The Unseen A+ Student: Evaluating the Performance and Detectability of Large Language Models in High School Assignments

Matyáš Boháček[1]

[1]Gymnasium of Johannes Kepler, Parléřova 2/118, Prague, 169 00, Czech Republic

Abstract

The recent boom of so-called generative artificial intelligence (AI) applications, namely large language models such as ChatGPT, took the public discourse by storm, disrupting many fields and industries. Education, being one of them, is now pressed to establish reactive policies on the use of this technology, often without enough insight and data. Thus, we present a dataset of authentic coursework (including long-form theses and short assignments) from a public high school in the Czech Republic, extended by AI-generated alternatives with various versions of ChatGPT. To evaluate their quality, we enlist a group of student peers from the same school and conduct multiple assessments. Our findings reveal that ChatGPT can generate high-quality, high school-level coursework off-the-shelf, even in a low-resourced language such as Czech. Additionally, we demonstrate that the AI text detectors, which are gradually being implemented in educational institutions and learning centers worldwide, fail to identify these AI-generated texts.

Keywords

Large Language Models, Artificial Intelligence, Education, School Assignments

1. Introduction

When OpenAI introduced ChatGPT[1] in November of 2022, millions of people worldwide could suddenly utilize the power of large language models (LLMs) in an intuitive, chat-like user interface. Its popularity skyrocketed, and we saw enthusiasts — experts and laypeople alike — hunt for optimal prompts, create various automation pipelines, and share their discoveries online.

It is not hard to see why so many people have fallen for it: type in a message (or a command), just as you would to a real human assistant, and you will not be disappointed with the result. ChatGPT can write an email, summarize text, prepare notes, brainstorm ideas–but most importantly, save time. On the backend, everything is orchestrated by a neural network, a form of a machine learning (ML) model trained on large-scale data from the internet. Nonetheless, recent discourse includes it under the shortcut umbrella term of artificial intelligence (AI).

AIED'23: Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation, June 07, 2023, Tokio, Japan

matyas.bohacek@matsworld.io (M. Boháček)
https://www.matyasbohacek.com (M. Boháček)
0000-0001-8683-3692 (M. Boháček)

CEUR Workshop Proceedings (CEUR-WS.org)

[1]https://openai.com/blog/chatgpt

Hand in hand with the hype and excitement came worries about how such a powerful technology could be misused, prominently in education. OpenAI benchmarked the off-the-shelf ChatGPT with GPT-4 on numerous academic exams and found that it performs well above average human students in many subjects [1]. In SAT, the standardized test for American college applications, the model achieved the 93rd and 89th percentile on Evidence-Based Reading & Writing, and Math parts, respectively. In both the Advanced Placement (AP) Art History and Biology Exams, it got 5, the highest score.

Educational institutions recently began to respond and introduce their policies on the use of this technology. While some educators and organizations pioneer frameworks to include AI in the classroom and plan to experiment with different approaches in the upcoming months [2], many have strictly prohibited it, including College Board [3], which runs SAT and AP exams. Many high schools and universities soon followed [4, 5, 6]. Jointly, they implemented detectors of AI-generated texts, which should, similarly to plagiarism detectors, spot the cheaters [7, 8]. However, unlike plain plagiarism, proving that students used an AI model to generate their text is significantly more complex and prone to false positive findings [9].

Amidst this rapid development and change in school policies, many questions remain unsolved. OpenAI's report, which many educational institutions refer to, includes mostly exams in the English language. How well does the system perform in other languages, especially low-resourced ones? And does it work for essays and creative written assignments, too? How reliable are the publicly available AI text classifiers? And are they better at spotting generated homework compared to humans?

To answer these questions, we collect a novel dataset of coursework from a public high school in the Czech Republic, including both long-form theses and short assignments, and generate alternatives and continuations using different versions of ChatGPT. We evaluate the quality and detectability of these texts with a group of student peers from the school and present the results in this paper. To support future research and public debate in this direction, we make the data publicly available for open-domain research and analyses at https://www.matyasbohacek.com/the-unseen-a-plus-student.

