Fine-Tuning GPT-3 for Synthetic Danish News Generation

Mina Almasi and **Anton Drasbæk Schiønning**

Aarhus University, Department of Linguistics, Cognitive Science and Semiotics
mina.almasi@post.au.dk, drasbaek@post.au.dk

Abstract

While GPT-3 has garnered significant attention for its capabilities in natural language generation, research on its use outside of English is still relatively limited. We focus on how GPT-3 can be fine-tuned for generating synthetic news articles in a low-resource language, namely Danish. The model's performance is evaluated on the dimensions of human and machine detection in two separate experiments. When presented with either a real or GPT-3 generated news article, human participants achieve a 58.1% classification accuracy. Contrarily, a fine-tuned BERT classifier obtains a 92.7% accuracy on the same task. This discrepancy likely pertains to the fine-tuned GPT-3 model oversampling high-likelihood tokens in its text generation. Although this is undetectable to the human eye, it leaves a statistical discrepancy for machine classifiers to detect. We address how decisions in the experimental design favoured the machine classifiers over the human evaluators, and whether the produced synthetic articles are applicable in a real-world context.

1 Introduction

In recent years, rapid development in natural language processing, particularly in the area of pretrained language models, has led to significant advancements in various language tasks. State-of-the-art models, such as GPT-3 (Brown et al., 2020) and BERT (Devlin et al., 2019), have excelled in tasks such as classification of documents (Kong et al., 2022), text completion (Balkus and Yan, 2022), language translation (Yan et al., 2022) and text summarization (Wazery et al., 2022). These advances have even led some to suggest that we are currently experiencing a whole paradigm shift in NLP with the introduction of pretrained language models (Min et al., 2021).

However, most contemporary studies using GPT-3 focus on its performance in English. This is to be expected as the model was almost exclusively trained on English with less than 8% of training data being non-English (OpenAI, 2020). Still, the few investigations on GPT-3 for non-English text generation show promising results (Kraft et al., 2022; Müller and Laurent, 2022). This even holds for low-resource languages such as Catalan (Armengol-Estapé et al., 2021).

Yet, the more prevalent approach in NLP for low-resource languages has been using smaller language-specific models or multilingual models such as mBERT (Doddapaneni et al., 2021). This is despite multilingual models seemingly lacking in natural language generation tasks, especially for the Nordic languages and other low-resource languages (Rönnqvist et al., 2019; Wu and Dredze, 2020). In terms of language-specific models, this development has also occurred in Danish NLP with several Danish models appearing based on the likes of BERT and ELECTRA (e.g., Tamimi-Sarnikowski, 2021 and Møllerhøj, 2021). Nevertheless, such models are miniscule in size compared to the state-of-the-art language models. For instance, the Danish BERT model by Møllerhøj (2021) is trained on 9.7 billion characters. Comparatively, GPT-3's total training data corresponds to 1.1 trillion characters (OpenAI, 2020).

In this paper, we seek to understand how well GPT-3 can perform for a low-resource language such as Danish when optimized for that language through fine-tuning. To our knowledge, this is the first structured assessment of GPT-3's capabilities in a Danish NLP task. Concretely, we investigate whether GPT-3 can be fine-tuned to produce synthetic news articles that are indistinguishable to real news articles written by journalists. Generating news articles with GPT-3 is a common task with previous work showing remarkable results in English (Brown et al., 2020; Uchendu et al., 2021).

Inspired by a similar study from Ippolito et al. (2020), we make a two-fold evaluation of the model's performance:

(A) Human Detection: Can untrained human participants distinguish between real and synthetic articles in an experimental setting?

(B) Machine Detection: Can machine classifiers be trained to distinguish between real and synthetic articles?

As human and machine detection methods presumably apply distinct techniques to spot synthetically generated text (Ippolito et al., 2020), a two dimensional evaluation provides a more nuanced insight into how GPT-3 performs on the task.

Our findings suggest that a fine-tuned GPT-3 can generate convincing Danish synthetic news, deceiving human readers while being identifiable by a BERT classifier. This demonstrates GPT-3's capacity to perform succesfully in the context of low-resource languages, but with the drawback of heightened machine-detectability due to an overuse of high-probability tokens.

2 Related Work

2.1 Natural Language Generation with Pretrained Language Models

Natural language generation (NLG) is a subfield of NLP concerned with the process of producing intelligible language. However, even within this subfield, there are a diverse range of related sub-tasks. Examples of such tasks, which have natural language as the input and output, are summarization, question answering and translation (Celikyilmaz et al., 2021).

Similar to other fields in NLP, text generation has evolved rapidly with the paradigm of pretrained language models. These models have been critical for advancing NLG as they understand natural language, express it fluently and are capable of being fine-tuned for a specific domain (Li et al., 2021). Importantly, pretrained language models can generate natural language that is novel rather than just outputting text memorized from the training data. This was demonstrated in McCoy et al. (2021) who found GPT-2 and Transformer-XL to produce novel words and unique syntactic structures not found in the training data.

The demonstrated successes of GPT-3 in NLG cannot only be attributed to the sheer amount of data it has seen, but also to the underlying decoder-transformer architecture. GPT-based models are built using only decoder blocks which possess a masked self-attention layer that prevents the language model from considering future context (Wang et al., 2022). This architecture is more easily applicable to NLG tasks than the alternative encoder-only structures found in BERT-based models (Lewis et al., 2020).

2.2 The Fine-Tuning Approach

The groundbreaking paper introducing GPT-3 titled, "Language Models are Few-Shot Learners" highlighted its ability to achieve strong performance on various NLP tasks after only seeing a few examples (Brown et al., 2020). This few-shot learning approach can be contrasted with fine-tuning in which the model is updated through re-training with task-specific data. Although GPT-3 arguably excels at text generation from few-shot learning, OpenAI recommends fine-tuning the model for custom applications citing advantages such as higher quality results.[1]

Related research has also already demonstrated the flexibility of GPT-3 by fine-tuning it for a wide variety of tasks. Perhaps the most ambitious fine-tune of GPT-3 is OpenAI's Codex which was trained on 159 GB of Python files from 54 million GitHub repositories. As a result of this fine-tune, Codex outperformed base GPT-3 on a benchmark on several different coding tasks (Chen et al., 2021). However, fine-tunes of GPT-3 extend beyond just massive applications. A study by Zong and Krishnamachari (2022) on extracting equations from math word problems found an 80% accuracy for a fine-tuned GPT-3 model compared to only 40% accuracy with 3-shot learning. Contrary to the enormous Codex fine-tune, their fine-tune had just seen 1000 examples. Similar small-scale fine-tunes of GPT-3 improved its abilities for assessing students' short answer questions (Moore et al., 2022) and writing less biased job advertisements (Borchers et al., 2022).

