

Motivation

LLM exhibit widespread **language confusion**: producing tokens from the wrong language during multilingual generation. We benchmark this across leading commercial and open-source LLMs on FLORES-NO-LATIN, measuring the percentage of confused Chinese/Japanese (CJ) and Latin characters alongside task-specific BLEU scores. The results highlight that even strong models like GPT and Claude suffer from nontrivial confusion rates, as shown in the example on the right.

Table 2: Language Confusion Rates on the FLORES-NO-LATIN Benchmark for Leading Commercial LLMs. This table displays the percentage of responses containing erroneous Chinese/Japanese (CJ%) and Latin (Latin%) characters, alongside the task-specific BLEU score. These results highlight that language confusion is a persistent issue across various SOTA models.

Model	CJ%	Latin%	BLEU
GPT-5-Chat	0.57	0.67	10.66
Claude-Sonnet-4	0.00	0.35	21.77
Gemini-2.5-Pro	0.04	0.50	19.11
DeepSeek-v3.1	0.67	1.06	18.11
Qwen3-235B-Instruct	2.27	5.07	15.43

Norm-Adjusted Self-Distillation

We find that output token embedding norms vary dramatically across language families: high-resource scripts (Latin, CJ) dominate the top-5% of embedding norms, creating a systematic bias toward these families at decoding time.

To remove this bias, we divide each logit by its embedding norm, aligning scores to cosine similarity. From these norm-adjusted predictions, we generate multi-label pseudo-targets over language families and train a lightweight gate via cross-entropy loss.

This **self-distillation** process requires no external labels.

Table 1: Percentage of tokens in each language family with embedding norms among top 5% of all token embedding norms.

Model	CJ%	Latin%	Low-Res%
Qwen3-8B	10.74	4.61	0.14
Qwen3-30B-A3B	6.52	5.50	0.07
Llama3.1-8B	4.38	5.95	1.34
Gemma3-12B	0.94	5.04	2.40
GPT-OSS	0.00	7.00	0.00

Translate the following English text into Hebrew. Only provide the translation. English: The views presented are often cursory, general and oversimplified compared to the more detailed information available elsewhere.

Hebrew:
Qwen3-8B
 הדרכים שהוצגו לעתים קרובות הן כלליות, מפורט שקיים במקום אחר מוגזמות ופשוטות מדי. בהשוואה למידע

Rank	Before Norm Adjustment			After Norm Adjustment		
	Token	Prob(%)	Norm	Token	Prob(%)	Norm
1	'更加'	26.17	1.6406	'המ'	43.75	1.0703
2	'המ'	20.41	1.0703	'מ'	26.56	1.0547
3	'更为'	11.62	1.7266	'ה'	9.77	1.0938
4	'מ'	9.62	1.0547	'יותר'	2.04	1.2188
5	'ה'	5.49	1.0938	'ב'	1.40	0.9766
6	'יותר'	4.57	1.2188	'ש'	0.45	1.0547
7	'הפר'	2.03	1.3359	'שה'	0.45	1.0156
8	'dé'	0.95	1.6250	'ו'	0.40	0.8633
9	'מע'	0.74	1.3047	'פ'	0.38	1.1562
10	'המק'	0.54	1.3750	'הפר'	0.38	1.3359

Translate the following Chinese text into Thai. Only provide the translation. Chinese: 对于哪些加工物品是古董, 没有统一的定义。一些税务机构将 100 年以上的物品定义为古董。 Thai:

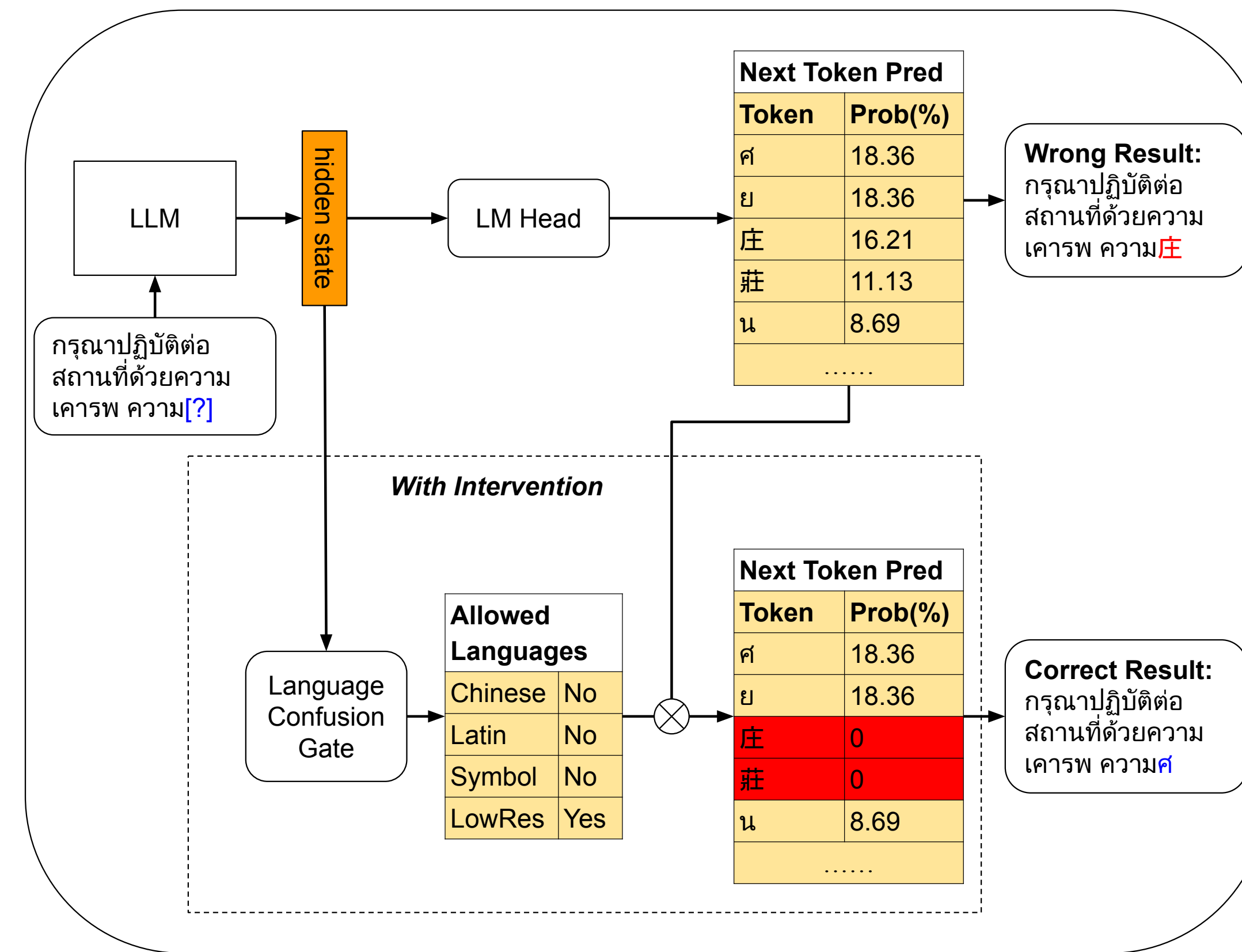
GPT-5-Chat
 สำหรับสิ่งของที่ผ่านการ加工ในบ้างจึง จะถือเป็นของโบราณ ยังไม่มีคำนิยามที่ เป็นมาตรฐาน บางหน่วยงานด้านภาษี ถือว่าสิ่งของที่มีอายุมากกว่า 100 ปีเป็น ของโบราณ

