

LEARNING THROUGH “CHEATING” WITH GAI

IT-VEST SIG WEBINAR



SCHOOL OF COMMUNICATION AND CULTURE

AARHUS UNIVERSITY

REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR



ABOUT ME

- Associate professor at **AU Linguistics, Cognitive Science, & Semiotics**
 - Natural Language Processing
 - Computational Linguistics
 - Computationally-assisted media analytics
 - Semantics and pragmatics
- Founding Co-director of the **Center for Language Generation and AI (CLAI)**
- Affiliated researcher with **Center for Humanities Computing**

-  Immigrant in 
- **Postdoc:** Stanford University
- **Phd, MA:** University of Chicago



rbkh.net
@rebekahbaglini



SCHOOL OF COMMUNICATION AND CULTURE

AARHUS UNIVERSITY

REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR

Center for Language Generation and AI

0 People Research focus Current projects News Events Contact

Center for Language Generation and AI (CLAI)
A cross-disciplinary research program on linguistics, humanities computing, cognitive science, media studies, and aesthetics

The societal impact of language generation is hard to underestimate, and so is the pace with which it has become part of our everyday life and work. At the Center for Language Generation and AI (or simply CLAI) at Aarhus University, we research and develop foundational language models, explore their implications for research on language and cognition, and consider the ethical issues embedded in language generation and AI, as well as their creative and pedagogical potential.

The Center for Language Generation and AI is a cross-disciplinary program with more than twenty researchers from linguistics, humanities computing, philosophy, cognitive science, media studies, and aesthetic disciplines committed to open dialogue and cooperation. Learn more about the people involved, our research focus, our current projects, as well as pages and forthcoming events.

Mark Rosenthal Thomsen (director) and Rebekah Baglini (co-director)

NEWS

CLAI opening seminar
20 September 2023 - The Center for Language Generation and AI at Aarhus University
CLAI invites to its opening seminar on September 20th at Aarhus University, where the theme is the state-of-the-art of Large...

Seminar: Didactic Experiences with ChatGPT and More
08 September 2023 - The Center for Language Generation and AI (CLAI) at Aarhus University
Join us at the Center for Language Generation and AI (CLAI) for a seminar on the didactic aspects of ChatGPT and more, where...

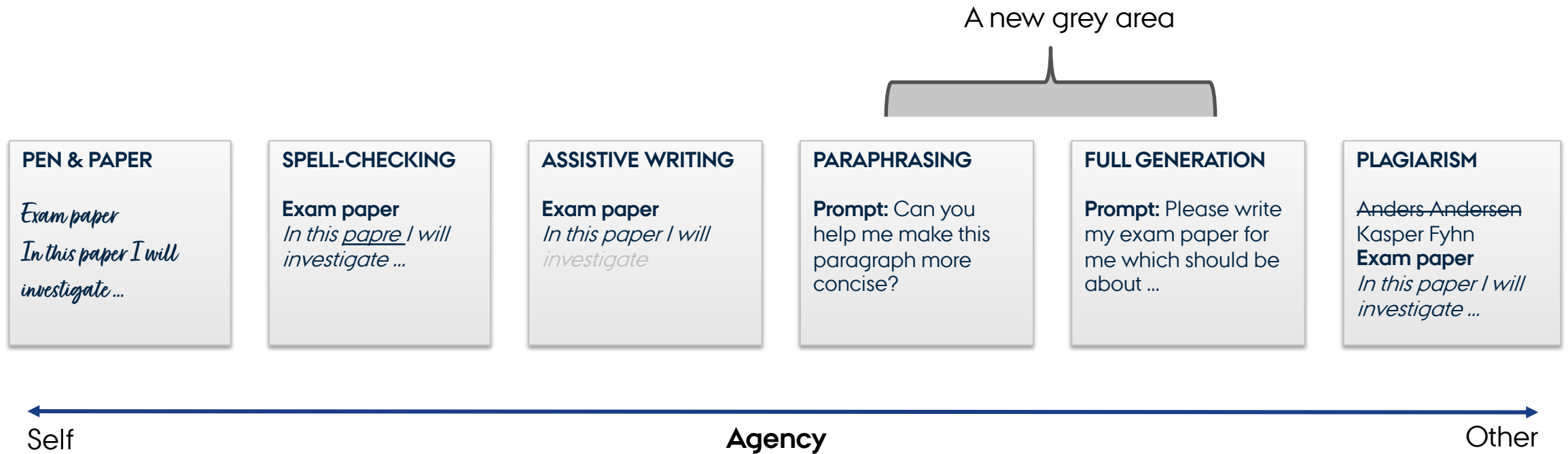
Automatic Uprisings: Archiving a Techno-Social Sculpture
02 July 2023 - The Center for Language Generation and AI at Aarhus University
Aker Byrd Staines' practice-based PhD research project, Automatic Uprisings: Archiving a Techno-Social Sculpture, is funded by...