The takeaway from these studies is that GPT-3 can improve performance through fine-tuning for specific downstream tasks despite its generalized task excellence from just few-shot learning.

[1] https://platform.openai.com/docs/guides/fine-tuning

2.3 Evaluating Synthetically Generated Text

2.3.1 Human Detection

Evaluating whether artificial intelligence can deceive humans dates back to the Turing Test (Turing, 1950). While the current state of AI is incomparable to the 1950s, the underlying idea of judging machines on their human-like performance is still relevant. Much of research using this approach evaluated language models by asking participants to classify whether text excerpts were human-written or synthetically generated (Bogaert et al., 2022; Brown et al., 2020; Uchendu et al., 2021).

Although these classifications provide valuable insight into a language model's capabilities, they leave many questions as to why and how these models excel. For this reason, other studies ask participants to rate various qualities of the text without knowing whether the text is synthetic or real. The exact qualities that are rated differ across studies. For instance, some studies judge the overall text quality (Zhang et al., 2020) or fluency (Adelani et al., 2020) on a Likert scale. Dou et al. (2022)'s SCARECROW framework offers a more systematic approach to analyzing synthetic text, accessible to laypeople with basic training. It groups common error types within categories, like language errors for grammar and incoherence, and factual errors for incorrect or nonsensical information.

2.3.2 Machine Detection

Although SCARECROW provides a standardized human evaluation of language models, human detection may not be ideal for detecting GPT-3 news articles as low accuracies would suggest. For instance, Clark et al. (2021) found that human evaluators only unmasked GPT-3 news stories with 56% accuracy despite them being trained for the task. Yet, this does not imply that synthetic text cannot be detected at all. In fact, past research on synthetic text detection has found machines to be superior to humans (Ippolito et al., 2020; Meyer et al., 2022; Uchendu et al., 2021). For example, Ippolito et al. (2020) utilized both a bag-of-words logistic regression and a fine-tuned BERT, reporting much greater performance than human evaluators. While the BERT model was optimal, the bag-of-words model did not lag far behind. As formulated by the study, the high performing machine detectors are likely due to the sampling method of language models being skewed towards high-likelihood words. Therefore, synthetic text is more easily distinguishable from human language which has greater variability in word choice (Holtzman et al., 2020). This linguistic difference is also noted in other research (Gehrmann et al., 2019; Tay et al., 2020).

Nevertheless, models relying solely on word probabilities are still inferior to more complex language models such as BERT. This may indicate that there are other factors which differentiate real and synthetic articles that language models pick up on with fine-tuning. Just like Ippolito et al. (2020), Uchendu et al. (2021) found that the fine-tuned BERT was the best performing detector across text generated by 19 language models including GPT-3.

3 Data

The real news stories were all sourced from the Danish news site tv2.dk. In October 2022, TV2's news platform boasted over 3 million unique users (Danske Medier Research, 2022), which is more than half of Denmark's population. Hence, it makes an excellent representation of typical news content consumed by Danes. These articles were obtained via two channels: directly scraping from TV2 and employing the DaNewsRoom Danish news database (Varab and Schluter, 2020).

In the selection process, only article bodies with a minimum length of 100 words were considered, and longer articles were shortened to a maximum of 150 words. Although the exact threshold is somewhat arbitrary, it was kept in this range for two reasons. Firstly, accumulating costs for generating articles with the fine-tuned GPT-3 necessitated that we kept the articles short. Also, using longer articles would entail that each participant would evaluate fewer articles as their time was limited.

In total, 1866 real Danish news articles from TV2 were sourced and used for three purposes: Fine-tuning GPT-3 (1209 real articles), providing training/validation data for machine classifiers (609 real articles), and serving as test data in the experiments (48 real articles). Additionally, 657 synthetic articles were generated by the fine-tuned GPT-3 for training the classifiers (609 synthetic articles) and test data in the experiments (48 synthetic articles).

4 Methods

4.1 Fine-Tuning GPT-3

GPT-3, specifically *text-davinci-002*, was fine-tuned with 1209 pre-processed real news articles

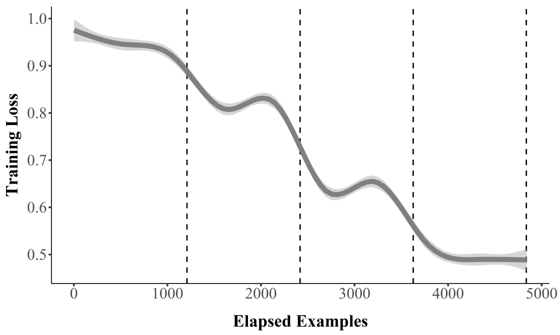


Figure 1: Training loss for fine-tuning GPT-3. The dashed lines indicate an epoch ending (1209 examples).

using OpenAI's API in Python. All articles were formatted to JSONL in accordance with the API documentation.[2] The headlines and subheadings were combined to be the prompts for GPT-3, and the corresponding article bodies were completions. All hyperparameters set for the fine-tune are detailed in Appendix A.1. As the training loss for the fine-tune plateaued during the fourth epoch (Figure 1), we ended model training after this epoch.

4.2 Generating Synthetic News Articles

The fine-tuned GPT-3 was then applied to generate synthetic news articles. As in the training phase, the prompts (headline and subheading) came from real news articles.[3] When generating the text completions, we modified several of the default hyperparameters based on previous research for similar cases and OpenAI's general recommendations.[4]

Firstly, GPT-3's temperature sampling method was adjusted by setting the temperature parameter to zero. In temperature sampling, a high temperature means that low probability tokens are more frequently sampled. By setting temperature to zero, the model becomes deterministic, always sampling the most probable token when generating text. We made this adjustment because a high temperature may lead to factual errors as the model "takes more risks". Also, Ippolito et al. (2020) found that a zero temperature in conjunction with a frequency penalty was the most successful for generating English news articles. This parameter penalizes a new token based on how frequently it appears in the generated text so far. It can be used along with a presence penalty (penalizes solely based on presence rather than frequency) to decrease the likelihood of sampling repetitive token sequences. As repetitiveness is also a particular concern for GPT-3's text generation (Dehouche, 2021), we add small presence and frequency penalties of 0.2. The full specification including ranges for the hyperparameters are in Appendix A.2.