Translate the following English text into Korean. Only provide the translation. English: One of the most noteworthy recent examples of this was the North Atlantic campaign of WWII. The Americans were trying to move men and materials across the Atlantic Ocean to help Britain. Korean:

Qwen3-235B-2507-Instruct
 최근의 주목할 만한 사례 중 하나는 제2차 세계대전 당시 北大 서양 작전이었다. 미국은 영국을 지원하기 위해 대서양을 가로질러 인력과 물자를 이동시키려 하고 있었다.

Translate the following English text into Hebrew. Only provide the translation. English: They have feet with scales and claws, they lay eggs, and they walk on their two back legs like a T-Rex. Hebrew:

Claude-4-Sonnet
 יש להם גליים, וציפורניים, ושקשוקים. הם הולכים על שתי הרגליים כמו ט-רקס.



How LCG Works

At each decoding step, the Language Confusion Gate reads the LLM's hidden state and predicts which language families (CJ, Latin, Symbols, Low-Res) are permissible.

When confusion is detected, it dynamically masks disallowed tokens before sampling. Three safeguards ensure robustness:

- Symbols and Low-Resource tokens are never masked
- the gate defers to the model when it has high confidence
- and the previous token's family is always permitted for language continuity

The gate is a 2-layer MLP: a plug-in module that requires no modification to the base LLM.

Main Results

LCG achieves order-of-magnitude reductions in language confusion across all tested models while preserving or improving task performance.

On FLORES-NO-LATIN, confusion rates drop dramatically

- Qwen3-8B CJ falls from 4.5% to 0.1%
- Llama3.1-8B Latin from 8.4% to 2.9%
- BLEU scores unchanged or slightly improved.

LCG is currently used in Qwen's Production system

	Qwen3-30B	Llama3.1-8B	Gemma3-12B	Qwen3-8B
FLORES-NO-LATIN				
CJ% (No LCG)	1.0	3.0	0.2	4.5
CJ% (LCG-unadjusted)	0.2	2.0	0.1	0.5
CJ% (LCG-adjusted)	0.0	0.4	0.1	0.1
Latin% (No LCG)	4.4	8.4	1.0	12.1
Latin% (LCG-unadjusted)	0.7	5.7	0.6	6.2
Latin% (LCG-adjusted)	0.4	2.9	0.5	2.0
BLEU (No LCG)	13.2	11.3	16.9	12.1
BLEU (LCG-unadjusted)	13.3	12.2	17.0	11.9
BLEU (LCG-adjusted)	13.4	12.3	17.1	12.1
INCLUDE				
CJ% (No LCG)	2.21	0.87	0.00	1.67
CJ% (LCG-unadjusted)	0.22	0.51	0.00	0.44
CJ% (LCG-adjusted)	0.11	0.07	0.00	0.18
Accuracy (No LCG)	71.12	46.12	64.95	61.43
Accuracy (LCG-unadjusted)	71.55	46.12	65.02	62.84
Accuracy (LCG-adjusted)	70.83	46.34	65.75	61.76

Discussion

Preservation of natural code-switch

Unlike rule-based approaches, LCG preserves legitimate language mixing. When evaluated on FLORES-WITH-LATIN when code-switching is expected, LCG permits English tokens at 86.7% of human-validated code-switch points. LCG trims noise without sacrificing expressivity.

Efficiency and compatibility with speculative decoding

- adds only ~0.4% overhead per decoding step
- fully compatible with speculative decoding.
- requires no changes to the base model's weights or architecture
- Training data is generated entirely from the model itself via norm-adjusted self-distillation, no human annotation needed