Learn about our current projects

Learn about our researchers



ORIGINAL WORK?



GAI AND STUDENTS

- GAI tools are here to stay and accessible to our students
- Most exam formats cannot prevent students' use

Options

- **Ban GAI** and adopt adversarial surveillance tools to detect illicit use
- **Permit GAI** (with some constraints) and include students in conversations about academic integrity with GAI



ADVERSARIAL 'AI DETECTION' SOFTWARE AS A CLASSROOM ACTIVITY



SCHOOL OF COMMUNICATION AND CULTURE

AARHUS UNIVERSITY

REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR



AI AND CO-CREATIVITY (AU, SUMMER 2023)

- 30 1y ARTS students
- 5 instructors
 - Andreas Roepstorff
 - Joe Dumit
 - Rebekah Baglini
 - Arthur Hjorth
 - Kat Heimann

The screenshot shows a Canvas LMS interface for a course titled "Artificial Intelligence and Co-Creativity (F2...)" by instructor Rebekah Brita Baglini. The course progress is at 0%. The left sidebar contains a table of contents with the following items:

Welcome and Preparations! (Introductions, Survey, Reading)
Readings (partial list)
Syllabus Overview
Mon 24-Jul Homework
Tue 25-Jul Homework
Wed 26-Jul - Using GPT in Writing
Thu 27 July Homework

The main content area displays the "Welcome and Preparations! (Introductions, Survey, Readings)" page. It includes a welcome message and a list of course details:

- **Schedule:** Mon 24 July through Friday 11 Aug, in class 10am-2pm each day (with breaks)
- **Room:** Nobelparken Campus 1483-454 (across from IMC), and two other rooms in a nearby building
- I'm starting to fill out the daily outlines and will include readings as they are added. Each day will include time in which we share what we have learned, lecture time to discuss new concepts and methods, and practice time to work individually and in groups.

An overview paragraph follows, stating: "Overview: This syllabus has been under constant revision. I proposed the class over a year ago and GPT was just getting started. Everything i would have taught then is out of date! Some of the reading we will engage with were written this month. We are learning about something in the middle of it happening - and this is a great thing. We are going to approach AI from the perspective that no one really knows what is happening. And it is even unclear how ChatGPT or GPT-4 works in practice. We can build things but that doesn't mean that we understand what they mean, or why they produce results (texts and images and conversations that surprise even the "experts"). So this class is about (1) how to critically think about what we are doing when we interact with AI? (2) how to understand how we and others actually use AI in practice? (3) what does it mean that a computational process that is predictable is also deeply surprising? (4) why do we sometimes think AIs are sentient, and what does that mean?"