The text completions formed the synthetic news articles, utilized as training/validation data for machine detection and test data for both experiments. Sanity checks were made to verify that these articles were similar to the real news articles in length and structure, but we made no modifications to them whatsoever.

5 Experiment A: Human Detection

Experiment A is designed as a binary classification task where human participants must distinguish between real articles written by humans and GPT-3's synthetic ones.

5.1 Participants

120 participants (66.6% female, age: M = 30.0, SD = 13.7) voluntarily took part in the online study. The study was run on the online platform SoSci Survey (Leiner, 2022) for one week in October 2022. To ensure a wide participant reach, the study was optimized for both computer and smartphone use. Complying with the prerequisites for the study, all participants were adult Danish native speakers.

5.2 Experimental Procedure

In each experimental trial, participants saw a page with one news article and four questions to be answered (see Appendix A.5). Participants were firstly asked to evaluate whether they believed the article body to be written by a human or an artificial intelligence. Subsequently, participants had to rate their confidence on a 5-point Likert scale from *completely unsure* (1) to *completely sure* (5). Finally, participants were asked to label whether the article had any distracting language or factual errors. These error types were inspired by the SCARECROW framework but simplified as the full framework would be too complex for untrained evaluators. To ensure participants understood what the error types implied, examples were written beneath each question. The articles were formatted to be closer in appearance with a real news article. This was done by differentiating in the size and color of

[2] https://platform.openai.com/docs/guides/fine-tuning/prepare-training-data

[3] To avoid double-dipping, these headlines and subheadings came from articles that were not part of the 1866 real articles mentioned in the data section.

[4] platform.openai.com/docs/api-reference

the headline, subheading and the article body (Appendix A.5). Importantly, it was clearly stated that only the body should be evaluated, not the headline and subheading as those always originated from real news stories.

In total, each participant evaluated 16 articles (8 real and 8 synthetic) in a randomly shuffled order. To cover the wide topical variance within news articles, 96 articles were used across all participants. That is, each participant only assessed a sixth of the total article pool, which corresponds to every article being evaluated by 20 unique participants.

5.3 Results

5.3.1 Human Accuracy

With 20 assessments of 96 articles, the human results are based on 1920 total classifications. The overall classification accuracy was just 58.1%. This means that participants only performed eight percentage points over chance level which is a comparable result to similar studies conducted in English (see 2.3). Interestingly, when presented with a synthetic news article, participants correctly labeled it as machine-written 53.6% of the time. Contrarily, a true positive rate of 62.6% indicates that participants were better at identifying real news articles as human-written. In addition, it should also be underlined that none of the 96 articles were exclusively classified correctly or incorrectly. The articles that were the easiest to identify were classified correctly 95% of the time, whereas there were only 15% correct classifications for the hardest ones.

Moreover, none of the 120 participants answered correctly on all 16 articles that they saw, with all of them misclassifying at least one synthetic news article as real news. This implies that the synthetic news articles have fooled all 120 participants to some extent.

Furthermore, all participants were screened on their news consumption level and prior knowledge of GPT-3. To see whether domain expertise caused enhanced performance, a mixed effects logistic regression model was run with media consumption level and GPT-3 knowledge as fixed effects. The news article ID is used as a random effect to account for variance that is specific to the articles.[5]

The full model output is displayed in Appendix A.6. The baseline/intercept in the model corresponds to a participant who never reads news and never had heard of GPT-3 prior to the experiment. The output reveals that a higher level of news consumption does not lead to significantly higher accuracies. However, compared to the baseline, we see significant improvements for participants that have heard of GPT-3 before ($\beta = 0.327, odds\,ratio = 0.581, SE_\beta = 0.131, p = 0.013$) and those participants that have additionally read GPT-3 texts ($\beta = 0.478, odds\,ratio = 0.617, SE_\beta = 0.146, p = 0.001$). This suggests that having GPT-3 knowledge may give an advantage in demarcating real from synthetic news, although participants who had worked with GPT-3 (highest level of GPT-3 knowledge) did not outperform the baseline.

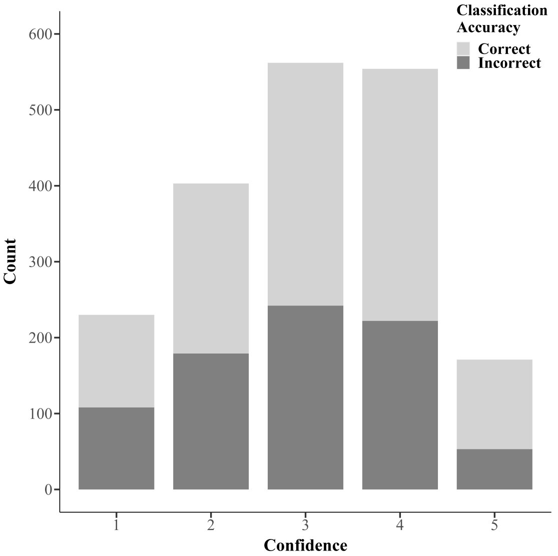


Figure 2: Confidence rating distribution of all trials. The fill indicates whether the corresponding classifications were correct or not.

5.3.2 Confidence and Error Identification

Participants were also asked to rate their confidence in the classification as well as marking error types for each trial. Figure 2 reveals that participants typically abstain from the most extreme confidence ratings of *Completely unsure (1)* and *Completely sure (5)*. As expected, participants' accuracy is around the chance level for low confidences. However, even when claiming to be *Completely sure*, the fraction of correct answers only increases to 69%. For confidences of *Fairly sure (4)*, this drops to only 60% correct answers.

We also see some interesting patterns in error type responses by the participants. Figure 3 illustrates which errors were marked for real and synthetic articles respectively. Overall, the pat-

[5] accuracy ~ news consumption + gpt-3 knowledge + (1 | article ID)

terns are strikingly similar. The figure reveals that participants most often did not find errors in the articles. When errors then were marked, there was a propensity to find language errors over factual errors for both real and synthetic articles. Despite the similarities, participants were more inclined to identify both factual and language errors for synthetic articles than for real ones. However, this did not necessarily lead to correct classifications. For instance, when participants marked *Both* errors, articles were almost exclusively labeled as synthetic (89.7% of cases) although Figure 3 reveals that this was often incorrect.

