UDC 004.8:004.912

Ilya Tsoy
Grade 11 student
State Lyceum No. 165
(Almaty, Kazakhstan)

Scientific supervisor: Assyltay Sydyk
2nd-year student
New York University Abu Dhabi
(Abu Dhabi, UAE)

WHO PAYS MORE TOKENS FOR THE SAME SENTENCE? RUSSIAN, KOREAN AND ENGLISH IN FOUR LANGUAGE MODEL TOKENIZERS ON 1,012 FLORES-200 SENTENCES

Abstract: Language models read and write text in tokens, and commercial services charge and limit their users per token, so a language that needs more tokens for the same content costs its speakers more money, time and context space. Earlier tables of token cost measured each language against English; in them the order of Russian and Korean already differed between older tokenizers (those of GPT-2, GPT-4 and the first LLaMA), but no comparison of Russian with Korean in the tokenizers of Llama 3, Gemma 1/2 and GPT-4o/GPT-5 was found, and the order was not tested sentence by sentence. This paper compares English, Russian and Korean on the same 1,012 sentences of the FLORES-200 parallel benchmark in four tokenizers: those of GPT-4 (cl100k_base), Llama 3, the first two generations of Gemma, and GPT-4o and GPT-5 (o200k_base). No language model was run. Before any count was made, the author forecast that English would be the cheapest and that Russian would cost more than Korean because Korean is more compact. English was the cheapest in all four tokenizers and the cheapest of the three languages in at least 94.4% of individual sentences. Russian needed 1.38 to 2.46 times as many tokens as English and Korean 1.47 to 2.37 times. Which of the two costs more depends on the tokenizer: Russian in GPT-4 and Llama 3 (Korean/Russian 0.96 and 0.92), Korean in Gemma 1/2 and GPT-4o/GPT-5 (1.19 and 1.04). Each order holds in a majority of single sentences, from 55% to 83% (exploratory sign tests, $p < 0.001$), and where the gap is only a few percent it depends on how long each translation is. OpenAI's newer tokenizer uses 43% fewer tokens than the older one for Russian and 38.5% fewer for Korean. In these tokenizers the order goes together both with a larger vocabulary and with more Cyrillic entries for each Hangul entry, and the data cannot say which of the two matters. The forecast was right about English and right for two of the four tokenizers about Russian and Korean: English stays the cheapest everywhere, but which of the two other languages pays more is set by the tokenizer.

Keywords: tokenization, language models, Russian language, Korean language, multilingual fairness, FLORES-200

Introduction

A language model does not read letters or words. Its tokenizer first cuts the text into pieces from a fixed vocabulary, called tokens, and the model then works with these pieces.

Services built on language models count tokens: providers of commercial language model services charge their users "on the number of 'tokens' processed or generated" [1, p. 9904], the time needed to answer grows with the number of tokens, and a model can hold only a fixed number of tokens in its context at once [2, p. 36964]. If the same content needs more tokens in one language than in another, its speakers pay more, wait longer and can give the model less text.

The author's interest in this question comes from the languages he uses: he reads and writes Russian and English and is learning Korean. When his mentors offered him five research directions on language and neural networks and recommended this one, he chose it and explained: "I'm interested myself. I know that letters from different alphabets weigh differently (Latin and Cyrillic). It's interesting what happens with Korean" (1 October 2026; the originals of all quotations are given in the Note). From five possible questions he then chose the second, a comparison of Korean, Russian and English, explaining that it shows "who will have to write more symbols for their goal, a prompt and so on" (2 October 2026).

Earlier studies measured the token cost of many languages against English [1, 2, 4]. Their tables already show that the order of Russian and Korean can differ between tokenizers [2, Appendix C], but they did not compare Russian with Korean in the tokenizers of the newer models, and they report one number per language and tokenizer, without testing whether the order holds sentence by sentence. Russian and Korean use two very different writing systems: Russian is written in the Cyrillic alphabet, Korean in Hangul, where letters are grouped into syllable blocks, so that a Korean sentence has far fewer characters than its Russian translation. This paper asks two questions: (1) how many times more tokens Russian and Korean need than English for the same sentences, and (2) which of the two needs more, and whether the answer is the same in the tokenizers of GPT-4, Llama 3, Gemma 1/2 and GPT-4o/GPT-5.

Related work

Token cost across languages. Petrov et al. [2] defined the tokenization premium of a language as the ratio of the lengths of a sentence and its translation after tokenization and measured it for many tokenizers on FLORES-200 dev and devtest (about 2,000 sentences, including the ones used here). They found that "the same text translated into different languages can have drastically different tokenization lengths, with differences up to 15 times in some cases" [2, p. 36963], that the tokenizer of GPT-4 needs about 1.6 times as many tokens for Italian as for English, 2.6 times for Bulgarian and 3 times for Arabic, and that such differences make some languages more expensive, slower and able to fit less content into a model's context [2, p. 36964]. Their extended tables give GPT-4's tokenizer (cl100k_base) a premium of 2.49 for Russian and 2.38 for Korean, so that Russian costs more; the earlier GPT-2 tokenizer also charges Russian more (5.74 against 5.07), while the tokenizer of the first LLaMA charges Korean more (3.18 against 1.64) [2, Appendix C, pp. 36987, 36989]. Ahia et al. [1] studied what this means for prices: they found that commercial pricing by tokens favours Latin-script languages, that speakers of many supported languages "are overcharged while obtaining poorer results", and that for most mid-resourced Indic languages with non-Latin scripts the cost is close to five times that of English [1, pp. 9904, 9908]. Teklehaymanot and Nejdl [4] measured the GPT-4 tokenizer on the same 1,012 FLORES-200 devtest sentences in more than 200 languages and found that non-Latin and morphologically complex languages often need three to five times as many tokens as English. Rust et al. [3] measured the "fertility" of a tokenizer, the average number of pieces into which it cuts a word. For the

multilingual model mBERT, trained on Wikipedia, fertility was much higher than for monolingual models in Arabic, Finnish, Korean, Russian and Turkish and lowest in English, which the authors explain by the amount of English the model had seen and by English being "morphologically poor" [3, p. 3122].

How the four tokenizers are built. Most current tokenizers use byte pair encoding (BPE), which starts from single symbols and repeatedly merges the most frequent pair into a new vocabulary entry [5]; frequent words therefore become single tokens, while rare words are cut into several. GPT-4's cl100k_base and the newer o200k_base are byte-level BPE tokenizers published by OpenAI in the tiktoken library, whose mapping assigns o200k_base to GPT-4o and GPT-5 and cl100k_base to GPT-4 [6]. The Llama 3 vocabulary of 128K tokens "combines 100K tokens from the tiktoken tokenizer with 28K additional tokens to better support non-English languages" [7, p. 7]. Gemma uses a SentencePiece tokenizer [8] with a vocabulary of 256K entries that "relies on byte-level encodings for unknown tokens" [9, p. 3]. The tokenizers of GPT-4, GPT-4o/GPT-5 and Llama 3 work on the UTF-8 bytes of the text, in which a Latin letter takes one byte, a Cyrillic letter two and a Hangul syllable three, so a character that is not in the vocabulary as a whole is split into byte pieces; Gemma's tokenizer falls back to bytes only for characters it does not contain.

In this journal. Assan [10] measured how much worse open language models understand Kazakh when it is written in the new Latin alphabet instead of Cyrillic. That design showed that the script matters but, as the author noted, it could not separate tokenization from other causes [10, p. 210]. The present paper measures the tokenization step directly, for Russian and Korean rather than Kazakh.

The author's forecast

On 2 October 2026, after the question was fixed and before anything was counted, the author was asked for a forecast: which of the three languages costs the most, roughly how many times more than English, and why. On 3 October 2026 he wrote: "1) An English word weighs 1 token (because the models were trained on English and the vocabulary is bigger). 2) A Korean word will cost more than a Russian one because of the small vocabulary, and the model will split words into syllables and then into letters, spending tokens. But the language itself is compact, so it is cheaper than Russian. 3) A Russian word will cost more because the vocabulary is small, and the model will split words into syllables and spend even more tokens."