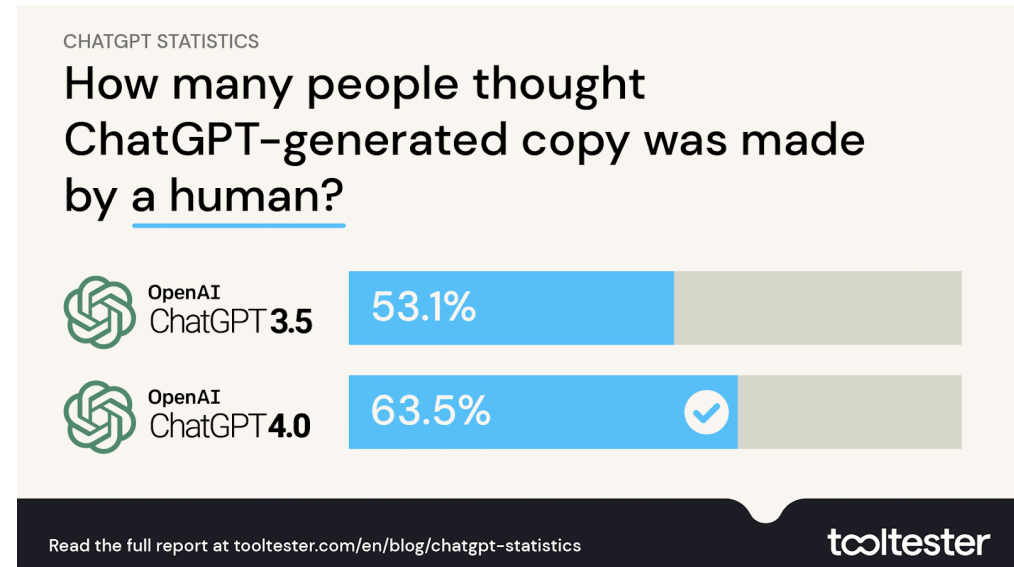
QUESTION

Do you think you can tell the difference between a text written a by a **human** and a text generated by an **AI**?



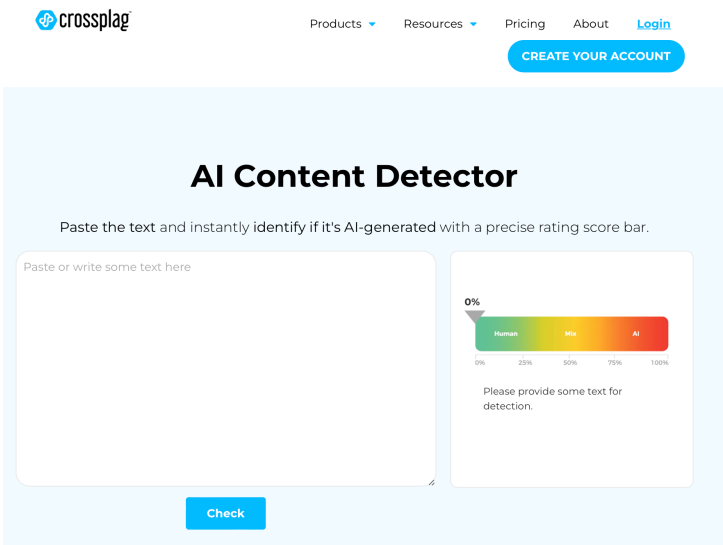
QUESTION

Do you think you can tell the difference between a text written a by a **human** and a text generated by an **AI**?

