

What the Same Sentence Costs in Tokens: A Measured Comparison Across About 100 Languages

Cuelara Token Counter

Abstract

Language models are priced and limited by tokens, and a tokenizer does not split every language into the same number of tokens. We counted the same 20 sentences, translated into 108 languages across 28 writing systems, with OpenAI's two published encodings, o200k_base and cl100k_base. Relative to English, the median of the 107 other languages needs 1.80× as many tokens under o200k_base and 2.53× under cl100k_base. The spread is wide: Santali needs 15.29× the English count under o200k_base and Portuguese needs 1.29×. 27 languages need at least twice the English count under o200k_base. The counts are exact for these two encodings on this sample; the study makes no claim about the tokenizers of other model families.

Keywords: *tokenization; multilingual; large language models; token cost; context window; FLORES-200; o200k_base; cl100k_base*

1. Introduction

Language models read and write tokens, not characters or words. A tokenizer splits text into pieces from a fixed vocabulary learned from a training corpus, commonly with subword methods such as byte-pair encoding [1]. Providers bill per token and give each model a context window measured in tokens, so the number of tokens a text needs is the quantity that sets both its price and how much of it fits in a request.

A vocabulary learned mostly from some languages covers those languages in few, long pieces and covers others in many short ones. Earlier work has reported that this makes the same content cost different amounts in different languages: Ahia et al. [2] tested OpenAI's API on 22 languages and found that speakers of many languages pay more for the same content, and Petrov et al. [3] found large differences in token counts for identical translated text. A widely repeated rule of thumb, that a token is about three-quarters of an English word, describes English only.

This report adds a small, reproducible measurement of that effect for the two OpenAI encodings used by GPT-4o and GPT-4. It asks one question: for the same 20 sentences, how many tokens does each of 108 languages need, relative to English? It is written to inform practical budgeting of tokens and context for non-English text, and it is the data behind the language comparison in the Cuelara Token Counter.

2. Methodology

2.1 Data

We used the devtest split of FLORES-200 [4, 5], a collection of sentences professionally translated from English into more than 200 languages, so that line *n* says the same thing in every language. It is released under CC BY-SA 4.0. We took 20 sentences: the devtest lines numbered 10, 60, 110, 160, 210, 260, 310, 360, 410, 460, 510, 560, 610, 660, 710, 760, 810, 860, 910, 960, that is, every 50th line starting at line 10. The selection was fixed in advance and no sentence was chosen by hand.

2.2 Languages

We selected 108 of the languages FLORES-200 covers, aiming for a spread of writing systems (28 in total) and of resource levels, with English as the baseline. The selection is a hand-picked sample of the dataset, not the whole of it, and it is listed in full in Appendix A.

2.3 Tokenization

Each sentence was tokenized with gpt-tokenizer 4.0.0 using two encodings: o200k_base, used by GPT-4o and newer OpenAI models, and cl100k_base, used by GPT-4 and GPT-3.5 [6]. Both are published, so every count in this report is exact for its encoding. Special tokens are not added.

2.4 Measures

- Total tokens: the sum of tokens over a language's sentences, per encoding.
- Ratio to English: a language's total tokens divided by English's total tokens for the same sentence ids. English is therefore exactly 1.00 in both encodings.
- Tokens per character: total tokens divided by total characters, where characters are UTF-16 code units of the trimmed sentences.
- Tokens per word: total tokens divided by total words, where words are counted with the JavaScript engine's Intl.Segmenter in word mode. Splitting on spaces would count a whole Chinese, Japanese or Thai sentence as one word.
- Summary statistics (medians, counts above a ratio, and the group medians by script) are computed over languages, with English excluded where stated.

2.5 Reproducibility

The sentence selection and the counting are two scripts in the project repository (npm run tokens:data and npm run tokens:report). The text of this report is generated from the resulting data file, so its figures can be regenerated exactly.

3. Results

3.1 Summary

Table 1. Summary of the measurement.