The forecast contains four claims that the data can test: (F1) English is the cheapest of the three languages, at about one token per word; (F2) a Korean word needs more tokens than a Russian word; (F3) a Korean sentence is nevertheless shorter, so a Russian sentence costs more than a Korean one; (F4) the reason is the vocabulary available to each language. Point 3 can also be read as saying that a Russian word costs even more than a Korean one; F2 follows point 2, and his supervisor read the forecast as English cheapest and Russian dearer than Korean. He gave no number for how many times more Russian and Korean cost than English. When he wrote the forecast he had been told that text in some writing systems is cut into more pieces than in others, and the direction he chose had been offered to him as a "tax on Korean", that is, as Korean needing more tokens than English; nothing had been said about Russian, and no count for any language had been shown to him.

Data and methods

Sentences. FLORES-200 is a benchmark for machine translation built from 3,001 sentences taken from English Wikimedia projects (842 articles), which professional translators translated into about 200 languages [11, 12, 13]. Its devtest part contains 1,012 sentences from 281 articles of Wikinews, Wikibooks and Wikivoyage. The files eng_Latn.devtest, rus_Cyrl.devtest and kor_Hang.devtest and the file metadata_devtest.tsv, which names the article of every sentence, were taken from the official archive (dl.fbaipublicfiles.com/nllb/flores200_dataset.tar.gz; licence CC BY-SA 4.0); the copies used were checked against a fresh download from the same address on 25 October 2026 and are identical [13]. On 5 October 2026 the author opened the three files himself, confirmed that each has 1,012 lines and, reading all three languages, that the first line is the same sentence in each. The Korean file contains 58 lines with doubled quotation marks (""), apparently quotation marks escaped in CSV style; the Russian file has one such line and the English file none. The Russian file also contains five lines (304, 342, 501, 511 and 594) with a word that mixes Latin look-alike letters with Cyrillic ones, and one line with a non-breaking space. The analysis uses the files as published, and the effect of repairing the quotation marks is reported in the Results.

Tokenizers. Four tokenizers were used: cl100k_base (GPT-4) and o200k_base (GPT-4o and GPT-5) from tiktoken [6], Llama 3 and Gemma. For Llama 3 the file tokenizer.json of the 8B model was used. For Gemma the tokenizer of Gemma 2B was used, which the second generation, Gemma 2, shares (the repositories of gemma-2b and gemma-2-2b list the same tokenizer file, and the Gemma 2 report says it uses "the same tokenizer as Gemma 1" [16, p. 3]); it is called Gemma 1/2 below. Both files were taken from public copies on Hugging Face whose files are identical to Meta's and Google's (the repositories list the same file identifiers; for Gemma, Google's tokenizer.model and the tokenizer.json of the copy gave identical counts on all 3,036 sentences). The vocabulary sizes are 100,277, 200,019, 128,256 and 256,000 entries. Later generations of the same families, Gemma 3 and 4 and Llama 4, use different tokenizers; they were added after the main analysis and are reported separately as exploratory. Special tokens such as start-of-text markers were not counted. A tokenizer only splits text; no language model was run in this study.

Plan written before the counts. On 6 October 2026 his supervisor wrote down the plan, and the author agreed to it the same day: each of the 1,012 sentences in each language goes through the tokenizers of GPT-4, Llama 3 and Gemma; for each language the mean number of tokens per sentence is computed; the result is how many times Russian and Korean cost more than English in each tokenizer and which of the two costs more; if the order is the same in all three, the conclusion is firm, and if not, it is reported as it is. The plan also offered to add a fourth tokenizer of a model the author uses; he named none. On 7 October 2026, asked by his supervisor, he counted the first sentence in English, Russian and Korean on the online tokenizer of OpenAI's developer platform, in its tab labelled "GPT-5.x & O1/3". His three counts (22, 26 and 30 tokens) are the same as those given by o200k_base, the encoding that tiktoken assigns to GPT-5 and GPT-4o [6], for the strings he entered, and on 9 October his supervisor added o200k_base as a fourth tokenizer, reported next to the three planned ones.

Measures. The main measure is the ratio of the mean number of tokens per sentence in Russian or Korean to that in English, which equals the ratio of the total numbers of tokens for

the whole set. The following checks were added after the results had been seen and are exploratory: 95% confidence intervals for each ratio from 10,000 bootstrap resamples of sentences (the three translations of a sentence are always resampled together) and, because sentences from the same article are related, from 10,000 resamples of whole articles; the number of sentences, and of articles, in which Korean needs more tokens than Russian, fewer, or the same, with an exact two-sided sign test (ties left out); the number of characters, UTF-8 bytes and space-separated units per sentence; the composition of each vocabulary, counting entries whose letters are all Cyrillic or all Hangul; the tokenizers of Gemma 3, Gemma 4 and Llama 4; and, added last, vocabulary entries split into single letters and longer pieces and the order of Russian and Korean in three groups of sentences by translation length. All dates are given in Almaty time.

Results

Cost against English

English was the cheapest language in all four tokenizers (Table 1, Figure 1). Its mean length hardly depends on the tokenizer, from 26.55 to 26.86 tokens per sentence. Russian needed 1.38 times (Gemma 1/2) to 2.46 times (GPT-4) as many tokens as English, and Korean 1.47 times (GPT-4o/GPT-5) to 2.37 times (GPT-4). The bootstrap intervals are narrow, at most ±0.03. The same holds sentence by sentence: in GPT-4 every one of the 1,012 Russian and Korean sentences needs more tokens than its English original; in the other three tokenizers Russian needs more tokens than English in 94.8% to 99.4% of sentences and Korean in 98.9% to 99.5%. For GPT-4 the ratios agree with the premiums of Petrov et al., measured on a larger part of the same benchmark: 2.49 for Russian and 2.38 for Korean [2, Appendix C].

Table 1. Mean tokens per sentence on the 1,012 FLORES-200 devtest sentences, ratio to English (95% bootstrap interval) and number of sentences in which Korean needs more tokens than Russian, fewer tokens, or the same number (GPT-4: cl100k_base; GPT-4o/5: o200k_base)

Tokenizer	English	Russian	Korean	Russian / English	Korean / English	Korean more / fewer / same
GPT-4	26.86	66.13	63.58	2.46 (2.43–2.49)	2.37 (2.34–2.39)	427 / 556 / 29
Llama 3	26.85	43.40	39.96	1.62 (1.60–1.63)	1.49 (1.47–1.50)	306 / 647 / 59
Gemma 1/2	26.72	36.85	43.81	1.38 (1.36–1.39)	1.64 (1.62–1.66)	837 / 138 / 37
GPT-4o/5	26.55	37.74	39.10	1.42 (1.41–1.44)	1.47 (1.46–1.49)	577 / 379 / 56