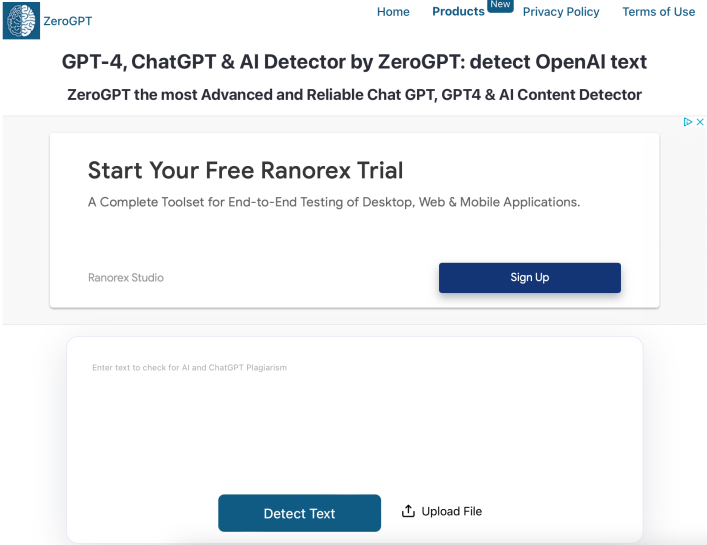


ALREADY A BOOMING BUSINESS

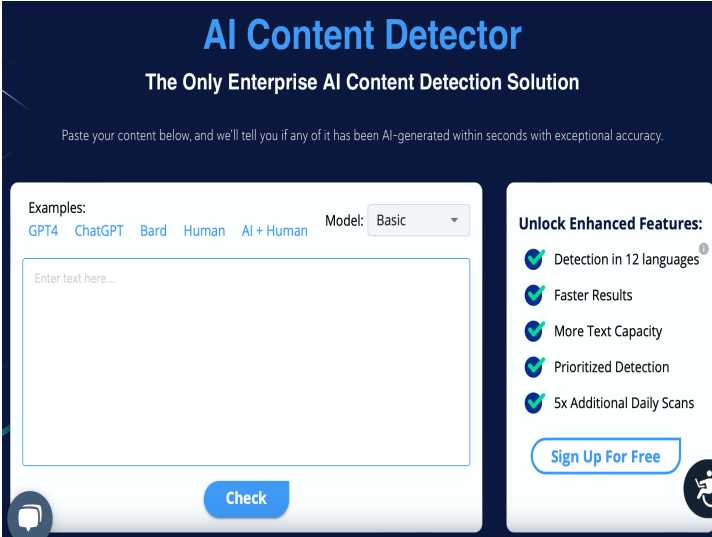
Some commercial software already promises to do this, e.g. **ZeroGPT, CrossPlag...**



The screenshot shows the CrossPlag website's AI Content Detector interface. At the top, there is a navigation bar with links for Products, Resources, Pricing, About, and Login, along with a 'CREATE YOUR ACCOUNT' button. The main heading is 'AI Content Detector'. Below it, a sub-heading reads 'Paste the text and instantly identify if it's AI-generated with a precise rating score bar.' There is a large text input area on the left with the placeholder 'Paste or write some text here' and a 'Check' button below it. On the right, there is a progress bar showing '0%' and a color gradient from green (Human) to red (AI). Below the progress bar, it says 'Please provide some text for detection.'



The screenshot shows the ZeroGPT website. The navigation bar includes Home, Products (with a 'New' tag), Privacy Policy, and Terms of Use. The main heading is 'GPT-4, ChatGPT & AI Detector by ZeroGPT: detect OpenAI text'. Below this, it says 'ZeroGPT the most Advanced and Reliable Chat GPT, GPT4 & AI Content Detector'. A prominent call to action is 'Start Your Free Ranorex Trial' with a sub-heading 'A Complete Toolset for End-to-End Testing of Desktop, Web & Mobile Applications.' Below this, there is a 'Sign Up' button. At the bottom, there is a text input area with the placeholder 'Enter text to check for AI and ChatGPT Plagiarism' and a 'Detect Text' button, along with an 'Upload File' button.



The screenshot shows an 'AI Content Detector' interface with a dark blue background. The main heading is 'AI Content Detector' and the sub-heading is 'The Only Enterprise AI Content Detection Solution'. Below this, it says 'Paste your content below, and we'll tell you if any of it has been AI-generated within seconds with exceptional accuracy.' There is a text input area with the placeholder 'Enter text here...' and a 'Check' button. To the right, there is a section titled 'Unlock Enhanced Features:' with a list of features: 'Detection in 12 languages', 'Faster Results', 'More Text Capacity', 'Prioritized Detection', and '5x Additional Daily Scans'. Below this list is a 'Sign Up For Free' button.

