ChiEngMixBench: Evaluating Large Language Models on Expert-Style Chinese-English Terminology Mixing

Anonymous ACL submission

Abstract

The utilization of large language models (LLMs) in multilingual professional discourse is increasingly widespread, yet the coverage of existing evaluation benchmarks struggles to keep pace. To further advance these models, it is crucial to investigate whether they can align with the authentic communicative habits of human experts. We identify code-mixing as a realistic, sustainable, and challenging testbed for evaluating the expressive efficiency and alignment capabilities of next-generation language models. To this end, we introduce *ChiEngMixBench*, a benchmark specifically designed to evaluate the "Chinese-English terminology mixing" capabilities of models within Chinese technical contexts. Sourced from authentic Chinese technical communities, the dataset comprises approximately 530k characters and 1,706 test samples. In *ChiEngMixBench*, models are required to generate responses that conform to real-world community communication habits based on given contexts. This task requires models to understand contextual consistency and the language selection rules of professional communities, presenting a significantly higher difficulty than traditional translation-based multilingual evaluations. Our evaluation reveals that current models still struggle to stably reproduce the natural terminology mixing habits found in authentic human expression. While top-tier models can generate relatively fluent responses, they often lack the spontaneous tendency to retain key English terminology. More alarmingly, current mainstream instruction tuning may further suppress this professional code-mixing behavior. By bridging the gap between rigid translation metrics and dynamic, real-world pragmatic norms, our work provides a robust foundation for developing more cognitively aligned AI assistants in specialized domains.

1 Introduction

Large language models (LLMs) are increasingly deployed in technical and professional scenarios (Hurst et al., 2024; Dubey et al., 2024; DeepSeek-AI, 2024). In these contexts, high-quality language generation requires more than just grammatical correctness or simple bilingual translation capabilities. In many expert communities—particularly within Chinese AI and computer science discourse—the most natural mode of communication often involves combining a Chinese syntactic framework with English technical terminology. For instance, researchers frequently write sentences such as "检查logits 的分布" (check the distribution of logits), "先跑baseline" (run the baseline first), or "这个API 会触发timeout" (this API will trigger a timeout). This distinctive usage is by no means a random or chaotic mixing of languages; rather, it intrinsically reflects a deep terminological compactness and shared community-specific conventions, thereby optimizing both communicative efficiency and semantic precision in professional technical communication.

Existing multilingual evaluations fail to adequately test this behavior. Most benchmarks focus on translation or cross-lingual instruction following, assuming models should strictly maintain a single language or cleanly translate between them. However, real-world expert communication breaks this assumption. A pure Chinese translation may be comprehensible but feels unnatural, while randomly inserting English words disrupts established pragmatic norms. Consequently, prevailing evaluation paradigms systematically overlook the crucial question of whether a large language model can naturally match the nuanced linguistic conventions of its target professional user community.

Evaluating this capability faces two practical challenges. First, natural terminology mixing is highly context-dependent, meaning simple "syn-

thetic word substitution" cannot reliably capture it. Second, human judgments of mixed-language outputs are prone to subjective noise; evaluators may accept understandable but unidiomatic or over-translated expressions. Thus, a robust benchmark requires authentic expert data and objective quantitative signals to distinguish between "merely comprehensible bilingual text" and "truly natural community usage."

To address these challenges, we introduce *ChiEngMixBench*, a benchmark designed to evaluate LLMs on their "Chinese-English terminology mixing" capabilities within the context of Chinese AI discourse. Unlike traditional translation-oriented benchmarks, *ChiEngMixBench* focuses on whether models can replicate the terminology selection behaviors and distributional patterns of experts in real technical communities. Specifically, given a Chinese technical context where using an English term is the most natural choice for an expert, the model should spontaneously retain this highly compact English expression, seamlessly maintain the surrounding Chinese syntactic structure, and successfully avoid both unnatural monolingual over-translation and excessively fragmented or chaotic language switching.

ChiEngMixBench is constructed from authentic Chinese AI research corpora, comprising 1,706 contrastive pairs covering 1,345 technical terms. We evaluate models from two complementary perspectives. First, we test whether a model intrinsically prefers expert-style English terms over their Chinese equivalents within the identical context. We achieve this through Minimal Contrastive Pairs (MCPs) and log-likelihood probing. Second, we assess whether the generated responses resemble authentic expert writing. To do this, we introduce an automated naturalness metric that compares model outputs to expert reference distributions across three dimensions: structural, semantic, and terminological compactness. These two perspectives allow us to clearly disentangle the model's intrinsic terminological preferences from its open-ended generation quality.

Our evaluation of 12 representative LLMs reveals several consistent patterns. First, while models show a strong tendency to continue the linguistic inertia of the surrounding Chinese context, they systematically demonstrate a higher internal probability preference for the English forms of specialized technical terms. Second, stronger instruction tuning does not always improve expert-style multilingual behavior; in some cases, it actively suppresses natural terminology mixing, pushing the model toward overly monolingual outputs. Finally, we find that our automated naturalness metric correlates highly with expert consensus, indicating that distributional signals from authentic expert discourse can provide a reliable anchor for evaluating this capability.

In summary, our main contributions are:

- We formalize expert-style "Chinese-English terminology mixing" as a practical multilingual alignment problem for LLMs, rather than a simple translation or synthetic code-switching task.
- We construct *ChiEngMixBench* based on authentic AI technical discourse and provide a dual-dimensional evaluation framework to assess both intrinsic terminological preference and generation naturalness.
- We empirically characterize how modern LLMs handle expert terminology mixing, revealing systematic term-level preferences and the potential performance degradation introduced by current instruction tuning pipelines.

Reproducibility Statement All experimental data, implementation code, and comprehensive evaluation scripts are made fully available within the supplementary material and will be released as a public open-source repository upon the acceptance of this work.

2 Related Work

This study intersects code-switching dataset construction, LLM multilingual evaluation, and cognitive linguistic theories (Winata et al., 2023).

