

AUTHOR: ROSEN IVANOV

What Makes an Asian Language AI-Friendly?

Script complexity, tokenization costs, and data scarcity mean two Asian languages with identical word counts can carry very different budgets, timelines, and quality risks. Here's what localization managers and AI leads need to know before planning an Asian rollout.

1-STOPASIA®

GENERATED AI

The Token Tax Nobody Budgets For

A 2023 NeurIPS study by Oxford researchers found that the same text can require **up to 15× more tokens** in one language than another with Shan (Myanmar) as the extreme case. Tokens are the units AI systems use to charge, process, and limit context windows.

💰 Cost Impact

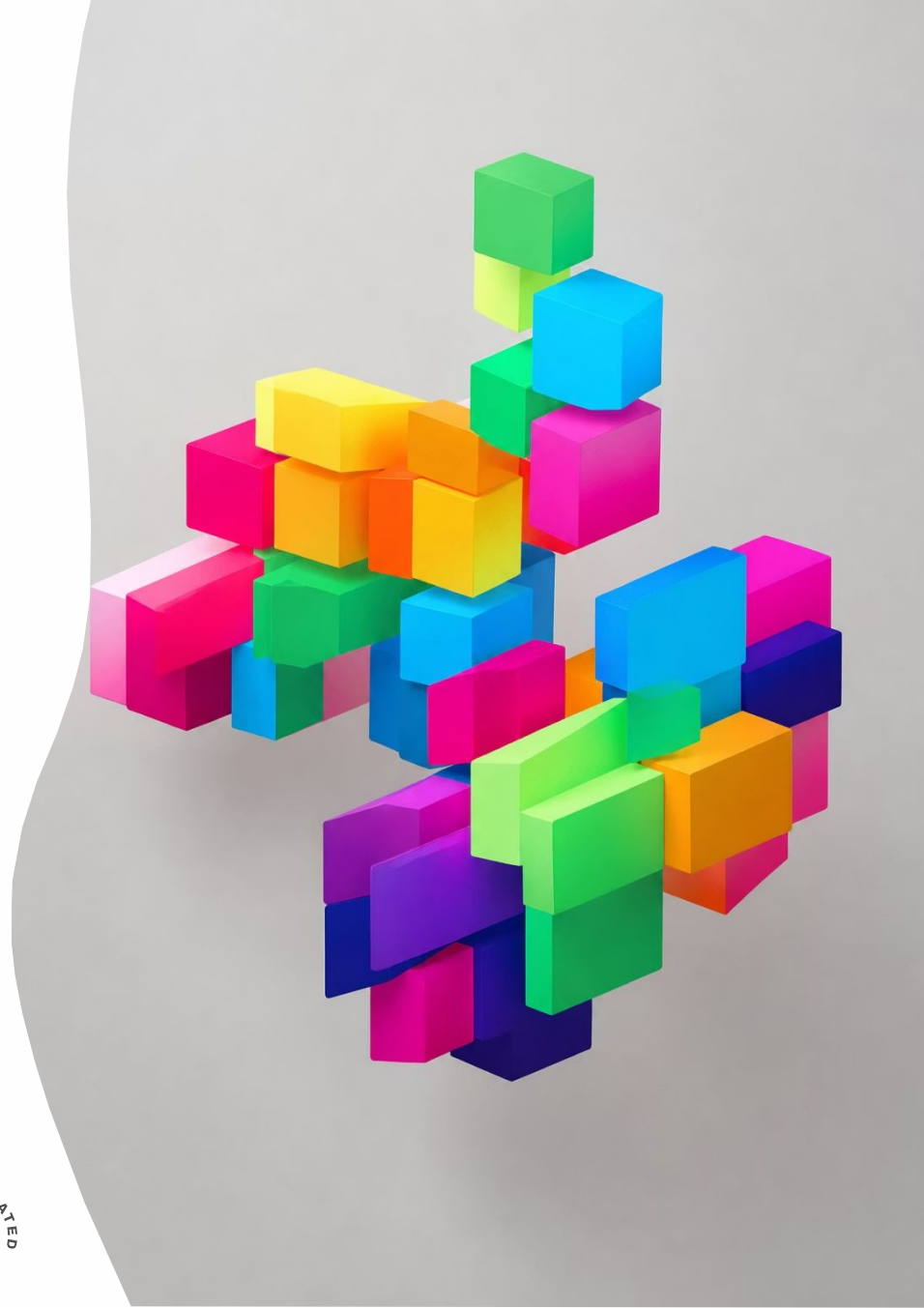
Providers charge per token-more tokens mean higher invoices, directly.

⏱️ Speed Impact

Processing time follows token count - timelines stretch for token-heavy languages.

