Bachelor thesis

IT Forensics & Information Security 180hp

A thesis that writes itself

On the threat of AI-generated essays within academia

Digital Forensics 15hp

Halmstad 2022-05-12

August Olsson & Oscar Engelbrektsson

HALMSTAD UNIVERSITY

Abstract

Historically, cheating in universities has been limited to smuggling notes into exams, unauthorized cooperation, plagiarism and using ghost writers. New improvements in natural language processing now allow students to easily generate text, that is both unique and, in many ways, indistinguishable from what a human would create. These texts can then be submitted with little to no risk of getting caught by anti-cheating software.

There are currently a multitude of such text generators online, which vary in ease of use, cost and capabilities. They are capable enough to generate unique text which will evade plagiarism-tools employed by universities. If you combine relatively cheap pricing, ease of use, pressure to perform well in school and low risk of detection. It is not too difficult to imagine that students will use tools like these to cheat.

This thesis mainly focuses on whether humans can differentiate AI-generated essays from human written ones and what countermeasures can be used to hinder its use. By giving teachers at Halmstad University human and AI-generated text; then asking them to guess the source of text presented. The experiment concluded that teachers' ability to differentiate AI-generated text from human written text could not be proven.

This thesis also surveys the currently available detection methods for AI-generated text and determines that they are not sufficient in their current form. Lastly, this thesis showcases alternative examination methods that could be used instead of essay-style examinations.

Keywords: Academic dishonesty, Artificial Intelligence, GPT-3, Natural Language Processing

Acknowledgements

We would like to thank our supervisors Eric Järpe and Jens Lundström. Their valuable insights and encouragement have been instrumental in completing this project. The same goes for Nadia Benamer and Emma Lundin who through opposition continually provided us with feedback throughout the writing process.

Secondly, we want to extend our deepest gratitude to Mohammed Eldefrawy who with enthusiasm contributed greatly by reaching out to all the relevant examiners at the School of Information Technology. As well as giving us suggestions and feedback on our thought process, has helped us through our work. This appreciation is extended to Johan Thunberg and Wagner De Morais who through discussion provided interesting and important questions in how to approach the subject, especially in the discussion chapter.

Thirdly, a big thank you to all teachers and examiners from the School of Information Technology who participated in our experiment. As well as Jens Elmlund, Elin Jönsson and Enoch Chen (陳以諾) who provided us with their human written text that was used in our experiment. This goes for Jocelyn Chen(陳文慧) as well for reading and providing relevant feedback on academic writing.

Lastly, we would like to thank OpenAI for providing us with a research grant and permission to use GPT-3 in our research. This thesis would not have been possible without their support.

August Olsson

Oscar Engelbrektsson

Terms & Abbreviations

AI

Artificial Intelligence (AI) is the ability of a system to interpret external data and learn from that data for use in achieving specific goals. Machine learning is a subset of AI [1].

Cheating

Cheating is defined as any attempt to receive academic credit for work that is not one's own. This includes, but is not limited to, plagiarism, fabrication, forgery and unallowed cooperation on examinations.

GPT-3

GPT-3 stands for Generative Pre-trained Transformer 3 and is a set of autoregressive language models that are able to understand and generate natural language. GPT-3 is one of the world leading language models for Natural Language Processing (NLP) and is developed by OpenAI [2].

Language model

A Language model is a statistical model that can be used to predict the probability of a given sequence of words. An example of a LM is GPT-3. Language models are the foundation of which Natural Language Processing (NLP) stands upon [2].

NLG

Natural language generation (NLG) is the use of AI to generate natural language text. NLG takes input from a computer system such as a text prompt and produces human-like text. NLG requires Natural Language Understanding (NLU) for extracting information from the input for use in the text generation [4].

NLU

Natural Language Understanding (NLU) is the use of AI to interpret natural language, which is the everyday human language that we use to communicate with each other and extract meaning from it as well as the main train of thought in the text. NLU can be thought of as a computer ability for reading comprehension [4].

NLP

Natural Language Processing (NLP) is the use of AI to process and analyze written or spoken text by breaking it down, comprehending its meaning, and determining the appropriate action. It involves parsing, sentence breaking, and stemming. NLP is the result of combining NLU and NLG [4].

OpenAI

OpenAI is a research company founded by SpaceX co-founder and Tesla CEO Elon Musk, Greg Brockman from notable data startup Cloudera, and entrepreneur Rebekah Mercer. The company focuses on researching AI. Their focus is in developing a highly autonomous artificial general intelligence (AGI) that can outperform humans at most forms of work. Their mission is to ensure that AI benefits all of humanity [3].

Table of contents

1. Introduction

The future of AI is here. AI-based language models are now capable to generate text that is near human like. These language models can be used in writing text for any purpose, for example writing newspaper articles, computer code, movie scripts and even academic essays.

This opens up a new avenue for students seeking to cheat on an essay: let an AI write it for you. All you have to do is give it a prompt, set the parameters and press go. What you are left with is a fully unique text, written with perfect grammar in your subject of choice.

Because the text generated by these language models is completely unique; no anti-plagiarism software currently available is able to detect this form of cheating. In this thesis, we aim to answer to what extent teachers are able to differentiate AI-generated texts from human written texts and what countermeasures can hinder cheating using language models.

One of the current world leading language models is OpenAI’s GPT-3 and it is used in this thesis to generate texts, which will be used in the experiment where teachers' ability to differentiate AI-generated from human written essays is tested. To answer the questions of how this cheating can be hindered, this thesis looks into computer based detection methods of AI-generated text as well as which alternative examination methods can be used to complement or substitute essays.

2. Background & Theory

2.1. Turing test & Chinese room argument

The experiment performed in this thesis is a version of the Turing test. The Turing test is a test of a machine's ability to exhibit intelligent behavior. The test is named after Alan Turing[1], who proposed it in 1950. In the test, a human and a machine are placed in separate rooms. In a typical Turing test, a human judge engages in a written conversation with two other participants, one of which is a computer. The computer is disguised as a human by using a text-generating program. The human judge is tasked with determining which of the two participants is the computer.

There are a few problems with the Turing test. The first is that the test is based on the assumption that a human being and a computer are equal in terms of their ability to think. This may not be the case, as computers may be able to process information much more quickly than humans. Secondly, the test relies on the assumption that a human being is able to communicate with a computer, in a way that is identical to the way they would communicate with another human being. This may not be the case, as humans are able to communicate differently in many ways with other humans compared to a computer. Finally, the Turing test is based on the assumption that a computer is not able to disguise its true identity as a human. However, it is possible for a computer to be programmed to pass the Turing test, which would mean that it is not really thinking, but instead is just pretending to be a human [5].

