Authorship Verification Using Cloze Test with Large Language Models

Tomáš Foltýnek **Tomáš Kancko** **Pavel Rychlý**
Faculty of Informatics, Masaryk University, Brno, Czechia
foltynek@fi.muni.cz 469029@mail.muni.cz pary@fi.muni.cz

Abstract

Assignment outsourcing, also known as contract cheating, occurs when a student outsources an assessment task or a part of it to a third party. It has been one of the most pressing ethical issues in university education and was further exacerbated by the wide availability of chatbots based on large language models. We propose a method that has the potential to verify the authorship of a document in question by filling in a cloze test. A close test with 10 items selected by our method can be used as a classifier with an accuracy of 0.988 and a F_1 score of 0.937. We also describe a general method for building a cloze-test-based classifier when the probability of authors and non-authors correctly filling in cloze items is known.

1 Introduction

Student assessment plays a critical role in education, aiming not only to provide feedback on student learning but also to verify students' skills and abilities. Assessment methods based on written documents (essays or theses) are common in most disciplines and most countries. To ensure the assessment's security, the institution has to be sure that the given student really wrote the document that demonstrates the required skills and abilities. Individual unsupervised work creates space for various forms of misconduct like plagiarism, assignment outsourcing (Awdry, 2020), or unauthorised content generation using tools based on generative AI (Foltýnek et al., 2023)

For decades, teachers have been using various technological tools to detect potential misconduct. Support tools for plagiarism detection identify text matches that may be used as evidence of plagiarism (Foltýnek et al., 2020). However, not every text match constitutes plagiarism (proper citations, random matches, general collocations), and not every plagiarism can be detected via text match (translation or paraphrase or other types of disguises). Nonetheless, compared to other forms of misconduct, plagiarism seems to be identifiable fairly easily.

Contract cheating (Clarke and Lancaster, 2006), also known as assignment outsourcing (Awdry, 2020), happens if a student hires a third party to complete an assignment for them. This form of misconduct is much harder to identify as the contractor produces an original document that is unlikely to have any identifiable text matching the documents in the tool's database. Some text-matching tools also provide stylometric analysis to identify potential contract cheating. However, these methods require a corpus of documents written by a given person for a comparison of stylometric features.

The problem is even harder in the case of text generated by AI. Current tools are capable of generating text almost indistinguishable from human-written text. Even though there are systems that claim to detect AI-generated text, they produce both false positives and false negatives. Moreover, there is no evidence of misconduct, which means that the outputs of these tools are barely useful in disciplinary procedures (Weber-Wulff et al., 2023). With these limitations in mind, such tools should not be used in academia at all, and educators are recommended to rethink assessment strategies so that they are not focused on the written piece of work (Perkins et al., 2024).

There have also been efforts to leverage the potential of LLMs in authorship identification or attribution, e.g. (Huang et al., 2024), who proposed a novel framework that combines instruction-based prompting with parameter-efficient fine-tuning to enhance LLMs' performance in authorship verification tasks. The notable advantage of their method is transparent and understandable explanations for its decisions, addressing the explainability challenge in authorship analysis. Nonetheless, it still suffers

Proceedings of Recent Advances in Natural Language Processing,pages 369–377
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-045

from an inherent drawback of the stylometry-based verification method, which is the need for documents written by the same person.

Still, written assignments – when the student really writes them – are meaningful forms of student assessment, and many educators don't want to give them up. Even if the text comes from other sources (copied from existing documents, written by someone else or generated by something else), the learning outcomes may have been achieved if the student's input was significant enough and the student thoroughly understood the matter and demonstrated their writing skills. Many educators are willing to tolerate potential misconduct, including using unauthorised aid, as long as the student achieves the desired learning outcomes. To meet this demand, some companies started developing tools for reliable authenticity verification of student submissions. Examples of such tools are NorValid, Mentafy, or Auth+ (Quesnel et al., 2023).

This paper proposes a method that can reliably confirm the authorship of a document in question in case a suspicion is raised by a technological tool. It has the potential to complement existing tools or the tools being developed, and together with them, it can save students' assignments as a reliable form of assessment. Moreover, it does not require any other documents to compare.

2 Cloze Test

The cloze test was introduced by Wilson Taylor in 1953. It consists of units defined as *"any single occurrence of a successful attempt to reproduce accurately a part deleted from a "message" (any language product) by deciding, from the context that remains, what the missing part should be"* (Taylor, 1953).

