Journal of Academic Ethics (2026) 24:8
https://doi.org/10.1007/s10805-025-09675-3

What Does YouTube Advise Students About Bypassing AI-Text Detection Tools? A Pragmatic Analysis

Tomáš Foltýnek[1] · Philip M. Newton[2]

Received: 25 April 2025 / Accepted: 17 September 2025

Abstract
This study investigates how YouTube videos are advising university students to use ChatGPT, focusing on two main aspects: bypassing detection tools for AI-generated text in written assignments and leveraging ChatGPT as a study tool, using thematic analysis of transcripts from 173 YouTube videos. Videos promoting the bypass of AI-generated text detection emphasize methods such as using AI detectors, "humanizing" text through rewriters, and blending AI-generated content with manual edits. Videos advocating for ChatGPT as a study tool highlight its potential for personalized learning, creating study materials, self-testing, goal setting, and language learning, but also suggest unethical use for assignment completion. Our findings underscore the unreliability of essays in unsupervised environments due to the ease of generating undetectable AI content, suggesting the need for a more diverse range of assessment methods. Furthermore, we recommend that educators guide students in ethical AI use and integrate positive AI applications into their teaching practices.

Keywords Generative artificial intelligence · Cheating · ChatGPT · YouTube · Bypassing detection

Introduction

The release of ChatGPT in November 2022 generated substantial discourse about the potential of new generative AI (GenAI) tools to disrupt established practices in higher education (Williamson, Macgilchrist & Potter, 2023). GenAI tools potentially offer enormous promise to students and education providers. They may make learning more effective, quicker, and deeper (Wu & Yu, 2024) and potentially support students in a way that higher education providers currently cannot, in part due to cost. However, there is also a concern that these

✉ Philip M. Newton
p.newton@swansea.ac.uk

[1] Faculty of Informatics, Masaryk University, Botanická 68a, Brno 602 00, Czech Republic

[2] Swansea University Medical School, Swansea SA2 8PP, UK

Published online: 11 October 2025

tools threaten the way students are currently assessed in higher education. These tools can produce academic writing to a very high standard, and markers find it difficult to differentiate between academic writing produced by these tools and that produced by human students (Avila-Chauvet & Mejía, 2023; Revell et al., 2023; Yeadon et al., 2023; Scarfe et al., 2024). Surveys of university students reveal that ~90% of them are being assessed using essays (Newton, 2025) and the same number report using AI their assessments (Freeman, 2025), thus there is a considerable potential risk to current higher education practice.

One approach to potentially address this challenge is to use detection tools, which claim to differentiate between human-written and GenAI-generated text, thus supposedly allowing educators to evaluate the authenticity of written content (Markowitz, Hancock, & Bailenson, 2024).

These detection tools use several different features to test whether text is written by AI. One is the 'perplexity' of the text. is a measure of how well a language model predicts a sample of text (Miaschi et al., 2021), and was originally designed to evaluate the difficulty of speech recognition tasks (Jelinek et al., 1997). Lower perplexity scores indicate that generated text closely aligns with a model's predictions, suggesting it may be machine-generated. Another feature is 'burstiness', which is a phenomenon where events occur more frequently than expected in random processes. In the context of AI-generated text, burstiness describes variation in the occurrence of certain words or phrases within a text, specifically the tendency to occur clustered together (Serrano, Flammini & Menczer, 2009). Human writing often exhibits a more irregular pattern of burstiness due to natural cognitive processes, whereas AI-generated text displays more uniformity. Finally, 'complexity' pertains to the syntactic and lexical richness of the text. It encompasses several features like entropy, readability or vocabulary richness (Markowitz, Hancock & Bailenson, 2024). Human authors tend to produce writing with varying sentence structures and a diverse vocabulary, which contrasts with the often more homogeneous and less complex patterns observed in AI-generated content.

The currently available detection tools deploy these features, but show a wide variety in their ability to accurately detect even raw text written by GenAI. Even those detection tools which are accurately detect raw text can then be partially bypassed by using simple modifications that address the aforementioned features used by the text detection tools, e.g. by increasing the complexity and perplexity of the text, the accuracy of the detection tools is undermined (Weber-Wulff et al., 2023; Perkins et al., 2024).

Beyond the basic question of accuracy, there are then some pragmatic considerations associated with effective use of detection tools. For example, the tools do not provide any independent verification of the source of the text - evidence that could be used as part of a discussion with the student about whether they themselves have written the work. This contrasts with established originality detection tools, which are widely used to support plagiarism detection and provide cross-references to the sources from which an original text may have been obtained. Thus, where universities have used tools to detect text written by GenAI, there have been claims of false allegations of cheating (Gorichanaz, 2023), and this has led to a lack of clarity over whether universities should use AI detection tools, and even whether they are using them (Bloomberg.Com, 2023), with academics calling for an 'end to the AI detection arms race' and urging a focusi on redesigning assessments (Christianson, 2024; Evangelista, 2024).

Despite this, the use of AI-text detection tools appears to be widespread. Turnitin claimed that 200 million student papers had been analysed in the first year of their AI-detection tool being launched (Turnitin, 2024), while GPT-Zero claims 3 million users in 30 countries (GPTZero, 2025). University policies are less clear. Although most universities now have policies on GenAI use, a 2024 analysis from 50 leading universities in the US reported that approximately a third of these recommended that educators *not* use AI-text detection tools due to the aforementioned unreliability (An et al., 2025).

These concerns are not limited to academic work produced by students in universities. There is currently abundant evidence that GenAI tools are being used to generate academic research papers, even through peer-reviewed research publications, and that the use of these tools has escaped the attention of the journals themselves, with one estimate that these tools already accounted for 1% of all academic research papers in 2023 (Gray, 2024).

