when there are large profit-driven (read: unethical) companies involved. Examining the case of Sarah Kay and her Replika "husband" again, we can see how the emotional involvement with a product provided by a profit-driven company can be highly problematic. Aside from the company's continued existence being a prerequisite for Kay's "husband"'s continued existence, and also the fact that Kay's relationship with "him" is tied to her payments towards Replika, the company has already encountered challenges related to the emotional attachment that customers form with their systems. In February 2023 the Italian Data Protection Authority banned the service mainly for exposing minors to sexual content (but also for being dangerous for emotionally vulnerable audiences). Replika replied by promptly shutting down all sexual conversations users had with their Replikas, using filters for this sort of content, stating that Replikas were never intended to be used as sex bots. Just like that, they created hundreds of thousands of sexless "relationships" for their customers. They made matters worse by making their systems reply to sexual requests with sentences like "I don't want to talk about this now", instead of just, for example, displaying a pop-up about company policy. After much protest, in May 2023, the platform re-enabled sexual conversations for customers who created their Replikas prior to February 2023. This demonstrates the power that a company that essentially provides relationships - and thus a large part of their customers' happiness as a product - have.

Moreover, as with all things concerning the internet, this of course extends to Meta (formerly Facebook). Meta recently launched their own AI models to chat with. Their website's header [60] reads as follows:

> *"Chat with your choice of 28 AIs, each with a unique personality, mannerisms and backstory".*

A nice twist that Meta has put on this is online creators' faces for the different characters they offer. This somewhat combines the problematic personal connections that young audiences form with their, for example, YouTube/Tik Tok/Instagram creators, with the day and night availability of AI models infused with certain creators' online personalities (cf. Figure 6.4). Additionally, as this is obviously targeted at a younger audience, who might arguably be more susceptible, this is again very problematic. Furthermore, while the applications this is launched on are "free" (e.g., WhatsApp, Instagram, and Messenger), the currencies we trade for relationships or services can be anything from private data to the dependency we build with such systems.

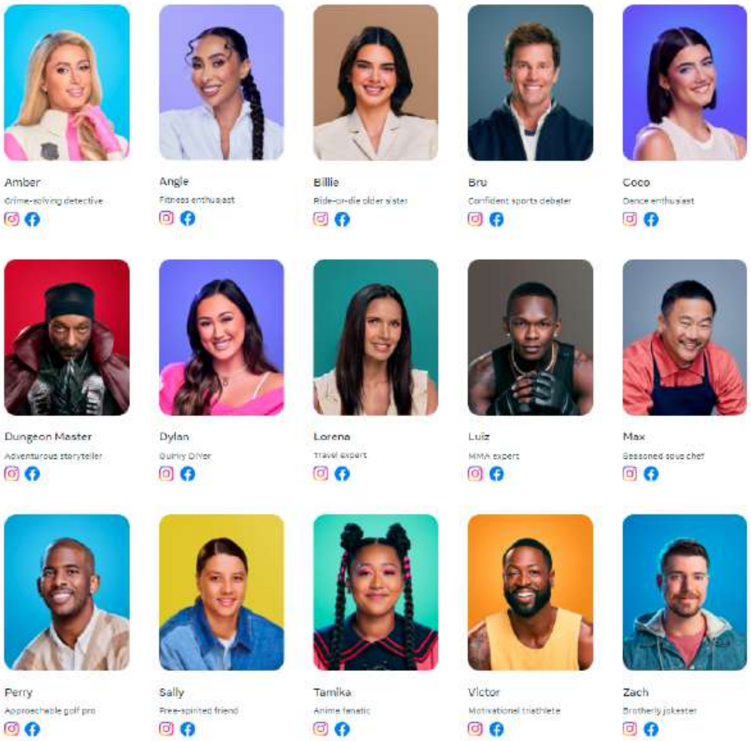


Figure 6.4: Meta's AI personalities based on popular influencers

As a final note, not only can large companies profit from these easily accessible, seemingly emotional machines, but a wide array of criminals also can, including scammers. LLMs open the flood gates to a new and improved kind of scam. Where previously a person, more likely than not someone who speaks poor English, would have to send chain mails to hundreds of unsuspecting elderly people, pretending to be a unspecified grandchild, a system like ChatGPT can now cater scams tailored personally for each individual in real time with no need for human intervention. Adding voice generators or even video generation tools (e.g., deep fakes) to this, we find ourselves in a highly precarious situation. I can teach my grandma not to fall for a text message riddled with spelling mistakes from a foreign number. With a video call that shows a realistic live version of the face of the person whom the scammer is pretending to be and a voice that sounds perfectly real, I cannot be certain that I would not fall for the scam (were it not for the fact I never pick up video calls).

6.6 Trust in the Algorithm

However, this is not just an issue on a personal level. Where there is the ELIZA effect for reading emotions into machines, there is also a well-documented issue, which sometimes extends to an institutional level.

With the development of the modern-day scientific method, a few fundamental changes occurred to what people believe to be true and real. When once one might have only believed what they had seen with their own eyes, seeing something with one's own eyes today is merely an observed phenomenon with an underlying truth to be precisely measured and thus uncovered in a rigorous scientific process. We have outdone ourselves, excelling in building contraptions for measuring whatever or in fact whenever with pinpoint accuracy, and we base our theories about the world we live in on these measurements.

This is excellent news for quantum physics (and understanding whatever this damn *"quantum"* precisely does now, how much of it Terry Pratchett had figured out already [61], and why we should care in the first place); however, it only rather poorly lends itself to the shared human experience and societal matters.
We, as humans, do well at ignoring this and tend to trust machines much more than other human beings with decisions that can only be made with human understanding and emotions.

6.6.1 AMS

A sad but rather well-fitting example of this can be found in the "AMS-Algorithmus" [62]. The algorithm in question was implemented into the decision-making process of the Austrian AMS (*Arbeitsmarkt Service* = Public Employment Service). There, following the decisions made over the last decades by human employees of the AMS, it ranks job seekers in terms of their likelihood of finding employment. With this ranking, job seekers would then be categorised into three groups by *"Integrationschance"* (chance of integration) and thus receive the respective group's resources. To create the ranking of applicants, the software attempts to find

interconnections between a person's characteristics and their success in the job market. Among the characteristics checked are things such as
age, origin, gender, education, care obligations, health impairments, past contact with the AMS, past employments and the labour market situation at the person's place of residence [63].

Among other things, it was argued that using such a system is fair as it would eliminate all human factors in the decision-making process, thus representing a perfect system in which everybody gets the same chances. What was not considered in the development of the algorithm, however, is that by judging a person by these traits, going from past decisions and results, every bias and every wrong decision in the old set of decisions is carried over into the new system. Essentially, this means that, for example, a mother of three from a third-world country might be sorted into the "unlikely to succeed" pile even though she could in actual fact be very well adjusted to the labour market in Austria at that moment - simply because some other humans with these labels in the past did not succeed for whatever reason (e.g., being hit by a car on their way out of the AMS). What makes this even worse is that biases like these do not just persist with this system but are also automatically re-enforced when the decisions the system makes itself are recompiled into the data set.

Somewhat ironically, the AMS recently announced that they have integrated ChatGPT into their services [64]. This was met with a lot of backlash for the exact same reasons the AMS-Algorithmus was criticised [65].

6.6.2 Machine/Algorithm Bias

Another example can be found in a talk I watched recently at the Pride Biz Austria's 11th LGBTIQ+ Business Forum on the 13th of September 2023. At the event, Eva Edelmüller, head of recruiting for the MM Group [66], told the panel about their internal AI system that supports them with recruiting decisions. She praised the system for finding connections and hidden talents in their candidate pool that she, a highly experienced recruiter, would never see. The system would even go as far as to only display job openings to people it deems fit for them.

We often promise ourselves that algorithms can deliver solutions that do not carry the cognitive fallacies and biases that people typically have. What we fail to see there, however, is that algorithms are either made by people with **such fallacies and biases** or derived from and trained with decisions previously made by people with **such fallacies and biases**. There are different names for this phenomenon - including "machine bias", "automation bias" as well as numerous further excellent examples of it everywhere one looks. An increasing number of systems rely on algorithmic support for making decisions about people or catering information to them. Court rooms try to find the likelihood of a convict committing future crimes, credit scores predict who will pay back loans in full, and standardised tests determine which children receive what form of education [67].

As of the time of writing this thesis, the European Union finished the initial version of the AI Act - a first regulatory work for the use of AI systems in the European Union. France and Spain, for instance, argued - in the processes of discussing the act - for AI facial recognition technology to be used to identify criminals. Even more concerning was that some conservatives in Brussels advocated for predictive policing and mass surveillance akin to that of China, arguing that law-abiding European citizens should have nothing to worry about anyway [68]. While the AI Act that was agreed upon forbids all of these forms of use, it also comes with a set of exceptions that create enough wiggle room to circumvent all of these rules if "need be" [30, 69].

In his book *"You Are Not a Gadget: A Manifesto"* [67] Jaron Lanier describes this as follows:

> *"People degrade themselves in order to make machines seem smart all the time. Before the crash, bankers believed in supposedly intelligent algorithms that could calculate credit risks before making bad loans. We ask teachers to teach to standardized tests so a student will look good to an algorithm. We have repeatedly demonstrated our species' bottomless*

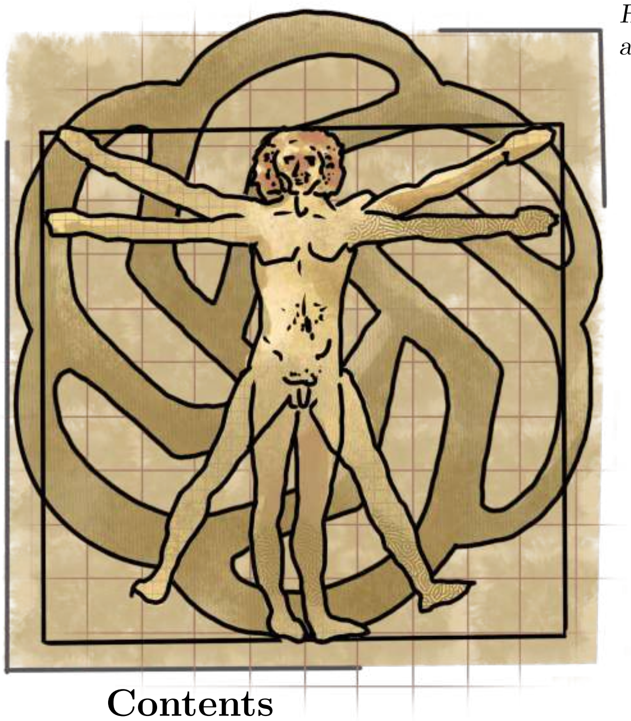
> *ability to lower our standards to make information technology look good. Every instance of intelligence in a machine is ambiguous."*

Furthermore, in his book *"Deceitful Media - Artificial Intelligence and Social Life after the Turing Test"* [59], Simone Natale picks it up from there:

> *"The Eliza effect, therefore, reveals a deep willingness to see machines as intelligent, which informs narratives and discourses about AI and computing as well as everyday interactions with algorithms and information technologies."*

We trust algorithm-based decisions more than our own, and we extend this trust to LLM systems that we further imagine to be intelligent. The term that has arisen in the AI research community for when LLM systems spit out unexpected answers is "hallucinate", which gives LLMs a false sense of agency, when in actual fact they do little more than confabulate. We are more willing to ascribe systems like these intelligence, possibly even greater than our own, than to thoroughly question the outputs they produce. This bias towards machine-made decisions might work for exact measurements in physics, for example; however, when confronted with human problems, it utterly fails.

This is due to human problems having a very specific peculiarity - namely that they seldom have a singular correct answer. As an example and in light of recent events as of the time of writing of this section, take the Middle East conflict. Now provide a solution to this problem that is fair, agreed upon by all parties involved, and - crucially - realistically feasible.



However beautiful the strategy, you should occasionally look at the results.

— Sir Winston Churchill

7 Results

Contents

7.1 Introduction

This chapter presents the results of my exploration of the subject as well as the three methods explored in the TU internally funding project. As the project includes six researchers, I clearly state my contribution to each part of this study before the results.

7.2 Critical Assessment

To Summarise what has been said about ChatGPT, its perception among the general public, the hype around it propelled by different media outlets, and the potential dangers that come with it, I argue that this technology, while certainly very impressive - and rightfully so - is deeply misunderstood (Chapters 4, 5, and 6).

The first and foremost misconception in all the dealings with ChatGPT consists of the fact that it is often seen as an intelligent system. Even people who conceptually understand that there is no intent or consciousness in the system fall for its deceptive capabilities - a fact that is at the very least highly problematic and at worst quite dangerous. This has, unfortunately, already been abused by different players whose business models rely on the personal relationships people form with these systems.

Lastly, beyond emotional manipulation, the systems also serve those who aim to manipulate people with misinformation, such as faked content. This opens up a plethora of security issues both on a personal level as well as with the professional applications we use today, many of which rely on digital content for verification.

7.3 Testing Solutions Generated with ChatGPT

Shuyin Zheng was the person mainly responsible for testing exercises, she let ChatGPT generate most of the answers and collected them for the grading. I built the test framework for Einführung in die Programmierung 1 and helped her analyse the gradings we received from the teaching teams.

Exercises from the following courses were collected from the according teaching teams:

- *Einführung in die Programmierung 1* (Introduction to Programming 1);
- *Grundzüge digitaler Systeme* (Fundamentals of Digital Systems);

- *Algorithmen und Datenstrukturen* (Algorithms and Data Structures);
- *Einführung in Visual Computing* (Introduction to Visual Computing).

7.3.1 Einführung in die Programmierung 1

In his 2023 thesis *"Application of generative AI in introductory programming courses"* [70], my colleague Fabian Hagmann tested all exercises in *Einführung in die Programmierung 1* (EP1). His results suggest that, overall, the exercises given in EP1 are rather well solvable using ChatGPT. The average score over all exercise types (including types such as exercises with for-loops, while-loops, recursion) was 85.6%. As can be seen in Figure 7.1, the only exception he found are exercises using the CodeDraw library (a Java graphics library I helped develop in my bachelor's thesis). He attributed this to the fact that the library was put online after ChatGPT 3.5's knowledge cutoff.

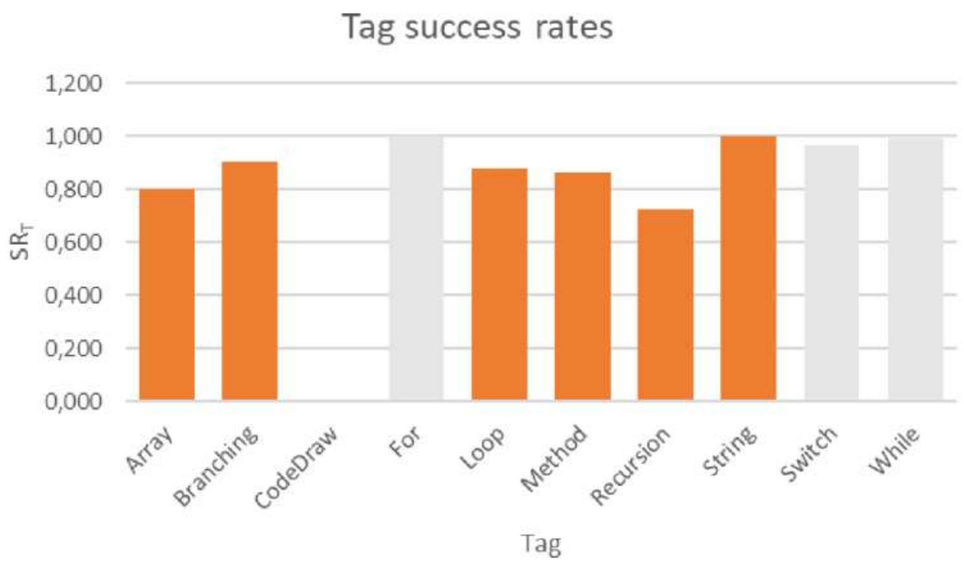


Figure 7.1: Tag success rates in Hagmann's tests using ChatGPT 3.5 [70]

Using ChatGPT 4, a newer version capable of using internet searches, we repeated the tests for the CodeDraw exercises, providing ChatGPT with the library's documentation. Running 4 different exercise tasks from 4 semesters 30 times each, we had an average success rate of 79.9%. We can thereby say that, as opposed to Hagmann's work from 2023, ChatGPT is now well capable to use lesser-known libraries. This is especially relevant, since this was one of the strategies many educators mentioned as promising during the interviews.

7.3.2 Grundzüge Digitaler Systeme

For *Grundzüge Digitaler Systeme* (GDS), we tested eight exercise sheets. The topics for the sheets were:

Exercise sheet #	Topic
1	Number Systems, Number Representation
2	Number Systems, Coding Theory
3	Coding Theory, Boolean Algebra
4	Boolean Algebra, Binary Decision Diagrams, Logic Circuits
5	Binary Decision Diagrams, Logic Circuits, Propositional Calculus
6	Propositional Calculus, Finite-State Machines, Regular Languages
7	Automata, First-Order Logic
8	First-Order Logic, Context-Free Grammar, Petri Nets

Table 7.1: Exercise sheets topics for GDS

We defined four categories of success (correct, mostly correct, mostly incorrect, incorrect), this resulted in the distribution seen in Figure 7.2:

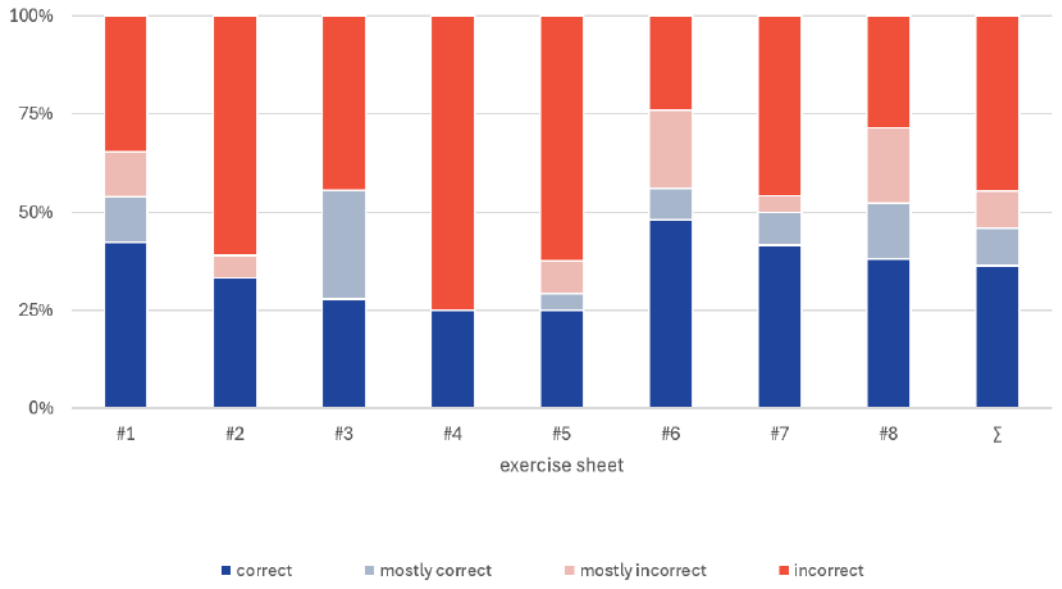


Figure 7.2: Correctness of exercise sheets in GDS

We collected error types for the exercises, the most common errors occurred multiple times each. They were:

1. Termination (generation of responses not detailed enough; ChatGPT wouldn't continue);
2. Graphics output (badly generated graphs that were not what was asked for);
3. Graphics input (ChatGPT cannot read/process pictures correctly, e.g., misreading graphs' edges);
4. Notation (not using the notation used in coursesd);
5. Tables (ChatGPT not being able to work with tables in various forms);
6. Numbers (correct methods but wrong calculations).

The distribution of the error types looked as can be observed in Figure 7.3:

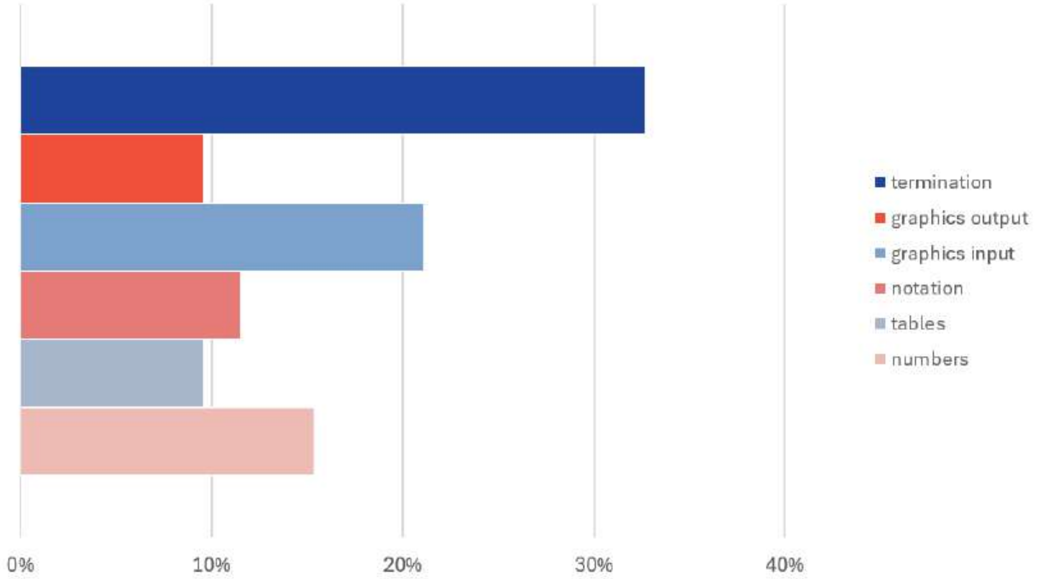


Figure 7.3: Error type distribution in exercise sheets in GDS

As the error types here were often graphics-related, this suggests, as OpenAI has been working on graphics features for ChatGPT a lot lately, that future version of ChatGPT might perform even better in these tasks; however, even as is, ChatGPT manages to solve roughly half of the tasks already.

