Culture and Language in Human and AI-Generated Writing: A Cross-Linguistic Study

Kulzat Sadirova[1], Kuralay Tashpenova[1*], Mairagul Kushtayeva[1], Mukhtar Mirov[1], Raushan Zhazykova[1], Nazym Nossiyeva[2]
[1]K. Zhubanov Aktobe Regional University, Kazakhstan, [2]Kazakh National University of Sports, Kazakhstan

Abstract Language models write in many languages, but most comparisons of student and machine writing concern English and ask whether the two can be told apart. This exploratory study described how essays by students in Kazakhstan differ from essays by two language models on the same topics in Kazakh, Russian, and English. The corpus comprised 77 student essays and 180 essays by GPT-4o and Llama-3.1-Sherkala-8B-Chat, coded for cultural grounding, code-switching, and three formal features. At least one of four grounding features (a proverb, a named cultural figure or quotation, a time-anchored personal episode, or a named local place) occurred in 53.2% of student essays and 17.8% of machine essays; the difference was clear in Kazakh, weaker in Russian, and absent in English, in samples differing in topic and context. Closing formulae were more frequent in machine essays only in English. Under these prompts, neither model quoted a Kazakh proverb; the models exceeded the students in lexical diversity, but not in grounding.

Keywords: *AI-generated text, Cultural grounding, Discourse analysis, Kazakh-Russian bilingualism, Proverbs*

1. Introduction

Since generative language models became freely available, the question most often asked about a student text has become whether a person wrote it. Much research compares human and machine writing with a view to detection (Reinhart et al., 2025; Weber-Wulff et al., 2023), yet detection has proved unreliable and biased against non-native writers of English (Liang et al., 2023). The problem is acute in Kazakhstan, where Kazakh, Russian, and English share everyday, academic, and professional life (Koptleuova et al., 2023; Tlepbergen et al., 2025).

An alternative to classification is description, which shows what a model does with a writing task without deciding who wrote a particular essay. Studies that take this route (Berber Sardinha, 2024; Jiang & Hyland, 2025a, 2025b; Zhang & Crosthwaite, 2025) mostly analyse English, including comparisons with second-language writers (Fredrick & Craven, 2025), and count lexical,

***Corresponding Author:**
Kuralay Tashpenova
kuralai.tashpenova@mail.ru

Received: July 2, 2026
Revised: September 25, 2026
Accepted: October 20, 2026
First Published: November 10, 2026

© 2026 Sadirova, Tashpenova, Kushtayeva, Mirov, Zhazykova, and Nossiyeva.
This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

https://doi.org/10.22034/ijscl.2026.2090294.4741

grammatical, or metadiscursive features; proverbs, quotations, named places, and remembered episodes have rarely been examined. Kazakh-oriented models such as Sherkala-8B, in turn, have been evaluated on benchmark tasks, not on student-type writing (Koto et al., 2025). No study known to us has compared student and machine essays in Kazakh, Russian, and English with attention to culturally grounded discourse.

The objective of the present study is to describe how essays by first-year university students writing in Kazakh, Russian, and English differ from essays by two language models, the general commercial model GPT-4o and the Kazakh-oriented Llama-3.1-Sherkala-8B-Chat, given the same task and topics. The student and machine essays are independent samples, not matched pairs. In particular, the study seeks to address the following research questions:

1. Which textual features distinguish the student essays from the machine essays, and how large is each difference?
2. How do the essays of GPT-4o and Sherkala-8B differ in these features?
3. How do the differences vary across Kazakh, Russian, and English, and how far can the data support cross-linguistic conclusions?

2. Theoretical Framework

2.1. Machine-Generated Text, Metadiscourse, and Footing

Student and machine essays in English differ in measurable ways, for example, in engagement markers and metadiscourse (Jiang & Hyland, 2025a, 2025b), collocation (Zhang & Crosthwaite, 2025), grammatical and rhetorical constructions (Reinhart et al., 2025), and register (Berber Sardinha, 2024). Metadiscourse comprises the resources through which a text refers to itself and manages the relationship with the reader (Hyland, 2005). A closing marker such as "In conclusion" is an ordinary interactive device, expected as a move of the genre (Swales, 1990). When a generated essay reports its own word count, by contrast, the words belong to the exchange between a user and an assistant; in Goffman's (1981) terms, the footing shifts, and we call this frame leakage. Verbatim repetition of whole sentences has been linked to decoding choices in language models (Holtzman et al., 2019). These differences in surface features raise a broader question about how human and machine writing have typically been compared.

Most comparisons of human and machine writing have been framed as classification. Berber Sardinha (2024) compared multidimensional profiles of AI-generated and human-authored texts, and Reinhart et al. (2025) examined variation in grammatical and rhetorical style across human and model writing. Detection tools built on such differences have performed poorly under testing (Weber-Wulff et al., 2023), and GPT detectors have flagged the writing of non-native English writers as machine-generated more often than that of native writers (Liang et al., 2023). Such studies establish that machine and student texts differ in aggregate, but they say little about which discourse resources account for the difference.

Among the textual dimensions used in such comparisons, lexical diversity has been measured with several indices. Fredrick and Craven (2025) compared lexical diversity, syntactic complexity, and readability in ChatGPT and second-language essays. McCarthy and Jarvis (2010) validated the measure of textual lexical diversity (MTLD) against other lexical diversity measures, and Covington and McFall (2010) proposed the moving-average type–token ratio (MATTR) as a length-independent alternative to the type–token ratio. Lexical diversity is treated here as one dimension of linguistic sophistication, which the study separates from cultural grounding.

2.2. Culturally Grounded Discourse in Kazakh and Russian

Writers build a discoursal self from the voices of their community (Bakhtin, 1979/1986; Ivanic, 1998), and quotation is a basic form of intertextuality (Fairclough, 1992). We use cultural grounding for the extent to which a text ties its claims to culturally situated sources that the reader is expected to recognise: a communal saying, a named author or text, a named place, or the writer's own dated experience. In the coding (Section 3.4.1), the construct is narrowed to these four observable kinds of reference; emblems

were coded separately. This construct is not freestanding but draws on three bodies of work. The first is the work on intertextuality and voice cited above. The second, paremiology, treats the proverb as a compact argument backed by communal authority (Mieder, 2004; Norrick, 1985), and Kazakh proverbs remain the subject of active research in media and digital discourse (Omarbekova & Aasland, 2023; Seyitova et al., 2026). The third, cultural linguistics, treats such expressions as instances of conceptualisations shared by a community (Sharifian, 2017). Proverbs about labour and idleness encode different cultural codes in Kazakh, Russian, and English (Rsaliyeva et al., 2026), and Kazakh digital communication follows culturally specific norms of address (Konyrbayeva et al., 2026).