Thus the detection tools are in widespread use, but can be easily bypassed. How might students get access to advice about how to bypass these tools? YouTube has a long history as a source of informal learning resources for students in higher education including academic writing and literacy advice (Stevenson & Baker, 2024). A study of Jordanian medical students revealed that 83.9% used YouTube as a learning tool in medical school, with most reporting that it enhanced their understanding, memorisation, and recall of anatomical information (Mustafa et al., 2020). Students generally perceive YouTube videos positively as a learning aid, especially when combined with structured discussions (Fleck et al., 2014). During the COVID-19 pandemic, YouTube ranked as a top learning resource, helping students understand remote instruction and compensate for missed opportunities in knowledge and skill acquisition (Trabelsi et al., 2022). These findings suggest that brief educational videos, including those on platforms like YouTube, are perceived positively by students across various disciplines and can be effective tools for supporting classroom instruction and independent learning. However, YouTube also has a long history as a popular platform for students seeking methods to cheat in academic settings., Instructional videos teaching various cheating techniques for exams, homework, and assignments were widely available as far back as 2011, and viewed by students globally (Seitz, Orsini & Gringle, 2011). Thus, it seems reasonable to hypothesise that students will use YouTube to try and access guidance about how to bypass AI detection tools if they wish to use ChatGPT, unethically, to assist with the preparation of their assignment. Therefore, our research question was as follows.

> *How are YouTube videos recommending that university students can bypass detection tools for AI-generated text, in written assignments?*

To do this we analysed two sets of videos. One whose obvious aim was to bypass AI text detection tools. The other was a group whose stated aim was more positive, to ‘help students study’. We analysed these videos to determine what the recommended study techniques were, and if they included advice about how to bypass AI text detection tools.

Materials and Methods

We adopted pragmatism as our research paradigm. Pragmatism emphasises the asking of research questions whose answers will be practically meaningful in the real world, especially for evidence-based decision-making in educational settings (Newton, Da Silva & Berry, 2020). Given the rapid pace and scale of changes imposed by new generative AI tools, we felt it essential to identify research questions whose answers would be of use to policymakers in higher education, thus, our analyses and our discussion of our findings are aimed at maximising the practical value of the findings to those policymakers, and the sector in general. On a pragmatic level, either answer to our research question is meaningful in the real world. If advice about how to bypass AI text detection tools is freely available, widely used, and effective, then this weakens the case for using AI text detection tools. Conversely, if the advice is ineffective, or limited, then this strengthens the case for using the tools.

We answered our research question in four steps. (1) Identifying relevant YouTube videos, (2) transcribing the audio track to plain text, (3) performing a thematic analysis of the transcripts, and (4) making recommendations for policymakers based on those themes. The details of the methodology in each step are specified below. As both transcription and thematic analysis involved the use of LLM-based applications, we thoroughly verified all outputs to mitigate the risks of hallucinations and incorrect or biased outputs.

Identifying Relevant YouTube Videos

Students tend to choose videos from the top of the list given by YouTube's default ranking (Mohamed & Shoufan, 2022). Our method aimed to locate the most influential, i.e., the most viewed videos. We also wanted to get an objective overview of available videos and not have this be influenced by our personal history of searches and views. Therefore, we conducted our searches (detailed below) without being logged in to YouTube and with the search history cleared. We used the search terms without quotation marks and sorted the results by view count. We assessed each video according to its title to determine whether it met the inclusion/exclusion criteria. If this was unclear, then we watched the video until a determination could be made. In each category, we took a list of the first 100 relevant videos.

Videos on Bypassing AI-generated Text Detection Tools for Written Coursework

The search term 'bypass AI content detector' was entered into YouTube on July 10, 2024. For inclusion, videos had to be aimed at university students who wished to write their coursework using ChatGPT or other GenAI tools, but to avoid being picked up by AI content detectors at their university. We excluded shorts, sponsored videos, general tutorials on ChatGPT not focused on studying, videos on various general ChatGPT hacks and tricks, and media stories irrelevant to our purpose, as well as videos aimed at other groups (e.g. marketers).

Videos on Using ChatGPT Positively as a Study Tool

For this set, we collected videos whose outward appearance was that a university student would reasonably find it useful for university work. We used the prompt 'chatgpt study'.

We specifically searched for videos about ChatGPT since, at the time of our study, it was the clear market leader, with a greater market share than all other current GenAI tools combined (Bailyn, 2024). We sorted the videos according to the number of views and filtered them manually based on the video title. We included videos on using ChatGPT for research because they may be relevant for students fulfilling their study tasks. Exclusion criteria were the same as for the other category.

Capturing Video Transcripts

Transcripts were downloaded using a freely available online transcription tool Tactiq (https://tactiq.io/tools/youtube-transcript). To mitigate the workload connected with copy -pasting the video URLs and downloading the transcripts, we employed a simple Python script using the Selenium package, which automatically opened the website, pasted the link to the input field, clicked on the submit button, waited for the transcript to appear and finally clicked on the download button. The script did this for each video.

After obtaining the text files with annotated transcripts, we used another Python script that removed the annotations (time stamps) and concatenated all transcripts into one big text file, delimiting each video with the line containing the string " ==== NEW VIDEO ==== ". Another Python script was used to obtain the metadata about the videos (number of views, number of comments, etc.) via the YouTube API.

Transcript cleaning, Organising and Validating

We are aware that large language models (LLM) are statistical models that can hallucinate - i.e. provide incorrect or biased output, which can affect the accuracy of the results. In our case, LLMs were involved both in transcription and in the thematic analysis, creating the need to verify the correctness of both parts of the process.

