INLG 2023

Fine-Tuning GPT-3 for Synthetic Danish News Generation

Mina Almasi and **Anton Drasbæk Schiønning**

AARHUS UNIVERSITY

Department of Linguistics, Cognitive Science and Semiotics

Overview

00 **Who Are We?**

01 **Research Question**

02 **Text Generation with GPT-3**

03 **Experiment A: Human Detection**

04 **Experiment B: Machine Detection**

05 **Limitations & Main Takeaways**

Who Are We?

🇩🇰 **Mina Almasi**

Junior Developer, ML Team
@ Center for Humanities Computing,
Aarhus University

🇩🇰 **Anton D. Schiønning**

Data Analyst @ Off The Pitch

2023 - 2025	**MSc. in Cognitive Science**
2020 - 2023	**BSc. in Cognitive Science**

AARHUS
UNIVERSITY

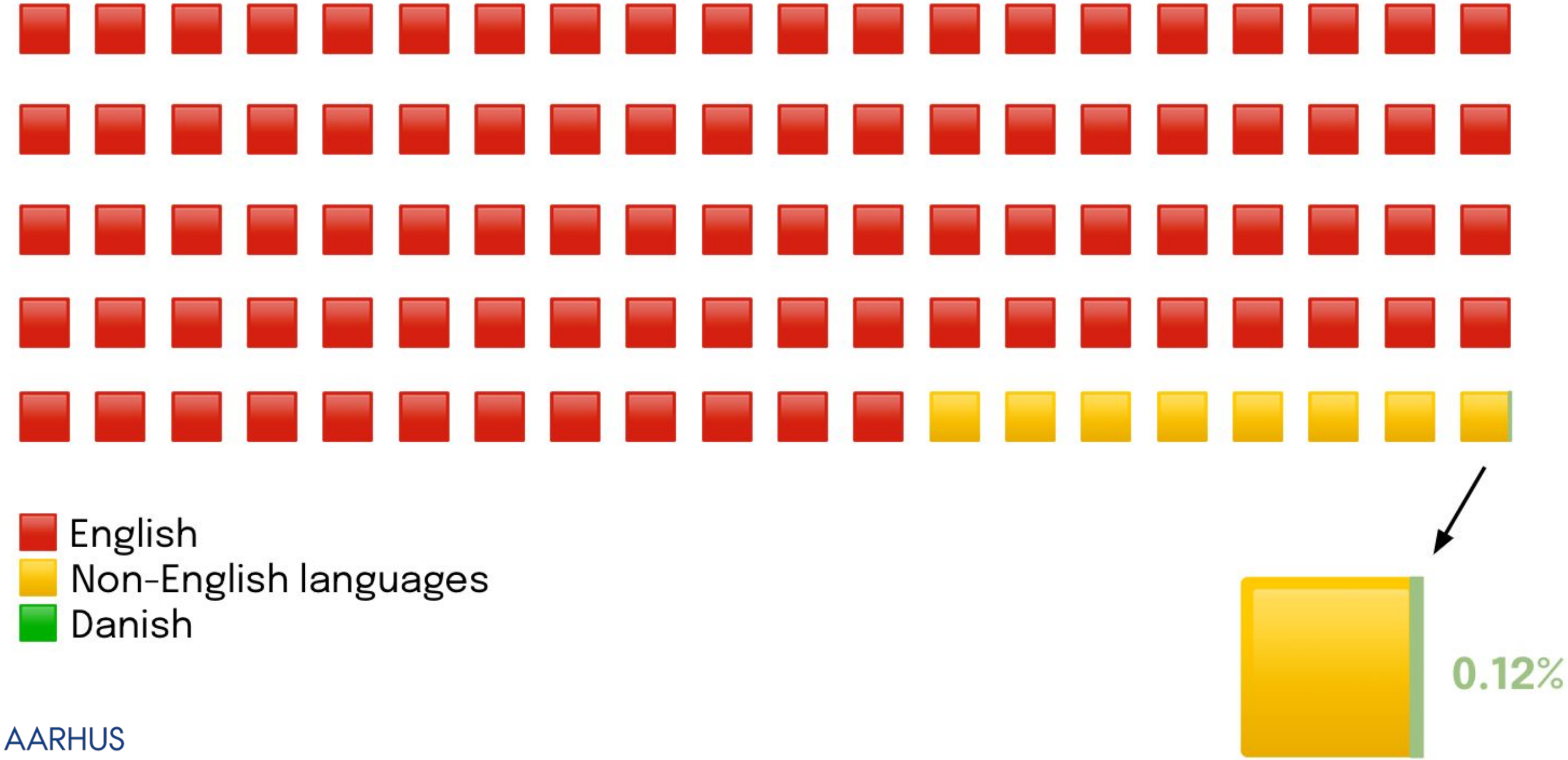