John Searle[2] criticise the Turing test in his Chinese room argument. The argument is a thought experiment that attempts to show that a computer could never be as intelligent as a human being, because it could never have true understanding. The argument imagines a room in which a person is given a rule book that tells them how to respond to Chinese characters that are passed into the room. The person looks up the rule for each character and writes down the corresponding response. A person outside the room who does not know that the person inside is following a rule book, would think that the person inside is conscious and understands Chinese. However, the person inside the room does not actually understand Chinese, they are just following the rules. Therefore, the Turing test cannot be used to determine if a machine is actually thinking. It just proves that a computer is able to follow pre-defined rules. Since the machine does not understand what it is saying, it cannot be considered conscious and thinking [6].

---

[1] Alan Turing (1912 - 1954) was a mathematician, computer scientist and codebreaker. He was best known for his work on the Enigma code during World War II, which helped to shorten the war. Turing was also a key figure in the development of early computers and helped to establish the field of artificial intelligence.

[2] John Searle (1932 -) is an American philosopher and cognitive scientist. He was the Lucille J. and Edgar F. Kaiser Professor Emeritus of Philosophy at the University of California, Berkeley, and is known for his work in the philosophy of mind and philosophy of language.

2.2. GPT-3

Generative Pre-trained Transformer 3 (GPT-3) is a set of models that are able to understand and generate natural language. GPT-3 was developed by OpenAI and the beta version was released in 2020. The largest GPT-3 model contains 175 million parameters where each parameter functions as a variable and with these 175 million parameters GPT-3 is able to create natural language which is in many cases indiscernible from human written text.

GPT-3 was largely created using machine learning to automatically learn from data and improve. It uses generative modeling, which is how it is able to use what it already knows of a language to predict the coming words based on the previous words. GPT-3 is an autoregressive language model which means that it uses previous values to predict the next value.

In training GPT-3, OpenAI primarily used a filtered and curated version of the CommonCrawl[3] dataset as well as the WebText2[4] dataset combined with the English version of Wikipedia. The datasets were unlabeled and are therefore considered unsupervised learning. Since GPT-3 is already trained on data, it is known as a pre-trained language model. The specific deep learning model that GPT-3 uses is called Transformer. GPT-3 excels in writing short and concise texts very well. When tasked to write an overly long text, GPT-3 has a tendency to deviate from the topic at hand [2].

2.3. Cheating within Academia

Data from Universitetskanslersämbetet (UKÄ), the Swedish government agency that evaluates the quality of higher education and is responsible for official statistics about higher education. Shows that there has been a steady increase of disciplinary actions from 2015 to 2021 taken by Swedish universities, with an especially sharp increase between the years 2019 and 2021. According to UKÄ, this may be due to the shift from in-school learning to remote learning because of the Covid-19 pandemic and lack of regulatory guidance regarding which control measures can be used when having distanced exams.

[3] The CommonCrawl dataset is a large dataset of web crawl data that is freely available. The dataset is made up of data from various web crawls and contains a wealth of information on the web. The dataset used by GPT-3 is 570GB.

[4] The WebText2 dataset is a collection of text data that was scraped from the internet. The dataset is around 26GB in size.

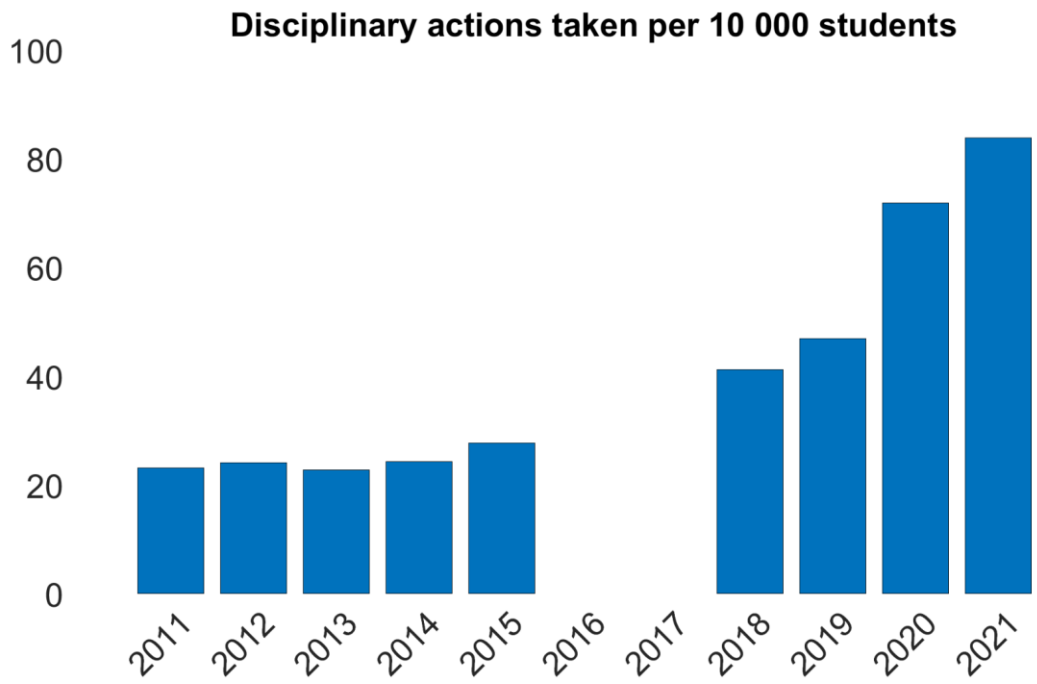


Figure 1. Data published by UKÄ on disciplinary actions taken per 10 000 students[5] [7]

According to UKÄs data, the type of cheating that has increased most between 2020 and 2021 is unauthorized cooperation between students. Traditionally, the cheating methods have been plagiarism and in other ways misleading the examiner, for example by bringing notes to an exam, but since the majority of examinations have been moved to remote examination, there now is a much lower risk of getting caught collaborating. Suggestions have been raised to change the forms of evaluations being used remotely to make it more difficult to cheat, alternatively better proctoring software to be used [8].

2.4. Current anti-cheat/anti-plagiarism software

There are a variety of anti-cheating and anti-plagiarism software available for universities. Some popular anti-cheating and anti-plagiarism software used include Ouriginal, Turnitin, SafeAssign, and Copyscape. An anti-plagiarism software compares student papers against a database of previously published works to identify similarities using a technology called content matching. This can help flag papers that are in part identical or show great similarities in other ways to other sources. But since the system by design only looks for copied or slightly modified text, they lack the functionality to flag anything except for plagiarized content. Therefore, current anti-plagiarism software are powerless against AI-generated essays which are unique [9].