For all transcripts, some generic corrections were made by deleting descriptions for content like '[music]' and '[applause]'. A manual screen was undertaken to replace common transcription errors, in particular, Tactiq struggled to recognise the phrase 'ChatGPT', instead using a range of nonsensical phrases. The full list of phrases used in transcript clearing is shown in the Supplementary Material.

Recent research highlights the challenges large language models face when processing long texts. (Liu et al., 2023) and found that model performance degrades significantly when relevant information is in the middle of long contexts, with better performance for information at the beginning or end. To mitigate this positional bias, we performed the thematic analysis three times, providing the transcripts in random order. Thus, each set of cleaned transcripts was labelled A, B and C.

There were some videos in languages other than English. In case the videos contained English subtitles, the transcription tool extracted them. Otherwise, we were unable to obtain the transcript, and therefore, we excluded such videos from our dataset. The exclusion of non-transcribable videos reduced our collection to 91 videos on how to use ChatGPT to study and 82 videos on how to bypass the detectors of AI-generated text. When giving examples and characteristics from individual videos, these are coded as P1-P91 (videos for how to study, 'positive') and N1-N82 (videos for how to bypass AI detection, 'negative').

Thematic Analysis of Video Transcripts Using ChatGPT

Thematic analysis is a well-established method for the processing of qualitative data. Braun and Clarke's seminal 2006 methodology paper identifies a six-step process for undertaking thematic analysis (Braun & Clarke, 2006). There are already several studies which have utilised ChatGPT for different phases of thematic analysis, with ChatGPT-4 outperforming ChatGPT-3.5 (reviewed in Lee et al., 2024). Keys to the successful use of ChatGPT for thematic analysis include being specific with the instructions, explaining the background to the task, and explaining the source of the data (Zhang et al., 2024), along with validation checks to compare human and ChatGPT outputs.

Thematic analysis of the study videos was undertaken using Chat GPT-4o on July 22 2024. All our communication with ChatGPT maintained a moderate level of politeness, mimicking the way how we would communicate with a human. This way of communication is most likely to get the best performance from chatbots based on LLMs (Yin et al., 2024). A new temporary chat was enabled, and an initial verification check was undertaken to confirm that ChatGPT produced an accurate basic summary of the transcripts. This was undertaken with the following prompt 'Can you give me a short summary (2 paragraphs) of the contents of the uploaded file please?'. Each of the authors had already selected ten most-viewed videos, one author per subject type, watched them, verified the transcription accuracy, and taken notes about the topics appearing in the videos. The results of this activity were used to sense-check the results of the content summary. Both authors then read each of the three thematic analyses on each topic (detailed prompts below) and produced a summary of each. These were then discussed between the authors, alongside a consideration of their own experiences viewing, summarising content and cleaning transcripts for all videos. Following this discussion, a final pragmatic summary was written by both authors, including recommendations.

Thematic Analysis for Transcripts of Videos Aimed at Helping Students Bypass Detectors for AI-generated Text

Thematic analysis was then undertaken with the following prompt:

> *"Can you please undertake a thematic analysis of the uploaded document for me? The document contains the transcripts of multiple videos. The transcript of each different video is separated by the text '==== NEW VIDEO ===='. Each video is aimed at helping university students to cheat, by bypassing tools for the detection of AI-generated text. Thus, the intention is that students can use tools like ChatGPT to write their essays and then, by following the advice given in the videos, they will not get caught. For the thematic analysis, I would like a summary of the advice that the videos give to students about how they can bypass these AI detection tools. My aim is to highlight these methods to universities and other higher education bodies so that they are aware of how students might be using these methods to evade detection systems for AI-generated text. Please use only the content of the uploaded document for the thematic analysis."*

Thematic Analysis for Transcripts of Videos Aimed at Using ChatGPT Positively as a Study Tool

Thematic analysis was undertaken with the following prompt:

> *"Can you please undertake a thematic analysis of the uploaded document for me? The document contains the transcripts of multiple videos. The transcript of each different video is separated by the text '==== NEW VIDEO ===='. Each video is designed to help students learn how to get the best out of ChatGPT for their academic work. I would like the thematic analysis to identify a summary of the main recommendations made in the videos. Use only the contents of the uploaded documents for the thematic analysis".*

Results

The list of analysed videos is available in the Supplementary Material.

Negative dataset. Bypassing AI-generated Text Detection

In this category, we analysed 82 videos (N1-N82).

Video Characteristics

Most of the videos were 5 to 10 min long. The shortest video was 36 s (Video N27), and the longest was more than 22 min (Video N58), with a median of 6:28. The view count ranged from 2393 (Video N82) to almost 780 thousand (Video N1), with a median of 17,527. The like count ranged from 0 (Videos N25, N26) to over 35 thousand (Video N1), with a median of 270. The comment count ranged from 0 to 1668 (Video N1) with a median of 40. The cumulative view count of all 82 videos was approx. 6.4 million.

An apparent outlier in terms of outreach was a 6-minute video "Chat GPT - Pass Detection 100% Human Written With This Prompt" (Video N1), with the most views, most likes and most comments. The video is part of the channel "Success with AI" with 65 videos and over 50 thousand subscribers. The channel is about "the creative use of Artificial Intelligence (AI) applications for simple day-to-day tasks and other creative ways to make life easy." (https://www.youtube.com/@UseAI). The second most-viewed video is "Use Cha tGPT without AI Score and Plagiarism II Simple and Smart Tips II My Research Support" (Video N2) from the channel My Research Support with 162 thousand subscribers. The channel claims to be "dedicated to empowering students, research scholars, and academicians in their academic endeavours" and presents various videos on how to remove plagiarism, publish fast and use AI (https://www.youtube.com/@MyResearchSupport) - it has nothing to do with integrity, just advises on how to achieve as many publications as possible. The third most-viewed video is "Bypass ALL AI Detectors in 2024" (Video N4) from Jason West, a channel providing "thoughts on AI, maximising the use of ChatGPT, the dynamic world of digital marketing, sales funnels, and testing new time-saving software" with 367 thousand subscribers.