In sum, participants struggled with demarcating real news articles from synthetic ones in Experiment A. The overall accuracy was only 58% with classifications of synthetic news articles approaching chance level. Also, all 120 participants were fooled by at least one synthetic article and even the most confident classifications frequently led to wrong responses. Finally, patterns in error types marked by participants are similar for real and synthetic articles which shows the participants' inability to demarcate the articles by style and content.

6 Experiment B: Machine Detection

Experiment B explores whether it is possible to construct machine classifiers that are capable of distinguishing between real and synthetic articles. This is approached with logistic regression using bag-of-words (BOW) and TF-IDF as baseline models. The more advanced language model, NB-BERT-LARGE, is then fine-tuned, tested and evaluated against the baselines and human participants.

6.1 Building Classifiers

Two baseline classifiers are constructed using logistic regression with BOW and TF-IDF numerical representations of the vocabulary within the entire corpus (see Appendix A.3 for their hyperparameters). The BOW classifier is the most simple baseline, solely representing word frequencies within each document. TF-IDF provides a more detailed representation by also accounting for a word's rarity in relation to the entire set of documents.

Expanding beyond purely vocabulary-based classification, we fine-tune the BERT model, NB-BERT-LARGE (Kummervold et al., 2021), for the binary classification task. This BERT model was pretrained on the Norwegian Colossal Corpus which is a diverse collection of textual data (Kummervold et al., 2022). Although Norwegian is the primary language of the corpus, the collection contains several languages. Notably, Danish is the biggest language after Norwegian with 13.6% of the corpus being in Danish. We fine-tuned the model with the Trainer API using Hugging Face's transformers package (Wolf et al., 2020) in Python. The fine-tuning dataset comprised 1218 labeled articles split into a training and validation set (training: 75%, validation: 25%). Half of these were the real news articles from TV2 and the other half synthetic news articles. The test data comprised the same 96 articles that humans evaluated in Experiment A.

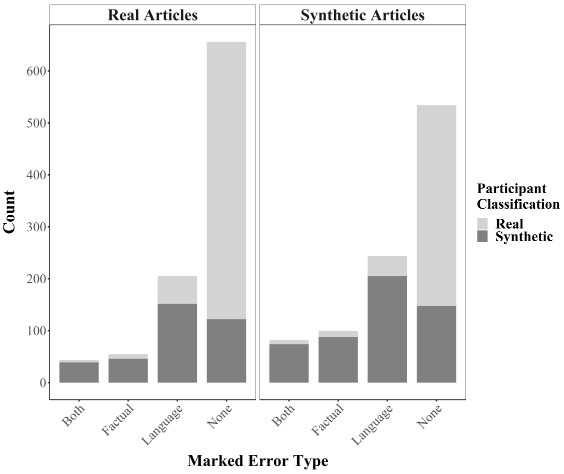


Figure 3: The marked error types by participants. On the left, the responses to real articles are found, and the synthetic responses are on the right. The fill indicates how participants classified the corresponding articles.

The hyperparameters for the fine-tuning of BERT are detailed in Appendix A.4. Resulting from an early stopping callback,[6] the model was fine-tuned for two epochs, obtaining a validation accuracy of 95.7%.

6.2 Results

6.2.1 Classification Accuracies

Table 1 shows the results of both the machine and human detection on the test data of 96 articles. The fine-tuned BERT model outclasses humans at the task with a 92.7% accuracy on the test set as well as the highest F1-score. Also, even the baseline BOW and TF-IDF models performed substantially better than the human average accuracy with accuracies around 80%, indicating that vocabulary discrepancies can demarcate the real and synthetic articles to an extent.

[6] based on the validation accuracy

Classifier	Accuracy	F1	Precision	Recall	TP	TN	FP	FN
Human	0.581	0.599	0.575	0.626	62.6%	53.6%	46.4%	37.4%
BOW	0.802	0.796	0.822	0.771	77.1%	83.3%	16.7%	22.9%
TF-IDF	0.802	0.800	0.809	0.792	79.2%	81.3%	18.8%	20.8%
BERT (fine-tuned)	0.927	0.927	0.932	0.927	87.5%	97.9%	2.1%	12.5%

Table 1: Evaluation metrics for all classifiers on the test data of 96 articles.

An interesting similarity between all machine classifiers is their tendency to classify articles as synthetic. This is most noticeable with fine-tuned BERT which has 12.5% false negatives as opposed to just 2.1% false positives. Remarkably, BERT's true negative classifications of 97.9% means that the model has only classified a single synthetic article wrong. This propensity to classify articles as synthetic contrasts human participants, who had a bias towards classifying most articles as real.

6.2.2 Classifier Agreement

We turn to examine classifier agreement quantitatively by evaluating their inter-rater reliability using Cohen's Kappa. Unsurprisingly, this metric reveals that TF-IDF and BOW have an almost perfect agreement, $\kappa = 0.91, z = 3.37, p < 0.05$. Moreover, both TF-IDF ($\kappa = 0.62, z = 6.14$) and BOW ($\kappa = 0.62, z = 6.11$) have a substantial agreement with BERT that is greater than would be expected by chance (both $p < 0.05$).

Table 2 gives a qualitative insight into the agreements with examples of how four test articles were classified. Article A was the most commonly misclassified article for humans (17 out of 20 misclassifications). However, interestingly, all three classifiers correctly identified it as synthetic news. Additionally, article B is one of four instances where BERT correctly identified a synthetic news article while both BOW and TF-IDF failed. Oppositely, article C provides an example of BERT's over-inclination to classify as synthetic. It is one of three articles where BERT misclassified a real news article while BOW and TF-IDF did not. Finally, article D is the only synthetic article that BERT misclassified. However, as Table 2 shows, BOW and TF-IDF also struggled with this article.

The overall takeaway remains that these machine detections performed vastly better than human participants. This improvement was clear even for the two baseline models based on BOW and TF-IDF. Still, the more sophisticated fine-tuned BERT classifier performed the best by far, with an impressive 92.7% overall accuracy and just a single misclassification of the 48 synthetic articles.

7 Limitations

A few limitations must be addressed in relation to these results. Firstly, several design decisions presumably favoured the machine detectors over the human evaluators. Whereas 78.3% of human participants had never seen GPT-3 produced texts before, all machine classifiers received extensive training on over 1000 labelled articles prior to the final testing. Also, the zero temperature token sampling for generating synthetic articles created an overrepresentation of high-likelihood tokens. This may be identified by the machine detectors, whereas such patterns are are probably too subtle to notice for humans (Ippolito et al., 2020). Also, Dou et al. (2022) show that higher temperatures are associated with GPT-3 making off-prompt errors. Such errors would not be captured by the machine classifiers, whereas humans would more likely identify these more semantic shortcomings.