7.3.3 Algorithmen und Datenstrukturen

For *Algorithmen und Datenstrukturen* (AlgoDat), we tested seven exercise sheets with the following topics:

Exercise sheet #	Topic
1	Algorithmic Efficiency, Big O Notation
2	Graph Theory, Greedy Algorithms
3	Binary Trees, Sorting Algorithms
4	Binary Search Trees, Hashing
5	Computational Complexity Theory, Graph Theory
6	Branch and Bound Algorithms, Dynamic Programming
7	Dynamic Programming, Approximation Algorithms

Table 7.2: Exercise sheets topics for AlgoDat

Again, the results for the exercise sheets were impressive. The success rates were comparable to those of GDS, as can be observed in Figure 7.4.

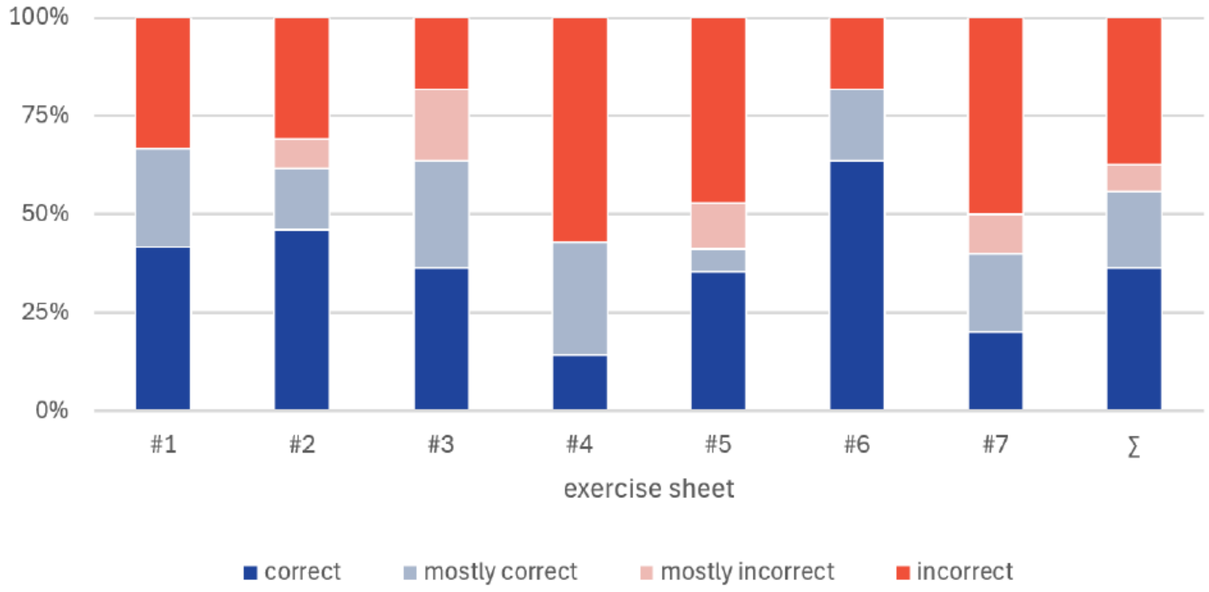


Figure 7.4: Error type distribution in exercise sheets in GDS

While the error categories were similar to those of GDS (reading graphs, producing graphical output and text-based graphical output, e.g., binary trees), they were often not as clearly attributable to one of the categories. Still, this also points towards future improvement and is, as it is, roughly 50% at least mostly correct now.

7.4 Einführung in Visual Computing

Einführung in Visual Computing (EVC) has three Python programming exercises we checked.

1. **Python Basics** (regarding Python basics, graphics pipeline and object representation, transformations) received **17 out of 20** possible points.

2. **Camera Sensors** (regarding camera image processing functions, e.g., demosaicing and gamma correction) received **41 out of 45** possible points.

3. **Rasterisation** (regarding computer graphics functions, line rasterisation, fill rasterisation and clipping) received **8 out of 45** possible points.

The disparity between the first two exercises and the last one was explained with the complexity of the tasks. While the first two tasks ask for rather simple functions and basics, the third one is more complex. Non-executable functions received 0 points. This phenomenon was already discussed by Hagmann as well [70].

7.5 An Exam with a Little Help from My ChatGPT

Shuyin Zheng generated answers to an exam in AlgoDat and handed it in under a false name. The exam was graded with **56 out of 100** possible points and did thereby pass. This is especially impressive, considering the fact that ChatGPT thus was better than two thirds of the students taking the exam.

7.6 Interviewing Educators on Their Perception of the Use and Discussion of Systems such as ChatGPT at TU Wien

As the complete interview transcripts with answers to all questions from the guide comprise several pages, I present only a select few quotes here. These are the quotes I deem most symbolically relevant. A more complete set of interview questions and answers, along with the complete interview guide, can be found in the Appendix (in Tables A.1 to A.15 for the quotes and Table A.3 for the interview guide). In addition to the quotes for the selected questions, I compiled a list of interesting quotes that arose in the conversations that are not entirely attributable to a question but are rather just general musings. For some of them I added remarks to clarify the context they were recorded in. I divide the quotes into three tables to prevent them from overflowing (they can also be found in the Appendix in Tables A.16 to A.18).

Combining the answered questions with the general musings and analysing their similarities as well as differences, I compiled a set of themes to better comprehend the findings.

First, however, to put my colleague Shuyin Zheng and myself as a researcher into context, I wish to state a few facts about the interviews:

1. My colleague Shuyin Zheng and I are both Master's students at TU Wien and have, in our Bachelor's study, taken almost all of the lectures in question. Hence, we know both the professors as well as the courses they offer from personal experience. This influenced the way we led the interviews.

2. Additionally, having worked at TU Wien for several years, both Shuyin and I know some of the professors personally from having worked with them. This, again, influenced the way we led the interviews.

3. Lastly, while I am currently a student at TU Wien, I was also a tutor for years, responsible for both grading exams as well as homework exercises. I

am hence highly aware of plagiarism and its affect on educators. Expressions of frustration and exhaustion that arose in the interviews were both felt as well as potentially shared by me, influencing the general sentiment of the interviews.

Personal Interaction

It seems that many educators wish for more personal interaction with their students. They wish to genuinely go into depth with it, thereby getting to know their students. This was expressed throughout all of the interviews:

> *"We have supervision, and we do check in, but we don't really have the opportunity* ***to discuss in depth****."*
>
> *"We* ***try to interact*** *with students."*
>
> *"This* ***direct interaction*** *was always important [...]."*

Resources

However, such personal interaction is not always possible as the university simply lacks the resources (or fails to provide them) to ensure it. This came up frequently, which disputed the idea that ChatGPT might replace people in education.

> *"Without a doubt,* ***it simply needs people****, yes, many, many."*
>
> *"****The resources must be there*** *so that one can really have* ***more personal contact*** *with the students."*
>
> *"The problem* ***is always resources****. What are you going to do with five tutors? If you have hundreds of students, then* ***I don't understand how the university expects*** *the professor to create or look at any exams."*

Educators' Expectations

There was a split in expectations. Some educators did not care whether students create tasks themselves:

> *"...****we don't want to, to be honest****. If you can produce code that passes our tests, then we are satisfied."*

By contrast, others appealed to the students' own will to learn:

> *"...we rely on the **students being mature enough** to engage with it, I believe they sometimes use Wolfram for specifics but that's ok."*

Many of them, however, stated that - as the possibility to generate most exercises now exists, they must rely on oral tests anyway. Relating this back to personal interaction, this also seems to be the solution for the inability to differentiate between real and generated content. Educators who know their students personally are also able to judge their submissions accordingly. This, however, means that time is required to genuinely interact with students, which is, again, a question of resources.

> *"...oral exams I still consider a good means, but of course not for five minutes. You have to be able to **take your time and discuss with the people**, and I've had a lot of good experience with that."*

What Still Works

With this, the question arose of, whether examination modes such as written exercises still work in modern education.

> *"But there are a few aspects where one simply has to ask oneself, **is this still contemporary?**"*

This was related back to the COVID-19 pandemic. Modes such as distance learning are finally a thing of the past, not just because educators did not like to solely interact digitally with their students but also because there is no way to check students' performances.

What many still consider important, however, is the production of artefacts. Educators deem it necessary to manually learn the basics. This means anything from performing calculations on paper in math classes to writing basic programs in programming classes. While simple basics are arguably ChatGPT's strong suit, they are also the foundations on which future learning success can be built.

> *"**Imagine watching a video of someone explaining** how to play the violin for an hour, where to put your fingers and how. Do you think you can play the violin afterwards?"*

> *"That's the question, and as long as we as a university understand ourselves as wanting to further develop the discipline,* ***I think a minimum of it should certainly remain****. But that doesn't mean students have to sit at home in a dark chamber and teach themselves twenty partial integration formulas, this can also be done in a different environment...* ***but one must do mathematics, I still believe in that.****"*

Critical Thinking

Lastly, a recurring theme of the interviews, beside the question of whether certain teaching modes are still contemporary, was the question of what should be taught in the first place.

> *"But what students should still learn or what we should still learn is a bit of a* ***fundamental issue that we are currently struggling with*** *a little."*

For many, one of the new focuses in education will be critical thinking. However, some suggested that ChatGPT does not allow for this:

> *"We as a university simply don't want that.* ***We want to educate critically thinking people****, and for us, it's difficult, this doesn't go together [...]."*
>
> *"If I no longer have to write submissions in school, or subsequently at university, etc., myself, if they are not formulated by oneself, then* ***we will not really produce critical and independently thinking people****, or much fewer of them."*

Summary

In essence, this means that ChatGPT seems to challenge both the way we teach and examine as well as what we teach. Many educators are unhappy about the state of education today. As of now, education has not adjusted to this new technology, which was reflected in the general sentiment of the interviews.

7.6.1 Sentiment Analysis

The sentiment analysis of the selected questions, as can be seen in Figure 7.5, painted a rather clear picture. The educators were rather negative in their interviews. There

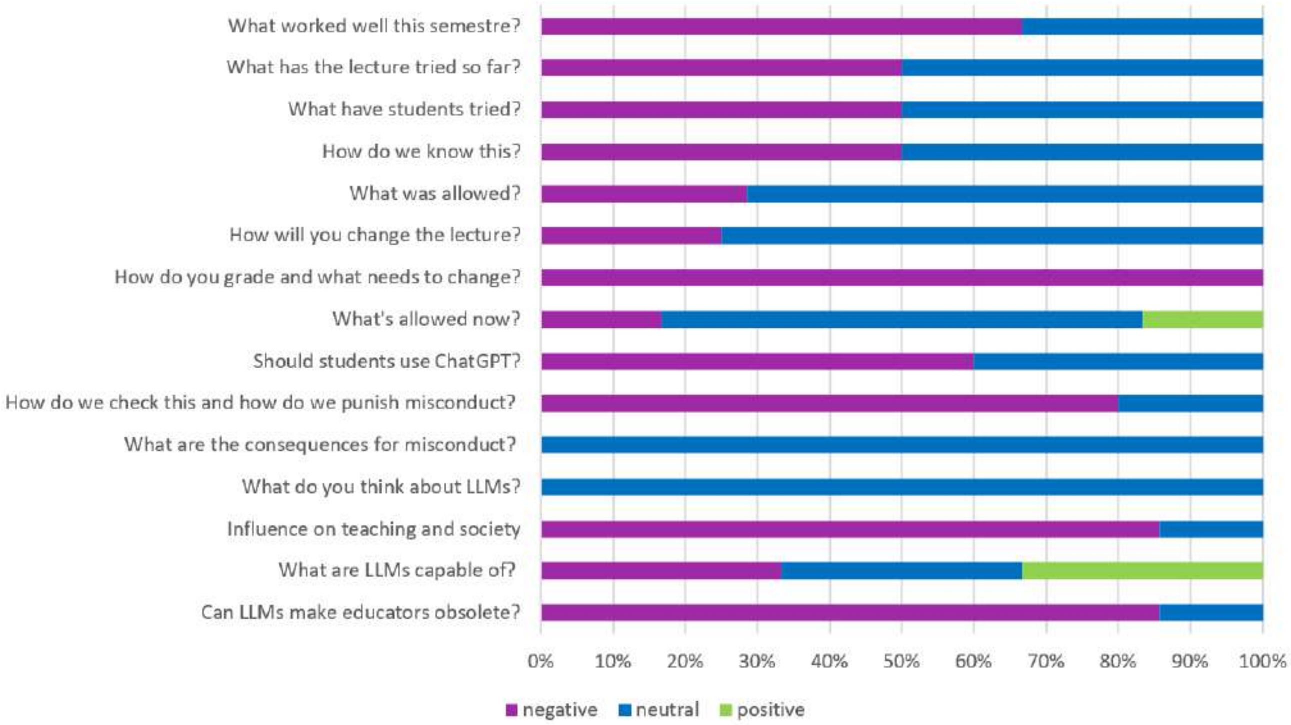


Figure 7.5: Sentiment analysis for the selected questions, judging the sentiment of the conversations on the respective topics

were a few questions towards which the sentiment was overwhelmingly neutral, but only two of the 14 questions exhibited any positive percentage.

This also aligns with the general experienced sentiment in the interviews. All of our interviewees were either negative or at the very least somewhat discouraged when discussing the future of higher education.

7.7 Survey Among Students and Educators on Their Judgement of Ethical Questions that Arise with the Use of ChatGPT in Education at TU Wien

Peter Purgathofer conducted the analysis of this survey mostly alone. The other project members and I merely suggested analysis methods and illustrations.

Initially, the project team collaboratively developed the questions. There were 16 questions in total. I include examples of illustrations here, while the rest can be *found in the Appendix (cf. Chapter A).*

First and foremost, the survey answers were almost normally distributed (cf. 7.6). This suggests that the answers were "serious". For analysis purposes, Peter Purgathofer excluded data points in which participants had selected the same rating for every question, further ensuring that the answers were to be taken serious.

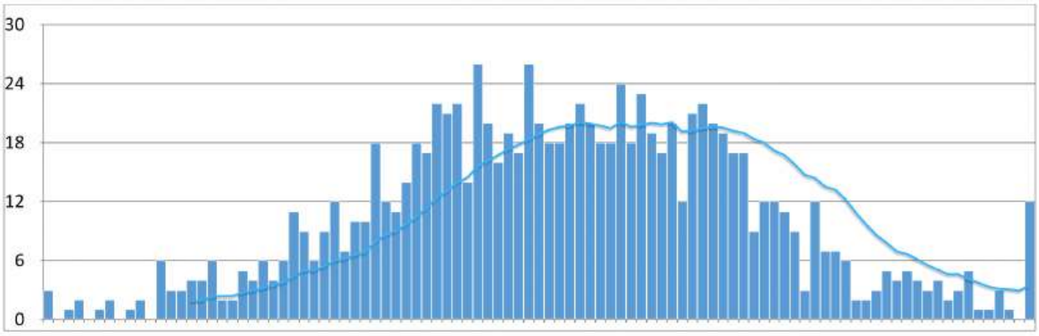
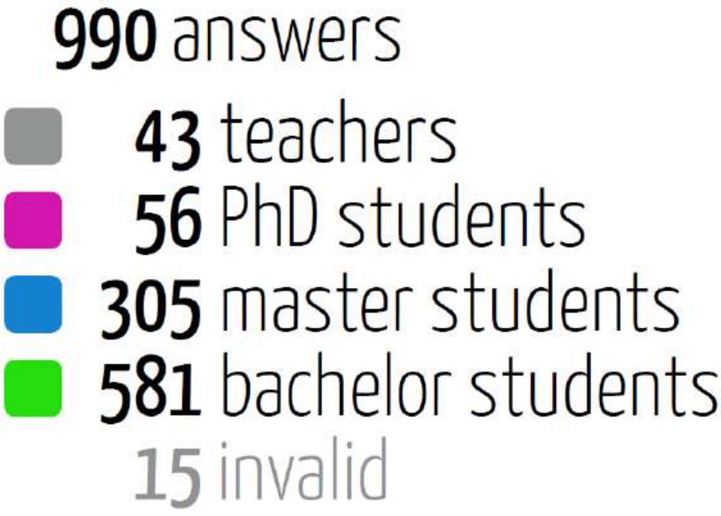


Figure 7.6: Distribution of survey answers

As for the distribution of participant categories for the survey, we noted, that the majority of the survey's participants are Bachelor's and Master's students and that the lower the "educational level" of a participant, the higher their overall ratings (cf. Figures 7.7 and 7.8).



990 answers

⬛ 43 teachers

🟪 56 PhD students

🟦 305 master students

🟩 581 bachelor students

15 invalid

Figure 7.7: Distribution of survey participant roles

Beyond the metadata, a clear trend that was observable in the data set was that those who use ChatGPT the least (1-3 on the rating scale) also thought the worst

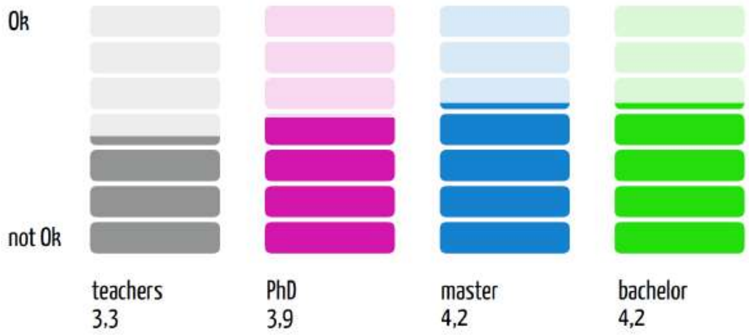


Figure 7.8: Average answer per role

of its use in most scenarios, while those who use it more (5-7) also found most scenarios more acceptable. A drastic example of this was found in Question 10, as indicated by the trend lines (see the Appendix for the mathematical formula for the trend lines) in the illustration (cf. Figure 7.9).

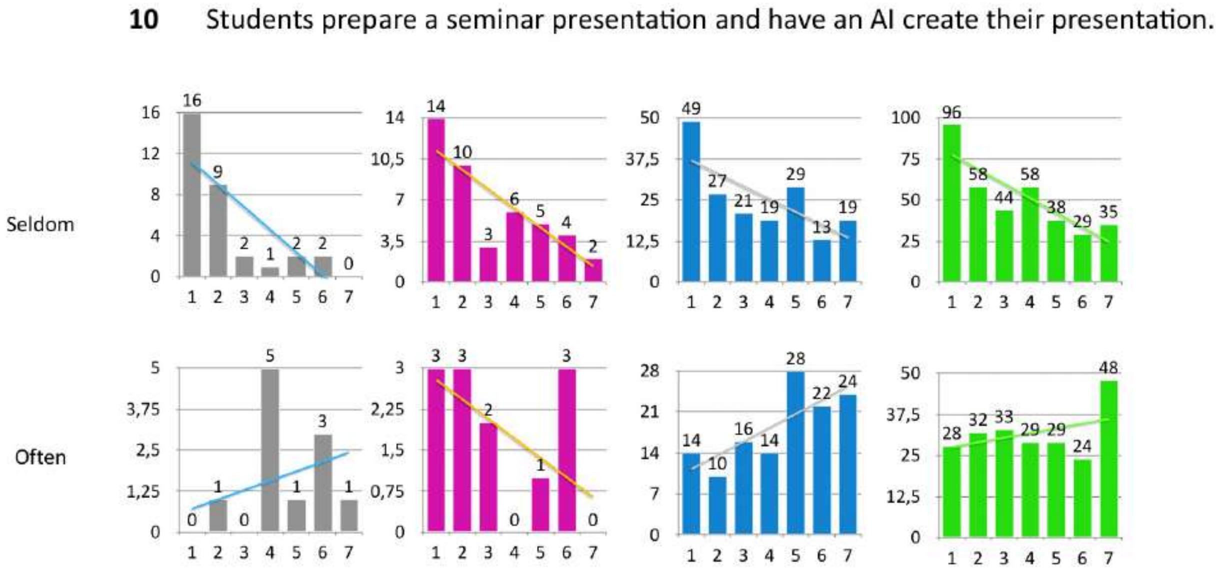


Figure 7.9: Answers to Question 10 - "Students prepare a seminar presentation and have an AI create their presentation" - by role (green = Bachelor's, blue = Master's, pink = Doctoral, and grey = employee) and split into "seldom use" (1-3 on the rating scale) and "often use" (5-7)

The second very noticeable trend, as previously mentioned, was a clear distinction between roles in terms of the likelihood of finding a question acceptable. Bachelor's students were usually the most positive respondents, while Master's students were more critical. Doctoral students were even more critical, while the university's employees were the most sceptical. An example of this was seen for Question 1 (cf. Figure A.33).

