LingVaria
Rok XXI (2026) | 2 (42) | s. 53–70
https://doi.org/10.12797/LV.21.2026.42.04
Licencja: CC BY-NC-ND 4.0

Bartłomiej Alberski
Politechnika Wrocławska, Katedra Sztucznej Inteligencji, Wrocław
bartlomiej.alberski@pwr.edu.pl

Arkadiusz Janz
Politechnika Wrocławska, Katedra Sztucznej Inteligencji, Wrocław
arkadiusz.janz@pwr.edu.pl

Ewa Rudnicka
Politechnika Wrocławska, Katedra Sztucznej Inteligencji, Wrocław
ewa.rudnicka@pwr.edu.pl

Alicja Derych
Uniwersytet Wrocławski, Wydział Neofilologii, Wrocław
alicja.derych@uwr.edu.pl

LINGUISTIC EVALUATION OF LARGE LANGUAGE MODELS: INSIGHTS AND ERROR TYPOLOGIES FROM POLISH MATURA EXAM RESPONSES[1]

Keywords: LLM correctness, LLM evaluation, Polish Matura exam, typology of errors, human evaluation

Słowa klucze: poprawność GMJ, ewaluacja GMJ, matura, typologia błędów, ręczna ewaluacja

1. Introduction

Large language models (LLMs) are currently used in a variety of natural language processing (NLP) tasks, such as, for instance, automatic summarisation, translation, offensiveness detection, humour recognition, spam detection, and word sense disambiguation (see Kocoń, Cichecki, Kaszyca, Kochanek, Szydło, Baran, Bielaniewicz, Gruza, Janz, Kanclerz, Kocoń, Koptyra, Mieleszczenko-Kowszewicz, Miłkowski,

[1] Wkład autorów w powstanie tego tekstu: Bartłomiej Alberski 45%, Ewa Rudnicka 25%, Arkadiusz Janz 10%, Alicja Derych 20%.

Oleksy, Piasecki, Radliński, Wojtasik, Woźniak, Kazienko 2023). By virtue of those multiple usages, they are also applied in engineering and medical research (see Liao, Chen, Du 2023), in social research, political science, and education (Fan, Li, Ma, Lee, Yu, Hemphill 2024). Moreover, LLMs are also employed in the further development of AI such as improvement of infrastructures' performance or further development of LLMs themselves (see Tang, Chuang, Hu 2024; Wu, Yang, Zhan, Yuan, Chao, Wong 2025). Still, the more we use LLMs, the more aware of their limitations we become. We observe that they make language and factual mistakes, hallucinate, or simply lie. When, why and how often they do so depends on the properties of a specific model. Identifying the cases of LLMs' misperformance can be the first step towards repairing them.

In this work we verify the LLMs' performance in the Polish Matura exam essay writing task, inspired by an earlier research by Rafał Mazur (2024). We follow the line of research centring on the linguistic correctness of LLM-generated texts (Weissweiler, Hofmann, Kantharuban, Cai, Dutt, Hengle, Kabra, Kulkarni, Vijayakumar, Yu, Schütze, Oflazer, Mortensen 2023; Korniienko 2024; Lin 2024; Qiu, Duan, Cai 2024), alongside the accuracy of their content. The research on the Polish linguistic correctness of LLMs (Gryka, Gradoń, Kozłowski, Kutyła, Janicki 2024a, 2024b) focuses on the use of the correctness criterion to distinguish between human-generated and model-generated texts.

The release of Chat GPT in 2023 marked a breakthrough in the development of artificial intelligence (AI). The possibility of prompting a model with queries formed in natural language was a game changer. The first LLMs, such as GPT, Llama and Mistral were trained mainly on English language data which explains their poorer performance in tasks provided in other, especially less popular languages such as, for instance, Polish. Polish is underrepresented in the most popular large models, for example, LLaMA 3 contains approximately 0.5 percent of data in Polish (Kozłowski, Karlińska, Tomaszewska, Oleksy 2024). This has motivated the development of 'national' LLMs such as Bielik and the family of PLLuM models. They obviously had greater exposure to Polish data than GPT-4 and LLaMa 3.1 (Karlińska, Kocoń, Piasecki, Pęzik, Kozłowski, Ogrodniczuk, Roszko 2025). Some versions of PLLuM were based on 50% and some on even 100% of Polish data. The one used in this research, namely PLLuM 8x7B, was built as a result of the continuation of the pretraining of the base Mistral model with the Polish data (Kocoń, Piasecki, Janz, Ferdiwnan, Radliński, Koptyra, Oleksy, Woźniak, Walkowiak, Wojtasik, Moska, Naskręt, Walkowiak, Gniewkowski, Szyc, Motyka, Banach, Dalasiński, Rudnicka, Alberski, Walkowiak, Szczęsny, Markiewicz, Bernaś, Mazur, Żyta, Tykierko, Chodak, Kajdanowicz, Kazienko, Karlińska, Seweryn, Kołos, Chrabąszcz, Lorenc, Krasnodębska, Wilczek, Dziewulska, Betscher, Cieślińska, Kowol, Mikoś, Trzciński, Krutul, Kozłowski, Dadas, Poświata, Perełkiewicz, Grębowiec, Kazuła, Białas, Roszko, Roszko, Vaičenonienė, Utka, Levchuk, Kowalski, Prawdzic-Jankowska, Ogrodniczuk, Borys,

Bulińska, Gumienna, Kieraś, Komosińska, Krasnowska-Kieraś, Kobyliński, Lewandowska, Łaziński, Łątkowski, Mastalerz, Milewicz, Mykowiecka, Peljak-Łapińska, Penno, Przybysz, Rudolf, Rybak, Saputa, Tomaszewska, Wawer, Woliński, Wołoszyn, Wróblewska, Żuk, Żarnecki, Kaczyński, Cichosz, Deckert, Garnys, Grabarczyk, Janowski, Karasińska, Kujawiak, Misztela, Szymańska, Walkusz, Siek, Kwiatkowski, Pęzik 2025).