01

Research Question

GPT-3's Training Data Languages*

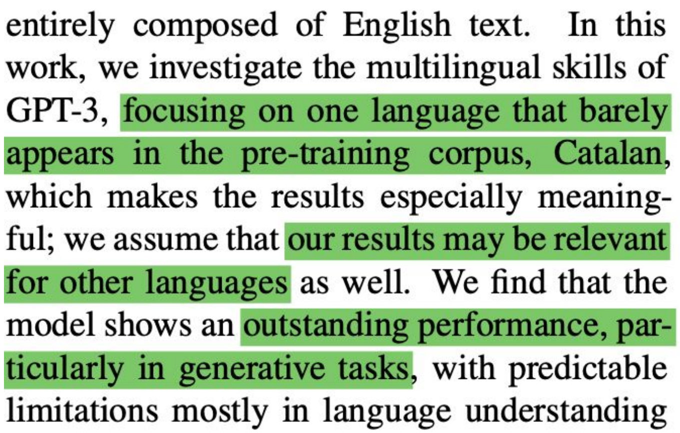
(OpenAI, 2020)

**based on character counts*



GPT-3 for Low-Resource Languages

GPT-3's Performance in Catalan
(Armengol-Estapé et al., 2021)



entirely composed of English text. In this work, we investigate the multilingual skills of GPT-3, focusing on one language that barely appears in the pre-training corpus, Catalan, which makes the results especially meaningful; we assume that our results may be relevant for other languages as well. We find that the model shows an outstanding performance, particularly in generative tasks, with predictable limitations mostly in language understanding