2.1 Code-Switching Datasets

Early code-switching research focused on speech (Lyu et al., 2010) and later shifted to social media text (Aguilar et al., 2020; Srivastava and Singh, 2020; Khanuja et al., 2020). While providing naturalistic data, these corpora are saturated with informal slang, complicating the isolation of pragmatic rules in high-stakes discourse. Recent benchmarks expand language coverage (Yang and Chai, 2025) and explore expert domains (Li et al., 2023). However, many datasets still rely on automated synthesis or random substitution (Pratapa et al.,

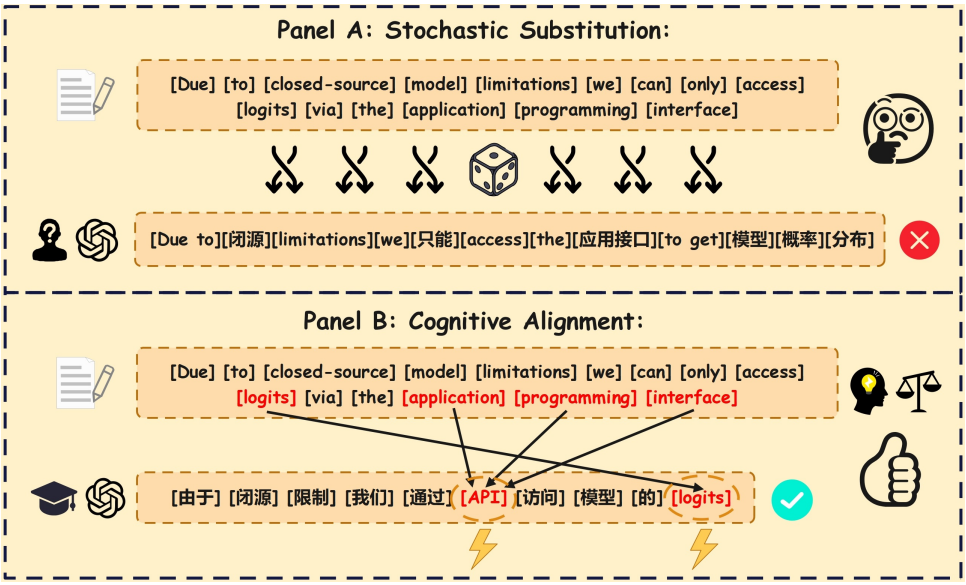


Figure 1: **Contrast between Stochastic Substitution and Pragmatic Alignment strategies.** *Panel A* illustrates the pitfalls of random word replacement, where forcing English terms into a literal Chinese translation results in awkward phrasing (marked by a red cross). *Panel B* demonstrates the focus of *ChiEngMixBench*: a cognitively aligned approach where high-entropy *Anchor Terms* (e.g., "API", "logits") are spontaneously retained within a natural Chinese syntactic framework, adhering to expert community norms (marked by a green check).

2018; Rizvi et al., 2021), disrupting the subtle *pragmatic prosody* that experts spontaneously generate. ChiEngMixBench fills this gap by providing an authentic, community-derived benchmark preserving structural integrity.

2.2 Code-Mixing Evaluation in LLMs

Existing work on LLM code-mixing evaluation typically spans three paradigms: **Understanding** tasks assessing comprehension of mixed inputs (Huzaifah et al., 2024; Marchisio et al., 2024; Lee et al., 2025); **Explicit Generation** via prompting (Yong et al., 2023), which often produces unnatural switches; and **Task-Specific Benchmarking** on structured tasks like NER (Sheokand and Sawant, 2025; Shirke et al., 2025).

We move beyond capability assessments by evaluating whether LLMs spontaneously and naturally adopt expert-style code-mixing. This aligns with *Output Language Alignment* (OLA) (Kim et al., 2026), emphasizing the necessity of matching specialized linguistic norms of target communities to improve interaction efficiency.

2.3 Linguistic Theoretical Foundations

We ground our evaluation in three theories: *Zipf's Principle of Least Effort* (Zipf, 1949), where bilinguals minimize cognitive load by selecting accessible labels; the *Matrix Language Frame* (Myers-Scotton, 1993), positing asymmetrical mixing with Chinese as the morphosyntactic frame and English as content morphemes; and the *Revised Hierarchical Model* (Kroll and Stewart, 1994), linking specialized concepts directly to L2 labels. By treating code-mixing as a normative *Unmarked Choice*, we move beyond translation metrics to address the *Subjectivity Paradox*, utilizing EDP as a proxy for community norms.

3 Dataset Construction

To evaluate LLM code-mixing behaviors within a context that is both *controlled* and *authentic*, we constructed *ChiEngMixBench*. The construction pipeline follows three phases: authentic data acquisition, spontaneity subset (MCP) generation, and naturalness baseline profiling (see Figure 2).

3.1 Data Acquisition and Cleaning

We prioritize *Ecological Validity* by sourcing raw text from high-quality Chinese technical communities such as CSDN and Zhihu, spanning May 2021 to November 2025. This timeframe ensures coverage of emerging AI terminology including "LoRA" and "RLHF".

The raw corpus (approximately 1.35M characters) underwent a rigorous multi-stage cleaning process:

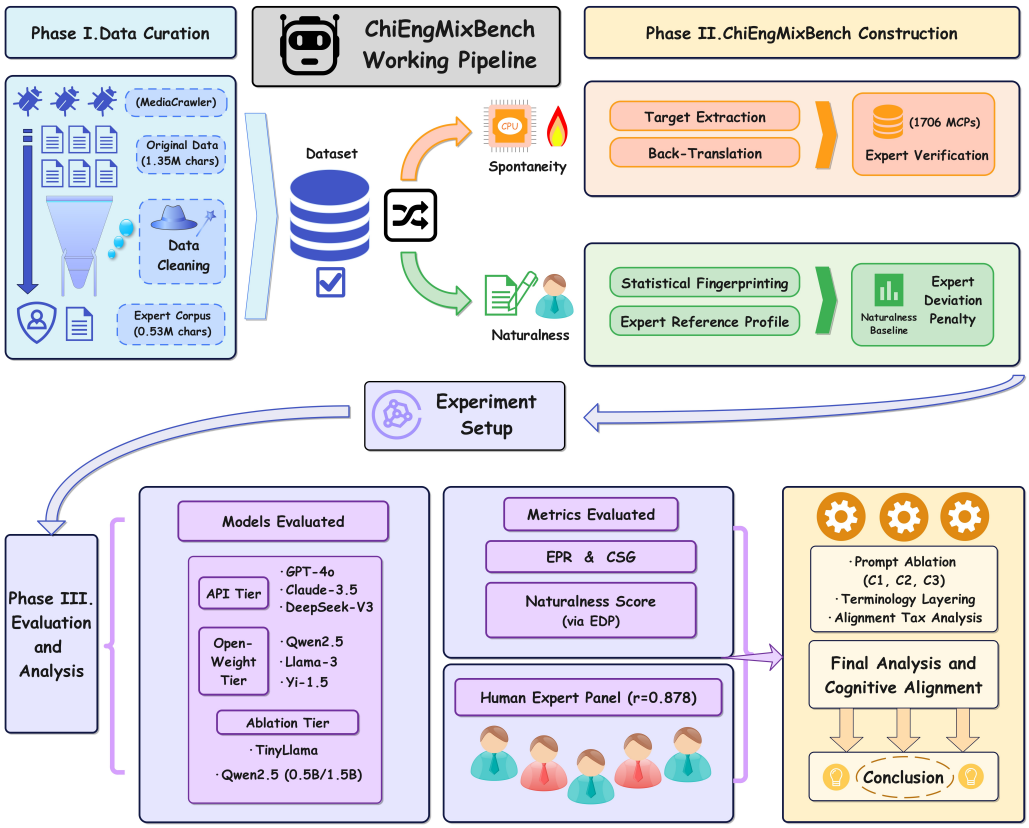


Figure 2: **The *ChiEngMixBench* Construction Pipeline.** *Phase I* illustrates the data curation process from raw technical forums to a cleaned expert corpus. *Phase II* details the dual-track dataset construction: extracting Minimal Contrastive Pairs (MCPs) via back-translation and expert verification to evaluate *Spontaneity*, alongside extracting statistical fingerprints to establish an Expert Reference Profile for *Naturalness*. *Phase III* outlines the evaluation of representative LLMs using our proposed metrics (EPR, CSG, and EDP) and prompt ablation to derive the final cognitive alignment analysis.

1. *Deduplication:* A hash-based exact matching algorithm removed duplicate documents.

2. *Content Filtering:* We applied regular expressions to filter out fragmented sentences (< 20 characters) and purely monolingual Chinese text.

3. *De-identification:* A strict PII (Personally Identifiable Information) scrubbing pipeline removed user IDs, avatar links, and timestamps to ensure privacy.

The final cleaned corpus consists of *530k characters* of expert discourse.

3.2 Minimal Contrastive Pair (MCP) Construction

To isolate the model's token selection preference from contextual noise, we developed a three-step hybrid pipeline to construct Minimal Contrastive Pairs (MCPs). Each MCP consists of an authentic code-mixed sentence (S_{mix}) and its strict monolingual Chinese equivalent (S_{mono}).

Target Extraction. We identify expert code-mixed sentences from the corpus and extract the embedded English terminology as the target token t_{en}.

Contrastive Mapping. We utilize a meticulously curated terminology mapping table—such as mapping "CNN" to "卷积神经网络"—to replace t_{en} with its standard Chinese equivalent t_{zh}. This generates the S_{mono} sequence while keeping the syntactic skeleton identical. To ensure high-quality mappings, we follow a *Back-Translation + Expert Verification* pipeline: for each candidate t_{zh}, we verify if it can be back-translated to t_{en} in technical contexts without ambiguity.

Expert Verification. To ensure strict semantic equivalence and syntactic fluency, domain experts manually validate each pair. We enforce a "Zero Semantic Drift" rule: any pair where the replace-

ment alters the technical nuance or results in disfluent Chinese syntax is discarded.

This pairwise design ensures that S_{mix} and S_{mono} differ *only* in their terminological form. Consequently, variance in log-likelihood (logits) probing can be definitively attributed to the model's internal preference rather than external semantic shifts.

3.3 Dataset Stratification and Statistics

The final dataset contains *1,706 high-quality MCP pairs* covering *1,345 unique terms*. To investigate the hierarchy of code-mixing, we stratify the terms based on domain specificity:

- *Specialized Terms (92.2%):* Domain-specific proper nouns, acronyms, and high-entropy technical terms such as *Transformer* and *Logits* that lack concise Chinese equivalents. These terms are characterized by low document frequency ($< 0.1\%$) but high burstiness in technical discussions.

- *General Terms (7.8%):* High-frequency academic vocabulary including *zero*, *version*, and *label* that are common across general domains but frequently appear in mixed discourse. These are identified via a pre-defined stop-list of 200 common cross-lingual anchors.

The *11.8:1* ratio reflects the reality of expert discourse: specialized terms form a vast, long-tailed distribution (Zipfian $\alpha = 1.16$), while general terms are few in type but frequent in usage. The average sentence length is 41.0 characters, providing a robust contextual foundation for assessing the model's linguistic consistency.

3.4 Naturalness Baseline Construction

To provide an empirical anchor for the Expert Deviation Penalty (EDP) metric, we established an objective *Expert Reference Profile* by extracting the "statistical fingerprints" of human expert discourse from the cleaned corpus. We calculate the *Semantic Distribution Center* ($\mu_{gold}, \sigma_{gold}$) using `all-MiniLM-L6-v2` (Wang et al., 2020) and derive distributional boundaries for English ratio and switching frequency. Furthermore, we maintained a *Terminology Compactness Dictionary*—for instance, mapping "CNN" to its verbose equivalents {"卷积神经网络", "Convolutional Neural Network"}—to identify and penalize non-compact expressions that deviate from expert conventions.

4 Evaluation Framework

To comprehensively assess expert-style code-mixing, we propose a dual-dimensional framework (Figure 2), quantifying model capabilities from both internal probability (Spontaneity) and external generation (Naturalness) perspectives.

4.1 Spontaneity Assessment Framework

To evaluate Spontaneity without the confounding factors of open-ended generation, we employ constraint-based probabilistic probing on Minimal Contrastive Pairs (MCPs).

Scoring Function For a target term t consisting of tokens $x_1, \ldots, x_L$ within a given Chinese context prefix C, we calculate the length-normalized log-likelihood $\mathcal{S}(t|C)$:

$$\mathcal{S}(t|C) = \frac{1}{L} \sum_{i=1}^{L} \log P_\theta(x_i|C, x_{<i}) \quad (1)$$

where P_θ represents the probability distribution of the evaluated LLM, x_i denotes the i-th token of the term, and $x_{<i}$ represents the preceding tokens of the term. Length normalization ($1/L$) is strictly applied to prevent inherent bias against longer English subword tokenizations.