Quantity	Value
Languages (including English)	108
Writing systems	28
Sentences per language	20
English tokens, o200k_base	457
English tokens, cl100k_base	463
English characters / words	2,414 / 386
English tokens per word, o200k_base	1.18
Median ratio to English, o200k_base	1.80×
Median ratio to English, cl100k_base	2.53×
Languages at 2× English or more, o200k_base / cl100k_base	27 / 87
Languages at 5× English or more, o200k_base / cl100k_base	7 / 29
Languages at 10× English or more, o200k_base / cl100k_base	1 / 9

The 20 English sentences total 457 tokens under o200k_base and 463 under cl100k_base. The median of the 107 other languages is 1.80× the English count under o200k_base and 2.53× under cl100k_base (Table 1).

3.2 Highest and lowest ratios

Santali		15.29×
Dzongkha		9.49×
Tibetan		8.98×
Lao		8.78×
Tigrinya		6.20×
Amharic		6.12×
Odia		5.51×
Khmer		3.79×
Burmese		3.43×
Central Kurdish		3.01×

Figure 1. The ten languages with the highest ratio to English under o200k_base. English = 1.00.

Figure 1 and Table 2 show the ten languages with the highest ratio; Table 3 shows the ten with the lowest, English excluded. The highest is Santali at 15.29× and the lowest is Portuguese at 1.29×.

Table 2. The ten languages with the highest ratio to English.

Language	Script	o200k tokens per sentence	Ratio (o200k)	Ratio (cl100k)
Santali	Ol Chiki	349.4	15.29×	14.17×
Dzongkha	Tibetan	216.9	9.49×	12.77×
Tibetan	Tibetan	205.3	8.98×	12.47×
Lao	Lao	200.7	8.78×	10.22×
Tigrinya	Ethiopic	141.7	6.20×	8.12×
Amharic	Ethiopic	139.9	6.12×	8.06×
Odia	Odia	125.8	5.51×	13.72×
Khmer	Khmer	86.7	3.79×	10.04×
Burmese	Myanmar	78.4	3.43×	12.57×
Central Kurdish	Arabic	68.8	3.01×	5.18×

Table 3. The ten languages with the lowest ratio to English (English excluded).

Language	Script	o200k tokens per sentence	Ratio (o200k)	Ratio (cl100k)
Portuguese	Latin	29.4	1.29×	1.57×
Chinese (Simplified)	Han (Simplified)	29.5	1.29×	1.97×
Indonesian	Latin	29.8	1.30×	1.65×
Dutch	Latin	30.4	1.33×	1.73×
Danish	Latin	31.4	1.38×	1.62×
Swedish	Latin	31.5	1.38×	1.64×
Malay	Latin	31.6	1.39×	1.75×
Sundanese	Latin	31.9	1.40×	1.85×
Javanese	Latin	31.9	1.40×	1.81×
German	Latin	31.9	1.40×	1.68×

3.3 By writing system

Table 4 groups the languages by writing system. Several groups contain only one or two languages, so a group's median describes this sample and not the script in general.

Table 4. Median ratio to English by writing system, o200k_base, highest first.

Script	Languages	Median	Lowest	Highest
Ol Chiki	1	15.29×	15.29×	15.29×
Tibetan	2	9.24×	8.98×	9.49×
Lao	1	8.78×	8.78×	8.78×
Ethiopic	2	6.16×	6.12×	6.20×
Odia	1	5.51×	5.51×	5.51×
Khmer	1	3.79×	3.79×	3.79×
Myanmar	1	3.43×	3.43×	3.43×
Gurmukhi	1	2.97×	2.97×	2.97×
Sinhala	1	2.86×	2.86×	2.86×
Greek	1	2.35×	2.35×	2.35×
Kannada	1	2.15×	2.15×	2.15×
Tamil	1	2.11×	2.11×	2.11×
Thai	1	2.11×	2.11×	2.11×
Malayalam	1	2.07×	2.07×	2.07×
Telugu	1	2.05×	2.05×	2.05×
Bengali	2	2.00×	1.88×	2.13×
Georgian	1	1.94×	1.94×	1.94×
Armenian	1	1.94×	1.94×	1.94×
Devanagari	6	1.90×	1.73×	2.25×
Cyrillic	10	1.90×	1.45×	2.17×
Gujarati	1	1.88×	1.88×	1.88×
Arabic	9	1.79×	1.43×	3.01×
Han + Kana	1	1.71×	1.71×	1.71×
Latin	54	1.69×	1.29×	2.57×
Hebrew	1	1.57×	1.57×	1.57×
Hangul	1	1.57×	1.57×	1.57×
Han (Traditional)	2	1.49×	1.49×	1.50×
Han (Simplified)	1	1.29×	1.29×	1.29×

