scientific reports

Check for updates

OPEN

Predicting autism from written narratives using deep neural networks

Izabela Chojnicka[1]✉ & Aleksander Wawer[2]

Despite the heterogeneity of language and communication abilities within the autistic population, challenges associated with the pragmatic (social) use of speech remain consistently observable across the entire spectrum of autism. Therefore, the study of narrative competence is particularly relevant, and there has been a considerable rise in research on narrative skills in autism. Most studies have focused on spoken narratives, with some describing the potential use of automated computational methods. In this study, we analyzed written narratives collected in a standardized manner during a national exam. We gathered 363 essays from students in the final eighth grade of primary school: 193 from autistic students (ASD Group) and 168 non-autistic peers (non-ASD Group). We tested several deep neural models to predict whether an essay was written by an autistic student or a student from the non-ASD Group. Several models achieved promising results, exceeding values of 0.85 for sensitivity, specificity, and accuracy coefficients. In addition to studying narrative competence, the data from national exams and their utility in distinguishing autistic individuals may potentially pave the way for large-scale and cost-effective epidemiological studies on autism in the future.

Autism, clinically referred to as Autism Spectrum Disorder (ASD), belongs to the group of neurodevelopmental disorders of complex and heterogeneous etiopathogenesis related to brain development[1]. Over the past several years, the prevalence of autism has significantly increased and is estimated at 1% worldwide[2] and 2.8% in the US[3], making it a global matter in communities around the world. While in recent decades, numerous genetic, physiological, neurological, and other markers associated with ASD have been identified, they have not yet been translated into a diagnostic process[4]. ASD is defined behaviorally - the symptoms include persistent deficits in initiating and sustaining social communication and reciprocal social interactions, developing, maintaining, and understanding social relationships, and a range of inflexible, repetitive, and restricted patterns of behavior, interests, or activities[5].

Although autism is described behaviorally, atypical cognitive profiles are commonly observed, particularly in relation to social cognition and perception, executive functions, and perceptual, informational, and language processing[6]. Autism is characterized by significant heterogeneity in terms of language capacity, ranging from the lack of speech to fluent use of speech[5]. Autistic individuals struggle with both verbal and nonverbal communication skills. Despite the heterogeneity of language and communication abilities within the autistic population, challenges related to the pragmatic (social) use of speech are consistently observable across the entire spectrum of autism in individuals with different levels of intellectual and linguistic capacity and varying severity of symptoms typical of ASD throughout their lives[7]. Pragmatic language refers to using communication in social situations in everyday interactions with different people. It encompasses what we say, how we say it, our nonverbal communication, and how adequate our interaction and messages are for a given context[8]. Pragmatic competence is essential for effectively communicating our thoughts, ideas, and feelings. The pragmatic use of language involves the use of language for various purposes, such as greetings, sharing information, making declarations, and asking questions. It also encompasses adjusting the language used to the listener and situational context while acknowledging the perspective of others. Difficulties with pragmatic speech use can affect the daily lives of autistic individuals, potentially resulting in educational underachievement, difficulties in forming social relationships, and various psychological challenges[9].

One aspect of pragmatic competence that autistic individuals struggle with is narrative discourse[10]. Bruner [11] describes narrative discourse as the ability to think, communicate, and share experiences to adjust and reconstruct them for better understanding. Narrative ability is a complex cognitive task that involves the integration of information, remembering the details of a story, and using the collected knowledge about the world to create a coherent narrative describing a series of actions and events that unfold over time. Narrative skills help express

[1]Faculty of Psychology, University of Warsaw, 00-183 Warsaw, Poland. [2]Institute of Computer Science, Polish Academy of Sciences, 01-248 Warsaw, Poland. ✉email: izabela.chojnicka@psych.uw.edu.pl

one's thoughts and experiences using language in communication contexts. The development of narrative skills begins in the early stages of life and is associated with cognitive, social, and linguistic development[12]. Storytelling abilities have a significant impact on various aspects of the individual's development, such as planning, organizing, and ordering thoughts and actions, and most importantly, developing a sense of identity.

Previous studies on narration in autism have primarily focused on analyzing spoken storytelling abilities. Studies on writing skills, particularly narrative writing and writing development in autistic individuals, are more limited. Zajic and Wilson[13], in their systematic review, concluded that little is known about the writing development of autistic children and youth. They categorized writing challenges into lower-order (handwriting) and higher-order (written expression) difficulties. Handwriting is a perceptual-motor activity requiring the coordination of several simultaneous tasks and may contribute to higher-order story generation[14]. Price and colleagues[15] argue that pragmatic competence underlies written language ability and likely contributes to the writing difficulties reported in autistic adolescents. They investigated the writing skills of autistic adolescents (n =14) and their typically developing peers (n = 12) across the persuasive, expository, and narrative genres, after controlling for IQ. They reported no differences between groups in the length of narrative writings, but autistic adolescents used less dialogue and greater variety of words in their narratives. A study of autistic adolescents by Baixauli *et al.*[16] revealed challenges with writing skills, particularly in productivity, lexical diversity, and overall story coherence, similar to findings in spoken narratives.

Shevchuk-Hill *et al.*[17] compared stories written by autistic (n = 19) and non-autistic (n = 23) university students using automated methods. They concluded that writing may be considered a strength for autistic students, as their stories were rated at a higher reading level, contained fewer grammatical errors, and demonstrated a more positive writing affect. In our previous study[18], we employed natural language processing techniques to compare stories written by autistic adolescents and their non-autistic peers in terms of emotional valence, language abstraction and readability. To evaluate the readability of a story, we calculated the Gunning Fog Index, which estimates the years of formal education necessary for a person to comprehend the text upon their initial reading. Stories written by autistic students were shorter than those written by their peers, but more complex in terms of readability. Autistic students used words with positive evaluative meaning and verbs describing emotional and mental states statistically less frequently than non-autistic adolescents. The calculated language abstraction for the entire narratives was lower in the autistic than non-autistic group.

In our previous research, we explored the ability of several types of deep neural network-based text representation models to detect ASD based on spoken narratives[19,20]. We also analyzed the linguistic characteristics of the written stories[18]. Here, we evaluate the utility of several deep neural networks for identifying autism based on written narratives. We analyzed 195 essays written by autistic students and 168 written by non-autistic peers. The essays were written as part of the nationwide eighth-grade school exam in Poland.