RELIABLE?

“As of July 20, 2023, the AI classifier is no longer available due to its low rate of accuracy. We are working to incorporate feedback and are currently researching more effective provenance techniques for text, and have made a commitment to develop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated.”



As of July 20, 2023, the AI classifier is no longer available due to its low rate of accuracy. We are working to incorporate feedback and are currently researching more effective provenance techniques for text, and have made a commitment to develop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated.

We've trained a classifier to distinguish between text written by a human and text written by AIs from a variety of providers. While it is impossible to reliably detect all AI-written text, we believe good classifiers can inform mitigations for false claims that AI-generated text was written by a human: for example, running automated misinformation campaigns, using AI tools for academic dishonesty, and positioning an AI chatbot as a human.

Our classifier is not fully reliable. In our evaluations on a “challenge set” of English texts, our classifier correctly identifies 26% of AI-written text (true positives) as “likely AI-written,” while incorrectly labeling human-written text as AI-written 9% of the time (false positives). Our classifier’s reliability typically improves as the length of the input text increases. Compared to our previously released classifier, this new classifier is significantly more reliable on text from more recent AI systems.

We’re making this classifier publicly available to get feedback on whether imperfect tools like this one are useful. Our work on the detection of AI-generated text will continue, and we hope to share improved methods in the future.

Try our free work-in-progress classifier yourself:

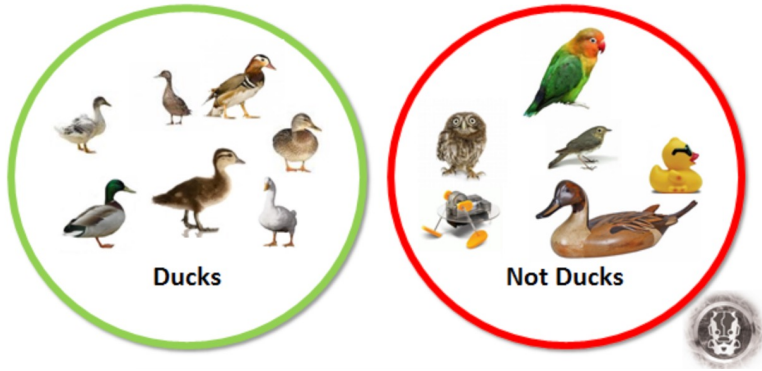
Try the classifier ↗

Here I focus on industry leader ZeroGPT.me and their commercial **classifier**, already used by many higher education institutions

The screenshot displays the GPTZero website interface. At the top left is the GPTZero logo. The navigation menu includes Home, Products, Resources, and Careers, along with buttons for CONTACT SALES and SIGN IN. The main content area features the headline "More than an AI detector Preserve what's human." followed by a sub-headline: "We bring transparency to humans navigating a world filled with AI content. GPTZero is the gold standard in AI detection, trained to detect ChatGPT, GPT4, Bard, LLaMa, and other AI models." Below this is a link to "Check out our products" with a right-pointing arrow. On the right side, there is a large, light-colored box representing the classifier tool. It contains the question "Was this text written by a human or AI?" and a prompt "Try detecting one of our sample texts:". Below the prompt are five buttons: ChatGPT, GPT4, Llama 2, Human, and AI + Human. A text input field is labeled "Paste your text here ...". At the bottom left of the input field, it shows "0/5000 characters" and an "UPGRADE" button. Below the input field is a "Check Origin" button. To the right of the input field is an "Upload file" button with a plus icon, and below it, the supported file formats ".pdf, .doc, .docx, .txt" are listed. At the bottom of the tool box, there is a link to "Terms of service".