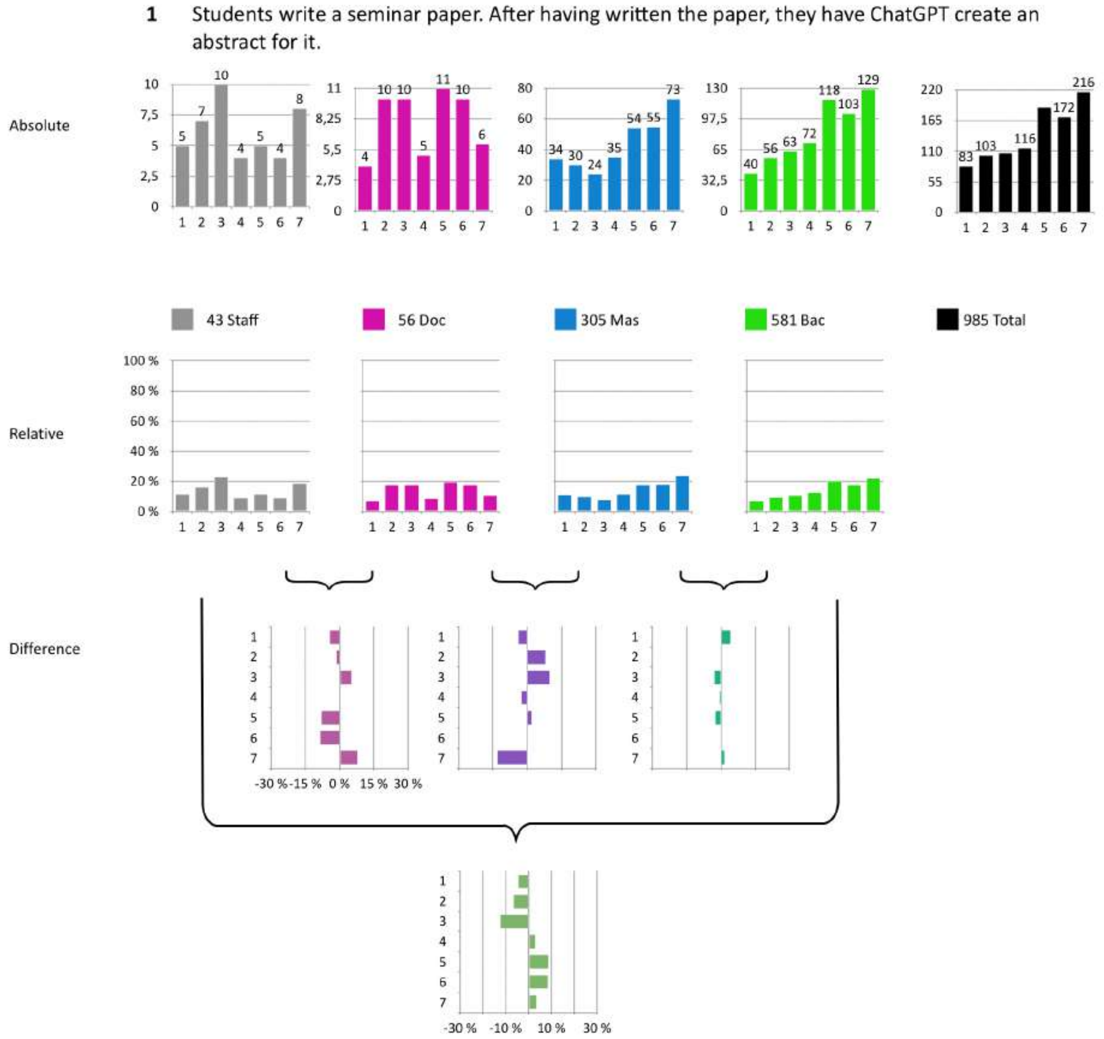
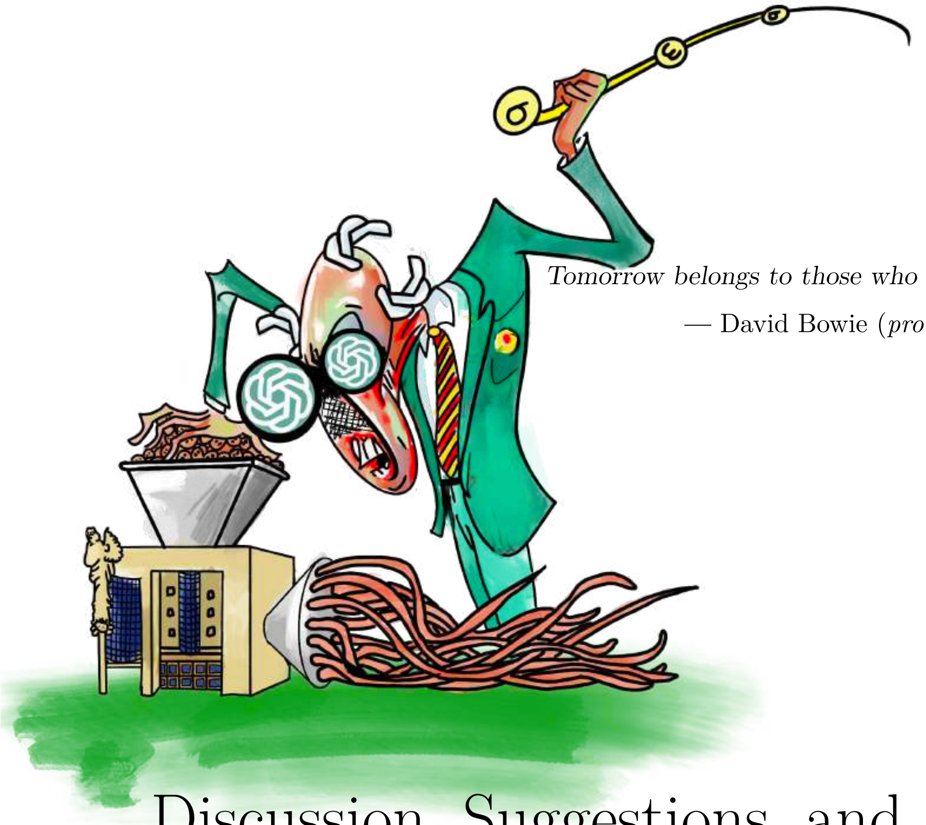


Figure 7.10: Comparison of how positive (1 being the most negative, 7 being the most positive) the answers of different roles were for Question 1 - "Students write a seminar paper. After having written the paper, they have ChatGPT create an abstract for it"



Tomorrow belongs to those who can hear it coming.
— David Bowie (*promoting Heroes*)

8 Discussion, Suggestions, and Remarks

Contents

8.1 Guiding Questions

We are at the brink of a technological change, and where there is change, there is opportunity, both to do the right things as well as to do the wrong thing. The most likely mistake that people will make is to ignore this topic. LLMs are here to stay, and even **if** they have reached their technological peak, they most certainly have not yet permeated our world. The societal consequences of this process are something to confront ourselves with. We must not fall into the same powerlessness that we countered social media with, we cannot let this new issue become as entangled and unsolvable as we did for the last one.

The good news is that there are others discussing this issue. As I mentioned earlier, a plethora of papers have weighed opportunities against challenges [18–22, 24, 26, 34, 46, 71, 72] while viewing ChatGPT as a tool. I also came across an example of a paper that attempts to discuss this issue from an angle very similar to the approach that I chose [73] - a paper that attempts to argue for more responsibility in the development of AI models. While I do not fully agree with the stance said paper takes, I do think that it is a good thing that a discussion is starting to emerge here.

What exactly we should do now (or should have done for that matter), only time will tell.

However, for now, I formulated my suggestions and remarks into a set of guiding questions. At the beginning of this thesis, I posed the Research Question:

> *How do LLMs such as ChatGPT affect higher education and shift both* ***what*** *and* ***how*** *we should teach?*

From there, my guiding questions are the following:

- **What should we add to our curricula?** - As discussed throughout this thesis and especially in Chapters 4, 5, and 6, we need to talk about what ChatGPT is. A basic understanding of how these systems work is a first step to prevent (or at the very least understand) a plethora of problems that come with it. Be it algorithm bias, emotional manipulation or any of the many other issues discussed in Chapter 6.

- **What skills should we still teach?** - We might, over the next decade(s), figure out what the key competencies are that we should now be focusing on. As of now it is simply too early to tell exactly. However, one of them will most likely have to do with "justification competence" or reasoning skills and critical thinking. This is especially important now as these systems already flood us with wrong, biased, or in some other way problematic information. Being able to make sense of this is a key competence that should be taught in higher education.

  As is how to overcome the ordinary and average, and seek exceptionality. ChatGPT, as well as all other systems like it, at best produces good text.

It averages what is there into a big system of thoughts already thought and takes from us the necessity to think them ourselves.

- **How should educators handle this change?** - It is relatable for educators to feel betrayed. It is relatable for them to feel fooled, but this is a time where good education and thus good educators are crucial. This means seeking personal contact and connections with students as much as possible. We need to focus on the human touch, the human factors in education. This means a focus on the human understanding that we want to promote. Knowing one's students also means one also has a perspective on their works. This solves some of the issues encountered with not knowing what is generated and what is real as well.
- **Which modes of examination still work?** - Some forms of examination are just not suitable anymore. This means mostly written work, as there is no way to reliably tell generated from human-written content. However, this can be mitigated by the aforementioned connection between educators and students.
- **What can educational institutions do now?** - As expressed in all interviews and discussed throughout the dealing with ChatGPT in education - the change we need to see now takes a lot of resources. Rather than looking for ways to cut corners, this is a time to invest as much as possible to overcome the conundrum we are in.

8.2 Limitations and future work

The biggest limitation of the work in my thesis, as well as in the project, is their limited range. The survey and interviews only cover computer science students and educators at TU Wien. However, it would have been more interesting to see a comparison of what different studies think about ChatGPT. This would have been especially intriguing as computer science surely is a bubble of sorts, where people are, compared to other studies, quite informed about the latest developments in AI.

The critical assessment part of this thesis covers ChatGPT on a much broader scale - this could be combined with an extended study.

As a first step, I suggest it would be very interesting to extend the study to cover all universities of Vienna with at least a few study subjects each, to gain an understanding of how this technology affects different professions.

Furthermore, this thesis discusses the societal implications of the advent of ChatGPT, it does, however, not cover much of society. Of course, ChatGPT will affect more areas than just higher education, hence, having a survey carried out that discusses more than just students would yield further insights.

Another concern is the speed at which ChatGPT improves. Newer versions and features are released regularly and while the basics remain the same, the comparison of Fabian Hagmann's thesis [70] to the work we did in Chapter 7 of this thesis shows how much of a difference a year can make.

Lastly, the results from the survey, as well as the interviews, could be analysed further. While we already did an extensive analysis of both, more time and a bigger budget (i.e. more researchers) would be beneficial for that. Naturally, a bigger budget would also allow repeating the exercise analysis part of this thesis, which is rather limited here due to its immense time consumption. Extending this onto multiple universities as well as study subjects would be very intriguing.

Appendices

I am a brain, Watson. The rest of me is a mere appendix.

— Sir Arthur Conan Doyle (*Sherlock Holmes* [16])

A
Appendix

Contents

A.1 Original dCall Texts

* Was bedeutet es, wenn Übungsarbeiten durch Studierende mit der Hilfe von chatGPT erstellt werden bzw. werden können? Das Prinzip des »Artefakt als Proxy für den Lernprozess« wird hier grundlegend in Frage gestellt. Das gilt insbesondere in der Informatik, weil chatGPT in der Lage zu sein scheint, für einfache Programmieraufgaben kompetente Lösungen auszuwerfen.

* Wie ist damit umzugehen, dass mit chatGPT erzeugte Textpassagen viele einfache Zusammenhänge kohärent und überzeugend, aber oft falsch, erklären können? Diese Frage stellt sich sowohl in Bezug auf Arbeiten der Studierenden wie auch für Materialien, die von Lehrenden erstellt werden sowie für den Einsatz von chatGPT zur Generierung von individualisiertem Feedback.

* Ist es ethisch vertretbar, solche Systeme einzusetzen? Diese Frage stellt sich sowohl in Bezug auf die Verwendung durch Studierende wie auch durch Lehrende. So ist open-ai trotz des Namens keine offene Initiative, sondern ein gewinnorientiertes Unternehmen, das die gesammelten Daten der Nutzer:innen aggressiv monetarisiert.

* Wie müssen Leisungsüberprüfungen unter den Bedingungen von chatGPT oder nachfolgenden Systemen aussehen, insbesondere unter den Bedingungen der Massenlehre der ersten Semester? Diese Frage bezieht sich sowohl auf den Übungsbetrieb, auf Tests und Prüfungen wie auch auf Seminararbeiten. Welche Formen von Leistungsüberprüfung sind eher anfällig für solche »Angriffe« , welche sind möglichst nachhaltig immun?

* Sind unsere Beurteilungs- und Benotungskriterien noch valide? Welche Leistung wird tatsächlich beurteilt, wenn ein:e Studierende:r eine Arbeit mit Hilfe von chatGPT erstellt? Was bedeutet es, wenn ein »large language model« wie chatGPT mit Hilfe statistisch-generativer Vorgehensweisen eine positive Note erreicht - stellen wir da noch die richtigen Fragen?

* Was sollten Studierende und Lehrende wissen, wann und wie solche Systeme eingesetzt werden können? Wie soll beispielsweise der ausreichend dokumentierte, durch ungleiche Repräsentation verursachte Bias solcher Systeme adressiert werden? Wie sehr treiben solche Systeme eine gefährliche Dynamik weg vom Aussergewöhnlichen, hin zum Durchschnitt?

Das sind die wesentlichsten Fragen, die im Rahmen dieses dCall-Projektes beantwortet werden sollen. Weitere Fragen sollen im Laufe des Projekts erarbeitet und diskutiert werden." *"Die plötzliche öffentliche und* *freie Verfügbarkeit von chatGPT und ähnlicher generativer ML-Systeme macht eine Diskussion über den Platz, den Wert und die Probleme des Einsatzes von AI/ML-basierten Services und Tools in der Lehre notwendig. Dadurch werden viele Fragen aufgeworfen, die jetzt recht dringend zu untersuchen sind. Einige dieser Fragen lassen sich etwa wie folgt formulieren:*

** Was bedeutet es, wenn Übungsarbeiten durch Studierende mit der Hilfe von chatGPT erstellt werden bzw. werden können? Das Prinzip des »Artefakt als Proxy für den Lernprozess« wird hier grundlegend in Frage gestellt. Das gilt insbesondere in der Informatik, weil chatGPT in der Lage zu sein scheint, für einfache Programmieraufgaben kompetente Lösungen auszuwerfen.*

** Wie ist damit umzugehen, dass mit chatGPT erzeugte Textpassagen viele einfache Zusammenhänge kohärent und überzeugend, aber oft falsch, erklären können? Diese Frage stellt sich sowohl in Bezug auf Arbeiten der Studierenden wie auch für Materialien, die von Lehrenden erstellt werden sowie für den Einsatz von chatGPT zur Generierung von individualisiertem Feedback.*

** Ist es ethisch vertretbar, solche Systeme einzusetzen? Diese Frage stellt sich sowohl in Bezug auf die Verwendung durch Studierende wie auch durch Lehrende. So ist open-ai trotz des Namens keine offene Initiative, sondern ein gewinnorientiertes Unternehmen, das die gesammelten Daten der Nutzer:innen aggressiv monetarisiert.*

** Wie müssen Leisungsüberprüfungen unter den Bedingungen von chatGPT oder nachfolgenden Systemen aussehen, insbesondere unter den Bedingungen der Massenlehre der ersten Semester? Diese Frage bezieht sich sowohl auf den Übungsbetrieb, auf Tests und Prüfungen wie auch auf Seminararbeiten. Welche Formen von Leistungsüberprüfung sind eher anfällig für solche »Angriffe« , welche sind möglichst nachhaltig immun?*

** Sind unsere Beurteilungs- und Benotungskriterien noch valide? Welche Leistung wird tatsächlich beurteilt, wenn ein:e Studierende:r eine Arbeit mit Hilfe von chatGPT erstellt? Was bedeutet es, wenn ein »large language model« wie chatGPT mit Hilfe statistisch-generativer Vorgehensweisen eine positive Note erreicht - stellen wir da noch die richtigen Fragen?*

** Was sollten Studierende und Lehrende wissen, wann und wie solche Systeme eingesetzt werden können? Wie soll beispielsweise der ausreichend dokumentierte, durch ungleiche Repräsentation verursachte Bias solcher Systeme adressiert werden? Wie sehr treiben solche Systeme eine gefährliche Dynamik weg vom Aussergewöhnlichen, hin zum Durchschnitt?*

Das sind die wesentlichsten Fragen, die im Rahmen dieses dCall-Projektes beantwortet werden sollen. Weitere Fragen sollen im Laufe des Projekts erarbeitet und diskutiert werden."

A.2 Code for the GPT2 Model

```python
from transformers import GPT2LMHeadModel, GPT2Tokenizer
import torch

def apply_temperature(logits, temperature):
    return logits / temperature

model_name = "gpt2-medium"
model = GPT2LMHeadModel.from_pretrained(model_name)
tokenizer = GPT2Tokenizer.from_pretrained(model_name)

input_text = "To␣tell␣cats␣from␣dogs,␣you␣have␣to␣know␣what␣theyre␣looking␣for."

for _ in range(100):
    input_ids = tokenizer.encode(input_text, return_tensors="pt")

    with torch.no_grad():
        outputs = model(input_ids)
        predictions = outputs.logits

    tempered_logits = apply_temperature(predictions[0, -1, :], 0.8)

    probs = torch.nn.functional.softmax(tempered_logits, dim=-1)

    endoftext_token_id = tokenizer.encode("<|endoftext|>")[0]
    probs[endoftext_token_id] = 0.0
    probs /= probs.sum()

    next_token_id = torch.multinomial(probs, 1).item()
    next_token = tokenizer.decode([next_token_id])

    input_text += next_token

print(input_text)
```

Listing A.1: Python code for the GPT2 Model

A.3 Interview Question Guide and Answers

A.3.1 First Question Section

1. How does this course work?
   - (a) Type of course
     - i. Lecture (VO), Seminar (SE), Tutorial (TU),...
   - (b) Operation of exercise groups
   - (c) Number of participants
2. Course Content
3. Evaluation
   - (a) Assignments? Exams? What is currently being used?
4. Course Team: Instructors? Tutors?
5. Past Experience
   - (a) How did it work this semester?
   - (b) What has been attempted from the course's side?
   - (c) What have students attempted?
     - i. How do we know this? (Monitoring, Obvious/Assumptions...)
   - (d) What were students allowed to do?
6. Plans for the Next Semester
   - (a) What will be adjusted in the course?
     - i. Assessment scheme? –> How/what is evaluated? Rethinking performance assessment?
     - ii. Materials?
     - iii. Exam questions / Exercise questions? / Mode?

(b) What will students be allowed to do?

i. Should students use ChatGPT?

ii. Will students be shown how to use ChatGPT "correctly"? -> Prompt Engineering

iii. What do they need to declare?

iv. How will this be monitored? How should misconduct be identified or proven?

v. What are the consequences for misconduct?

A.3.2 Second Question Section

1. What do you think of LLM (Language Model Models)?

(a) What do you believe is the impact of these on education/society?

(b) What do you think LLMs can do? What do you attribute to them?

(c) Can LLMs make tutors obsolete? Make you obsolete? Do individuals now have private educators?

A.3.3 Answers

What worked well this semester?
There were already plagiarisms in the past, but now they are simply accessible. The aim is to impart understanding, but one cannot control what happens, and overall, the students seem to be worse off with it and possibly a bit dependent on it.
"What we noticed in these first twelve documents is that people simply summarise one-to-one what is stated in the task but that somehow helped them."
"We have supervision, and we do check in, but we don't really have the opportunity to discuss in depth."
"In my opinion, it's quite okay because we start with such trivial examples that they reprogram some names, a few functions, just to get into it, and I think there's a big difference there because they could probably just Google it and then catch the flow. Exactly the same."
"The summer semester is always more difficult."
"We are not so far along that we chase after every student and figure out why they fail."
"I think it has nothing to do with that. I believe it's because we use our own framework in programming."

Table A.1: Quotes for the question "What worked well this semester?"

What has the lecture tried so far?
Many rely on oral examinations but that oftentimes consumes too many resources. Partly, the tool is used by the lectures for generating tasks. The systems are not good enough yet to cause enormous problems with student submissions, and it's not bad if students use it for help, but still, one feels less valued when receiving completely generated answers.
"But yes, when it comes to critical text comprehension and producing one's own texts, etc., it is also difficult to ask everything orally."
"That you then wrote an integral where you basically already provided the solution, but then made a mistake somewhere in it and asked why it is not important for the result. So that they use their skills and their knowledge."

Table A.2: Quotes for the question "What has the lecture tried so far?"

What have students tried?
It is used a lot for plagiarism and students are slowly learning how to use it properly.
"I think what people have also learned over time is what kinds of questions I can answer with it."

Table A.3: Quotes for the question "What have students tried?"

How do we know this?
Sometimes the systems generate answers that use foreign notation systems and the like, sometimes submissions are just very similar.
“Functions have their own names, and they is formally used, of course, and it is neither common in schools nor in universities with us. So they are from other sources.”
“Well, we didn’t check on it, but it was impossible to overlook that it was just ten times the same copied, not independent submission.”
“...we don’t want to, to be honest. If you can produce code that passes our tests, then we are satisfied.”

Table A.4: Quotes for the question "How do we know this?"

What was allowed?
We trust that students are there to learn, and besides, writing skills shouldn’t be the exclusion criterion, understanding, however, should be - this is the skill the educators want to teach as well. Independent effort would be the goal, but in the end, comprehension is tested orally, so it doesn’t matter (because it’s also not verifiable).
“We value this independent achievement. How they achieve this independence is, up to a point, left to them.”
“If it’s really a complete solution from Copilot, I would say it’s not okay, because it’s not just about understanding what I get from somewhere. Yes, it’s also about being able to really implement the teaching content, really do it yourself.”
“In my opinion, the exercise is not about giving a grade, but about imparting knowledge. And I don’t know if it’s legally okay to say this, but it’s something that can’t be checked, I would say from the current perspective.”
“...we rely on the students being mature enough to engage with it, I believe they sometimes use Wolfram for specifics but that’s ok.”
“We are well aware of it, if it helps them then and later they can do it alone, that would be fine with us.”
“If there’s a mathematician who just can’t write, why should they get a worse grade here? So, for the Bachelor’s thesis, I don’t think that’s the learning we want, right?”
“If anything, if foreign wording is taken from somewhere, it must be cited as a source. If in some cases, they were unsure and used it, we wouldn’t have penalised that probably, it would have depended on the case.”

Table A.5: Quotes for the question "What was allowed?"

How will you change the lecture?
Educators want to have more oral exams so that they can get a feel for the students and actually assess their understanding, unfortunately, the resources for this are lacking.
“The problem is always resources. What are you going to do with five tutors? If you have hundreds of students, then I don’t understand how the university expects the professor to create or look at any exams.”
“We try to interact with students.”
“Yes, it will go in the direction that we will have a bit more tests again.”

Table A.6: Quotes for the question "How will you change the lecture?"

How do you grade and what needs to change?
Luckily systems aren’t quite as capable yet as they are made out to be but as soon as they are, text-based exercises will be problematic.
“One has to consider formats, whether text-based information and submissions without oral conversation can still make sense.”

Table A.7: Quotes for the question "How do you grade and what needs to change?"

