The Token Tax: Why AI Speaks English Fluently and Bahasa Expensively

Hafiz Hanif

Faculty of Human Development, Sultan Idris Education University · ORCID 0000-0001-9943-6557

CITE AS	Hanif, H. (2026, September 4). The Token Tax: Why AI Speaks English Fluently and Bahasa Expensively. Hack(edu).tech Chronicle. https://hackedu.tech/articles/the-token-tax-why-ai-speaks-english-fluently-and-bahasa-expensively
PUBLISHED	4 September 2026
DOI	10.5281/zenodo.22303127
LICENCE	CC-BY-4.0
AVAILABLE AT	https://hackedu.tech/articles/the-token-tax-why-ai-speaks-english-fluently-and-bahasa-expensively

ABSTRACT

Before AI reads a single word, text is chopped into tokens by machinery calibrated on English, and other languages pay for it. This article traces the token tax on Bahasa Melayu through three costs: degraded performance on affixes and syllables, the foundations of early literacy; higher prices and shorter memory for the same content; and cultural defaults imported with the language. Left unaudited, the tax becomes an unlegislated language policy. Regional models and vigilant procurement offer the correction.

KEYWORDS

Artificial Intelligence, AI Token, Large Language Model

Here is an experiment you can run tonight. Ask an AI chatbot a question in English. Then ask the exact same question in Bahasa Melayu. You will get two answers, probably both decent, and you will notice nothing unusual at all.

But under the hood, something unequal just happened, and it happens every single time. To the machine, your Malay question was more expensive to read, more expensive to answer, and heavier to carry in its memory. Not because Malay is harder. Because the machine's meter was calibrated somewhere else.

Nobody decided this. No committee voted on it. It fell out of the machinery, which is exactly why it deserves a column: the biases nobody chose are the ones nobody audits.

Let me explain the mechanism once, gently, and then we can leave the engine room and never return. Before an AI reads anything, your sentence is chopped into pieces called tokens, and the machine pays its costs, in money, speed, and memory, per piece. Crucially, the chopping is not done by a linguist. It is done by frequency statistics learned mostly from English text, because English dominates the Internet the machine learned from. The consequence: English gets carved into big, meaningful, efficient chunks, while many other languages get shredded into small, arbitrary fragments. Researchers at Oxford who measured this found that the very same sentence, translated across languages, can cost up to fifteen times more tokens in some

languages than in English, an inequality baked in before the AI even begins to think, affecting price, speed, and how much the system can hold in its head at once (Petrov et al. 2023). Malay sits nowhere near the worst extreme, but it reliably pays more than English. Think of it as a tax on the language, levied by a meter that was never calibrated for us.

That is the entire technical content of this article. Everything that follows is about what the tax does.

The first cost is performance, and it lands exactly where our youngest learners live. Because the machine's chopping ignores real word structure, it routinely slices Malay words at points no teacher would. Our language builds meaning through affixes: 'sembah' becomes 'mempersembahkan' by wearing 'mem', 'per', and 'kan'. A cikgu teaching imbuhan needs those parts to be exactly right. But the machine may carve 'mempersembahkan' into fragments where the root 'sembah' never appears intact, which is why AI-generated materials for imbuhan exercises, suku kata drills, pantun syllable counts, and kata ganda like kanak-kanak deserve a suspicious eye every single time. The machine is at its weakest precisely at the level of language that Tahun Satu is built on. For essays and explanations, meaning-level work, it performs admirably in Malay. For letter-level and syllable-level work, the foundation stones of early literacy, it is a confident amateur. Teachers of young children, in other words, need the most scepticism, not the least.

The second cost is economic, and it compounds quietly. Because AI services meter by the token, serving the same lesson, the same feedback, the same chatbot conversation costs more in Malay than in English. The same subscription buys less. The same free-tier allowance runs out faster. The machine's working memory, also counted in tokens, fills up sooner, so long Malay documents get truncated earlier than long English ones. For an individual user this is pocket change. For a ministry serving five million students, or an EdTech startup pricing a product for national schools, the token tax scales into real money, and it tilts every business case toward serving users in English.