Results

Predicting autism using deep neural networks

We evaluated the models in 10-fold cross-validation with the same seed. In each fold, the validation set was created from a randomly selected 5% of the train split. The best-performing model was subsequently evaluated on the test split. We report results averaged over all the folds. Table 1 presents an evaluation of these models using psychological/medical test metrics, and Table 2 shows the evaluation using machine learning metrics, along with information on the number of overall and trainable parameters.

Figure 1 illustrates the effectiveness of the evaluated models, considering the number of trainable parameters. The size of each circle represents the corresponding parameter count, with larger circles indicating a higher number of trainable parameters, as shown in Table 2.

Table 3 contains results of the computationally intensive randomization version of the paired sample (matched-pair) t-test. Under the null hypothesis, the results of the two models are not really different, so labels produced by one of the models could have just as likely come from the other. These responses are shuffled, and each is reassigned to one of the two models to compute how likely such a shuffle produces a difference in the metric(s) of interest (we used F1). This permutation test has several advantages: it relaxes the requirement for independence of samples that is often violated and does not underestimate the significance[21]. Most of the differences between the model pairs turned out to be significant. The exceptions are the two best-performing models: HerBERT large and GPT-2 large + LORA, and three other pairs containing the LaBSE model.

Model	Sensitivity	Specificity	PPV	NPV
HerBERT large	0.86	0.89	0.90	0.85
LaBSE	0.85	0.86	0.88	0.83
GPT-2 large	0.80	0.85	0.86	0.79
GPT-2 large + LORA	0.91	0.87	0.89	0.89
GPT-2 large + P-tuning	0.76	0.51	0.64	0.65
OpenAI ada + SVM	0.81	0.88	0.88	0.80
USE + SVM	0.77	0.82	0.83	0.76

Table 1. Results of models used for ASD prediction, medical test metrics. *NPV* negative predictive value, *PPV* positive predictive value, *SVM* support vector machines classifier.

Model	Accuracy	Overall #	Trainable #
HerBERT large	0.88 ± 0.02	330M	330M
LaBSE	0.86 ± 0.02	470M	470M
GPT-2 large	0.82 ± 0.02	775M	775M
GPT-2 large + LORA	0.89 ± 0.02	775M	875K
GPT-2 large + P-tuning	0.64 ± 0.03	775M	376K
OpenAI ada + SVM	0.84 ± 0.02	N/A	324K
USE + SVM	0.79 ± 0.02	85M	129K

Table 2. Results of models used for ASD prediction. Accuracy ± standard error of cross-validation estimation, number of overall and trainable parameters.

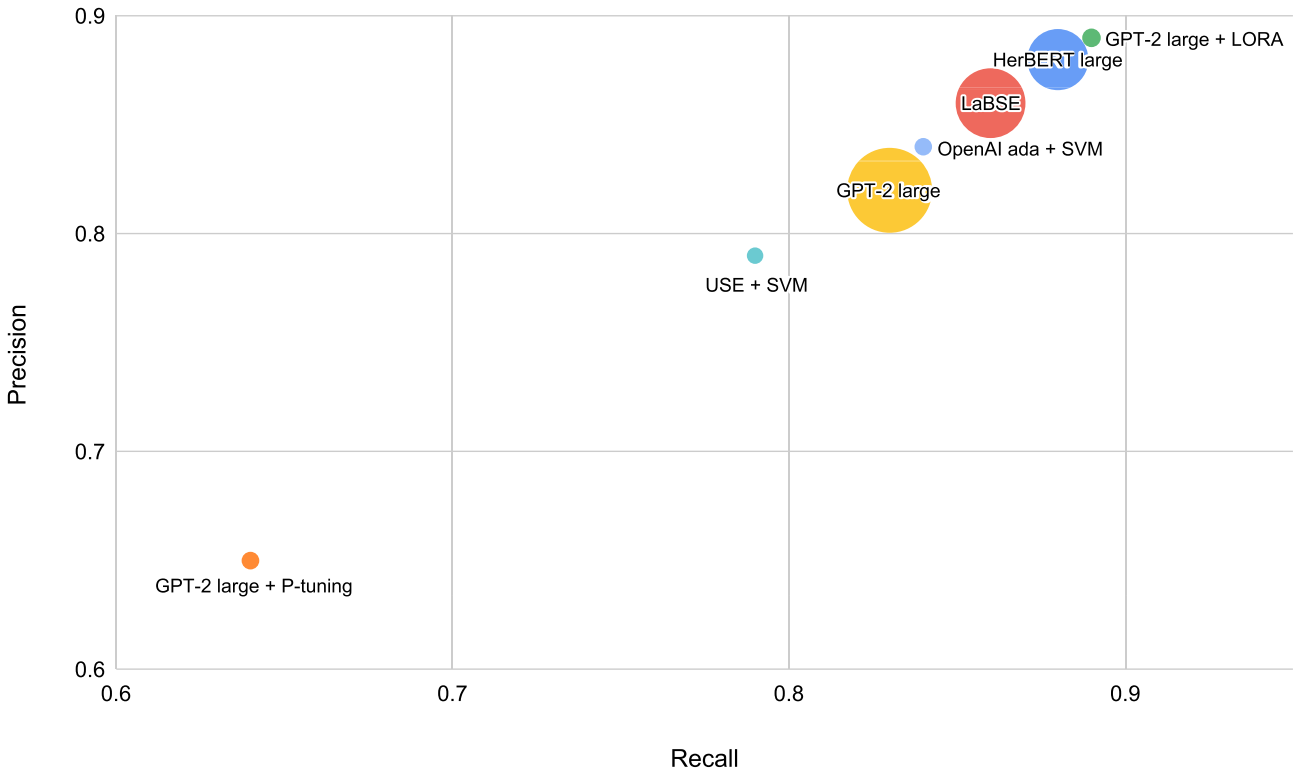

Fig. 1. Precision and recall of evaluated models. The size of a circle represents the number of trainable parameters. The larger the circle, the greater the number of trainable parameters, although the proportions between the circles do not accurately reflect the actual differences in the number of parameters (e.g., HerBERT large has 330M trainable parameters, while the SVM computed on OpenAI ada embeddings has 324K as in Table 2).