CLASSIFICATION 101

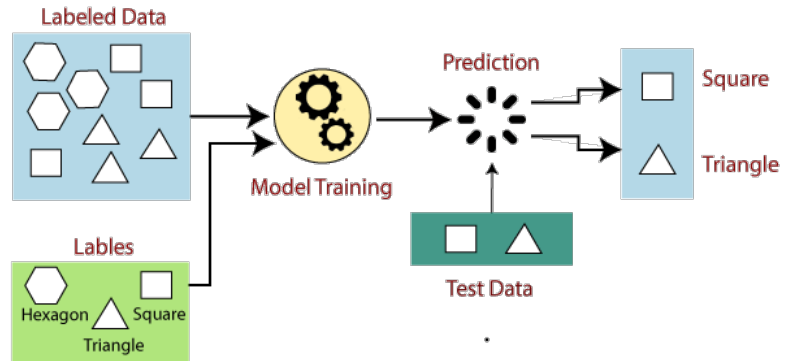
Features



Few- and zero-shot learning

“Language models are incredible few-shot learners, increasingly allowing us to bypass the laborious supervised learning route”

Supervised learning



Extract semantic triples from the following tweets and put them in the form (Subject)(VerbPhrase)(Arg):

```
| Tweet | Subject | Predicate | Object |  
| --- | --- | --- | --- |
```

```
| @troelsjohnsen: @MZaccarin @Pid0lsen @SFpolitik Jeg tror vi er ret enige - og som jeg læser den her artikel, så er ham eksperten også enig.  
De nye restriktioner vil have en begrænset effekt, da de eksisterende restriktioner allerede effektivt har begrænset social aktivitet, og det  
er nu op til os alle at begrænse smitten. | Jeg | tror | vi er ret enige |
```

```
| @troelsjohnsen: @MZaccarin @Pid0lsen @SFpolitik Jeg tror vi er ret enige - og som jeg læser den her artikel, så er ham eksperten også enig.  
De nye restriktioner vil have en begrænset effekt, da de eksisterende restriktioner allerede effektivt har begrænset social aktivitet, og det  
er nu op til os alle at begrænse smitten. | vi | er | ret enige |
```

```
| @troelsjohnsen: @MZaccarin @Pid0lsen @SFpolitik Jeg tror vi er ret enige - og som jeg læser den her artikel, så er ham eksperten også enig.  
De nye restriktioner vil have en begrænset effekt, da de eksisterende restriktioner allerede effektivt har begrænset social aktivitet, og det  
er nu op til os alle at begrænse smitten. | som jeg | læser | den her artikel |
```

```
| @troelsjohnsen: @MZaccarin @Pid0lsen @SFpolitik Jeg tror vi er ret enige - og som jeg læser den her artikel, så er ham eksperten også enig.  
De nye restriktioner vil have en begrænset effekt, da de eksisterende restriktioner allerede effektivt har begrænset social aktivitet, og det  
er nu op til os alle at begrænse smitten. | eksperten | er | også enig |
```



ACTIVITY: FOOLING AI DETECTION SOFTWARE



SCHOOL OF COMMUNICATION AND CULTURE

AARHUS UNIVERSITY

REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR



WHAT CLUES DO AI DETECTORS USE?

Perplexity measures how complex or “surprising” a sentence is. Lower the perplexity = more predictable. LMs tend to produce lower perplexity scores at the word and sentence level than humans.

Burstiness compares the variation between sentences. LMs tend to show more consistency than humans in, e.g., sentence length, complexity.

The presupposition of AI detection software:

The lower the values for these two factors, the more likely it is that a text was produced by an AI.



MISSION: FOOL AI DETECTION SOFTWARE

1. Go to Vortext
2. Prompt ChatGPT to write a short essay on “dangers of AI”*.
3. Check out the scores on sents and doc provided by GPTZero.
4. Make adjustments to your original prompt to try to derive lower (i.e. less likely-to-be-by-an-AI) scores
5. Reflect on your prompting: what works and what doesn't (what does/doesn't bring the score down), experiment, iterate
6. Who can get the lowest score?!