Which produces the third and most serious cost, the cultural one. The tax is not only about how text is chopped; it is about what the machine has read. These systems learned from an Internet overwhelmingly written by and about a narrow slice of humanity, and researchers warned early that the resulting models would encode the values and viewpoints of those overrepresented populations while rendering others nearly invisible (Bender et al. 2021). The machine knows Shakespeare in oceans and Usman Awang in drops. Ask it for a story to teach honesty and you will meet American children in American suburbs unless you insist otherwise. None of this is malicious. It is statistical. But statistics, left unexamined, become defaults, and defaults, in classrooms, become curriculum.

Now connect the three costs and watch the quiet consequence emerge. If the AI works better in English, costs less in English, and feels more natural in English, then every teacher, student, and app developer faces a gentle, permanent nudge toward using it in English. Multiply that nudge by millions of daily interactions and you have a de facto language-of-instruction policy that no one in Putrajaya ever drafted, debated, or gazetted. And it lands unevenly: reviews of language-model use in education keep flagging equity as an unresolved challenge (Yan et al. 2023), and here is its linguistic face: the student in an English-speaking home gets the machine at its best, while the B40 child in a Malay-medium national school gets it at its weakest and priciest. The digital divide we spent two decades narrowing acquires a new linguistic layer, invisible in any infrastructure audit because every school has the same devices and the same apps. The inequality is inside the language.

So what do we do? Not despair, because this is one of the rare technology problems where our region is actually fighting back, and the fight is worth knowing about. Singapore's national AI programme has built SEA-LION, a family of open models deliberately trained on Southeast Asian languages, Malay included, precisely because mainstream models underrepresent our

region's languages and cultures (Ng et al. 2025). At home, Malaysian developers released MaLLaM, Malay language models trained on tens of billions of tokens of Malaysian text, and in 2025 the country launched ILMU, a domestically developed model tuned for Bahasa Malaysia. I hold no brief for any of these systems, and their classroom worth must be proven, not presumed. But their existence changes the conversation: the token tax is not a law of nature. It is a design choice, and design choices can be remade by people who care about our language.

While the builders build, educators and institutions have three moves available now:

- First, test in Malay before adopting: any tool being considered for a Malaysian school should be trialled in the language it will actually be used in, not the language of the vendor's demo.
- Second, keep human eyes on every Malay language-learning artefact, especially for early literacy, where the machine is structurally weakest; this is not busywork, it is the professional judgement that the foundational research on AI literacy says must precede use, not follow it (Ng et al. 2021).
- Third, ask vendors the impolite questions: how does your system handle Bahasa Melayu, what does the same task cost in Malay versus English, and what Malaysian content did your model actually learn from? If the salesperson cannot answer, that is an answer.

Bahasa jiwa bangsa, we say. Language is the soul of the nation. That saying is about to face its strangest test, because for the first time the dominant machinery of knowledge charges our soul a surcharge. The response is not to abandon the language for the cheaper one; souls are not priced per token. The response is to see the tax clearly, name it in our procurement meetings, support the builders correcting it, and keep our teachers' judgement wrapped around every output until the machinery finally learns to read us properly.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2, 100041. https://doi.org/10.1016/j.caeai.2021.100041

Ng, R., Nguyen, T. N., & Huang, Y. (2025). SEA-LION: Southeast Asian languages in one network. https://arxiv.org/abs/2504.05747

Petrov, A., La Malfa, E., Torr, P. H. S., & Bibi, A. (2023). Language model tokenizers introduce unfairness between languages. In Advances in Neural Information Processing Systems (Vol. 36, pp. 36963–36990). NeurIPS 2023.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. British Journal of Educational Technology, 55(1), 90–112. https://doi.org/10.1111/bjet.13370