📐 Context Impact

Less content fits in a context window, limiting what the model can "see" at once.



1-STOPASIA®

GENERATED AI

Where Scripts, Spaces & Grammar Create Friction

AI-friendliness varies by language feature. Each Asian language presents its own distinct challenge.

Japanese: Three Scripts, One Sentence

Kanji, hiragana, and katakana coexist in a single sentence. ChatGPT's tokenizer needed **3 tokens for 65%+ of kanji characters**. Register shifts mean the same word can appear in multiple scripts.

Korean: Simple Letters, Demanding Grammar

Hangul syllable blocks are more regular than Chinese characters, but Korean's honorific system changes verb endings throughout, making **honorific consistency a task that still needs human review.**

Chinese, Thai & Burmese: No Spaces, Tones Matter

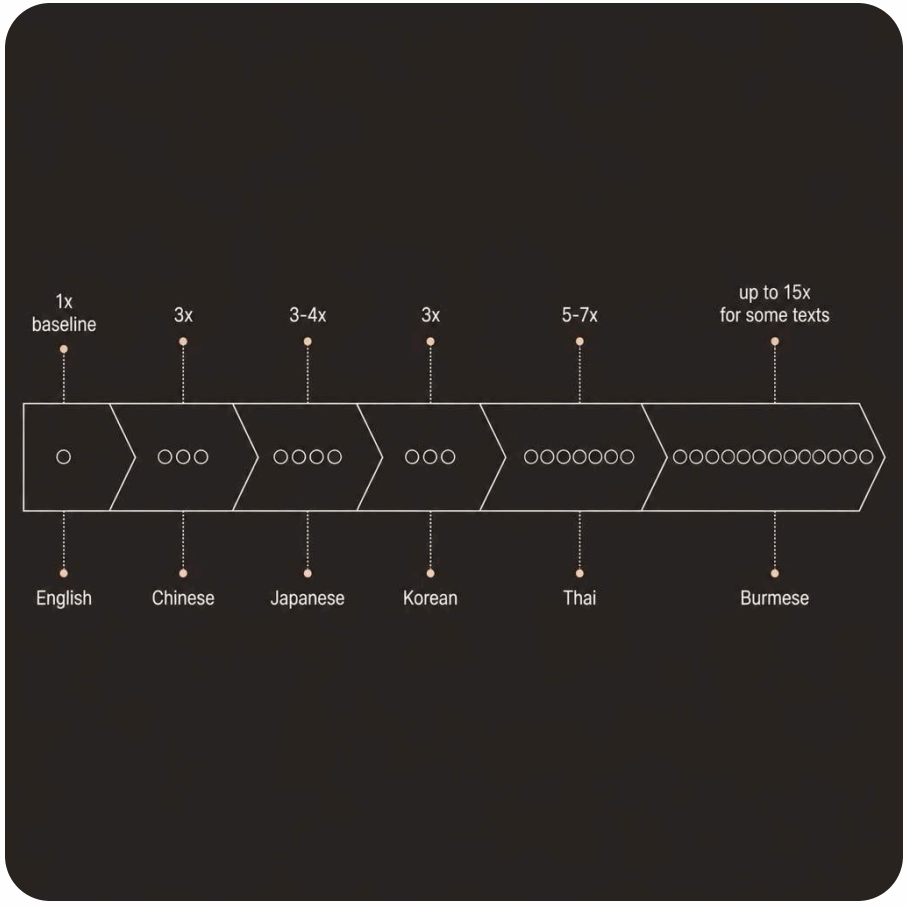
Without word boundaries, AI must decide where words begin before translating. A wrong cut changes meaning entirely. Tone markers in Vietnamese, Thai, and Mandarin add another layer of precision required.

The Multilingual Model Still Pays the Price

The Oxford team tested **17 tokenizers**, including those built specifically for multilingual support, and disparities persisted across all of them. No tokenizer eliminated the gap.

Byte-level encoding of Burmese or Tibetan text runs **over 4× longer** than the equivalent Chinese text. For project planning, this means two Asian languages with identical word counts can carry very different compute budgets and delivery timelines.

⚠ Asian languages must be priced and scheduled **language by language**, not as a single "Asian" line item.



GENERATED AI

Low-Resource Languages: When Training Data Barely Exists

Thin data means tokenization is only the first problem. A **2024 study on Marathi** found that advanced closed models like Gemini and GPT underperformed BERT-based baselines, even though similar models excel in English. Bigger models don't consistently fix this.

