Benchmarking Large Language Models for Chinese and Japanese IMEs

Phonetic-to-Character Generation and Textual Error Correction

Yuchun Zou[1] • Tedd Lee[2] • Xiaodi Fan[3] • Jun Li[1,4]

[1,4]CUNY Graduate Center / Queens College [2]CUNY Hunter College [3]Meta Inc., New York

LREC-COLING 2026

1. INTRODUCTION & MOTIVATION

Typing in **Chinese** and **Japanese** requires an **Input Method Editor (IME)**: users type phonetic forms (Pinyin for Chinese; Romaji for Japanese), and the IME converts them into logographic characters.

Core difficulty —Homophones:

- *"yì yi"* → 意义 (significance) or 异议 (objection)?
- *"tā men"* → 他们 (he) / 她们 (she) / 它们 (it)?
- *"ki"* → 木 (tree) / 気 (spirit) / 機 (machine)?

Traditional **n-gram IMEs** are fast but limited to a short local window and cannot handle phonetic typos.

LLMs offer:

- Full-sentence contextual reasoning
- Correction of phonetic & semantic errors
- Multilingual coverage across Chinese & Japanese



Real IME candidate list: sentence context should push 上海 to rank 1.

2. EXPERIMENTAL SETUP

Task	Subtask	Lang.
Phonetic → Char.	Word-level; Sentence-level	ZH + JA
Error Correction	Phonetic errors; Character errors	ZH + JA

Two tasks, four subtasks:

17 LLMs in 4 categories:
Online: GPT-4, GPT-4o/-mini, Gemini 2.0/2.5 Flash, Claude 3 Sonnet, Claude 3.5 Haiku
Offline: Llama3-8B, Llama3.2-3B, Mistral-7B, GLM4-9B
Chinese-specific: Qwen2-7B, Baichuan2-7B
Japanese-specific: RakutenAI-7B/-mini, Nekomata-7B, Swallow-8B
Traditional baselines: macOS Pinyin/Romaji, Pinyin2Hanzi (HMM/DAG), pycorrector, OpenNMT, WanaKana

8 datasets across generation and correction in ZH & JA. Phonetic-error sets were **synthetically created** by substitutions, deletions, and insertions.

Metrics: SimHash, BERTScore, BLEU, ROUGE-L, CER, Completion Time, TTFT, TPS

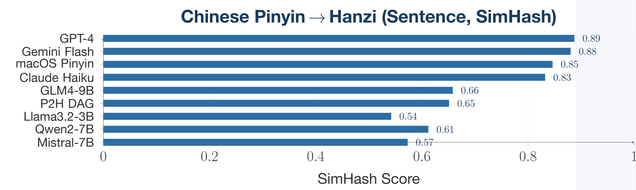
3. FINDING 1 —Word-Level Generation

Traditional baselines win at isolated word conversion. macOS Pinyin leads Chinese words (SimHash 0.820, BLEU 0.637) with near-instant response; Pinyin2Hanzi DAG is also near-instantaneous (<0.01s).

- For Japanese, macOS Romaji achieves the best BERTScore (0.952), BLEU (0.849), and CER (0.085)
- Gemini 2.0 Flash is the best LLM on Japanese words (SimHash **0.831**)
- Word-level conversion is still mostly a **lookup problem**

4. FINDING 2 —Sentence-Level Generation

At the **sentence level**, LLMs dominate by leveraging full-sentence context. Pinyin2Hanzi collapses —n-gram models cannot resolve long-range disambiguation.



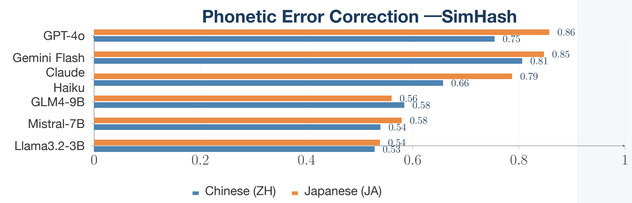
Data from Table 10/3 (Zou et al., 2026). GPT-4 achieves highest SimHash, BLEU (0.906); Gemini Flash highest BERTScore (0.977).

Japanese surprise: Language-specific models (Nekomata 0.549, RakutenAI 0.557) *designed for Japanese* fall far below macOS Romaji (0.802) and Gemini Flash (**0.890**). **Scale beats specialization.**

Key insight: Gemini 2.0 Flash delivers accuracy nearly matching GPT-4 but completes in **0.759 s vs. 2.649 s** (3.5× faster) with the best TPS (**92.30** tok/s).

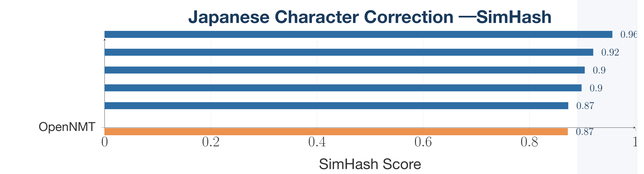
5. FINDING 3 —Phonetic Error Correction

No traditional baseline exists for corrupted phonetic input —standard IMEs cannot operate on noisy Pinyin/Romaji. This is an **entirely new capability** provided only by LLMs.



Gemini 2.0 Flash leads Chinese (SimHash 0.806) while being 2× faster than GPT-4o (1.082 s vs. 2.316 s). **GPT-4o leads Japanese** (0.857). Offline models struggle but are not zero —even Llama3.2-3B produces partial output (CER ≈1.0).

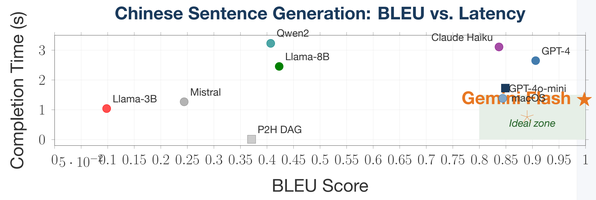
6. FINDING 4 —Character Error Correction



GPT-4o achieves near-perfect performance (SimHash 0.956, BERTScore 0.995, CER 0.013) in **1.665 s** —while OpenNMT takes **4.175 s** for comparable accuracy. Notably, **RakutenAI-7B** excels here (0.898) despite poor generation scores.

7. ACCURACY—LATENCY TRADE-OFF

Ideal model: **high BLEU, low latency** (bottom-right). Latency is a hard constraint for real-time text input.



Gemini 2.0 Flash ★ sits closest to the ideal zone: BLEU 0.890 at **0.759 s**. GPT-4 leads accuracy (0.906) but at 3.5× higher latency.

8. CONCLUSIONS

Word gen: Traditional baselines win —LLMs often worse

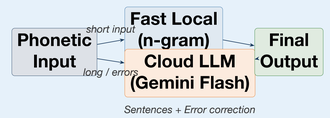
Sentence gen: LLMs dominate via full-sentence context

Error corr: LLMs only —no traditional system can even attempt phonetic correction

Best model: Gemini 2.0 Flash balances accuracy & speed

Scale wins: Frontier LLMs beat language-specific fine-tuned models even in their native language

Proposed Hybrid IME Architecture



Instant response for simple inputs; near-perfect accuracy for complex ones.

References Zou, Lee, Fan & Li (2026). *Benchmarking LLMs for Chinese and Japanese IMEs.* LREC-COLING 2026. Sun et al. (2021). *ChineseBERT.* ACL. Zhang et al. (2020). *BERTScore.* ICLR. Papineni et al. (2002). *BLEU.* ACL.

Code & Data `github.com/junli-cuny/llm-ime-benchmark`

Scan for code & data

junli@qc.cuny.edu