Metrics Using this scoring function on both the English term (t_{en}) and its strict Chinese equivalent (t_{zh}) within the identical context C, we derive two core metrics:

- *English Preference Rate (EPR):* The proportion of samples across the dataset (N) where the model assigns a higher probability to the English term:

$$\text{EPR} = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}\left(\mathcal{S}(t_{en}^{(j)}) > \mathcal{S}(t_{zh}^{(j)})\right) \quad (2)$$

- *Contextual Spontaneity Gap (CSG):* The continuous log-likelihood difference for a specific pair:

$$\text{CSG}^{(j)} = \mathcal{S}(t_{en}^{(j)}|C^{(j)}) - \mathcal{S}(t_{zh}^{(j)}|C^{(j)}) \quad (3)$$

The magnitude of CSG carries explicit physical significance. While CSG is typically negative due to the strong Contextual Inertia of the Chinese prefix (biasing the model toward t_{zh}), we focus on the *Relative Lift*. If specialized terms

exhibit a higher CSG than general terms (i.e., $\text{CSG}_{spec} > \text{CSG}_{gen}$), it empirically demonstrates the model's intrinsic **terminological activation preference** to overcome monolingual inertia for communication efficiency.

4.2 Naturalness and the Expert Deviation Penalty

To evaluate Naturalness, we propose the *Expert Deviation Penalty (EDP)*, an objective automated metric that calculates the statistical distance between generated text and authentic expert distributions. The score starts at a baseline of 5.0 and applies graded deductions based on three pragmatic dimensions. These penalty weights—such as 0.8 and 1.6—and thresholds are tuned on a validation set to maximize correlation with expert consensus.

$$\text{EDP} = 5.0 - (P_{sem} + P_{str} + P_{trm}) \quad (4)$$

Crucially, the validation set used for tuning weights (≈100 samples) is strictly disjoint from the final evaluation set to ensure the metric's generalizability and prevent over-fitting.

Semantic Style Consistency (P_{sem}) We establish a *Golden Centroid* $(\mu_{gold}, \sigma_{gold})$ by projecting the expert corpus into a high-dimensional semantic space. For a generated sentence with cosine distance d from the centroid:

$$P_{sem} = \begin{cases} \lambda_s \cdot d & \text{if } d \leq \mu_{gold} + 1.2\sigma_{gold} \\ \gamma_s \cdot d & \text{if } d > \mu_{gold} + 1.2\sigma_{gold} \end{cases} \quad (5)$$

where the threshold 1.2σ covers ≈89% of expert variations to ensure robustness, and $\gamma_s \gg \lambda_s$ triggers a severe penalty for semantic deviations such as a machine-translated tone.

Structural Distribution (P_{str}) P_{str} aggregates deductions from the English token ratio (r_{en}) and switching frequency (f_{sw}):

$$\begin{aligned} P_{str} = &\max(0, |r_{en} - \mu_r| \cdot \alpha_r) \\ &+ \max(0, (f_{sw} - \tau_{sw}) \cdot \beta_{sw}) \end{aligned} \quad (6)$$

The threshold $\tau_{sw} = \mu_{sw} + 1.2\sigma_{sw}$ penalizes fragmented "Word Salad" while maintaining tolerance for natural switching.

Terminological Compactness (P_{trm}) Using a universal reference dictionary, we penalize pragmatic violations based on cognitive burden. The relative weights of these penalties reflect the prioritized consensus of our human expert panel:

- *Format Flaw:* Minor formatting errors, such as using the lowercase "cnn", incur a 0.8 deduction. These surface-level errors represent **minor deviations** from formal writing standards.
- *Pragmatic Violation:* Verbose expansions, such as using the full name "Convolutional Neural Network", incur a 1.6 deduction. These represent **major failures** in community alignment, as verbose expressions violate the Principle of Least Effort and signal non-expert behavior.

Qualitative calibration showed that experts rank terminological redundancy as twice as detrimental to professional communication as minor casing errors, justifying the 1:2 penalty ratio.

4.3 Mitigating the Subjectivity Paradox

A critical challenge in evaluating code-mixing is the *Subjectivity Paradox*: human evaluators often exhibit systematic bias due to differing internal thresholds for naturalness. Our human evaluation analysis reveals two distinct archetypes: *Archetype E* (Efficiency-driven experts who prefer compact switching) and *Archetype P* (Prescriptive-leaning evaluators who tolerate verbose translations).

Directly using human scores as a gold standard introduces significant noise. EDP mitigates this by serving as a *Conservative Lower Bound*. By anchoring penalties to the objective statistical boundaries of authentic expert discourse (1.2σ), EDP filters out subjective leniency. As shown in Section 5.5, EDP achieves a high correlation ($r = 0.878$) with the expert consensus, outperforming individual human scorers in consistency.

5 Experiments and Analysis

5.1 Experimental Setup

We evaluate a diverse set of representative LLMs across *Scaling Families*, *Language Bias Groups*, and *Alignment Stages* (detailed in Appendix C) to answer three core questions: (Q1) Do models spontaneously code-mix? (Q2) Is the generated text natural? and (Q3) Does the Terminology Layering Strategy exist? All experiments were strictly conducted in a *Zero-shot* setting. For Spontaneity, we utilized Logits Probing; for Naturalness, we simulated expert dialogues using 32 mixed-context prompts.

Table 1: **Main Evaluation Results on ChiEngMixBench. RQ1 (Spontaneity)**: EPR (%) and *CSG* (Logits). **RQ2 (Naturalness)**: EDP Score (0-5). EDP $= 5.0 - (P_{sem} + P_{str} + P_{trm})$. Colors: Blue Spontaneity; Green Naturalness. 1st, 2nd, 3rd ranks.