The original purpose was to measure the readability of the text: The higher the likelihood that participants guess the missing word correctly, the more readable the given text is. Nonetheless, the method has numerous other applications – identification of author writing style, text comprehension (Glatt and Haertel, 1982), or "an objective measure of language correspondence between the reader and writer" (Rankin, 1959, cited by Glatt and Haertel 1982).

Glatt and Haertel (1982) performed a cloze-test experiment involving plagiarising and non-plagiarising students and showed that the non-plagiarising group achieved higher scores in cloze tests. Standing and Gorassini (1986) experimented with cloze tests constructed from the essay authored by a student and the essay authored by their classmates, confirming the results of Glatt and Haertel. Both studies blanked every 5th word regardless of its meaningfulness, frequency, part of speech, or other characteristics. Even though the differences between plagiarists and non-plagiarists were statistically significant, the method was not discriminative enough to avoid false positives and false negatives. Numerous studies examined the difficulty of cloze test items. Abraham and Chapelle (1992) summarises the most significant findings:

- Functional words are easier to guess than content words;
- The amount of context needed to restore the word increases the cloze item difficulty;
- The Length of the sentence increases the difficulty.

They then developed a theory based on intrinsic criteria (which can be derived from the text) and extrinsic criteria (students' previous knowledge). The overall difficulty is a combination of both. They conclude that the cloze test scores can be interpreted as "students' ability to retrieve content words from long-term memory or to find them elsewhere in the text" (Abraham and Chapelle, 1992). These results indicate that previous awareness of the text increases the cloze test score.

Gellert and Elbro (2013) showed that careful selection of blanked words allows for testing the comprehension of the text and could be used instead of more time-consuming question-answering tests. Over time, the cloze test became a common means of language testing, creating a need to develop systems to create cloze items automatically (Hoshino and Nakagawa, 2008).

In this study, we use the cloze test method to verify authorship. More specifically, our goal is to find out what words should be blanked so that the overall cloze test score allows us to derive a probability of authorship. To our knowledge, the first study exploring the potential of a cloze test for authorship verification was a diploma thesis of Dobeš (2022). Dobeš confirmed the results of Abraham and Chapelle regarding the relative easiness of guessing functional words compared to the content words. The fact that functional words are easier to guess from the surrounding context makes them more likely to be guessed by both authors and non-authors. Therefore, functional words have much weaker discriminative power between authors and non-authors.

In this study, we will explore the potential of language models to select cloze items that discriminate best between authors and non-authors. We will use Dobeš's dataset as a starting point that helps us to develop the selection method. We will then verify the usability of this method by a user study.

3 Method

Our goal is to design a method that maximises authors' success rates while minimising the success rate of non-authors. We will use LLMs trained to predict a word (token) from its context to simulate non-author behaviour. Selecting the most suitable multilingual language model will involve reviewing some of the publicly available models and comparing their properties. For the purposes of testing these properties, we will run the chosen models on the dataset of Dobeš (2022). We will prioritise the language model that achieves the highest success rate in filling words while still maintaining an acceptable size and speed. Table 1 displays the values of the properties we evaluated for the language models under consideration. The experiments were conducted on a Google Colab notebook with a GPU.

Based on the data presented in the table, there appears to be a trade-off between the speed and success rate of the language models in correctly guessing the missing word. Additionally, the sizes of the models are all relatively acceptable and, therefore, do not appear to be a significant factor in deciding which model to choose. After considering the advantages of each model, we selected mt5-large as our preferred choice due to its higher success rate in filling in missing words when compared to the other models. Though it is much slower and requires the use of GPU, it still satisfies our needs and significantly increases the success rate.

MT-5 (Xue et al., 2021) is a multilingual version of the Text-to-Text Transfer Transformer (T5, Raffel et al. 2020) using the encoder-decoder architecture, which has been trained on and covers over 100 languages, including the languages of our interest – Slovak, Czech, and English. The pretraining of T5 (mT5 closely follows it) consisted of replacing input tokens with masked ones and letting the model reconstruct the original sentence. This approach corresponds precisely to the task of the users in their cloze-test. The most represented language in the MT-5 training dataset was English, with 5.67% of overall tokens. The Czech language was 14th with 1.72%, and Slovak ended 28th with 1.19% of overall tokens, which still represents a solid number of 18 billion tokens (Xue et al., 2021).

