Language Confusion Gate: Language-Aware Decoding Through Model Self-Distillation

Collin Zhang[2] Fei Huang[1] Chenhan Yuan[1] Junyang Lin[1]

[1]Qwen Team, Alibaba Group [2]Cornell University

Motivation

LLMs sometimes output the **wrong script**—Chinese in Thai, Latin in Korean. This **language confusion** is widespread:

Model	CJ%	Latin%	Low-Res%
Qwen3-8B	10.74	4.61	0.14
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

Goal: **suppress confusion** while **preserving legitimate code-switching**. Prior fixes (prompting, greedy, ORPO) either degrade performance or block valid mixing.

Key Observation

At confusion points, the confused token is top-1 in 56.7% of cases, yet the correct token is in the **top 3** in 99.3%. Confusion is rare but predictable—**targeted masking** suffices.

Output embedding **norm imbalance** biases toward high-resource scripts, motivating **norm-adjusted** supervision.

Method: Language Confusion Gate

LCG is a **lightweight plug-in** (2-layer MLP, <1M params) that reads hidden states and predicts **allowed language families** at each step, masking disallowed tokens *only when needed*.

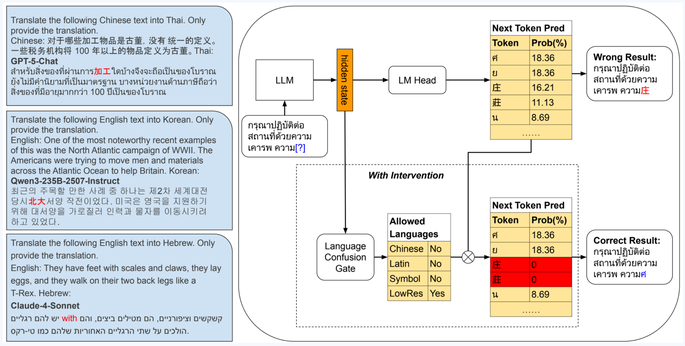


Figure 1. LCG reads hidden states, predicts allowed families, and masks confused tokens—without modifying the base LLM.

Token Family Classification

Every token is classified into four families via Unicode analysis of decoded BPE: **CJ** (27,658), **Latin** (94,666), **Symbols** (10,355), **Low-Resource** (19,257). This deterministic mapping enables **decoding-time** control with zero overhead.

Norm-Adjusted Self-Distillation

Standard logits are biased by embedding norms. We compute **norm-adjusted logits**: $\text{logit}_i^{\text{adj}} = h^\top e_i / \|e_i\|$, aligning scores to cosine similarity.

	Before Adj.		After Adj.	
Model	Top-1	Top-3	Top-1	Top-3
Qwen3-8B	43.26	0.71	14.18	0.71
Qwen3-30B	35.46	0.00	19.86	0.00
Llama3.1-8B	56.74	0.71	26.24	0.00

From adjusted predictions, we form **multi-label pseudo-targets** over families and train via **BCE loss**—self-distillation without external labels.

Smart Masking Rules

Three rules balance precision with robustness:

- **Never mask Symbols/Low-Resource**: always available.
- **Respect confidence**: if a strong candidate contradicts the gate, defer to the model.
- **Persistence**: allow the previous token's family for script continuity.

Intervention rate: only 0.33–0.38% of tokens.

Effect of Norm Adjustment

	Before Norm Adjustment			After Norm Adjustment		
Rank	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

Figure 2. Top-10 logits at a confusion point (Qwen3-8B). **Before**: high-norm Chinese tokens dominate despite Hebrew being the target. **After** dividing by $\|e_i\|$, correct Hebrew tokens rise to the top—norm imbalance, not model intent, drives confusion.

Main Results

LCG delivers **order-of-magnitude** confusion reductions with **no performance loss**:

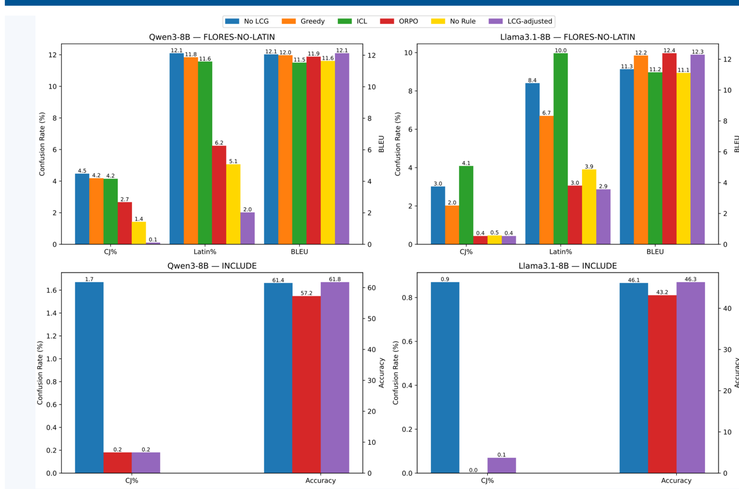
Model	Method	CJ%	Latin%	BLEU
Qwen3-8B	No LCG	4.50	12.10	12.10
	LCG-adj.	0.10	2.00	12.10
Qwen3-30B	No LCG	1.00	4.40	–
	LCG-adj.	0.00	0.40	–
Llama3.1-8B	No LCG	3.00	8.40	11.30
	LCG-adj.	0.40	2.90	12.30

On HumanEval-XL, GPT-OSS CJ drops 0.38%→0.06%. LCG-adjusted consistently beats unadjusted.

Preserving Code-Switching

LCG permits English tokens at 86.7% of human-validated code-switch points. Code-switch rates moderate naturally (Qwen3-8B: 46.3% → 25.9%) while staying close to references.

Baselines Comparison



LCG vs. baselines on FLORES-NO-LATIN and INCLUDE. ICL/Greedy barely help; ORPO may hurt accuracy.

Efficiency & Limitations

- **Overhead**: ∼0.4%/step; compatible with speculative decoding.
- **Limitation**: script-level granularity (CJ vs. Latin), not per-language. Future: fine-grained gates.