	HerBERT large	LaBSE	GPT-2 large	GPT-2 large + LORA	GPT-2 large + P-tuning	USE + SVM
LaBSE	–					
GPT-2 large	*	*				
GPT-2 large + LORA	–	–	*			
GPT-2 large + P-tuning	*	*	*	*		
USE + SVM	*	–	*	*	*	
OpenAI ada + SVM	*	*	*	*	*	*

Table 3. Results of permutation tests comparing the differences in results between pairs of models[21]. '∗' denotes significant differences between two models at p value=.005, '–' denotes no significant differences.

We calculated sensitivity, specificity, PPV, and NPV values for Human raters experienced in autism spectrum disorder diagnosis. They are fundamental metrics used to evaluate the performance of diagnostic tests[22]. Experts assigned randomly selected 45 essays, 20 from ASD group and 25 from non-ASD group to one of the two groups based solely on the content of the essays, with no additional clinical information. Human raters achieved significantly lower effectiveness: sensitivity was 0.56, specificity 0.66, positive predictive value (PPV) 0.58, and negative predictive value (NPV) was 0.59. The Krippendorff's alpha inter-rater reliability was 0.158 and indicates a low level of agreement between the human experts.

Discussion

Previous studies on narrative skills in autism have primarily focused on spoken narratives[10]. Here, we present the results related to written narrations collected in a standardized manner during a national exam. This allowed us to analyze a large sample—a frequent challenge in studies of clinical groups using deep learning methods. We compared stories written by autistic students to those from a control group, which included typically developing peers, peers with developmental learning disorder (often referred to as dyslexia, dysorthographia, and dyscalculia), and, most likely, ADHD and undiagnosed adolescents with neurodevelopmental or psychiatric conditions present in the population. We tested various deep neural models to assess their effectiveness in the automated identification of autism based solely on essay content and compared their performance with that of human raters.

In our previous inquiries on spoken narratives[19,23], participants in the autistic and control groups were matched for non-verbal and verbal intelligence quotients. In the current project, we did not have access to data regarding the intellectual functioning of the participants, particularly their non-verbal IQs. However, it is reasonable to assume that students from both groups were within the typical range of intellectual functioning (see Participants and Data section). Furthermore, the groups did not differ significantly statistically in terms of writing skills as measured by the number of points awarded by teacher-examiners in the categories: *Topic Development, Creative Elements, Literary Skills, Style, Language, Spelling, Punctuation*, and Total score. Hence, there was no significant difference between groups in the writing competencies assessed by teachers in a school-oriented manner. Essays by autistic participants were shorter compared to those written by peers from the control group ($p < 0.001$). This is consistent with previous studies indicating reduced productivity in spoken[10] and written narratives[14] among autistic individuals.

As the non-ASD group included individuals with developmental learning disorder, it is worth considering how this may have influenced the process of predicting autism by deep neural models. Developmental learning disorder refers to difficulties primarily affecting reading, writing, and/or mathematical skills, with typical intelligence range and intact sensory abilities. However, atypical cognitive profiles are prevalent in this population, particularly concerning executive functions, working memory, perceptual-motor integration, and information processing[24]. Difficulties characteristic of dyslexia or dysorthography may manifest as omitting, adding, or rearranging letters or words, substituting one word for another, distortions, and spelling or punctuation errors. Developmental Learning Disorder may co-occur with Autism Spectrum Disorder[5]. Therefore, taking into account the scores assigned by teachers, we hypothesize that errors of this kind might not influence the classification decisions of the models.

The two models yielded the best results: sensitivity 0.91, specificity 0.87, and accuracy 0.89 for GPT-2 large + LORA and sensitivity 0.86, specificity 0.89, and accuracy 0.88 for HerBERT large. Additionally, three other models—GPT large, LaBSE, and OpenAI ada + SVM—achieved accuracies above 0.80. There is still a long way to go before these types of models can be used for clinical purposes in the same way as the standardized instruments used today, but the results obtained are promising. Values of sensitivity, specificity, and accuracy like these would be considered good to excellent for a screening tool for autism spectrum disorder[25]. Most standardized instruments for an autism screening are designed to assess toddlers and young children[26]. There are far fewer screening tools with good psychometric properties for evaluating older children, adolescents, and adults. In their meta-analysis, Hirota and colleagues[27] identified three screening tools that proved relatively effective in assessing autism in older participants: the Autism-Spectrum Quotient, the Social Communication Questionnaire, and the Social Responsiveness Scale. However, other authors suggest the limited utility of the Social Communication Questionnaire due to the low specificity (25.7%) for screening purposes in school-aged children[28]. In this light, the possibilities of utilizing automated computational methods are worth exploring. Previous studies have investigated narrative writing in autistic individuals, typically developing individuals, and non-autistic individuals, highlighting several differences worth considering in this context. Autistic adolescents were reported to write shorter texts with less varied syntactic structures[16]. Baixauli *et al.*[16] also showed that autistic students struggled with the resolution component of coherence—specifically, mentioning what happened at the end of the conflict presented in the story. Hilvert *et al.*[29] investigated personal narrative writing skills among autistic children and adolescents in comparison to their neurotypical peers. They reported that autistic individuals wrote less syntactically diverse texts with a higher number of grammatical errors. In our previous study, we examined measures of language abstraction and emotional tones in stories written by autistic and non-autistic adolescents[18]. We found that autistic adolescents used fewer verbs related to mental and emotional states, fewer words with positive evaluative meanings, and their stories were less abstract than those written by non-autistic peers. The models we tested may utilize these types of text characteristics in the classification process. However, we do not have information on which specific features were considered or the weight assigned to them. Future research could help identify these features.

As expected, the best performing models (HerBERT large and GPT-2 large + LORA) were Polish monolingual. Interestingly, good results were achieved also by selected multilingual models (LaBSE and OpenAI ada + SVM). The question remains to what extent the information obtained by the model about the features of texts written by

autistic youths depends on the Polish language. In particular, how such a multilingual model trained on Polish texts will cope with recognizing autism in texts written in other languages.

Two solutions using embeddings and a separate classifier show the significant progress that has been made over the last few years in the field of Natural Language Processing. The USE model[30] with the SVM classifier achieves an accuracy of 0.79, while the ada embeddings from OpenAI achieve accuracy of 0.84 with the SVM classifier.

