Universal Byte-Level Encoding: UTF-8/UTF-16 Routing to Reduce Cross-Script Token-Budget Disparities

Lower cross-script disparity for BPE tokenizers, and up to one-third shorter inputs for byte-level models.

Hyunsik Kim · **Youngmoon Jung**
Samsung Research

Project page

1 UTF-8 charges some scripts more

Byte-level BPE (BBPE) [2] starts from UTF-8 bytes. Where no learned merge covers a character, it costs one base symbol per byte: the **encoding floor**.

	ASCII	Cyrillic, Arabic, Hebrew	Korean, Chinese, Japanese, Devanagari, Thai	Non-BMP
BBPE	1	2	3	4
BBPE16	2	2	2	4
UBE	**1**	2	**2**	4

Base symbols per character before BPE merges. BMP: Basic Multilingual Plane.

BBPE tokens per FLORES-200 sentence, 101-lang/16K

308.5
LC-BMP (low-coverage BMP): Tibetan, Myanmar, Ol Chiki, Tifinagh scripts

34.9
English

A fixed token budget fits about **one-ninth** as many parallel sentences as English.

2 UTF-8 is less compact for 3-byte BMP

`1` format bit · `0` leading zero · `x` code-point bit · UBE route

1-byte U+0000–007F
UTF-8 `0xxxxxxx` — **1 byte**
UTF-16 `00000000 0xxxxxxx` — 2 bytes

2-byte U+0080–07FF
UTF-8 `110xxxxx 10xxxxxx` — **2 bytes**
UTF-16 `00000xxx xxxxxxxx` — 2 bytes

3-byte U+0800–FFFF
UTF-8 `1110xxxx 10xxxxxx 10xxxxxx` — 3 bytes
UTF-16 `xxxxxxxx xxxxxxxx` — **2 bytes**

4-byte U+10000–10FFFF
UTF-8 `11110xxx 10xxxxxx 10xxxxxx 10xxxxxx` — 4 bytes
UTF-16 `110110xx xxxxxxxx 110111xx xxxxxxxx` — **4 bytes**

A 3-byte UTF-8 character spends **8 of its 24 bits** on length and continuation markers, the price of self-synchronization; UTF-16 stores these **16 code-point bits** in 2 bytes [1]. UTF-16 adds a byte for ASCII and saves one only for 3-byte BMP characters.

UBE resynchronizes at each alphabet switch: the decoder pairs UTF-16 bytes from the start of each run.
UTF-16 code units are drawn most significant bit first; UBE writes them little-endian.

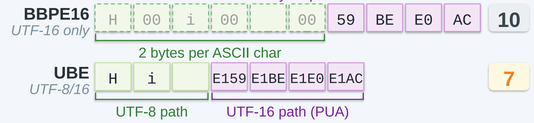
3 UBE: route each character

One deterministic rule on each character's UTF-8 length:

1–2 bytes	UTF-8 bytes	GPT-2 byte symbols
3–4 bytes	UTF-16LE bytes	PUA U+E100–E1FF

Two disjoint 256-symbol alphabets tell the decoder which path each symbol took: **no marker tokens, no language ID, no script table.**

Example: Hi 빙고 before BPE merges



Each box is one base symbol, the input unit of a byte-level model.
PUA: Private Use Area.

Standard BPE, same model
Only the initial symbols change: same BPE trainer, merge rule, and Transformer.

Exact decoding
Round-trips every Unicode 17 scalar and all tested strings exactly.

Small overhead: 256 extra base symbols (<1% for V ≥ 32K).

4 LM quality and speed

Matches BBPE's LM quality (BPB and downstream tasks).
182M–1.3B parameters; 32K–128K vocabularies. BPB: bits per byte.

Matched UTF-8 byte exposure; same test text · 101-lang/32K · 182M · 1 seed
BPB ↓: UBE **1.5298** · BBPE 1.5439 · BBPE16 1.5418

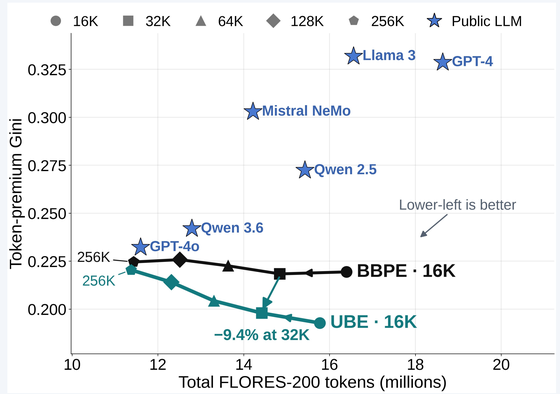
+45.0%
Amharic prompt-processing throughput

53%
more Santali content

FLORES-200 speed (same text): 6-lang/32K, 182M, A100 80 GB, batch 1, FP32.
Context: 101-lang/32K, FLORES-200, fixed token budget (token-count estimate).