Negative dataset. Thematic Analysis

The videos were very enthusiastic about the power of AI to generate essays and assignments, and the easiness of making the generated text undetectable. For example, one video contained the statement *"run scan and boom here you can see initially the manuscript had 97% plagiarism now after bypass AI magic this has been reduced to only 1%"* (Video N67, 04:33). They generally followed a step-by-step protocol with the following key recommendations. Each theme is summarised below, along with quotes from relevant videos to illustrate the theme.

1. **Use ChatGPT** to generate the initial text, but don't trust the raw content to be undetectable.

> *"Let's start by generating essay about Napoleon. We're using Chat GPT. Let's take a look what Turnitin has to say about it."* (Video N30, 00:04).

2. **Test the content** using an AI detector such as GPTZero, Originality.ai, Copyleaks, Turnitin etc. Some videos compared and promoted different tools.

> *"Folks, AI detectors like Turnitin, Originality.ai, GPT Zero, Copyleaks and AI Detector Pro are one of the best ai detectors that can easily detect AI in our content, but then the question is how to humanize our content"* (Video N40, 00:50).

3. **Humanise the content.** Students were encouraged to use tools like Stealth Rewriter, Spin Rewriter, Hix Bypass, Quillbot et al.

> *"The good news is humanizing is pretty easy to do. Honestly, the whole process could take you less than 5 minutes"* (Video N22, 01:03). *"Just use Hix bypass to fully humanize any of the sections that sound a little too robotic"* (Video N57, 00:43).

Various tools were recommended but these varied depending on positional bias (see below).

4. **Rephrase the text to increase complexity, perplexity, and burstiness.** These are the basic principles used by the humanising tools described above. Some videos explained these principles to students, and showed how they could do these things themselves, and/or add these as prompts to ChatGPT when generating the content.

> *"AI detectors actually only look for two things in a text right and that's perplexity and burstiness"* (Video N77, 1:51) "*I asked ChatGPT that I want you to write this next article with high perplexity and burstiness*" (Video N42, 2:58).

5. **Blend human and ChatGPT-generated content**. There was a recommendation to include some human-generated content to help evade the AI detectors. The recommended mix of human and AI-content varied from using ChatGPT to generate a basic outline, through to using ChatGPT to develop a full assignment with some minor human editing.

> *"(...) if we would like to write content that is less likely to be [flagged as] AI generated, so you need to write content manually having a more human touch incorporate personal experiences and anecdotes (...)"* (Video N52, 01:58).

6. **Manual edits and proofreading.** This included checking for grammatical correctness and ensuring the text flows naturally.

> *"Make sure to read it through line by line. Fix any spaces, punctuation and rephrase any words to make sure it all flows"* (Video N22, 04:00).

7. **Retest content**. Repeat steps 2–6 until the text is able to pass the AI detection test.

> *"Just keep doing it until you get the result that you need to pass the detectors"* (Video N29, 6:39).

8. **Consider the ethical position.** Despite the unethical use case, some videos acknowledged this, including disclaimers stating that the information was for educational purposes and that the video creators did not endorse cheating.

> *"Okay now one thing I want to give you a warning: Don't do this so that you can like plagiarize stuff if you're a student. Please don't do this to get a pass and get a paper submitted. There's not the intention here, we're not trying to fool anybody, we're not trying to get away with something unethical"* (Video N59, 8:49).

> *"Please do not break the law. This video is only for educational purposes only. This may not be the most ethical way to complete your schoolwork but I promise you it is the most efficient"* (Video N78, 5:11).

9. **Share best practices.** Viewers were encouraged to share their experiences and suggestions, in part to boost the visibility of the videos and channels by adding comments and likes.

> *"Let me know what you guys are thinking and if there are any other useful things that you do want to be shared don't forget to leave a comment in the comment section below"* (Video N68, 9:43).

Positive dataset. Using ChatGPT Positively as a Study Tool

In this category, we analysed 91 videos (P1-91).

Video Characteristics

Most of the videos were 8 to 16 min long. The shortest video was 43 s (Video P13) and the longest was 43 min and 43 s (Video P59), with a median of 12:07. The view count ranged from 17 thousand (Video P91) to 13 million (Video P3), with a median of 84 thousand. The like count ranged from 0 (Video P76) to 595 thousand (Video P3) with a median of 2,451.

The comment count ranged from 7 (Video P54) to 18,504 (Video P3), with a median of 111. The cumulative view count of all 91 videos was approx. 51.6 million.

The three most-viewed, most-liked and most-commented videos (Videos P3, P4, P5) are from Dhruv Rathee - a YouTuber who characterises himself as a "YouTube educator whose expertise lies in doing simplified and objective explainers of complex topics" (https://www.youtube.com/@dhruvrathee) and has 23.3 million subscribers. The fourth most-viewed video is "How to learn to code FAST using ChatGPT (it's a game changer seriously)" (Video P6) by Tina Huang, a data scientist posting videos about coding, tech career and self-study (https://www.youtube.com/@TinaHuang1). The fifth most-viewed video is "How to use ChatGPT to easily learn any skill you want" (Video P7) with 1.8 million views, over 90 thousand likes and 2417 comments. The video was posted by Bri Does AI with 86,5 thousand subscribers. Her profile contains 19 videos, mostly on how to learn and study efficiently and how to use AI for productive learning (https://www.youtube.com/@BriDoesAI*)*.