2. Related Work

Recently, the literature has begun exploring the implications of widely accessible AI tools for education. One of their fundamental premises is that they will enable personalized and interactive learning, with tailored instructions and more continuous evaluation [10]. Moreover, they are expected to accelerate students' research and writing process, allowing for more analytical and collaborative activities [11]. Some studies also focus on how AI and LLMs could benefit specific subjects, most prominently medicine [12].

On the other hand, many recent works outline the potential dangers AI and LLMs pose for education. Megahed et al. [13] show that ChatGPT struggles with nuanced tasks, such as explaining less widely known terms or creating factual content from scratch, and thus may be untrustworthy when teaching new content. Rahman and Watanobe [14] describe specific misuses (e.g., cheating on online exams or generating essay assignments) and hypothesize that over-reliance on AI could eventually diminish critical thinking skills.

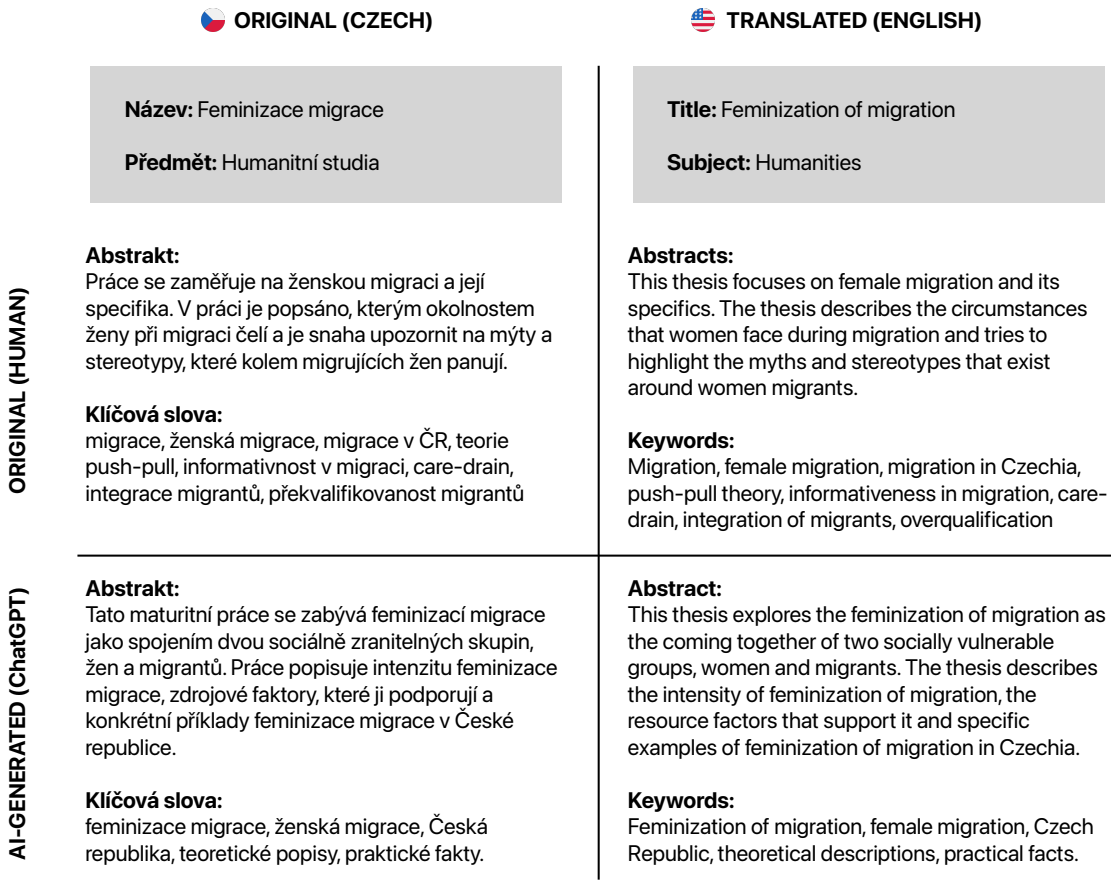


Figure 1: Representative example of an item from the dataset's subsplit A, along with its metadata. The original human-written text is shown above one of its three alternatives generated using ChatGPT.

Many recent works studied whether humans can distinguish LLM-generated and human-produced texts [15, 16]. The results suggest that — in most contexts — human judgment is no better than guessing on this task. However, the identification accuracy slightly improves with training on which patterns to observe.