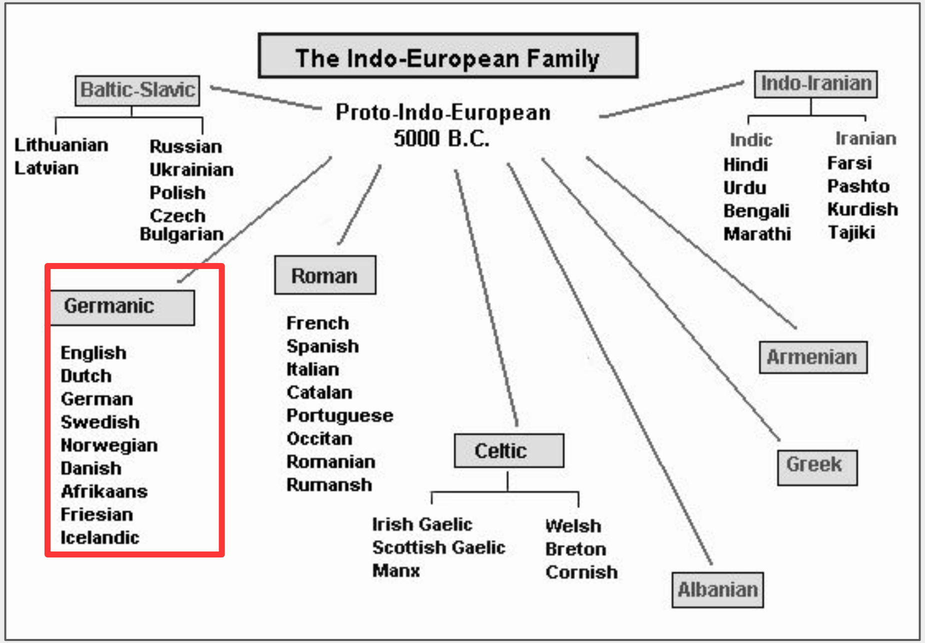
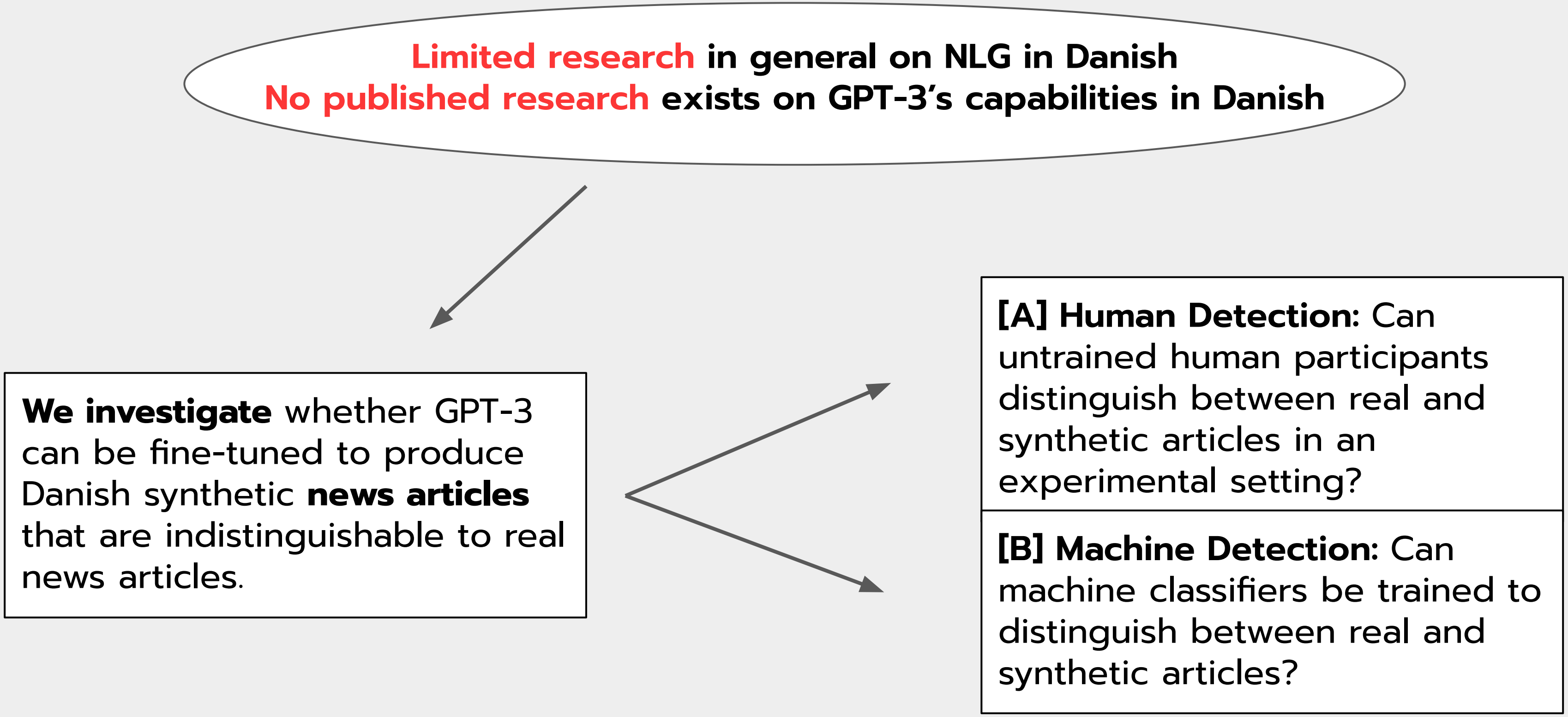


Figure by University of Ottowa (n.d.)

Research Question



Limited research in general on NLG in Danish
No published research exists on GPT-3's capabilities in Danish

We investigate whether GPT-3 can be fine-tuned to produce Danish synthetic **news articles** that are indistinguishable to real news articles.

[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?