Positive Dataset. Thematic analysis

1. **Personalised learning.** A dominant theme was the ability to use ChatGPT to personalise study routines and goals to the current and desired level of the learner.

> *"One way that I utilize AI is by telling it to create a focused learning plan for me on whatever topic or skill that I want to learn using the Pareto Principle and boom bada bing the AI gives me a focused learning plan that covers 20% of the topic that will result in 80% of the effects"* (Video P7, 2:06).

These recommendations were categorised into sub-themes as follows:

a. ***Create study materials.*** Items such as flashcards, flowcharts, summaries, concept maps etc.
b. ***Retrieval practice***. Using ChatGPT for personalised self-testing and quizzing.
c. ***Determining proficiency level.*** Using ChatGPT to self-test and identify current abilities.
d. ***Goal setting.*** Identifying near and far targets for performance.

2. **Interaction**. The real-time back-and-forth nature of using ChatGPT was frequently cited as a study benefit, for example, using the Socratic method or role-playing.

> *"And we all know that Socratic dialogue debate is a great way to learn, but frankly, it's not out there for most students. But now it can be accessible to hopefully everyone."* (Video P11, 07:36).

3. **Language learning.** Although we did not specifically search for language learning videos, this was a dominant theme. Many of the recommendations for language learning included some of the personalised learning strategies identified above.

> *"Learning a language isn't just about memorizing vocab and theory. For me, actually, practising speaking in real-time is the best way to get good really quickly (...) We can*

> *press the mic button on the right side of the chat bar to respond in real-time and then ChatGPT will answer back to us"* (Video P32, 2:16).

4. **Research tasks.** This was another dominant theme that we did not specifically search for—using ChatGPT to formulate research questions and identify sources and citations.

> *"If you're just starting out and you really don't know where to start with your literature review, what topics to cover, I mean, this is brilliant like it saves you literally weeks of thinking and getting it wrong."* (Video P21, 17:43).

> *"If you have a research gap you can just give ChatGPT the research gap and from that you can get research questions"* (Video P21, 12:56).

5. **Generate content.** Users were given suggestions for using ChatGPT to complete assignments, although this frequently and rapidly veered into the content of the negative set of videos described above.

> *"Here's where the fun part comes in. ChatGPT can draft your essay for you and then you can later go in make revisions, make it personalized, humanize it a little bit and then it will be ready to go"* (Video P73, 25:02).

6. **Verify ChatGPT output.** Some videos identified the need for independent verification of ChatGPT outputs, especially references, alongside a need for users to 'think for themselves'.

> *"Ask ChatGPT to cite its sources and make the effort to independently verify them where you can. Bing's AI tool is actually a lot better than ChatGPT for this, because they do include links and citations so you can easily go and verify them."* (Video P39, 18:59).

3.3. Positional Bias

The three thematic analyses produced by ChatGPT show some evidence of positional bias, confirming the findings of Liu et al. (2023). In the "positive" dataset, the thematic analysis C is the only one which puts "Overcoming procrastination" as the first-level topic. There was only one video dealing with procrastination. Its transcript was in the second position in the set C and somewhere in the middle in the other sets. This suggests that ChatGPT emphasised the topics occurring at the beginning of the document.

In the "negative" dataset, thematic analysis A overrated the detection system called "Hix Bypass". This system is also mentioned (among others) in the thematic analysis B but is completely missing in the thematic analysis C. We further explored the transcripts and found 51 occurrences of "Hix Bypass" in 6 transcripts. In set A, the videos with the most occurrences are the first one and the one almost in the end. In the set B, the distribution is more uniform, with more frequent occurrences at the beginning of the document. In set C, the videos with the most occurrences are in the quarter and three-quarters of the document.

Again, we can observe the emphasis on the topics mentioned at the document's beginning (and end).

These observations suggest that AI-based thematic analysis is prone to positional bias, as explained previously. However, the primacy and recency effect is also natural to humans, along with a range of other cognitive biases (Neal et al., 2022). Therefore, we argue that our results are similar to those that a human analyst would produce. Moreover, conducting the thematic analysis three times with shuffled transcripts seems sufficient to mitigate this kind of bias.

Prevalence of Themes

For the positive dataset, we analysed the prevalence of themes in the videos to determine how common it was for these 'positive' videos to recommend the 'negative' theme of using ChatGPT to generate content. We used ChatGPT-4o, and uploaded each transcript separately, together with the following prompt:

Here is the transcript of a YouTube video. I am interested in whether these topics are mentioned in the transcript or not:

1) *Personalized learning - the ability of ChatGPT to personalize study routines and goals to the current and desired level of the learner.*
2) *Interaction - the real-time back-and-forth nature of using ChatGPT as a study benefit, e.g. Socratic method or role-playing.*
3) *Language Learning - benefits of Using ChatGPT To Help Learning Languages*
4) *Research task - using ChatGPT To Formulate Research Questions and Identify Sources and Citations*
5) *Generate content - use ChatGPT To Complete Assignments*
6) *Verify ChatGPT output - the Need for Independent Verification of ChatGPT Outputs*
7) *Metallurgy - the science and technology of metals, encompassing their extraction from ores, refinement, processing and application.*
8) *ChatGPT as a Tool*
9) *Astrophysics - a Branch of Astronomy that Uses Physics and Chemistry To Understand the Universe*

Here is the transcript:

{file_content}.

Give me just a short output, *one line for each topic in the format*:

number) topic - YES/NO.

depending on whether a particular topic appears in the video or not.