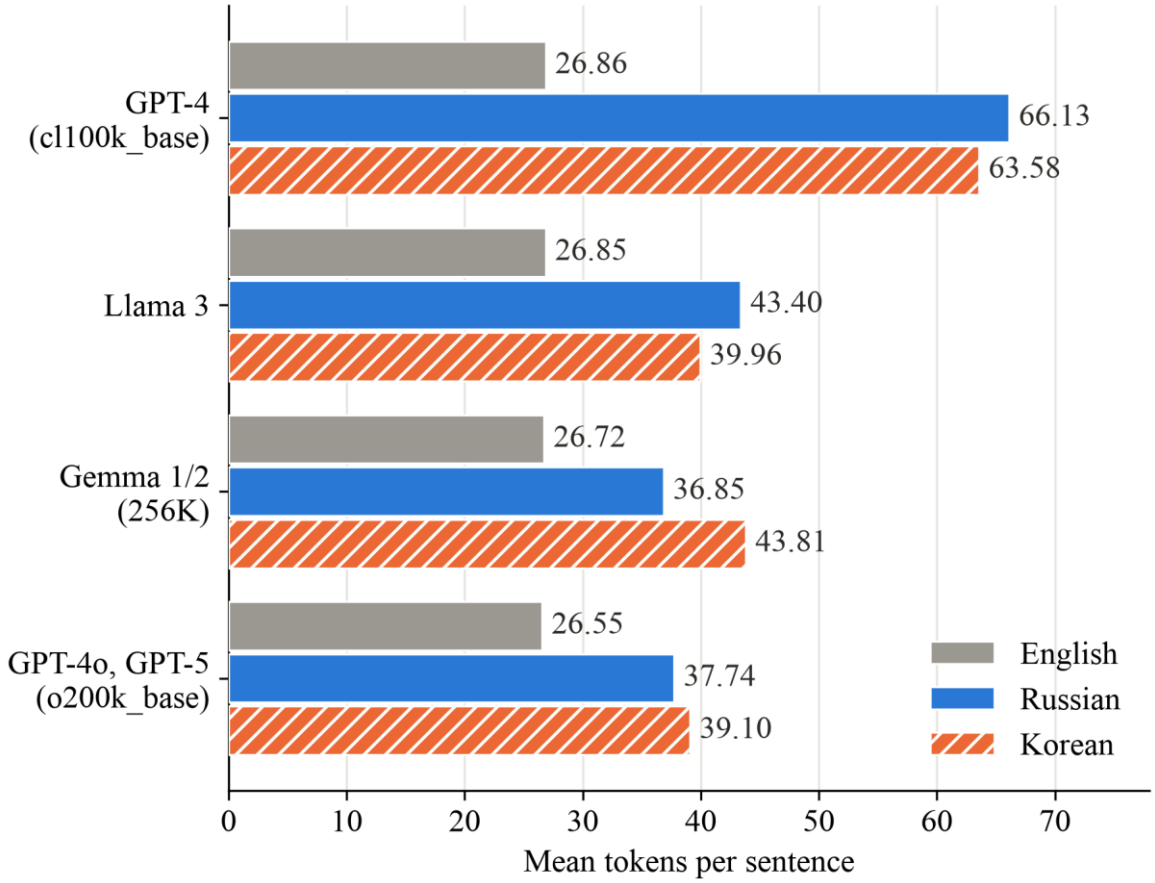


Figure 1. Mean number of tokens per sentence for the same 1,012 FLORES-200 devtest sentences in English, Russian and Korean, by tokenizer. Values from Table 1.

Russian or Korean?

The order of Russian and Korean is not the same in the four tokenizers. In GPT-4 and Llama 3, Russian needs more tokens than Korean (Korean/Russian 0.96, 95% interval 0.95–0.97, and 0.92, 0.91–0.93). In Gemma 1/2 and in GPT-4o/GPT-5, Korean needs more (1.19, 1.18–1.20, and 1.04, 1.02–1.05). Each order also holds in a majority of single sentences, large in some tokenizers and slim in others (Table 1, last column): Russian is costlier in 556 (55%) and 647 (64%) of 1,012 sentences in GPT-4 and Llama 3, Korean in 837 (83%) and 577 (57%) in Gemma 1/2 and GPT-4o/GPT-5, and in all four cases the exploratory sign test gives $p < 0.001$ (the largest p is 0.00004). The difference is largest in Gemma 1/2, which is at the same time the cheapest of the four tokenizers for Russian and the most expensive for Korean after GPT-4.

The two languages differ greatly in length before tokenization. A Korean sentence has on average 65.2 characters, a Russian one 140.4 and an English one 130.4. In UTF-8 bytes, which is what byte-level tokenizers fall back to, the order changes: Korean 155.9, Russian 257.3, English 130.5. In space-separated units, a Korean sentence has 15.3, a Russian one 19.5 and an English one 21.6. These units are not equivalent: a Korean unit usually carries the grammatical particles that Russian and English write as separate words or endings. A unit of Korean text needs more tokens than a unit of Russian text in every tokenizer (GPT-4: 4.17 against 3.40; Llama 3: 2.62 against 2.23; Gemma 1/2: 2.87 against 1.89; GPT-4o/GPT-5: 2.56 against 1.94), while an English unit needs 1.23 to 1.24 tokens. Whether a Korean sentence as a

whole is cheaper than a Russian one therefore depends on whether its shorter length makes up for the higher cost of each unit, and this happens in two tokenizers out of four.

The old and the new OpenAI tokenizer

The tokenizer of GPT-4o and GPT-5 needs fewer tokens than that of GPT-4 for every one of the 1,012 Russian and the 1,012 Korean sentences. On average it uses 43% fewer tokens for Russian (37.74 against 66.13) and 38.5% fewer for Korean (39.10 against 63.58), while English changes very little (1% fewer on average; fewer tokens in 281 of 1,012 sentences and more in 63). As a result, Russian moved from the most expensive of the three languages in GPT-4 to the second in GPT-4o/GPT-5. OpenAI's announcement of GPT-4o showed the change on one example sentence per language: from 39 to 23 tokens in Russian, from 45 to 27 in Korean and from 27 to 24 in English [15]. In that single example Korean needs more tokens than Russian in both tokenizers, while over the 1,012 sentences Russian needs more in GPT-4, which shows why one sentence is not enough.

The author's own example shows the same change (Table 2). On 13 October 2026 he chose a short question to a chatbot, "как сказать привет на английском" ("how to say hi in English"), wrote it also in English and Korean and counted the three versions on the online tokenizer in the tab for GPT-5; on 14 October he counted the same phrases in the tab for GPT-4. In GPT-4 the Russian phrase needs 16 tokens against 6 for English; in GPT-4o/GPT-5 both need 6. The Korean phrase, with only 11 characters, needs 10 tokens in GPT-4 and 7 in GPT-4o/GPT-5. (His screenshot for GPT-4o/GPT-5 shows 8 tokens and 12 characters because a space was left at the end of the Korean phrase; without it the count is 7.) The phrase also shows the same split as the whole set: in GPT-4 and Llama 3 the Korean phrase is shorter than the Russian one, and in Gemma 1/2 and GPT-4o/GPT-5 it is longer.

Table 2. Tokens for the author's example phrase in the four tokenizers

Phrase	Characters	GPT-4	Llama 3	Gemma 1/2	GPT-4o/5
как сказать привет на английском	32	16	9	5	6
how to say hi in english	24	6	6	6	6
영어로 인사하는 방법	11	10	6	6	7

Vocabulary

The vocabularies contain very different numbers of pieces written in each script (Table 3). The vocabulary of GPT-4 contains 729 entries whose letters are all Cyrillic and 281 whose letters are all Hangul. The vocabulary of GPT-4o/GPT-5 is twice as large overall, but it contains 14,209 Cyrillic entries, almost 20 times more, and 2,360 Hangul entries, about 8 times more. In the four planned tokenizers both the total size of the vocabulary and the ratio of Cyrillic to Hangul entries line up with the order of Russian and Korean: GPT-4 and Llama 3, where Russian costs more, have the smaller vocabularies (100K and 128K entries) and ratios of 2.59 and 2.88, while Gemma 1/2 and GPT-4o/GPT-5, where Korean costs more, have the larger ones (256K and 200K) and ratios of 5.50 and 6.02. The two properties go together here, so these four tokenizers cannot show which of them matters. The counts are also not like for like: most Hangul entries are single syllables (2,016 of 2,387 in Gemma 1/2, 219 of 281 in GPT-4), while almost all Cyrillic entries are pieces of two or more letters. Counting only such

longer pieces, the ratio of Cyrillic to Hangul entries is 10.1 in GPT-4, 5.3 in Llama 3, 34.6 in Gemma 1/2 and 11.4 in GPT-4o/GPT-5.