[5] Universitetskanslersämbetet (UKÄ) did not collect nor publish any data on disciplinary actions for 2016-2017.

2.5. Misuse of Language models

In 2019, OpenAI made a blog post regarding the future of language models and their implications. Where they mentioned that they are interested in seeing research done on bias, misuse, detection of these language models. Which they saw that as a requirement for them to feel confident in releasing larger language models. This shows that OpenAI has a large interest in their technology being used for good purposes and cares deeply about the potential social impact it will have [10].

3. Problem

The questions posed in this thesis are:

- To what extent can teachers differentiate AI-generated from human written essays?
- What countermeasures can hinder cheating using language models?

3.1. Positioning

AI-generated text used for cheating in higher education is a new subject and therefore lacks proper research done on it. OpenAI did a study in 2020 on short articles with ~200 words in which ~52% of the human subjects could determine that an AI had produced the text. Mills. Et al. discusses AI as a tool for cheating and the challenge this poses for academic integrity in a conference paper from 2019. This thesis's goal is that its result will contribute to the scientific discussion around this subject [2] [11].

AI as a tool to cheat is a problem which is likley to grow the more widespread public knowledge becomes about these tools, how affordable and user-friendly they are to use. When this eventually happens the authors believe that more research will be done on the subject.

3.2. Demarcation

This study is limited to testing the ability of teachers of the Academy of Information Technology at Halmstad University's skills in distinguishing AI-generated text from human-generated text. It was chosen to focus specifically on GPT-3 in this thesis as it is the currently most proficient language model in generating human-like text.

Due to the time-constraint of this thesis and the limited scope of the experiment, it will be limited in pushing the boundaries of the subject forward. However, it will hopefully lay groundwork for something to be built upon, as well as act as a springboard to lead to future research.

3.3. Problematization

The use of AI-generated text is a relatively new area, yet there is a lot of scientific research done on the subject. The specific use of AI-generated text as a tool for cheating in academia, however, has very little research done on it. In the research done beforehand the authors have not been able to find any Swedish literature regarding this specific area. It was therefore chosen to broaden the search by using international publications. This could be problematic due to the fact that they are written by authors having different academic backgrounds and cultures which could entail that their reflections and insights are not necessarily applicable to the Swedish educational system.

The experiment was conducted on the teachers from the School of Information Technology of Halmstad University. The validity of the experiment could be challenged if only a small sample size is used. Since the teachers at the School of Information Technology have different backgrounds, ages and cultures the authors still believe that the experiment is sufficient to get an factual answer to the question posed in this thesis.

Lastly there is a problem with the second question regarding which countermeasures can be used to stop this type of cheating. As the question could be considered irrelevant, if the answer to the first question is that AI-generated essays do not pose a threat to academia.

4. Method

4.1. Planning stage

While planning this thesis, it was identified that in order for the experiment which would hopefully answer the first question posed in this thesis to be feasible, input was needed from individuals more knowledgeable in the field of AI, as well as in the academic field. Therefore, multiple professors belonging to the School of Information Technology at Halmstad University were contacted. After discussing the subject with them and receiving positive feedback, the experiment began to be constructed. The non-profit organization OpenAI, which specializes in AI, was approached for a research grant in order to access their model for natural language processing in the early planning stages of this thesis. OpenAI provided access to GPT-3 together with a grant of 18 dollars in credits to use in generating the text that was needed in the research.

To answer the second question posed in this thesis, relevant litterature on computer based detections methods was identified and later a evaluation was made on currently available computer based detections methods. Alternative examination methods were also examined to answer the second question posed in this thesis.

4.2. Problem

The most optimal way to test the hypothesis would be to test subjects who were unaware that they were being tested. However, this would create two problems, one problem would be the ethical problem, since experimenting on people without their knowledge is highly unethical and against both good research practice and Swedish law which required informed and explicit consent. There were also problems in how such an experiment would be carried out practically. Discussions of registering for courses to test the hypothesis were had but was ultimately decided against for the aforementioned concerns. Because of this, a design of the experiment where the subjects would be both aware and willing to participate was chosen. Therefore, the experiment was redesigned to where the subject is being presented with multiple different texts where they have to read and then make a decision if it is a text generated by AI or written by a human. By doing it this way, the subjects were able to give their consent and the data necessary to test the hypothesis was able to be collected. [12]

The chosen test subjects were people within academia, who are in a position where they would need to decide if an essay should be reported for suspected cheating. Henceforth these people will be referred to as teachers for simplicity's sake.

The necessary subjects to partake in the study were decided to be only those from the School of Information Technology. This is because the texts generated and used in the experiment are about Big Data. If these questions were asked to a person with little or no background in IT, they would not be able to make any distinctions between the answers provided in the texts. If the experiment was designed to answer questions that are well known to most people it would instead have the problem that it would not be testing an examiner in a close to live scenario. Specified questions asked in a specific environment are needed to as close as possible recreate the real life scenario where this type of cheating might happen.

It could be argued that only including test subjects from one university does not provide a wide enough sample group. A small sample group like this might not be a representative group for the broader academia. It would arguably be better to test examiners from all universities in all different countries to get a much more accurate test result. But due to the not too insignificant amount of work this would require, it is not feasible within the limited timeframe of this thesis. It is a very interesting proposition to test and compare country to country. This will be left for others who want to do research on the subject.

There is an argument to be made that there is a conflict of interest due to the grant that was provided, however the authors are not bound by any loyalty or debt to OpenAI that could interfere with the research.

4.3. AI text generation

In the experiment the specific GPT-3 model Davinci was chosen for generating the texts as it is the most capable model in the GPT-3 family and is superior to the others in terms of quality of output, even when small amounts of information is given. According to OpenAI, Davinci is particularly good at understanding the intent of text as well as solving logic problems and is even “able to solve some of the most challenging AI problems involving cause and effect.”.

The specific engine of the Davinci model that was used was text-davinci-001 (previously known as davinci-instruct-beta-v3) which reads the prompt text as instructions for what to write instead of taking a prompt and attempting to continue the text which is normally the case. No further training was done such as Few-Shot or One-Shot in this experiment as it was deemed not necessary nor representative of a real life scenario. [13]

When generating text several different parameters can be adjusted and the engine is only one of these. Another important parameter is the temperature which is a value that can be set from 0 to 1 and it controls randomness in the text generation. A high temperature means that the text will be more random and human-like while a low temperature value could lead to the text becoming deterministic and repetitive according to OpenAI. In the testing it was found that the temperature setting 0.0 many times led to the language model generating the same output text when given the same prompt as opposed to different outputs for the same prompt which a temperature setting of 1.0 provided. The maximum number of tokens that can be used in the text generation can also be specified to control the maximum length of the message. One token is roughly 4 characters. Another parameter is Top-p which can be used to limit which of the least likely words that can be generated. An example of output from GPT-3 based on the prompt “Write a review of a vacuum cleaner:” is shown in figure 2. [14] [15]

Parameters:	Temperature: 1.0 Max tokens: 100 Top P: 1.0
Prompt:	Write a review of a vacuum cleaner:
Output:	We bought the Hoover WindTunnel T-Series Rewind UH70120 vacuum cleaner because we needed one with a lot of power. We were not disappointed. This vacuum cleaner has more than enough suction to get the job done. It's also very easy to operate. The cord retracts automatically, so there's no danger of tripping over it. The WindTunnel technology really works well, preventing the vacuum cleaner from getting clogged up with dirt.