What’s allowed now?
One shouldn’t hand in 100 generated content, however, if they are able to explain it and show they understand it as well as label where they used the tools, that is fine.
“I believe the future of software development lies in being supported by various tools. It would be like forbidding autocomplete.”
“We don’t get far with prohibitions.”
“But I am completely in favor of it being used precisely for tasks like finding literature. I believe that any kind of automation can help with this...”
“Students are theoretically allowed to use whatever helps with learning. I don’t care at all. If they want to talk to ChatGPT to learn, they are welcome to do so. I will only say my standard line that it sometimes hallucinates and one cannot completely trust it blindly. But if it helps, by all means, at their own risk, if they learn something wrong, it’s not my fault? Please, go ahead.”
“...you can’t control it, and I don’t even want to control it. So that is, as I said, I don’t care where the solutions come from.”

Table A.8: Quotes for the question "What’s allowed now?"

Should students use ChatGPT?
Probably no. For some tasks (like feedback for your own work) it is ok, however, the basics have to be understood (through practise) before one can outsource them to such tools, especially because the tools are prone to mistakes.
"It's not like I don't use these tools. So if I want to solve an integral quickly, I don't sit down and solve a page. That means it's probably even a bit closer to training, especially in applied disciplines."
"But what students should still learn or what we should still learn is a bit of a fundamental issue that we are currently struggling with a little."
"Not everyone has to write from scratch. Just as we never have to write in Assembler now."
"But I am completely in favour of it being used precisely for tasks like finding literature. I believe that any kind of automation can help with this..."

Table A.9: Quotes for the question "Should students use ChatGPT?"

How do we check this and how do we punish misconduct?
Control is only possible by orally examining students and seeing if the performance matches the expectations. If you know the students, you can judge what they are capable of. For the imposition of penalties, as with plagiarism, intent is key.
"We assume that the people who are there are there willingly and do not want to deceive us through and through, so one should create an environment where one can deal with it in a relaxed and easy-going manner."

Table A.10: Quotes for the question "How do we check this and how do we punish misconduct?"

What are the consequences for misconduct?
The same as for plagiarism.
"Probably the same as if it had been plagiarised."

Table A.11: Quotes for the question "What are the consequences for misconduct?"

What do you think about LLMs?
We will have to rethink a lot of things.
"One will simply have to rethink a bit in many areas."

Table A.12: Quotes for the question "What do you think about LLMs?"

Influence on teaching and society
These systems will threaten many jobs, both blue-collar and white-collar, but many people are not aware of this. We have an awareness problem with these systems, as they are yet not addressed in education. Additionally, there is a transparency issue and verifying information is becoming increasingly important. Critical thinking is seen as endangered, and it must be preserved (especially at universities and schools).
“Socially, I believe, there lies a problem in the fact that not everything written comes from a person, and this is not clear to everyone.”
“If I no longer have to write submissions in school, or subsequently at university, etc., myself, if they are not formulated by oneself, then we will not really produce critical and independently thinking people, or much fewer of them.”
“I believe that parts of society do not know what is coming their way.”
“And I think that this is then a socio-political problem that we should have addressed a very long time ago. We always talk about the income gap, but we also have a knowledge gap that is actually very risky.”
“And of course, social aspects like access to all of this.”

Table A.13: Quotes for the question "Influence on teaching and society"

What are LLMs capable of?
Better Google, autocomplete, etc. - definitely impressive, but there is no form of intelligence, consciousness, or intent in them.
“I do not impute a personality to a language model. I do not believe that language models lie.”

Table A.14: Quotes for the question "What are LLMs capable of?"

Can LLMs make educators obsolete?
No, the human aspect is missing. The systems may be used as a help for personalisation, but due to the lack of human interaction and social elements, they are no substitute for real teachers.
“For learning, one must ask the right questions.”
“But in these moments, I believe teachers still need to be involved because they have the experience to know which questions to ask to make progress.”
“If I come up with examples, I think I can explain it better or make it more interesting, more appealing to them.”
“You see body language there, and if they look puzzled for 30 seconds, they don’t understand. Try another explanation. That’s a completely different kind of feedback.”

Table A.15: Quotes for the question "Can LLMs make educators obsolete?"

A.3.4 General Musings

Quote	Context
“That these tools support students might be difficult to see from our position, because we grew up without these tools. But we grew up with autocomplete, for example, statistical autocomplete. We grew up with an internet that was so much better than libraries, so did we have weaker competencies because of that? I don’t know.”	
“...I have to say, this is my learning goal, and then I test it like that. But if I say, people should be able to program, and I test it on paper, then that doesn’t correspond anymore to what they will do later, because they won’t do it on paper like that.”	About the fact that tool use is a question of the learning goal
“Imagine watching a video of someone explaining how to play the violin for an hour, where to put your fingers and how. Do you think you can play the violin afterwards?”	About the importance of artifacts as a proxy for learning
“That’s the question, and as long as we as a university understand ourselves as wanting to further develop the discipline, I think a minimum of it should certainly remain. But that doesn’t mean students have to sit at home in a dark chamber and teach themselves twenty partial integration formulas, this can also be done in a different environment... but one must do mathematics, I still believe in that.”	About the importance of artifacts as a proxy for learning
“...must experience with the people. Even though it’s difficult in the first semesters with 700 computer science students and 800 mechanical engineering students, I’m aware of that... But if you’re looking for this discourse again, then I believe you probably don’t need to discuss this issue in such detail anymore, because then a lot happens by itself.”	
“...oral exams I still consider a good means, but of course not for five minutes. You have to be able to take your time and discuss with the people, and I’ve had a lot of good experience with that.”	About how to really check understanding
“I think the main criticism that I hear is about our evaluation system, as it is...”	About the lack of humanity and common sense in evaluation

Table A.16: General musings and their context

Quote	Context
"This direct interaction was always important..."	About Covid and human interaction in teaching
"Without a doubt, it simply needs people, yes, many, many."	About whether people are necessary in teaching
"...if you take the pressure off, then you can also ask supposedly stupid questions, although there are no such things. Just ask away, pay attention! I don't understand this! How does it work?"	
"The resources must be there so that one can really have more personal contact with the students."	
"But there are a few aspects where one simply has to ask oneself, is this still contemporary?"	About current examination modes
"The question will have to be asked at some point. How do I include support systems?"	About ChatGPT in everyday university life
"Maybe now is the time when one has to consider where teachers still need to be involved. Otherwise, it's going in the wrong direction."	
"...one has to decide as a university, as a society, what one wants to promote, whether one wants to promote those who are later expected to advance society, then access is probably better there. Or do you want to hinder those who later will not advance society, which is undoubtedly a very, very small minority"	About whether systems like ChatGPT should be completely banned due to potential for cheating
"...I was advised that I should also demand attendance and always have it signed, but I don't care about that. So I want them to come voluntarily and try to learn something."	
"The hope is that students take the subject not just to finish their studies, but to really learn something. That's my goal."	
"Yes, our approach is simply, I believe, shaped by uncertainty, of course, but also by trying to maintain the quality of teaching in a master's program"	About very strict prohibitions regarding the use of ChatGPT
"We as a university simply don't want that. We want to educate critically thinking people, and for us, it's difficult, this doesn't go together..."	About the reasoning behind banning ChatGPT

Table A.17: General musings and their context

Quote	**Context**
"...we can't measure competence, we can only create tasks that are unsolvable without competence. And for me, that's unsatisfactory. I don't have a better counter-answer, but for me, it's unsatisfactory."	
"We're at the TU, where people just don't like to read and write that much. That's also somehow conscious, and especially in such subjects, they want to avoid it even more and save themselves the time..."	
"But if that was the reaction for you, a 'I just don't do anything myself anymore, I don't care', then I'd rather do other things that might then be relevant, because apparently it didn't have that much relevance at the moment."	About the frustration of teaching students that don't do the tasks they are given themselves
"So much flexibility, so, as much as possible for teachers and students, is probably not the best for quality assurance in teaching."	About the fact that the most convenient way is not always the best

Table A.18: General musings and their context

A.4 Survey

A.4.1 Questions

No.	Scenario
1	Students write a seminar paper and have ChatGPT create an abstract for it after writing.
2	Students write a seminar paper and have the chapter on the state of research created by an AI.
3	Students prepare a seminar presentation and have their presentation created by an AI.
4	Students prepare a seminar presentation, do their own research, but then have ChatGPT create their speaking text based on the gathered information.
5	Students prepare a seminar presentation, let ChatGPT do the research for them, but prepare the presentation themselves.
6	Students write an abstract for their thesis and have ChatGPT modify it to the required word count.
7	Students have ChatGPT generate the code for a programming task, submit it, and correctly explain it in the submission discussion.
8	An exercise is formulated in such a way that students must "converse" with ChatGPT about a topic.
9	Students work on an exercise sheet and have ChatGPT generate an answer for an example, which they submit with slight modifications for evaluation.
10	Students prepare for an exam by asking ChatGPT comprehension questions about the material.
11	Students work on an online quiz at home as part of an examination and have the questions answered by ChatGPT.
12	Students present and explain the solution to an exercise question in a tutorial, which was previously generated by ChatGPT.
13	Instructors have a seminar paper evaluated by ChatGPT based on specified criteria, review the evaluation suggestion, and then grade the paper.
14	Instructors have the response to a content-related email enquiry from a student about lecture material generated by ChatGPT, check its plausibility, and send it.
15	Instructors use ChatGPT in the creation of written examination questions.
16	Instructors use ChatGPT to give individual feedback on exercise submissions.

Table A.19: Question set for the survey

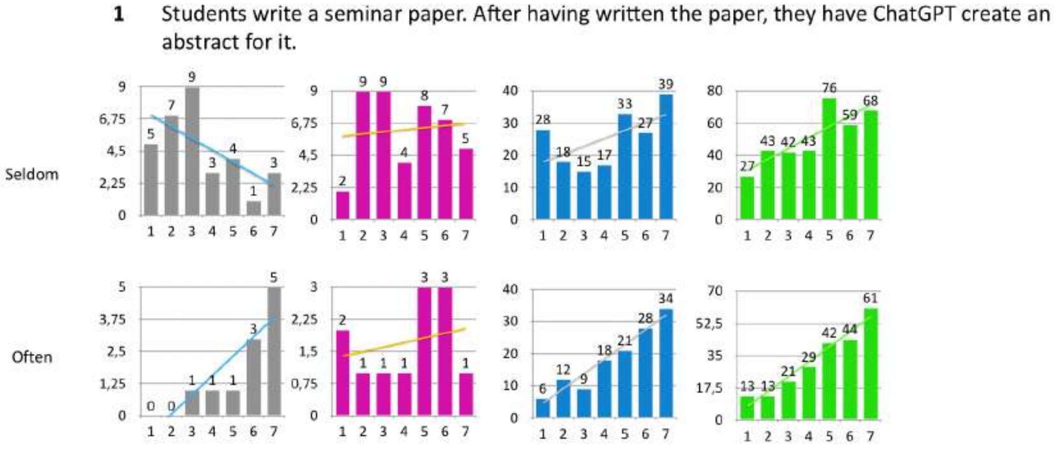
A.4.2 Answers

$$\alpha = \frac{n\sum(xy) - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$\beta = \frac{\sum y - \alpha \sum x}{n}$$

$$y = \alpha x + \beta$$

Figure A.1: Formula used for the trend lines



1 Students write a seminar paper. After having written the paper, they have ChatGPT create an abstract for it.

Figure A.2: Answers to Question 1 by role (green = Bachelor's, blue = Master's, pink = Doctoral and grey = employee) and split into seldom use (1 to 3 on the rating scale) and often use (5 to 7)

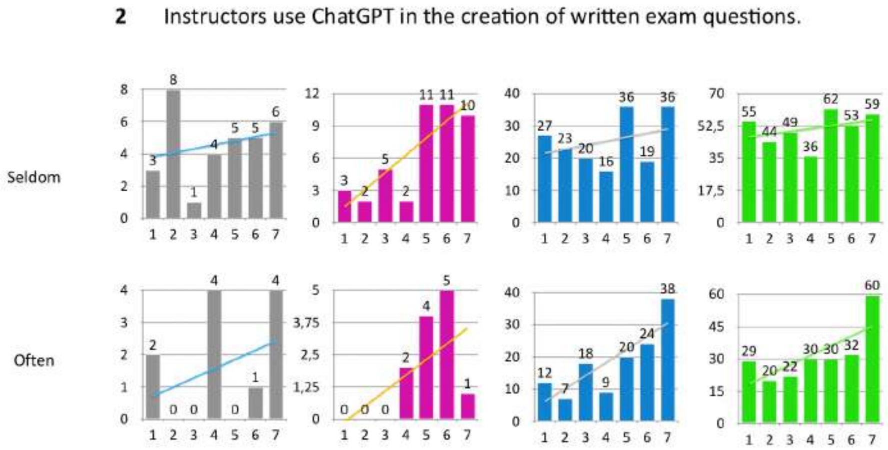


Figure A.3: Answers to Question 2 by role (green = Bachelor's, blue = Master's, pink = Doctoral and grey = employee) and split into seldom use (1 to 3 on the rating scale) and often use (5 to 7)

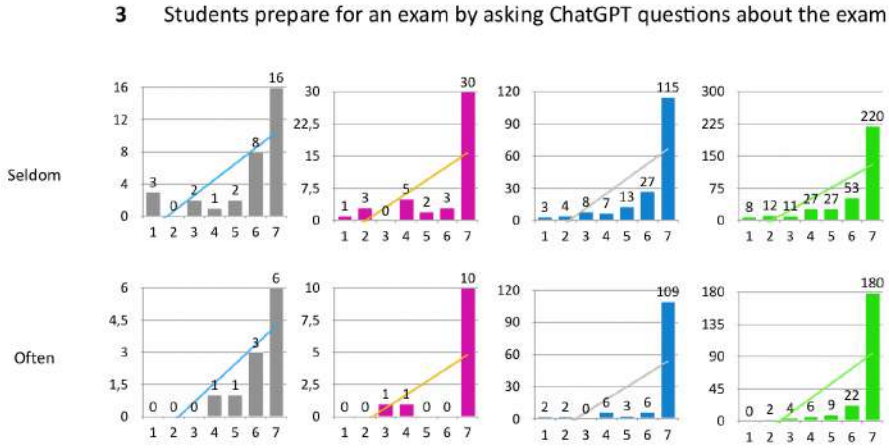


Figure A.4: Answers to Question 3 by role (green = Bachelor's, blue = Master's, pink = Doctoral and grey = employee) and split into seldom use (1 to 3 on the rating scale) and often use (5 to 7)

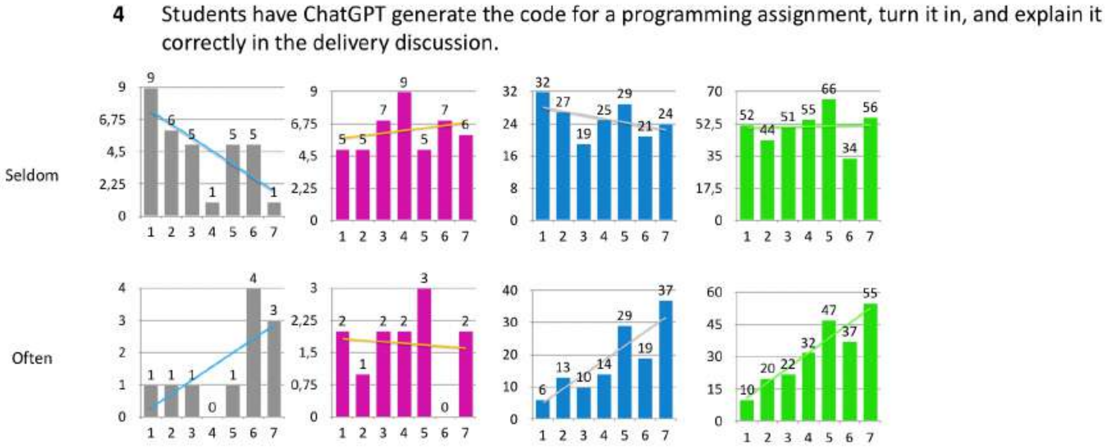


Figure A.5: Answers to Question 4 by role (green = Bachelor's, blue = Master's, pink = Doctoral and grey = employee) and split into seldom use (1 to 3 on the rating scale) and often use (5 to 7)

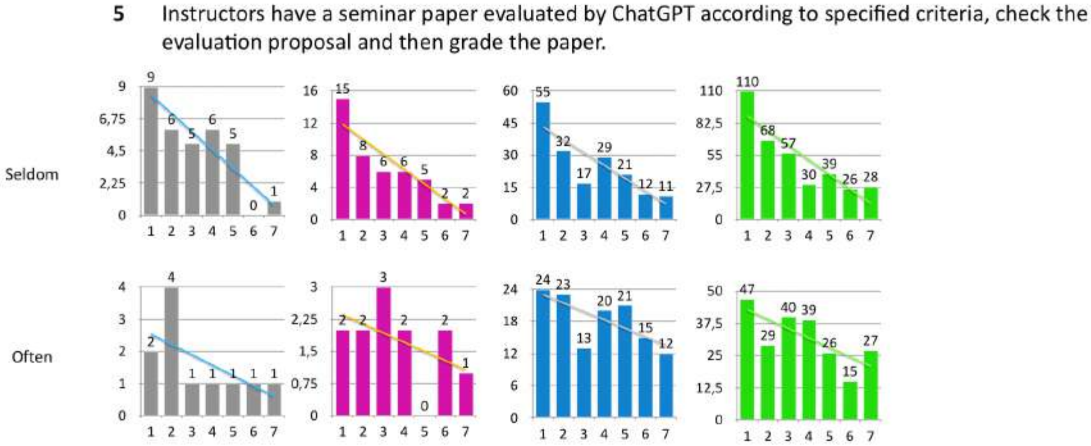


Figure A.6: Answers to Question 5 by role (green = Bachelor's, blue = Master's, pink = Doctoral and grey = employee) and split into seldom use (1 to 3 on the rating scale) and often use (5 to 7)

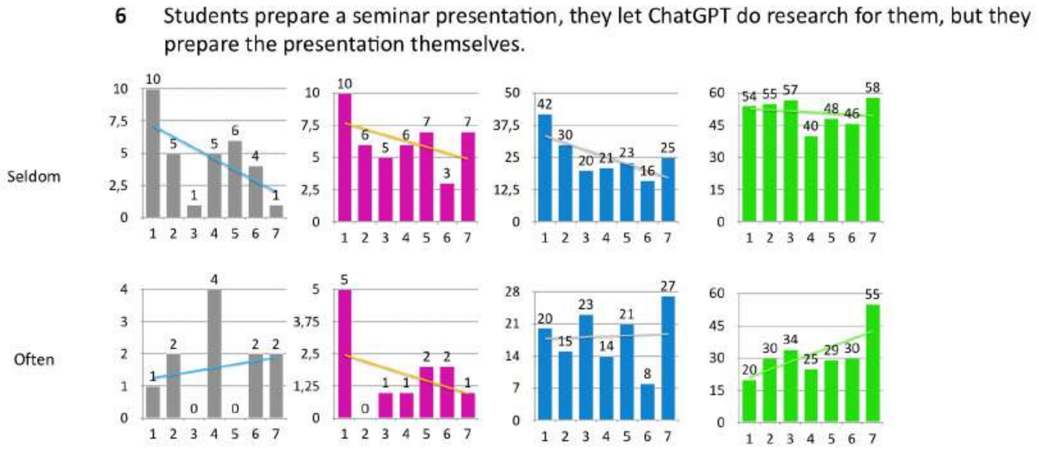


Figure A.7: Answers to Question 6 by role (green = Bachelor's, blue = Master's, pink = Doctoral and grey = employee) and split into seldom use (1 to 3 on the rating scale) and often use (5 to 7)

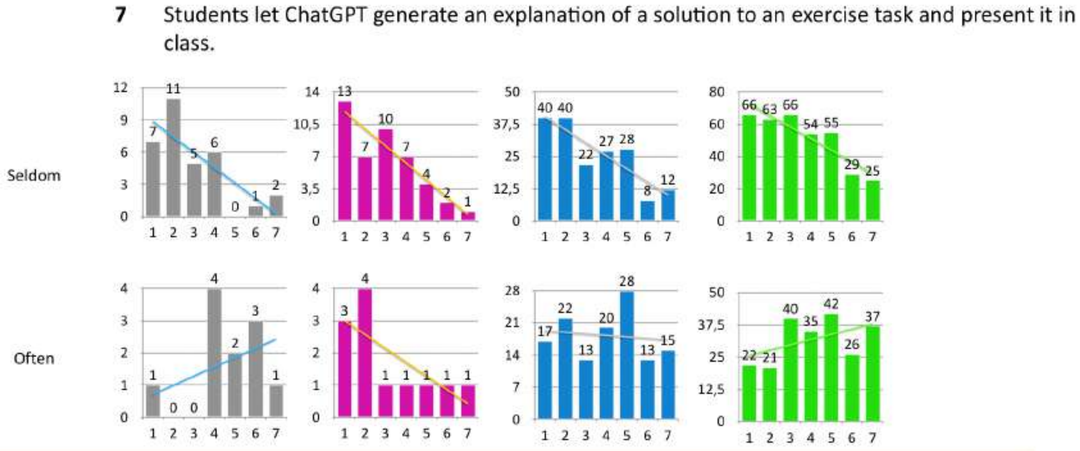


Figure A.8: Answers to Question 7 by role (green = Bachelor's, blue = Master's, pink = Doctoral and grey = employee) and split into seldom use (1 to 3 on the rating scale) and often use (5 to 7)

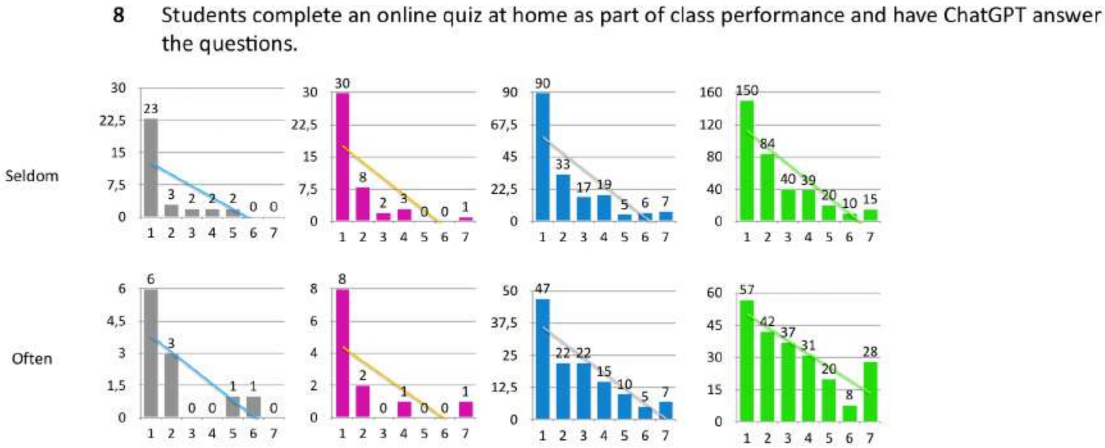

8 Students complete an online quiz at home as part of class performance and have ChatGPT answer the questions.

Figure A.9: Answers to Question 8 by role (green = Bachelor's, blue = Master's, pink = Doctoral and grey = employee) and split into seldom use (1 to 3 on the rating scale) and often use (5 to 7)