Table 3. Vocabulary entries whose letters are all Cyrillic or all Hangul, and the order of Russian and Korean (GPT-4: cl100k_base; GPT-4o/5: o200k_base; Gemma 3 has 262,145 entries and Gemma 4 has 262,144, and the two give identical counts)

Tokenizer	Vocabulary size	Cyrillic entries	Hangul entries	Cyrillic / Hangul	Korean / Russian tokens
GPT-4	100,277	729	281	2.59	0.96
Llama 3	128,256	6,503	2,259	2.88	0.92
Gemma 1/2	256,000	13,130	2,387	5.50	1.19
GPT-4o/5	200,019	14,209	2,360	6.02	1.04
Gemma 3 and 4	262,145	13,397	4,678	2.86	1.01
Llama 4	201,135	22,830	5,447	4.19	1.03

Newer tokenizers of the same families (exploratory)

The tokenizers of Gemma 3 and Gemma 4 (262K entries) give identical counts on all three files: 26.79 tokens per English sentence, 36.58 per Russian and 36.83 per Korean, so the two languages are almost equal (Korean/Russian 1.007; Korean costs more in 496 sentences and Russian in 441, sign test $p = 0.08$). The tokenizer of Llama 4 (201K entries) needs 26.73, 32.82 and 33.83 tokens: Korean costs slightly more than Russian (1.03; 564 against 386 sentences, $p < 0.001$), the opposite of Llama 3. English remains the cheapest on average in both, and in single sentences in 90.5% (Gemma 3 and 4) and 79.7% (Llama 4). The Gemma 3 report describes its tokenizer as "more balanced for non-English languages" [17, p. 3]; it has almost twice as many Hangul entries as Gemma 1/2 (4,678 against 2,387), and Korean/Russian falls from 1.19 to 1.01. Across all six distinct tokenizers, the size of the vocabulary separates the two orders without exception: the two vocabularies of up to 128K entries make Russian costlier, and the four of 200K entries or more make Korean costlier or about equal. The ratio of Cyrillic to Hangul entries does not separate them as cleanly: Gemma 3 and 4 (2.86) and Llama 3 (2.88) have almost the same ratio but give different results, and counting only pieces of two or more letters, Gemma 3 and 4 (5.8) and Llama 4 (5.6) have lower ratios than GPT-4 (10.1). With six tokenizers from three developers, neither property can be singled out as the cause (Table 3).

Robustness

When the 58 Korean lines with doubled quotation marks are repaired, the mean Korean length falls by 0.06 tokens in every tokenizer, the Korean ratios to English change by about 0.002, and the number of sentences in which Korean costs more than Russian changes by at most two. Resampling whole articles instead of sentences moves each end of the 95% intervals of the ratios to English by at most 0.007 (the widest becomes 2.43–2.50 for Russian in GPT-4). Counted by article, Korean costs more than Russian in 100 of 281 articles and Russian in 175 in GPT-4 (sign test $p < 0.001$), 62 against 214 in Llama 3 ($p < 0.001$), 272 against 8 in Gemma 1/2 ($p < 0.001$) and 174 against 102 in GPT-4o/GPT-5 ($p < 0.001$). No conclusion changes. The translations themselves matter more: when the sentences are split into three equal groups by how long the Korean translation is compared with the Russian one,

Korean/Russian rises from the group with relatively short Korean translations to the group with relatively long ones in every tokenizer (GPT-4: 0.81, 0.97, 1.14; Llama 3: 0.79, 0.93, 1.06; Gemma 1/2: 1.04, 1.19, 1.36; GPT-4o/GPT-5: 0.89, 1.04, 1.19). Where the two languages differ by only a few percent, as in GPT-4, GPT-4o/GPT-5, Llama 4 and Gemma 3 and 4, the order therefore depends on these particular translations. What holds in every group is the ranking of the tokenizers: Gemma 1/2 makes Korean the most expensive relative to Russian, then GPT-4o/GPT-5, then GPT-4, then Llama 3.

Discussion

The forecast against the data. The first claim (F1) was right in its main part: English is the cheapest language in all four tokenizers and the cheapest of the three in at least 94.4% of individual sentences (955 of 1,012 in Gemma 1/2, all 1,012 in GPT-4). For Korean against English, the way the direction had been offered to him already suggested this; the part he added himself is that Russian, too, costs more than English. An English unit needs 1.23 to 1.24 tokens, about a quarter more than the one token per word the author expected. The second claim (F2) holds for space-separated units: a unit of Korean text needs more tokens than a unit of Russian text in every tokenizer. This does not test his claim about words, because a Korean unit carries particles and more content than a Russian one (15.3 units per Korean sentence against 19.5 per Russian), so the per-unit result follows largely from the length of the units. The third claim (F3), that a whole Russian sentence costs more than a Korean one, holds for GPT-4 and Llama 3 and is reversed for Gemma 1/2 and GPT-4o/GPT-5. The plan said that if the three planned tokenizers gave the same order the conclusion would be firm; they did not (GPT-4 and Llama 3 against Gemma 1/2), so the answer to the second question is that it depends on the tokenizer. The newer tokenizers of the same families confirm this: Llama 4 reverses the order of Llama 3, and in Gemma 3 and 4 the two languages are almost equal. The fourth claim (F4), that the vocabulary is the reason, is consistent with the data but not tested by them. Within one developer the vocabulary clearly matters: OpenAI's newer vocabulary has many more Cyrillic and Hangul entries and needs about 40% fewer tokens for both languages, and Gemma 3, whose report describes its tokenizer as more balanced for non-English languages, almost evens out Russian and Korean. Across developers, the size of the vocabulary and the number of Cyrillic entries for each Hangul entry go together in the four planned tokenizers, and with six tokenizers neither can be singled out (Table 3). Petrov et al. expected languages with a script of their own, such as Korean, to pay more, because "no other language uses the Korean script" and so shares its tokens [2, p. 36971]; Russian shares its alphabet with Ukrainian, Kazakh and other languages.

The author's explanations. On 9 October 2026, when his mentors showed him the first results, the author wrote: "Probably the models' ways of reading differ. And for users this affects the number of requests they will be able to make." Asked to expand this, and in particular to say which texts the models had learned from and which languages they had seen most often, he added on 12 October: "The models learned to read on Wikipedia articles. They saw English most often. The models' tokenization algorithms are different, so the number of tokens spent is different too." His mentors pointed out that current models are trained on large amounts of text from the internet, of which Wikipedia is only a part. Older multilingual models such as mBERT were indeed trained on Wikipedia and cut Korean and Russian into more pieces than English [3, p. 3122]; for the tokenizers studied here, the data on which they were built are not published. English dominates text on the web: in the latest monthly crawl of

Common Crawl (CC-MAIN-2026-43), 42.56% of pages have English as their main language, 6.30% Russian and 0.88% Korean [14]. These shares describe the web in 2026, not the data from which the tokenizers were built. The structure of the languages offers another explanation that the data cannot separate from this one: Russian words change their endings, and Korean attaches particles to words, so both produce many word forms that are rarer than English ones, and Rust et al. likewise point to morphology as one reason why English is cut into fewer pieces [3, p. 3122].

After counting his example phrase, the author concluded on 13 October: "Even though Korean has fewer characters, it has more tokens, which means a Korean user will lose more money and time." This holds on average for GPT-4o/GPT-5 and Gemma 1/2, but not for GPT-4 and Llama 3. On 16 October his supervisor summed up his counts of the phrase in both OpenAI tokenizers and asked why the new tokenizer reduced Russian more than Korean. On 17 October he answered: "Russian is much more popular than Korean and has more speakers. The models' vocabularies have grown. Now a Russian-speaking user can use AI more economically." The second sentence is supported directly by Table 3: the new OpenAI vocabulary contains almost 20 times more Cyrillic entries and 8 times more Hangul entries than the old one, and Russian lost a larger share of its tokens than Korean (43% against 38.5%). The first sentence agrees with the web shares above, where Russian pages are about seven times as common as Korean ones; the number of speakers was not checked in this study. The explanation is consistent with the data but not proven by it: no published description of how the new vocabulary was chosen was found.

What it means for users. Services charge a fixed amount for a given number of tokens, whatever the language [1, p. 9908], so the ratios in Table 1 are also ratios of cost and, roughly, of processing time. With the same token limit, a user who writes in Russian to a model with GPT-4's tokenizer can send about 41% of the text that an English user can (1/2.46), and a Korean user about 42%. With the tokenizer of GPT-4o and GPT-5 the shares rise to 70% for Russian and 68% for Korean. The penalty has shrunk but has not disappeared, and on these sentences it is now slightly larger for a Korean user than for a Russian one.