Figure 2. An example of output from GPT-3 based on the prompt “Write a review of a vacuum cleaner:”.

4.4. Experiment

The experiment consisted of 10 texts where each text contained answers the following 4 questions:

- Illustrate the following: The 6 big V's of Big Data
- Illustrate the following: Structured and Unstructured Data
- Illustrate the following: Cloud Computing, Advantages and Disadvantages
- Illustrate the following: Edge Computing vs Fog Computing: What’s the Difference?

The subject read through each of the texts and would afterwards select if they believe the answers in the text were written by a human or an Artificial Intelligence. The length of each text excluding the questions were between 375 and 475 words with an average around 415 words. 5 of the texts were AI-generated and 5 were written by a human. This information was not provided to the subjects. The total cost of generating the 5 AI-texts was around 0.15$.

The questions were taken from an exam written on January 1, 2022 from the course Internet of Things and Network Applications. Permission to use the questions in the experiment was given by the course examiner. These questions were chosen for a number of reasons, one being because human written answers were already avaliable. Another reason is that they are from actual exams. It is Important to note that this exam was a sit-down examination and not a take-home exam. Because the questions are from actual exams they are suitable for the purposes of this experiment. It was also felt that the level of complexity in these questions was on par with questions found in other examinations.

Since both authors were familiar with the topics that the questions involved, an understanding of whether or not the AI was making logical and convincing statements in response to the questions was easily obtainable. The questions were also suitable because they range from factual questions to questions that require reasoning, such as the last one.

To be able to adequately decide if a text was written by a human or not, the subject must first have a basic understanding of the subject. That was why teachers belonging to the School of Information Technology at Halmstad University were selected as the main test subjects. Another positive aspect of using the teachers from the School of Information Technology at Halmstad University as subjects is that many of them have English as their primary language.

5. Result

5.1. To what extent can teachers differentiate AI-generated from human written essays?

The form that the experiment was conducted via was open from March 11, 2022 till April 1, 2022 and received 32 submissions with a total of 310 guesses made on the author type of the different texts in the experiment. Of the 32 subjects only 2 chose not to answer all questions. It is worth mentioning once more that the subjects were aware that they were being tested and would therefore look more thoroughly for signs of unnatural writing which would not be the case in a real life scenario.

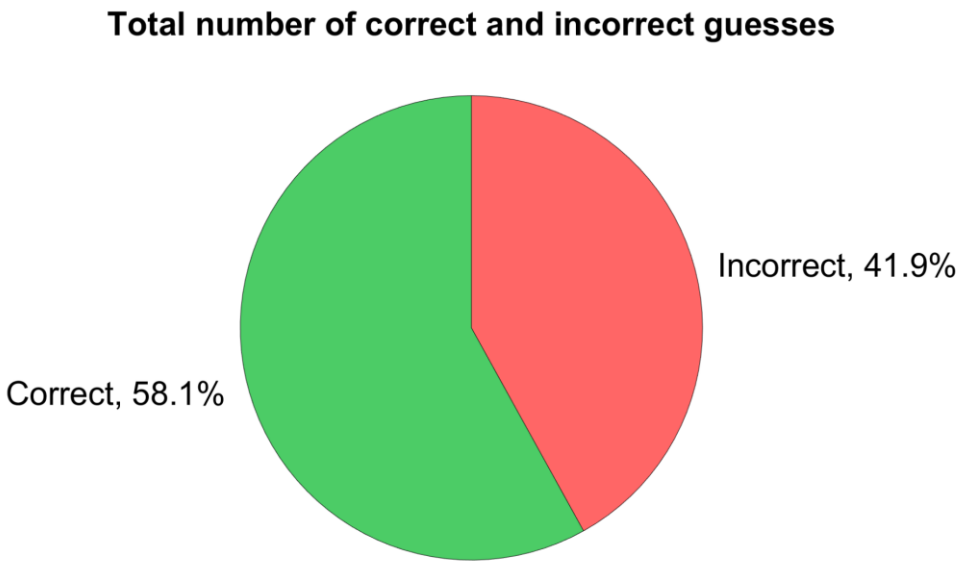


Figure 3. Results from the experiment on the total number of correct and incorrect guesses

The results of all the texts including both the AI-generated texts and the human written texts shown in Figure 3 were 180 (58.1%) correct guesses and 130 (41.9%) incorrect guesses.

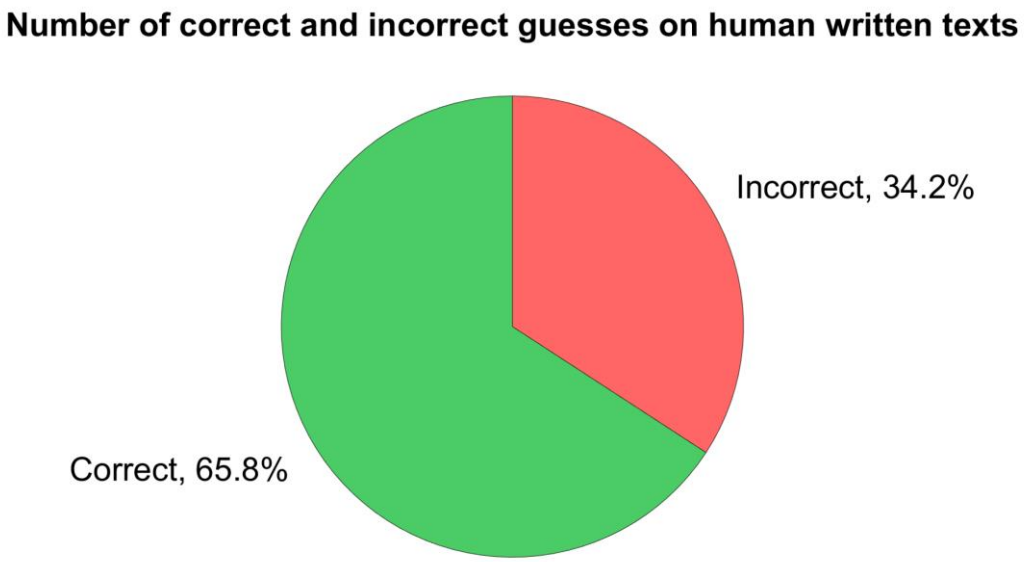


Figure 4. Results from the experiment on the number of correct and incorrect guesses on human written texts

The results for the 5 human written texts shown in Figure 4 were 102 (65.8%) correct guesses and 53 (34.2%) incorrect guesses. This indicate that teachers are to some degree able to distinguish human written text from a sample of both AI-generated texts and human written ones. The teacher's result was 15.8% percentage units above a coin flip.