Inspired by similar framework by Ippolito et al. (2020)

02

Text Generation with GPT-3

Fine-Tuning GPT-3

Previous Examples of Performance Enhancements

(Chen et al., 2021)

Codex: Solving various coding tasks

(Zong and Krishnamachari, 2022)

Extracting equations from math word problems

(Moore et al., 2022)

Assessing students' short answer questions

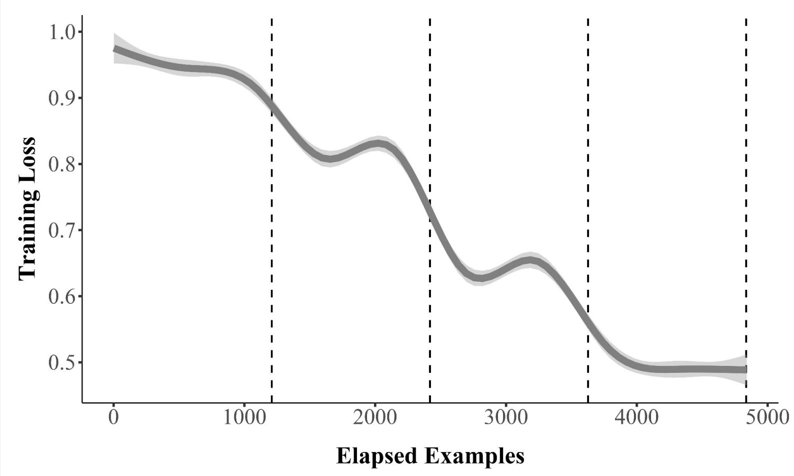
(Borchers et al., 2022)

Writing less biased job advertisements

Fine-Tuning GPT-3

How our Danish News Writing Model was Fine-Tuned

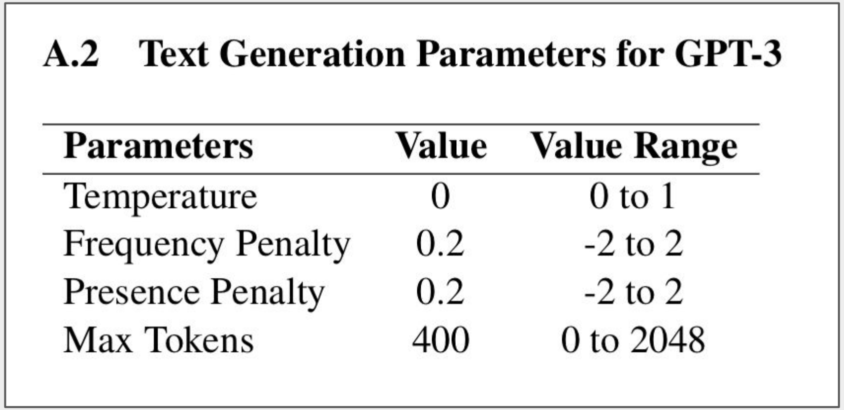
Using **1209 real news articles** from a Danish news site (TV2.dk)



		Example
Prompt	Headline + Subheading	Threatened Danish small animals delay giant project across Denmark. The Complaints Board demands new environmental studies before natural gas pipeline can be completed.
Completion	First 100-150 words of article body	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 (...)

Generating Synthetic News Articles

Setting Generation Parameters



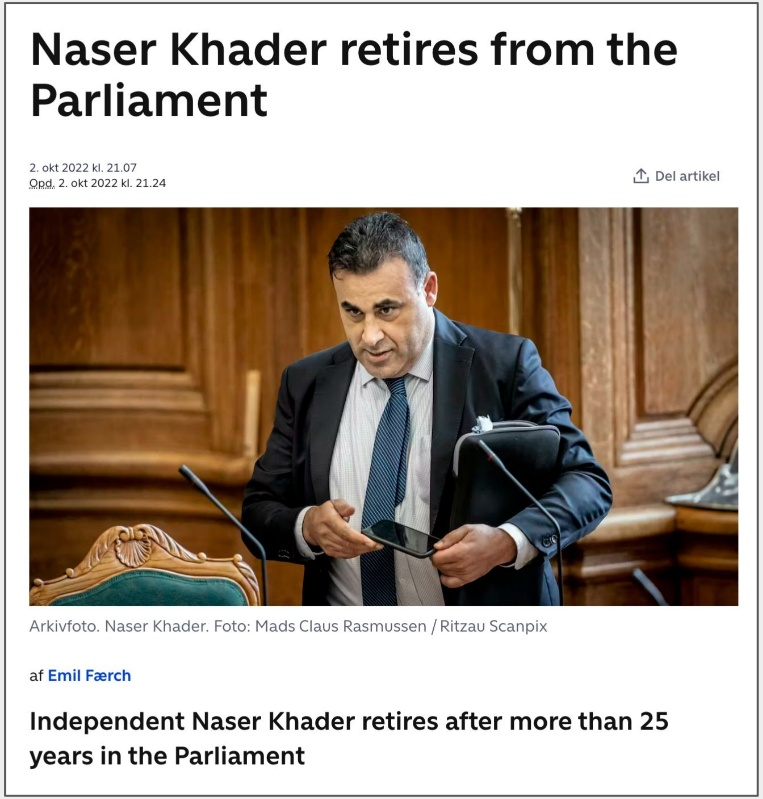
A.2 Text Generation Parameters for GPT-3