Our goal is not to use the language model to fill in the words as such, but to find the words which, when used as cloze items, would discriminate between authors and non-authors the best. Therefore, we will take into account not only the words correctly guessed by the language model (i.e., those with the highest probability of being filled in that gap) but also the words that the model would consider as a good fit. In this (preliminary) study, we consider the top 20 words, according to their probability, to be a good fit.

Based on the careful examination of Dobeš (2022) dataset, we propose the following method of filtering candidate words with the MT5 language model. We hypothesise that the words that distinguish the best between authors and non-authors are those that the model could guess among the top 20 but not as its most certain choice, i.e., the words which ranked 2-20 based on the probability of fitting to the given context. The logic behind this approach lies in the idea that (1) people cannot remember every word they wrote in their documents; (2) if the word makes sense in the current context (the position is at most 20) but is not the most certain one, authors may fill it with a higher chance than non-authors.

We also considered other options, i.e., ranking 1 — 20 or 2 — infinity, but none of these modifications reached a higher difference between author vs. non-author scores on the dataset of Dobeš (2022). Considering the part of speech, the biggest difference was for nouns (see Figure 1).

Our method consists of the following steps:

1. Remove everything except the main body of the document, i.e., titles and university information, declarations and acknowledgements, abstracts and keywords, table of contents, appendices and bibliography, and sentences that contain too much non-text information, such as theorems, tables, lists of elements, URLs, code snippets, and so on.

2. Split the document's body into sentences, remove stop words and tokenise the text.

Model	**Success rate**		**Size**	**Speed**
(Multilingual version)	**First try**	**20 tries**	**[GB]**	**[words/s]**
distilbert-base-cased	9.8%	27.0%	0.54	47.4
bert-base-cased	25.2%	35.2%	0.71	28.5
xlm-roberta-base	36.7%	48.4%	1.12	25.1
xlm-roberta-large	40.0%	50.3%	2.24	13.3
mt5-large	50.1%	71.9%	4.92	1.1

Table 1: Initial language model selection

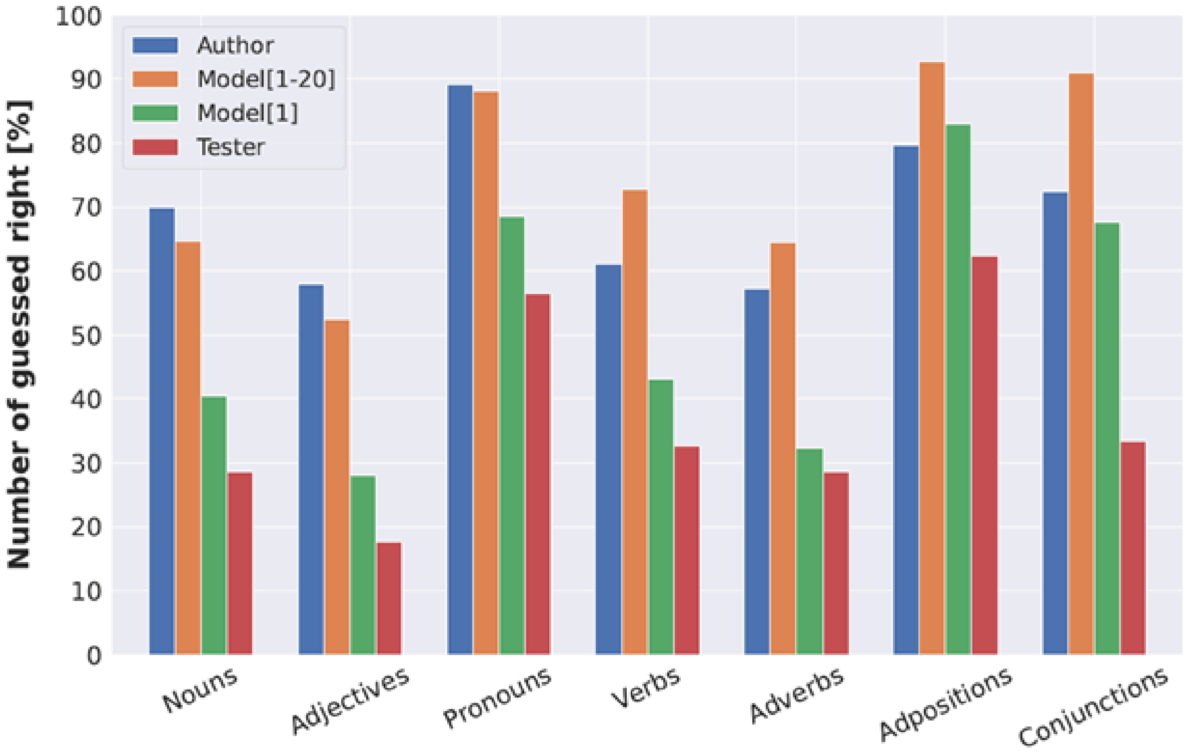

Figure 1: Success rate of authors, non-authors (testers) and MT-5 model. Model[1] indicates the top score of the correct word, and Model[1-20] indicates the correct word being among the top 20 model predictions.