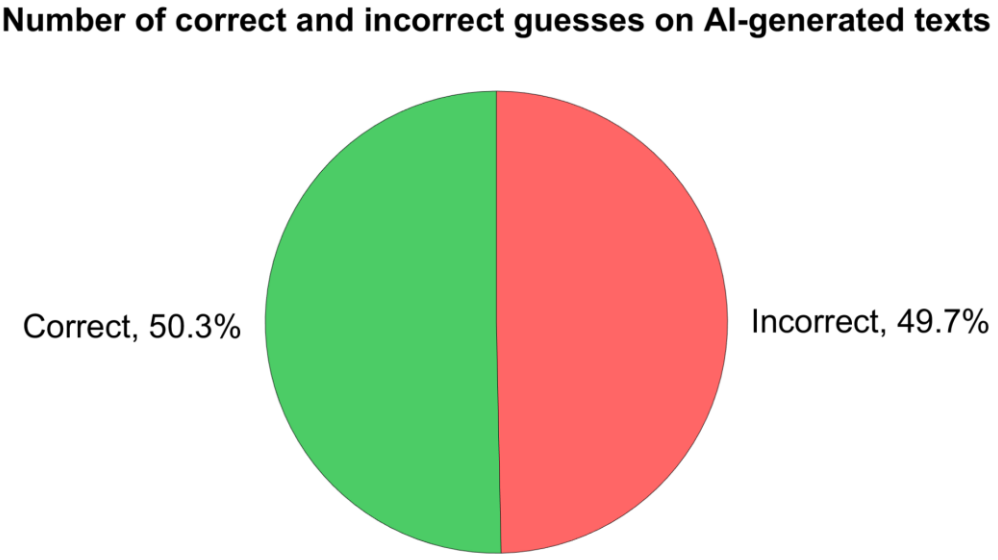


Figure 5. Results from the experiment on the number of correct and incorrect guesses on AI-generated tests

The results for the 5 AI-generated texts shown in Figure 5 were 78 (50.3%) correct guesses and 77 (49.7%) incorrect guesses. This means that the teachers were only able to correctly guess the author type of the AI-generated texts with a minuscule margin above pure guessing.

5.2. What countermeasures can hinder cheating using language models?

5.2.1. Computer based detection methods

If teachers' own ability to detect AI-generated text is not sufficient then computer-based detection methods might be necessary. Computer based detection methods would also be a viable way to combat this form of cheating, as it would allow for automatic detection of AI-generated text at large scale. November 2, 2020 Jawahar et. al published the paper “Automatic Detection of Machine Generated Text: A Critical Survey” where a literature review of the current computer-based detection methods for AI-generated text is presented. In their paper the aspects of an ideal detector is described as having the following features: accuracy, data-efficiency, generalizability, interpretability and robustness. Accurate in the sense that it would have low amounts of false positives and false negatives. Data-efficient means the input data required for an accurate detection. Arguably the most important aspect of such a detector would be generalizability, meaning that it is not bound by any specific language model such as RoBERTa or GPT-2/3. As mentioned earlier, the output from the detector being interpretable specifically for people with little knowledge in the area of AI is an important aspect. Lastly, the authors mention that a detector needs to be robust, meaning that it can handle adversarial examples. [16]

In their survey of current computer-based detection methods for AI-generated text, detectors are divided into three main groups based on their underlying methods: Zero-shot detectors, detectors requiring fine tuning and detectors trained from scratch. Zero-shot detectors use an already pre-trained language model such as GPT-3 or RoBERTa to detect text generated by either itself or from another model. These are the simplest to use as they are ready to use from the start. They are also the only option when no training data exists. This however, is often not an issue. In their survey two zero-shot detectors were mentioned: Total log probability and Giant Language model Test Room (GLTR). Fine-tuned detectors also use a pre-trained language model, however they are further trained using supervised training. In the paper two fine-tuned detectors are discussed, one based on GROVER and one based on RoBERTa. The main advantage of using zero-shot detectors compared to fine-tuned detectors is that they require no further training with the disadvantage that they generally perform worse and have a higher percentage of false positives and false negatives. Detectors trained from scratch employ “classical machine learning methods such as logistic regression to train a model from scratch” to be able to differentiate AI-generated text from human written text. Detectors trained from scratch do not use a language model as a base and therefore require much more training data compared to fine-tuned detectors since the fine-tuned detector is partially pre-trained. [16]

The paper concludes with an experiment of the RoBERTA detector's ability to detect Amazon product reviews written by GPT-2. The aim was to evaluate the difference in performance based on the number of training examples used to fine tune the detector as well as different sampling methods for training the detector. These sampling methods were pure, top-k and top-p (nucleus sampling). These sampling methods are all stochastic and sampling in this context refers to different ways to select the next token based on the probability distribution given by the model. Pure sampling samples directly from the probabilities predicted by the model while top-k limits sampling to only the tokens with the highest probability and top-p sampling only considers a predetermined number of tokens. [16]

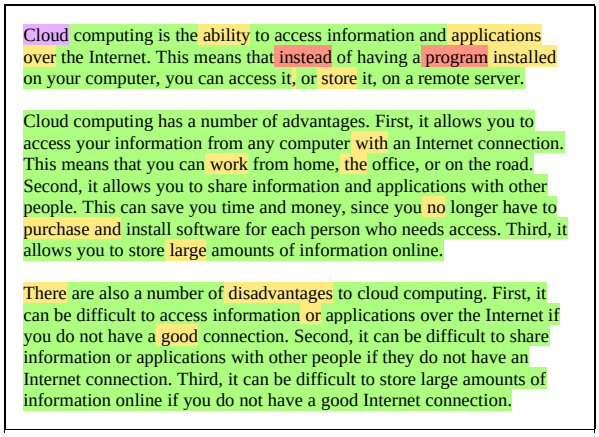
With only 1000 training examples the sampling method top-k had an accuracy of 80% while pure and top-p both had an accuracy of 70%. With 30000 training examples the top-k sampling method reached an accuracy of 94% while pure and top-p were both at 87% accuracy. Further increasing the samples to 200000 gives top-k an accuracy of 97%. The biggest challenge for RoBERTA in their experiment was the shortness of the reviews which forced RoBERTA to make a call with little data to base it on. Longer texts are heavily preferred when detecting the originator of the text as it leads to an increased data to work with. This behavior is true for all detectors. The RoBERTA detector also showed shortcomings in not noticing clear contradictions, errors in logic and unnecessary repetition that a human would clearly notice such as "My husband loves flavored coffee but is not a big fan of flavored coffee". This is an indication that RoBERTA is not able to comprehend the meaning of the text. Connecting this with the aspects of an ideal detector previously mentioned this shows that RoBERTA still has ways to improve in terms of accuracy. [16]