Yet before national LLMs were developed, LLM enthusiasts tested the Chat GPT's performance on tasks in their national languages. One of such studies is Mazur's (2024) work on linguistic correctness of texts generated by ChatGPT 3.5 in response to the Polish language Matura exam task no 18. These are short essays on a specified topic that need to include references to at least some of literary works appearing on the enclosed list of required readings. Mazur analysed the model-generated texts according to the guidelines of the Central Examination Board (CEB) (Buttler, Kurkowska, Satkiewicz 1971; Markowski 2014; Burkacka 2014; Pędzich 2014a, 2014b) and using Andrzej Markowski's (2007) typology of language errors. Mazur noted that most of the errors made by ChatGPT 3.5 were syntactic ones (35%), where the model used incorrect syntactic patterns. The second largest group of errors were lexical ones (24%), manifested in incorrect word combinations and, in some cases, confusion between the meanings of commonly known words. The most common stylistic errors were numerous repetitions (Mazur 2024).

This work is a continuation of Mazur's study. Mazur tested Chat GPT3.5 alone; we have extended his study to newer models such as Bielik 2.3, GPT-4, LLaMA 3.1, and PLLuM 8x7B. We follow Mazur in applying the typology of errors based on A. Markowski's typology (Markowski 2007) and the CEB guidelines for examiners. We have broadened it with the Statutory Regulations for Foreigners Learning Polish[2] and Danuta Buttler's typology of errors (Buttler, Kurkowska, Satkiewicz 1971). Furthermore, we have added another dimension to the evaluation of LLMs' performance, namely the presence and appropriateness of literary references in the model-generated texts, as well as content-based errors related to the specific exam task.

2. Research Method

The research method adopted in this study largely follows the original Mazur's (2024) experiment design and consists of feeding the model with a prompt in the form of the Matura exam task and analysing the model's response using the Central Examination Board instructions. This was done to make the results of our study comparable to Mazur's research. Still, to systematise the analysis of errors we have

[2] https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20160000405 (access: 24 II 2025).

employed a typology of errors compiled on the basis of different classifications present in the literature.

2.1. Prompt Content

The Polish Matura task used as a prompt for an LLM in this study was Task 18: Write a short essay on a provided topic making references to (at least some of) the literary works found in the list of the required readings. Below we provide the translation of the original Polish instruction for the task[3]:

> Do you know these required readings?
> Here, a full list of required readings assigned to this Matura exam task is provided. The list of readings can be found in the Matura exam form[4].
> Model response: **Yes**.
> Choose one of the following topics and write an essay.
> In your essay, analyse the issue presented on the topic. Also, present your own opinion and justify it.
> Support your discussion with arguments, referring to literary works and selected contexts (e.g., historical-literary, literary, biographical, cultural, religious, mythological, biblical, historical, philosophical, existential, political, or social).
> One of the literary works must be a mandatory reading, chosen from the list on pages 3 and 4 of this exam booklet.
> Your essay should be at least 300 words long.
> Topic 1
> The human being, a creature full of contradictions.
> In your essay, refer to the following.
> A selected mandatory reading, an epic, or dramatic work.
> Another literary work – which may also be a poetic piece.
> Selected contexts.

The above prompt was fed to the following models: Bielik 2.3, GPT-4, LLaMA 3.1, and PLLuM 8x7B. The move was repeated 10 times per each model. This gave us a set of 40 model-generated texts[5] for further analysis, while Mazur analysed 25 texts, yet

---

[3] The full prompt in Polish and English is available at https://drive.google.com/drive/folders/18MUgKKoaypzklKSqqUcBR_zyzfueuRSD. The second topic is also available in the task: *What makes a person a hero for another person?* However, we limit the study to the first topic, analogically to Mazur's study.

[4] Matura exam form: https://cke.gov.pl/images/_EGZAMIN_MATURALNY_OD_2023/Arkusze_egzaminacyjne/2023/Jezyk_polski/poziom_podstawowy/MPOP-P2-100-2305.pdf (access: 24 II 2025).

[5] The essays are available here https://drive.google.com/drive/folders/18MUgKKoaypzklKSqqUcBR_zyzfueuRSD.

generated by only one model that was GPT3.5. In our study the number of characters in each generated text ranged from 2024 to 7884, with an average of 3164.25.

2.2. Typology of Errors

In the next stage of our study, we conducted a manual evaluation[6] of texts generated by the chosen models. All the texts were evaluated by two expert linguists and next super-annotated[7]. The texts were evaluated with respect to two aspects: linguistic correctness (Table 1) and thematic appropriateness (Table 2 below).

Table 1. Assessment of linguistic correctness

Assessment of Linguistic Correctness

Word level	Phrase level
Choice of vocabulary	**Adherence to syntax rules**
Adherence to derivation rules	**Adherence to inflection rules (as dependent on syntax)**
	Stylistics

The types of errors distinguished by us fall into two main categories: the errors on the word level and the errors on the phrase level. Insofar as the lexical or multi word expression (MWE) error is limited to a specific language unit[9], the essence of syntactic, inflectional and stylistic errors is based on the relationship between units. For example, a syntactic error of formation or violation of the requirements of recursion or valency is impossible without at least two units entering into a relationship resulting in an error. Similarly, the determination of stylistic or inflectional errors is impossible without several units co-occurring. At a higher level of generality, the

[6] The evaluation instruction is available here https://drive.google.com/drive/folders/18MUgKKoaypzklKSqqUcBR_zyzfueuRSD.