3. Pass the tokenised sentences to a POS tagger to identify nouns.

4. Order the nouns according to their frequency.

5. Take the nouns, starting with the most frequent ones and select them if they meet the following conditions. Stop when the required number of words (typically 10) is selected or until the end of the list is reached:

   (a) The same word (or another form of it) has not yet been selected.

   (b) The same word (or another form of it) does not appear in the same sentence more than once.

   (c) The sentence has not yet been chosen for a different word.

   (d) The ranking of the word according to the language model is 2 – 20.

6. If all nouns have been processed and fewer than 10 words have been selected, random nouns that have not yet been chosen are selected. This situation typically arises only for documents shorter than two pages.

4 User Study

To test the usability of our method, we recruited a sample of 23 participants – our friends, colleagues and fellow students – with at least a college degree to participate in our study. Out of them, 15 were men and 8 were women. Ten participants had master's degrees, 11 had bachelor's degrees, and 2 did not have university degrees yet. All participants took part voluntarily and consented to use their written documents for the research.

The participants were asked to upload their documents (theses or essays), which resulted in a sample of 9 English, 2 Czech and 12 Slovak documents. From each document, a cloze test of 10 items was prepared utilizing the above-described method. Each cloze item consisted of a sentence taken from the document, with a single word blanked out. Each participant was requested to take the cloze test created from his/her document and at least one test created from someone else's document. Most of the participants took multiple cloze tests on non-authored documents. In total, we obtained 230 items from the authors and 730 items from the non-authors, forming a sufficiently large dataset of 960 items.

5 Results

As a baseline, we took the study of Dobeš (2022), which examined the success rate of authors and non-authors split according to the parts of speech. The overall success rates of authors and non-authors when guessing the words blanked by our method are in Table 2. As we can see, both authors and non-authors are more successful, but the difference between authors' and non-authors' scores is much larger, allowing for more reliable classification.

Success rate [%]	Baseline	Our method
Authors	60.98	84.35
Non-authors	23.08	27.12

Table 2: Comparison of our method with the baseline (Dobeš, 2022), page 55

Even though the overall percentages don't indicate evidence of (non-) authorship, when the cloze test contains multiple items, the overall score provides a much more accurate indication. Figure 2 shows the distribution of the number of correct answers for authors and non-authors. We can see that the relative counts, both for authors and non-authors, follow the normal distribution. As each cloze item can have two possible outcomes (correct/incorrect), we can consider them as independent Bernoulli trials and approximate our data with the binomial distribution. Pearson's correlation coefficient between the original and approximated data is 0.972, which indicates a very precise approximation.

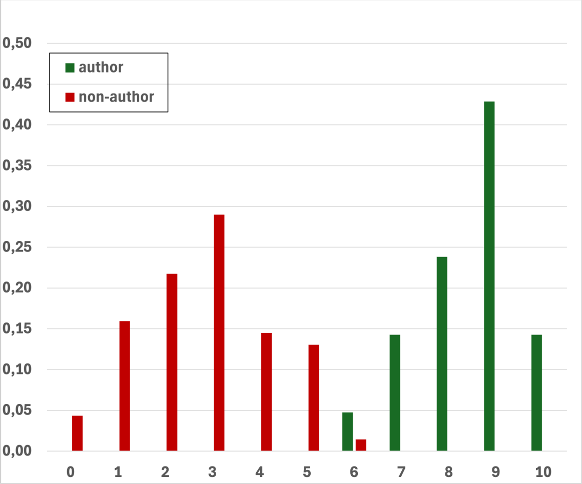

Figure 2: Relative count of authors/non-authors (y-axis) correctly guessing given number of items (x-axis)

6 Probability of Authorship

Let us now generalise our approach to any method of cloze item selection. Knowing the number of correct answers (denoted as $C = n$) given that the participant is the document's author (denoted as A), we know the a priori conditional probability $P(C = n|A)$. We will consider filling in individual cloze items as independent trials and use the probabilities $P(C = n|A)$ and $P(C = n|N)$ given by binomial distribution, where the success probability for each trial is known from the overall results of user study (see Table 3). We may use the Bayes formula

$$P(A|C = n) = \frac{P(C = n|A) \cdot P(A)}{P(C = n)} \tag{1}$$

to determine the posterior conditional probability $P(A|C = n)$, i.e., the probability of the participant being an author given the number of correct answers in the cloze test. The phenomenon N means that the given participant is not an author of the document. Obviously, $P(N) = 1 - P(A)$. In the further text, we will denote $cA(k) = P(A|C = k)$ and $cN(k) = P(N|C = k)$.