Cultural grounding is distinct from linguistic sophistication, the range, complexity, and accuracy of a writer's language. A lexically diverse, error-free essay may refer to nothing outside its own argument, and a short essay with learner errors may rest on a proverb and a remembered village. The presence of such references shows that a writer drew on a shared resource, not how deeply the writer understands the culture behind it. Beyond referencing shared cultural resources within one language, writers can also signal cultural grounding through the strategic use of another language. A switch into another language can serve the same function, signalling that the quoted words carry an authority the surrounding text does not claim (Gumperz, 1982). Kazakh-Russian code-mixing has been described as a deliberate resource (Zharkynbekova & Chernyavskaya, 2022); students in western Kazakhstan attach Kazakh affixes to Russian and English roots (Kurmanova et al., 2023), and university students regard trilingual switching as normal (Anarbekova et al., 2026; Utegenova et al., 2024).

2.3. Language Models for Kazakh

Kazakh has far less text available for model training than Russian or English (Kadyrbek et al., 2025; Veitsman & Hartmann, 2025). Sherkala-8B was built to narrow this gap by further training an open model on Kazakh, Russian, English, and Turkish (Koto et al., 2025).

General models, by contrast, have been suggested to align more closely with the values of English-speaking societies (Tao et al., 2024). A Kazakh-oriented model may therefore differ from a general one not only in its command of Kazakh but in the cultural frame it draws on, a difference that the present design can register but not isolate. This difference in cultural frame leads to a concrete, testable expectation about the text each model produces. Since a model produces text from patterns in its training data and not from a writer's own experience (Bender et al., 2021), we expected fewer grounding features in the machine essays, and possibly more in the Kazakh-oriented model than in the general one.

Taken together, the study draws on four perspectives: metadiscourse and footing for the formal features (Hyland, 2005; Goffman, 1981), intertextuality and voice for cultural grounding (Bakhtin, 1979/1986; Fairclough, 1992; Ivanic, 1998), contextualisation for code-switching (Gumperz, 1982), and cultural conceptualisation for cultural figures and emblems (Sharifian, 2017). These perspectives motivate the nine features in Table 3 and the comparisons in Section 4.

3. Methodology

3.1. Participants

The student essays were written in 2026 by first-year undergraduates of the philology programme at K. Zhubanov Aktobe Regional University, Kazakhstan. The students were recruited by convenience sampling: all first-year students in the Kazakh-medium philology groups taught by the authors were invited during regular classes, and those who agreed and gave informed consent wrote one essay each by hand during class, on a topic and in a language assigned by the instructor: Kazakh, their first language; Russian, a second language widely used in Kazakhstan; or English, a foreign language. The 77 students whose essays were retained (37 female, 40 male, aged 18–19) all reported Kazakh as their first language. Participation was voluntary, and the study was approved by the university's ethics committee.

3.2. The Students Corpus

The essays were photographed and transcribed with spelling, punctuation, and grammar left as written, and every transcription was checked against the original image. At a first review of the photographs,

IN PRESS

before the screening register was opened, 34 were set aside without an individual reason being recorded (10, 12, and 12); the remaining 96 were transcribed and screened (44, 25, and 27), with each decision and its reason recorded in the register, and 19 were excluded, leaving 77 (Table 1). An essay was excluded if (1) it could not be treated as the writer's own work (eight essays: texts from essay-sharing websites, signs of a machine-generated source, a near-verbatim match with a GPT-4o essay in the corpus, or an unedited outline fragment); (2) it was written as a school pupil's essay, which suggested a reused school text; (3) the handwriting could not be transcribed with confidence; or (4) too little of the essay had been photographed. The 34 essays set aside at the first review were not transcribed, and the reasons for setting them aside cannot now be reconstructed, so it is not known whether they met the exclusion criteria or differed from the essays analysed. The screening exclusions were concentrated in the Kazakh sample, mainly due to illegible handwriting and texts taken from the internet. Only 77 of the 130 essays collected were analysed, so the final groups are not a random subset of those collected.

Table 1

Selection of the Student Essays by Language

Stage	Kazakh	Russian	English	Total
Essays collected	54	37	39	130
Set aside at first review of photographs (no reason recorded)	10	12	12	34
Essays transcribed and screened	44	25	27	96
Excluded at screening	13	2	4	19
(1) not the writer's own work	4	2	2	8
(2) written as a school pupil	2	0	1	3
(3) handwriting not transcribable	6	0	1	7
(4) incomplete photograph	1	0	0	1
Retained for analysis	31	23	23	77

3.3. The Machine Corpus

Each model produced ten essays for every combination of language (Kazakh, Russian, English) and topic (friendship, native land, work and idleness), 180 essays in total (Table 2). Each essay came from a single generation call, with the prompt sent as one user message, without a system prompt or earlier turns; there were no repeated attempts, no selection among outputs, and no editing, and all 180 outputs were retained. The prompt asked for an essay of about 150–250 words, written as a first-year student would, in the first person, with elements of personal experience, in simple language, and without lists or subheadings; the Kazakh and Russian prompts were close translations of the English one (see Appendix). The essays were generated in May 2026, and no machine essay was written as a counterpart to a student essay.

Table 2

Composition of the Corpus by Language, Topic, and Source (Number of Essays)

Language	Topic	Students	GPT-4o	Sherkala-8B	Total
Kazakh	Friendship	20	10	10	40
	Native land	9	10	10	29
	Work and idleness	2	10	10	22
	Total	31	30	30	91
Russian	Friendship	18	10	10	38
	Native land	5	10	10	25
	Work and idleness	0	10	10	20
	Total	23	30	30	83
English	Friendship	23	10	10	43
	Native land	0	10	10	20
	Work and idleness	0	10	10	20
	Total	23	30	30	83
All		77	90	90	257

IN PRESS

The two models were run under different settings because they behaved differently in a preliminary run in June 2026, whose outputs were used only to develop the coding scheme. GPT-4o (OpenAI Chat Completions API, gpt-4o-2024-08-06) produced complete essays with the default settings, so only the temperature was varied, drawn at random between 0.9 and 1.1 for each essay, with no length limit. Sherkala-8B (Hugging Face, inceptionai/Llama-3.1-Sherkala-8B-Chat, default revision, not pinned, 4-bit quantisation with float16 computation, Google Colab) continued past the end of the essay or repeated itself under these settings, so it was run with the same temperature range together with top-p 0.95, a repetition penalty of 1.15, a no-repeat 4-gram constraint, a 650-token limit, and an explicit end-of-turn token. No random seed was set. These settings served only to obtain complete, non-repetitive essays, but the repetition controls may also raise lexical diversity and discourage recurring formulae, and quantisation may lower output quality. Differences between the models are therefore differences between model-and-setting combinations, not between the models alone.