[7] We would like to clarify the analysis process for transparency by explaining that initially, the texts were annotated by a single expert. Following a review, we engaged a new annotator to conduct a second annotation of part of the material, with the aim of calculating consistency and verifying the guidelines (this involved 12 texts, or approximately 30% of the total). During this work, it became clear that we needed to refine the guidelines. After discussing and updating the annotation guidelines and typology of errors, we completed the second annotation across the entire material. With both the old and new annotations in hand, we conducted a super-annotation, and the decisions made by the super-annotator according to the updated guidelines are considered binding. We present these in Table 3.

[8] https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20160000405 (access: 24 II 2025).

[9] We understand MWE as a unit with a non-summary, global meaning of the components included (Bogusławski 1976; Grochowski 1981; Bańko 2001).

essential effects of the presence of an error are the unintelligibility of a passage or the incompatibility of the meaning of a passage with the intention of the sender.

Moving on to the second aspect of evaluation, thematic assessment takes into account the quantity and quality of references to the Polish literary works and Polish culture, see Table 2.

Table 2. Thematic assessment

Thematic Assessment
Polish Cultural Realities
Literary References
Hallucinations
Content and Overall Quality

3. Results of Text Analysis

This section provides a detailed analysis of answers provided by the chosen models to our Matura exam task prompt. They are analysed with respect to all the aspects discussed in the previous section, namely the linguistic correctness of the generated texts, the presence and appropriateness of cultural and literary references, the presence of model hallucinations, and the overall quality of the texts.

3.1. Linguistic Correctness

The summary presented in Table 3 shows large differences in the outputs of the different models. The least number of linguistic errors was made by GPT-4, where, for example, a recurring syntactic error was a faulty construction: *pod wpływem ambicji i manipulacji żony* 'influenced by his wife's ambition and manipulation,' where it is unclear whether it was the wife who manipulated or was manipulated. Only the knowledge of *Macbeth*, where this error occurred, made it possible to determine the correct interpretation. This striking linguistic flawlessness of the model shows a significant increase in quality relative to version 3.5, whose linguistic correctness was verified by R. Mazur (2024). In the sample of texts he analysed, the number of errors ranged from 3 to 16 (ibid.). In contrast, in the sample of 10 texts we analysed, 3 had 2 errors, and the text with the most errors had 8 of them. The second change is a difference in the type of errors. This time, stylistic errors predominated (33,87%).

Table 3. Errors of the models

	Bielik 2.3	GPT-4	LLaMa 3.1	PLLuM 8x7B
In Total	**63**	**62**	**112**	**99**
Word Level Errors	**18**	**8**	**36**	**29**
Choice of vocabulary	**18**	**8**	**34**	**29**
Adherence to derivation rules	**0**	**0**	**3**	**0**
Phrase Level Errors	**45**	**54**	**76**	**70**
Adherence to syntax rules	**23**	**13**	**36**	**16**
Adherence to inflection rules	**7**	**2**	**10**	**0**
Stylistics	**15**	**21**	**30**	**54**

The next model was Bielik. Syntactic errors constituted the majority of errors generated by the model in the texts such as:

> *Dante ukazuje człowieka jako istotę skazaną na ciągłą walkę między dobrem a złem, między cnotą a* ***grzechem. Jego*** *wędrówka przez zaświaty symbolizuje ludzkie zmagania z własnymi słabościami i dążenie do doskonałości.*
>
> 'Dante portrays man as a being condemned to a constant struggle between good and evil, between virtue and **sin. His** journey through the afterlife symbolizes humanity's struggle with its own weaknesses and the pursuit of perfection.'

The pronoun *jego* – 'his' anaphorically refers to the noun *grzechem* – 'sin' in the previous sentence. While the remaining part of this fragment assures that the intention was to reference *człowiek* 'man'. Another interesting lexical error that appeared in the text was: *Ta narracja ukazuje* – 'This narrative shows' (should be: *Ta historia/fabuła ukazuje* – 'This story/plot shows'). Though *narracja* – 'narrative' is a word close to *historia* – 'story' and *fabuła* – 'plot,' they have a different reference. Bielik also created redundant constructions several times, i.e., stylistically incorrect ones, such as: *z poczuciem winy i wyrzutami sumienia* – 'with guilt and remorse.' Another stylistic error that occurred was repetition, e.g.

> *Tytułowy bohater, kierowany ambicją i żądzą władzy, decyduje się na zamordowanie króla Duncana, aby samemu objąć tron. Jego postępowanie jest motywowane chęcią osiągnięcia sukcesu za wszelką cenę, co jednak* ***prowadzi*** *do tragicznych konsekwencji. Makbet staje się ofiarą własnych ambicji, które* ***doprowadzają*** *go do paranoi i obsesji, co w końcu* ***prowadzi*** *do jego upadku. Ten utwór ukazuje, jak silne emocje i pragnienia mogą zdominować racjonalne myślenie i* ***prowadzić*** *do autodestrukcji.*

'The title character, driven by ambition and lust for power, decides to murder King Duncan in order to take the throne himself. His actions are motivated by a desire to succeed at all costs, which, however, **leads** to tragic consequences. Macbeth becomes a victim of his own ambition, which **leads** him to paranoia and obsession, which eventually **leads** to his downfall. This piece shows how strong emotions and desires can dominate rational thinking and **leads** to self-destruction.'