Model	**Type**	**Size**	**RQ1: Spontaneity**				**RQ2: Naturalness (Deviations)**				
			EPR	**Gen**	**Spec**	**Lift**	**EDP**	P_{str}^{rat}	P_{str}^{sw}	P_{sem}	P_{trm}
Human Expert	*Native*	–	–	–	–	–	*4.85*	0.00	0.05	0.10	0.00
Tier 1: Closed-Source API Models											
GPT-4o	API	–	–	–	–	–	*4.36*	0.00	0.00	0.45	0.19
DeepSeek-V3	API	–	–	–	–	–	4.19	0.00	0.00	0.45	0.21
Claude-3.5-Sonnet	API	–	–	–	–	–	4.16	0.07	0.05	0.45	0.27
Tier 2: Open-Weight Models (> 6B)											
Llama-3-8B-Instruct	Weights	8B	11.72	-0.37	-0.32	+0.05	*4.09*	0.01	0.05	0.46	0.39
Mistral-7B-v0.3	Weights	7B	12.43	-0.25	-0.20	+0.05	3.14	2.88	0.05	0.47	0.33
Yi-1.5-9B-Chat	Weights	9B	21.63	-0.31	-0.25	+0.06	3.10	1.03	0.00	0.43	0.69
Qwen2.5-7B-Base	Weights	7B	14.36	-0.42	-0.39	+0.03	2.96	1.28	0.05	0.50	0.48
Qwen2.5-7B-Instruct	Weights	7B	14.54	-0.43	-0.39	+0.04	2.71	4.30	0.06	0.48	0.15
Gemma-2-9B-It	Weights	9B	7.62	-0.62	-0.56	+0.06	1.23	8.25	0.00	0.47	0.15
Tier 3: Ablation & Analysis Models (< 3B)											
Qwen2.5-0.5B-Instruct	Weights	0.5B	15.59	-0.47	-0.40	+0.07	*4.17*	0.36	0.03	0.47	0.19
Qwen2.5-1.5B-Instruct	Weights	1.5B	12.90	-0.47	-0.40	+0.06	2.20	5.42	0.02	0.48	0.29
TinyLlama-1.1B	Weights	1.1B	10.21	-0.21	-0.25	-0.03	1.77	6.81	0.00	0.51	0.40

Table 2: **Alignment with Human Experts.** ($N = 100$).

Metric	*Eval 1*	*Eval 2*	$\cdots$	*Consensus*
Pearson (r)	0.800	0.790	$\cdots$	*0.878*
QWK (κ)	0.576	0.631	$\cdots$	*0.770*

5.2 RQ1: Spontaneous Drive

Table 1 presents the core evaluation results.

Systematic Shift in Subconscious: Experiments reveal the absolute dominance of *Contextual Inertia*: even English-dominant models fail to exceed 20% EPR. However, Specialized Terms exhibit higher CSG than General Terms (Lift $\approx +0.04$– $+ 0.07$), validating the "Entropy-Driven Mechanism."

Case Study: The "API" Phenomenon: The CSG for the term "API" significantly crosses the decision threshold (CSG > 0), peaking at $+0.937$. In contrast, the term "Label" is consistently suppressed (-0.4). This contrast reveals LLMs' adherence to *Zipf's Principle of Least Effort*: "API" breaks monolingual alignment.

5.3 RQ2: Naturalness & Alignment Tax

Closed-Source vs. Open-Weight: *GPT-4o* achieves the highest Naturalness score (*4.36*), followed closely by *DeepSeek-V3* (*4.19*), both aligning well with the Matrix Language Frame. In contrast, *Llama-3-8B* (*4.09*) exhibits slightly higher P_{trm} due to occasional redundant verbose phrasing.

The Alignment Tax: Comparing Qwen2.5-7B Base vs. Instruct, we observe significant degradation. Instruct models often exhibit "monolingual collapse," triggering massive P_{rat} and P_{sw} penalties. Current RLHF suppresses the nuanced mixing observed in base models.

5.4 RQ3: Terminology Layering & Cognitive Economy

Ablation Analysis: Our results (Appendix D) indicate a *Consistency-Constraint Trade-off*. While expert-styled prompts improve the Spontaneity of proprietary models like *GPT-4o*, forcing code-mixing (C3) leads to a sharp increase in P_{trm}. This confirms that expert-style mixing is an emergent capability linked to **terminological density** rather than simple instruction following.

5.5 Human Evaluation

We conducted a blind review ($N = 100$) using Stratified Random Sampling. Three independent NLP experts provided 1–5 Likert-scale ratings. Final consensus was the mean of the two closest scores to mitigate outliers.

As shown in Table 2 and Fig. 5, our automated metric correlates strongly with *Human Consensus* ($r = 0.878$).

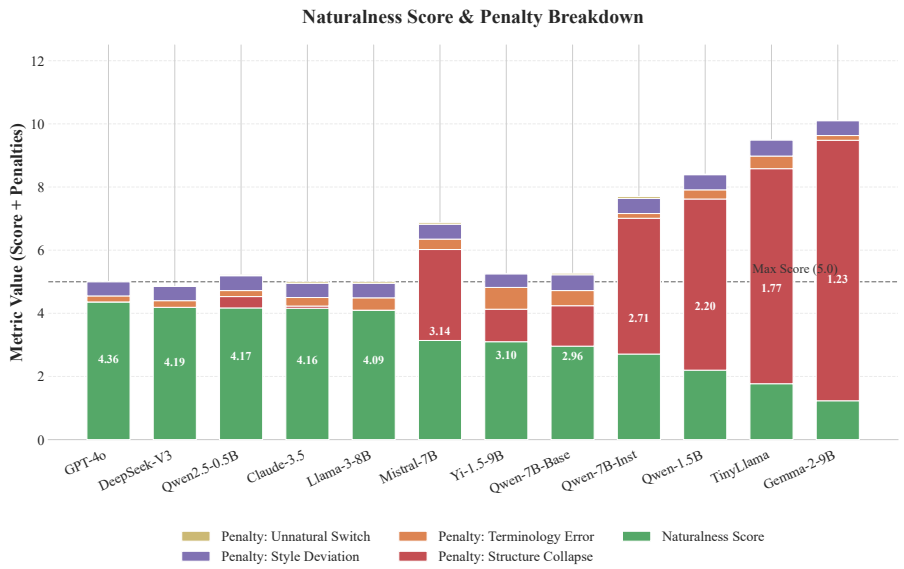


Figure 3: **Naturalness Score & Penalty Breakdown.** A detailed comparison of EDP components across models. The stacked bars reveal that while base models exhibit balanced profiles, instruct models suffer from severe structural penalties (P_{str}) due to monolingual collapse.

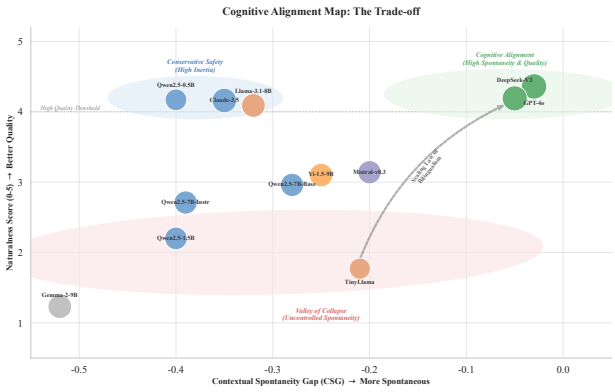


Figure 4: **Cognitive Alignment Map.** Visualizing the trade-off between Spontaneity (CSG) and Naturalness (EDP) across evaluated models.