5 Token level: lower disparity

FLORES-200 devtest, 204 languages.



Gini of token premiums (a language's token count relative to English).
UBE/BBPE: 101-language corpus, 16K–256K. Stars: public tokenizers with different training data.

−9.4%
Gini at 101-lang/32K: 0.2183 → 0.1979

189/204
languages with fewer tokens (101-lang/32K)

Mean tokens per sentence, 101-lang/32K

Language (script)	UBE	BBPE	Δ
Santali (Ol Chiki)	222.944	341.287	−34.7%
Central Atlas Tamazight (Tifinagh)	183.305	269.166	−31.9%
Tibetan	286.737	336.804	−14.9%
Arabic	60.407	60.600	−0.3%
English	31.068	31.160	−0.3%

UBE lowers English tokens at every tested vocabulary size (6-lang, 101-lang).
Why: 3-byte BMP characters use 2 symbols, freeing merges for Latin-script text.
Gini change: UDHR **−11.1%**, SIB-200 **−8.5%** (101-lang/32K).

Composability: BBPE → UBE (101-lang/32K)

Boundary policy	Tokens Δ	Gini
SuperBPE-style [3]	−2.8%	0.1922 → 0.1725
SCRIPT-style [4]	−2.8%	0.2036 → 0.1824

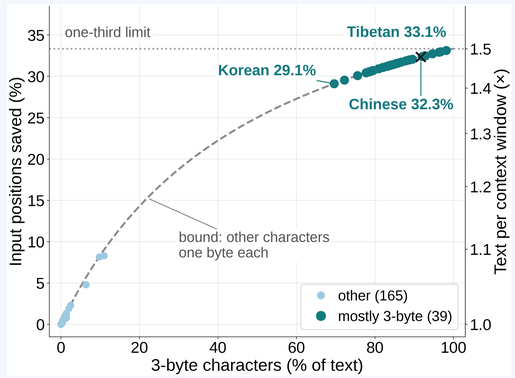
Emulated commercial mixtures · 10 UBE/BBPE pairs, 32K–200K.
30 GiB mC4: inferred GPT-4o/Mistral NeMo shares [5] + a Chinese-centric mix.

1.9–4.6% fewer tokens · **12.5–25.3%** fewer LC-BMP tokens · **5.5–11.7%** lower Gini

English tokens also fall in every pair.

6 Byte-level models gain more

Byte-level models such as ByT5 [6] read one input position per base symbol, with no merges.



UBE vs. UTF-8 bytes, one point per FLORES-200 language; mostly 3-byte: at least half of the characters; Chinese: mean of Chinese (Simplified), Chinese (Traditional), and Cantonese.

29–33%
shorter in all 39 languages with mostly 3-byte text

−28%
Gini at the byte level: 0.2077 → 0.1500

202/204
languages shorter, 2 tied, none longer

Mean base symbols per sentence, FLORES-200 (no merges)

Languages	UBE	UTF-8 bytes	Δ
Korean, Chinese, Japanese	91.252	133.665	−31.7%
LC-BMP	277.056	406.632	−31.9%
All 204	176.088	196.106	−10.2%
English	130.481	130.530	−0.04%

- English is also shorter: ASCII keeps 1 symbol, and 3-byte characters such as em dashes and curly quotes take **2 symbols instead of 3**.
- **1.41–1.50×** text per context window in these 39 languages.

UTF-8 where it is compact, UTF-16 where it is not.

[1] Kim et al., BBPE16, ICASSP 2026. [2] Wang et al., Neural machine translation with byte-level subwords, AAAI 2020. [3] Liu et al., SuperBPE, COLM 2025. [4] Land and Arnett, SCRIPT-BPE, Tokenization Workshop 2025. [5] Hayase et al., Data mixture inference, NeurIPS 2024. [6] Xue et al., ByT5, TACL 2022.