3.4 Effect of the encoding

107 of the 108 languages need fewer tokens under o200k_base than under cl100k_base, with a median reduction of 26.6% across all languages. The median ratio to English falls from 2.53× to 1.80×. Table 5 lists the five languages with the largest reduction.

Table 5. The five languages with the largest reduction in tokens from cl100k_base to o200k_base.

Language	cl100k tokens	o200k tokens	Reduction
Armenian	5,047	886	82.4%
Georgian	4,862	888	81.7%
Kannada	4,615	981	78.7%
Malayalam	4,405	947	78.5%
Gujarati	3,903	859	78.0%

4. Discussion

Token cost varied widely across the sample. Under o200k_base the highest ratio is 11.9 times the lowest non-English ratio, and 27 of 107 non-English languages need at least twice the English count. For a user, this means the same prompt, document or conversation can use up a context window, or a budget, several times faster in some languages than in English.

Writing system is a strong correlate in this sample. The 54 Latin-script languages other than English have a median ratio of 1.69× under o200k_base, against 1.94× for the 53 languages in other scripts. The ten lowest ratios are in 2 scripts (Latin, Han (Simplified)). We do not test why; a vocabulary trained on text dominated by some scripts is the usual explanation in the literature [2, 3], but this study measures the outcome and not the cause.

The newer encoding narrows the gap on this sample, but not uniformly. The reduction is largest for Armenian, where tokens fall by 82.4% from 5,047 to 886 (Table 5), and it differs a great deal between languages. Estimates made with one encoding should not be carried over to another.

Practically, token and context budgets for non-English text should be measured on the text itself instead of scaled from an English estimate. The ratios here are a way to see roughly how far apart languages are, not a substitute for counting.

5. Limitations

- **Sample size.** 20 sentences per language support a comparison between languages on this text, not a promised ratio for any other text. Domain, formality, numbers, names and code all change the result.
- **Single translation.** Each sentence has one professional translation per language, and some languages are written in more than one standard or script.
- **Two encodings.** The tokenizers measured are OpenAI's published ones. Other providers, including Anthropic and Google, use tokenizers that are not available to run in a browser; their per-language costs may differ, and nothing here describes them.
- **Tokens are not prices.** Providers price input and output differently and change prices over time; a ratio of token counts is not a ratio of money.
- **Language selection.** 108 languages were chosen by hand for coverage of scripts and resource levels. Medians over this set describe this set, not all languages.
- **Language detection.** The Token Counter's detection step is separate from this measurement. It identifies 79 of the 108 languages and can confuse close relatives, so a figure shown next to pasted text is a lookup into this study's data, not a measurement of that text.

6. Conclusion

On 20 parallel sentences in 108 languages, the number of tokens needed for the same meaning differs strongly by language. The median non-English language needs 1.80× the English count under o200k_base and 2.53× under cl100k_base, and the extremes are far apart. The counts are exact for the two encodings measured and reproducible from the scripts and data described above.

Data and code availability

The sentence selection and counting scripts (tokens:data, tokens:report) are in the Cuelara project repository, and the measured results are published with this report. The underlying sentences are from FLORES-200 [4, 5] and are redistributed under CC BY-SA 4.0 (<https://creativecommons.org/licenses/by-sa/4.0/>); material adapted from them remains under that licence. We selected 20 sentences per language and trimmed whitespace; nothing else was altered.