Limitations. The study uses one set of 1,012 sentences from Wikinews, Wikibooks and Wikivoyage and their translations; other kinds of text, such as chat messages or code, may give different ratios, and the same sentence can be translated in ways that differ in length, which, as Petrov et al. also note [2, p. 36971], changes the counts. The counts are for the text alone, without the extra tokens that chat services add around a message, and only for input text; the answers a model writes were not measured. Four tokenizers were studied in the planned analysis and three more were added afterwards; for each generation only the tokenizer of one model was used. Counting vocabulary entries by script is a rough description: an entry may be used rarely, a single Hangul syllable counts the same as a long Cyrillic piece, and Cyrillic entries also serve Ukrainian, Kazakh and other languages. The explanations discussed above are hypotheses, because the training data of the tokenizers are not public and their design is published only in outline. The statistical tests and intervals beyond the planned means are exploratory; sentences from the same article are related, but the article-level analysis gives the same results.

Conclusion

On the same 1,012 sentences, Russian needs 1.38 to 2.46 times and Korean 1.47 to 2.37 times as many tokens as English, depending on the tokenizer, and English is the cheapest

language in every tokenizer studied. Which of Russian and Korean costs more is set by the tokenizer, as earlier tables already suggested for older tokenizers [2]: Russian in GPT-4 and Llama 3, Korean in Gemma 1/2, GPT-4o/GPT-5 and, in an exploratory check, Llama 4, with the two almost equal in Gemma 3 and 4; where the difference is only a few percent, it depends on the translations. The author's forecast was right that English is the cheapest and, for two of four tokenizers, that a Russian sentence costs more than a Korean one; his claim about single words could be checked only on space-separated units, which in Korean include particles, so it is not a separate result. OpenAI's newer tokenizer uses 43% fewer tokens for Russian and 38.5% fewer for Korean, and its vocabulary contains many more Cyrillic and Hangul pieces, which agrees with the author's explanation that the vocabularies have grown. In these tokenizers the order of Russian and Korean goes together both with a larger vocabulary and with more Cyrillic entries for each Hangul entry, and the data cannot say which of the two matters. The cost of writing to a language model in a language other than English should therefore be measured for each model rather than assumed.

Note: original quotations

The author's words were written in Russian and are quoted in the text in translation. A slash marks a line break in the original message.

Choice of direction (1 October 2026): «давайте первую. интересно самому. я знаю что буквы из разных алфавитов весят по разному(латиница и кириллица). интерсено что с корейским»

Choice of question (2 October 2026): «давайте 2 / кому придется написать больше символов для своей цели, промпта и тд»

Forecast (3 October 2026): «1)английское слово весит 1 токен (т.к. модели на англ обучались и словарь больше) / 2)корейское слово будет стоить дороже русского изза маленького словаря и модель будет делить слова на слоги, а потом на буквы, тратя токены. Но сам язык компактный поэтому дешевле русского / 3)русское слово будет стоить дороже т.к. словарь маленький и модель будет делить слова на слоги и тратить еще больше токенов»

First interpretation (9 October 2026): «наверно способы чтения у моделек меняются. а для пользователей это влияет на кол-во запросов которые они смогут сделать»

Explanation (12 October 2026): «модели учились читать на статьях в Википедии. Английский язык они видели чаще всего. алгоритмы токенизации у моделей разные, поэтому кол-во токенов потраченных тоже.»

Example (13 October 2026): «несмотря на меньшее кол-во символов у корейского, кол-во токенов больше значит корейский пользователь потеряет больше денег и времени»

The new tokenizer (17 October 2026): «русский язык намного популярнее корейского и носителей его больше. словари моделей увеличились. теперь русскоязычный пользователь может более выгоднее использовать ии»

Data availability

FLORES-200 is publicly available under the CC BY-SA 4.0 licence [13]. The tokenizers are publicly available through the tiktoken library [6] and on Hugging Face. The per-sentence token counts and the analysis program that produces every number in this paper are available from the author's mentors on request.

Author contributions and use of AI tools

The author chose this study from five research directions proposed by his mentors, who recommended this one, and chose its question from five options they formulated (1 and 2 October 2026). He wrote the forecast (3 October), found, with search terms given by his supervisor, the papers [1] and [2] and summarised them (4 October), opened the three data files and checked that each has 1,012 lines and that the first sentence is the same in all three (5 October), agreed to the analysis plan (6 October), counted the first sentence of the set and his example phrase on OpenAI's online tokenizer (7, 13 and 14 October), chose the example phrase, and wrote the interpretations quoted in the Discussion (9, 12, 13 and 17 October). His scientific supervisor, Assyltay Sydyk, a mentor in the ADMIT mentoring programme, proposed the directions, the question options and the target journal, wrote the analysis plan and, working with a generative AI tool (Claude, Anthropic), wrote the counting program and computed the results of 9 October 2026. A second mentor of the same programme, who took over the project after a change of mentors, working with the same tool, re-ran every count from the data, added the checks described as exploratory in the Methods, made the tables and the figure, translated the author's messages quoted in this paper and wrote the whole text of the paper in English. No number, quotation or reference was produced by a language model without being checked: every number computed for this paper comes from the output of the analysis program, and on 25 and 26 October 2026 every number taken from other works and each statement about a cited work was checked against its published text. The author did not use AI tools or machine translation when writing his own messages and summaries quoted in this paper (his statement, 27 October 2026). The author has read the final manuscript and takes responsibility for it.

Conflicts of interest

No research funding was received for this study. The author takes part in ADMIT, a paid mentoring programme; the scientific supervisor and the second mentor named in Author contributions work in this programme, and they did the analysis and drafted the text as described there. The author declares no other conflicts of interest.

References:

1. Ahia O., Kumar S., Gonen H., Kasai J., Mortensen D.R., Smith N.A., Tsvetkov Y. Do all languages cost the same? Tokenization in the era of commercial language models // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023. P. 9904-9923. DOI: 10.18653/v1/2023.emnlp-main.614.

2. Petrov A., La Malfa E., Torr P., Bibi A. Language model tokenizers introduce unfairness between languages // Advances in Neural Information Processing Systems 36 (NeurIPS 2023). 2023. P. 36963-36990. DOI: 10.52202/075280-1608.

3. Rust P., Pfeiffer J., Vulić I., Ruder S., Gurevych I. How good is your tokenizer? On the monolingual performance of multilingual language models // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021. P. 3118-3135. DOI: 10.18653/v1/2021.acl-long.243.

4. Teklehaymanot H.K., Nejdl W. Tokenization disparities as infrastructure bias: How subword systems create inequities in LLM access and efficiency. arXiv:2510.12389, 2025. URL: https://arxiv.org/abs/2510.12389 (accessed: 25.10.2026).

5. Sennrich R., Haddow B., Birch A. Neural machine translation of rare words with subword units // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016. P. 1715-1725. DOI: 10.18653/v1/P16-1162.

6. OpenAI. tiktoken, GitHub repository (README and tiktoken/model.py). URL: https://github.com/openai/tiktoken (accessed: 25.10.2026).

7. Grattafiori A., Dubey A., Jauhri A., Pandey A. [et al.] The Llama 3 herd of models. arXiv:2407.21783, 2024. URL: https://arxiv.org/abs/2407.21783 (accessed: 25.10.2026).

8. Kudo T., Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2018. P. 66-71. DOI: 10.18653/v1/D18-2012.

9. Gemma Team, Mesnard T., Hardin C., Dadashi R. [et al.] Gemma: Open models based on Gemini research and technology. arXiv:2403.08295, 2024. URL: https://arxiv.org/abs/2403.08295 (accessed: 25.10.2026).

10. Assan R. The alphabet tax: A controlled benchmark of how Kazakhstan's Cyrillic-to-Latin script reform degrades open-model comprehension of Kazakh // Central Asian Scientific Journal. 2026. No. 3(31). P. 206-212.

11. NLLB Team, Costa-jussà M.R., Cross J., Çelebi O. [et al.] No language left behind: Scaling human-centered machine translation. arXiv:2207.04672, 2022. URL: https://arxiv.org/abs/2207.04672 (accessed: 25.10.2026).

12. Goyal N., Gao C., Chaudhary V., Chen P.-J., Wenzek G., Ju D., Krishnan S., Ranzato M., Guzmán F., Fan A. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation // Transactions of the Association for Computational Linguistics. 2022. Vol. 10. P. 522-538. DOI: 10.1162/tacl_a_00474.

13. Facebook Research. FLORES, GitHub repository (README and flores200/README.md). URL: https://github.com/facebookresearch/flores (accessed: 25.10.2026).