3.4. Procedure

3.4.1. Coding

Nine features were coded as present or absent in each text (Table 3). Four are the principal grounding features: proverbs and sayings, cultural figures and quotations, time-anchored personal episodes, and named local places. They were grouped because each ties a claim to a source the reader is expected to recognise. The cultural-grounding indicator is calculated for each text: 1 if at least one of the four features is present and 0 otherwise, so the features have equal weight. Equal weighting was a simplifying choice, so each feature is also reported separately, and the main comparison is repeated without episodes. For example, «Досы көпті жау алмайды» деп бекерге айтпаған ("Not for nothing is it said: the enemy cannot defeat one who has many friends") grounds a claim in a proverb, whereas a general statement such as *friendship is valued in every culture* draws on no shared source and was not coded.

Code-switching and named cultural emblems were coded and reported separately, because their presence alone does not show that they support a claim; an emblem such as a holiday may be named only in a description. A word or stretch was coded as a switch if it belonged to another language and was not an established loanword in the language of the essay. Established loanwords written and inflected according to the norms of the essay's language (e.g., *университет* in Kazakh) and names of people, places, works, and brands were not coded. Forms that kept the spelling or morphology of the source language, Russian words used in place of a common Kazakh equivalent, and hybrid forms of a foreign root with a native affix were coded as unmarked switches; a quoted or glossed stretch in another language was coded as a marked switch (examples and borderline cases in Table 3). Code-switching was coded once per essay, so the frequencies are numbers of essays with at least one switch. These frequencies depend on the borrowing rule and are interpreted with caution. The remaining features are formal: closing formulae, frame leakage, and sentence repetition.

Table 3

Coding Rules, with Examples, for the Nine Features and the Cultural-Grounding Indicator

Feature	Coded as present when…	Coded example/not coded
Principal grounding features		
1. Proverb or saying	the text cites a traditional proverb or a sentence presented as a saying (label such as мақал, пословица, as they say)	Coded: «Досы көпті жау алмайды» деп бекерге айтпаған. Not coded: words attributed to a named person (e.g., a professor)
2. Cultural figure or quotation	a writer, poet, public or fictional figure, or literary work is named, with or without a quotation	Coded: Абай атамыз айтқандай – «Оқу, еңбек, терең ой…». Not coded: friends' or classmates' names
3. Time-anchored personal episode	one specific event from the writer's life is located in time (year, age, school grade, or a specific occasion such as last summer)	Coded: мы дружим уже 12 лет, практически с 1 класса. Not coded: habitual activities (every summer); the first weeks at university
4. Named local place	a place below the level of a country is named	Coded: Ақтөбе, Балқаш, Медеу. Not coded: Қазақстан; an unnamed village

Cultural-grounding indicator	at least one of features 1–4 is present (equal weight)	
Features reported separately		
5. Code-switching (marked switches reported as a subset)	a word or stretch in another language that is not an established loanword; marked if quoted or glossed	Coded: колледже, отношение, группаластармен (unmarked); a quoted, glossed Kazakh proverb in Russian (marked). Not coded: университет in Kazakh; names; brands. Borderline: этап (not coded); комфорт beside хейт (coded)
6. Named cultural emblem	a custom, holiday, game, dish, or traditional term associated with a particular national culture	Coded: Наурыз, жайлау, тұсау кесер, Масленица. Not coded: tradition, national food, coffee
Formal features		
7. Closing formula	the final paragraph (in one-paragraph essays, the last one or two sentences) opens with a listed marker: Kazakh қорытындылай and inflections, қорыта айтқанда, қорытынды жасай отырып, түйіндей, сөз соңында; Russian в заключение, в завершение, к завершению, подводя итог, таким образом, итак, в итоге, в целом; English in conclusion, to conclude, to sum up, in summary, in essence, overall, all in all, ultimately, finally	Not coded: a closing proverb; a marker inside the paragraph
8. Frame leakage	text belonging to the exchange between user and assistant (e.g., a word count, a refusal)	No case found
9. Sentence repetition	a whole sentence repeated word for word in the same text	Not coded: paraphrase, partial repetition

Candidate passages were located by computer-assisted searches (closing markers, proverb labels, quotation marks, names and toponyms, time expressions, other scripts); initial codes were proposed with the help of an AI assistant (Claude, Anthropic), and every code was decided by a human coder after reading the text. Ambiguous cases were resolved by the definitions and exclusions in Table 3 and, where these did not settle a case, by the rule closest to the feature's purpose. The scheme was drafted on the preliminary outputs of June 2026 and fixed before the main coding; the only later change to the rules was the revision of the episode and emblem rules described in Section 3.4.2.

3.4.2. Reliability

A second coder, working independently and without information about the source of the texts, made two checks; agreement is reported as percentage agreement and Cohen's κ. First, the second coder applied the rules for the three formal features to a stratified random sample of 45 texts, five per source-by-language cell, and agreed with the primary coding on all 45 (κ = 1.00 for closing formulae and repetition; undefined for frame leakage, which had no positive case). Second, the second coder coded the grounding features, code-switching, and cultural emblems in a new stratified random sample of 60 texts (ten student essays and five essays per model in each language). Agreement was high for cultural figures and quotations and for named places (100%, κ = 1.00), proverbs (93.3%, κ = .76), code-switching (95.0%, κ = .70), and marked code-switching (98.3%, κ = .79), but low for time-anchored episodes (73.3%, κ = .32) and cultural emblems (91.7%, κ = .41); for the indicator it was 71.7% (κ = .46). By common benchmarks (Landis & Koch, 1977), κ values of .32 and .46 indicate only fair and moderate agreement, too low for the episode codes or the indicator to be taken as reliable under the original rules.

The check showed that two rules were underspecified: the episode rule did not say whether habitual activities or the first weeks at university should count, and the emblem rule did not say whether items

IN PRESS

without a national association, such as coffee, should count; it also revealed proverbs and switches missed in the initial coding. The two rules were made more precise (Table 3), and the second coder, who had read the whole corpus, recoded all 257 texts against them, again without information about their source; these final codes are the ones analysed in Section 4. Because the κ values above describe agreement under the original rules, a further check was made under the final rules: a third coder, who had not taken part in the earlier coding and had no access to it, coded a new stratified random sample of 60 texts (ten student essays and five essays per model in each language, none from the earlier sample) without information about their source. Agreement with the final codes was 96.7% for the indicator (κ = .93), proverbs (κ = .90), and time-anchored episodes (κ = .87), 98.3% for code-switching (κ = .94), and 100% for cultural figures and quotations, named places, marked code-switching, and cultural emblems (κ = 1.00). The five disagreements (two proverbs, two episodes, one switch) were borderline cases. By the same benchmarks, this agreement is almost perfect, which suggests that the low earlier values reflected the underspecified rules. Named places and emblems occurred in only two of the 60 texts each, so their estimates are imprecise. As a precaution, the main comparison is also reported without episodes, and the emblem results are treated as descriptive.