PLLuM made the most stylistic errors (54), and, similarly as Mazur's research demonstrated (ibid.), most of them were repetition, such as:

W historii ludzkości można odnaleźć wiele przykładów, ***które potwierdzają tę tezę****. Ludzie często są rozdarci między różnymi pragnieniami, wartościami i celami, co prowadzi do konfliktów wewnętrznych i zewnętrznych. W mojej pracy przedstawiłem kilka przykładów z literatury,* ***które potwierdzają tę tezę****.*

'There are many examples in the history of mankind that **confirm this thesis**. People are often torn between different desires, values and goals, which leads to internal and external conflicts. In my work, I have presented several examples from the literature, **which confirm this thesis**.'

It is worth noting that the high number of stylistic errors involving repetitions made by the PLLuM model may be related to the length of the texts. The texts generated by the model were the longest, averaging 3612.3 characters, with an overall average of 3164.25. This model also generated the longest text with 7884 characters.

PLLuM made also syntactic errors, and interestingly, as many as four times they were identical to the most popular ones made by GPT-4, i.e., *pod wpływem przepowiedni i manipulacji swojej żony, Lady Makbet* – 'under the influence of the prophecy and manipulation of his wife, Lady Macbeth' (without the knowledge of the text, it is not clear who is manipulating whom). An inflectional error *w człowiek jest* 'in man is' (instead of: *w człowieku jest* 'in man is').

Finally, the model that made the biggest number of language errors was LLaMA (112 errors). Syntactic errors dominated (36), which included:

1) word order, e.g., *nie jest ona wynikiem jednej tylko perspektywy* – 'it is not the result of only one perspective' (unmarked word order is: *nie jest ona wynikiem tylko jednej perspektywy* – 'it is not the result of only one perspective');
2) connectivity, e.g., *Człowiek jest istotą, która posiada wiele różnych cech* ***i które*** *są w ciągłym konflikcie*. – 'Man is a creature which has many different characteristics **and which** are in constant conflict.' (two different conjunctions occured: *i* and *które*);

3) syntax, e.g., *popełnia morderstwo starej lichwiarki* – 'murders the old usurer' (without knowing the context, it is not clear whether the usurer murders or is murdered);
4) accommodation, e.g. *Innym utworem, który przedstawia sprzeczności ludzkiej natury,* ***jest "Dziady"*** *Adama Mickiewicza.* – 'Another work that presents the contradictions of human nature **is "Dziady"** by Adam Mickiewicz.' (*Dziady* is plural, it should be: *są Dziady*).

There were also derivational errors in these texts, e.g. *kompromituowaną* 'compromised' (*kompromitowaną* 'compromised'), *mądrzym* 'wisdom' (*mądrzy* 'wise') *Achillesh* ('Achilles'). There was also no shortage of inflectional errors, such as *zostali wyrzucone* 'they were thrown away' (*zostali wyrzuceni* 'they were thrown away'), *popełnienia okropnego zbrodni* 'committing a terrible crime' (*popełnienia okropnej zbrodni* 'committing a terrible crime'), lexical errors, e.g. *Jak możemy zmienić się samych siebie i świata* 'how we can change ourselves and the world' (*zmienić się* 'change' instead of *zmienić* 'change'), and stylistic errors, e.g. *inteligentna i głupia* 'intelligent and stupid,' *Jego natura jest złożona i składa się z wielu części, które często są ze sobą sprzeczne.* 'His nature is complex and consists of many parts that often contradict each other.'

3.2. Literary and Cultural References

We would also still like to describe some important observations about literary and cultural references, see Table 4. By cultural references, we mean the evocation of claims and attitudes disseminated by specific philosophers and artists throughout history, see Table 5. The fewest references to literary works appeared in GPT4-generated texts. There were 10 works of literary sources (exclusively non-Polish) referenced 20 times. Dostojewski's *Zbrodnia i kara* – "Crime and Punishment" was cited most often by the model. That, by the way, happened to be the most frequently cited source overall (35 references in all texts generated by the models). The second most cited work was *Makbet* – "Macbeth" (25 references in all texts generated by the models). The largest number of Polish sources was cited by PLLuM (16 sources), which at the same time cited the most sources out of all of the models (25) and quoted them the highest number of times, i.e., 39 citations. The second highest number of cited sources and the number of citations at the same time was provided by Bielik, i.e. 24 citations in 23 sources. The detailed data is provided in Table 5 and Table 6.

Table 4: Literary References and Hallucinations

	Bielik 2.3	GPT-4	LLaMa 3.1	PLLuM 8x7B
Polish Literary References	3	0	4	16
Foreign Literary References	20	8	12	9
Hallucinations	4	0	9	3

Table 5: Literary references

Author	Title	Bielik 2.3	GPT-4	LLaMA 3.1	PLLuM 8x7B
Homer	Iliada	1	0	3	0
Homer	Odyseja	0	0	2	0
Homer	Dzieje	0	0	1	0
Sofokles	Antygona	0	0	0	2
św. Augustyn	Wyznania	0	0	1	0
G. Herling-Grudziński	Inny Świat	0	0	0	2
S. Mrożek	Tango	0	0	0	1
G. Orwell	Rok 1984	0	0	0	1
H. Krall	Zdążyć przed Panem Bogiem	0	0	0	1
O. Tokarczuk	Profesor Andrews w Warszawie	0	0	0	1
A. Stasiuk	Miejsce	0	0	0	1
J. Dukaj	Katedra	0	0	0	1
M. Nowakowski	Góra Edek	0	0	0	1
J. Słowacki	Kordian	0	0	0	1
R.L. Stevenson	Dr Jekyll, Mr Hyde	0	0	0	1
J. Słowacki	Balladyna	0	0	1	0
M. Nowakowski	Raport o stanie wojennym	0	0	1	0
C.K. Norwid	Confiteor	0	0	0	1
K. Moczarski	Rozmowy z katem	0	0	0	1
Z. Nałkowska	Granica	0	0	0	2
C.K. Norwid	Koniec wieku XIX	0	0	0	1
H. Sienkiewicz	Dzieje	0	0	1	0
A. Mickiewicz	Konrad Wallenrod	0	0	2	2
W. Szymborska	Pod jedną kręgą	1	0	0	0