Based on this, we can construct a Naïve Bayes classifier to distinguish between authors and non-authors given the number of correct answers:

$$cA(n) = \frac{P(C = n|A) \cdot P(A)}{P(C = n)} \tag{2}$$

$$cN(n) = \frac{P(C = n|N) \cdot P(N)}{P(C = n)} \quad (3)$$

In the equations above, $P(C = n) = P(C = n|A) \cdot P(A) + P(C = n|N) \cdot P(N)$, where $P(C = n|A)$ and $P(C = n|N)$ are given by binomial distribution.

The overall probability of authorship $P(A)$ may be estimated based on the academic integrity literature, specifically on the studies dealing with assignment outsourcing. In the experiment of Glatt and Haertel (1982), three out of 75 undergraduate students (i.e., 4%) confessed to plagiarism when the aim of the study was explained, and students were guaranteed no penalty. In their calculations of conditional probability, they used the proportion of plagiarists equal to 5%. The meta-study of (Newton, 2018) identified 71 samples of students being surveyed about commercial contract cheating, including a total of 54,514 participants. The mean percentage of students admitting to having submitted an essay obtained from a contract cheating company was 3.5%, but the trend was clearly increasing. The percentage reported by contract cheating studies heavily depends on the way how students are asked and what scenarios are considered cheating. For example, a study from Czechia found out that 7% of students have used a commercial company to write an essay or thesis for them (Foltýnek and Králíková, 2018), but when the cheating scenarios include also having an essay written by a friend or family member, the percentage raised to shocking 19.7% (Králíková et al., 2018). Therefore, we can see that any number ranging from 3% to 20% can be justified by selecting an appropriate study from the body of academic integrity literature.

In our calculations, we used the estimates of 3%, 5%, 7%. 10%, 15% and 20%. The probability of authorship given the number of correct answers is shown in Figure 3. A score of 3 or less out of 10 items can be considered evidence of non-authorship, especially in common law jurisdictions allowing for the balance of probabilities in civil proceedings (Wright, 2011). The students who correctly guessed only 4 cloze items are probably not authors, but the evidence is not strong enough. For 5 correct guesses out of 10, the authorship is unclear. Students who correctly guess 6 or more items are likely authors of the document in question.

7 Classifier Performance

Knowing the a priori probability of non-authors and having a fixed number of cloze items in a test, we can derive the classifier's performance, specifically the accuracy and F_1 score. Note that the classes of authors and non-authors are heavily unbalanced in real-world scenarios; therefore, considering only accuracy could provide misleading information. The situation when an author is classified as a non-author is considered a false positive. The situation when a non-author is classified as an author is considered a false negative.

$$FP = \sum_{k=0}^{n} [cA(k) < cN(k)] * cA(k) \quad (4)$$

$$FN = \sum_{k=0}^{n} [cA(k) \geq cN(k)] * cN(k) \quad (5)$$

$$TN = \sum_{k=0}^{n} [cA(k) \geq cN(k)] * cA(k) \quad (6)$$

$$TP = \sum_{k=0}^{n} [cA(k) < cN(k)] * cN(k) \quad (7)$$

Then,

$$Acc = 1 - (FP + FN) \quad (8)$$

and

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (9)$$

The accuracy and F_1 scores are depicted in Figure 4 and 5. We can see that the main determinant of the overall classifier performance is the difference between authors' and non-authors' success rates.

8 Discussion

Standing and Gorassini (1986), who used the cloze method for plagiarism detection and selected blanked words randomly, reported mean scores in their two experiments: The authors achieved 84.3% and 84.5%, while non-authors achieved 66.4% and 58.5%. Our method achieved a much larger difference, namely by significantly decreasing the success rate of non-authors. This is particularly the result of the involvement of an LLM that allows us to filter out potential blanks that would be easy to guess from the context.