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

Temperature: Sampling Tokens
Set to **0** as high temperatures lead to the model "taking more risks"

Frequency and **Presence Penalty**
Penalizing the sampling of repetitive tokens

Generating Synthetic News Articles



Naser Khader retires from the Parliament

2. okt 2022 kl. 21.07
Opd. 2. okt 2022 kl. 21.24

Del artikel

Arkivfoto. Naser Khader. Foto: Mads Claus Rasmussen / Ritzau Scanpix

af **Emil Færch**

Independent Naser Khader retires after more than 25 years in the Parliament

Generation with a Temperature of 1

“A second and a third morning, Naser Khader [Danish politician] has stood up in parliament and yelled ‘F*ck’ to the Environmental Committee. Secondly, he has not slept in parliament for two days, he explains (...) And thirdly, Khader drank a double-espresso for lunch a single time, as far as he recalls.”

Not Ideal....

(Original text in Danish, English translation displayed)

03

Experiment A

Human Detection

Can untrained human participants distinguish between real and synthetic articles in an experimental setting?

AARHUS UNIVERSITY

[A] Experimental Design

120 participants
Danish native speakers

16 articles
*Assessed by **each** participant*
(8 real and 8 synthetic)

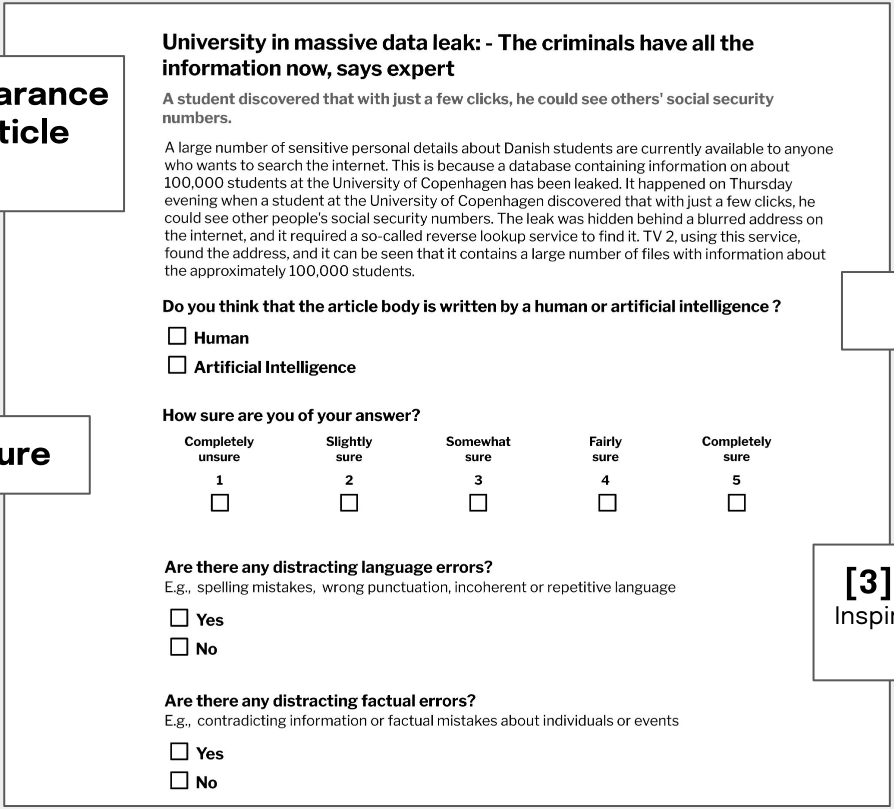
96 articles
In total

Each article
*assessed by **20** participants*

[A] Experimental Design

Illustration of a trial

[0] Formatting the appearance to resemble a news article



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

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

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**

[1] Binary question

[2] Confidence measure

[3] Language & factual errors
Inspired by the SCARECROW framework (Dou et al., 2022)