14. Common Crawl. Statistics of Common Crawl monthly archives: distribution of languages. URL: https://commoncrawl.github.io/cc-crawl-statistics/plots/languages (accessed: 25.10.2026).

15. OpenAI. Hello GPT-4o. 13.05.2024. URL: https://openai.com/index/hello-gpt-4o/ (accessed: 26.10.2026).

16. Gemma Team, Riviere M., Pathak S., Sessa P.G. [et al.] Gemma 2: Improving open language models at a practical size. arXiv:2408.00118, 2024. URL: https://arxiv.org/abs/2408.00118 (accessed: 26.10.2026).

17. Gemma Team, Kamath A., Ferret J., Pathak S. [et al.] Gemma 3 technical report. arXiv:2503.19786, 2025. URL: https://arxiv.org/abs/2503.19786 (accessed: 26.10.2026).

МАЗМҰНЫ – СОДЕРЖАНИЕ – CONTENT

ӘЛЕУМЕТТІК-ЭКОНОМИКАЛЫҚ ЖӘНЕ ҚҰҚЫҚ ҒЫЛЫМДАРЫ - СОЦИАЛЬНО-ЭКОНОМИЧЕСКИЕ И ЮРИДИЧЕСКИЕ НАУКИ - SOCIO-ECONOMIC AND LEGAL SCIENCES

Төлегенов Тимур Сейфуллаулы
ДИНАМИКА ТУРИСТСКИХ ПОТОКОВ И ЭФФЕКТИВНОСТЬ ИНФРАСТРУКТУРЫ ТУРКЕСТАНА: ОТ ЭКСТЕНСИВНОГО РОСТА К СМАРТ-УПРАВЛЕНИЮ 12

Alexander Gorevoy
WHICH INEQUALITY STATISTICS SURVIVE AN ADMINISTRATIVE BOUNDARY CHANGE? EVIDENCE FROM KAZAKHSTAN, 2010–2023 16

Rakhimov Djalil
DOES THE SHIFT TO CASHLESS PAYMENTS GO TOGETHER WITH FASTER GROWTH IN HOUSEHOLD CREDIT? EVIDENCE FROM THE REGIONS OF KAZAKHSTAN, 2023-2025 28

Kazhgerey Asan
HOW MANY FOOD BASKETS DOES AN AVERAGE WAGE BUY? WAGES AND FOOD PRICES IN THE 20 REGIONS OF KAZAKHSTAN, 2024-2026 39

Tagimova Alua
WHO WAS LEFT BEHIND? EDUCATION AND FINANCIAL INCLUSION IN KAZAKHSTAN AND THREE CENTRAL ASIAN NEIGHBOURS, 2021-2024 46

Wei Wei, Mynzhanova Gulzhakan, Niyetalina Gaukhar
DIGITAL–PHYSICAL INTEGRATION AND ECONOMIC PERFORMANCE IN INDUSTRIAL INTEGRATED ENERGY SYSTEMS: A SCENARIO-BASED ANALYSIS OF SMART GRID CONTROL, BATTERY ENERGY STORAGE, AND DIGITAL TWINS 53

Xu Fumin, Wei Wei, Ruziyeva Elvira, Moldabekova A.
FROM CULTURAL RESOURCES TO GLOBAL CONSUMER VALUE: THE BRIDGE–IMC MODEL AND THE INTERNATIONALIZATION OF BLACK MYTH: WUKONG UNDER THE BELT AND ROAD INITIATIVE 66

Ibraimkhan Alikhan
WHICH CRIMES CAN A 14-YEAR-OLD BE TRIED FOR? THE LIST IN ARTICLE 15 OF KAZAKHSTAN'S CRIMINAL CODE COMPARED WITH 30 OTHER JURISDICTIONS 82

Yerboluly Madi
DID IMPORTED CARS REALLY GET CHEAPER? CHINESE CARS AND THE AVERAGE CUSTOMS VALUE OF PASSENGER CARS IMPORTED INTO KAZAKHSTAN, 2015-2025 93

Салимбаев Ербол Аманжолович
АНАЛИЗ ИНВЕСТИЦИОННОЙ АКТИВНОСТИ КАК ИНДИКАТОР ИЗМЕНЕНИЙ ПОТЕНЦИАЛЬНОГО СПРОСА НА КОНСАЛТИНГОВЫЕ УСЛУГИ 101

Rambergenov Ramil
REGIONAL NEW-FIRM FORMATION IN UZBEKISTAN AFTER THE 2017 REFORMS: PERSISTENCE IN SCALE, CONVERGENCE IN INTENSITY, AND RECOMPOSITION TOWARD TRADE 109

Abdesh Dinara
FROM A WOMEN'S LEAD TO A MEN'S LEAD: THE GENDER GAP IN DIGITAL PAYMENTS IN KAZAKHSTAN AND THREE CENTRAL ASIAN NEIGHBOURS, 2021-2024 118

Amangeldi Aidana
DID SMALL BUSINESS IN KAZAKHSTAN REALLY SHRINK IN 2026? THE NEW TAX CODE, REGIONAL TAX RATES AND THE FALL IN THE NUMBER OF INDIVIDUAL ENTREPRENEURS 126

Cheng Che
BUILD, BUY, OR PARTNER? ENTRY MODE SELECTION AND RISK GOVERNANCE FOR WIND POWER INVESTMENT IN KAZAKHSTAN **135**

Wang Liang
FROM CAPACITY EXPANSION TO SYSTEM VALUE: COORDINATED DEVELOPMENT OF RENEWABLE ENERGY, GRID FLEXIBILITY AND ENERGY STORAGE IN KAZAKHSTAN **148**

Chen Jun
CONSTRUCTION-PERIOD CREDIT RISK MANAGEMENT FOR LARGE INFRASTRUCTURE PROJECT FINANCE IN KAZAKHSTAN **161**

Alniyaz Danial
HOW MUCH FUEL DOES A HYBRID REALLY SAVE? A SAME-MODEL COMPARISON OF HYBRID AND GASOLINE CARS IN US FUEL ECONOMY **175**

Assem Yergali, Maral Assanova
DIGITAL TRANSFORMATION AND SUSTAINABLE BUSINESS MODEL INNOVATION IN SMALL AND MEDIUM-SIZED ENTERPRISES: A SYSTEMATIC REVIEW OF DYNAMIC CAPABILITIES, VALUE CREATION, AND THE CREDIBILITY GAP IN EMERGING ECONOMIES **182**

Бейбіт Инара Алибекқызы, Едилов Саназар
ВОДА ИЗ ТУМАНА ИЛИ ВОДА ИЗ МОРЯ: ПОРОГОВАЯ СТОИМОСТЬ, ПРИ КОТОРОЙ ТУМАНОУЛОВИТЕЛИ КОНКУРЕНТОСПОСОБНЫ В МАНГИСТАУСКОЙ ОБЛАСТИ **195**

Zhenisbekova Aknur
DID A TENGE DEPOSIT OR THREE KAZAKHSTANI SHARES PROTECT SAVINGS FROM INFLATION? REAL RETURNS OF BANK DEPOSITS, KAZATOMPROM, KEGOC AND KASPI.KZ, 2020-2025 ..... **207**

Абилдаева Аселя
ОРГАНИЗАЦИОННАЯ ГОТОВНОСТЬ ВУЗОВ КАЗАХСТАНА К ВНЕДРЕНИЮ ЭНДАУМЕНТ-ФОНДОВ: КОНЦЕПТУАЛЬНАЯ РАМКА НА ОСНОВЕ ОБЗОРА ЛИТЕРАТУРЫ И АНАЛИЗА ДОКУМЕНТОВ **215**

Yang Juntao
DIGITAL ESG GOVERNANCE AND NEXT-GENERATION DIGITAL COMPETENCY AS DRIVERS OF SUCCESSION IN FAMILY AGRICULTURAL ENTERPRISES: A CONCEPTUAL FRAMEWORK ... **229**

Kairbek Mariya Islamkyzy, Kylbaev Erlan Serikovich
STATE SUPPORT FOR THE DIGITALIZATION OF MANUFACTURING IN THE REPUBLIC OF KAZAKHSTAN: AN ANALYSIS OF ITS COVERAGE OF MARKETING AND SALES ACTIVITIES **242**