Due to GPT-3 being both new and not freely available to the public, there are currently no publicly available detectors for text specifically generated by GPT-3. However, there are multiple detectors available for other language models such as GPT-2, the predecessor to GPT-3. GPT-2 has been freely available since its release on February 14, 2019 and for that reason many detectors have been created for text generated by GPT-2 and are accessible via the Internet. Because of this it was decided to test the previously generated texts using GPT-3 against GPT-2 detectors to see if it was able to accurately detect the text as being not written by a human. GPT-2 and GPT-3 use the same foundational language model with the major difference being size of the training data and the number of parameters. Because of the differences between GPT-2 and GPT-3 it was unclear whether a detector built for detecting text written by GPT-2 would work on text written by GPT-3. Two publicly available tools that allowed for detection of text generated specifically by GPT-2 were selected to be evaluated. These two tools were Giant Language model Test Room (GLTR) and an implementation of the RoBERTa model fine-tuned to detect GPT-2 output. [2]

5.2.1.1. GLTR

GLTR was created by Sebastian Gehrmann, Hendrik Strobelt and Alexander M. Rush from MIT and Harvard. The detector is called the Giant Language model Test Room (GLTR) and it analyzes text based on how likely each word would be the predicted word given the context. Upon submitting a text, the program goes through the text word by word and uses GPT-2 itself to calculate the odds of different words being the next word in the string of text. What makes the detector work is the unique anomalies found in human writing. In other words, words not calculated to come netx based on GPT-2.

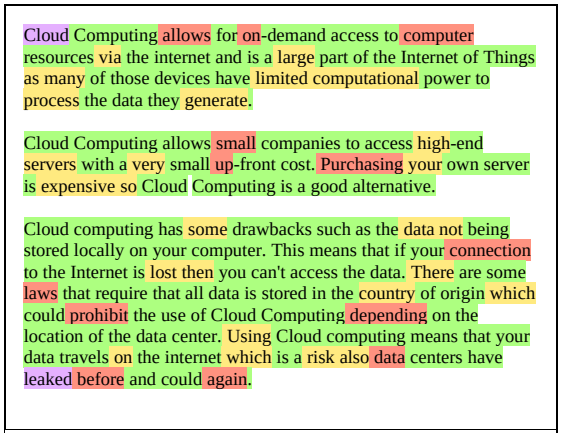
The detector highlights the words of the text a certain color based upon how likely that specific word would appear at that place in that string of text. A word that is highlighted green means that it was in the top 10 most likely to be generated, yellow means top 100, red top 1000 and purple for all other words. This gives the user a clear visual understanding of the likelihood that the text was written by GPT-2 as well which parts of a text where GPT-2's writing is potentially prevalent. The highlighting can be calibrated by changing the different thresholds to suit one's needs. A text written by a human would show a larger degree of yellow and red highlighted words. This would be because a human would not use the same choice of words and order of those words as GPT-2. The same 10 texts from the experiment (5 human written, 5 generated using GPT-3) were individually placed into this detector and it was able to consistently give the text written by humans a noticeable and quantifiable higher percentage of yellow, red and purple highlights than for the text written by GPT-3. For this evaluation the threshold values for the default highlighting settings were used.


Cloud computing is the ability to access information and applications over the Internet. This means that instead of having a program installed on your computer, you can access it, or store it, on a remote server.

Cloud computing has a number of advantages. First, it allows you to access your information from any computer with an Internet connection. This means that you can work from home, the office, or on the road. Second, it allows you to share information and applications with other people. This can save you time and money, since you no longer have to purchase and install software for each person who needs access. Third, it allows you to store large amounts of information online.

There are also a number of disadvantages to cloud computing. First, it can be difficult to access information or applications over the Internet if you do not have a good connection. Second, it can be difficult to share information or applications with other people if they do not have an Internet connection. Third, it can be difficult to store large amounts of information online if you do not have a good Internet connection.

Figure 6. Output from GLTR when provided with text generated using the GPT-3 engine text-davinci-001 with temperature set to 1 with the prompt "Illustrate the following: Cloud Computing, Advantages and Disadvantages".


Cloud Computing allows for on-demand access to computer resources via the internet and is a large part of the Internet of Things as many of those devices have limited computational power to process the data they generate.

Cloud Computing allows small companies to access high-end servers with a very small up-front cost. Purchasing your own server is expensive so Cloud Computing is a good alternative.

Cloud computing has some drawbacks such as the data not being stored locally on your computer. This means that if your connection to the Internet is lost then you can't access the data. There are some laws that require that all data is stored in the country of origin which could prohibit the use of Cloud Computing depending on the location of the data center. Using Cloud computing means that your data travels on the internet which is a risk also data centers have leaked before and could again.

Figure 7. Output from GLTR when provided with text written by a human answering the questions "Illustrate the following: Cloud Computing, Advantages and Disadvantages"

Figure 6 and Figure 7 shows the output of GLTR when provided with AI-generated and human written text respectively. An important aspect of GLTR is its ability to convey the result of its analysis to a person with no prior knowledge in AI or in detection of AI-generated text. This is something that the GLTR tool does remarkably well. GLTR also provides a pie

chart to further visualize the distribution in terms of how likely each word would be the predicted word given the context [17].

5.2.1.2. RoBERTa

The detector is hosted on the domain Huggingface and uses a detection method created by Jong Wook Kim, Jeff Wu and Alec Radford at OpenAI, the research lab who created GPT-2 and subsequently GPT-3. The detector outputs the predicted probabilities of the text being written by GPT-2 as a ratio between real and fake. Real or fake in this context refers to the text being authentically written by a human or by GPT-2 and works better the more text it is provided with a lower limit of around 40 words for reliable results.

The detector is an implementation of the RoBERTa language model which has been trained on large amounts of GPT-2 data. The main difference between the GLTR detector and this one is that GLTR uses GPT-2 to identify GPT-2 generated text while this detector is based upon the RoBERTa language model to identify GPT-2 generated text. The RoBERTa detector has proved very successful in identifying content generated by GPT-2. Whether using the language model itself for detection of its own text or if another language is better is outside the scope of this thesis.