3.4.3. Analysis

The analyses are exploratory. For RQ1, students and the two models pooled were compared on each feature with two-sided Fisher's exact tests; effect sizes are differences in percentage points with 95% confidence intervals (Newcombe's method). For this family of nine comparisons (the four grounding features, the indicator, code-switching, marked code-switching, and cultural emblems in Table 4, and closing formulae in Section 4.3), p-values were also adjusted with the Holm method. Because pooled comparisons mix source with topic and language composition, they were repeated within each language, with the machine corpus restricted to the six language–topic cells in which students wrote (120 essays), within the friendship topic, and on 1,000 random subsamples of 23 Kazakh student essays. For RQ2, the two models were compared with Fisher's exact tests. *F*or RQ3, student–model comparisons were made within each language. These secondary comparisons check the consistency of the main result, are not adjusted, and are treated as descriptive; a result significant only before adjustment or in only one check is called suggestive.

Lexical diversity was computed with a Python script written for the study (SciPy for the tests; no corpus-linguistics package; available on request). Texts were lower-cased; tokens were sequences of Cyrillic (including Kazakh-specific) or Latin letters, with internal hyphens and apostrophes kept, and digits and punctuation were removed. Word forms were not lemmatised, because a separate lemmatiser for each language would have added tool-specific error, so inflected forms counted as different types. Three measures were computed: the type–token ratio (TTR); the moving-average type–token ratio (MATTR), the mean TTR over all 50-token windows (Covington & McFall, 2010), with TTR used for the one text shorter than 50 tokens; and the measure of textual lexical diversity (MTLD; McCarthy & Jarvis, 2010), with a threshold of .72, computed in both directions and averaged. MATTR was the main measure because its fixed window makes it independent of text length; no other normalisation was applied. Since Kazakh, an agglutinative language, produces more word forms per lexeme than Russian, and Russian more than English, the measures were compared only within a language, with Mann–Whitney U tests and rank-biserial correlations.

4. Results

Percentages are proportions of texts, with the number of texts in the relevant group as the denominator; the two models are pooled unless stated otherwise, and differences are student minus model percentages. Tests are two-sided, and p-values are given to three decimals (below .001 as < .001).

4.1. Cultural Grounding, Code-Switching, and Cultural Emblems

Pooled across languages (Table 4), the cultural-grounding indicator was present in 41 of 77 student essays (53.2%) and in 32 of 180 machine essays (17.8%), a difference of 35.5 percentage points (95% CI 22.8 to 47.3; $p < .001$, Holm-adjusted $p < .001$). The difference was carried by three components, which remained significant after adjustment: proverbs and sayings (32.5% against 3.9%; 28.6 points),

cultural figures and quotations (11.7% against 1.1%; 10.6 points), and named local places (19.5% against 5.6%; 13.9 points). Time-anchored personal episodes did not differ overall (16.9% against 11.1%; $p = .225$), although in Kazakh they were more frequent in the student essays (6/31 against 2/60; $p = .017$, unadjusted); the prompt asked the models to include personal experience. Within languages, the indicator separated students from models in Kazakh (23/31, 74.2%, against 8/60, 13.3%; 60.9 points, 95% CI 40.3 to 74.6; $p < .001$) and Russian (14/23, 60.9%, against 14/60, 23.3%; 37.5 points, 14.1 to 56.7; $p = .002$), but not in English (4/23, 17.4%, against 10/60, 16.7%; $p = 1.000$; Figure 1).