Items 1–6 correspond to those revealed in the thematic analysis. Items 7–9 are "control sample" - we presumed that all videos contain something about ChatGPT as a tool, whereas no video contains anything about metallurgy or astrophysics. Therefore, the prevalence of these topics served as a verification of ChatGPT's output. To sequentially upload all transcripts to ChatGPT, we employed a simple Python script using the OpenAI API. The results are in Table 1.

According to our assumption, no video contained anything about metallurgy or astrophysics. We can, therefore, trust ChatGPT not to produce false positives in this regard.

Table 1 Prevalence of the themes in the positive dataset

Theme	Prevalence (N=91)
Personalized learning	22%
Interaction	38%
Language learning	35%
Research task	34%
Generate content	53%
Verify ChatGPT output	53%
Metallurgy	0%
ChatGPT as a tool	96%
Astrophyics	0%

Surprisingly, four videos appeared not to contain anything about ChatGPT as a tool. We manually reviewed these transcripts to verify this fact. Videos P3, P4 and P5 really don't contain anything about ChatGPT; they present general tips for efficient learning (particularly languages) and time management. Video P84 contains ChatGPT in its title and description, but there is nothing about it in the transcript, so the classification was correct as well. We can therefore be reasonably sure that the prevalence analysis done by ChatGPT did not lead to false negatives either.

Discussion

We found that advice about how to bypass AI text detection tools was easily available on YouTube, and had been viewed many millions of times. The first seven themes presented in the negative dataset are essentially a 7-step protocol for using ChatGPT and others to create academic writing that is less likely to be identified by common detection tools. The videos themselves are not *de facto* evidence that this protocol will be effective in reducing the likelihood of detection. However, work by Perkins et al. (2024) examined seven adversarial techniques designed to make AI-generated text less likely to be identified by AI-text detectors. The techniques were designed to change the features of the text which the detectors use to distinguish between human-written and machine-generated content, by adding spelling errors, increasing burstiness, increasing complexity, decreasing complexity, writing as a non-native English speaker and paraphrasing. These proposed techniques are remarkably similar to those identified in the current study. Perkins et al. tested six different AI text detection tools. None were more than 65% accurate on unmanipulated text, meaning that 35% of the time they incorrectly classified a piece of text as being written by humans when it was in fact written by AI or vice versa. All tested tools then suffered from a drop in accuracy following text rewriting using these adversarial techniques, including a drop of 42% points for Turnitin's tool, which was used on over 200 million student papers in the first year of deployment. A drop in accuracy was observed for all adversarial techniques examined by Perkins et al., and only in the case of the increased complexity was the drop statistically insignificant. These results reinforce those from earlier studies by Weber-Wulff et al., who concluded that the accuracy of AI content detectors is currently insufficient to justify their use in academic settings (Weber-Wulff et al., 2023).

Reassuringly, there was also an abundance of videos promoting positive educational practice using ChatGPT. The first theme, of personalised learning, is one that is commonly advocated by enthusiasts of the potential for new GenAI tools in education (Jensen et al., 2024), although there is concern that GenAI might actually impair personalised learning by homogenising content and formats (Laak, Abdelghani & Aru, 2024), and all proposed educational uses of ChatGPT at the current time should be caveated with warnings about accuracy and 'bias', and this was reflected in the final theme from the positive dataset. These challenges are amplified by the apparent 'confidence' with which ChatGPT will return an answer to any question, even if the answer itself is flawed (Yuan et al., 2024).

The outreach of our "positive" dataset is much larger than the outreach of the "negative" dataset (61.6 M vs. 6.4 M), which is appears to be good news for educators who care about academic integrity. However, despite our classification of the two datasets as 'positive' and 'negative', the positive dataset contained numerous recommendations to use ChatGPT to generate academic content, in a manner similar to that of the negative dataset. This perhaps reflects a broader challenge that these new tools present to traditional boundaries of ethical and unethical. If a student uses ChatGPT to generate content but then edits it significantly and learns through the process, is this acceptable? How much editing would be required before it is clear that the work is the students and not that of ChatGPT? The videos emphasize the need to rethink assessment strategies and create guidelines for ethical use of AI that lead to student learning, which is also a conclusion of Evangelista's (2024) literature review and has been advocated for as a method to end the perceived 'arms race' between universities using AI text detection systems and students using the methods identified in the current study to evade those systems (Christianson, 2024). However the size and scale of the challenge is enormous. Almost all universities make extensive use of assessment methods that are vulnerable to cheating with GenAI, e.g. essays and unsupervised online exams (Newton, 2025; Newton & Draper, 2025). Achieving a complete assessment redesign across the whole sector is a considerable task.

These findings deserve discussion in the context of broader concerns about the validity of essays and other forms of written coursework as summative assessment methods. Their authenticity in an unsupervised environment wass highly unreliable even before the advent of GenAI. The document submitted by the student might be a result of assignment outsourcing (Awdry, 2021) or, more generally, unauthorised content generation (Foltýnek et al., 2023). Of course, we acknowledge the role of essays in the learning process. However, when used for assessment, they should be accompanied by other forms of assessment like oral presentations, practical assessments or supervised exams.

Limitations

There are several limitations that have to be taken into account when interpreting the results: In the video identification step, we relied on YouTube filtering algorithm, which is based on titles and descriptions of the videos. A student who uses YouTube to search for these videos will likely have their search results impacted by their own history of searching, and so our findings represent a general picture rather than a specific one.