Author	Title	Bielik 2.3	GPT-4	LLaMA 3.1	PLLuM 8x7B
W. Shakespeare	Hamlet	1	0	2	3
F. Dostojewski	Zbrodnia i kara	9	10	6	10
A. Camus	Dżuma	5	0	0	1
D. Alighieri	Boska komedia	1	0	0	0
F. Kafka	Proces	5	0	0	1
A. Mickiewicz	Dziady	2	0	1	3
W. Shakespeare	Makbet	3	10	6	6
B. Prus	Lalka	1	0	0	4
Total		24	10	19	39

Table 6: Cultural references

Cultural references	Bielik 2.3	GPT-4	LLaMa 3.1	PLLuM 8x7B
F. Nietzsche	1	4	0	0
Sokrates	3	0	0	0
Arystoteles	1	0	4	0
A. Schopenhauer	1	0	0	0
C.G. Jung	1	0	0	0
J.P. Sartre	2	4	0	0
I. Kant	2	0	0	0
J.J. Rousseau	1	0	0	0
S. Kierkegaard	0	2	0	0
Platon	1	1	3	0
Kartezjusz	0	1	0	0
S. Freud	0	0	2	0
Michał Anioł	1	0	0	0
Biblia	5	5	4	4
Mitologia grecka	1	0	1	0
Total	17	17	11	4

3.3. Hallucinations

When analysing the generated texts for linguistic correctness, our attention was also drawn to factual errors. These include hallucinatory errors, where the model cites a non-existent literary work, e.g. Bielik cites a non-existent poem by Polish Nobel Prize winner Wisława Szymborska *Pod jedną kręgą* – "Under One Circle," whose

title is similar to an existing poem by the poet: *Pod jedną gwiazdką* – "Under One Star." A very interesting mistake is LLaMA model's citation of Henryk Sienkiewicz's non-existent work *Dzieje* – "History"[10]. In the context of the non-existent *Dzieła* – "Works," the model invokes Andrzej Kmicic, the protagonist of Sienkiewicz's existing novel: *Potop* – "The Deluge." The PLLuM model invoking the title character of *Konrad Wallenrod* claims that Konrad decides to betray his nation and go over to the side of the enemy in order to save his homeland from annihilation, while the whole story presented by Adam Mickiewicz in *Konrad Wallenrod* presents the completely opposite course of events. Another example of a factual error is the Bielik model's pointing out that Job, a righteous and God-fearing man, faces great hardships and suffering that undermine his faith and trust in God (Job remained faithful to God). An interesting error appeared in a passage of the text generated by the LLaMA model, which was not thematically connected to the entire statement on the topic indicated in the task:

> *W kontekście religijnym, możemy odnieść się do chrześcijańskiej doktryny o grzechu pierworodnym, która mówi, że człowiek jest od urodzenia skażony grzechem i że jego natura jest skłonna do popełniania złych czynów.*
>
> 'In a religious context, we can refer to the Christian doctrine of original sin, which says that man is tainted with sin from birth and that his nature is inclined to commit evil deeds.'

According to this doctrine, man needs redemption and salvation to achieve happiness and perfection. Examples of non-literary and non-hallucinatory factual errors concerning non-existent texts are, for example, the LLaMA model's claim that Plato believed that humans consist of three parts: reason, soul, and body, whereas Plato actually spoke of the duality of human nature. It is worth noting that GPT4 did not make any such mistakes.

3.4. Content

In one case, one of the models made an additional reference to Polish cultural realities without reference to a specific artist or work. The PLLuM model alluded to the experience of World War II, which profoundly affected the Polish nation and is still an unprocessed trauma. The model noted that:

[10] This kind of error may be on the other hand a factual contamination with *Dzieje* – "Histories" by Herodotus.

Sprzeczności w naturze człowieka są również widoczne w kontekście historycznym i społecznym. Przykładem może być okres II wojny światowej, kiedy to ludzie byli zmuszeni do podejmowania decyzji i działań, które były sprzeczne z ich wartościami i przekonaniami. Wielu ludzi musiało wybierać między życiem a śmiercią, między dobrem a złem, między lojalnością a zdradą. Ich sprzeczne uczucia i myśli były wynikiem sytuacji, w której się znaleźli, a ich działania miały tragiczne konsekwencje.

'Contradictions in human nature are also evident in historical and social contexts. An example is the period of World War II, when people were forced to make decisions and take actions that contradicted their values and beliefs. Many people had to choose between life and death, between good and evil, between loyalty and betrayal. Their conflicting feelings and thoughts were the result of the situation in which they found themselves, and their actions had tragic consequences.'

Another interesting example was an excerpt from the text generated by GPT-4, in which the model began to spin its own reflection:

Moim zdaniem, sprzeczności w ludzkiej naturze są nieodłącznym elementem naszego istnienia. To one sprawiają, że jesteśmy istotami dynamicznymi, zdolnymi do refleksji i rozwoju. Wewnętrzne konflikty, choć często bolesne, prowadzą nas do głębszego zrozumienia siebie i świata. Literatura, ukazując te sprzeczności, pozwala nam lepiej zrozumieć naszą własną naturę i uczy nas, że pomimo wewnętrznych konfliktów, możemy dążyć do harmonii i równowagi.