Figure 1 and Table 2 illustrate the connection between the number of parameters, both trainable and overall, and model performance. The positive effect of reducing the number of trainable parameters works in the case of GPT-2 with 775M parameters and the LORA approach. This combination yields the best result among all tested models; however, the difference compared to the second-best model, HerBERT large, is not significant (see Table 3). The conclusion that can be drawn from the results is that successfully solving the problem of diagnosing autism from school essays requires hundreds of thousands of trainable parameters, not hundreds of millions. However, the best results are achieved using models with hundreds of millions of parameters. Importantly, their parameters can mostly remain constant and do not have to be fully tuned.

In our previous study on spoken narratives in autism[19], we analyzed narratives from the ADOS-2[31,32] picture book task. The only model tested then, which we investigated in the current project, is USE. Intuition would suggest that spoken narratives may contain valuable meta-information for deep neural models. Moreover, in the study of spoken narratives, the control group consisted of neurotypical individuals carefully matched on age, gender, and verbal and non-verbal IQs. Nonetheless, USE achieved better sensitivity, specificity, and accuracy in written narratives than in spoken ones. This may be due to the sample size - in the case of spoken narratives, it was much lower (50 participants) than in the current study. However, it cannot be ruled out that there are differences between spoken and written narratives that influenced the obtained results.

To examine how the different wording of the narrative topic for the two groups might affect the results, we asked three experts in autism diagnosis to classify 40 written stories and provide feedback on any potential, obvious differences between the essays. The experts did not identify such differences, and their classifications were close to random. Moreover, as the inter-rater reliability analysis showed, the experts' classifications had a low level of agreement. Therefore, it was not the specific essays that were difficult to classify, but rather that all the raters' verdicts were inconsistent. We concluded that, while the different wording may influence the model predictions, this effect is neither obvious nor detectable by humans.

Similar to our previous study on spoken narratives[19], in this study, even to a greater extent, deep neural models outperformed human raters. In contrast to deep neural networks, an experienced clinician does not learn to diagnose patients solely based on their isolated written statements. We want to emphasize that our results do not indicate that AI models can replace clinical judgment and a recommended[33] extensive diagnostic process, which involves various methods, different sources of information, and a differential diagnosis. Instead, the results indicate the potential usefulness of automated tools in studying narrative competence in autism and as quantitative language differential between ASD and non-ASD participants. We would like to further investigate in the future the underlying sources of decision-making in deep neural models, specifically which aspects of the essays and language have the highest discriminatory power.

We want to address both the limitations and strengths of the study. The first limitation relates to our sample: a lack of demographic and clinical characteristics of participants except for writing skills; not perfect matching of participants based on sex, and unequal group sizes. We lacked the characteristics that would allow us to study identification rates while accounting for individual differences among participants. There is a possibility that an unequal gender ratio influenced the model results as a confounding variable. Unfortunately, precisely estimating the effect of this confounding variable in large neural networks such as BERT or GPT is not feasible. At the same time, results from human experts suggest that this influence is not straightforward to interpret—different gender proportions do not appear to affect classification effectiveness. In the future, we aim to explore possible sex differences—assessing the utility of deep neural models in identifing autistic girls and linguistic characteristics that differ among autistic girls, boys, and typically developing peers. On the other hand, a definite strength of the study was the relatively large sample size, with 363 participants randomly selected from different regions of Poland. We must emphasize again[19] the challenge for text-only predictions. As our study involved most likely participants without disorders of intellectual development, the findings may not be universally applicable to all individuals across the autism spectrum. From this standpoint, and considering the heterogeneity of both the autistic and the non-autistic groups, the results can be considered valuable, potentially revealing the upper limit of screening capabilities based solely on closed-text data. The analyzed essays were written in Polish, characterized by rich morphology and relatively flexible word order. Nevertheless, narrative challenges in autism appear to be universal across languages with different typologies (e.g.,[34–36]). Additional limitation relates to the slight difference in prompts provided to the two groups: autistic participants were asked to describe an adventure in a fictional world, whereas non-autistic participants were asked to describe a meeting with a character from that world. Although both prompts allowed for a wide range of narrative choices—including both social and non-social content—this difference may have encouraged greater use of social language in one group and non-social language in the other. Since reduced use of social language is itself a characteristic often associated with autism[37], it is challenging to fully disentangle the effect of the task wording from the underlying group differences. Nevertheless, we acknowledge this as a potential measurement artifact that may have influenced some of the patterns detected by neural models, even though it was not apparent to human raters nor identified through a comparison of personal pronoun usage (as one of the methods of quantifying social language) between the two groups. We consider this a valuable direction for future research. While the obtained results are promising, we do not know the characteristics of the essays that underlie these findings. As mentioned previously, we would like to pursue this in future research, employing an explainable AI approach.

We have presented a promising deep-learning approach to identify autism based on written narratives using text representation models. For several models, we achieved good and excellent test metrics for autism identification efficiency. Although encouraging, the results are preliminary and certainly require further research. They do, however, offer promise for future automated, objective, time, and labor-effective solutions for studying language and narrative competence in autism and perhaps also for developing methods useful for screening purposes in older children, adolescents, and adults with autism.

Methods

Participants and data

We collected 363 essays from students in the final eighth grade of primary school. Among them, there were 195 essays written by autistic students (ASD Group; average age 14.85 years, SD = .62; 25% girls) and 168 non-autistic students (non-ASD Group; average age 14.64 years, SD = .48; 39% girls). The non-ASD group included not only neurotypical students but also peers with developmental learning disorder (commonly referred to as dyslexia, dysorthographia, and dyscalculia), as well as peers with ADHD and undiagnosed adolescents with neurodevelopmental or psychiatric conditions present in the population. We had no control over the composition of the non-ASD group—it was determined by the type of examination sheet (1) used during the eighth-grade exam (please see below). It seems reasonable to assume that the frequencies were similar to those in the general population. The sex bias reflects the difference in autism prevalence between boys and girls[3].

In Poland, compulsory education applies to children from the 7th year of age until their 18th birthday. During the exam year, most students (94%) in Poland, including neurodivergent students, attended public schools funded by the state[38]. Some students attended community or private schools, while a small number were homeschooled. The curriculum was consistent across all types of schools. The essays were written as part of the nationwide eighth-grade exam in the Polish language, which takes place in Poland at the conclusion of primary school education. The exam is taken by all students across the country simultaneously, under standardized conditions, and using standardized examination sheets. The sheets are adapted for students with special educational needs. Thus, we distinguish ten versions of examination sheets: (1) for students without disabilities and students with specific learning difficulties; (2) for students with autism, including Asperger's syndrome; (3) for visually impaired students, font size 16 pt; (4) for visually impaired students font size 24 pt; (5) for deaf and hard of hearing students; (6) for students with disorder of intellectual development, mild; (7) for students with aphasia; (8) for students with motor disabilities caused by cerebral palsy; (9) for students whose limited knowledge of the Polish language hinders understanding of the text being read; and (10) since the outbreak of the war in Ukraine, also an examination sheet for students who are citizens of Ukraine.