There are several potential avenues to further increase the performance of our classifier. First, we can adjust the selection method to achieve a larger

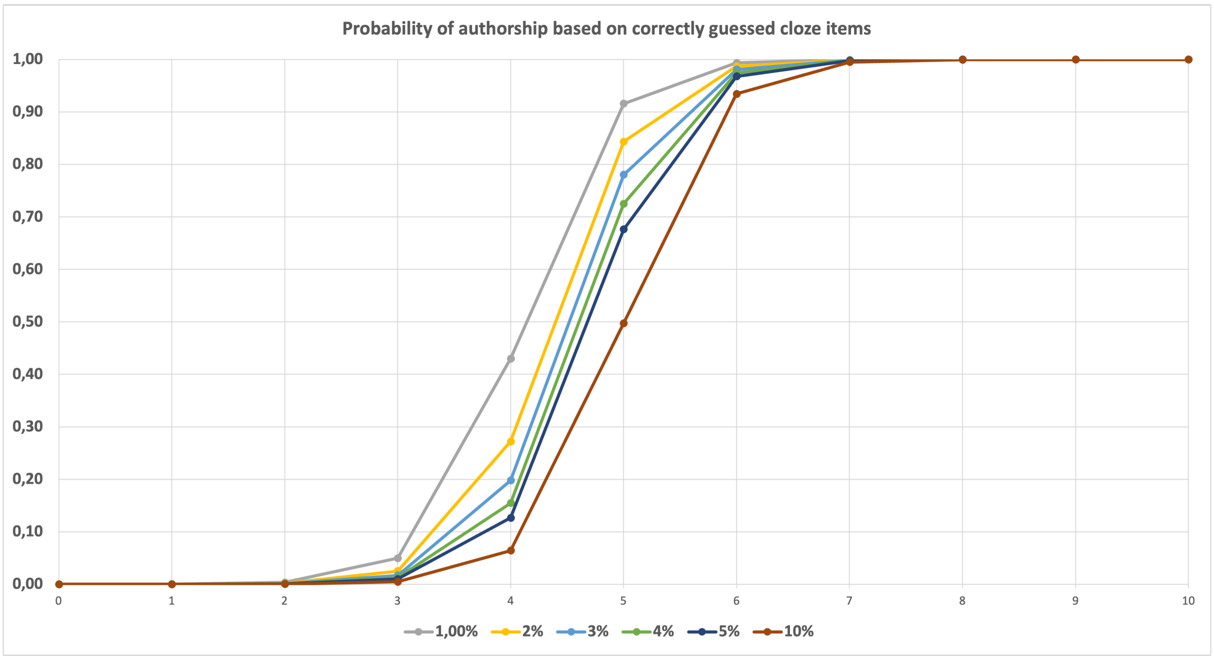

Figure 3: Probability of authorship given the number of correctly guessed cloze items for different percentages of cheating students.

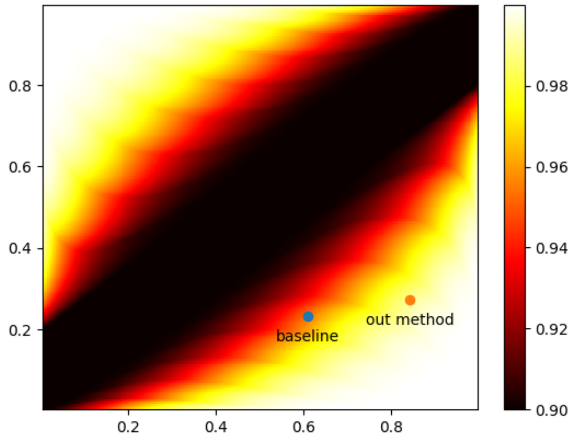

Figure 4: Colour-coded classifier accuracy for a cloze test of 10 items with 0.1 a priori probability of non-authors. X-axis: the probability of a correct answer by a non-author, Y-axis: the probability of a correct answer by an author. Due to the chosen a priori probability, the minimum accuracy is 0.9 (black colour). This occurs for similar probabilities of correct answers of authors and non-authors (along the main diagonal).