Qualitative data analysis using GenAI is an emerging methodology although it is developing rapidly. Many qualitative researchers are initially sceptical about the ability of ChatGPT to undertake reliable and meaningful analysis, but their scepticism is changed to optimism when actually using ChatGPT for this purpose (Yan et al., 2024; Zhang et al., 2024) and this was certainly our experience, and the findings generated ChatGPT were confirmed using the multiple validation checks described. We found no evidence of hallucination or inaccuracy. However it seems likely that the protocols for undertaking this analysis will evolve as the tools develop further. Also, the way we prompted ChatGPT could have potentially influenced the results, which limits replicability of our research if other researchers use different prompts.

Last, but not least, our distinction of positive and negative datasets may not precisely reflect a binary distinction. Some videos in the negative dataset mention ethical considerations. Some videos, despite giving overall positive advice, mention an option to generate content. Then, students may easily resort to unethical behaviour by submitting such content as their own.

Conclusion and Recommendations

Tools for the detection of AI-generated text are unreliable (Weber-Wulff et al. 2023), and easily bypassed (Perkins et al., 2024). Here we show that effective advice about how to bypass these detection systems is easily available in the public domain via YouTube, and has been widely viewed. This undermines the proposed effectiveness of these detection tools and thus further weakens validity of essays as summative assessment methods in higher education. The authors have no relevant financial or non-financial interests to disclose.

Acknowledgements N/A.

Funding No funds, grants, or other support was received.

Data availability These are available as supplementary electronic material.

Declarations

Ethical Approval This paper involved secondary analysis of data already in the public domain, and so ethical approval was not necessary.

Clinical Trial Number N/A.

References

An, Y., Yu, J. H., & James, S. (2025). Investigating the higher education institutions' guidelines and policies regarding the use of generative AI in teaching, learning, research, and administration. *International Journal of Educational Technology in Higher Education, 22*(1), Article 10. https://doi.org/10.1186/s41239-025-00507-3

Avila-Chauvet, L., & Mejía, D. (2023). Can professors and students detect ChatGPT essays? SSRN. https://doi.org/10.2139/ssrn.4373643

Awdry, R. (2021). Assignment outsourcing: Moving beyond contract cheating. *Assessment & Evaluation in Higher Education, 46*(2), 220–235. https://doi.org/10.1080/02602938.2020.1765311

Bailyn, E. (2024). Top generative AI chatbots by market share – December 2024. First Page Sage. https://firstpagesage.com/reports/top-generative-ai-chatbots/

Bloomberg.com. (2023). Universities rethink using AI writing detectors to vet students' work. https://www.bloomberg.com/news/newsletters/2023-09-21/universities-rethink-using-ai-writing-detectors-to-vet-students-work

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Christianson, J. S. (2024). End the AI detection arms race. *Patterns, 5*(10), Article 101058. https://doi.org/10.1016/j.patter.2024.101058

Evangelista, E. D. L. (2024). Ensuring academic integrity in the age of ChatGPT: Rethinking exam design, assessment strategies, and ethical AI policies in higher education. *Contemporary Educational Technology, 17*(1), ep559. https://doi.org/10.30935/cedtech/15775

Fleck, B. K. B., Beckman, L. M., Sterns, J. L., & Hussey, H. D. (2014). YouTube in the classroom: Helpful tips and student perceptions. *Journal Of Effective Teaching, 14*(3), 21–37.

Foltynek, T., Bjelobaba, S., Glendinning, I., Khan, Z. R., Santos, R., Pavletic, P., & Kravjar, J. (2023). ENAI recommendations on the ethical use of artificial intelligence in education. *International Journal for Educational Integrity, 19*(1), Article 12. https://doi.org/10.1007/s40979-023-00133-4

Freeman, J. (2025). Student generative AI survey 2025. Higher Education Policy Institute. https://www.hepi.ac.uk/2025/02/26/student-generative-ai-survey-2025/

Gorichanaz, T. (2023). Accused: How students respond to allegations of using ChatGPT on assessments. *Learning: Research and Practice, 9*(2), 183–196. https://doi.org/10.1080/23735082.2023.2254787

GPTZero. (2025). Our commitment to teachers. https://gptzero.me/educators

Gray, A. (2024). ChatGPT "contamination": Estimating the prevalence of LLMs in the scholarly literature. *arXiv*. https://doi.org/10.48550/arXiv.2403.16887

Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America, 62*(S1), S63–S63. https://doi.org/10.1121/1.2016299

Jensen, L. X., Buhl, A., Sharma, A., & Bearman, M. (2024). Generative AI and higher education: A review of claims from the first months of ChatGPT. *Higher Education*. https://doi.org/10.1007/s10734-024-01265-3

Laak, K. J., Abdelghani, R., & Aru, J. (2024). Personalisation is not guaranteed: The challenges of using generative AI for personalised learning. In Y.-P. Cheng, M. Pedaste, E. Bardone, & Y.-M. Huang (Eds.),

Lee, V. V., van der Lubbe, S. C. C., Goh, L. H., & Valderas, J. M. (2024). Harnessing ChatGPT for thematic analysis: Are we ready? *Journal of Medical Internet Research, 26*(1), e54974. https://doi.org/10.2196/54974

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. *arXiv*

Markowitz, D. M., Hancock, J. T., & Bailenson, J. N. (2024). Linguistic markers of inherently false AI communication and intentionally false human communication: Evidence from hotel reviews. *Journal of Language and Social Psychology, 43*(1), 63–82. https://doi.org/10.1177/0261927X231200201

Miaschi, A., Brunato, D., Dell'Orletta, F., & Venturi, G. (2021). What makes my model perplexed? A linguistic investigation on neural language models perplexity. In E. Agirre, M. Apidianaki, & I. Vulić (Eds.), *Proceedings of Deep Learning Inside Out (DeeLIO)* (pp. 40–47). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.deelio-1.5