The students were tasked with writing an essay on a given topic relating to an adventure in the world of a selected mandatory reading from the list. In the case of the examination sheet for students without disabilities and students with specific learning difficulties, the task was: "*Write a story about meeting one of the characters from the required readings list. The shared adventure led you to reflect that it was worthwhile to immerse yourself in the world presented in this literary work. Your essay should demonstrate your good knowledge of the chosen required reading.*". For autistic students, the same task was: *"Imagine that you have the opportunity to travel through time to the world of one of the mandatory readings. Write a story about your adventure in this world. Your essay should demonstrate your good knowledge of the chosen required reading.*" Although the task descriptions varied slightly between groups, human raters (who, when classifying essays, were not familiar with the task instructions), after conducting assessments, reported no noticeable differences between the essays that could be attributed to the slight variations in the wording of the task. A change in task wording could lead to differences in social language use, so we chose to quantify this by comparing personal pronoun usage between the two groups using automated part-of-speech tagging, performed by the Spacy library[39]. This aspect of social language did not prove useful for distinguishing the groups, as the difference in personal pronoun usage (calculated as the proportion of the number of personal pronouns used to the total number of words in the story) was not statistically significant (p value = 0.101). The reasoning behind the experts' decision to use different wording for the same story topic is not publicly known, nor is it obvious. However, we hypothesize that their intention was to phrase the instructions in a slightly simpler and more unambiguous language.

The students were assigned to the ASD Group based on a clinical diagnosis of Autism Spectrum Disorder, confirmed by a psychiatrist and a commission responsible for making decisions regarding the need for special education. The Non-ASD Group consisted of students without sensory and motor impairments, without Statements of Special Educational Needs, or with a diagnosis of a developmental learning disorder.

We received data from the District Examination Boards responsible for storing documentation from the nationwide eighth-grade examination. The Examination Boards were asked to prepare essays that, as much as possible, they reflected a normal distribution of grades among students, but the sampling otherwise being random. We obtained anonymized scans of essays in PDF files along with information about the student's sex, year of birth, and diagnosis (or lack thereof). Most of the scans contained handwritten narratives. However, in 25 cases from the ASD group and in 2 cases from the non-ASD group, the essays were typed on a computer. The essays from PDF files were transcribed into text files, maintaining the original spelling, errors, typos, etc. The paragraph structure was preserved. However, deleted words or text fragments were not considered, and the task instructions were not included.

As the data was analyzed only collectively and was received in anonymized format from the District Examination Boards, which provided it in accordance with the provisions of the Polish Education System Act (https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20240000750) written consent was not required. The project received approval from the Ethics Committee of the University of Warsaw Faculty of Psychology. The authors assert that all procedures contributing to this work comply with the Helsinki Declaration, as revised in 2013.

Evaluation of essays by teachers

As part of the exam, essays written by students were subject to assessment conducted by teachers trained to follow standardized examination guidelines. The assessment included the following categories: *Topic Development, Creative Elements, Literary Skills, Text Composition, Style, Language, Spelling*, and *Punctuation*. The total score for the task was determined by summing the points in each category. The essays did not differ statistically significantly between the two groups with respect to the mentioned categories and the total score, except in the category of *Text Composition*. In the *Text Composition* category, autistic students scored fewer points (M = 1.25, SD = 0.894) than their peers from the non-autistic group (M = 1.51, SD = .750), with a small effect size (U = 13966.50, Z = −2.684, p = 0.007, r = 0.141). Therefore, it appears that, overall, the essays analyzed using the Mann-Whitney U test did not differ significantly in terms of writing competencies assessed in a school-oriented manner by teachers. The basic linguistic and literary characteristics of the narratives are presented in Table 4.

Deep neural models

Distinguishing between essays of autistic and non-autistic students was posed as a binary text classification problem. We tested several models, both generative of decoder architecture and BERT-like encoders. We selected models with support for the Polish language. This section contains their brief overview.

Universal sentence encoder (USE) embeddings

The model is trained and optimized for short texts, such as sentences, phrases, or short paragraphs. It is trained on a variety of data sources and a variety of tasks to dynamically accommodate a wide variety of natural language understanding tasks. The Universal Sentence Encoder (USE) is often applied to text classification, semantic similarity, and information retrieval. We used the multilingual large variant with support for the Polish language. Based on the transformer architecture, it consists of 85M parameters. The input is variable-length text, and the output is a 512-dimensional vector[30]. Since this model is mostly aimed at computing text embeddings, we used the SVM algorithm with a radial kernel as a classifier. We did not attempt to fine-tune the USE model.

OpenAI ada embeddings

We tested the second-generation embedding model from OpenAI referred to as `text-embedding-ada-002`. The embeddings have as many as 1536 dimensions. Unfortunately, no details are provided as to what is the architecture and model training method. Unofficial developer remarks point out that the model is based on GPT-3 with certain elements of GPT-3.5[40]. No paper is provided, and the only explanation describes use cases and API[41]. As was the case with the USE embeddings, we used the SVM algorithm with a radial kernel as a classifier.

LaBSE

Language-agnostic BERT sentence embedding model (LaBSE)[42] supports 109 languages, among them Polish. It is a 12-layer transformer based on the BERT model with a 500k token vocabulary. Unlike BERT, it combines masked language model (MLM) and translation language model (TLM) pretraining with a translation ranking task using bi-directional dual encoders. Specifically, the model is trained and optimized to produce similar representations exclusively for bilingual sentence pairs that are translations of each other. It can be used for mining for translations of a sentence in a multilingual corpus.

HerBERT

HerBERT is a monolingual, Polish-only BERT-based Language Model trained using Masked Language Modelling (MLM) and Sentence Structural Objective (SSO) with dynamic masking of whole words[43]. HerBERT was trained on six different corpora available for the Polish language: CCNet Middle and Head, the National Corpus of Polish, Open Subtitles and Wikipedia in Polish, and Wolne Lektury (a collection of school books). The training dataset was tokenized into subwords using a character-level byte-pair encoding with a vocabulary size of 50k tokens. We used the large variant (`allegro/herbert-large-cased`) of 330M parameters with batch size 4. We decreased the learning rate to 1e-5.