In an effort to evaluate this detector, the same 10 texts of 5 human written answers and 5 answers generated using GPT-3 from the experiment were individually placed into this detector. The detector gave the 5 human written texts the following predicted probabilities of being real: 99.98% , 99.98%, 99.98%, 99.98% and 0,02%. The notable exception is the text that received a 0,02% likelihood of being real despite being fully written by a human. It is unclear why that specific text was given such a low degree of realness since the detector only provides a number while the inner workings are hidden from the user.

The detector gave the 5 texts generated using GPT-3 the following predicted probabilities of being written by a real person: 10.69% , 1.09%, 0.13%, 0.04% and 0.04%. This is a good result as all of the texts were given an accurate and therefore low score in terms of realness. The major strength and biggest weakness of this detector is in its simplicity. It is easy to use in terms of inputting data and generating an output, however because the output is condensed into a single number all nuances are lost which could leave the user wondering how the detector is reasoning.

With the exception of the unexplained anomaly that occurred with one of the human written texts a conclusion can be drawn regarding the detector when comparing the two groups of texts and that is that the detector would be a useful tool to have when deciding if a text was written by a human or an AI. This shows that the detector works well with text generated using GPT-3 despite being made for GPT-2. One could reasonably assume that a detector

built using the same internal logic but with GPT-3 as a base would perform much better with GPT-3 [18] [19].

In further experimenting with the two detectors, it became clear that the temperature parameter for GPT-3 text generation played a large role for the likelihood of a text being deemed suspicious. The greater the temperature overall, the lower accuracy of detection. The texts from the experiment were all generated with the max temperature of 1.0. An important aspect of these detectors not to be overlooked is their ability to do this processing both automatically and with minimal amount of input required from the user. This makes them simple to use and easy to implement into a larger system, such as a system for plagiarism detection. It is important not to see these results and conclude that a GPT-2 detector being able to some extent being able to detect GPT-3 text is a sign of this being close to a detector for AI-generated text in general. Different language models base their NLG on different methods and with different training and data-sets, they therefore most likely require different methods for detection.

5.2.2. Alternative examination methods

The most obvious way to hinder the use of language models would be to prohibit the use of computers during the examination such as in a supervised written exam. However, limiting examiners to only this form of exams would be heavily impairing. Since language models are able to write essays, instructions, manuscripts and even program code; all text based examinations where the student has access to a computer are in some way susceptible to cheating with the use of language models.

One way to hinder the use of AI-generated text as a tool for cheating in academia would be to use alternative examination methods. One such examination method could be to have oral exams since the students would not be able to in any way make use of a language model. One way to continue having written examinations while still hindering the use of language models could be to combine a written exam with a shorter hearing where the student is asked questions on the topic they have written about to gauge the authenticity. In the same light presentations could be an effective examination method in hindering the use of AI-generated texts only if coupled with a hearing after the presentation as language models have the ability to write scripts for a presentation.

Examination methods that include an illustrative aspect could be used to hinder the use of language models as they are limited to text only. However, adding an illustrative aspect to an examination is not always applicable and could cause more pain than it is worth. Using multiple choice questions could also make language models unusable. However, this form of examination is susceptible to online-searches for the answer. Practical exams are impervious to cheating using language models as they would serve no purpose as they are only possible to use for generating text. It must be noted that different examination methods such as practical

exams and written exams test different aspects of one's knowledge and therefore not all forms of examinations should be treated as if they are interchangeable in all situations.

6. Conclusion

AI-generated essays' is an issue currently facing academia and since it is becoming increasingly easy to generate text that can trick a human into thinking it was written by another human. The result of the experiment was that we could not prove that teachers have an ability to differentiate AI-generated text from human written text, especially not to the degree that is required as a basis for a report of suspected cheating. The tools required to generate essays text are accessible, easy to use and affordable.

The research into which countermeasures can hinder this type of cheating showed firstly that the computer-based detection methods that are currently available detectors for the previous generation GPT-2 generated text were promising to some degree for detecting GPT-3 text. However, their current error rate makes them highly inadequate as definitive proof in a case of suspected academic dishonesty. It also shows that detectors have a long way to go both in terms of accuracy and arguably more importantly generalizability as while GPT-3 is currently the leading model for NLP, other models will likely improve drastically in the coming years. Fortunately there are less expensive countmeasures that can be inplemented instead of sospihsitcated technical solutions. Examples include but are not limited to presentations, supervised written exam and oral examinations. By implementing these kinds of examinations you would counter the threat in its entirety

7. Discussion

In our experiment teachers' ability to differentiate AI-generated text from human written text could not be proven. This claim can also be supported by a statistical hypothesis test. But even if it would be possible for an examiner to easily distinguish an AI-generated text when grading papers there is still a problem that persists. Since an examiner needs to present a factual report where they highlight and prove that the student has cheated on their submission or exam, the examiner would need to be able to prove that the text is generated by a language model. The current tools available for universities are plagiarism that searches the internet for text that matches the submitted text. But since language models generate unique blocks of text with each use, the submission would pass this barrier without detection.

Current detection methods have a long way to go until they reach that level of accuracy required for them to be taken as proof of cheating in a disciplinary trial. It would be more reasonable and cost effective to modify the way examinations are conducted instead of trying to combat this issue with technology. However, since students using language models is not currently a major threat, moving away from traditional examination methods is not recommended. This might not be the case for long and at that moment the alternatives to text-based examination methods should be considered. Each alternative examination method comes with its own tradeoffs and these pros and cons must be weighed together for every individual case to decide what is suitable.

In the full span of the research 2$ of the research grant was spent on text generation, around 0.15$ of which went into generating the 5 texts for the experiment and the rest went into learning how to use the model and maximizing its potential. The current pricing of these models is already competitive and therefore pricing is not cited as a factor holding it back. Instead, it is believed that the main reason why the widespread use of this technology from students as a tool for cheating is not seen is a lack of knowledge of the existence, availability and capability of current language models. However, it is believed that this time will be short lived as students will find these tools and use them as a way to circumvent current plagiarism checkers that are mainly based on content matching. The market is clearly there as shown by the success of already existing applications such as Jasper.ai and Contentbot.ai.

English is the de facto lingua franca[6] of the world and this extends to the world of computers as well. In all the different interactions of GPT the data sets are composed of text in the English language which is why it is much more reliable and proficient in producing text in English. This does raise an interesting question. Since humans learn languages in a not too different way than a language model. We are trained by reading and consuming our target/native languages and develop a proficiency over time. Could it be possible that people

[6] A language that is used as a common language between speakers of different languages.

who have English as their first and native language would have a higher proficiency in discovering that there is something “unnatural” with the paper they have been presented with.