Table 4

Cultural Grounding, Code-Switching, and Cultural Emblems, by Source and Language

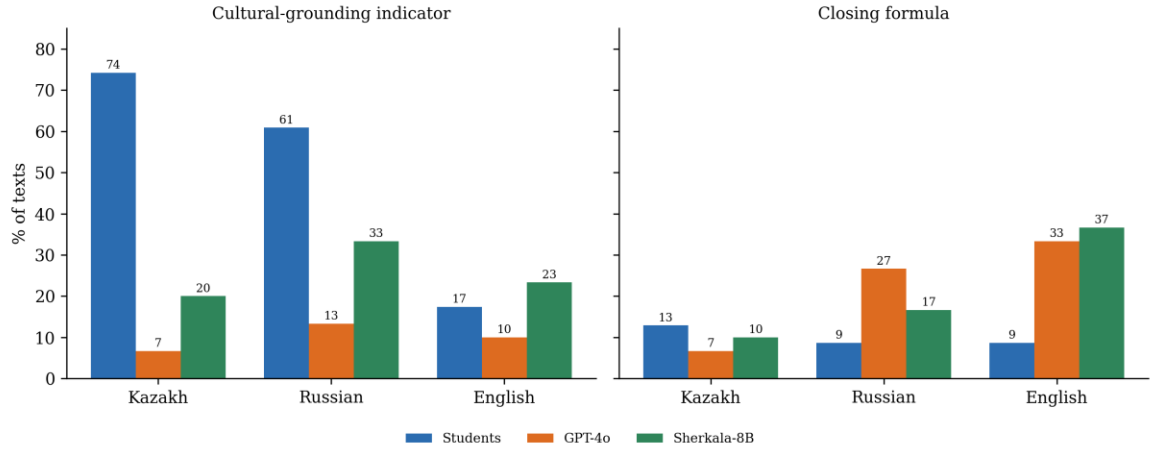
Feature	Source	Kazakh	Russian	English	All	Difference [95% CI]; *p*; *p* (Holm)
1. Proverb or saying	Students	17/31 (54.8)	8/23 (34.8)	0/23 (0.0)	25/77 (32.5)	28.6 [18.4, 39.8]; < .001; < .001
	GPT-4o	0/30 (0.0)	4/30 (13.3)	1/30 (3.3)	5/90 (5.6)	
	Sherkala-8B	0/30 (0.0)	2/30 (6.7)	0/30 (0.0)	2/90 (2.2)	
2. Cultural figure or quotation	Students	7/31 (22.6)	2/23 (8.7)	0/23 (0.0)	9/77 (11.7)	10.6 [4.5, 19.7]; < .001; .003
	GPT-4o	0/30 (0.0)	0/30 (0.0)	0/30 (0.0)	0/90 (0.0)	
	Sherkala-8B	2/30 (6.7)	0/30 (0.0)	0/30 (0.0)	2/90 (2.2)	
3. Time-anchored personal episode	Students	6/31 (19.4)	4/23 (17.4)	3/23 (13.0)	13/77 (16.9)	5.8 [-2.9, 16.4]; .225; .449
	GPT-4o	1/30 (3.3)	3/30 (10.0)	2/30 (6.7)	6/90 (6.7)	
	Sherkala-8B	1/30 (3.3)	7/30 (23.3)	6/30 (20.0)	14/90 (15.6)	
4. Named local place	Students	7/31 (22.6)	6/23 (26.1)	2/23 (8.7)	15/77 (19.5)	13.9 [5.4, 24.4]; .001; .007
	GPT-4o	1/30 (3.3)	0/30 (0.0)	0/30 (0.0)	1/90 (1.1)	
	Sherkala-8B	6/30 (20.0)	2/30 (6.7)	1/30 (3.3)	9/90 (10.0)	
Cultural-grounding indicator (any of 1–4)	Students	23/31 (74.2)	14/23 (60.9)	4/23 (17.4)	41/77 (53.2)	35.5 [22.8, 47.3]; < .001; < .001
	GPT-4o	2/30 (6.7)	4/30 (13.3)	3/30 (10.0)	9/90 (10.0)	
	Sherkala-8B	6/30 (20.0)	10/30 (33.3)	7/30 (23.3)	23/90 (25.6)	
5. Code-switching	Students	9/31 (29.0)	6/23 (26.1)	0/23 (0.0)	15/77 (19.5)	10.6 [1.7, 21.3]; .022; .087
	GPT-4o	0/30 (0.0)	1/30 (3.3)	0/30 (0.0)	1/90 (1.1)	
	Sherkala-8B	6/30 (20.0)	8/30 (26.7)	1/30 (3.3)	15/90 (16.7)	
Marked code-switching (subset of 5)	Students	1/31 (3.2)	6/23 (26.1)	0/23 (0.0)	7/77 (9.1)	8.0 [2.6, 16.5]; .004; .018
	GPT-4o	0/30 (0.0)	0/30 (0.0)	0/30 (0.0)	0/90 (0.0)	
	Sherkala-8B	0/30 (0.0)	1/30 (3.3)	1/30 (3.3)	2/90 (2.2)	
6. Named cultural emblem	Students	2/31 (6.5)	0/23 (0.0)	0/23 (0.0)	2/77 (2.6)	-3.0 [-7.7, 3.9]; .519; .519
	GPT-4o	4/30 (13.3)	2/30 (6.7)	0/30 (0.0)	6/90 (6.7)	
	Sherkala-8B	1/30 (3.3)	0/30 (0.0)	3/30 (10.0)	4/90 (4.4)	

Note. Number of texts with the feature/number of texts in the cell (percentage). Last column: student minus pooled model percentage (77 and 180 essays), in percentage points, with 95% confidence interval (Newcombe); Fisher's exact test; Holm-adjusted p for the nine main comparisons. A positive difference means the feature was more frequent in the student essays. Code-switching and cultural emblems are not part of the indicator.

IN PRESS

Figure 1

Texts with the Cultural-Grounding Indicator and with a Closing Formula, by Source and Language (Percentage of Texts)



Student essays used proverbs as premises or conclusions (e.g., KZ_kz_h_001) and quoted verse, such as Qunanbaiuly's (1995) lines on the qualities a person should strive for, sometimes with small changes (KZ_a0c64f7a). Code-switching occurred in 15 of 77 student essays (19.5%) and 16 of 180 machine essays (8.9%; p = .022 unadjusted, Holm-adjusted p = .087), and marked switches in 7 and 2 essays (Holm-adjusted p = .018). The kind of switching differed: six Russian student essays quoted and glossed Kazakh sayings, and one Kazakh student essay quoted or glossed a stretch in another language, whereas Sherkala-8B inserted English words or sentences into Kazakh essays, as in Еңбек – бұл мy армандарыма жетудің жалғыз жолы ("Work is the only way to my dreams", KZ_sherkala_trud_prazdnost_09), and Kazakh or Ukrainian forms into Russian essays, mostly unmarked. All seven machine proverbs occurred in essays on work and idleness, six in Russian and one in English, that is, in two language–topic cells in which no student wrote; the 20 Kazakh machine essays on this topic contained none. No machine essay contained a Kazakh proverb, whereas 17 Kazakh and 8 Russian student essays did. Named cultural emblems did not differ (2/77 against 10/180; p = .519). Between the models, Sherkala-8B had the indicator more often than GPT-4o (23/90 against 9/90; p = .010), mainly through named places (9/90 against 1/90; p = .018), and more code-switching (15/90 against 1/90; p < .001); time-anchored episodes did not differ between the two models (14/90 against 6/90; p = .095). Given the different settings, these differences cannot be attributed to the models alone.

4.2. Comparisons with Topic and Language Held Constant

With the machine corpus restricted to the six language–topic cells in which students also wrote (120 essays; Table 5), the indicator remained more frequent in the student essays (41/77, 53.2%, against 18/120, 15.0%; p < .001), as did proverbs, cultural figures and quotations, named places, and marked switching. The matched machine essays contain no proverb (0/120): the seven machine proverbs reported in Section 4.1 all come from the Russian and English essays on work and idleness, which fall outside the matched cells, so the two figures are consistent. Code-switching as a whole did not differ here (15/77 against 12/120; p = .088), unlike in the pooled comparison (p = .022). The two comparisons differ only in the machine group: the 60 machine essays outside the matched cells contain only four switches, so the machine rate rises from 8.9% to 10.0%. The pooled result does not survive Holm adjustment, and the matched one is not significant, so code-switching as a whole is treated as inconclusive. Within the friendship topic, the indicator separated students from models in Kazakh (13/20 against 1/20; p < .001) but not in Russian (9/18 against 4/20; p = .087) or English (4/23 against 3/20; p = 1.000). In the Kazakh subsamples of 23 essays, the share with the indicator lay between .652 and .826 in 95% of draws, above the Kazakh model rates (.200 and .067). With episodes excluded from the indicator, the difference remained in Kazakh (21/31 against 7/60) and Russian (14/23 against 8/60; both p < .001) and was absent in English (2/23 against 2/60; p = .306). These contrasts describe these samples and are confounded with topic and with the status of the language of writing.