Measure	ASD group M (SD)	Non-ASD group M (SD)	p value	Effect size r
Topic development	1.81 (0.43)	1.76 (0.44)	0.200	0.07
Creative elements	3.11 (1.15)	3.29 (1.13)	0.144	0.08
Literary skills	1.47 (0.68)	1.57 (0.64)	0.175	0.07
Text composition	1.25 (0.89)	1.51 (0.75)	0.007	0.14
Style	1.77 (0.61)	1.77 (0.55)	0.502	0.04
Language	0.92 (1.31)	0.99 (1.31)	0.401	0.04
Spelling	0.80 (0.91)	0.81 (0.90)	0.876	0.01
Punctuation	0.16 (0.37)	0.18 (0.39)	0.625	0.03
Total score	11.28 (4.27)	11.88 (4.18)	0.180	0.07
Tokens	369.49 (135.56)	421.26 (146.65)	< .001	0.18
Characters	1993.08 (705.16)	2243.09 (740.55)	< .001	0.17

Table 4. Linguistic and literary characteristics of narrations.

Polish GPT-2

GPT-2 is a unidirectional transformer-based language model trained with an auto-regressive objective originally introduced in[44]. The original English GPT-2 was released in four sizes differing by the number of parameters: Small (112M), Medium (345M), Large (774M), and XL (1.5B). The capacity of the language model is essential to the success of zero-shot task transfer, and increasing it improves performance in a log-linear fashion across tasks. The largest GPT-2 variant is a 1.5B parameter Transformer that achieves competitive results on multiple language modeling datasets in a zero-shot setting without any task adaptation. Authors demonstrate[44] that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. It contains scraped content from all outbound links from Reddit, which acts as a filter.

We used the Polish GPT-2 model from the Polish NLP resources repository[45]. Unfortunately, no details are provided about training data. The released checkpoints support longer contexts than the original GPT-2 by OpenAI. For example, small and medium models support up to 2048 tokens. To perform text classification, the GPT2 Model transformer was extended with a sequence classification head on top (linear layer), using the last token to do the classification.

Model fine-tuning

It is challenging to finetune large language models for downstream tasks because of the huge number of parameters, reaching hundreds of millions or billions. For this reason, pre-trained language models typically need large amounts of data to learn effectively. Fine-tuning on small datasets is potentially difficult, as it can lead to overfitting and, thus, insufficient generalization capability. To address this issue, we tested two approaches to reducing the number of trained parameters, described below, and compared them to fine-tuning the full model.

Fine-tuning with fixed embeddings

In the first approach, to minimize the number of training parameters, the processing is divided into two parts.

First, a pre-trained language model is used to compute text embedding vectors. It is not fine-tuned, the embeddings are multi-purpose and do not contain autism-specific knowledge. We tested embeddings computed using OpenAI Ada and Universal Sentence Encoder, described in Section 4.3.

Second, a classifier is trained on these embeddings to recognize autism. We used a different type of classifier than gradient-based learning and a neural network, namely a Support Vector Machine (SVM) with a radial kernel. This combination was found well-performing in the context of classifying text embeddings in low-data regime of autism identification from picture book narratives[19].

Parameter-efficient fine-tuning

In the second approach, the adaptation of pre-trained language models to the downstream application of detecting autism is performed by fine-tuning only the selected model's (extra) parameters. Unlike the previous embedding-based approach, where only the parameters of the last classification part of the model were trainable, here fine-tuning can also occur in lower layers. Because the number of trainable parameters remains low, fine-tuning in the low-data regime is likely to be more successful than full model training. Below, we describe two specific techniques used in our experiments, namely P-tuning [46] and Low-Rank Adaptation of Large Language Models (LORA) [47].

The goal of P-tuning [46] is improved usage of prompts to steer the model toward a particular downstream task without fully fine-tuning a model. Typically, the prompts are handcrafted, which may not be practical because it can take a lot of effort to find the best prompts. P-tuning is a method for automatically searching and optimizing for better prompts; it employs trainable continuous prompt embeddings in concatenation with discrete prompts.

The idea behind Low-Rank Adaptation of Large Language Models (LORA) is to freeze the pre-trained model weights and inject trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks[47]. As the model to run LORA, we selected the Polish GPT-2 described in Section 4.3, variant `sdadas/polish-gpt2-large` with 770M parameters. The larger variant, `sdadas/polish-gpt2-xl` with 1,550M parameters, was too large to fit into the 40 GB memory of the A100 Tesla GPU card we used to run the experiments.

We used the GPT-2 specific hyper-parameters provided in Table 11 of the LORA article for the E2E and WebNLG datasets. The only exceptions were setting of warmup steps to 80, as the value of 500 in the paper is tailored to a larger amount of data, and the batch size to 1 due to the available memory of the A100 Tesla GPU. We applied LORA only to the attention heads without affecting other parts of the GPT-2 model.

Full model fine-tuning

This is a standard approach, which involves training all parameters (weights) of the model using the backpropagation technique. If not explicitly stated otherwise, we used the learning rate of 2e-5 and set warmup steps to approximately 20% of the total number of steps. We applied it to train the following models: LaBSE[42], HerBERT large[43], and Polish GPT-2 large[44].

Human raters

To further evaluate the utility of deep learning models and determine whether a slightly different formulation of the task in both groups could have affected the content of the essays, we asked three experts, psychologists experienced in autism diagnosis, to classify the essays. One expert was a researcher, academic teacher and practicing clinician working in the field of autism diagnosis; while the other two were clinicians regularly involved in diagnostic processes for autism. We randomly selected 45 essays, 20 from ASD and 25 from non-ASD group. The experts were unaware of how many essays were drawn from each group. They were asked to

		ASD group *M* (SD)	Non-ASD Group *M* (SD)	*p* value
% of females		25	40	0.289
Teachers' scores	Topic development	1.84 (0.38)	1.68 (0.48)	0.224
	Creative elements	2.95 (0.91)	3.12 (1.09)	0.542
	Literary skills	1.37 (0.83)	1.48 (0.71)	0.726
	Text composition	1.16 (0.96)	1.32 (0.85)	0.607
	Style	1.74 (0.65)	1.76 (0.60)	0.985
	Language	0.63 (0.90)	0.72 (1.02)	0.817
	Spelling	0.89 (0.99)	0.80 (0.96)	0.758
	Punctuation	0.11 (0.32)	0.16 (0.37)	0.604
	Total Score	10.68 (3.64)	11.04 (4.19)	0.803
Tokens		325.79 (113.29)	365.00 (113.17)	0.303
Characters		1772.00 (572.88)	1950.52 (564.85)	0.349

Table 5. Characteristics of narratives classified by human raters.

read the anonymous essays and determine whether a given essay was written by an autistic student or a student from the non-autistic group. The format of the essays was identical, as all essays were transcribed into text files (see the Participants and Data section). Characteristics of the narratives analyzed by the raters are summarized in Table 5.