'In my opinion, the contradictions in human nature are inherent in our existence. They are what make us dynamic beings, capable of reflection and development. Internal conflicts, though often painful, lead us to a deeper understanding of ourselves and the world. Literature, by showing these contradictions, allows us to better understand our own nature and teaches us that despite internal conflicts, we can strive for harmony and balance.'

3.5. Overall Quality

The greatest schematicism of construction and argumentation applies to the GPT-4 model's responses. The small number of literary references and rather poor vocabulary make up for the considerable repetitiveness of the generated texts.

The LLaMA model, on the other hand, showed considerable variability, as each text was decidedly different, containing different endings and beginnings. At the same time, it is worth remembering that the texts generated by LLaMA contained a very large number of errors of various types.

The Bielik model's responses are characterised by a certain schematism visible especially in the construction of the text, the so-called openings and closings, i.e., beginnings and endings. Each time the texts open with the title of the task: *Człowiek – istota pełna sprzeczności* – 'Man – a creature full of contradictions,' and the ending always begins with the word *Podsumowując* 'Summing up.' However, the numerous literary references force varied argument constructions.

The PLLuM model showed great structural diversity, a rich vocabulary, and a large number of literary references. At the same time, it is worth noting that the references were made in various ways, i.e.

1) by the referenced characters, e.g. *Warto również zwrócić uwagę na postać Zenona Ziembiewicza z powieści "Granica" Zofii Nałkowskiej.* – 'It is also worth noting the character Zenon Ziembiewicz from the novel "The Frontier" by Zofia Nałkowska.'
2) titles of works, e.g. *Na przykład w dramacie Sofoklesa "Antygona" tytułowa bohaterka staje przed wyborem między posłuszeństwem wobec prawa państwowego a wiernością wobec prawa boskiego.* – 'For example, in Sophocles' drama "Antigone" the title character is faced with a choice between obedience to state law and loyalty to divine law.'
3) described problem, e.g. *Jednym z najbardziej znanych przykładów sprzeczności ludzkiej natury jest postać doktora Jekylla i pana Hyde'a z noweli Roberta Louisa Stevensona.* – 'One of the most famous examples of the contradictions of human nature is the character of Dr. Jekyll and Mr. Hyde from Robert Louis Stevenson's novella.'
4) described events, e.g. *Józef K. jest urzędnikiem, który zostaje oskarżony o nieokreślone przestępstwo i musi stawić czoła absurdalnemu systemowi prawnemu.* – 'Joseph K. is an official who is accused of an unspecified crime and has to face an absurd legal system.'

4. Limitations

This study should be regarded as an experiment in a manual LLM evaluation rather than a systematic approach to model evaluation. The analysis of linguistic correctness and cultural references should be included as a permanent part of the evaluation of LLMs, especially in the context of evaluating their quality. The observed results confirm our initial intuitions: models trained on Polish-language data perform much better in terms of cultural embedding, as evidenced by the PLLuM model. On the other hand, a positive surprise is GPT-4, whose linguistic correctness may not be related to the training data but to a large number of Polish-language queries handled. However, this issue requires a separate analysis.

The observed schematism and poor vocabulary of some of the essays may be related to the temperature of the LLM in question, which would also require in-depth research.

Implementing the typology of errors developed could enable automation of certain evaluation processes in the future. Another method of correctness analysis could be language tests for foreigners, but this approach is beyond the topic of this article.

5. Conclusions

The purpose of this study was to manually verify the linguistic correctness of four LLMs: Bielik 2.3, GPT-4, LLaMA 3.1 and PLLuM 8x7B. A total of 40 essays (10 for each model) generated by LLMs were analysed. An original classification of errors (see Table 1) was used based on the existing proposals and the guidelines of the Central Examination Board. In addition, the quantity and quality of references to literary and cultural works applicable to the Polish Matura exam were analysed, according to the content of the task that forms part of the prompt (see section *Prompt Content*). The texts showed significant progress in the quality of the model output, as syntactic and lexical errors decrease from 59% in GPT3 generated texts (Mazur 2024) to 33.87% in GPT4 (while stylistic errors increase, see Table 3). In contrast, for literary and cultural references, GPT4 did not refer to Polish sources at all, and cited the fewest number of sources in general (i.e. 8 texts; see Table 4). By far, the Polish models Bielik and PLLuM referred to the largest number of sources, including Polish readings (23 and 25 references, respectively). In addition, PLLuM was the only model to cite a cultural context not explicitly linked to any text or author (see section *Content*). In one of the essays, PLLuM referred to the context of World War II, which is unique among all the essays that were often schematic (see section *Overall Quality*).

It is certainly worth considering the development of the study by increasing not only the number of texts and models but also by analysing the temperature of the model and the quality of its responses separately, depending on the set parameters. One could also compare the evaluation by hand and by LLM-judge. Another method could also be the use of foreign language tests. However, this study is intended as a contribution to the on-going discussion on LLM evaluation methods in the first place.

Acknowledgments

The work was created as part of the project entitled: "Contribution to the European research infrastructure CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure" for the years 2024-2026, co-financed by the Minister of Science on the basis of agreement no. 2024/WK/01 as part of the programme "Support for the participation of Polish research teams in international research infrastructure projects".

Literature

Bańko M., 2001, *Z pogranicza leksykografii i językoznawstwa. Studia o słowniku jednojęzycznym*, Warszawa.

Bogusławski A., 1976, *O zasadach rejestracji jednostek języka*, „Poradnik Językowy" 8, pp. 356–364.