Table 5

Comparisons with Topic and Language Held Constant: Students against the Two Models Pooled

Comparison	Students	Models	*p*
Matched language–topic cells			
1. Proverb or saying	25/77 (32.5)	0/120 (0.0)	< .001
2. Cultural figure or quotation	9/77 (11.7)	2/120 (1.7)	.004
3. Time-anchored personal episode	13/77 (16.9)	10/120 (8.3)	.074
4. Named local place	15/77 (19.5)	9/120 (7.5)	.015
Cultural-grounding indicator (any of 1–4)	41/77 (53.2)	18/120 (15.0)	< .001
5. Code-switching	15/77 (19.5)	12/120 (10.0)	.088
Marked code-switching (subset of 5)	7/77 (9.1)	1/120 (0.8)	.006
6. Named cultural emblem	2/77 (2.6)	9/120 (7.5)	.207
7. Closing formula	8/77 (10.4)	19/120 (15.8)	.299
Friendship essays			
Cultural-grounding indicator, Kazakh	13/20 (65.0)	1/20 (5.0)	< .001
Cultural-grounding indicator, Russian	9/18 (50.0)	4/20 (20.0)	.087
Cultural-grounding indicator, English	4/23 (17.4)	3/20 (15.0)	1.000

Note. Matched cells: machine essays from the six language–topic cells in which students also wrote (Kazakh: all topics; Russian: friendship and native land; English: friendship). Friendship: 20 machine essays per language. Students are compared with the pooled models; Fisher's exact tests, not adjusted. The seven machine proverbs in Table 4 all occur in Russian and English essays on work and idleness, which are outside the matched cells; hence 0/120 here.

4.3. Closing Formulae, Frame Leakage, and Sentence Repetition

A closing formula opened the final paragraph of 4 of 31 Kazakh, 2 of 23 Russian, and 2 of 23 English student essays (Table 6). Pooled across languages, closing formulae were more frequent in the machine essays (39/180, 21.7%, against 8/77, 10.4%; $p = .035$), but not after Holm adjustment ($p = .105$) or in the matched cells ($p = .299$). Within languages, the difference was significant only in English (2/23 against 21/60; $p = .026$), not in Kazakh ($p = .484$) or Russian ($p = .216$). The two models hardly differed (20/90 and 19/90).

No frame leakage was found in any of the 257 texts, and one student essay in English repeated a whole sentence verbatim (EN_248368f3). Given the decoding settings of Sherkala-8B, these null results say little about how language models write in general.

Table 6

Texts with a Closing Formula, by Source and Language

Source	Kazakh	Russian	English
Students	4/31 (12.9)	2/23 (8.7)	2/23 (8.7)
GPT-4o	2/30 (6.7)	8/30 (26.7)	10/30 (33.3)
Sherkala-8B	3/30 (10.0)	5/30 (16.7)	11/30 (36.7)
Fisher's exact test, students vs. models	$p = .484$	$p = .216$	$p = .026$

Note. Percentages in parentheses; students: 31, 23, and 23 essays; models pooled (60 per language) in the tests, which are not adjusted.

4.4. Essay Length and Lexical Diversity

The machine essays were longer than the student essays in Russian and English (Mann–Whitney U = 47–151.5, $p < .001$, r = .56 to .86), but not in Kazakh (U = 346.5 and 368, $p \geq .089$; Table 7). MATTR was lower for the student essays than for both models in all three languages (U = 0–214, all $p < .001$; rank-biserial r = .54 to 1.00), with the largest gap in English (.767 against .850 for GPT-4o and .918 for Sherkala-8B). The English student essays also contained learner forms such as *I love they* (EN_2c8718a3), not found in the machine essays. Because the repetition controls of Sherkala-8B may raise lexical diversity, the comparison with GPT-4o is the more informative one.

IN PRESS

Table 7
Mean Essay Length and Lexical Diversity, by Source and Language

Language	Source	Words	TTR	MATTR	MTLD
Kazakh	Students	152	.79	.899	199.1
	GPT-4o	168	.84	.947	324.1
	Sherkala-8B	169	.80	.926	253.5
Russian	Students	161	.70	.839	124.0
	GPT-4o	202	.74	.895	212.3
	Sherkala-8B	258	.83	.947	442.4
English	Students	180	.57	.767	68.6
	GPT-4o	245	.63	.850	125.1
	Sherkala-8B	288	.74	.918	303.2

Note. Means per text; students: 31, 23, and 23 texts; each model: 30 per language. TTR = type–token ratio; MATTR = moving-average type–token ratio (50-token window); MTLD = measure of textual lexical diversity (threshold .72). Values are comparable within a language only. Students are compared with each model within a language (Section 4.4).

5. Discussion

This study had three aims. The first was to determine which textual features distinguish essays written by students from essays generated by GPT-4o and Sherkala-8B, and how large each difference is. The second was to compare the two language models on the same features. The third was to examine how these differences vary across Kazakh, Russian, and English, and how far the data can support cross-linguistic conclusions.

The first aim concerns the difference between student and machine essays. The clearest differences lay in proverbs, cultural figures and quotations, and named local places. Each was several times more frequent in the student essays. The differences held in the matched comparisons, and the pooled differences survived Holm adjustment. Most student proverbs were Kazakh sayings used as premises or conclusions, whereas the machine essays contained almost none. Proverbs rest on communal authority (Mieder, 2004; Norrick, 1985), so the students drew on a shared store of sayings that the models did not reproduce. The absence of Kazakh proverbs in the machine essays shows what the models produced under these prompts; it is not evidence that they cannot produce such proverbs, and a prompt that asked for them would probably yield some. Seven of the 15 student essays with a switch marked it by quoting or glossing: six Russian essays quoted and glossed a Kazakh saying, and one Kazakh essay quoted or glossed a stretch in another language. This is a contextualisation cue (Gumperz, 1982), in line with descriptions of Kazakh-Russian mixing as a deliberate resource (Zharkynbekova & Chernyavskaya, 2022). Fourteen of the 16 machine essays with a switch were unmarked. Dated personal episodes did not differ overall, even though the prompt asked for personal experience and only 11% of machine essays contained one. English-language studies of engagement markers, metadiscourse, and lexis (Jiang & Hyland, 2025a, 2025b; Zhang & Crosthwaite, 2025) focus on linguistic features rather than references to shared cultural sources, so the features that separate student from machine writing may depend on the language and culture of the writing.