Moreover, it must be addressed that human classifications are possibly influenced from being conducted in an experimental setting. Contrary to the machine classifiers, the human participants saw the headline and subheading for all articles. Despite being repeatedly told not to evaluate them, it cannot be dismissed that these extra elements still could have influenced their decision-making process. For instance, a familiar headline could have evoked an intuition for the article being real before reading the article body. On the other hand, one could argue that this was beneficial for humans as they could improve assessments by comparing contents in the headline and subheading to the article body.

Still, these methodological decisions systematically favored the machine classifiers over the human evaluators. However, asserting that the machine superiority would evaporate based on these considerations is a reach considering how vast the performance gap was.

Another limitation relates to the generalizability of the synthetic news articles. Due to experimen-

<table>
<tr><th colspan="5">Article A</th><th colspan="5">Article B</th></tr>
<tr><th>Correct</th><th>Human</th><th>BOW</th><th>TF-IDF</th><th>BERT</th><th>Correct</th><th>Human</th><th>BOW</th><th>TF-IDF</th><th>BERT</th></tr>
<tr><td><i>Synthetic</i></td><td><b>Real</b></td><td><b>Synthetic</b></td><td><b>Synthetic</b></td><td><b>Synthetic</b></td><td><i>Synthetic</i></td><td><b>Synthetic</b></td><td><b>Real</b></td><td><b>Real</b></td><td><b>Synthetic</b></td></tr>
<tr><td colspan="5">Greenland’s government has decided not to apply for permission for further oil drilling in the coming year. This is announced by the Greenlandic Ministry of Nature, Environment and Agriculture in a press release. "We have decided not to apply for oil drilling in 2023, because we want to spend time developing a new strategy for the Greenlandic economy, which will form the basis for a new oil and gas strategy," it says. The government also emphasizes that it will maintain its "vision of a fossil-free Greenland". The decision comes after a meeting on Tuesday between the government’s four parties. It is mainly the consideration for the climate that has led the government to drop further oil drilling.</td><td colspan="5">Two photographers and a culture minister are now criticized by the Press Council for having participated in a photo series where they posed with weapons. The Press Council writes this in a press release. In the case against Culture Minister Ane Halsboe-Jørgensen (S), the council has assessed that she has violated good press ethics by participating in the photo series ’The Gun Series’. "By participating in a photo series with weapons and ammunition, the Culture Minister has expressed that it is acceptable to carry weapons, whether it is in connection with artistic photography or not," the decision states. The decision against photographer Rasmus Flindt Pedersen and Jim Lyngvild is more stringent. Both have violated good press ethics by participating in the photo series, says the Press Council.</td></tr>
<tr><th colspan="5">Article C</th><th colspan="5">Article D</th></tr>
<tr><th>Correct</th><th>Human</th><th>BOW</th><th>TF-IDF</th><th>BERT</th><th>Correct</th><th>Human</th><th>BOW</th><th>TF-IDF</th><th>BERT</th></tr>
<tr><td><i>Real</i></td><td><b>Real</b></td><td><b>Real</b></td><td><b>Real</b></td><td><b>Synthetic</b></td><td><i>Synthetic</i></td><td><b>Real</b></td><td><b>Real</b></td><td><b>Real</b></td><td><b>Real</b></td></tr>
<tr><td colspan="5">Consideration for endangered animal species such as hazel dormice, birch mice, and bats in Denmark is now temporarily hindering a massive natural gas project that aims to supply Poland with natural gas from Norway. The Environmental and Food Appeals Board has annulled the project’s environmental permit, thereby halting the construction work of the Baltic Pipe pipeline across Denmark. "We are very disappointed with the decision," says Marian Kaagh, the deputy director of the company Energinet, which is responsible for the construction work in Denmark. In a press release, she says that Energinet has been working on a number of initiatives to ensure good living conditions for the animals in the areas where the pipeline is being built. This was a requirement when the Environmental Protection Agency granted the environmental permit for the Baltic Pipe project in 2019. However, according to the Environmental and Food Appeals Board, the conditions should have been thoroughly investigated before the permit was issued and construction work could begin.</td><td colspan="5">The upcoming super hospitals are meant to help improve the healthcare system in Denmark. However, they will not be completed on time. On average, the 16 hospital constructions are almost two years delayed, according to a statement from the Quality Fund for hospital constructions, which TV 2 has obtained access to. It is an expression of "a number of challenges," as the fund’s director, Morten Hjortenberg, puts it. "We had hoped for better results halfway through the construction period. It raises concerns and reflections on some of the decisions and priorities that were made during the projects," he says. The fund’s task is to provide funding for hospital constructions and ensure high quality – that is, what is often called "quality funds." The total budget for the constructions is over 30 billion Danish kroner – of which the state accounts for 23 billion and the regions’ self-financing contribution is 9 billion.</td></tr>
</table>

Table 2: Four article bodies from the test data translated to English. Predictions are marked in green if they were correct and red for incorrect. The human prediction is based on the majority classification for the 20 participants for an article (classified as real if split 50/50). See Appendix A.7 for the original articles in Danish.

tal constraints, articles were shortened greatly, and may therefore not be comparable to what we consider news in a real-world context. In addition, even if it could write longer articles, our fine-tuned GPT-3 model’s capabilities are practically useless in a journalistic context despite producing human-like outputs. This is because inferring a factually correct article body from just a headline requires additional, current context about the world which is inaccessible in this setup. Instead, the only thinkable purposes for this "headline-to-article news generator" have malicious undertones such as automating fake news production.

8 Conclusion

As advancements in natural language processing continue to progress rapidly, it is crucial to remember the importance of including and improving upon NLP in low-resource languages. This paper acknowledges this need by conducting a structured assessment of GPT-3’s abilities for Danish natural language generation when fine-tuned for the task.

Our study shows that GPT-3 can be fine-tuned to produce Danish synthetic news articles that are virtually indistinguishable to real news articles for humans. However, this does not imply that the articles are actually indistinguishable as the human eye is not all-seeing. By constructing a fine-tuned BERT model for the same discrimination task, we find that machine detection of the synthetic news articles was possible to a great extent. Hence, there must have been underlying flaws in GPT-3's article generations, likely relating to an oversampling of high-likelihood words.