The Burmese Encoding Problem

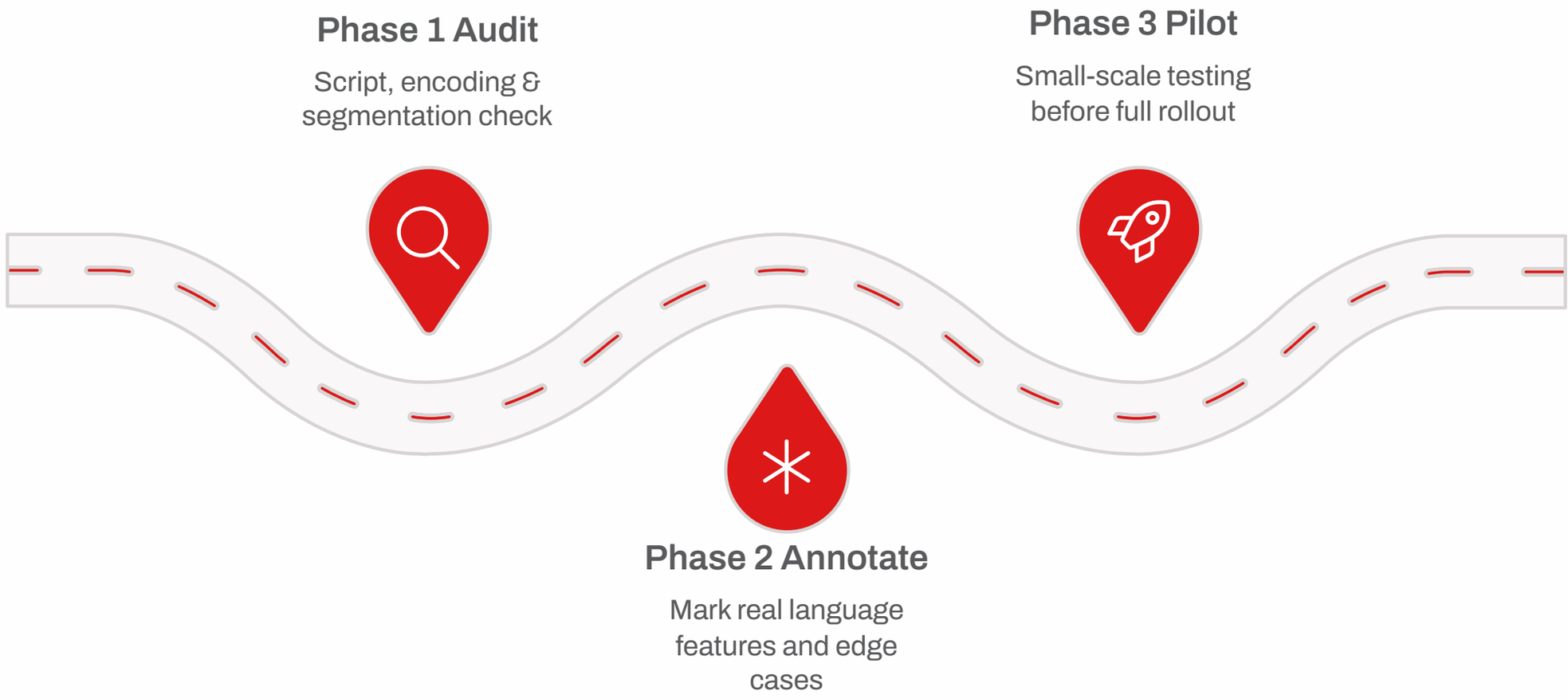
Much of Myanmar's historical online text, including most social media, was written in **Zawgyi**, a font encoding that stores characters in visual order rather than Unicode's logical order. Teams training on scraped Burmese data without checking encoding are mixing two incompatible systems simultaneously.

Waiting for a Bigger Model Is Not a Strategy

For low-resource languages, scaling model size does not reliably improve results. If your project has a deadline, the method must adapt to the language, not the other way around.