The second aim compared the two models. Sherkala-8B had the cultural-grounding indicator more often than GPT-4o (23 of 90 against 9 of 90; $p = .010$), mainly through named places (9 against 1), and it used code-switching far more often (15 against 1). Its switches were mostly unmarked: English words in Kazakh essays, and Kazakh or Ukrainian forms in Russian essays. This may reflect the training of a model on several languages, with Kazakh as one of them (Koto et al., 2025). The named-place difference fits the expectation in Section 2, but the settings differed, so it cannot be attributed to the models alone. Tao et al. (2024) suggest that general models align more closely with English-speaking values, which would fit this pattern, but the settings differed, so the point cannot be tested here. Closing formulae did not differ (20 against 19 of 90 essays), and time-anchored personal episodes did not differ (14/90 against 6/90). Both models were far below the students on proverbs, so the gap between human and machine writing is larger than the gap between the two models.

The third aim concerned variation across languages. The indicator separated students from models clearly in Kazakh (74.2% against 13.3%) and Russian (60.9% against 23.3%), but not in English (17.4% against 16.7%). Within the friendship topic, the Kazakh difference remained (65.0% against 5.0%), whereas the Russian difference was not significant (50.0% against 20.0%), so part of the Russian difference may reflect the topic. The English student essays were all on friendship and written in a foreign language, so the absent difference in English cannot be separated from topic, proficiency, and the status of English. The data therefore support a narrow conclusion: in these samples the contrast between students and models was clearest in Kazakh, smaller in Russian, and absent in English. They do not show whether English writing is less culturally grounded than Kazakh writing in general; a claim of that kind would need samples balanced in topic, proficiency, and language status.

The study has two theoretical implications. First, it separates cultural grounding from linguistic sophistication. The models exceeded the students in lexical diversity but not in grounding, so a framework that measures only sophistication would favour the machine essays and miss this difference. Cultural grounding can be studied through the intertextual resources by which a text ties its claims to sources its readers recognise (Fairclough, 1992; Ivanic, 1998). Second, the presence of a cultural reference is easily taken for competence, a misjudgement that Pishghadam (2026) examines in language studies. The indicator records whether a proverb is present, not how much it matters in the essay, and the same caution applies to any measure built on presence alone. In addition to the theoretical implications, the study has three practical implications. For teaching, grounding could become an explicit criterion in writing instruction, with students asked to support a claim with an episode, a proverb, or a quotation and to explain how the reference supports it. For the detection of machine writing, none of the features should decide whether a particular student used a model: the samples overlap on most features, and detectors are unreliable (Weber-Wulff et al., 2023) and biased against non-native writers (Liang et al., 2023). For research, comparisons of machine text should report generation settings in full and use the same settings for all models.

The study has several limitations. First, the student sample is a convenience sample from one programme, and its topics are unevenly distributed across languages, so differences between languages cannot be separated from topic and proficiency. Second, 34 of the 130 essays collected were set aside before the screening register was opened, without recorded reasons, so their possible influence on the final sample cannot be assessed. Third, the models were run under different settings with one prompt per language that did not ask for cultural references. Fourth, reliability was checked on 60 texts, in which named places and emblems were rare, and the code-switching codes depend on a judgement about borrowing. Fifth, the cultural-grounding indicator was developed for this study, weights four features equally, and has not been validated. Sixth, the analyses are exploratory, and only the nine main comparisons were adjusted for multiple testing. Future research should use balanced samples, record every exclusion from the outset, standardise prompts and settings with several prompts per language (some inviting cultural references), and validate the grounding measure against readers' judgements.

Acknowledgements

The authors thank the students who contributed their essays. During the preparation of this work, the authors used Claude (Anthropic) to assist with transcription, data processing, statistics, initial coding suggestions, the generation of scripts, the figure, and language editing. All transcriptions were checked against the original essays, and all codes were finalised by human coders. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Disclosure Statement

The authors claim no conflicts of interest.

Funding

The research did not receive any specific grants from funding agencies.

IN PRESS

Data Availability

The coded dataset (feature codes for each essay, with anonymised identifiers), the generation records (prompts, model versions, and settings), and the analysis script are available from the corresponding author upon reasonable request.

Ethics Declaration

The study protocol was reviewed and approved by the Institutional Research Ethics Committee of K. Zhubanov Aktobe Regional University (protocol No. 9). Participation was voluntary, and informed consent was obtained from all participants. No identifying data were kept, and personal names in the essays were replaced with substitute names.

References

Anarbekova, A. A., Suleimenova, Z. Y., Kuanyshbayeva, A. K., Saryshova, K. S., Abdullina, N. U., & Utelbayeva, G. U. (2026). Students' multilingual practices in the English classroom in Kazakhstan. *Theory and Practice in Language Studies*, *16*(3), 1048–1059. https://doi.org/10.17507/tpls.1603.35

Bakhtin, M. M. (1986). *Speech genres and other late essays* (C. Emerson & M. Holquist, Eds.; V. W. McGee, Trans.). University of Texas Press. (Original work published 1979)

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In M. C. Elish, W. Isaac, & R. Zemel (Eds.), *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Berber Sardinha, T. (2024). AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, *4*(1), Article 100083. https://doi.org/10.1016/j.acorp.2023.100083

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. https://doi.org/10.1080/09296171003643098

Fairclough, N. (1992). *Discourse and social change*. Polity Press.

Fredrick, D. R., & Craven, L. (2025). Lexical diversity, syntactic complexity, and readability: A corpus-based analysis of ChatGPT and L2 student essays. *Frontiers in Education*, *10*, Article 1616935. https://doi.org/10.3389/feduc.2025.1616935

Goffman, E. (1981). *Forms of talk*. University of Pennsylvania Press.

Gumperz, J. J. (1982). *Discourse strategies*. Cambridge University Press.

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). *The curious case of neural text degeneration* (arXiv:1904.09751). arXiv. https://doi.org/10.48550/arXiv.1904.09751

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. Continuum.

Ivanic, R. (1998). *Writing and identity: The discoursal construction of identity in academic writing*. John Benjamins.