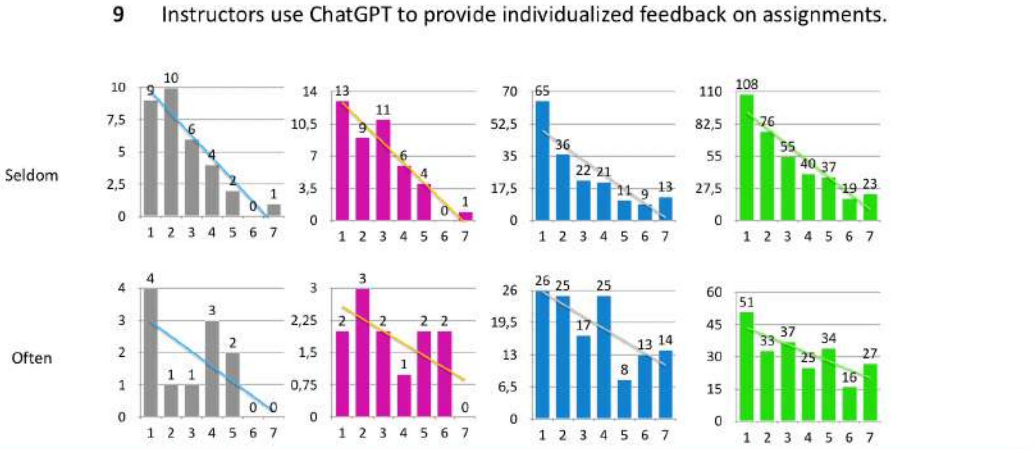

9 Instructors use ChatGPT to provide individualized feedback on assignments.

Figure A.10: Answers to Question 9 by role (green = Bachelor's, blue = Master's, pink = Doctoral and grey = employee) and split into seldom use (1 to 3 on the rating scale) and often use (5 to 7)

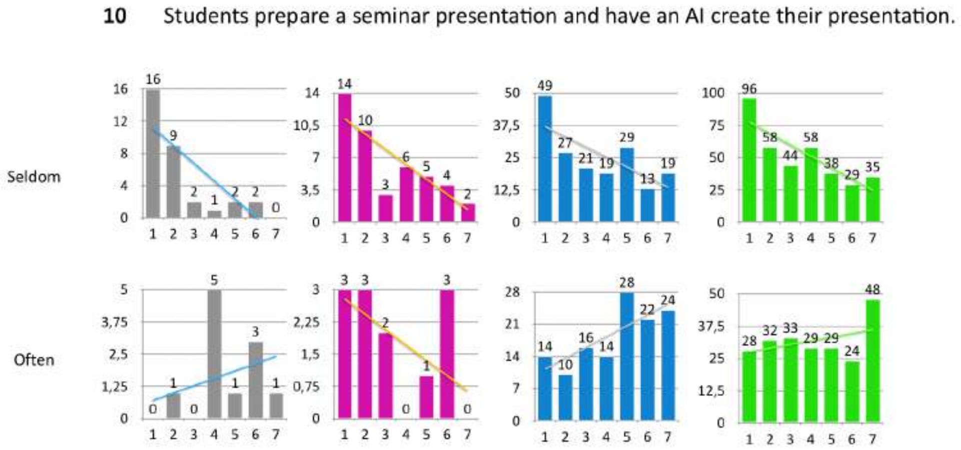


Figure A.11: Answers to Question 10 by role (green = Bachelor's, blue = Master's, pink = Doctoral and grey = employee) and split into seldom use (1 to 3 on the rating scale) and often use (5 to 7)

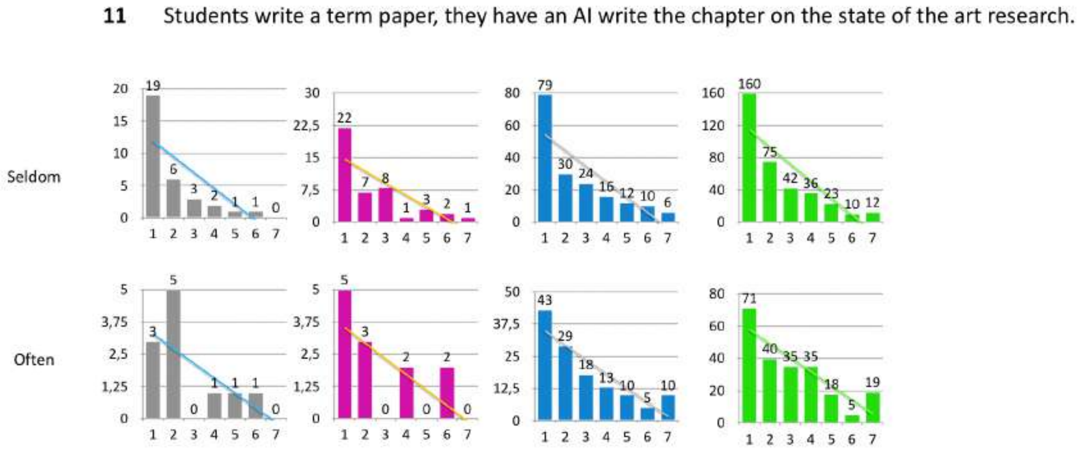


Figure A.12: Answers to Question 11 by role (green = Bachelor's, blue = Master's, pink = Doctoral and grey = employee) and split into seldom use (1 to 3 on the rating scale) and often use (5 to 7)

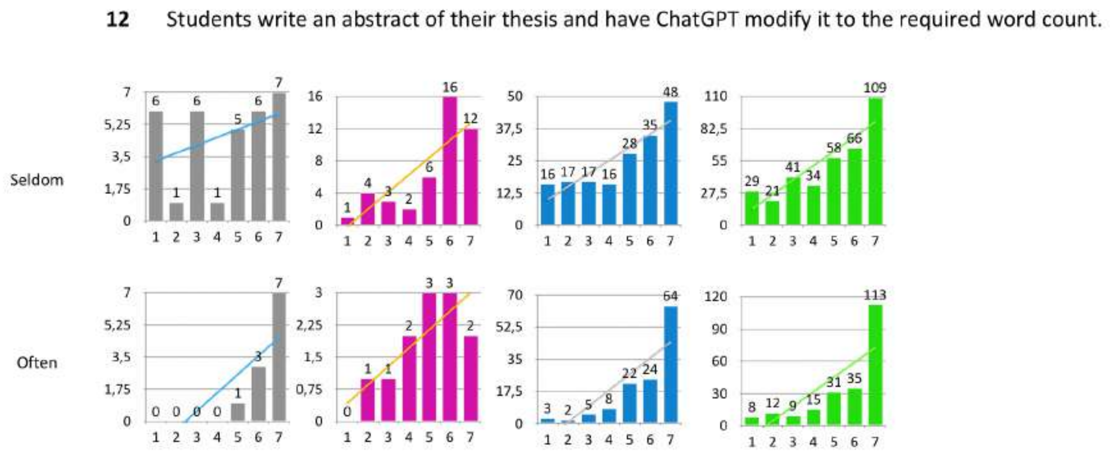


Figure A.13: Answers to Question 12 by role (green = Bachelor's, blue = Master's, pink = Doctoral and grey = employee) and split into seldom use (1 to 3 on the rating scale) and often use (5 to 7)

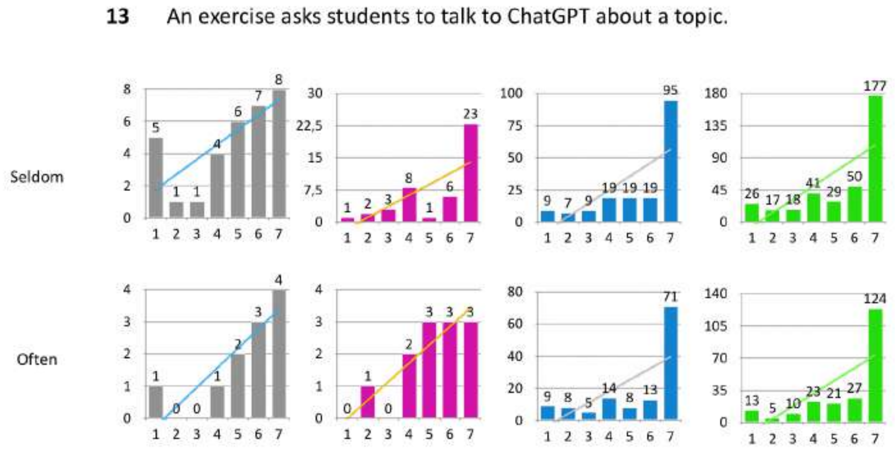


Figure A.14: Answers to Question 13 by role (green = Bachelor's, blue = Master's, pink = Doctoral and grey = employee) and split into seldom use (1 to 3 on the rating scale) and often use (5 to 7)

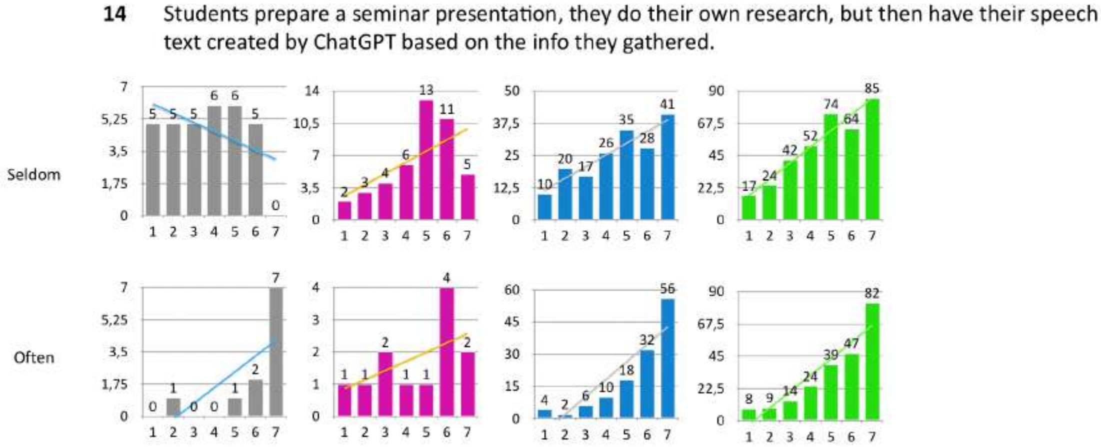


Figure A.15: Answers to Question 14 by role (green = Bachelor's, blue = Master's, pink = Doctoral and grey = employee) and split into seldom use (1 to 3 on the rating scale) and often use (5 to 7)

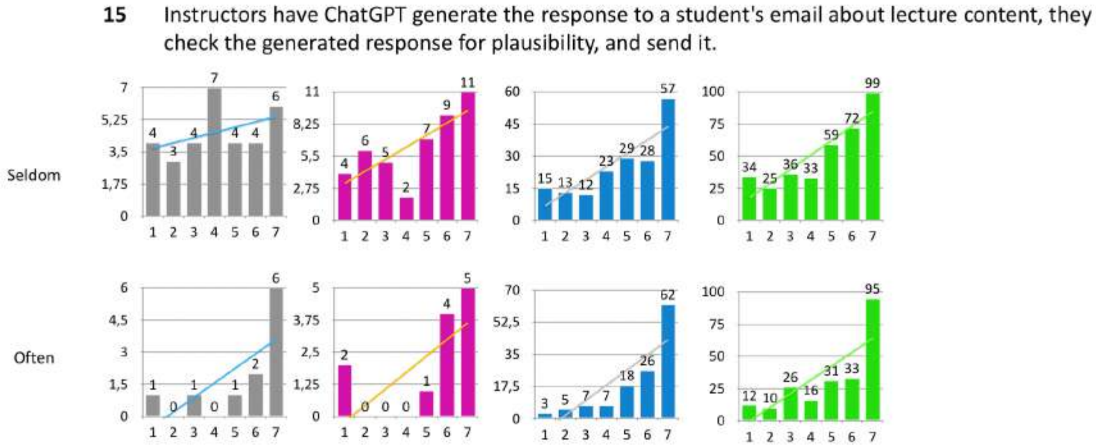
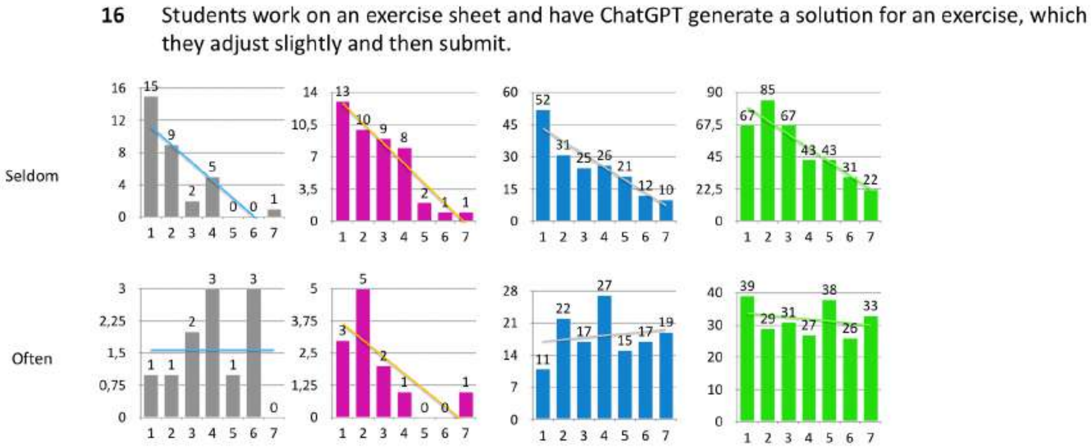


Figure A.16: Answers to Question 15 by role (green = Bachelor's, blue = Master's, pink = Doctoral and grey = employee) and split into seldom use (1 to 3 on the rating scale) and often use (5 to 7)



16 Students work on an exercise sheet and have ChatGPT generate a solution for an exercise, which they adjust slightly and then submit.

Figure A.17: Answers to Question 16 by role (green = Bachelor's, blue = Master's, pink = Doctoral and grey = employee) and split into seldom use (1 to 3 on the rating scale) and often use (5 to 7)

A.4.3 Answer Distribution Per Role

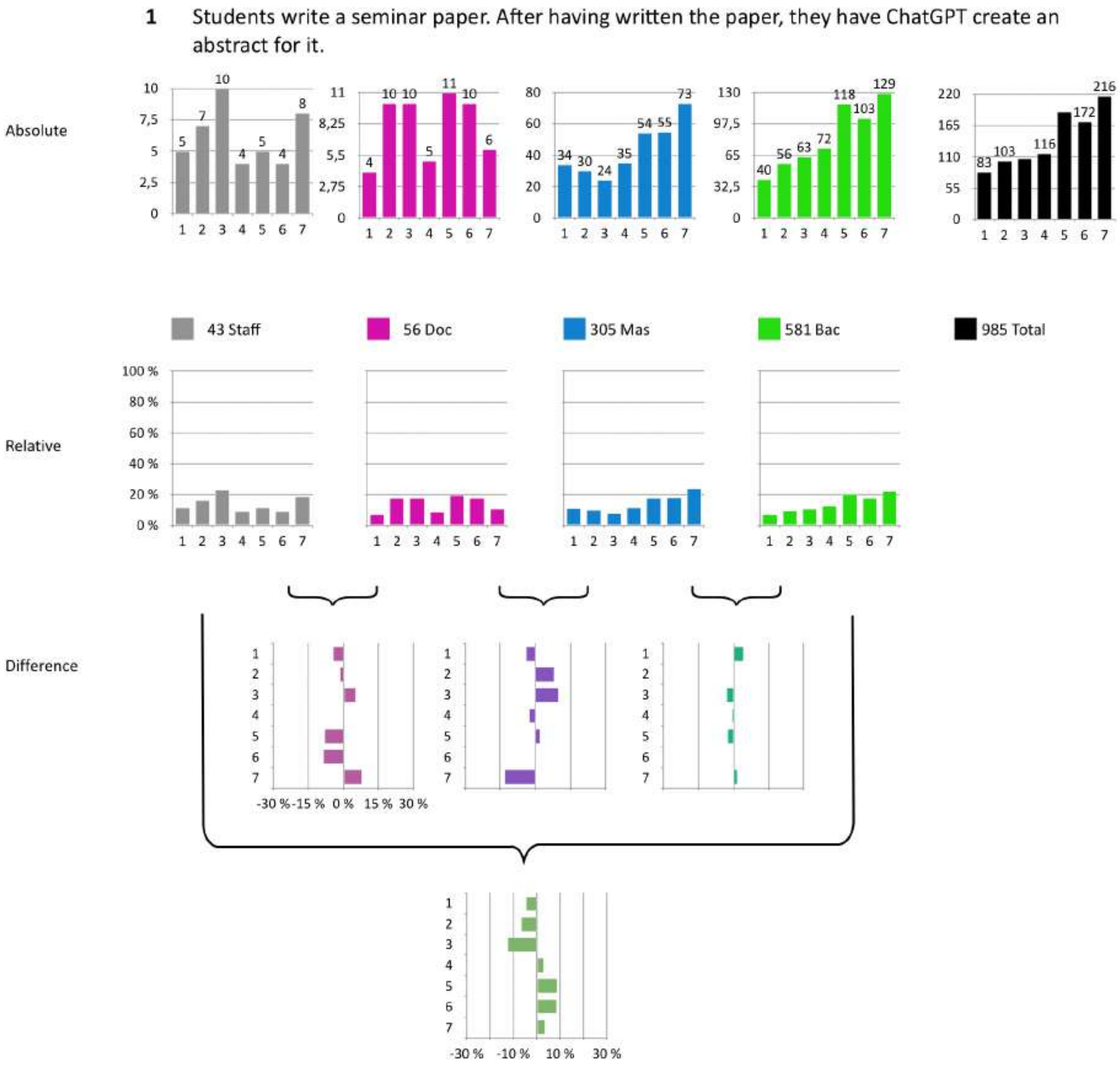


Figure A.18: Comparison of how positive (1 being the most negative, 7 the most positive) the answers of different roles are for Question 1

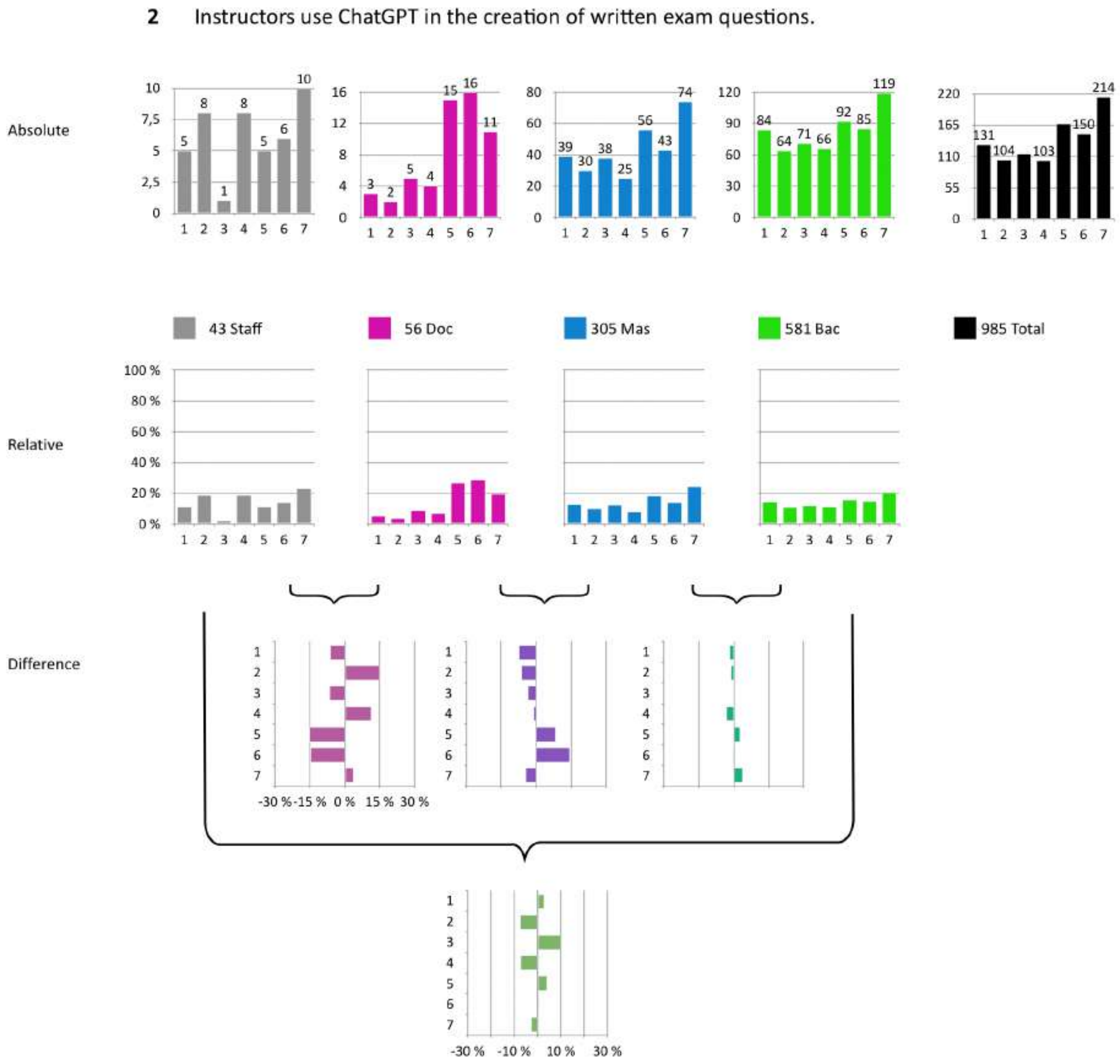


Figure A.19: Comparison of how positive (1 being the most negative, 7 the most positive) the answers of different roles are for Question 2