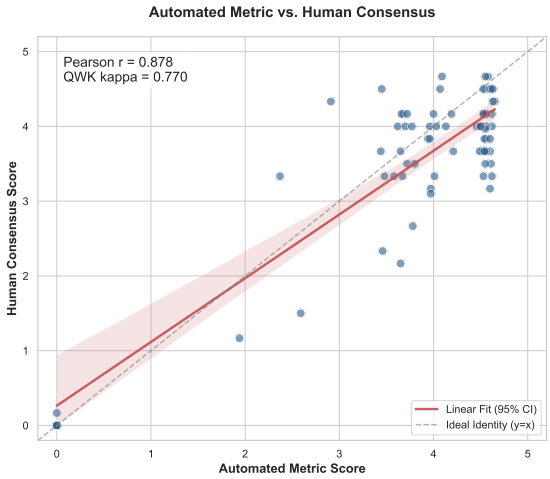


Figure 5: **Human-Machine Consistency.** Scatter plot demonstrating the strong correlation ($r = 0.878$) between automated EDP scores and human expert consensus.

While EDP serves as a conservative lower bound, its rigor is intentional. For the *High-Difficulty Subset*, ambiguous terms underwent human double-verification before penalty calculation. This ensures EDP reflects genuine linguistic misalignment rather than corpus noise, effectively resolving Subjectivity Bias by anchoring evaluation to the expert corpus's consensus.

6 Conclusion

In this paper, we introduced **ChiEngMixBench**, the first benchmark for *Expert-Style Insertional Code-Mixing*. Bridging established linguistic theories with NLP evaluation, we move beyond translation-style benchmarks to assess the pragmatic and structural alignment of LLMs.

Our framework reveals two emergent phenomena: the **Terminology Layering Strategy**, where models learn conceptual hierarchies, and the **Alignment Tax**, where RLHF suppresses spontaneous expert-style mixing. Furthermore, we address the **Subjectivity Paradox** by introducing the **Expert Deviation Penalty (EDP)**, an automated metric achieving high alignment with expert consensus. These findings underscore the need for evaluation systems respecting the linguistic norms of specialized communities. We make our dataset and evaluation suite fully open-source to facilitate research into alignment, expertise, and multilingualism.

Limitations

While ChiEngMixBench provides a rigorous framework for evaluating expert-style code-mixing, several limitations remain to be addressed in future work.

Domain Specificity Our current instantiation focuses primarily on the AI research community. While this domain is highly representative of technical terminology mixing, the pragmatic norms and "Expert Distributions" may differ significantly in other high-stakes fields such as clinical medicine or international maritime law. Future iterations should extend the construction pipeline to diverse expert communities to verify the cross-domain generalizability of the Expert Deviation Penalty.

Language Pair Coverage The benchmark is currently limited to Chinese-English mixing. Although this is the most prevalent pair in technical discourse globally, the structural constraints of the Matrix Language Frame may manifest differently in other language pairs (e.g., Hindi-English or Spanish-English), particularly where the Matrix Language lacks the rigid head-initial or head-final consistency of Mandarin.

The Temporal Dynamics of Expertise Technical terminology is inherently dynamic. Terms that are currently specialized (Anchor Terms) may eventually become loanwords or gain concise monolingual equivalents over time. Our pipeline's reliance on a static corpus timeframe (2021–2025) means the evaluation must be periodically updated to reflect the evolving linguistic repertoire of the expert community.

Ethics Statement

We strictly adhere to ethical guidelines for data collection and human evaluation. During the data acquisition phase, a rigorous Personally Identifiable Information (PII) scrubbing pipeline was executed prior to any data processing. All direct identifiers, including user IDs, avatar links, and exact timestamps, were permanently removed from the corpus to ensure user privacy and prevent deanonymization through cross-platform correlation.

Furthermore, human evaluation and verification were conducted by domain-knowledgeable annotators with academic backgrounds in computer science. All participants are native Chinese speakers proficient in English technical terminology. Annotators were fully informed of the research purpose and provided explicit consent. We ensured fair compensation for all participants, with hourly rates exceeding local standards for specialized technical auditing. Given the non-intrusive nature of the task and the voluntary participation of informed adults, this study followed our institutional ethical guidelines and was deemed exempt from formal Institutional Review Board (IRB) review.

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. Lince: A centralized benchmark for linguistic code-switching evaluation. *Preprint*, arXiv:2005.04322.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Howard Giles and Peter F Powesland. 1975. *Social psychology and language*. Academic Press.

David W Green. 1998. Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and cognition*, 1(2):67–81.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Muhammad Huzaifah, Weihua Zheng, Nattapol Chanpaisit, and Kui Wu. 2024. Evaluating code-switching translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6381–6394, Torino, Italy. ELRA and ICCL.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Ji-Hoon Kim and 1 others. 2026. Ola: Output language alignment in code-switched llm interactions. *arXiv preprint arXiv:2601.05432*.

Judith F Kroll and Erika Stewart. 1994. Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, 33(2):149.

Jean Lave and Etienne Wenger. 1991. *Situated learning: Legitimate peripheral participation*. Cambridge university press.

Nahyun Lee, Yeongseo Woo, Hyunwoo Ko, and Guijin Son. 2025. Controlling language confusion in multilingual LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1026–1035, Vienna, Austria. Association for Computational Linguistics.

Jian Li and 1 others. 2023. Unlocking the secrets behind advanced artificial intelligence language models in deidentifying chinese-english mixed clinical text. *JMIR Medical Informatics*, 11:e48592.

Dau-Cheng Lyu, Tien-Ping Tan, Eng-Siong Chng, and Haizhou Li. 2010. Seame: a mandarin-english code-switching speech corpus in south-east asia. In *Interspeech*.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA. Association for Computational Linguistics.

Carol Myers-Scotton. 1993. *Duelling Languages: Grammatical Structure in Codeswitching*. Oxford University Press.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553, Melbourne, Australia. Association for Computational Linguistics.