Data availability

Code (python jupyter notebooks) used in the experiments can be found at https://github.com/alexwz/autism-essays. The data that support the findings of this study are available from the corresponding author, [IC], upon reasonable request. The data are not publicly available due to the absence of participants' consent for their public disclosure.

Received: 5 December 2023; Accepted: 5 June 2025
Published online: 01 July 2025

References

1. Bölte, S., Girdler, S. & Marschik, P. B. The contribution of environmental exposure to the etiology of autism spectrum disorder. *Cell. Mol. Life Sci.* **76**, 12751297. https://doi.org/10.1007/s00018-018-2988-4 (2018).
2. Thapar, A. & Rutter, M. Genetic advances in autism. *J. Autism Dev. Disord.* **51**, 43214332. https://doi.org/10.1007/s10803-020-04685-z (2020).
3. Maenner, M. J. et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years autism and developmental disabilities monitoring network, 11 sites, United States, 2020. *MMWR Surveill. Summ.* **72**, 114. https://doi.org/10.15585/mmwr.ss7202a1 (2023).
4. Jensen, A. R. et al. Modern biomarkers for autism spectrum disorder: Future directions. *Mol. Diagn. Ther.* **26**, 483495. https://doi.org/10.1007/s40291-022-00600-7 (2022).
5. World Health Organization. International statistical classification of diseases and related health problems (2021).
6. Lai, M.-C., Lombardo, M. V. & Baron-Cohen, S. Autism. *Lancet* **383**, 896–910. https://doi.org/10.1016/S0140-6736(13)61539-1 (2014).
7. Parsons, L., Cordier, R., Munro, N., Joosten, A. & Speyer, R. A systematic review of pragmatic language interventions for children with autism spectrum disorder. *Plos One* **12**, e0172242. https://doi.org/10.1371/journal.pone.0172242 (2017).
8. Lam, Y. G. *Pragmatic Language in Autism: An Overview* 533–550 (Springer, New York, New York, NY, 2014).
9. Cummings, L. Pragmatic disorders and theory of mind. *The Cambridge Handbook of Communication Disorders* https://doi.org/10.1017/cbo9781139108683.036 (2013).
10. Baixauli, I., Colomer, C., Rosell, B. & Miranda, A. Narratives of children with high-functioning autism spectrum disorder: A meta-analysis. *Res. Dev. Disabil.* **59**, 234254. https://doi.org/10.1016/j.ridd.2016.09.007 (2016).
11. Bruner, J. The narrative construction of reality. *Crit. Inq.* **18**, 1–21. https://doi.org/10.1086/448619 (1991).
12. Leinonen, E., Letts, C. & Smith, B. *Children's Pragmatic Communication Difficulties* (Whurr, 2000).
13. Zajic, M. C. & Wilson, S. E. Writing research involving children with autism spectrum disorder without a co-occurring intellectual disability: A systematic review using a language domains and mediational systems framework. *Res. Autism Spect. Disord.* **70**, 101471. https://doi.org/10.1016/j.rasd.2019.101471 (2020).
14. Finnegan, E. G. & Accardo, A. L. Written expression in individuals with autism: A meta-analysis. *Curr. Dev. Disord. Rep.* **9**, 178186. https://doi.org/10.1007/s40474-022-00262-4 (2022).
15. Price, J. R., Martin, G. E., Chen, K. & Jones, J. R. A preliminary study of writing skills in adolescents with autism across persuasive, expository, and narrative genres. *J. Autism Dev. Disord.* **50**, 319332. https://doi.org/10.1007/s10803-019-04254-z (2019).
16. Baixauli, I., Rosello, B., Berenguer, C., Tllez de Meneses, M. & Miranda, A. Reading and writing skills in adolescents with autism spectrum disorder without intellectual disability. *Front. Psychol.* **12**, 646849. https://doi.org/10.3389/fpsyg.2021.646849 (2021).
17. Shevchuk-Hill, S., Szczupakiewicz, S., Kofner, B. & Gillespie-Lynch, K. Comparing narrative writing of autistic and non-autistic college students. *J. Autism Dev. Disord.* **53**, 39013915. https://doi.org/10.1007/s10803-022-05516-z (2022).
18. Chojnicka, I. & Wawer, A. Analysis of autistic adolescents essays using computer techniques. *J. Autism Dev. Disord.* https://doi.org/10.1007/s10803-024-06482-4 (2024).
19. Chojnicka, I. & Wawer, A. Detecting autism from picture book narratives using deep neural utterance embeddings. *Int. J. Lang. Commun. Disord.* **57**, 948962. https://doi.org/10.1111/1460-6984.12731 (2022).
20. Wawer, A., Chojnicka, I., Okruszek, L. & Sarzynska-Wawer, J. Single and cross-disorder detection for autism and schizophrenia. *Cogn. Comput.* **14**, 461473. https://doi.org/10.1007/s12559-021-09834-9 (2022).