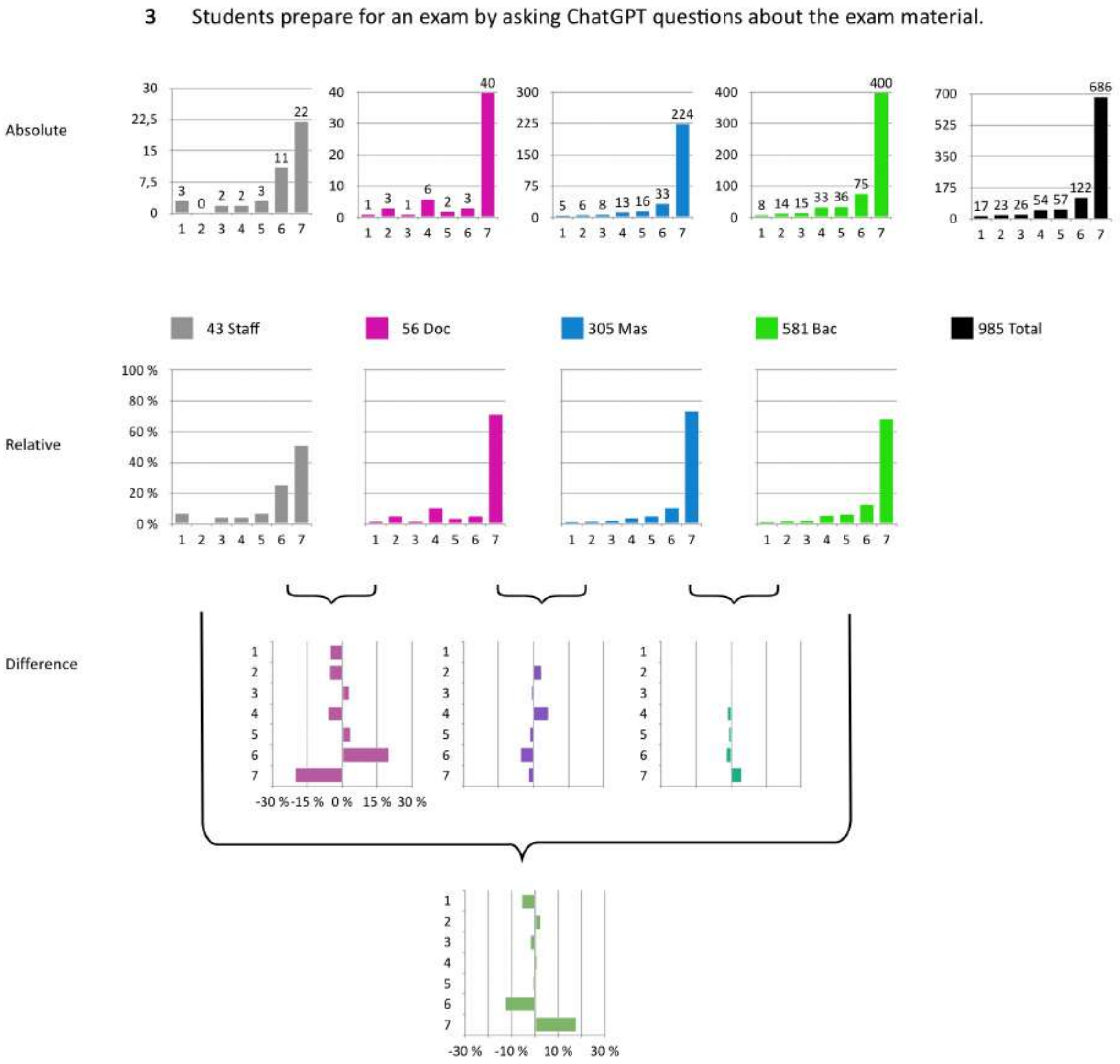


Figure A.20: Comparison of how positive (1 being the most negative, 7 the most positive) the answers of different roles are for Question 3

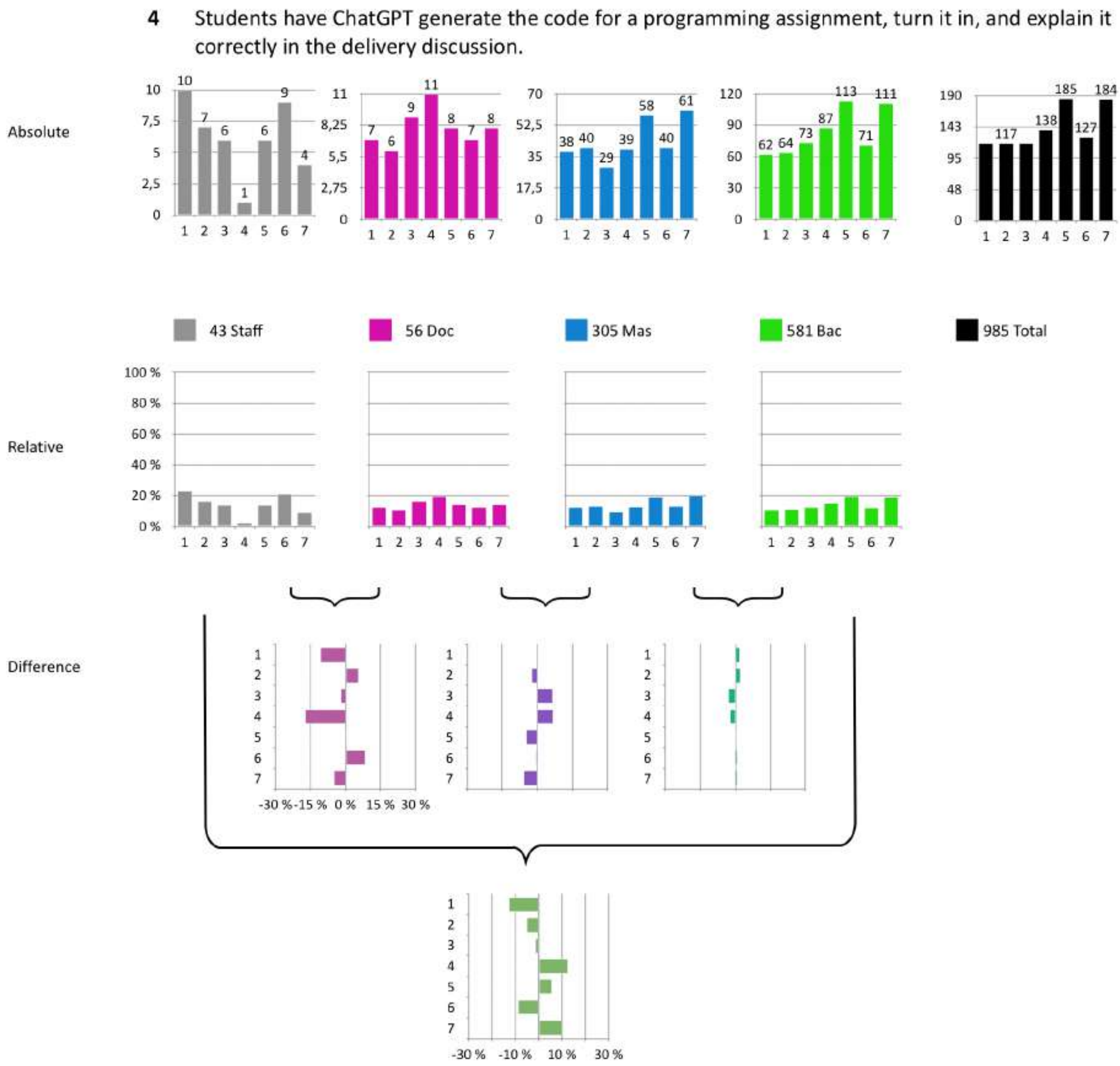


Figure A.21: Comparison of how positive (1 being the most negative, 7 the most positive) the answers of different roles are for Question 4

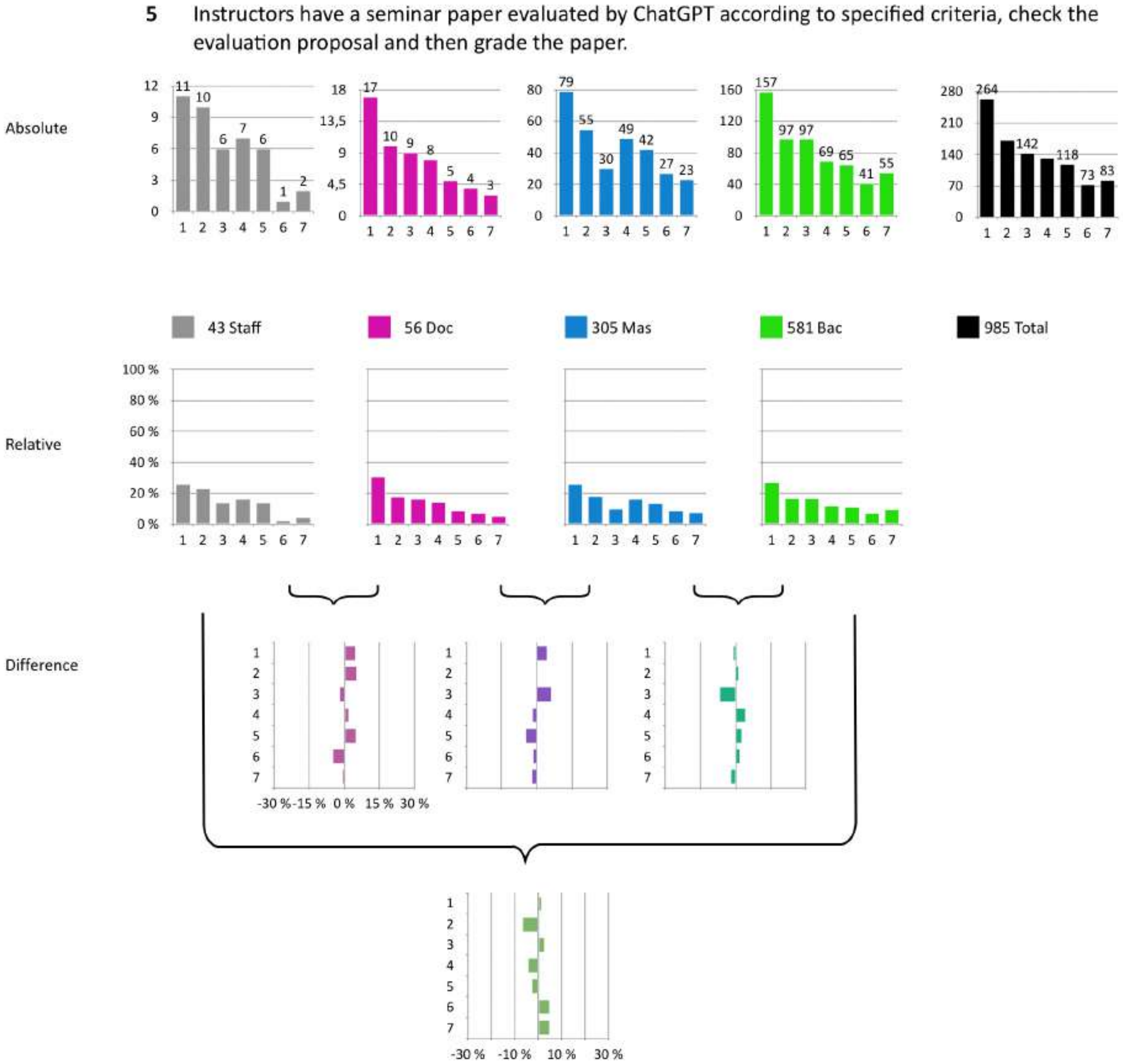


Figure A.22: Comparison of how positive (1 being the most negative, 7 the most positive) the answers of different roles are for Question 5

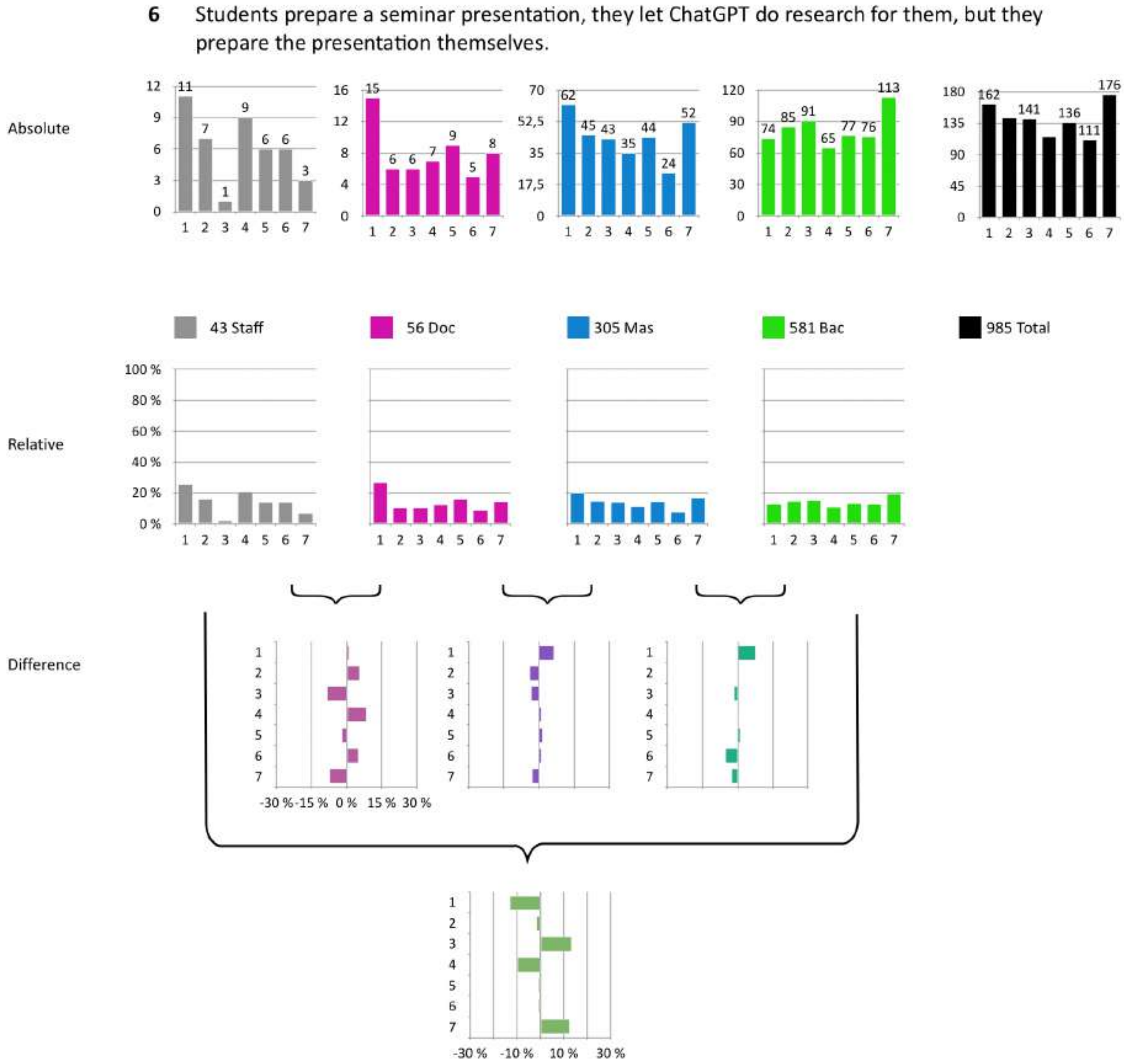


Figure A.23: Comparison of how positive (1 being the most negative, 7 the most positive) the answers of different roles are for Question 6

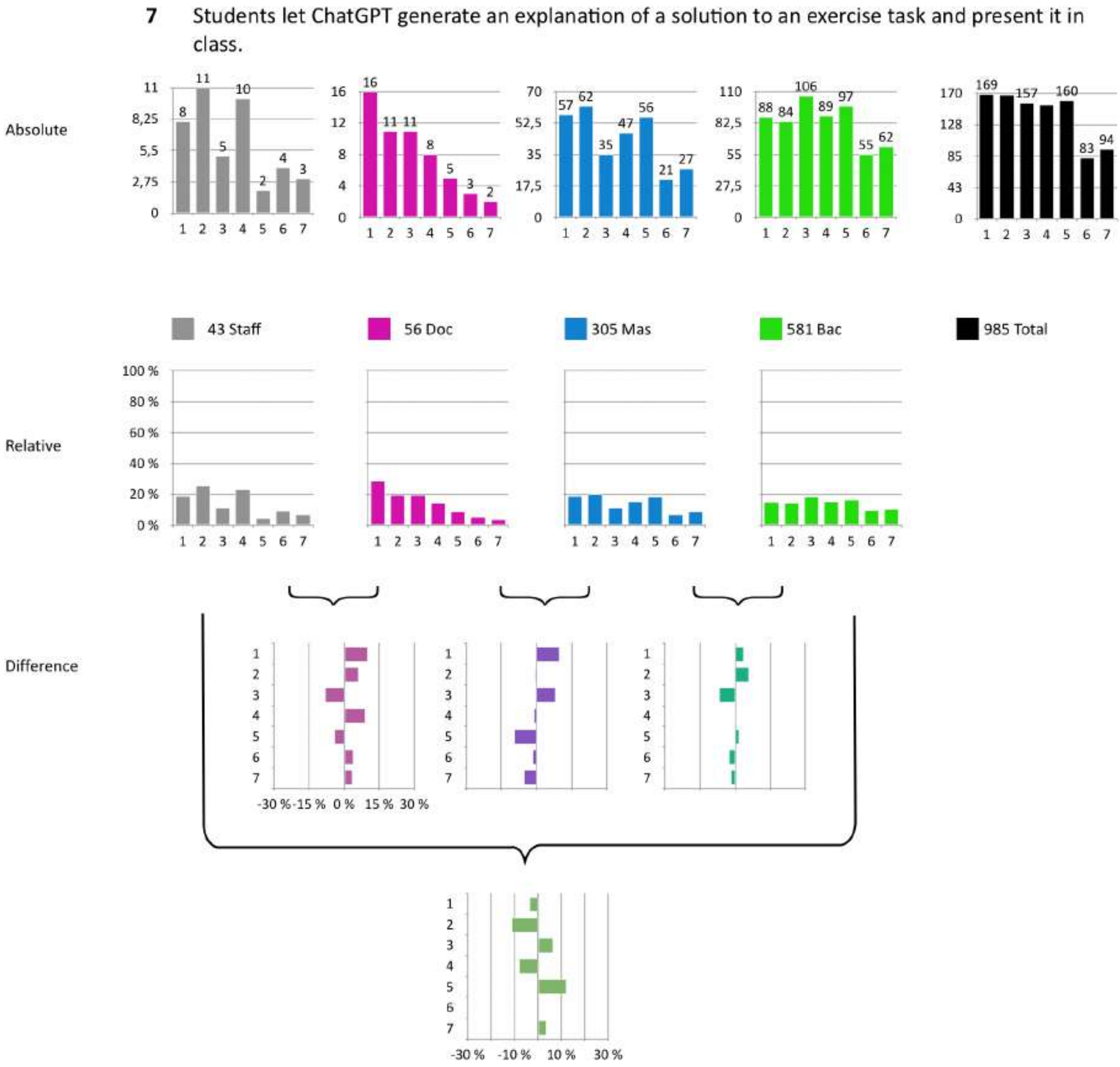


Figure A.24: Comparison of how positive (1 being the most negative, 7 the most positive) the answers of different roles are for Question 7

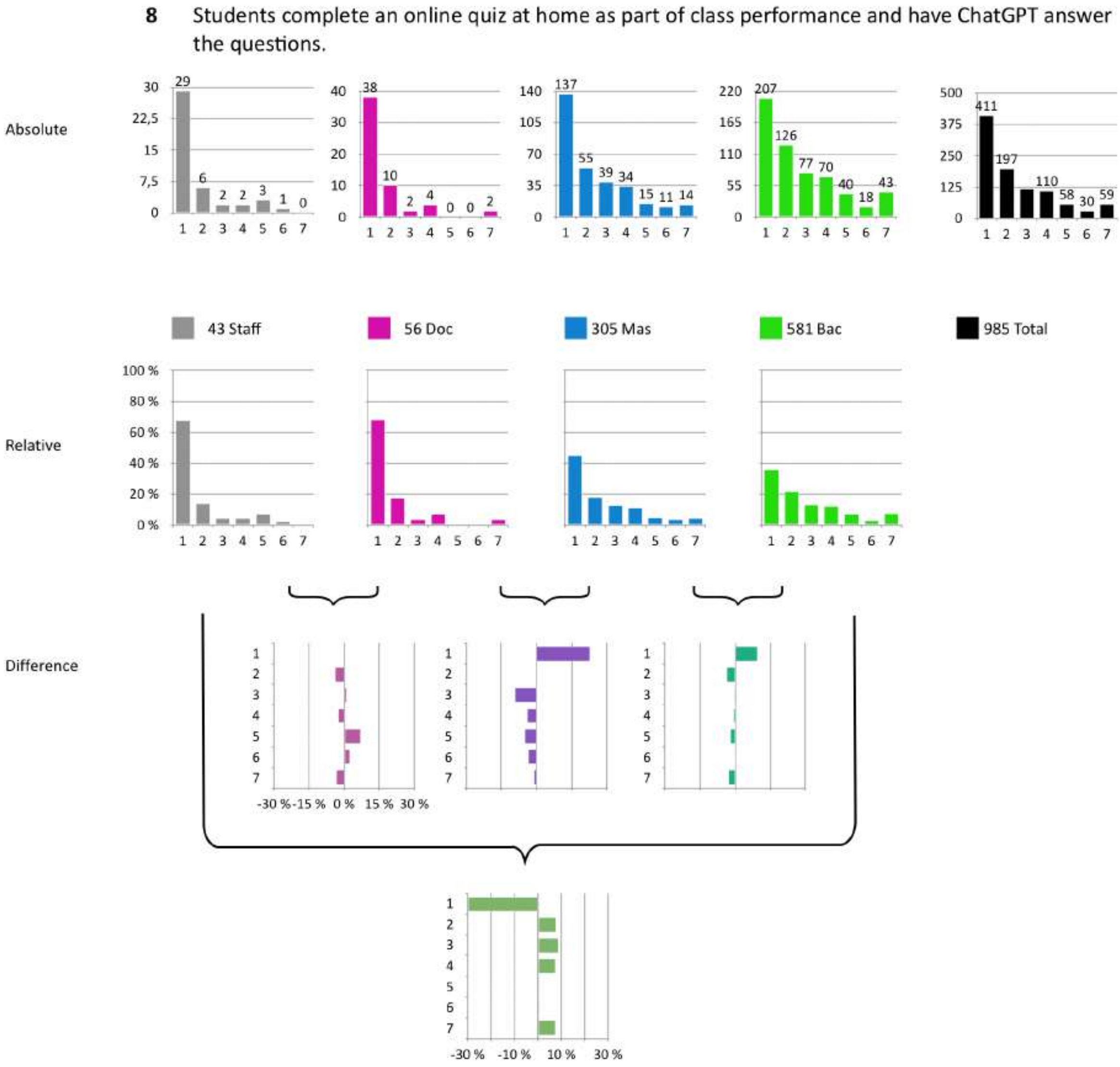


Figure A.25: Comparison of how positive (1 being the most negative, 7 the most positive) the answers of different roles are for Question 8