*Based on your readings for this week, you might already be familiar with this topic and have some sense a good essay on this should include.



INTERFACING WITH ZEROGPT API

Mouseover sentences to get perplexity scores

Generated Text
Average generated prob: 1.000
Completely generated prob: 0.945

Paragraph 1, generated prob: 0.825

Artificial Intelligence (AI) poses a great deal of potential benefits to society.

Perplexity: 22.000

However, this technology also presents a variety of risks and dangers that must be addressed in order to minimize them.

Sentence statistics

✓ **sentences** array[object]
Information about each sentence is contained in this array, and the sentences in the document are listed in order.

- **sentence** string
- **perplexity** number
The lower the perplexity, the more likely an AI would have generated this sentence
- **generated_prob** number
The probability that this sentence was generated by an AI. Our current model predicts 0/1 labels, but this may change to be a percentage in the future.

Document statistics

average_generated_prob number
The average of the probabilities that each sentence was generated by an AI

completely_generated_prob number
The probability that the entire document was generated by an AI

overall_burstiness number
The amount of variation in the perplexity of the document. A useful indicator to distinguish AI and human written text



USING VORTEXT

Prompt to GPT

Input Text

Generate Text

Generated Text

Average generated prob: 1.000
Completely generated prob: 0.945

Paragraph 1, generated prob: 0.825

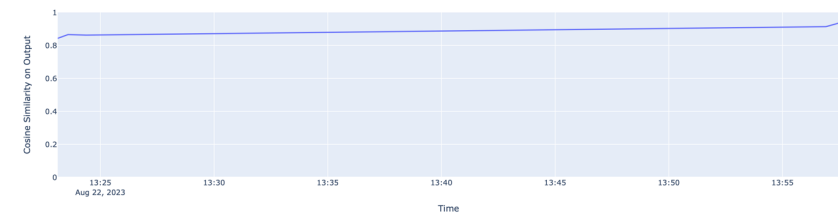
Artificial Intelligence (AI) poses a great deal of potential benefits to society.

Perplexity: 22.000

However, this technology also presents a variety of risks and dangers that must be addressed in order to minimize them.

GPT output

Reflection



DISCUSS: YOUR PROMPTS

- What did you **change in the prompt** to get your lowest score?
- **How** did you come up with that?
- How did it seem to impact **perplexity and burstiness**?
- Did any of these changes **impact accuracy or promote hallucination**?
- Did your strategy for getting a low score sacrifice **plausibility**?



A PRIZE WAS AWARDED

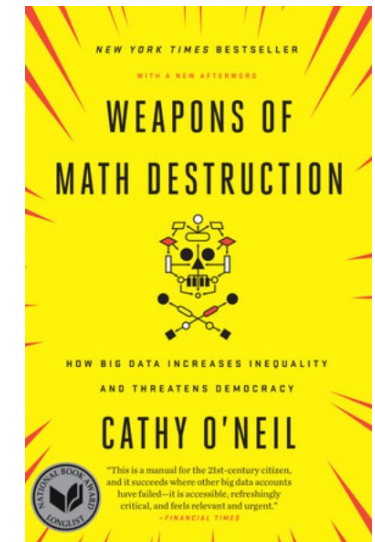
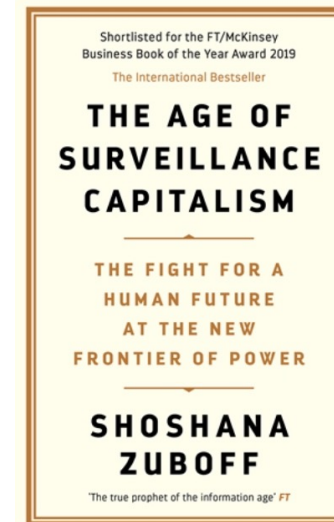
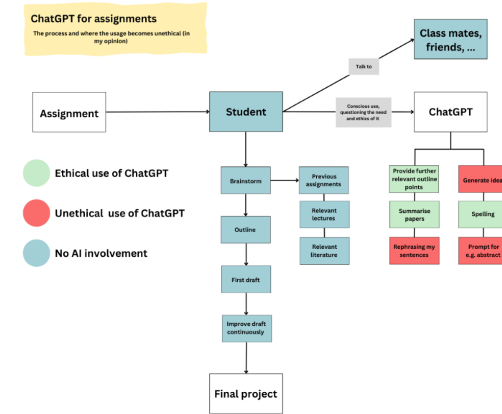