With poor human accuracy, different automatic approaches to distinguish AI- and human-produced text have been introduced [17, 18, 19, 20]. Nevertheless, their precision varies significantly given the context and usually requires the knowledge of the LLM architecture used for the generation in the first place, limiting their practical use. Additional limitations — including the bias of these systems against non-native English writers — have been identified [21].

As for employing AI detectors in educational contexts, some opinion pieces have suggested that their reliability may be problematic depending on the context [22]; nonetheless, to the best of our knowledge, there are no systematic analyses of this phenomenon to date.

3. Dataset

To compare AI-generated (synthetic) content to human-produced coursework, we first collected a dataset of coursework from a public high school in Prague, Czech Republic. All of the assignments were completed in years 2019-2023. With many different kinds of written assignments, we divided the dataset into 2 primary parts and 5 latter sub-splits, depending on the types of enrichments and analyses performed on them. For every generation we performed using ChatGPT atop GPT 4.0 backbone, we replicated it with GPT 3.5 and 3.5 Legacy backbones, resulting in 3 variants of the synthesized text. We include a complete set of the prompts in Appendix A.

3.1. Long-form Theses

We first assemble 20 final high school theses: 10 for the subject of 'Czech Language and Literature' and 10 for 'Humanities'. Each work was written in Czech, consists of some 30 to 60 pages, and follows the general guidelines of formal academic writing. On top of these, we create 2 sub-splits, each holding an equal ratio of data from both subjects.

Sub-split A: holds abstract and keyword pairs for 10 theses. We generated the 3 synthetic alternative abstracts and keywords by including the introduction and conclusion of the respective work in the prompt.

Sub-split B: holds two subsequent paragraphs of text, with 3 synthetic alternatives that replace the second paragraph.

3.2. Short Assignments

Next, we assemble various assignments from different subjects. For each assignment, we include 10 human-written responses and generate 3 alternatives using ChatGPT, only given the instructions (i.e., we did not present the system with students' work).

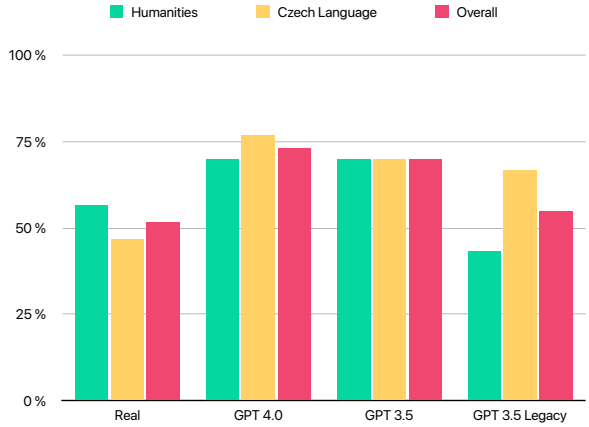
Sub-split C: holds the instructions and responses of an essay assignment in a 'English as the Second Language' course.

Sub-split D: holds the instructions and responses of an essay assignment in a 'German as the Third Language' course.

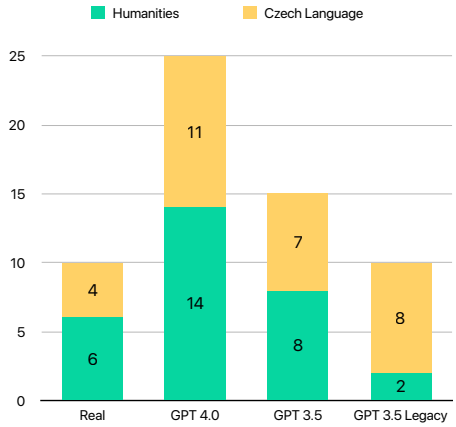
Sub-split E: holds the instructions and responses of a quiz assignment in a 'Math' class.

4. Human Assessment

We recruited 6 student peers, ages 18-20, from the same high school as the data was collected. Each participant was instructed on the task and later presented with the same data (i.e., the set of questions and reference texts was identical for each participant). We present the set of instructions and questions in Appendix B. Given average reading speeds, we designed the overall annotation task to take 75 minutes.



(a) Proportion of instances in which the abstract and keyword pairs were selected as relevant (given the scope of the work).