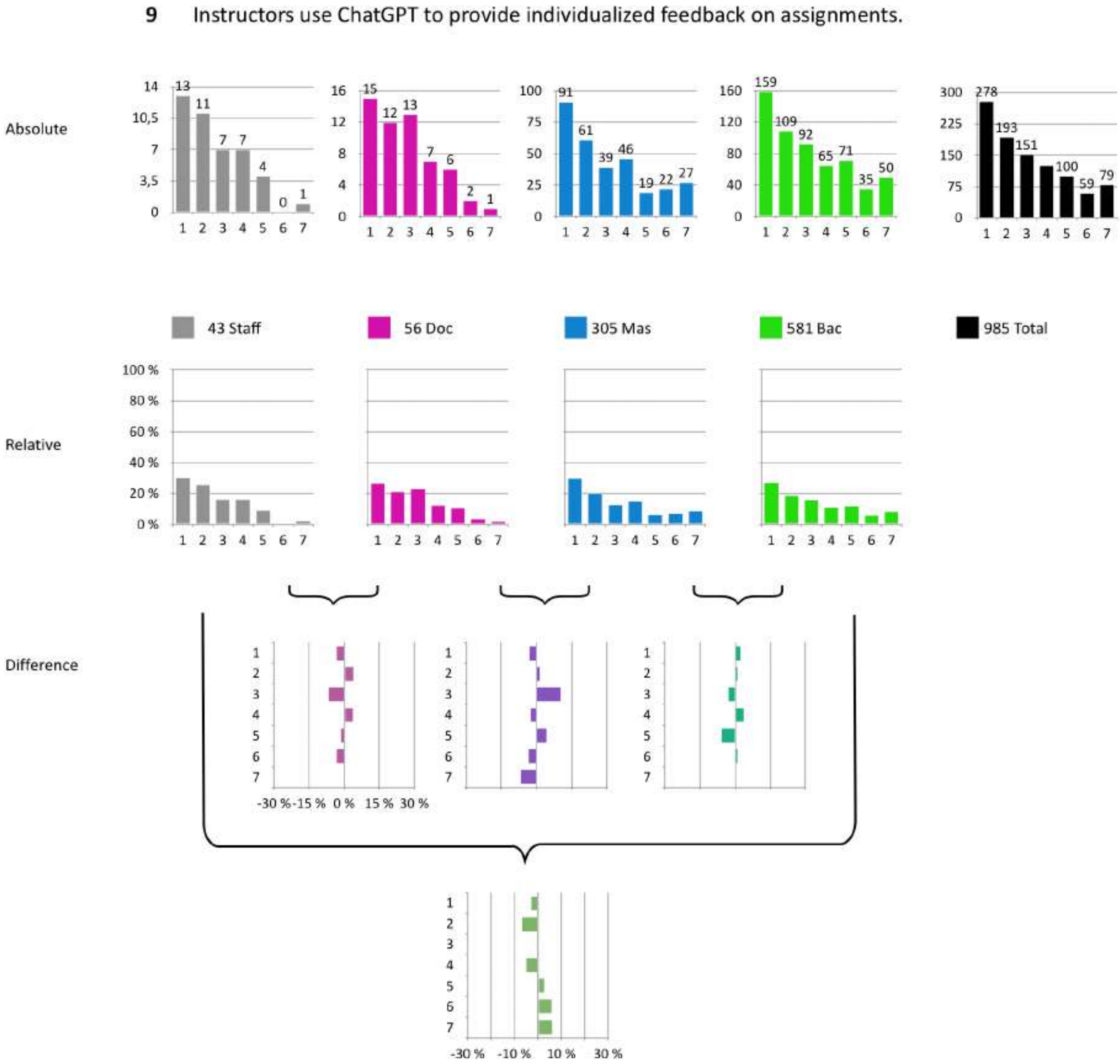


Figure A.26: Comparison of how positive (1 being the most negative, 7 the most positive) the answers of different roles are for Question 9

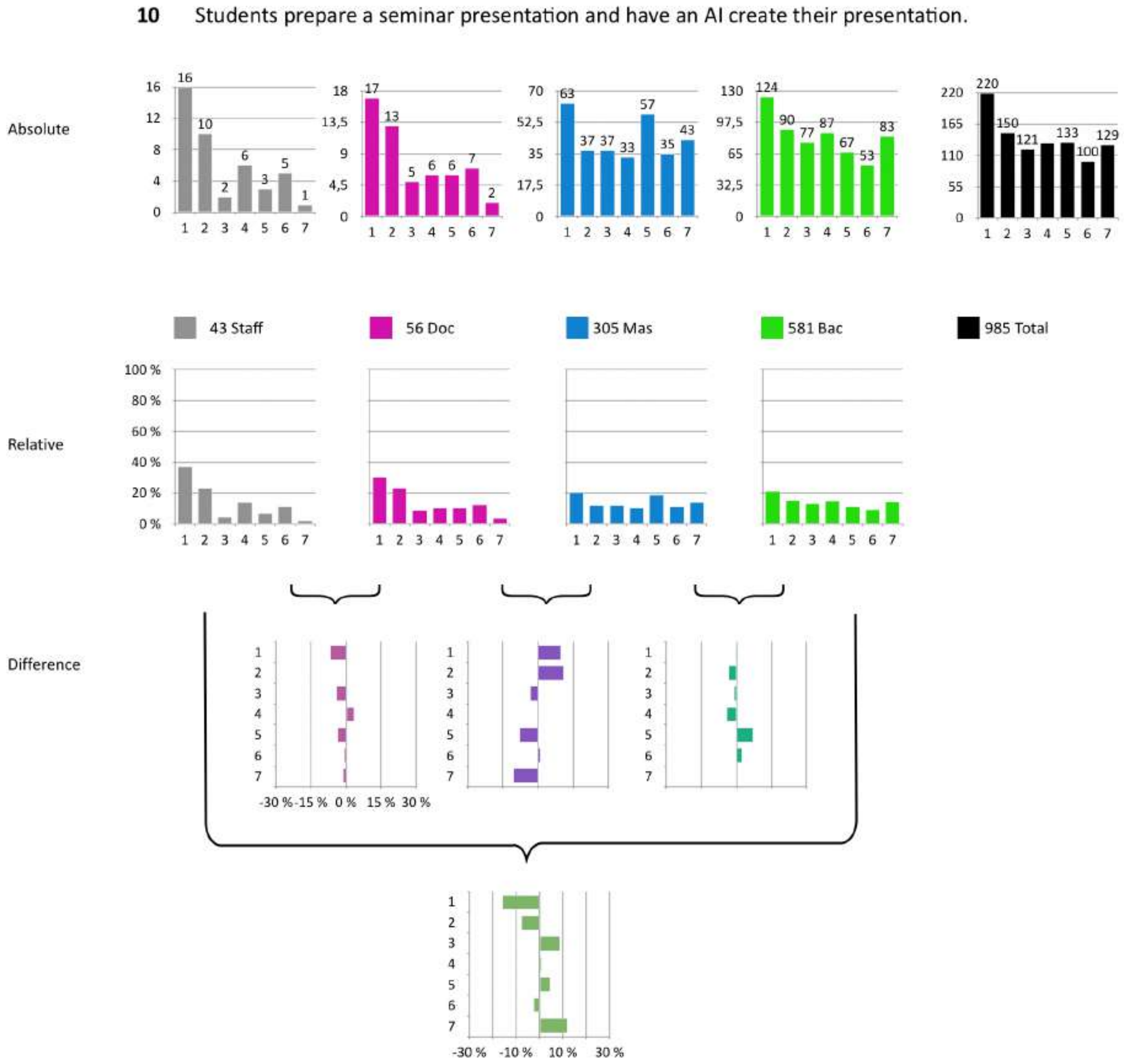


Figure A.27: Comparison of how positive (1 being the most negative, 7 the most positive) the answers of different roles are for Question 10

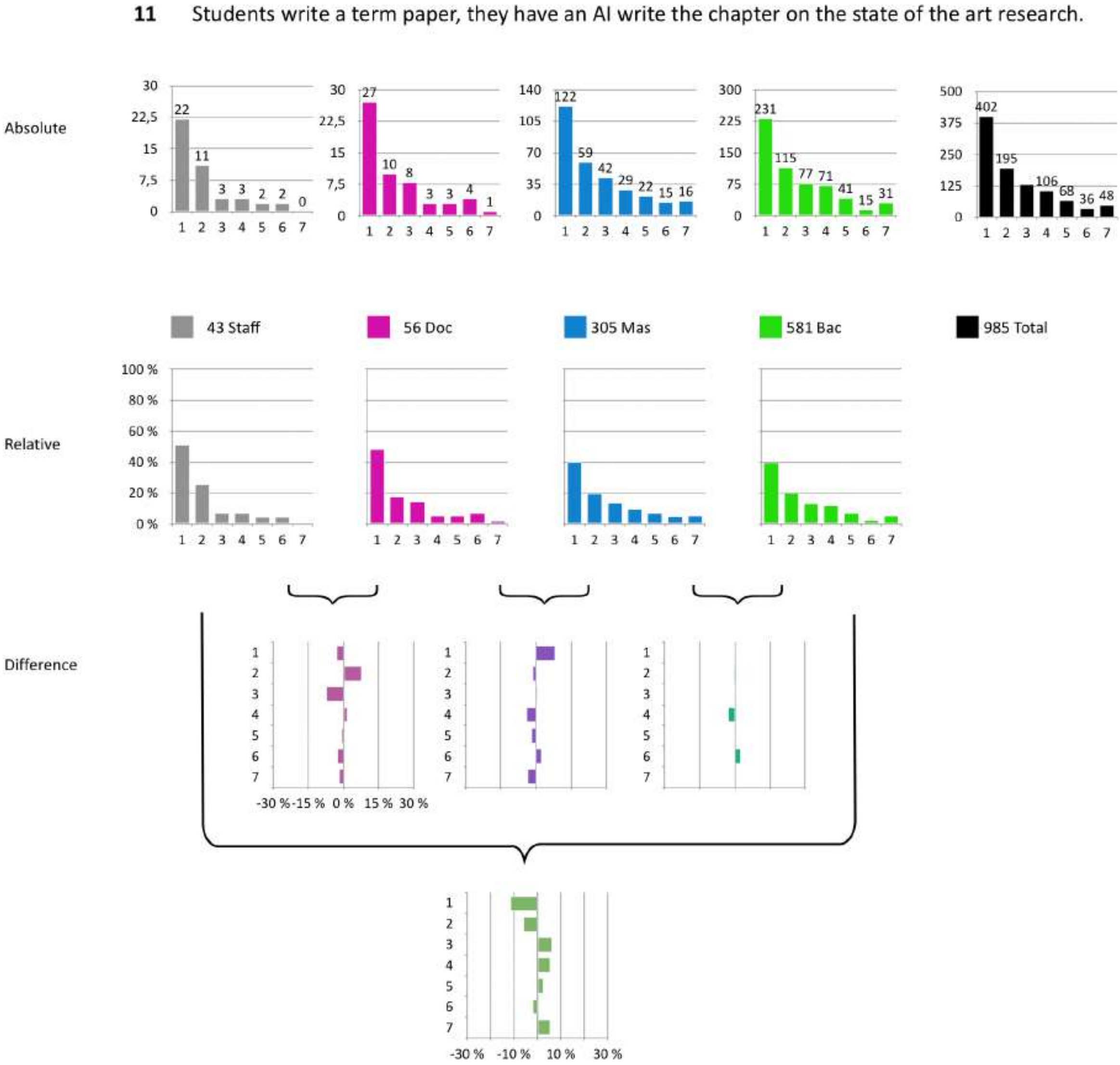


Figure A.28: Comparison of how positive (1 being the most negative, 7 the most positive) the answers of different roles are for Question 11

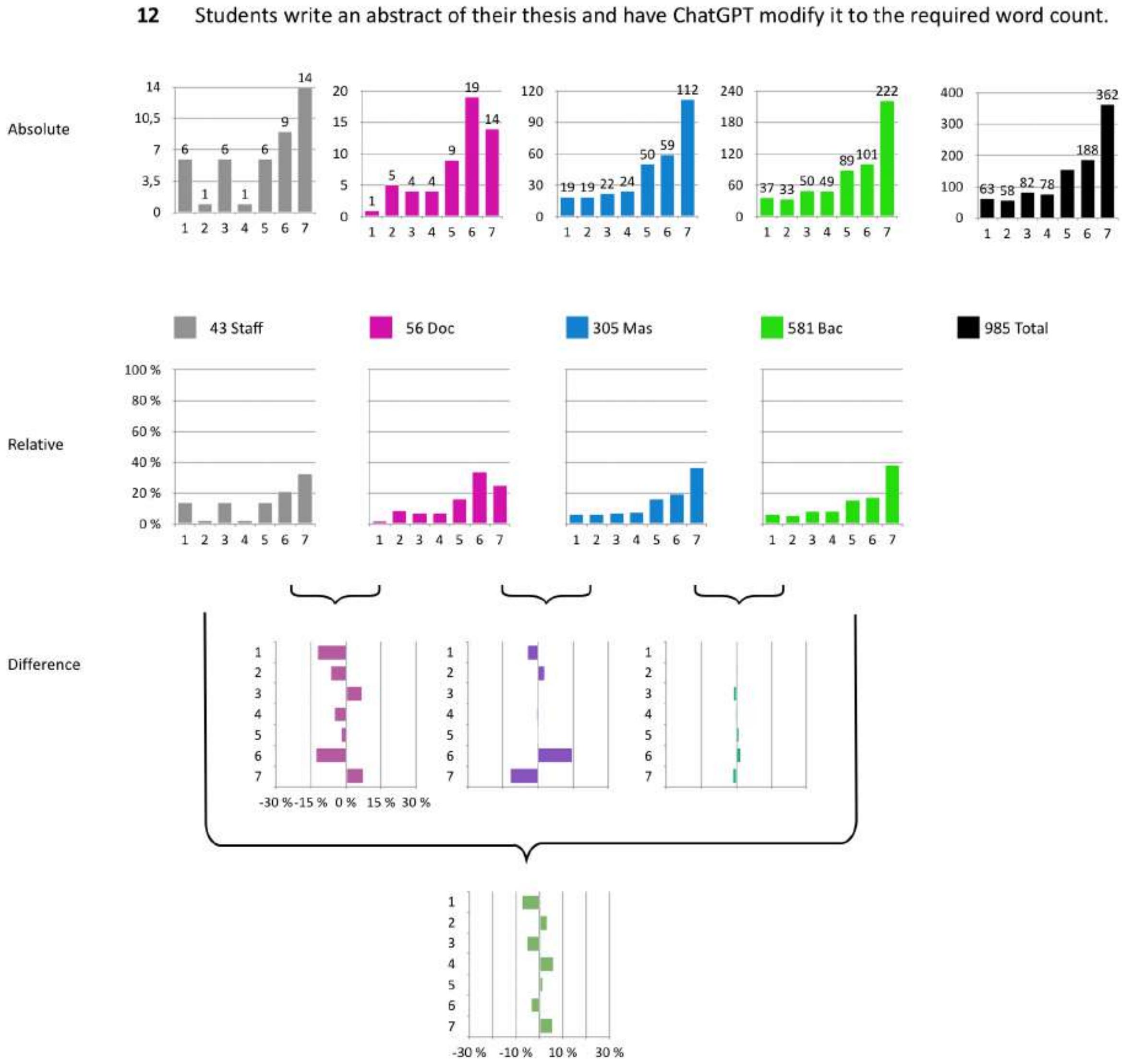


Figure A.29: Comparison of how positive (1 being the most negative, 7 the most positive) the answers of different roles are for Question 12

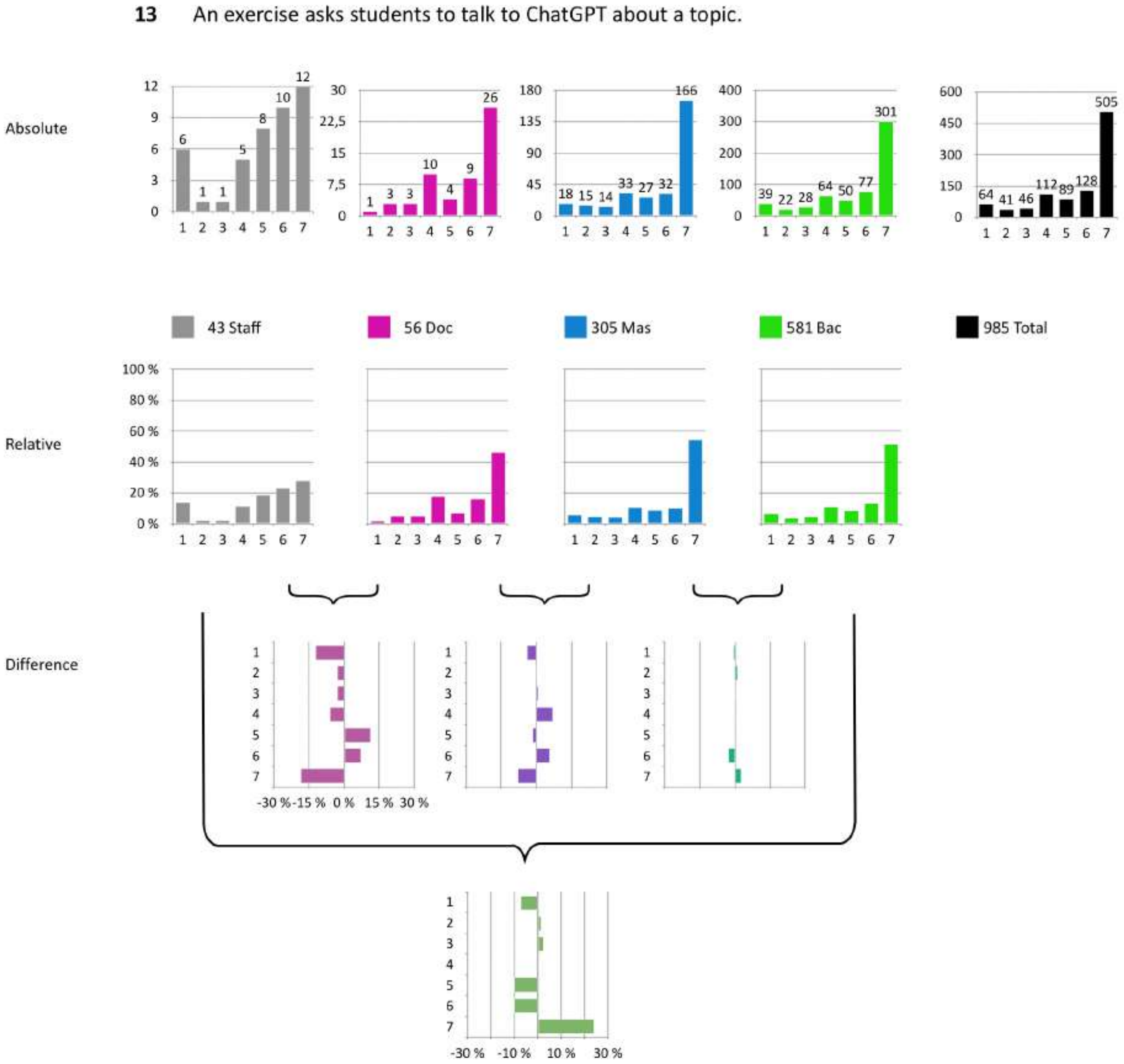


Figure A.30: Comparison of how positive (1 being the most negative, 7 the most positive) the answers of different roles are for Question 13