References

- [1] Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). arXiv:1508.07909.
- [2] Ahia, O., Kumar, S., Gonen, H., Kasai, J., Mortensen, D. R., Smith, N. A., and Tsvetkov, Y. (2023). Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models. arXiv:2305.13707.
- [3] Petrov, A., La Malfa, E., Torr, P. H. S., and Bibi, A. (2023). Language Model Tokenizers Introduce Unfairness Between Languages. Advances in Neural Information Processing Systems 36 (NeurIPS 2023). arXiv:2305.15425.
- [4] NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., et al. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672.
- [5] Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. (2021). The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. arXiv:2106.03193.
- [6] OpenAI tokenizers o200k_base and cl100k_base, as implemented by gpt-tokenizer 4.0.0. <https://github.com/nlpeani/gpt-tokenizer>

Appendix A. All languages

Table A1. Every language in the sample, highest ratio first.

Language	Script	o200k tokens per sentence	Ratio (o200k)	Ratio (cl100k)
Santali	Ol Chiki	349.4	15.29×	14.17×
Dzongkha	Tibetan	216.9	9.49×	12.77×
Tibetan	Tibetan	205.3	8.98×	12.47×
Lao	Lao	200.7	8.78×	10.22×
Tigrinya	Ethiopic	141.7	6.20×	8.12×
Amharic	Ethiopic	139.9	6.12×	8.06×
Odia	Odia	125.8	5.51×	13.72×
Khmer	Khmer	86.7	3.79×	10.04×
Burmese	Myanmar	78.4	3.43×	12.57×
Central Kurdish	Arabic	68.8	3.01×	5.18×
Punjabi	Gurmukhi	67.8	2.97×	9.00×
Sinhala	Sinhala	65.3	2.86×	9.34×
Uyghur	Arabic	62.0	2.71×	5.57×
Yoruba	Latin	58.7	2.57×	3.55×
Greek	Greek	53.7	2.35×	5.48×
Sanskrit	Devanagari	51.5	2.25×	5.34×
Belarusian	Cyrillic	49.5	2.17×	3.83×
Kannada	Kannada	49.0	2.15×	9.97×
Assamese	Bengali	48.7	2.13×	6.63×
Tamil	Tamil	48.2	2.11×	8.23×
Thai	Thai	48.1	2.11×	4.83×
Northern Kurdish	Latin	47.5	2.08×	2.30×
Maltese	Latin	47.4	2.07×	2.65×
Samoan	Latin	47.4	2.07×	2.51×
Malayalam	Malayalam	47.4	2.07×	9.51×
Telugu	Telugu	46.9	2.05×	9.10×
Tajik	Cyrillic	46.0	2.01×	3.92×
Maithili	Devanagari	45.5	1.99×	5.42×
Macedonian	Cyrillic	45.5	1.99×	3.16×