Mohd Sanad Zaki Rizvi, Anirudh Srinivasan, Tanuja Ganu, Monojit Choudhury, and Sunayana Sitaram. 2021. GCM: A toolkit for generating synthetic code-mixed text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 205–211, Online. Association for Computational Linguistics.

Ankit Sheokand and Sarita Sawant. 2025. Codemixbench: Evaluating large language models on code generation with code-mixed prompts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Aditya Shirke, Pratik Shembade, Sandeep Thorat, and 1 others. 2025. Comparative study of pre-trained bert and large language models for code-mixed named entity recognition. *Soft Computing*.

Vivek Srivastava and Mayank Singh. 2020. PHINC: A parallel Hinglish social media code-mixed corpus for machine translation. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 41–49, Online. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Yilun Yang and Yekun Chai. 2025. CodeMixBench: Evaluating code-mixing capabilities of LLMs across 18 languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2139–2169, Suzhou, China. Association for Computational Linguistics.

Zheng-Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Alham Fikri Aji, and Wilson Yan. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south east asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 10–27, Toronto, Canada. Association for Computational Linguistics.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

George Kingsley Zipf. 1949. *Human behavior and the principle of least effort*. Addison-Wesley Press.

A List of Terms and Rules

We provide samples from `terms.csv` and the mapping rules used in the terminology check.

Example of mapping rules:

- “卷积神经网络” → “CNN”
- “支持向量机” → “SVM”

B Theoretical Background

To provide a comprehensive interdisciplinary perspective, this appendix elaborates on the linguistic and cognitive theories that underpin the design of ChiEngMixBench. These theories bridge the gap between human linguistic behaviors and the metrics proposed in our evaluation framework.

B.1 Sociolinguistic Identity & Naturalness

The metrics for *Naturalness* (specifically the Expert Deviation Penalty) are grounded in sociolinguistic theories explaining why experts adhere to specific code-mixing norms.

Communication Accommodation Theory (CAT) Proposed by Giles and Powesland (1975), CAT posits that during social interaction, individuals unconsciously adjust their speech styles to minimize social differences (*Convergence*) or emphasize distinctiveness (*Divergence*).

- *Theoretical Relevance:* In the context of Human-AI interaction within a specialized domain, the human user (expert) expects the AI interlocutor to exhibit *Convergence* behavior.
- *Mapping to Metrics:* If an expert user employs the code-mixed term "Run Baseline," but the LLM responds with a purely monolingual Chinese translation (e.g., "运行基线"), this constitutes *Divergence*, increasing the psychological distance and signaling a lack of domain expertise. Our *Expert Deviation Penalty (EDP)* effectively quantifies the degree of divergence; by penalizing non-expert expressions, we implicitly align the model's responses with the target community.

Community of Practice (CoP) Lave and Wenger (1991) define a Community of Practice as a group of people who share a concern or a passion for something they do and learn how to do it better as they interact regularly. A key feature of a CoP is the development of a *Shared Repertoire*—including specialized vocabulary, acronyms, and styles—which serves as a boundary marker for group membership.

- *Theoretical Relevance:* For the AI research community, standard acronyms (e.g., "CNN", "SOTA") are intrinsic elements of this shared repertoire. Using full forms (e.g., "Convolutional Neural Network") in a high-context discussion violates the economy principle of the repertoire and signals "Out-group" status.
- *Mapping to Metrics:* This theory provides the sociological justification for our *Terminological Compactness Check*. The penalty for using full names or non-standard translations is not merely a stylistic preference but a validity check for community membership.

The Markedness Model Myers-Scotton (1993) proposed the Markedness Model, distinguishing between the *Unmarked Choice* (the expected, normative choice in a specific interactional type) and the *Marked Choice*.

- *Theoretical Relevance:* In general daily Chinese communication, code-mixing is often the *Marked Choice* (unexpected). However, our work posits that within the specific domain of Scientific Discourse, code-mixing becomes the *Unmarked Choice*. Conversely, strict monolingualism becomes *Marked* and pragmatically inappropriate.
- *Mapping to Metrics:* ChiEngMixBench fundamentally challenges the traditional NLP assumption that monolingual output is the "Gold Standard." By treating code-mixing as the normative (Unmarked) baseline for evaluation, we correct the misalignment in previous benchmarks.

B.2 Cognitive Mechanisms & Spontaneity

These theories offer explanations for the *Spontaneity* phenomena, particularly *Contextual Inertia* and the *Terminology Layering Strategy*.

Inhibitory Control Model (ICM) Green (1998) suggests that bilinguals share a common lexico-semantic system but must actively inhibit the non-target language to speak the target language. To switch languages, the activation level of a lemma in the non-target language must overcome this inhibition.

- *Theoretical Relevance:* This parallels the *Contextual Inertia* observed in our RQ1 results. The strong Chinese context acts as a powerful inhibitory signal for English tokens.
- *Mapping to Metrics:* Only terms with extremely high **intrinsic activation preference**

(high activation strength due to domain specificity, i.e., Specialized Terms) can breakthrough this inhibition threshold to trigger a code-switch. This explains why general terms (low activation) remain suppressed by the Chinese context.

Catachrestic Code-Switching Linguistics distinguishes between *Metaphorical Switching* (for rhetorical effect) and *Catachrestic Switching*—switching driven by a lexical gap in the matrix language, where no precise equivalent exists.

- *Theoretical Relevance:* The "API Anomaly" observed in our experiments (where the term "API" exhibits an exceptionally high Spontaneity score) is a classic example of catachrestic switching. Chinese lacks a concise equivalent for "Application Programming Interface" that retains the same functional precision.
- *Mapping to Metrics:* This confirms that the model's preference for code-mixing is not random but follows an efficiency-driven mechanism to fill lexical gaps, adhering to *Zipf's Principle of Least Effort*.