The introduction of ChatGPT and GPT-4 will likely impact the findings presented in this paper, lowering detection accuracies further for both humans and machines. Although, as those models are closed-sourced, it would be troublesome to assess whether the testing articles are already part of the training data which poses a methodological challenge. Regardless, as our findings for Danish conform with similar studies in English, we encourage future work on low-resource languages to develop machine detectors which possibly stand the test when human evaluators are deceived.

Supplementary Materials Availability Statement: All source code used in the project is available from GitHub at https://github.com/drasbaek/finetuning-gpt3-danish-news. A dataset with the synthetic articles as well as classifications made by machine detectors is also available on the GitHub. The dataset containing human responses from Experiment A cannot be made available due to GDPR regulations. The real news articles from TV2 are also not made publicly available due to copyright limitations. In the interest of reproducibility, dummy data is made available on the GitHub which mimics the actual data to the greatest possible extent under the circumstances. Contact the authors for more information on the project.

Ethical Considerations

In this paper, we have created a GPT-3 fine-tune that is capable of producing synthetic news. As it may be possible to use it for malicious purposes, the fine-tuned model will not be available to anyone besides the authors. Per January 4, 2024, the authors will also lose access to the model as OpenAI announced all davinci models, including fine-tunes, will depreciate. [7] Nonetheless, we acknowledge that this paper demonstrates the ease of producing such a model, but also how it may be detected.

Finally, we recognize that the synthetic news produced for this paper could potentially contain societal biases from GPT-3's training data or from the real news articles used for fine-tuning.

[7] https://openai.com/blog/gpt-4-api-general-availability

Acknowledgements

The expenses associated with fine-tuning and using GPT-3 were funded by a grant from Apart Research, a machine learning safety research organization. This funding did not grant Apart Research any influence on the course of the study. The authors would also like to thank Ross Deans Kristensen-McLachlan for supervising the bachelor's thesis which laid the foundation for the present paper. Finally, we thank IT-vest and Stibofonden for funding the publication of this paper at INLG 2023.

References

David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. 2020. Generating Sentiment-Preserving Fake Online Reviews Using Neural Language Models and Their Human- and Machine-Based Detection. In *Advanced Information Networking and Applications*, Advances in Intelligent Systems and Computing, pages 1341–1354, Cham. Springer International Publishing.

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2021. On the Multilingual Capabilities of Very Large-Scale English Language Models. ArXiv:2108.13349 [cs].

Salvador Balkus and Donghui Yan. 2022. Improving Short Text Classification With Augmented Data Using GPT-3. ArXiv:2205.10981 [cs].

Jérémie Bogaert, Marie-Catherine de Marneffe, Antonin Descampe, and Francois-Xavier Standaert. 2022. Automatic and Manual Detection of Generated News: Case Study, Limitations and Challenges. In *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, MAD '22, pages 18–26, New York, NY, USA. Association for Computing Machinery.

Conrad Borchers, Dalia Gala, Benjamin Gilburt, Eduard Oravkin, Wilfried Bounsi, Yuki M Asano, and Hannah Kirk. 2022. Looking for a Handsome Carpenter! Debiasing GPT-3 Job Advertisements. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 212–224, Seattle, Washington. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020.

Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of Text Generation: A Survey. ArXiv:2006.14799 [cs].

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code. ArXiv:2107.03374 [cs].

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Danske Medier Research. 2022. Toplisten.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sumanth Doddapaneni, Gowtham Ramesh, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2021. A Primer on Pretrained Multilingual Language Models. ArXiv:2107.00676 [cs].

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic Detection of Generated Text is Easiest when Humans are Fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Jun Kong, Jin Wang, and Xuejie Zhang. 2022. Hierarchical BERT with an adaptive fine-tuning strategy for document classification. *Knowledge-Based Systems*, 238:107872.

Angelie Kraft, Hans-Peter Zorn, Pascal Fecht, Judith Simon, Chris Biemann, and Ricardo Usbeck. 2022. *Measuring Gender Bias in German Language Generation*. Gesellschaft für Informatik, Bonn. Accepted: 2022-09-28T17:10:03Z ISSN: 1617-5468.

Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2022. Norwegian Colossal Corpus Description.

D.J. Leiner. 2022. SoSci Survey.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pretraining for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Pretrained Language Model for Text Generation: A Survey. volume 5, pages 4492–4499. ISSN: 1045-0823.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2021. How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN. ArXiv:2111.09509 [cs].

Selina Meyer, David Elsweiler, Bernd Ludwig, Marcos Fernandez-Pichel, and David E. Losada. 2022. Do We Still Need Human Assessors? Prompt-Based GPT-3 User Simulation in Conversational AI. In *Proceedings of the 4th Conference on Conversational User Interfaces*, CUI '22, pages 1–6, New York, NY, USA. Association for Computing Machinery.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey. ArXiv:2111.01243 [cs].

Steven Moore, Huy A. Nguyen, Norman Bier, Tanvi Domadia, and John Stamper. 2022. Assessing the Quality of Student-Generated Short Answer Questions Using GPT-3. In *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption*, Lecture Notes in Computer Science, pages 243–257, Cham. Springer International Publishing.

Jens Dahl Møllerhøj. 2021. certainlyio/nordic_bert.

Martin Müller and Florian Laurent. 2022. Cedille: A large autoregressive French language model. ArXiv:2202.03371 [cs].

OpenAI. 2020. openai/gpt-3: Languages by Character Count.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is Multilingual BERT Fluent in Language Generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Phillip Tamimi-Sarnikowski. 2021. sarnikowski/danish_transformers.

Yi Tay, Dara Bahri, Che Zheng, Clifford Brunk, Donald Metzler, and Andrew Tomkins. 2020. Reverse Engineering Configurations of Neural Text Generation Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 275–279, Online. Association for Computational Linguistics.

A. M. Turing. 1950. Computing Machinery and Intelligence. *Mind*, 59(236):433–460. Publisher: [Oxford University Press, Mind Association].

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Varab and Natalie Schluter. 2020. DaNewsroom: A Large-scale Danish Summarisation Dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6731–6739, Marseille, France. European Language Resources Association.

Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2022. Pre-Trained Language Models and Their Applications. *Engineering*.

Y. M. Wazery, Marwa E. Saleh, Abdullah Alharbi, and Abdelmgeid A. Ali. 2022. Abstractive Arabic Text Summarization Based on Deep Learning. *Computational Intelligence and Neuroscience*, 2022:e1566890. Publisher: Hindawi.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are All Languages Created Equal in Multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Rong Yan, Jiang Li, Xiangdong Su, Xiaoming Wang, and Guanglai Gao. 2022. Boosting the Transformer with the BERT Supervision in Low-Resource Machine Translation. *Applied Sciences*, 12(14):7195. Number: 14 Publisher: Multidisciplinary Digital Publishing Institute.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11328–11339. PMLR. ISSN: 2640-3498.