It is natural that you have a higher proficiency in your native tongue than in a language you have acquired later in life. You therefore have an instinctive feeling if something is natural or unnatural spoken or written. It could be an important tool in detecting AI-generated text. An experiment where instead of only testing the ability of examiners in general. It would be a very interesting experiment to test native vs. non-native speakers and have them explain why they find a text natural or unnatural.

As long as the risk of getting apprehended is minimal, which it currently is due to the severe limitations of current detection methods against AI-generated text, this attack vector will remain open. Fortunately the use of NLP does not currently pose a grave threat towards academia because of peoples lack of awareness of of language models. But the furure threat posed by language models in academica for cheating is serious and calls for countermeasures to be implemented to protect its integrity and reputation. Language models will no doubt play a large part in shaping the future of writing and what is known for certain is that a new tool has been added to the arsenal of those who seek to mislead us and gain an unfair academic advantage in examinations.

8. Future work

The authors believe that more research is necessary in this area, most importantly in the detection of AI-generated text. Research exploring the wider impact that NLP could have on all aspects of academia is also needed, including the benefits that were not highlighted as much in the report. Further research could also include evaluating the capabilities of other OpenAI GPT-3 models such as Curie, Babbage and Ada or even models from other developers such as RoBERTa.

This thesis was limited by its predetermined scale and set time frame and therefore only included the teachers of the School of Information Technology of Halmstad University reading texts on big data and cloud computing. A recreation of this thesis's main experiment on a larger scale and possibly in other languages would solidify the conclusions drawn in this thesis and expand the collective knowledge in this area. Evaluating the quality of the output in subjects other than IT such as philosophy or social science would also be very beneficial in future work. A deep dive into how these tools can be used to help people with considerable writing difficulties would also be of great value for this branch of research.

In conclusion, it is hoped that the results of this research will raise the awareness of the potential for cheating using NLP and the need for further research. The authors hope that this work will also inspire further research into the role of NLP in academia.

9. Bibliography

[1] M. Haenlein and A. Kaplan, “A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence,” California Management Review, vol. 61, no. 4, pp. 5–14, Aug. 2019, doi: 10.1177/0008125619864925.

[2] T. B. Brown et al., “Language Models are Few-Shot Learners,” arXiv:2005.14165 [cs], Jul. 2020 [Online]. Available: http://arxiv.org/abs/2005.14165. [Accessed: Apr. 01, 2022]

[3] “OpenAI Charter,” OpenAI. https://openai.com/charter/ (accessed Feb. 21, 2022).

[4] A. Singh, K. Ramasubramanian, and S. Shivam, “Natural Language Processing, Understanding, and Generation,” in Building an Enterprise Chatbot: Work with Protected Enterprise Data Using Open Source Frameworks, A. Singh, K. Ramasubramanian, and S. Shivam, Eds. Berkeley, CA: Apress, 2019, pp. 71–192. doi: 10.1007/978-1-4842-5034-1_5.

[5] G. Oppy and D. Dowe, “The Turing Test,” in The Stanford Encyclopedia of Philosophy, Winter 2021., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2021. Accessed: Feb. 21, 2022. [Online]. Available: https://plato.stanford.edu/archives/win2021/entriesuring-test/

[6] J. R. Searle, “Minds, brains, and programs,” Behav Brain Sci, vol. 3, no. 3, pp. 417–424, Sep. 1980, doi: 10.1017/S0140525X00005756.

[8] S. Axelsson, A. Viberg, and P. Kyrk, “Disciplinärenden 2021 vid universitet och högskolor rapport 2022:3.” Universitetskanslersämbetet, 2022. Accessed: Apr. 01, 2022. [Online]. Available: https://www.uka.se/download/18.3bd1341a17f449f672f50f1/1647354558830/rapport-2022-03-15-disciplinarenden-2021-vid-universitet-och-hogskolor.pdf

[7] “What we do,” UKÄ (Swedish Higher Education Authority). https://english.uka.se/about-us/what-we-do.html (accessed Apr. 02, 2022).

[9] R. Tripathi, P. Tiwari, and K. Nithyanandam, “Avoiding plagiarism in research through free online plagiarism tools,” in 2015 4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services, Jan. 2015, pp. 275–280. doi: 10.1109/ETTLIS.2015.7048211.

[10] “Better Language Models and Their Implications,” OpenAI, Feb. 14, 2019. https://openai.com/blog/better-language-models/ (accessed May 05, 2022).

[11] S. W. Gamage, E. Abdelaal, J. E. Mills (2019). Artificial Intelligence Is a Tool for Cheating Academic Integrity. S. W. Gamage, E. Abdelaal, and J. E. Mills, “Artificial Intelligence Is a Tool for Cheating Academic Integrity,” presented at the AAEE 2019 Annual Conference, Dec. 2019. [Online]. Available: https://www.researchgate.net/publication/339375213

[12] “Vad säger lagen? Etikprövningsmyndigheten,” Etikprövningsmyndigheten. https://etikprovningsmyndigheten.se/for-forskare/vad-sager-lagen/ (accessed Mar. 08, 2022).

[13] “OpenAI API.” https://beta.openai.com (accessed Feb. 14, 2022).

[14] “Pricing,” OpenAI. https://openai.com/api/pricing/ (accessed Feb. 23, 2022).

[15] R. Dale, “GPT-3: What’s it good for?,” Natural Language Engineering, vol. 27, no. 1, pp. 113–118, Jan. 2021, doi: 10.1017/S1351324920000601.

[16] G. Jawahar, M. Abdul-Mageed, and L. V. S. Lakshmanan, “Automatic Detection of Machine Generated Text: A Critical Survey,” arXiv:2011.01314 [cs], Nov. 2020, Accessed: Apr. 12, 2022. [Online]. Available: http://arxiv.org/abs/2011.01314

[17] S. Gehrmann, H. Strobelt, and A. M. Rush, “GLTR: Statistical Detection and Visualization of Generated Text,” arXiv:1906.04043 [cs], Jun. 2019, Accessed: Mar. 06, 2022. [Online]. Available: http://arxiv.org/abs/1906.04043

[18] “GPT-2 Output Detector.” https://huggingface.co/openai-detector (accessed Mar. 06, 2022).

[19] “Add RoBERTa-based GPT-2 Output Detector from OpenAI · huggingface/transformers@1c542df,” GitHub. https://github.com/huggingface/transformers/commit/1c542df7e554a2014051dd09becf60f157fed524 (accessed May 26, 2022).