(Note that the article body in the figure is synthetically generated)

(Original text in Danish)

AARHUS UNIVERSITY

[A] Results

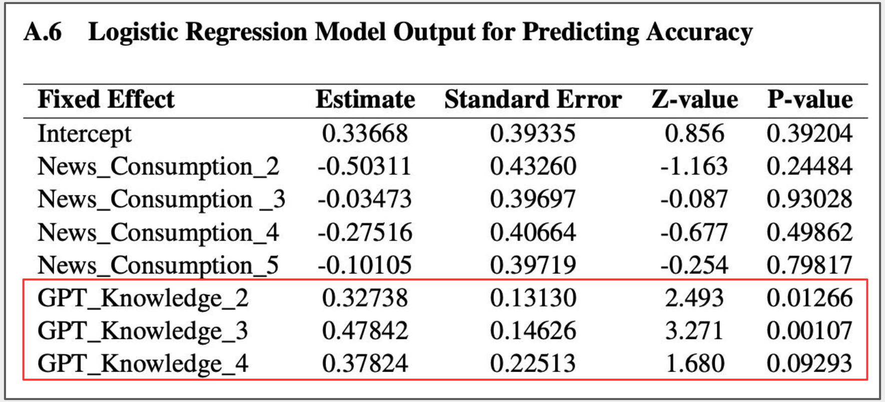
KEY RESULTS

1. **Overall Classification Accuracy: 58.1%**
(Based on 1920 classifications)

2. **None of the 96 articles were exclusively classified correctly / incorrectly**

3. **None of the 120 participants answered correctly on all articles they saw**

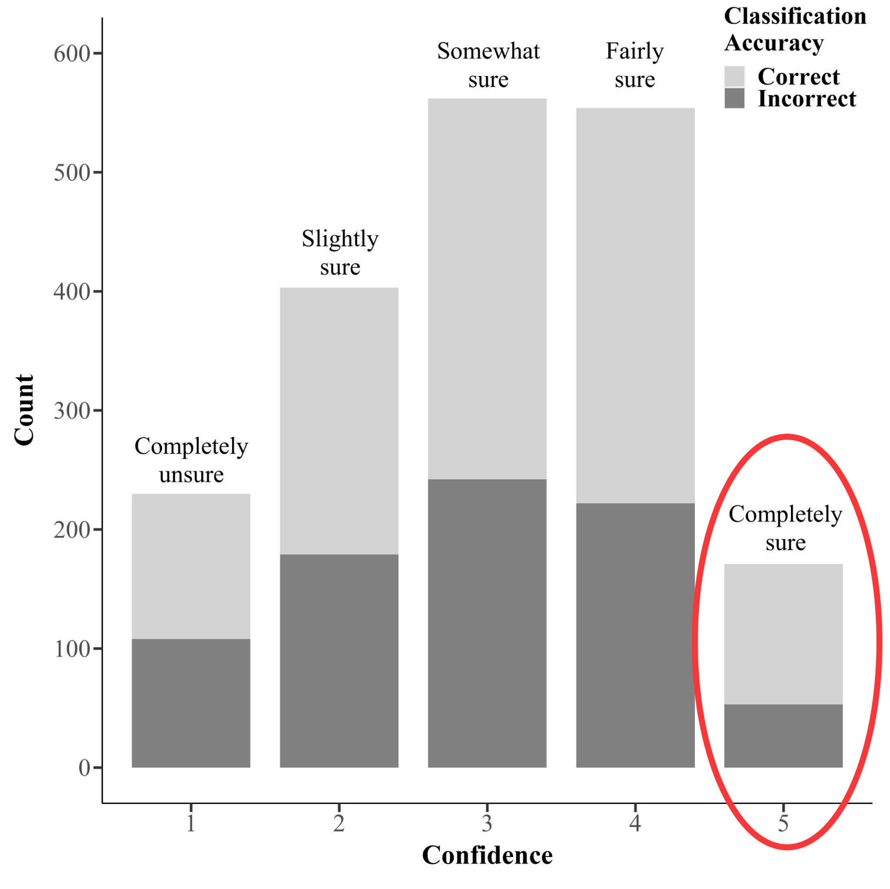
4. **Domain expertise enhanced performance slightly**