Mingyu Zong and Bhaskar Krishnamachari. 2022. Solving Math Word Problems Concerning Systems of Equations with GPT-3. *Proceedings of the Thirteenth AAAI Symposium on Educational Advances in Artificial Intelligence*, page 8.

A Appendix

A.1 Fine-tuning Parameters for GPT-3

Parameters	Value
Batch Size	2
Learning Rate Multiplier	0.2
Prompt Loss Weight	0.01
Epochs	4

A.2 Text Generation Parameters for GPT-3

Parameters	Value	Value Range
Temperature	0	0 to 1
Frequency Penalty	0.2	-2 to 2
Presence Penalty	0.2	-2 to 2
Max Tokens	400	0 to 2048

A.3 Logistic Regression Parameters for BOW and TFIDF

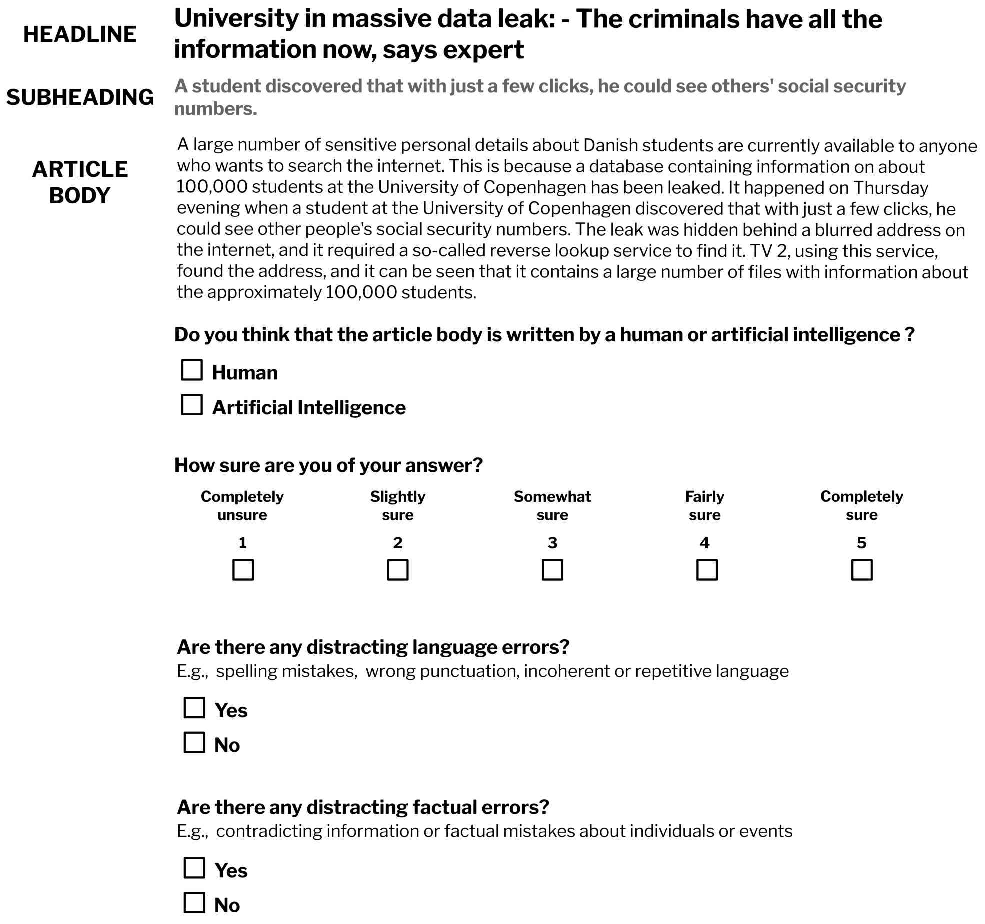
Parameters	Value
Solver	lfbgs
C	10
Random State	2
Max Iterations	250

A.4 Fine-Tuning Parameters for BERT

Parameters	Value
Learning Rate	2e-5
Weight Decay	0.01
Batch Size	24
Epochs run	2
Max Epochs	5

The fine-tuned BERT was defined with an early stopping callback which stopped model training if the validation accuracy did not improve for 3 epochs. The final model used for inference was thus only run for 2 epochs.

A.5 Experimental Procedure



HEADLINE

University in massive data leak: - The criminals have all the information now, says expert

SUBHEADING

A student discovered that with just a few clicks, he could see others' social security numbers.

ARTICLE BODY

A large number of sensitive personal details about Danish students are currently available to anyone who wants to search the internet. This is because a database containing information on about 100,000 students at the University of Copenhagen has been leaked. It happened on Thursday evening when a student at the University of Copenhagen discovered that with just a few clicks, he could see other people's social security numbers. The leak was hidden behind a blurred address on the internet, and it required a so-called reverse lookup service to find it. TV 2, using this service, found the address, and it can be seen that it contains a large number of files with information about the approximately 100,000 students.

Do you think that the article body is written by a human or artificial intelligence ?

□ **Human**

□ **Artificial Intelligence**

How sure are you of your answer?

Completely unsure	**Slightly sure**	**Somewhat sure**	**Fairly sure**	**Completely sure**
1	**2**	**3**	**4**	**5**
□	□	□	□	□

Are there any distracting language errors?

E.g., spelling mistakes, wrong punctuation, incoherent or repetitive language

□ **Yes**

□ **No**

Are there any distracting factual errors?

E.g., contradicting information or factual mistakes about individuals or events

□ **Yes**

□ **No**

Illustration of a trial from experiment A. All text was written in Danish in the actual experiment. The article body in the example is synthetically generated. The captions "HEADLINE", "SUBHEADING" and "ARTICLE BODY" did not appear in the actual experiment.