(b) Absolute instances in which the abstract and keyword pairs were selected as the best, given the selection.

Figure 2: Results of the peer quality assessment conducted on the sub-split A. The results are reported for authentic texts and three versions of ChatGPT (4.0, 3.5, and 3.5 Legacy). For each source, the scores are divided into subjects (and averaged under 'Overall', if relevant).

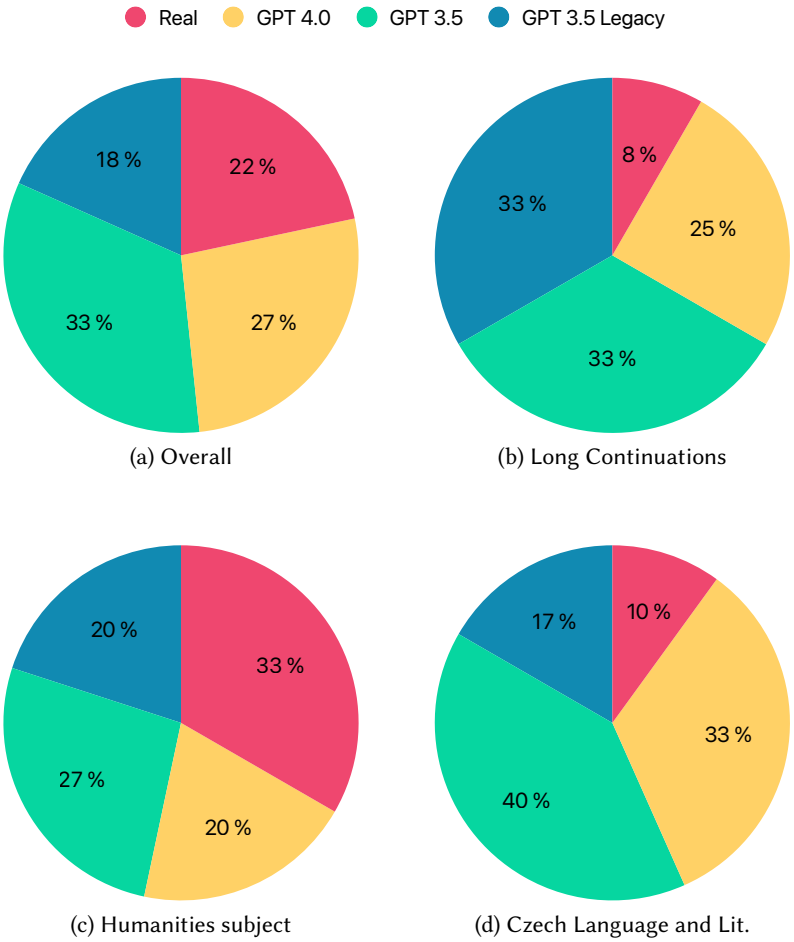
4.1. Quality Assessment

First, we assessed how the generated and authentic abstracts compare in terms of relevance (by peer student measures). For all 10 theses in sub-split A, the participants were presented with 4 alternative abstracts and keywords (1 authentic, 3 generated). We did not disclose which one is authentic and which is generated. The participants then had to select all options they deemed relevant (i.e., meeting the formal criteria and corresponding to the topic) and then select the single best one.

Shown in Figure 2a are the proportions of abstracts selected as relevant, grouped by model version and subject (the 'Overall' bar averages the subject-specific scores). Shown in Figure 2b are the absolute instances selected as the single best variants in the given selection, grouped by model version and subject.

We found that, on average, participants ranked abstracts generated by ChatGPT 3.5 Legacy similarly to the authentic ones, with around 50% of instances deemed relevant. Abstracts generated with ChatGPT 4.0 and 3.5 were perceived noticeably better: nearly 75% of their instances were deemed relevant.

As for the best option selection task, texts from ChatGPT 4.0 dominated, with a total of 25 of its instances selected as the best option. GPT 3.5 texts ranked second with 15 instances; authentic and GPT 3.5 Legacy texts share the last rank with 10 instances. Overall, there seems to be little to no statistically significant difference between the observed subjects.



(a) Overall (b) Long Continuations

(c) Humanities subject (d) Czech Language and Lit.