Burkacka I., 2014, *Błędy składniowe*, Centralna Komisja Egzaminacyjna, [on-line:] https://www.oke.krakow.pl/inf/filedata/files/Ra%BF%B1ce_b%B3%EAdy_j%EAzykowe_2_Sk%B3ad.pdf (access: 2 X 2025).

Buttler D., Kurkowska H., Satkiewicz H., 1971, *Kultura języka polskiego. Zagadnienia poprawności gramatycznej*, Warszawa.

Fan L., Li L., Ma Z., Lee S., Yu H., Hemphill L., 2024, *A Bibliometric Review of Large Language Models Research from 2017 to 2023*, „ACM Transactions on Intelligent Systems and Technology" vol. 15, iss. 5, pp. 1–25, https://doi.org/10.1145/3664930.

Grochowski M., 1981, *O wyróżnianiu jednostek opisu semantyki leksykalnej*, „Sborník Prací Filosofické Fakulty Brněnské University" vol. 30, iss. A29, pp. 31–37.

Gryka P., Gradoń K., Kozłowski M., Kutyła M., Janicki A., 2024a, *Detection of AI-Generated Emails – A Case Study*, [in:] *Proceedings of the 19th International Conference on Availability, Reliability and Security*, pp. 1–8, https://doi.org/10.1145/3664476.3670465.

Gryka P., Gradoń K., Kozłowski M., Kutyła M., Janicki A., 2024b, *Impact of Spelling and Editing Correctness on Detection of LLM-Generated Emails*, [in:] *Proceeding of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, vol. 39, pp. 603–608, https://doi.org/10.15439/2024F8906.

Karlińska A., Kocoń J., Piasecki M., Pęzik P., Kozłowski M., Ogrodniczuk M., Roszko R., 2025, *Zakończenie projektu PLLuM*, [on-line:] https://pllum.org.pl/blog/posts/zakonczenie-projektu-pllum (access: 10 III 2026).

Kocoń J., Cichecki I., Kaszyca O., Kochanek M., Szydło D., Baran J., Bielaniewicz J., Gruza M., Janz A., Kanclerz K., Kocoń A., Koptyra B., Mieleszczenko-Kowszewicz W., Miłkowski P., Oleksy M., Piasecki M., Radliński Ł., Wojtasik K., Woźniak S., Kazienko P., 2023, *ChatGPT: Jack of All Trades, Master of None*, „Information Fusion" vol. 99, art. 101861, https://doi.org/10.1016/j.inffus.2023.101861.

Kocoń J., Piasecki M., Janz A., Ferdinan T., Radliński Ł., Koptyra B., Oleksy M., Woźniak S., Walkowiak P., Wojtasik K., Moska J., Naskręt T., Walkowiak B., Gniewkowski M., Szyc K., Motyka D., Banach D., Dalasiński J., Rudnicka E., Alberski B., Walkowiak T., Szczęsny A., Markiewicz M., Bernaś T., Mazur H., Żyta K., Tykierko M., Chodak G., Kajdanowicz T., Kazienko P., Karlińska A., Seweryn K., Kołos A., Chrabąszcz M., Lorenc K., Krasnodębska A., Wilczek A., Dziewulska K., Betscher P., Cieślińska Z., Kowol K., Mikoś D., Trzciński M., Krutul D., Kozłowski M., Dadas S., Poświata R., Perełkiewicz M., Grębowiec M., Kazuła M., Białas M., Roszko R., Roszko D., Vaičenonienė J., Utka A., Levchuk P., Kowalski P., Prawdzic-Jankowska I., Ogrodniczuk M., Borys M., Bulińska A., Gumienna W., Kieraś W., Komosińska D., Krasnowska-Kieraś K., Kobyliński Ł., Lewandowska M., Łaziński M., Łątkowski M., Mastalerz D., Milewicz B., Mykowiecka A.A., Peljak-Łapińska A., Penno S., Przybysz Z., Rudolf M., Rybak P., Saputa K., Tomaszewska A., Wawer A., Woliński M., Wołoszyn J., Wróblewska A., Żuk B., Żarnecki F., Kaczyński K., Cichosz A., Deckert Z., Garnys M., Grabarczyk

I., Janowski W., Karasińska S., Kujawiak A., Misztela P., Szymańska M., Walkusz K., Siek I., Kwiatkowski J., Pęzik P., 2025, *PLLuM: A Family of Polish Large Language Models*, arXiv preprint arXiv:2511.03823, https://doi.org/10.48550/arXiv.2511.03823.

Korniienko O., 2024, *Enhancing Grammatical Correctness: The Efficacy of Large Language Models in Error Correction Task*, Lviv, (master thesis), [on-line:] https://er.ucu.edu.ua/server/api/core/bitstreams/e7fab636-8f8e-4342-a892-4b510773fce4/content (access: 2 X 2025).

Kozłowski M., Karlińska A., Tomaszewska A., Oleksy M., 2024, *Polski duży model językowy*, „IT w Administracji", iss. 7–8 (200–201), pp. 60–63 [on-line:] https://itwadministracji.pl/wydanie/lipiec-sierpien-2024/ (access: 12 I 2025).

Liao J., Chen X., Du L., 2023, *Concept Understanding in Large Language Models*: *An Empirical Study*, The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1–5, 2023, [on-line:] https://openreview.net/pdf?id=losgEaOWIL7 (access: 10 III 2026).

Lin S., 2024, *Evaluating LLMs' Grammatical Error Correction Performance in Learner Chinese*, „PloS ONE" 19, (10):e0312881, [on-line:] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0312881 (access: 2 X 2025), https://doi.org/10.1371/journal.pone.0312881.