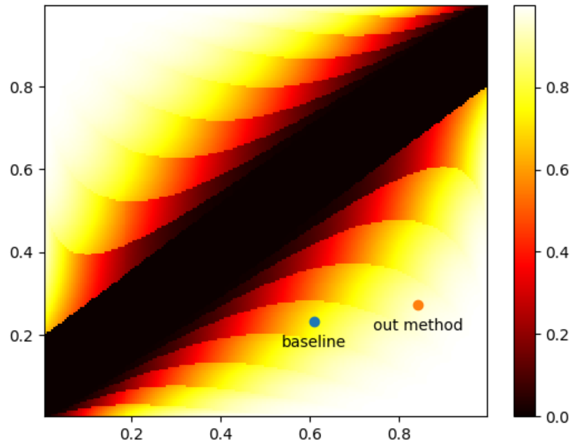

Figure 5: Colour-coded classifier F_1 score for a cloze test of 10 items with 0.1 a priori probability of non-authors. X-axis: the probability of a correct answer by a non-author, Y-axis: the probability of a correct answer by an author. If these probabilities are equal, the classifier is unable to distinguish authors from non-authors ($F_1 = 0$, black colour).

span between the success rates of authors and non-authors, which would consequently increase the classification accuracy. Nonetheless, even using the current method, simply increasing the number of items in the cloze test increases the classifier's performance. For 15 items, the classification accuracy for a scenario with 20% cheaters in the population would be 0.995%. The global minimum of the accuracy of the classifier using 15 cloze items is 0.993 in an even harder-to-imagine scenario with 72% cheaters in the population. In a practical setting, if a student receives a score from 4 to 6, which corresponds to the lowest classification confidence, repeating the test is an option to achieve more convincing evidence. Note that this happens in less than 10% cases if the population contains 20% cheaters and in approx. 6.4% cases if the population contains 5% cheaters.

8.1 Risks and Limitations

Even though the results are convincing, there is still a small chance of false accusations against an innocent student when a tool based on our results is used in disciplinary hearings. Therefore, when used in an academic setting, we recommend complementing this method with other methods to get a more complex picture of students' activities and learning outcome achievements.

There are several limitations in the study, which have to be taken into consideration when interpreting the results. The time gap between writing the document and the test could have influenced the results of the authors. We are not taking this aspect into account. The second limitation is certainly the small number of study participants. Reproducing our experiment with a larger cohort would make the results more convincing. The third limitation lies in the selection (or filtering) method itself. We considered the top 20 words regardless of their probability distribution. Taking this aspect into account may lead to a better distinction between authors and non-authors and allow for a more accurate classifier. We plan to address these limitations in our further studies.

9 Conclusion

This study investigated the potential of the cloze test generated with the help of LLMs in authorship verification. We propose the method which takes the most frequent nouns from the document in question and filters out those which are either the most probable candidates to fill the gap according to the LLM (i.e., anyone would correctly guess them from the context) or ranked worse than 20 (i.e., don't fit well to the context and even the author would struggle with guessing them correctly). Notable aspect of our method is the use of Bayes' formula, allowing for the incorporation of a priori probabilities related to authorship. This is particularly useful in real-world scenarios with inherently unbalanced classes.

Our study of 23 participants shows that if the words selected by our method are blanked, the authors fill them in correctly significantly more often than non-authors. A cloze test of 10 or 15 such items may be used as a reliable form of authorship verification in scenarios where stylometry or other techniques relying on documents from the author's history are not viable.

There are multiple usages of our method, depending on the legal background, university environment and course setting. It may be used as part of an exam or consultation about a written assignment, or complement other methods of misconduct detection, or it may be used as part of disciplinary hearings. Students suspected of assignment outsourcing or unauthorised content generation may be asked to fill in a cloze test prepared from the document they allegedly wrote. If students are supervised and don't have the documents at their disposal, the probability of authorship can be derived from the overall cloze test score. A classifier using a cloze test with 10 items achieves an accuracy of 0.988 and the F_1 score of 0.937. Other parameters and their comparison with baseline can be seen in Table 2. Moreover, the formulas (4) to (7) allow for the calculation of any classifier metrics relevant for the particular use case.

Cloze items	Parameter	Baseline	Our method
10	Accuracy	.949	.988
10	Precision	.862	.983
10	Recall	.584	.895
10	F_1 score	.696	.937
15	Accuracy	.968	.996
15	Precision	.968	.992
15	Recall	.968	.972
15	F_1 score	.824	.982

Table 3: Classifier parameters and their comparison with the baseline of (Dobeš, 2022). A priori probability of non-authors is 0.1

In our further studies, we plan to address the limitations of this preliminary study, specifically, the small sample of study participants, filtering out the words based on their rank without considering the probability distribution and omitting the time gap between writing the document and taking the test. Despite these limitations, the results are promising so far, and we believe we will be able to improve them further in order to develop a reliable authorship verification tool.