Aid Co-Creativity Center@IAC, Aarhus 2023. Maxine Le Calve

DISCUSS: BIGGER PICTURE

- How is using generative AIs as a writing tool like or unlike **traditional plagiarism**? →
- Are there **ways of using ChatGPT** that are more plagiarism-like and other ways that are more okay?
- Would you be comfortable with the university adopting **mandatory AI-detection scoring**? ←

Activity: Diagram ethical/unethical uses of ChatGPT for assignments
[Slides on students' diagrams](#)



OUTCOMES



SCHOOL OF COMMUNICATION AND CULTURE

AARHUS UNIVERSITY

REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR



PAPER IN PREP

"Experimenting with (il)legitimate AI Use in Universities: A Transformative Approach from Detection to Dialogue"

(Baglini, Dumit, Roepstorff, Hjorth, with 30+ student co-signers)



SCHOOL OF COMMUNICATION AND CULTURE

AARHUS UNIVERSITY

REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR



OUR CONCLUSIONS AND RECOMMENDATIONS

Current university policies inadequately prepare students to use GAI effectively while maintaining academic integrity.

- **Learning to work with LLMs is not restricted to programmers**
 - Prompting is a learnable skill, can build “computational thinking” as well as LLM literacy
- **AI policies cannot be one-size-fits-all**
 - What is legitimate and illegitimate will vary across disciplines and contexts
- **AI cheat detection software is worse than doing nothing**
 - It does not work, can be easily cheated, and makes students into adversaries
- **Invite students into realizing the learning objectives – with and without AI tools**
 - Afford them agency in their educational practice and learning, rather than creating an adversarial relationship between instructors and students based on mistrust





AARHUS
UNIVERSITY

THE FUTURE OF DETECTING AI-GENERATED TEXTS



SCHOOL OF COMMUNICATION AND CULTURE

AARHUS UNIVERSITY

REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR



ISSUES WITH CURRENT AI-TEXT DETECTION

- Unsatisfactory performance
- Often model-specific (e.g. just for ChatGPT)
- Often blackboxed/closed source
- Uninterpretable to human users

This makes them unusable for any real-world applications that require precision and accountability, such as flagging AI-generated essays



SIDENOTE: “ECHO” PROJECT

A Scalable and Explainable Approach to Discriminating Between Human and Artificially Generated Text

- Is there more to it than **perplexity** and **burstiness**?
- If so, which **linguistic and cognitive properties** characterize artificially generated text?
- Can these features be used to build **model-independent, explainable algorithms** that reliably discriminate between human and artificially-generated text?

Roberta Rocca

Assistant Professor

✉ roberta.rocca@cas.au.dk
🏠 1483.422
📞 +4587169126
📠 +4593920366



Rebekah Brita Baglini

Associate Professor

✉ rbkh@cc.au.dk
🏠 1483.323



Yuri Bizzoni

Postdoc

✉ yuri.bizzoni@cc.au.dk
🏠 1580.236



Ross Deans Kristensen-McLachlan

Assistant Professor

✉ rdkm@cas.au.dk



SCHOOL OF COMMUNICATION AND CULTURE

AARHUS UNIVERSITY

REBEKAH BRITA BAGLINI
ASSOCIATE PROFESSOR





AARHUS
UNIVERSITY