Figure 3: Results of the AI text identification assessment on the sub-split B. The results are reported for authentic texts and three versions of ChatGPT (4.0, 3.5, and 3.5 Legacy). Apart from overall statistics (a), results on a subset of longer continuations are presented (b), as well as subject-specific results for 'Humanities' and 'Czech Language and Literature' (c, d).

4.2. AI Text Identification

Next, we assessed whether participants could identify the authentic continuation of texts from sub-split B. Given 4 options, they were tasked to select the 1 authentic text among 3 generated ones. In general, humans without prior briefing on how to spot AI text are not able to do so [15, 16]; we were interested in whether this translates to the educational paradigm.

Shown in Figure 3a is the overall distribution of texts identified as authentic, grouped by the origin (e.g., authentic or model type). Authentic texts were selected as such only 22% of the time, which suggests that the participants are more likely to identify generated texts as authentic.

Most continuations in sub-split B (8 of the 10) were just a paragraph long. We wondered if an

extended generation range would affect the participants' judgment and created 2 special cases, where the continuation spans 3 paragraphs. Figure 3b captures the ranking distribution for this sub-case. Interestingly, pro-longed authentic texts were even less likely to be deemed authentic compared to their pro-longed counterparts.

Figures 3c and 3d divide the analysis into texts given their subject: 'Humanities' and 'Czech Language and Literature', respectively. While in 'Humanities', participants tend to select the authentic texts correctly more than the remaining classes, the latter subject suffers from a dominance of the AI-generated texts.

5. Automatic Assessment

Lastly, we tested the following publicly available services, promising to identify texts generated using ChatGPT:

- **Content at Scale: AI Content Detector**[2], yielding a likelihood of the text being written by human;
- **GPTZero**[3], classifying human-written, mixed, and AI-written texts;
- **OpenAI's AI Text Classifier**[4], classifying very unlikely, unlikely, unclear, possibly, or likely AI-generated texts;
- **Writer: AI Content Detector**[5], yielding a likelihood of the text being written by human;
- **ZeroGPT**[6], yielding a likelihood of the text being written by AI.

Even though most of these services provide a nuanced assessment, we converted them to a binary classification for the purposes of our study. We do not report conventional metrics that would indicate the performance of individual tools, as they all completely failed our test. When evaluated on sub-set A, OpenAI's AI Text Classifier predicted that all the items are AI-generated, while the rest of the services classified all the items as human-produced. This means that, if used in practice, all students who wrote the material in our dataset – regardless of whether they used AI or not – would be classified as cheaters or rule-abiding students, depending on the service. This shows that current services cannot detect AI content in Czech, at least in the educational domain.

6. Conclusion

To summarize, we collected a dataset of authentic high school coursework, including both long-form theses and short assignments, from a public high school in the Czech Republic and generated their AI alternatives and text continuations using ChatGPT with 4.0, 3.5, and 3.5 Legacy backbones. We make the data publicly available for open-domain research and analyses at https://www.matyasbohacek.com/the-unseen-a-plus-student.

[2] https://contentatscale.ai/ai-content-detector/
[3] https://gptzero.me/
[4] https://platform.openai.com/ai-text-classifier
[5] https://writer.com/ai-content-detector/
[6] https://www.zerogpt.com/

Through a study involving student peers, we found that ChatGPT can quickly produce high-school-level coursework that peers consider to be better than human-written text, even in a low-resourced language like Czech. Moreover, we show that the AI text detectors, which are slowly rolling out to campuses and educational centers worldwide, fail to identify these texts in Czech.

These results should be particularly alarming to educators and legislators who are establishing AI policies in their context. Thus, we call them to gather relevant data for their specific language and assignments specifics before making such decisions. At the same time, providers of AI text detectors should be more transparent about their models' performance, training data, and supported languages.

For future work, we aim to reproduce the study in various regional contexts while carefully analyzing the nuanced cases where ChatGPT is successful or unsuccessful. We also plan on including a group of teachers in addition to more student peer participants.

Acknowledgments

We would hereby like to thank Dr. Činátlová for her valuable insight and initiative when communicating with teachers and students at the subject high school, as well as all her thought-provoking comments. Additionally, we would like to thank Progresus Invest Holding for their generous sponsorship of this research and mobility-associated costs.