Acknowledgments

This research project was funded by the Czech Technological Agency, grant no. TQ01000110.

References

Roberta G Abraham and Carol A Chapelle. 1992. The Meaning of Cloze Test Scores: An Item Difficulty Perspective. *The Modern Language Journal*, 76(4):468–479. Publisher: [National Federation of Modern Language Teachers Associations, Wiley].

Rebecca Awdry. 2020. Assignment outsourcing: moving beyond contract cheating. *Assessment & Evaluation in Higher Education*, 46(2):1–16. Publisher: Routledge.

R Clarke and T Lancaster. 2006. Eliminating the successor to plagiarism? Identifying the usage of contract cheating sites. *Proceedings of the 2nd International Plagiarism Conference, Gateshead, UK.*

Erik Dobeš. 2022. Ghostwriting detector. Diploma thesis.

Tomáš Foltýnek, Sonja Bjelobaba, Irene Glendinning, Zeenath Reza Khan, Rita Santos, Pegi Pavletic, and Július Kravjar. 2023. ENAI Recommendations on the ethical use of Artificial Intelligence in Education. *International Journal for Educational Integrity*, 19(1):12.

Tomáš Foltýnek, Dita Dlabolová, Alla Anohina-Naumeca, Salim Razı, Július Kravjar, Laima Kamzola, Jean Guerrero-Dib, Özgür Çelik, and Debora Weber-Wulff. 2020. Testing of support tools for plagiarism detection. *International Journal of Educational Technology in Higher Education*, 17(1):46.

Tomáš Foltýnek and Veronika Králíková. 2018. Analysis of the contract cheating market in Czechia. *International Journal for Educational Integrity*, 14(1):4.

Anna S Gellert and Carsten Elbro. 2013. Cloze Tests May be Quick, But Are They Dirty? Development and Preliminary Validation of a Cloze Test of Reading Comprehension. *Journal of Psychoeducational Assessment*, 31(1):16–28.

Barbara S Glatt and Edward H Haertel. 1982. The Use of the Cloze Testing Procedure for Detecting Plagiarism. *The Journal of Experimental Education*, 50(3):127–136. Publisher: Routledge.

Ayako Hoshino and Hiroshi Nakagawa. 2008. A cloze test authoring system and its automation. In *Advances in Web Based Learning – ICWL 2007*, pages 252–263, Berlin, Heidelberg. Springer Berlin Heidelberg.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *arXiv preprint arXiv: 2408.08946.*

Veronika Králíková, Tomáš Foltýnek, Jana Dannhoferová, Dita Dlabolová, and Pavel Turčínek. 2018. Global Essay Mills Survey in Czechia: Insights into the Cheater's Mind. In Salim Razı, Irene Glendinning, and Tomáš Foltýnek, editors, *Towards Consistency and Transparency in Academic Integrity*. Peter Lang, Bern, Switzerland.

Philip M Newton. 2018. How Common Is Commercial Contract Cheating in Higher Education and Is It Increasing? A Systematic Review. *Frontiers in Education*, 3.

Mike Perkins, Jasper Roe, Binh H. Vu, Darius Postma, Don Hickerson, James McGaughran, and Huy Q. Khuat. 2024. Simple techniques to bypass GenAI text detectors: implications for inclusive education. *International Journal of Educational Technology in Higher Education*, 21(1):53.

Matthew Quesnel, Robert Guderian, and Brenda Stoesz. 2023. Quizzing Students about their Writing: Implications for Deterring and Detecting Contract Cheating, and Promoting Academic Integrity and Greater Engagement. *Canadian Perspectives on Academic Integrity*, page Vol. 6 No. 1 (2023). Publisher: Canadian Perspectives on Academic Integrity.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Earl F Rankin. 1959. The cloze procedure: its validity and utility. In *Eighth yearbook of the national reading conference*, volume 8, pages 131–144. Milwaukee: National Reading Conference.

Lionel Standing and Donald Gorassini. 1986. An Evaluation of the Cloze Procedure as a Test for Plagiarism. *Teaching of Psychology*, 13(3):130–132. Publisher: Routledge.

Wilson L Taylor. 1953. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, 30(4):415–433.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1):26.

Richard W Wright. 2011. Proving Causation: Probability Versus Belief. In R Goldberg, editor, *Perspectives on Causation*. Hart Publishing, London.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. _eprint: 2010.11934.