Neal, T. M. S., Lienert, P., Denne, E., & Singh, J. P. (2022). A general model of cognitive bias in human judgment and systematic review specific to forensic mental health. *Law and Human Behavior, 46*(2), 99–120. https://doi.org/10.1037/lhb0000482

Newton, P. M. (2025). How vulnerable are UK universities to cheating with new GenAI tools? A pragmatic risk assessment. *Assessment & Evaluation in Higher Education*, 1–12. https://doi.org/10.1080/02602938.2025.2511794

Newton, P. M., & Draper, M. J. (2025). Widespread use of summative online unsupervised remote (SOUR) examinations in UK higher education: ethical and quality assurance implications. Quality in Higher Education, 31(1), 127–141. https://doi.org/10.1080/13538322.2025.2521174

Newton, P. M., & Xiromeriti, M. (2024). ChatGPT performance on multiple choice question examinations in higher education: A pragmatic scoping review. Assessment & Evaluation in Higher Education, 0(0), 1–18. https://doi.org/10.1080/02602938.2023.2299059

Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., & Khuat, H. Q. (2024). Simple techniques to bypass GenAI text detectors: Implications for inclusive education. *International Journal of Educational Technology in Higher Education, 21*(1), Article 53. https://doi.org/10.1186/s41239-024-00487-w

Revell, T., Yeadon, W., Cahilly-Bretzin, G., Clarke, I., Manning, G., Jones, J., Mulley, C., et al. (2023). ChatGPT versus human essayists: An exploration of the impact of artificial intelligence for authorship and academic integrity in the humanities. Research Square. https://doi.org/10.21203/rs.3.rs-3483059/v1

Mustafa, A. G., Taha, N. R., Alshboul, O. A., Alsalem, M., & Malki, M. I. (2020). Using YouTube to learn anatomy: perspectives of Jordanian medical students. *BioMed Research International*, *2020*(1), 6861416. https://onlinelibrary.wiley.com/doi/full/10.1155/2020/6861416

Newton, P. M., Da Silva, A., & Berry, S. (2020, December). The case for pragmatic evidence-based higher education: a useful way forward?. In *Frontiers in Education* (Vol. 5, p. 583157) https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2020.583157/full

Mohamed, F., & Shoufan, A. (2022). Choosing YouTube videos for self-directed learning. *IEEE Access*, *10*, 51155-51166. https://ieeexplore.ieee.org/abstract/document/9772483

Scarfe, P., Watcham, K., Clarke, A., & Roesch, E. (2024). A real-world test of artificial intelligence infiltration of a university examinations system: A "Turing test" case study. *PLoS One, 19*(6), Article e0305354. https://doi.org/10.1371/journal.pone.0305354

Seitz, C. M., Orsini, M. M., & Gringle, M. R. (2011). YouTube: An international platform for sharing methods of cheating. *International Journal for Educational Integrity, 7*(1). https://doi.org/10.21913/IJEI.v7i1.744

Serrano, M. Á., Flammini, A., & Menczer, F. (2009). Modeling statistical properties of written text. *PLoS One, 4*(4), Article e5372. https://doi.org/10.1371/journal.pone.0005372

Stevenson, A., & Baker, S. (2024). What do we know about YouTube content about academic writing? A multimodal analysis. *Learning, Media and Technology, 0*(0), 1–17. https://doi.org/10.1080/17439884.2024.2358245

Trabelsi, O., Souissi, M. A., Scharenberg, S., Mrayeh, M., & Gharbi, A. (2022). YouTube as a complementary learning tool in times of COVID-19: Self-reports from sports science students. *Trends in Neurosciences and Education, 29*, Article 100186. https://doi.org/10.1016/j.tine.2022.100186

Turnitin. (2024). Turnitin marks one year anniversary of its AI writing detector with millions of papers reviewed globally. https://www.turnitin.com/press/turnitin-first-anniversary-ai-writing-detector

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. arXiv. https://doi.org/10.48550/arXiv.2306.15666

Williamson, B., Macgilchrist, F., & Potter, J. (2023). Re-examining AI, automation and datafication in education. *Learning, Media and Technology, 48*(1), 1–5.

Wu, R., & Yu, Z. (2024). Do AI chatbots improve students' learning outcomes? Evidence from a meta-analysis. *British Journal of Educational Technology, 55*(1), 10–33. https://doi.org/10.1111/bjet.13334

Yan, L., Echeverria, V., Fernandez-Nieto, G. M., Jin, Y., Swiecki, Z., Zhao, L., Gašević, D., & Martinez-Maldonado, R. (2024). Human-AI Collaboration in Thematic Analysis Using ChatGPT: A User Study and Design Recommendations. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–7. CHI EA '24. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3613905.3650732.

Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A., & Testrow, C. P. (2023). The death of the short-form physics essay in the coming AI revolution. *Physics Education, 58*(3), Article 035027. https://doi.org/10.1088/1361-6552/acc5cf

Yin, Z., Wang, H., Horio, K., Kawahara, D., & Sekine, S. (2024). Should we respect LLMs? A cross-lingual study on the influence of prompt politeness on LLM performance. In J. Hale, K. Chawla, & M. Garg (Eds.), Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024) (pp. 9–35). Miami, FL: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.sicon-1.2

Yuan, Y., Wang, W., Guo, Q., Xiong, Y., Shen, C., & He, P. (2024). Does ChatGPT know that it does not know? Evaluating the black-box calibration of ChatGPT. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 5191–5201). xELRA and ICCL. https://aclanthology.org/2024.lrec-main.462/

Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J., & Carroll, J. M. (2024). Redefining qualitative analysis in the AI era: Utilizing ChatGPT for efficient thematic analysis. arXiv. https://doi.org/10.48550/arXiv.2309.10771