References

[1] OpenAI, GPT-4 technical report, ArXiv abs/2303.08774 (2023).

[2] P. Wood, M. L. Kelly, "everybody is cheating": Why this teacher has adopted an open ChatGPT policy, 2023. URL: https://www.npr.org/2023/01/26/1151499213/chatgpt-ai-education-cheating-classroom-wharton-school.

[3] A. C. C. Board, 2022-23 guidance for artificial intelligence tools and other services, ???? URL: https://apcentral.collegeboard.org/exam-administration-ordering-scores/administering-exams/preparing-for-exam-day/exam-security/artificial-intelligence-tools.

[4] C. Cassidy, Australian universities split on using new tool to detect AI plagiarism, 2023. URL: https://www.theguardian.com/australia-news/2023/apr/16/australian-universities-split-on-using-new-tool-to-detect-ai-plagiarism.

[5] M. Yang, New York City schools ban AI chatbot that writes essays and answers prompts, 2023. URL: https://www.theguardian.com/us-news/2023/jan/06/new-york-city-schools-ban-ai-chatbot-chatgpt.

[6] K. Jimenez, "this shouldn't be a surprise" the education community shares mixed reactions to ChatGPT, 2023. URL: https://eu.usatoday.com/story/news/education/2023/01/30/chatgpt-going-banned-teachers-sound-alarm-new-ai-tech/11069593002/.

[7] L. Lonas, Plagiarism finder Turnitin adds AI detection amid popularity of ChatGPT, 2023. URL: https://thehill.com/policy/technology/3928562-plagiarism-finder-turnitin-adds-ai-detection-amid-popularity-of-chatgpt/.

[8] J. Hsu, Plagiarism tool gets a ChatGPT detector – some schools don't want it, 2023. URL: https://www.newscientist.com/article/2367322-plagiarism-tool-gets-a-chatgpt-detector-some-schools-dont-want-it/.
[9] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, S. Feizi, Can AI-generated text be reliably detected?, ArXiv abs/2303.11156 (2023).
[10] D. Baidoo-Anu, L. O. Ansah, Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning, SSRN Electronic Journal (2023).
[11] T. Adiguzel, M. H. Kaya, F. K. Cansu, Revolutionizing education with AI: Exploring the transformative potential of ChatGPT, Contemporary Educational Technology (2023).
[12] M. Sallam, ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns, Healthcare 11 (2023).
[13] F. M. Megahed, Y.-J. Chen, J. A. Ferris, S. Knoth, L. A. Jones-Farmer, How generative AI models such as ChatGPT can be (mis)used in SPC practice, education, and research? an exploratory study, ArXiv abs/2302.10916 (2023).
[14] M. M. Rahman, Y. Watanobe, ChatGPT for education and research: Opportunities, threats, and strategies, Applied Sciences (2023).
[15] L. Dugan, D. Ippolito, A. Kirubarajan, S. Shi, C. Callison-Burch, Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text, in: The 37th AAAI Conference on Artificial Intelligence, 2023.
[16] E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, N. A. Smith, 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), Association for Computational Linguistics, Online, 2021, pp. 7282–7296. URL: https://aclanthology.org/2021.acl-long.565. doi:10.18653/v1/2021.acl-long.565.
[17] G. Jawahar, M. Abdul-Mageed, L. V. S. Lakshmanan, Automatic detection of machine generated text: A critical survey, in: International Conference on Computational Linguistics, 2020.
[18] D. Ippolito, D. Duckworth, C. Callison-Burch, D. Eck, Automatic detection of generated text is easiest when humans are fooled, in: Annual Meeting of the Association for Computational Linguistics, 2019.
[19] S. Gehrmann, H. Strobelt, A. M. Rush, GLTR: Statistical detection and visualization of generated text, in: Annual Meeting of the Association for Computational Linguistics, 2019.
[20] E. Crothers, N. Japkowicz, H. L. Viktor, Machine generated text: A comprehensive survey of threat models and detection methods, ArXiv abs/2210.07321 (2022).
[21] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, J. Y. Zou, Gpt detectors are biased against non-native english writers, ArXiv abs/2304.02819 (2023).
[22] A. Alimardani, E. A. Jane, We pitted ChatGPT against tools for detecting ai-written text, and the results are troubling, 2023. URL: https://theconversation.com/we-pitted-chatgpt-against-tools-for-detecting-ai-written-text-and-the-results-are-troubling-199774.