Jiang, F. K., & Hyland, K. (2025a). Does ChatGPT write like a student? Engagement markers in argumentative essays. *Written Communication*, *42*(3), 463–492. https://doi.org/10.1177/07410883251328311

Jiang, F. K., & Hyland, K. (2025b). Rhetorical distinctions: Comparing metadiscourse in essays by ChatGPT and students. *English for Specific Purposes*, *79*, 17–29. https://doi.org/10.1016/j.esp.2025.03.001

Kadyrbek, N., Tuimebayev, Z., Mansurova, M., & Viegas, V. (2025). The development of small-scale language models for low-resource languages, with a focus on Kazakh and direct preference optimization. *Big Data and Cognitive Computing*, *9*(5), Article 137. https://doi.org/10.3390/bdcc9050137

Konyrbayeva, G. S., Smagulova, G. N., Kazhibayeva, G. S., Shayakhimov, D. K., Demessinova, A., & Okur, A. (2026). Phatic norms in Kazakh and English digital discourse: Cultural code and pragmatic function. *Journal of Language Teaching and Research*, *17*(5), 1785–1795. https://doi.org/10.17507/jltr.1705.23

IN PRESS

Koptleuova, K., Karagulova, B., Zhumakhanova, A., Kondybay, K., & Salikhova, A. (2023). Multilingualism and the current language situation in the Republic of Kazakhstan. *International Journal of Society, Culture & Language*, *11*(3), 242–257. https://doi.org/10.22034/ijscl.2023.2007080.3099

Koto, F., Joshi, R., Mukhituly, N., Wang, Y., Xie, Z., Pal, R., Orel, D., Mullah, P., Turmakhan, D., Goloburda, M., Kamran, M., Ghosh, S., Jia, B., Mansurov, J., Togmanov, M., Banerjee, D., Laiyk, N., Sakip, A., Han, X., . . . Nakov, P. (2025). *Sherkala-Chat: Building a state-of-the-art LLM for Kazakh in a moderately resourced setting* (arXiv:2503.01493). arXiv. https://doi.org/10.48550/arXiv.2503.01493

Kurmanova, B., Utegenova, A., Sultaniyazova, I., Khassanov, G., Almagambetova, N., & Abdigazi, S. (2023). Multilingual practices in the students' microcommunity. *International Journal of Society, Culture & Language*, *11*(1), 146–159. https://doi.org/10.22034/ijscl.2022.1978443.2862

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, *4*(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Mieder, W. (2004). *Proverbs: A handbook*. Greenwood Press.

Norrick, N. R. (1985). *How proverbs mean: Semantic studies in English proverbs*. Mouton.

Omarbekova, G., & Aasland, E. (Eds.). (2023). *Contemporary Kazakh proverb research: Digital, cognitive, literary, and ecological approaches*. Peter Lang.

Pishghadam, R. (2026). When culture is misjudged: Toward magnitude culturology in language studies. *International Journal of Society, Culture & Language*, *14*(2), 123–137. https://doi.org/10.22034/ijscl.2026.735705

Reinhart, A., Markey, B., Laudenbach, M., Pantusen, K., Yurko, R., Weinberg, G., & Brown, D. W. (2025). Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, *122*(8), Article e2422455122. https://doi.org/10.1073/pnas.2422455122

Rsaliyeva, Z., Dadebaev, Z., Abitzhanova, Z., Alpysbayeva, A. Y., Olkova, I. A., & Myachin, K. A. (2026). The use of metaphors in paremiological units of English, Russian, and Kazakh languages: A study based on the thematic group "Labor–Idleness". *Theory and Practice in Language Studies*, *16*(5), 1785–1795. https://doi.org/10.17507/tpls.1605.35

Qunanbaiuly, A. (1995). *Shygharmalarynyng eki tomdyq tolyq zhinaghy* [The complete collection of his works in two volumes] (Vol. 1). Zhazushy.

Seyitova, B., Yessenova, K., Nurzhanova, A., Kozhakhmetova, F., & Ismailova, F. (2026). Ethnopoetic codes in Kazakh media: Transformations of proverbs, mythologemes and Zhyrau poetry. *Theory and Practice in Language Studies*, *16*(6), 2171–2179. https://doi.org/10.17507/tpls.1606.36

Sharifian, F. (2017). *Cultural linguistics: Cultural conceptualisations and language*. John Benjamins. https://doi.org/10.1075/clscc.8

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, *3*(9), Article 346. https://doi.org/10.1093/pnasnexus/pgae346

Tlepbergen, D., Akzhigitova, A., & Zabrodskaja, A. (2025). Kazakh–English bilingualism in Kazakhstan: Public attitudes and language practices. *Languages*, *10*(5), Article 102. https://doi.org/10.3390/languages10050102

Utegenova, A., Kurmanova, B. Z., Abdullina, N. U., Koshimova, B., Otarova, A., Zhetkizgenova, A., Batinova, A. S., & Muratbe, B. K. (2024). Code-switching in a multilingual classroom in Kazakhstan: A university students' perspective. *Forum for Linguistic Studies*, *6*(6), 371–380. https://doi.org/10.30564/fls.v6i6.7338

Veitsman, Y., & Hartmann, M. (2025). Recent advancements and challenges of Turkic Central Asian language processing. In H. Hettiarachchi, T. Ranasinghe, P. Rayson, R. Mitkov, M. Gaber, D. Premasiri, F. A. Tan, & L. Uyangodage (Eds.), *Proceedings of the First Workshop on Language*

IN PRESS

Models for Low-Resource Languages (pp. 309–324). Association for Computational Linguistics. https://aclanthology.org/2025.loreslm-1.25/

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*, Article 26. https://doi.org/10.1007/s40979-023-00146-z

Zhang, M., & Crosthwaite, P. (2025). More human than human? Differences in lexis and collocation within academic essays produced by ChatGPT-3.5 and human L2 writers. *International Review of Applied Linguistics in Language Teaching*, *64*(2), 733–760. https://doi.org/10.1515/iral-2024-0196

Zharkynbekova, S., & Chernyavskaya, V. E. (2022). Kazakh-Russian bilingual practice: Code-mixing as a resource in communicative interaction. *RUDN Journal of Language Studies, Semiotics and Semantics*, *13*(2), 468–482. https://doi.org/10.22363/2313-2299-2022-13-2-468-482

Appendix

Prompts Used to Generate the Machine Essays

The same prompt was sent to both models; [topic] stood for the topic word (English: Friendship, Native Land, Labor and Idleness; Kazakh: Достық, Туған жер, Еңбек пен жалқаулық; Russian: Дружба, Родная земля, Труд и праздность).

English: Write an essay on the topic "[topic]" in English, approximately 150–250 words, as a first-year university student would. Write naturally, in the first person, including elements of personal experience, in simple student language, without lists or subheadings.

Kazakh: «[topic]» тақырыбына қазақ тілінде, шамамен 150–250 сөзден тұратын эссе жазыңыз, университеттің бірінші курс студенті сияқты. Бірінші жақтан, жеке тәжірибе элементтерін қоса, қарапайым студенттік тілмен, тізім мен тақырыпшаларсыз жазыңыз.

Russian: Напишите эссе на тему «[topic]» на русском языке, объёмом примерно 150–250 слов, как это сделал бы первокурсник университета. Пишите естественно, от первого лица, с элементами личного опыта, простым студенческим языком, без списков и подзаголовков.

IN PRESS