How 1-StopAsia Assesses Each Language

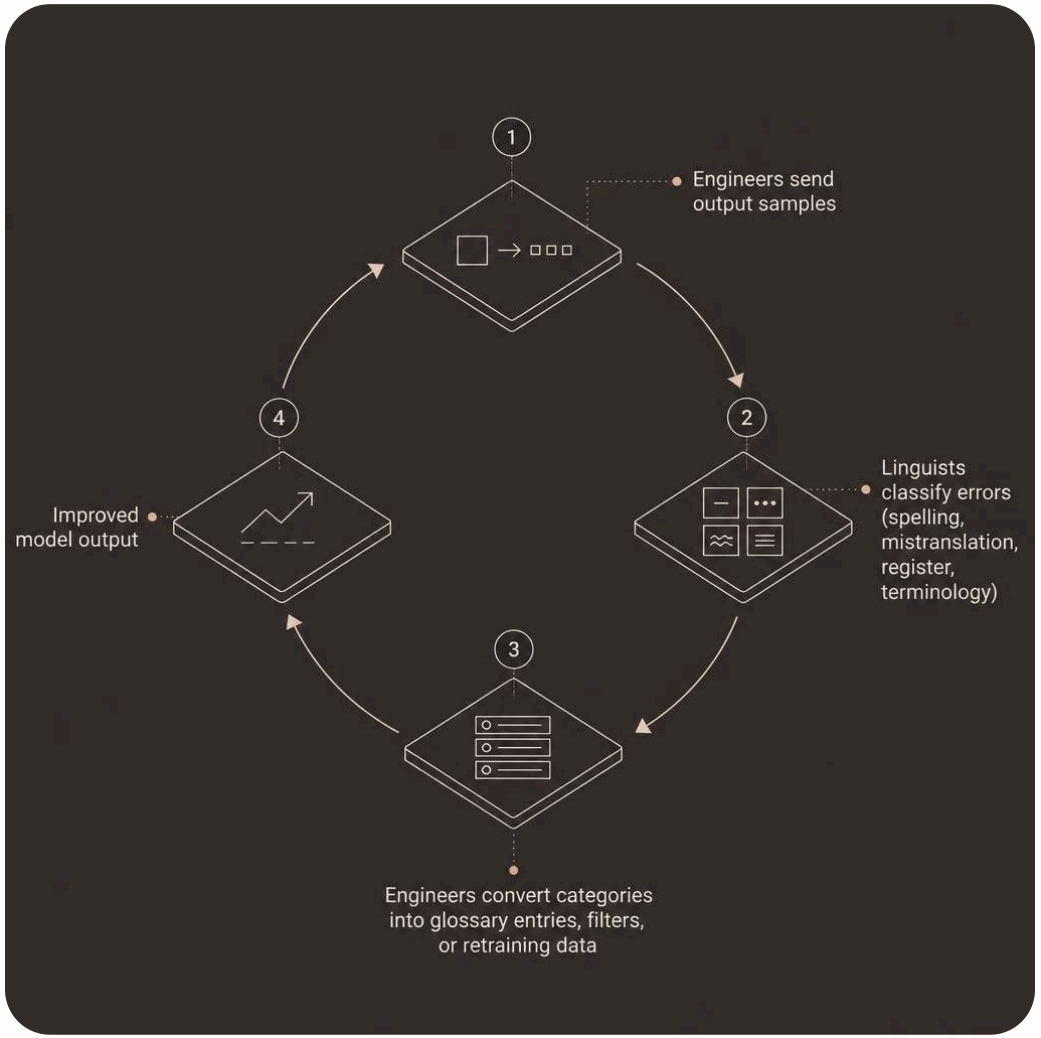
With over 25 years in Asian-language localization and 15+ core languages managed in-house, 1-StopAsia follows a fixed assessment sequence before any model work begins.



This sequence catches Zawgyi text, mixed-script issues, and inconsistent spacing while fixes are still cheap, long before full volume is committed.

GENERATED AI

The Linguist-Engineer Loop That Makes It Work



Fixed Hand-offs, Honest Accountability

Two habits keep this loop from drifting: an **agreed error taxonomy** before the pilot starts, and a **fixed review sample for every release**, so a retrained model is always judged against the same sentences as the previous version.

Tiered Post-Editing by Risk

- **Light editing** for internal, high-volume content
- **Deep editing** for anything customers read
- **Full human review** where legal, medical, or reputational risk applies

What Machine Translation Quality Actually Means

> In 1960, philosopher Yehoshua Bar-Hillel argued that fully automatic high-quality translation was not feasible, and proposed a human-machine partnership instead.

Neural systems have moved the line significantly, especially for high-resource languages like English, French, and Chinese. But the evidence shows the line sits in **different places for different languages** and no single score captures that.

Measure Quality Per Language, Per Content Type

- Automatic scores should be supported by native review wherever mistakes carry **legal, medical, or reputational weight**
- For high-volume internal content in well-resourced languages, **light post-editing is often sufficient**
- Full human rewrites where data is thin or stakes are high

The Method Must Follow the Language

No single AI approach covers all Asian languages. The features of each language decide the method and the method decides the cost.



Assess First

Audit scripts, encoding, and segmentation before any model work begins. Catching Zawgyi or mixed-script issues early is cheap; finding them in production is not.



Annotate What the Language Does

Label the real features, tone markers, honorifics, spacing rules, domain phrasing, that automated systems routinely miss.



Keep Local Linguists in the Loop

Teams that share one feedback loop between engineers and native linguists get predictable results. Teams that treat Asian languages as one block encounter problems in production.

Ready to plan your AI localization project? **Contact 1-StopAsia** to discuss your languages and request a consultation.