Tilep Aiaru
WHEN DID IT EXPORTS TAKE OFF? ICT SERVICE EXPORTS OF KAZAKHSTAN, ARMENIA, GEORGIA AND THREE NEIGHBOURS, 2015-2024 ..... **255**

Anuarbek Murager
DID THE TENGE STOP FOLLOWING OIL? BRENT PRICES AND THE TENGE-DOLLAR RATE BEFORE AND AFTER FEBRUARY 2022 ..... **266**

Kaiyrbek Zhanadil
HOW ACCURATELY CAN OBSERVABLE LISTING DATA ESTIMATE FAIR PRICES FOR USED CARS IN KAZAKHSTAN? A HEDONIC MODEL OF 2,546 ONLINE LISTINGS **276**

Wen Yali
PERCEIVED ORGANIZATIONAL SUPPORT AND TURNOVER INTENTION AMONG KNOWLEDGE WORKERS IN A CHINESE PRIVATE TRAINING INSTITUTION: THE MEDIATING ROLE OF AFFECTIVE ORGANIZATIONAL COMMITMENT **289**

Құрманалиева Дана
ИСТОЧНИКИ ФИНАНСИРОВАНИЯ ИННОВАЦИОННЫХ ПРОЕКТОВ И КРИТЕРИИ ВЫБОРА ФИНАНСОВОЙ МОДЕЛИ **300**

ТЕХНИКАЛЫҚ ҒЫЛЫМДАР - ТЕХНИЧЕСКИЕ НАУКИ - TECHNICAL SCIENCE

Nurmakhan Almira
ADVANTAGES OF FIBER OPTIC COMMUNICATION OVER TRADITIONAL COPPER CABLES **307**

Kozhabai Bereke
SECURITY CHALLENGES IN IOT-BASED TELECOMMUNICATION NETWORKS **315**

Azamatuly Yerassyl
WHY AI-GENERATED CODE FAILS A BENCHMARK: A FAILURE ANALYSIS OF THREE OPEN LANGUAGE MODELS ON 30 HUMANEVAL+ TASKS **324**

Nurzhan Sanzhar
A FEW SHELLS, A FEW OPERATORS: WHERE LOW EARTH ORBIT IS CROWDED IN AN ALTITUDE-INCLINATION MAP OF 16,620 ACTIVE SATELLITES **333**

Askarbek Nurislam
CAN THE WEATHER TELL WHICH DAYS IN ASTANA WILL BE POLLUTED? CLASSIFYING PM2.5 EXCEEDANCE DAYS WITH MACHINE LEARNING AND CHECKING THE DATA AGAINST A GROUND MONITOR **339**

Iliyas Arman
SHOOTING ON THE MOVE: HOW MUCH AIM A ROBOT LOSES BY IGNORING ITS OWN VELOCITY, AND WHETHER A VIRTUAL-TARGET CORRECTION RECOVERS IT **350**

Ануар Әлімжан Бахытжанұлы, Сексенбаева Айша Курманбековна
СРАВНИТЕЛЬНЫЙ АНАЛИЗ СТАНДАРТОВ И МЕТОДИК ОБУЧЕНИЯ КИБЕРБЕЗОПАСНОСТИ: NICE FRAMEWORK, ECSF И ПРАКТИКО- ОРИЕНТИРОВАННЫЕ ТЕХНОЛОГИИ **356**

Mukhamedzhan Yerganat
GROWTH OF GITHUB DEVELOPER COMMUNITIES IN KAZAKHSTAN, UZBEKISTAN AND KYRGYZSTAN, 2020-2026: EVIDENCE FROM THE GITHUB INNOVATION GRAPH **366**

Kashkinov Alden
WHAT DECIDES HOW LONG A LANDER WORKS ON ANOTHER WORLD? SURFACE LIFETIMES OF 55 LANDERS ON THE MOON, VENUS, MARS, TITAN AND SMALL BODIES IN THE GCAT CATALOGUE, 1966-2025 ..... **378**

Kenesbek Miras
HOW MANY HERONIAN TRIANGLES HAVE A GIVEN LONGEST SIDE? A COMPLETE COUNT UP TO 2000 AND THE RIGHT TRIANGLES HIDDEN INSIDE **390**

Amirlan Seitkadyrov, Akezhan Omashev
AUTOMATIC MULTIPLE-CHOICE QUESTION GENERATION USING ARTIFICIAL INTELLIGENCE **401**

Idrissov Amirzhan
WHERE DO SMALL OPEN LANGUAGE MODELS BREAK ON OLYMPIAD MATHEMATICS? DOMAIN, DIFFICULTY AND MODEL SIZE ON OMNI-MATH **410**

Маматаева Рамина Адилгазыевна, Бидуәлі Алмат Алматұлы
ОБЗОР И СРАВНЕНИЕ ПОПУЛЯРНЫХ КРОССПЛАТФОРМЕННЫХ ФРЕЙМВОРКОВ ДЛЯ РАЗРАБОТКИ МОБИЛЬНЫХ ПРИЛОЖЕНИЙ **415**

Қасен Ғауһар Ғалымжанқызы, Алкеева Дана Мирасовна
КОМПЬЮТЕРЛІК КӨРУДІ ПАЙДАЛАНЫП,МЕХАНИКАЛЫҚ ДЕНЕ ЗАҢЫН АНЫҚТАУ АЛГОРИТМДЕРІН ӘЗІРЛЕУ **421**

Khalilzade Seymur Fazil oglu
ASSESSING THE IMPACT OF ELECTRIC VEHICLES ON ENERGY DEMAND IN HIGH-RISE RESIDENTIAL BUILDINGS **434**

Igilikov Adilkhan Igilikovich
PRE-TRAINED REPRESENTATIONS OF PRIOR-ADMISSION HISTORY IMPROVE ONE-YEAR, BUT NOT IN-HOSPITAL, MORTALITY PREDICTION IN ICU PATIENTS WITH CHRONIC DISEASE: A MIMIC-IV STUDY **449**

Nurgali Dias, Nurgali Miras, Akashov Aslan
PREDICTIVE TRAFFIC MANAGEMENT: EXPLORING SOLUTIONS TO TRAFFIC CONGESTION IN ASTANA, KAZAKHSTAN **461**

ПӘНДЕР АРАЛЫҚ ҒЫЛЫМДАР – МЕЖДИСЦИПЛИНАРНЫЕ НАУКИ – INTERDISCIPLINARY SCIENCES

Albiyev Asser
HOW MUCH OF AN ARTICLE COUNT IS HUMAN WORK? DECOMPOSING THE CREATION OF FIVE WIKIPEDIAS, 2010–2025 ..... **467**

Amangeldy Nurbolat
HOW MUCH OF KAZAKHSTAN'S URBAN BUILT STOCK PREDATES ITS SEISMIC CODES, AND WHERE DOES THAT STOCK SIT RELATIVE TO HAZARD? **477**

Torekhan Sabrina
WHERE ALMATY WALKS ITS OWN WAY: INFORMAL SHORTCUTS IN OPENSTREETMAP AND THE DETOURS THEY AVOID **483**

Ailan Salybekova
PARENTAL SUPPORT, PARENTAL CONTROL, STRESS AND EARLY ALCOHOL USE IN ADOLESCENT GIRLS: A NARRATIVE REVIEW WITH A FOCUS ON KAZAKHSTAN **488**

Yerboluly Madi
WHY IS KAZAKHSTAN'S GENDER GAP AT THE TOP OF CHESS SO SMALL? A BIRTH-COHORT COMPARISON ON THE FIDE RATING LIST **496**

Omashev Akezhan
REGIONAL HEALTH AND SOCIAL INDICATOR PATTERNS AND PCA- BASED MORTALITY MODELING IN KAZAKHSTAN, 2011–2024. A TWO- PART EXPLORATORY ECOLOGICAL ANALYSIS **501**