Language	Script	o200k tokens per sentence	Ratio (o200k)	Ratio (cl100k)
Maori	Latin	45.0	1.97×	2.59×
Serbian	Cyrillic	44.9	1.96×	3.19×
Georgian	Georgian	44.4	1.94×	10.50×
Irish	Latin	44.4	1.94×	2.40×
Armenian	Armenian	44.3	1.94×	10.90×
Albanian	Latin	44.3	1.94×	2.39×
Mongolian	Cyrillic	44.1	1.93×	3.86×
Marathi	Devanagari	43.7	1.91×	5.28×
Pashto	Arabic	43.4	1.90×	4.19×
Bhojpuri	Devanagari	43.3	1.90×	5.06×
Sindhi	Arabic	43.0	1.88×	4.49×
Gujarati	Gujarati	43.0	1.88×	8.43×
Bengali	Bengali	42.9	1.88×	6.51×
Hungarian	Latin	42.8	1.87×	2.25×
Latvian	Latin	42.8	1.87×	2.42×
Ukrainian	Cyrillic	42.8	1.87×	3.18×
Uzbek	Latin	42.1	1.84×	2.26×
Kinyarwanda	Latin	42.0	1.84×	2.29×
Welsh	Latin	41.9	1.83×	2.20×
Somali	Latin	41.8	1.83×	2.31×
Slovak	Latin	41.5	1.81×	2.34×
Cebuano	Latin	41.4	1.81×	2.11×
Bulgarian	Cyrillic	41.3	1.81×	2.81×
Romanian	Latin	41.1	1.80×	2.08×
Luganda	Latin	41.0	1.80×	2.13×
Urdu	Arabic	41.0	1.79×	4.72×
Icelandic	Latin	40.9	1.79×	2.42×
Azerbaijani	Latin	40.7	1.78×	2.75×
Tagalog	Latin	40.5	1.77×	2.18×
Zulu	Latin	40.3	1.76×	2.32×
Nepali	Devanagari	40.0	1.75×	5.17×
Lithuanian	Latin	40.0	1.75×	2.25×
Wolof	Latin	39.9	1.75×	1.97×
Shona	Latin	39.8	1.74×	2.19×
Basque	Latin	39.5	1.73×	2.05×
Hindi	Devanagari	39.5	1.73×	5.16×
Kyrgyz	Cyrillic	39.5	1.73×	3.66×
Polish	Latin	39.1	1.71×	2.01×
Japanese	Han + Kana	39.0	1.71×	2.35×
Czech	Latin	38.8	1.70×	2.26×
Lingala	Latin	38.7	1.69×	1.96×
Igbo	Latin	38.5	1.69×	2.52×
Chichewa	Latin	38.2	1.67×	2.15×
Estonian	Latin	38.1	1.67×	2.06×
Xhosa	Latin	38.1	1.67×	2.11×
Persian	Arabic	37.7	1.65×	3.62×
Finnish	Latin	37.5	1.64×	2.08×
Kazakh	Cyrillic	37.4	1.64×	4.01×

Language	Script	o200k tokens per sentence	Ratio (o200k)	Ratio (cl100k)
Catalan	Latin	36.8	1.61×	1.89×
Croatian	Latin	36.5	1.60×	1.97×
Slovenian	Latin	36.3	1.59×	1.97×
Italian	Latin	35.9	1.57×	1.77×
Swahili	Latin	35.9	1.57×	2.08×
Hebrew	Hebrew	35.9	1.57×	3.98×
Korean	Hangul	35.8	1.57×	2.55×
Dari	Arabic	35.8	1.56×	3.45×
Hausa	Latin	35.2	1.54×	2.03×
Vietnamese	Latin	34.6	1.52×	2.53×
Afrikaans	Latin	34.4	1.51×	1.84×
Chinese (Traditional)	Han (Traditional)	34.3	1.50×	2.23×
Turkish	Latin	34.0	1.49×	2.00×
Cantonese	Han (Traditional)	34.0	1.49×	2.21×
Egyptian Arabic	Arabic	33.7	1.47×	3.26×
Norwegian (Bokmål)	Latin	33.4	1.46×	1.77×
Spanish	Latin	33.3	1.46×	1.70×
French	Latin	33.3	1.46×	1.69×
Russian	Cyrillic	33.0	1.45×	2.55×
Arabic	Arabic	32.8	1.43×	3.24×
German	Latin	31.9	1.40×	1.68×
Javanese	Latin	31.9	1.40×	1.81×
Sundanese	Latin	31.9	1.40×	1.85×
Malay	Latin	31.6	1.39×	1.75×
Swedish	Latin	31.5	1.38×	1.64×
Danish	Latin	31.4	1.38×	1.62×
Dutch	Latin	30.4	1.33×	1.73×
Indonesian	Latin	29.8	1.30×	1.65×
Chinese (Simplified)	Han (Simplified)	29.5	1.29×	1.97×
Portuguese	Latin	29.4	1.29×	1.57×
English	Latin	22.9	1.00×	1.00×