21. Yeh, A. More accurate tests for the statistical significance of result differences. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics* (2000).
22. Shreffler, J. & Huecker, M. R. *StatPearls, chap* (Sensitivity, Specificity, Predictive Values and Likelihood Ratios (StatPearls Publishing, Diagnostic Testing Accuracy, 2023).
23. Chojnicka, I. & Wawer, A. Social language in autism spectrum disorder: A computational analysis of sentiment and linguistic abstraction. *Plos One* **15**, e0229985. https://doi.org/10.1371/journal.pone.0229985 (2020).
24. Smith-Spark, J. H. & Gordon, R. Automaticity and executive abilities in developmental dyslexia: A theoretical review. *Brain Sci.* **12**, 446. https://doi.org/10.3390/brainsci12040446 (2022).
25. Cicchetti, D. V., Volkmar, F., Klin, A. & Showalter, D. Diagnosing autism using icd-10 criteria: A comparison of neural networks and standard multivariate procedures. *Child Neuropsychol.* **1**, 26–37. https://doi.org/10.1080/09297049508401340 (1995).
26. Petrocchi, S., Levante, A. & Lecciso, F. Systematic review of level 1 and level 2 screening tools for autism spectrum disorders in toddlers. *Brain Sci.* **10**, 180. https://doi.org/10.3390/brainsci10030180 (2020).
27. Hirota, T., So, R., Kim, Y. S., Leventhal, B. & Epstein, R. A. A systematic review of screening tools in non-young children and adults for autism spectrum disorder. *Res. Dev. Disabil.* **80**, 1–12. https://doi.org/10.1016/j.ridd.2018.05.017 (2018).
28. Staton, A., Dawson, D., Moghaddam, N. & McGrath, B. Specificity and sensitivity of the social communication questionnaire lifetime screening tool for autism spectrum disorder in a UK CAMHS service. *Clin. Child Psychol. Psychiatry* **28**, 952–964. https://doi.org/10.1177/13591045221137196 (2023).
29. Hilvert, E., Davidson, D. & Gmez, P. B. Assessment of personal narrative writing in children with and without autism spectrum disorder. *Res. Autism Spect. Disord.* **69**, 101453. https://doi.org/10.1016/j.rasd.2019.101453 (2020).
30. Yang, Y. *et al.* Multilingual universal sentence encoder for semantic retrieval. In Celikyilmaz, A. & Wen, T.-H. (eds.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 87–94, https://doi.org/10.18653/v1/2020.acl-demos.12 (Association for Computational Linguistics, Online, 2020).
31. Lord, C. *et al.Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) Manual (Part I): Modules 1–4* (WPS, 2012).
32. Chojnicka, I. & Pisula, E. Adaptation and validation of the ados-2, polish version. *Front. Psychol.* **8**, 1916. https://doi.org/10.3389/fpsyg.2017.01916 (2017).
33. Goldstein, S. & Ozonoff, S. *Assessment of Autism Spectrum Disorder* (Guilford Publications, 2018).
34. Engberg-Pedersen, E. & Christensen, R. V. Mental states and activities in Danish narratives: Children with autism and children with language impairment. *J. Child Lang.* **44**, 11921217. https://doi.org/10.1017/s0305000916000507 (2016).
35. Mkinen, L. et al. Characteristics of narrative language in autism spectrum disorder: Evidence from the Finnish. *Res. Autism Spect. Disord.* **8**, 987996. https://doi.org/10.1016/j.rasd.2014.05.001 (2014).
36. So, W.-C. et al. Robot-based play-drama intervention may improve the narrative abilities of Chinese-speaking preschoolers with autism spectrum disorder. *Res. Dev. Disabil.* **95**, 103515. https://doi.org/10.1016/j.ridd.2019.103515 (2019).
37. Cola, M. et al. Friend matters: Sex differences in social language during autism diagnostic interviews. *Mol. Autism* **13**, 1. https://doi.org/10.1186/s13229-021-00483-1 (2022).
38. GUS. Oswiata i wychowanie w roku szkolnym 2021/2022 — stat.gov.pl. https://stat.gov.pl/obszary-tematyczne/edukacja/edukacja/oswiata-i-wychowanie-w-roku-szkolnym-20212022,1,17.html. [Accessed 27-01-2025].
39. Honnibal, M., Montani, I., Van Landeghem, S. & Boyd, A. spaCy: Industrial-strength Natural Language Processing in Python, (2020). https://spacy.io, https://doi.org/10.5281/zenodo.1212303
40. OpenAI Developer Community. https://community.openai.com/t/what-version-of-gpt-is-text-embedding-ada-002-based-on/404462 (2023). [Accessed 30-09-2023].
41. OpenAI Platform. https://platform.openai.com/docs/guides/embeddings (2023). [Accessed 30-09-2023].
42. Feng, F., Yang, Y., Cer, D., Arivazhagan, N. & Wang, W. Language-agnostic BERT sentence embedding. In Muresan, S., Nakov, P. & Villavicencio, A. (eds.) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 878–891, https://doi.org/10.18653/v1/2022.acl-long.62 (Association for Computational Linguistics, Dublin, Ireland, 2022).
43. Mroczkowski, R., Rybak, P., Wrblewska, A. & Gawlik, I. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, 1–10 (Association for Computational Linguistics, Kiyv, Ukraine, 2021).
44. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
45. Dadas, S. A repository of polish NLP resources. Github (2019).
46. Liu, X. *et al.* GPT understands, too. *CoRR***abs/2103.10385** (2021). arXiv:2103.10385.
47. Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models. *CoRR***abs/2106.09685** (2021). arXiv:2106.09685.

Acknowledgements

The project was supported by a grant from the Polish National Science Centre (grant no. 2020/39/D/HS6/00809) and also from the funds awarded by the Ministry of Science and Higher Education in the form of a subsidy for the maintenance and development of research potential in 2024 and 2025 (501-D125-01-1250000 zlec.5011000228). The funders had no role in study design, data collection and analysis, decision to publish, or manuscript preparation. We want to express our gratitude for the openness and willingness to cooperate towards the Polish District Examination Boards. In particular, we would like to thank: Dr. Maria Krystyna Szmigel, the Deputy Director of the District Examination Board in Krakow; Mrs. Joanna Peter and Mrs. Anna Rappe from the District Examination Board in Krakow; Mrs. Grażyna Klimuszko, the Deputy Director of the District Examination Board in Łomża; and Mr. Grzegorz Dudzicki from the District Examination Board in Łomża; Mrs. Ragna Ślęzakowska, the Head of the General Education Examinations Department of the District Examination Board in Warsaw; We would like to thank the students who prepared essays for computer analysis: Patrycja Jaworska and Paweł Zubrzycki. We would also like to thank Piotr Stawiński for his support in the conceptualization of the study and analysis of the results.

Author contributions

Conceptualization, methodology, writing, review, and editing; analyzing and discussing the results: IC, AW; design, data collecting and project administration: IC; computational analyses: AW. All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests
The authors declare that there are no financial or non-financial competing interests.

Additional information
Correspondence and requests for materials should be addressed to I.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.