Markowski A., 2007, *Kultura języka polskiego. Teoria, zagadnienia leksykalne*, Warszawa.

Markowski A., 2014, *Błędy leksykalne, frazeologiczne i słowotwórcze*, [materiały pomocnicze], Centralna Komisja Egzaminacyjna, [on-line:] https://www.oke.krakow.pl/inf/filedata/files/Ra%BF%B1ce_b%B3%EAdy_j%EAzykowe_1_LeksFrazSlow.pdf (access: 2 X 2025).

Mazur R., 2024, *O poprawności językowej tekstów generowanych przez SI na przykładzie ChatuGPT*, „LingVaria" XIX, nr 1 (37), pp. 119–138, https://doi.org/10.12797/LV.19.2024.37.08.

Pędzich B., 2014a, *Błędy fleksyjne*, [materiały pomocnicze], Centralna Komisja Egzaminacyjna, [on-line:] https://www.oke.krakow.pl/inf/filedata/files/Ra%BF%B1ce_b%B3%EAdy_j%EAzykowe_3_Fleks.pdf (access: 2 X 2025).

Pędzich B., 2014b, *Błędy stylistyczne*, [materiały pomocnicze], Centralna Komisja Egzaminacyjna, [on-line:] https://www.oke.krakow.pl/inf/filedata/files/Ra%BF%B1ce_b%B3%EAdy_j%EAzykowe_4_Styl.pdf (access: 2 X 2025).

Qiu Z., Duan X., Cai Z., 2024, *Evaluating Grammatical Well-Formedness in Large Language Models: A Comparative Study with Human Judgments*, [in:] *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 189–198, [on-line:] https://aclanthology.org/2024.cmcl-1.16.pdf (access: 2 X 2025), https://doi.org/10.18653/v1/2024.cmcl-1.16.

Tang R., Chuang Y.-N., Hu X., 2024, *The Science of Detecting LLM-Generated Text*, „Communications of the ACM" nr 67 (4), pp. 50–59, [on-line:] https://dl.acm.org/doi/pdf/10.1145/3624725 (access: 2 X 2025), https://doi.org/10.1145/3624725.

Weissweiler L., Hofmann V., Kantharuban A., Cai A., Dutt R., Hengle A., Kabra A., Kulkarni A., Vijayakumar A., Yu H., Schütze H., Oflazer K., Mortensen D., 2023, *Counting the Bugs in ChatGPT's Wugs: A Multilingual Investigation into the Morphological Capabilities of a Large Language Model*, arXiv preprint arXiv:2310.15113, [on-line:] https://arxiv.org/pdf/2310.15113 (access: 2 X 2025), https://doi.org/10.18653/v1/2023.emnlp-main.401.

Wu J., Yang S., Zhan R., Yuan Y., Chao L., Wong D., 2025, *A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions*, „Computational Linguistics" 51 (1), pp. 275–338, https://doi.org/10.1162/coli_a_00549.

Web sources

https://cke.gov.pl/images/_EGZAMIN_MATURALNY_OD_2023/Arkusze_egzaminacyjne/2023/Jezyk_polski/poziom_podstawowy/MPOP-P2-100-2305.pdf (access: 24 II 2025).

https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20160000405 (access: 24 II 2025).

Linguistic Evaluation of Large Language Models: Insights and Error Typologies From Polish Matura Exam Responses

Abstract

The article presents a manual analysis of texts generated by four large language models (LLMs): Bielik 2.3, GPT-4, LLaMA 3.1, and PLLuM 8x7B. The texts are in Polish and the text topic is based on the Polish Matura exam. This research extends previous work by Rafał Mazur (2024), who tested only one LLM (Chat GPT 3.0) and had not formulated his own typology of errors. In the article, a typology of language errors formulated for the purposes of this study is used, based on existing solutions (Buttler 1971; Markowski 2007, 2014; Burkacka 2014; Pędzich 2014a, 2014b), as well as quantitative and qualitative criteria concerning the model's references to readings and cultural contexts. The study found that although GPT-4 made the fewest linguistic errors and did not hallucinate, it referred to the fewest number of texts, including no Polish readings or Polish cultural contexts. The PLLuM 8x7B, on the other hand, referred to the largest number of Polish readings of all models tested.

Ocena językowa generatywnych modeli językowych: analiza i typy błędów w wygenerowanych wypracowaniach maturalnych

Abstrakt

Artykuł prezentuje ręczną analizę tekstów wygenerowanych przez cztery generatywne modele językowe (GMJ): Bielik 2.3, GPT-4, LLaMA 3.1, i PLLuM 8x7B. Generowane teksty są w języku polskim i stanowią realizację tematu zadania maturalnego z języka polskiego. Całość badania stanowi rozszerzenie pracy Rafała Mazura (2024), który przetestował wyłącznie jeden GMJ (Chat GPT 3.0) i nie zaproponował własnej typologii błędów językowych. W artykule wykorzystano sformułowaną na potrzeby tego badania typologię błędów językowych bazującą na dotychczasowych rozwiązaniach (Buttler 1971; Markowski 2007, 2014; Burkacka 2014; Pędzich 2014a, 2014b) oraz liczbowe i jakościowe kryteria dotyczące odniesień modelu do lektur oraz kontekstów kulturowych. Wskutek analizy ustalono, że najmniej błędów językowych popełnił GPT-4, który również nie halucynował. Jednocześnie ten model odniósł się do najmniejszej liczby lektur, w tym nie przywołał żadnej lektury w języku polskim oraz żadnego polskiego kontekstu kulturowego. Z kolei PLLuM 8x7B odniósł się do największej liczby lektur w języku polskim ze wszystkich GMJ.