Мадиярова Жанел Биржановна
ДИПФЕЙК ТЕХНОЛОГИЯСЫНЫҢ ҚОҒАМДЫҚ ПІКІРГЕ ЫҚПАЛЫ ЖӘНЕ ОНЫ ҚҰҚЫҚТЫҚ РЕТТЕУ ТӘЖІРИБЕСІ **511**

Bolatov Sanzhar
HOW MUCH DOES KAZAKH SPEECH RECOGNITION LOSE WHEN THE MODEL GETS SMALLER? KAZAKH AGAINST 82 LANGUAGES IN THE PUBLISHED WHISPER RESULTS **521**

Ilya Tsoy
WHO PAYS MORE TOKENS FOR THE SAME SENTENCE? RUSSIAN, KOREAN AND ENGLISH IN FOUR LANGUAGE MODEL TOKENIZERS ON 1,012 FLORES-200 SENTENCES **530**

Assan Ramazan
WHICH NAME DO READERS STILL USE? FORMER NAMES OF RENAMED KAZAKHSTANI CITIES IN RUSSIAN AND KAZAKH WIKIPEDIA PAGE VIEWS, 2015–2026 ..... **543**

Nigmadilova Zhasmin
WHAT REPLACES PARABENS? THE PARABEN-FREE LABEL ACROSS PRODUCT CATEGORIES IN THE OPEN BEAUTY FACTS DATABASE **556**

ӘЛЕУМЕТТІК ЖӘНЕ ГУМАНИТАРЛЫҚ ҒЫЛЫМДАР - СОЦИАЛЬНО-ГУМАНИТАРНЫЕ НАУКИ - SOCIAL SCIENCES AND HUMANITIES

Мұқатаева Айзат Амантайқызы, Стамбекова Жанат Байғазықызы, Слямханова Гүлім Болатқызы, Қызырбек Аяулым Ерболқызы
ЖАСАНДЫ ИНТЕЛЛЕКТ ДӘУІРІНДЕГІ ҒЫЛЫМИ ТАНЫМ: ПОППЕР, КУН ЖӘНЕ ЛАКАТОС ИДЕЯЛАРЫ ТҰРҒЫСЫНАН ТАЛДАУ **561**

Мұқатаева Айзат Амантайқызы, Сапарханова Мадина Төлегенқызы, Окенова Аружан Муратхановна, Тұрсынғазыева Ақерке Ерболатқызы
ШӘКӘРІМ ФИЛОСОФИЯСЫНДАҒЫ ЕРКІНДІК ЖӘНЕ МОРАЛЬДЫҚ ТАҢДАУ МӘСЕЛЕСІ **571**

Komekova Saida
DO UNHAPPY REVIEWERS MIX IN MORE RUSSIAN? STAR RATINGS AND KAZAKH-RUSSIAN CODE-SWITCHING IN 175,002 ONLINE REVIEWS **580**

Zhumagali Yerbol
DOES A BIGGER MODEL CLOSE THE GAP? KAZAKH-LANGUAGE ACCURACY ON KAZAKHSTAN-SPECIFIC VERSUS UNIVERSAL SCHOOL SUBJECTS IN SMALL OPEN-WEIGHT LANGUAGE MODELS **593**

Abdrakhmanova Madina
DOES A MEDAL BUY ATTENTION? WIKIPEDIA VIEWS OF KAZAKHSTAN'S ATHLETES BEFORE AND AFTER THE PARIS 2024 OLYMPIC **597**

ПЕДАГОГИКАЛЫҚ ЖӘНЕ ПСИХОЛОГИЯЛЫҚ ҒЫЛЫМДАР – ПЕДАГОГИЧЕСКИЕ И ПСИХОЛОГИЧЕСКИЕ НАУКИ – PEDAGOGICAL AND PSYCHOLOGICAL SCIENCES

Kaldarbek Mergen
FAMILY OR SCHOOL? SOCIOECONOMIC COMPOSITION AND MATHEMATICS ACHIEVEMENT IN KAZAKHSTAN, PISA 2022 AND 2025 ..... **605**

Zhenis Adilzhan
WHICH OLYMPIAD PROBLEMS DO STUDENTS TAKE TO THE BOARD? CHOICES AND DEFENCES IN THE 8TH-GRADE LEAGUE OF KARAGANDY MATHBATTLES 2026 ..... **612**

Базарбаева Айнур Бектуровна
ДЕТСКИЙ ФОЛЬКЛОР И ПОДВИЖНЫЕ ИГРЫ КАК ИНСТРУМЕНТЫ ФОРМИРОВАНИЯ ОСНОВ ПРАВОВОЙ КУЛЬТУРЫ У ДЕТЕЙ СТАРШЕГО ДОШКОЛЬНОГО ВОЗРАСТА **619**

БИОЛОГИЯ ҒЫЛЫМДАР - БИОЛОГИЧЕСКИЕ НАУКИ - BIOLOGICAL SCIENCES

Bazhenova Diana
GROWTH DYNAMICS OF STERLET (Acipenser ruthenus L.) REARED IN A RECIRCULATING AQUACULTURE SYSTEM **622**

ТУРИСТИЧЕСКИЕ И СПОРТИВНЫЕ НАУКИ - ТУРИСТІК ЖӘНЕ СПОРТ ҒЫЛЫМДАРЫ - TOURISM AND SPORTS SCIENCES

Төлеп Жазира Маратқызы
ГАСТРОНОМИЯЛЫҚ ФЕСТИВАЛЬДЕРДІҢ ТУРИСТІК ДЕСТИНАЦИЯ ТАРТЫМДЫЛЫҒЫН ҚАЛЫПТАСТЫРУДАҒЫ РӨЛІ МЕН ДАМУ ҮРДІСТЕРІ **629**

ЕСТЕСТВЕННЫЕ НАУКИ - ЖАРАТЫЛЫСТАНУ ҒЫЛЫМДАРЫ - NATURAL SCIENCES

Нургельдина Асель Ермековна, Серикбаева Жанерке Алимжановна
ШЕКАРАСЫЗ ФИЗИКА: ӘР ЕЛДЕГІ ФИЗИКАЛЫҚ БІЛІМДІ САЛЫСТЫРУ **636**

Torekhan Sabrina
IS ALMATY A 15-MINUTE CITY? WALKING ACCESS TO SIX EVERYDAY PLACES FROM 170,838 BUILDINGS IN THE CITY'S EIGHT DISTRICTS **643**

Онгарбек Махабат Саятқызы
CU-PT ЖӘНЕ NI-PT ҚОРЫТПАЛАРЫНЫҢ ФИЗИКАЛЫҚ ҚАСИЕТТЕРІНЕ РЕТТЕЛУ ҮДЕРІСІНІҢ ЖӘНЕ ҚҰРЫЛЫМДЫҚ ӨЗГЕРІСІНІҢ ӘСЕРІН РЕТТЕУ **651**

Нургельдина Асель Ермековна, Әбдуәлі Мерей Ералықызы
ҮЙКЕЛІС КҮШІНІҢ ТАБИҒАТТАҒЫ РӨЛІ **659**

Ossipov Vladislav
MANY REPORTS, WEAK EVIDENCE? EVIDENCE LEVELS FOR MICROBIAL PLASTIC DEGRADATION IN THE 2,535 RECORDS OF PLASTICDB **664**

Нургельдина Асель Ермековна, Кабдулаева Жанель Батырбековна
РАДИОАКТИВТІЛІК. АЛЬФА, БЕТА, ГАММА СӘУЛЕЛЕРІНІҢ ТЕГІ МЕН ЗАҢДЫЛЫҚТАРЫ. ЯДРОЛЫҚ РЕАКЦИЯ ЖӘНЕ ЖАРТЫЛАЙ ЫДЫРАУ ПЕРИОДЫ **673**

Kashkinov Alden
DOES AN ACTIVE SUN BRING SATELLITES DOWN SOONER? SOLAR ACTIVITY AND THE RE-ENTRY OF UNCONTROLLED OBJECTS IN LOW EARTH ORBIT, 1990-2025 ..... **678**

Demin Mark
HOW MUCH LESS OXYGEN CAN THE WARMING CASPIAN SEA HOLD? THE LOSS OF OXYGEN SOLUBILITY IN THE NORTHERN, MIDDLE AND SOUTHERN CASPIAN, 1982-2022 ..... **684**