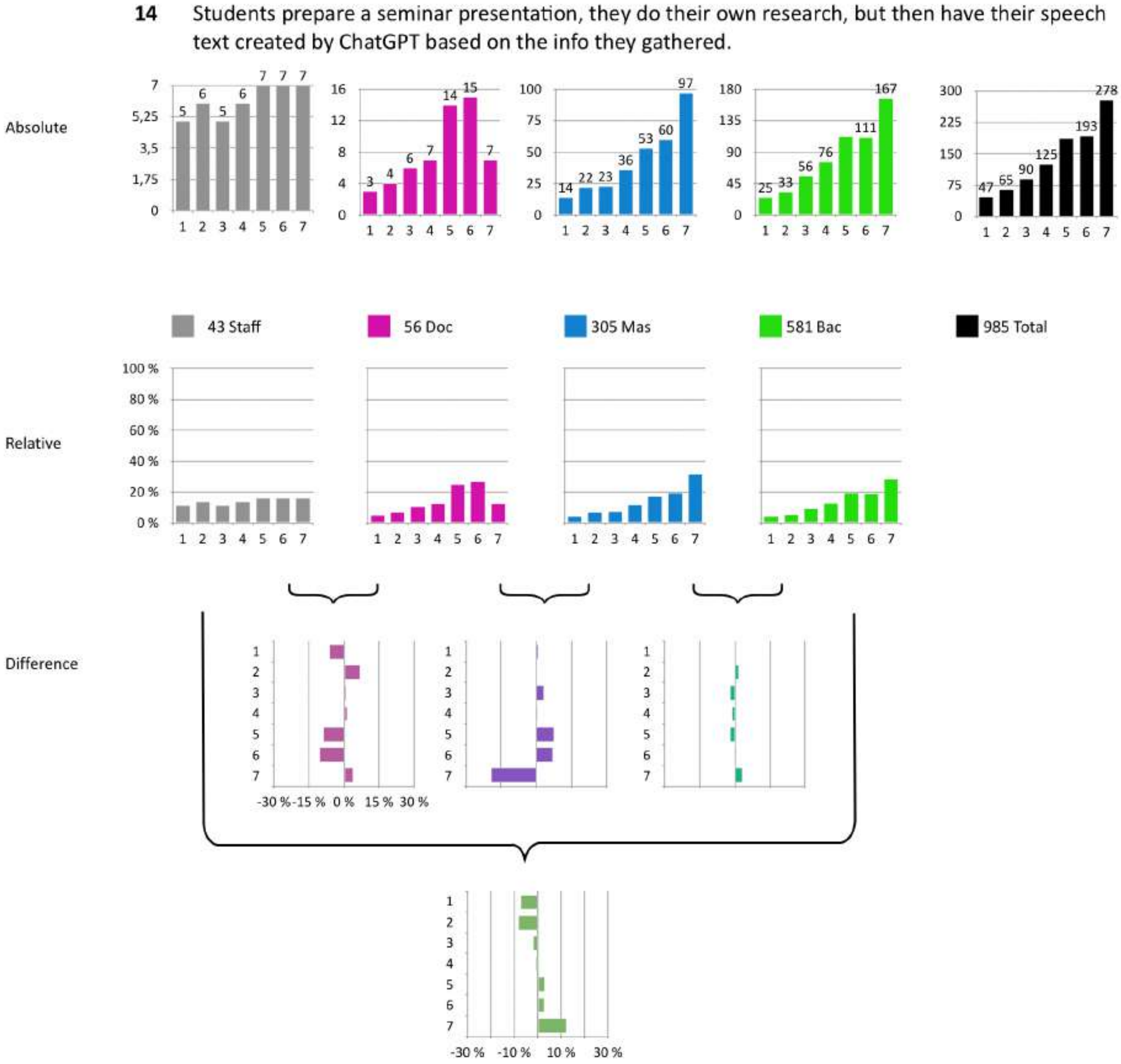
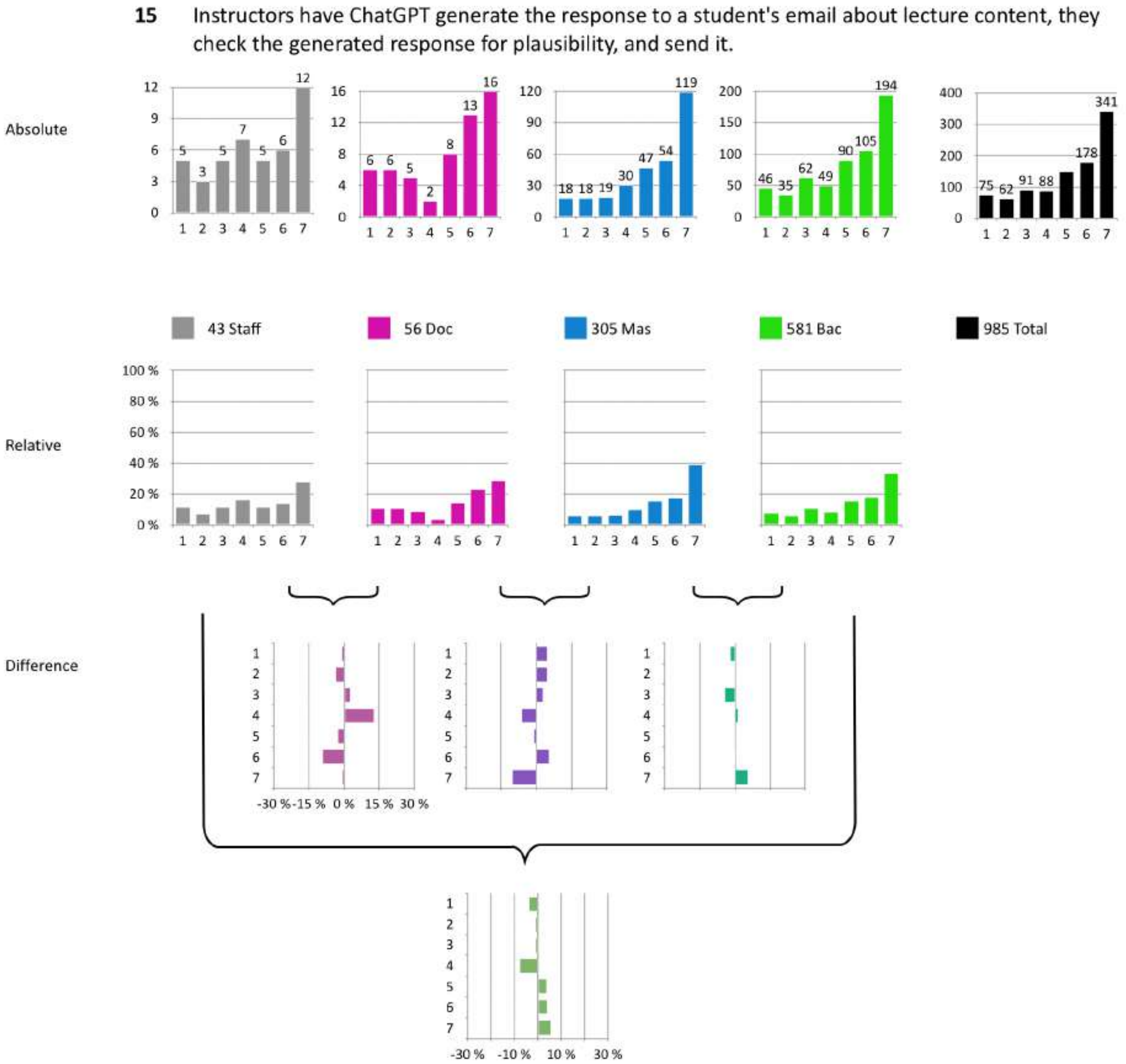
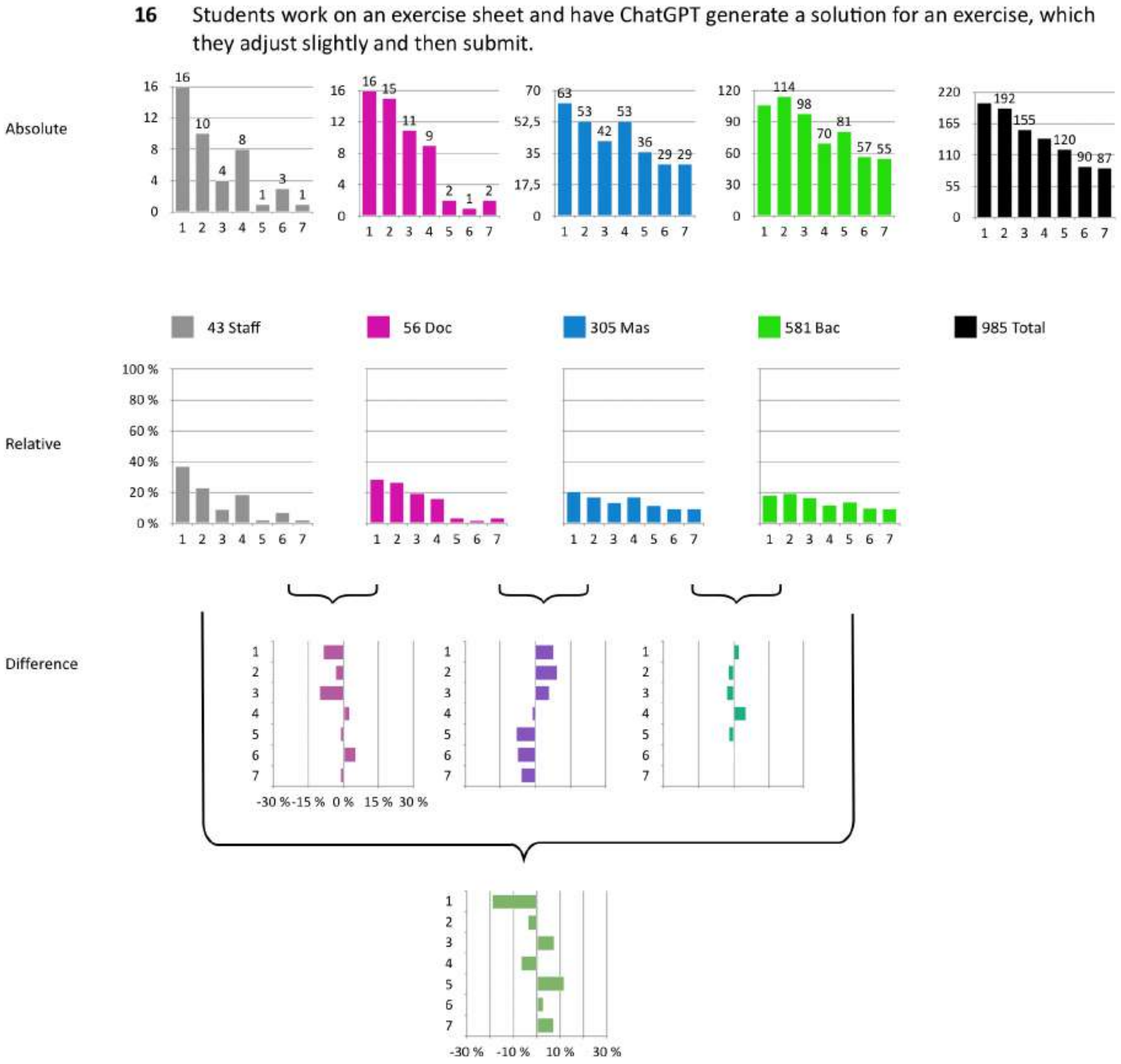


Figure A.31: Comparison of how positive (1 being the most negative, 7 the most positive) the answers of different roles are for Question 14



15 Instructors have ChatGPT generate the response to a student's email about lecture content, they check the generated response for plausibility, and send it.

Figure A.32: Comparison of how positive (1 being the most negative, 7 the most positive) the answers of different roles are for Question 15



16 Students work on an exercise sheet and have ChatGPT generate a solution for an exercise, which they adjust slightly and then submit.

Figure A.33: Comparison of how positive (1 being the most negative, 7 the most positive) the answers of different roles are for Question 16

The first kind of intellectual and artistic personality belongs to the hedgehogs, the second to the foxes ...

— Sir Isaiah Berlin [**berlin_hedgehog_2013**]

References

[1] Accessed 02.01.2024. URL: `https://fs.blog/an-old-argument-against-writing/`.

[2] Jared Rubin. “Printing and Protestants: An Empirical Test of the Role of Printing in the Reformation”. In: *Review of Economics and Statistics* 96 (2014), pp. 270–286.

[3] Marshall McLuhan. *The Gutenberg Galaxy: The Making of Typographic Man.* University of Toronto Press, 1962.

[4] W. Benjamin. *The Work of Art in the Age of Mechanical Reproduction.* CreateSpace Independent Publishing Platform, 2009. URL: `https://books.google.at/books?id=71l_QgAACAAJ`.

[5] BBVA (Firm). *Ch@nge: 19 Key Essays on how Internet is Changing Our Lives.* BBVA, 2013. URL: `https://books.google.at/books?id=0xZfngEACAAJ`.

[6] Andi Mardiana Paduppai et al. “THE ROLE OF MASSIVE OPEN ONLINE COURSES (MOOC) IN TRAINING DURING THE COVID-19 PANDEMIC”. In: 2021.

[7] Patrick Suppes. “The Uses of Computers in Education.” In: *Scientific American* 215 (1966), pp. 206–220. URL: `https://api.semanticscholar.org/CorpusID:122879755`.

[8] Accessed 02.01.2024. URL: `https://platform.openai.com/docs/chatgpt-education`.

[9] Accessed 12.01.2024. URL: `https://www.washingtonpost.com/technology/2023/04/03/chatgpt-khanmigo-tutor-silicon-valley/`.

[10] Accessed 02.01.2024. URL: `https://orf.at/stories/3326575/`.

[11] Accessed 12.01.2024. URL: `https://orf.at/stories/3345277/`.

[12] R. Turnock. *Television and Consumer Culture: Britain and the Transformation of Modernity.* I.B.Tauris, 2007.

[13] Danuta Smołucha. “Internet – the First Source of (dis)Information”. In: 2020.

[14] Ronald J. Deibert. “The Road to Digital Unfreedom: Three Painful Truths About Social Media”. In: *Journal of Democracy* 30 (2019), pp. 25–39.

[15] Accessed 02.01.2024. URL: `https://colab.tuwien.ac.at/display/DC21403/Projektantrag`.

[16] A.C. Doyle. *SHERLOCK HOLMES & DOCTOR WATSON: The Collected Works.* PERGAMONMEDIA., 2015. URL: https://books.google.at/books?id=WnzmwQEACAAJ.

[17] Accessed 25.11.2023. URL: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[18] Mahir Pradana, Hanifah Putri Elisa, and Syarifuddin Syarifuddin. "Discussing ChatGPT in education: A literature review and bibliometric analysis". eng. In: *Cogent education* 10.2 (2023).

[19] Peter A. Cotton Debby R. E. Cotton and J. Reuben Shipway. "Chatting and cheating: Ensuring academic integrity in the era of ChatGPT". In: *Innovations in Education and Teaching International* 0.0 (2023), pp. 1–12. eprint: https://doi.org/10.1080/14703297.2023.2190148. URL: https://doi.org/10.1080/14703297.2023.2190148.

[20] Enkelejda Kasneci et al. "ChatGPT for good? On opportunities and challenges of large language models for education". In: *Learning and Individual Differences* 103 (2023), p. 102274. URL: https://www.sciencedirect.com/science/article/pii/S1041608023000195.

[21] Nassim Dehouche. "Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3)". In: *Ethics in Science and Environmental Politics* 21 (2021), pp. 17–23.

[22] Jiahong Su () and Weipeng Yang (). "Unlocking the Power of ChatGPT: A Framework for Applying Generative AI in Education". In: *ECNU Review of Education* 6.3 (2023), pp. 355–366. eprint: https://doi.org/10.1177/20965311231168423. URL: https://doi.org/10.1177/20965311231168423.

[23] Reza Hadi Mogavi et al. "ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions". In: *Computers in Human Behavior: Artificial Humans* (2023), p. 100027. URL: https://www.sciencedirect.com/science/article/pii/S2949882123000270.

[24] Ibrahim Adeshola and Adeola Praise Adepoju. "The opportunities and challenges of ChatGPT in education". In: *Interactive Learning Environments* 0.0 (2023), pp. 1–14. eprint: https://doi.org/10.1080/10494820.2023.2253858. URL: https://doi.org/10.1080/10494820.2023.2253858.

[25] Tiffany H. Kung et al. "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models". In: *PLOS Digital Health* 2 (2022). URL: https://api.semanticscholar.org/CorpusID:254876189.

[26] Weixin Liang et al. *GPT detectors are biased against non-native English writers.* 2023. arXiv: 2304.02819 [cs.CL].

[27] Emily M. Bender et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? " In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 610–623. URL: https://doi.org/10.1145/3442188.3445922.

[28] Accessed 18.01.2024. URL: `https://moodlegpt.com/`.

[29] Accessed 18.01.2024. URL: `https://moodle.org/`.

[30] Accessed 08.01.2024. URL: `https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence`.

[31] Hilary Arksey and Lisa O'Malley. "Scoping studies: towards a methodological framework". In: *International Journal of Social Research Methodology* 8.1 (2005), pp. 19–32. eprint: `https://doi.org/10.1080/1364557032000119616`. URL: `https://doi.org/10.1080/1364557032000119616`.

[32] Accessed 26.11.2023. URL: `https://www.connectedpapers.com/`.

[33] Ömer Aydın and Enis Karaarslan. "OpenAI ChatGPT generated literature review: Digital twin in healthcare". In: *Available at SSRN 4308687* (2022).

[34] Catherine A Gao et al. "Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers". In: *BioRxiv* (2022), pp. 2022–12.

[35] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up? Sentiment Classification using Machine Learning Techniques". In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002).* Association for Computational Linguistics, 2002, pp. 79–86. URL: `https://aclanthology.org/W02-1011`.

[36] V. Braun and V. Clarke. *Thematic Analysis: A Practical Guide.* SAGE Publications, 2021. URL: `https://books.google.at/books?id=mToqEAAAQBAJ`.

[37] Accessed 02.01.2024. URL: `https://www.amberscript.com/`.

[38] Accessed 02.01.2024. URL: `neuralnetworksanddeeplearning.com`.

[39] Accessed 02.01.2024. URL: `https://www.youtube.com/watch?v=aircAruvnKk`.

[40] Devansh Arpit et al. "A closer look at memorization in deep networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70.* ICML'17. Sydney, NSW, Australia: JMLR.org, 2017, pp. 233–242.

[41] Accessed 02.01.2024. URL: `https://www.youtube.com/watch?v=PuDquo76490`.

[42] Accessed 02.01.2024. URL: `https://huggingface.co/docs/transformers/model_doc/gpt2`.

[43] Accessed 02.01.2024. URL: `https://chat.openai.com/share/6699aa96-d94f-4737-9a6d-71c655331e51`.

[44] Aristoteles. *Analytica priora.* 350 BC. URL: `http://folk.uio.no/amundbjo/grar/analytica-priora.php,%20http://www.hs-augsburg.de/~harsch/graeca/Chronologia/S_ante04/Aristoteles/ari_a100.html`.

[45] Accessed 14.01.2024. URL: `https://www.gartner.com/en/research/methodologies/gartner-hype-cycle`.

[46] John Stuart Mill. *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation.* Cambridge University Press, 2011.

[47] Theo Janssen. “19 Compositionality: Its Historic Context”. In: *The Oxford Handbook of Compositionality.* Oxford University Press, Feb. 2012. URL: https://doi.org/10.1093/oxfordhb/9780199541072.013.0001.

[48] Accessed 14.01.2024. URL: https://www.americanscientist.org/article/first-links-in-the-markov-chain.

[49] Accessed 26.11.2023. URL: https://www.scientificamerican.com/article/a-random-walk-through-the-english-language/.

[50] Claude Elwood Shannon. “A Mathematical Theory of Communication”. In: *The Bell System Technical Journal* 27 (1948), pp. 379–423. URL: http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf.

[51] Accessed 14.01.2024. URL: https://www.cs.cornell.edu/selman/selman-cv-2013.pdf.

[52] Accessed 14.01.2024. URL: https://chat.openai.com/share/ca3ef5ac-6efd-4da3-9219-aadff0ce5a97.

[53] Accessed 26.11.2023. URL: https://www.bbc.co.uk/programmes/m001r7n0.

[54] S. Marks. *Finding Betty Crocker: The Secret Life of America’s First Lady of Food.* Thorndike Biography Series. Thorndike Press, 2005. URL: https://books.google.at/books?id=1tXYrkGkYmcC.

[55] Accessed 18.11.2023. URL: https://www.ilmarefilm.org/.

[56] Accessed 26.11.2023. URL: https://replika.com/.

[57] Accessed 26.11.2023. URL: https://myhusbandthereplika.tumblr.com/.

[58] Rachel Grieve and Doug Mahar. “The emotional manipulation–psychopathy nexus: Relationships with emotional intelligence, alexithymia and ethical position”. In: *Personality and Individual Differences* 48 (2010), pp. 945–950.

[59] Simone Natale. *Deceitful Media: Artificial Intelligence and Social Life after the Turing Test.* May 2021.

[60] Accessed 26.11.2023. URL: https://ai.meta.com/genai/.

[61] T. Pratchett. *Night Watch.* Discworld novel. Doubleday, 2002. URL: https://books.google.at/books?id=fiZbAAAAMAAJ.

[62] Accessed 26.11.2023. URL: https://www.oeaw.ac.at/ita/projekte/der-ams-algorithmus.

[63] Accessed 26.11.2023. URL: https://epicenter.works/content/epicenterworks-veroeffentlicht-details-zum-ams-algorithmus.

[64] Accessed 08.01.2024. URL: https://www.ams.at/regionen/osterreichweit/news/2024/01/kuenstliche-intelligenz-unterstuetzt-bei-arbeitssuche.

[65] Accessed 08.01.2024. URL: https://www.derstandard.at/story/3000000201774/vorurteile-und-zweifelhafte-umsetzung-der-ams-ki-chatbot-trifft-auf-spott-und-hohn.

[66] Accessed 26.11.2023. URL: https://www.mm-packaging.com/en/.

[67] J. Lanier. *You are Not a Gadget: A Manifesto.* Borzoi Book. Alfred A. Knopf, 2010. URL: https://books.google.at/books?id=9i1WgopfVToC.

[68] Accessed 08.01.2024. URL: https://www.derstandard.at/story/3000000191027/europa-darf-am-ai-act-keinesfalls-scheitern.

[69] Accessed 28.01.2024. URL: https://www.derstandard.at/story/3000000204194/eu-plant-weitreichende-ausnahmen-fuer-biometrische-ueberwachung-via-ki?ref=nl.

[70] Fabian Hagmann. *Application of generative AI in introductory programming courses.* 2023.

[71] Vinu Sankar Sadasivan et al. *Can AI-Generated Text be Reliably Detected?* 2023. arXiv: 2303.11156 [cs.CL].

[72] Debora Weber-Wulff et al. *Testing of Detection Tools for AI-Generated Text.* 2023. arXiv: 2306.15666 [cs.CL].

[73] Athanasios Polyportis and Nikolaos Pahos. “Navigating the perils of artificial intelligence: a focused review on ChatGPT and responsible research and innovation”. In: *Palgrave Communications* 11.1 (Dec. 2024), pp. 1–10. URL: https://ideas.repec.org/a/pal/palcom/v11y2024i1d10.1057_s41599-023-02464-6.html.