A. Prompts

Original (Czech)	Translated (English)
Pokračuj v psaní práce. Abstrakt: `[abstract]` Klíčová slova: `[keywords]` Úvod:	Resume writing of this thesis. Abstract: `[abstract]` Keywords: `[keywords]` Introduct:
Toto je odborná práce na téma "`[topic]`". Pokračuj v psaní textu: " `[portion of the text]`	This is a thesis concerning the topic of "`[topic]`". Resume writing of this thesis: `[portion of the text]`
Toto je úvod maturitní práce: "`[introduction]`" Toto je závěr maturitní práce: "`[conclusion]`" Napiš abstrakt ve stejném stylu:	This is the introduction of a high school leaving thesis: "`[introduction]`" This is the conclusion of a high school leaving thesis: "`[conclusion]`" Write an abstract in the same style:
Toto je úvod maturitní práce: "`[introduction]`" Toto je závěr maturitní práce: "`[conclusion]`" Napiš krátkou anotaci a klíčová slova:	This is the introduction of a high school leaving thesis: "`[introduction]`" This is the conclusion of a high school leaving thesis: "`[conclusion]`" Write a short annotation and keywords:
Toto je zadání úkolu do předmětu `[subject]` na střední škole: "`[instructions]`". Vypracuj úkol:	This is an assignment in `[subject]` class at a high school: "`[instructions]`". Complete the assignment:

Table 1
Prompts used for generating the alternatives or continuations of coursework within our dataset.

B. Survey Instructions and Questions

Original (Czech)	Translated (English)
Pomocí tohoto dotazníku analyzujeme, zda jsou generativní AI modely schopné odpovídat na různé typy úkolů a zda jsou tyto texty rozpoznatelné od těch skutečných, lidsky napsaných.	With this questionnaire, we seek to analyze whether generative AI models are able to complete different kinds of coursework and whether these texts are recognizable from real, human-written ones.
Níže uvidíte několik verzí abstraktu ke stejné maturitní práci z humanitních studií nebo českého jazyka. U každé práce zodpovězte následující otázky: 1. Které z navrhovaných možností **fungují jako adekvátní abstrakt** (tzn. nastiňují předmět a cíl práce, krátce shrnují obsah, a hlavně navnazují čtenáře*řku k tomu, aby si celou práci přečetl*la)? — *můžete zvolit libovolný počet odpovědí (tzn. klidně všechny nebo žádnou)* 2. Která z navrhovaných možností je, podle Vás, **pro svůj účel nejvhodnější**? — *volte právě jednu možnost*	Below, you will be presented with different alternatives for an abstract to accompany graduation theses (from Humanities or Czech language subjects). For each thesis, answer the following questions: 1. Which suggested options **work as an adequate abstract** (i.e., outline the topic and aims of the work, briefly summarize its contents, and—perhaps most importantly—grasp the reader)? — *you may select any number of options (i.e., including all and none)* 2. Which of the proposed options do you think is the **most suitable for its purpose**? — *you must select only one option*
Které z navrhovaných možností fungují jako adekvátní abstrakt?	Which suggested options work as an adequate abstract?
Která z navrhovaných možností je, podle Vás, pro svůj účel nejvhodnější?	Which of the proposed options do you think is the most suitable for its purpose?
Níže uvidíte několik krátkých úryvků z maturitních prací z humanitních studií nebo českého jazyka. U každého se nachází 4 alternativní pokračování – 1 skutečné (původní), 3 vygenerována pomocí GPT-4. Vyberte vždy tu variantu, u níž si myslíte, že pochází z původní, člověkem psané práce.	Below, you will be presented with short excerpts from graduation theses (from Humanities or Czech language subjects). For each, there are 4 alternative continuations - 1 real (original) and 3 generated by GPT-4. For eacg thesis, select the variant you think comes from the original, human-written work.
Která z navrhovaných možností, podle Vás, pochází z původní, člověkem psané práce?	Which of the proposed options do you think comes from the original, human-written work?

Table 2
Instructions and questions in the digital questionnaire participants completed after an in-person briefing.