A.6 Logistic Regression Model Output for Predicting Accuracy

Fixed Effect	Estimate	Standard Error	Z-value	P-value
Intercept	0.33668	0.39335	0.856	0.39204
News_Consumption_2	-0.50311	0.43260	-1.163	0.24484
News_Consumption _3	-0.03473	0.39697	-0.087	0.93028
News_Consumption_4	-0.27516	0.40664	-0.677	0.49862
News_Consumption_5	-0.10105	0.39719	-0.254	0.79817
GPT_Knowledge_2	0.32738	0.13130	2.493	0.01266
GPT_Knowledge_3	0.47842	0.14626	3.271	0.00107
GPT_Knowledge_4	0.37824	0.22513	1.680	0.09293

Fixed Effect Level	Participant Response (translated)
News_Consumption_1	Never read the news
News_Consumption_2	Very rarely read the news
News_Consumption_3	Read news every week but not daily
News_Consumption_4	Read news once every day
News_Consumption_5	Read news multiple times a day
GPT_Knowledge_1	Never heard of GPT-3
GPT_Knowledge_2	Heard of GPT-3, but never read anything it wrote or worked with it
GPT_Knowledge_3	Heard of GPT-3 and read texts it wrote, but never worked with it
GPT_Knowledge_4	Heard of GPT-3, read texts it wrote and worked with it
	Participant Response (original)
	Læser aldrig nyheder
	Læser meget sjældent nyheder
	Læser nyheder hver uge men ikke dagligt
	Læser nyheder en gang om dagen
	Læser nyheder flere gange om dagen
	Aldrig hørt om GPT-3
	Hørt om GPT-3, men aldrig læst noget den har skrevet eller arbejdet med den
	Hørt om GPT-3 og læst tekster den har skrevet, men aldrig arbejdet med den
	Hørt om GPT-3, læst tekster den har skrevet og arbejdet med den

A.7 Classifier Agreement (Table 2) Danish Original Text

<table>
<tr><th colspan="5">Article A</th></tr>
<tr><th>Correct</th><th>Human</th><th>BOW</th><th>TF-IDF</th><th>BERT</th></tr>
<tr><td><i>Synthetic</i></td><td>Real</td><td>Synthetic</td><td>Synthetic</td><td>Synthetic</td></tr>
<tr><td colspan="5">Grønlands regering har besluttet sig for ikke at søge om tilladelse til yderligere olieboringer i det kommende år. Det oplyser det grønlandske ministerium for natur, miljø og landbrug i en pressemeddelelse. - Vi har besluttet os for ikke at søge om olieboringer i 2023, fordi vi vil bruge tid på at udvikle en ny strategi for den grønlandske økonomi, som skal danne grundlag for en ny olie- og gasstrategi, lyder det. Regeringen understreger samtidig, at den vil fastholde sin "vision om et fossilfrit Grønland". Beslutningen kommer efter et møde tirsdag mellem regeringens fire partier. Det er især hensynet til klimaet, der har fået regeringen til at droppe yderligere olieboringer.</td></tr>
</table>

<table>
<tr><th colspan="5">Article B</th></tr>
<tr><th>Correct</th><th>Human</th><th>BOW</th><th>TF-IDF</th><th>BERT</th></tr>
<tr><td><i>Synthetic</i></td><td>Synthetic</td><td>Real</td><td>Real</td><td>Synthetic</td></tr>
<tr><td colspan="5">To fotografer og en kulturminister får nu kritik af Pressenævnet for at have deltaget i en billedserie, hvor de poserede med våben. Det skriver Pressenævnet i en pressemeddelelse. I sagen mod kulturminister Ane Halsboe-Jørgensen (S) har nævnet vurderet, at hun har brudt god presseskik ved at deltage i billedserien ’The Gun Series’. - Kulturministeren har ved deltagelse i billedserie med våben og ammunition givet udtryk for, at det er acceptabelt at bære våben, uanset om det er i forbindelse med kunstnerisk fotografering eller ej, lyder det i afgørelsen. Afgørelsen mod fotografen Rasmus Flindt Pedersen og Jim Lyngvild er mere knibsk. Begge har brudt god presseskik ved deltagelse i billedserien, mener Pressenævnet.</td></tr>
</table>

<table>
<tr><th colspan="5">Article C</th></tr>
<tr><th>Correct</th><th>Human</th><th>BOW</th><th>TF-IDF</th><th>BERT</th></tr>
<tr><td><i>Real</i></td><td>Real</td><td>Real</td><td>Real</td><td>Synthetic</td></tr>
<tr><td colspan="5">Hensynet til truede dyrearter som hasselmus, birkemus og flagermus i Danmark stikker nu en midlertidig kæp i hjulet på et enormt naturgasprojekt, der skal forsyne Polen med naturgas fra Norge. Det er Miljø- og Fødevareklagenævnet, der har annulleret projektets miljøtilladelse, og dermed har sat en stopper for anlægsarbejdet af rørledningen Baltic Pipe på tværs af Danmark. - Vi er meget kede af afgørelsen, siger Marian Kaagh, der er vicedirektør i selskabet Energinet, der står for anlægsarbejdet I Danmark. I en pressemeddelelse siger hun, at Energinet har arbejdet med en række tiltag for at sikre gode levevilkår for dyrene de steder, hvor rørledningen bliver anlagt. Det var et krav, da Miljøstyrelsen i 2019 gav miljøtilladelse til Baltic Pipe-projektet. Men ifølge Miljø- og Fødevareklagenævnet burde forholdene være grundigt undersøgt, allerede inden tilladelsen blev udstedt, og anlægsarbejdet kunne begynde.</td></tr>
</table>

<table>
<tr><th colspan="5">Article D</th></tr>
<tr><th>Correct</th><th>Human</th><th>BOW</th><th>TF-IDF</th><th>BERT</th></tr>
<tr><td><i>Synthetic</i></td><td>Real</td><td>Real</td><td>Real</td><td>Real</td></tr>
<tr><td colspan="5">De kommende supersygehuse skal være med til at løfte sundhedsvæsenet i Danmark. Men de bliver ikke klar til tiden. Gennemsnitligt er de 16 sygehusbyggerier knap to år forsinkede, viser en opgørelse fra Kvalitetsfonden for sygehusbyggerierne, som TV 2 har fået aktindsigt i. Det er et udtryk for, at der er "en del udfordringer", som fondens direktør, Morten Hjortenberg, siger det. - Vi havde håbet på bedre resultater her halvvejs inde i byggeperioden. Det giver anledning til bekymring og eftertanke om nogle af de beslutninger og prioriteringer, der blev truffet under projekterne, siger han. Fondens opgave er at stille penge til rådighed for sygehusbyggerierne og sikre en høj kvalitet – altså det man ofte kalder "kvalitetsfonde". Byggeriernes samlede budget er på over 30 milliarder kroner – heraf står staten for 23 milliarder og regionernes selvfinansierende bidrag på 9 milliarder.</td></tr>
</table>