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

[A] Results



AARHUS UNIVERSITY

04

Experiment B

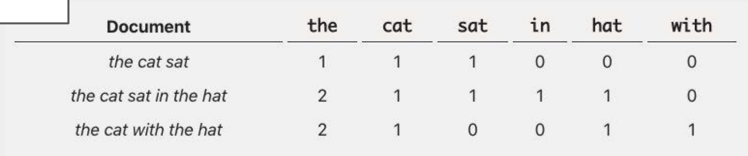
Machine Detection

Can machine classifiers be trained to distinguish between real and synthetic articles?

[B] Constructing Classifiers

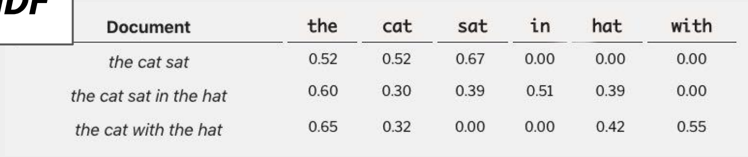
SIMPLE CLASSIFIERS
(Logistic Regression)

BOW



Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1

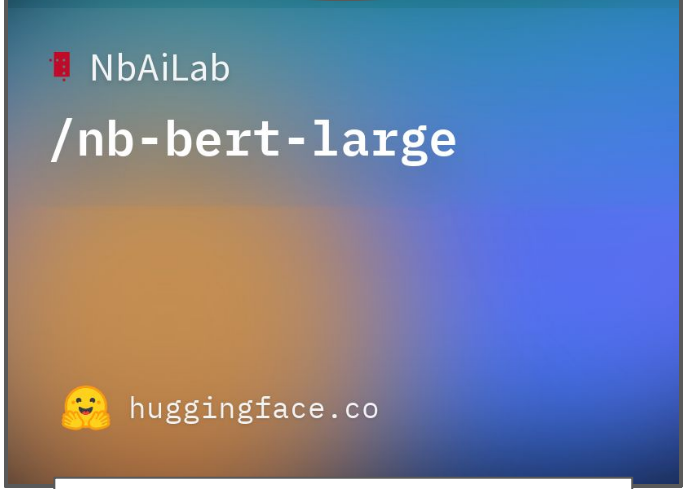
TF-IDF



Document	the	cat	sat	in	hat	with
the cat sat	0.52	0.52	0.67	0.00	0.00	0.00
the cat sat in the hat	0.60	0.30	0.39	0.51	0.39	0.00
the cat with the hat	0.65	0.32	0.00	0.00	0.42	0.55

BOW figure originally by Zhou (2019)

FINE-TUNING BERT
(Norwegian NB-BERT)



Fine-Tuned on on 1218 articles
Training 75%, Validation 25%

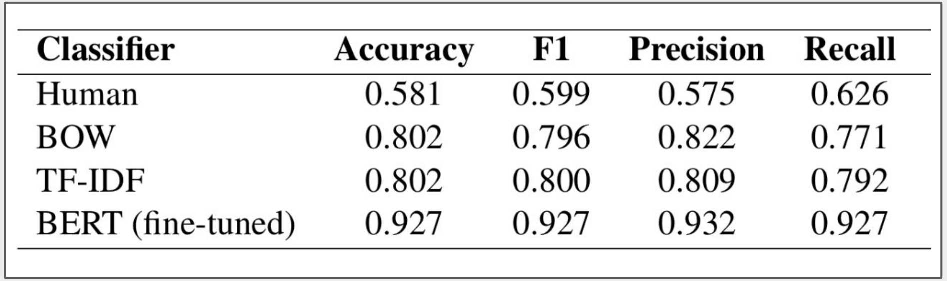
(Half real news articles & other half synthetic GPT-3 articles)

AARHUS UNIVERSITY

[B] Classification Accuracies

TEST DATA?

The same **96** articles for both human [A] and machine detection [B]



Classifier	Accuracy	F1	Precision	Recall
Human	0.581	0.599	0.575	0.626
BOW	0.802	0.796	0.822	0.771
TF-IDF	0.802	0.800	0.809	0.792
BERT (fine-tuned)	0.927	0.927	0.932	0.927

WHAT WAS BEST?

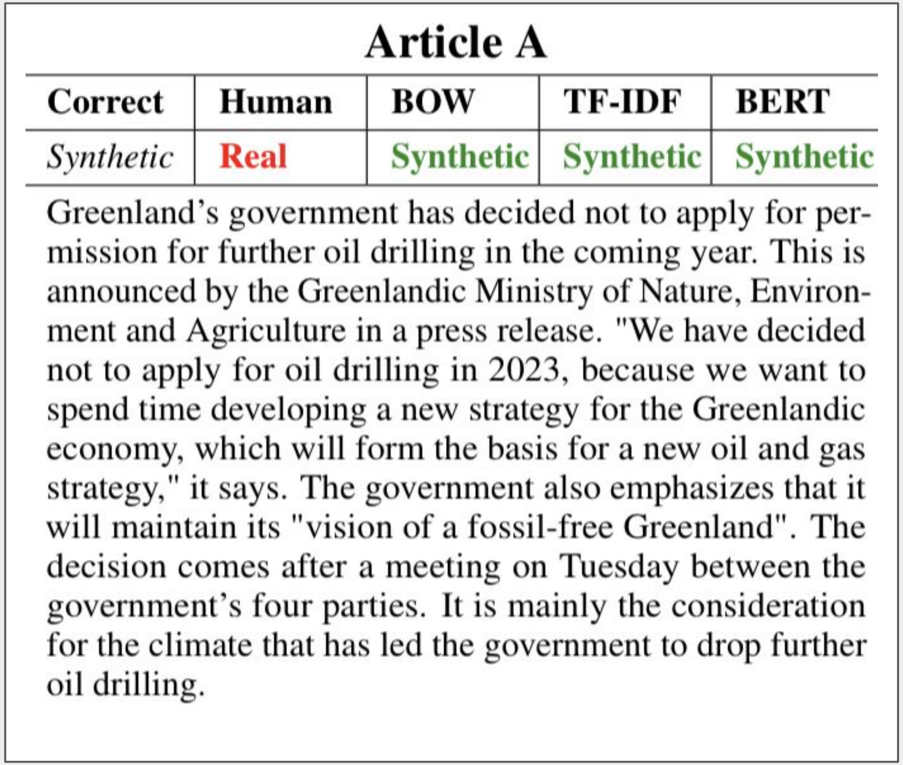
BERT (fine-tuned) with **92.7%** accuracy

But … even the simple baseline models BOW and TF-IDF performed *much better* than humans (**80.2%** accuracy versus **58.1%**)

CONCLUSION?

Machine detection of the fine-tuned GPT-3 model was possible to a great extent !

[B] Classification Accuracies



Article A

Correct	Human	BOW	TF-IDF	BERT
Synthetic	Real	Synthetic	Synthetic	Synthetic

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.

17 out of 20 human participants classified as Real

(Original text in Danish)

05 Limitations

DESIGN CHOICES FAVORING MACHINE DETECTORS

1. Zero temperature sampling → oversamples high-likelihood tokens

2. Humans received *no* training, classifiers were trained on +1000 labelled examples

3. Humans saw real headings and subheadings → evoked familiarity?

GENERALIZABILITY OF THE STUDY

1. Shortened articles → not comparable to a real-world context.

2. Useless generations? Inferring factual information from just a headline requires additional context in reality.

05 Main Takeaways

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.

BUT … the human eye is not all-seeing!

Constructing a machine detector for the same task, reveals that **machine detection of GPT-3 was possible to a great extent**

Likely related to underlying flaws in GPT-3's article generations

Different design could make results more favorable for humans

Questions?

/mina-almasi

@MinaAlmasi

mina.almasi@post.au.dk

/drasbaek

@drasbaek

drasbaek@post.au.dk

References

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.

Mark Chen et al. 2021. Evaluating Large Language Models Trained on Code. ArXiv:2107.03374 [cs].

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.

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.

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

University of Ottowa (n.d.). *The Indo-European Family*. Compendium of Language Management in Canada (CLMC)

Victor Zhou (2019). A Simple Explanation of the Bag-of-Words Model. TowardsDataScience.

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.