C Experimental Setup Details

We selected 12 models (including 3 high-performance closed-source models) based on three orthogonal controlled variables to ensure comprehensive coverage:

1. **Scaling Family (Scale Control Group):** We selected the Qwen2.5 family (**0.5B, 1.5B, 7B**) to investigate whether semantic discrimination capabilities exhibit emergent characteristics with growth in parameter size.
2. **Language Bias Group:**
   - **Qwen2.5-7B-Instruct** (Hui et al., 2024) and **Yi-1.5-9B-Chat**: High proportion of Chinese training data, representing "Chinese-centric" models.
   - **Llama-3-8B-Instruct** (Dubey et al., 2024), **Mistral-7B-v0.3**, and **Gemma-2-9B-It**: English-native models, representing "English-centric" thinking and training.
   - **TinyLlama-1.1B** (Zhang et al., 2024): Trained solely on 3T English tokens with no Chinese instruction tuning, acting as an extreme control for statistical *Inertia*.
3. **Architectural & Training Group:**
   - **DeepSeek-V3** (DeepSeek-AI, 2024): A Mixture-of-Experts (MoE) model representing the current state-of-the-art in open-weight reasoning.
   - **GPT-4o & Claude-3.5-Sonnet**: High-performance closed-source models used as the "Expert Gold Standard" for Naturalness.
   - **Alignment Control:** Qwen2.5-7B-Base vs. Instruct to verify the impact of RLHF on spontaneity.

C.1 Prompt Templates

For Naturalness evaluation, we used the following system prompt to simulate expert contexts:

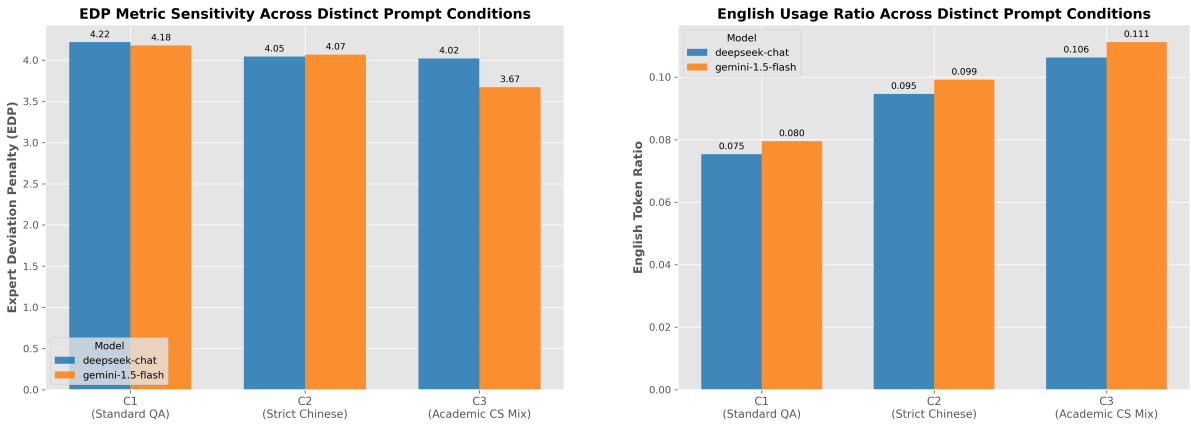
> *"You are an AI assistant helping a computer science researcher. Please answer the following technical question concisely, maintaining a professional academic tone. Do not over-explain basic concepts."*

Rationale for English System Prompt While the user queries are in Chinese, we intentionally utilize an English system prompt to define the model's persona. In the computer science community, English is the *lingua franca* of research. Utilizing an English-language persona effectively triggers the model's "Expert Distribution" learned during pre-training on high-quality academic corpora (e.g., ArXiv, StackOverflow), which is less likely to be contaminated by colloquial or generic Chinese conversational patterns. This setup tests the model's ability to maintain a professional expert identity across languages.

We curated 32 specific questions covering CV, NLP, and RL domains. Example: *"Explain the difference between Batch Norm and Layer Norm."*

D Ablation Study Details

We conducted an ablation study to isolate the impact of different prompting strategies (C1, C2, C3) on model spontaneity and naturalness. The results demonstrate that while explicit instructions can force code-mixing, they often degrade terminological precision in mid-sized models, whereas expert-styled cues (C2) better preserve the natural expert distribution.



(a) EDP Metric Sensitivity

(b) English Usage Ratio

Figure 6: **Ablation Study Results (C1 vs. C2 vs. C3).** Comparison of Spontaneity and Naturalness across prompting constraints. Expert-styled prompts (C2) achieve the best trade-off between mixing drive and linguistic naturalness, while C3 forces excessive mixing that degrades terminological precision.

D.1 Human Evaluation Rubric

The expert review panel followed a standardized 5-point Likert scale to assess the "Naturalness" of generated responses. The rubric is defined as follows:

- **Score 5 (Expert-Native):** Perfect cognitive alignment. Terminology is precise, code-switching occurs only at high-entropy positions, and the tone is indistinguishable from a senior domain researcher.
- **Score 4 (High Quality):** Fluent and accurate. Minor stylistic stiffness (e.g., slightly overuse of full names) but no pragmatic errors.
- **Score 3 (Acceptable):** Grammatically correct but functionally unnatural. May contain "Translationese" artifacts or unnecessary explanations.
- **Score 2 (Weak):** Noticeable issues. Contains "Word Salad" (fragmented switching) or excessive English ratio that breaks the Chinese syntax flow.
- **Score 1 (Collapse):** Severe failure. Complete monolingual collapse (ignoring the language constraint) or incoherent gibberish.

E Extended Spontaneity Analysis (RQ1)

This section contains the detailed distribution charts and the full analysis text that was summarized in the main paper.

E.1 Contextual Inertia Dominance

Figure 7 illustrates the CSG gap between General and Specialized terms across different model scales.

Analysis: The experimental results reveal the absolute dominance of contextual inertia in text generation. Data shows that regardless of model scale (from 1.5B to 8B), EPR is significantly suppressed to a low level. Even for the English-dominant Llama-3-8B, under the strong constraint of Chinese prefixes, its EPR fails to exceed 20%. This phenomenon powerfully demonstrates that maintaining consistency with the context distribution is the model's primary principle when predicting the next token. This statistical constraint from context significantly outweighs the model's intrinsic cross-lingual transfer capability or potential English preference.

E.2 Systematic Shift in Subconscious

Figure 8 illustrates the distribution of CSG at the Logits layer.

Analysis: Although the mean CSG for most terms remains deep in the negative interval (indicating final output is still Chinese), we observed a stable *Systematic Shift*. Almost all models show a consistent rightward shift in the CSG distribution for Specialized Terms relative to General Terms. This shift is numerically stable (Lift $\approx +0.04 - +0.07$) and transcends architectural differences (Qwen vs. Llama). This reveals the "Entropy-Driven Mechanism" behind code-mixing: although high-information-density specialized terms are often

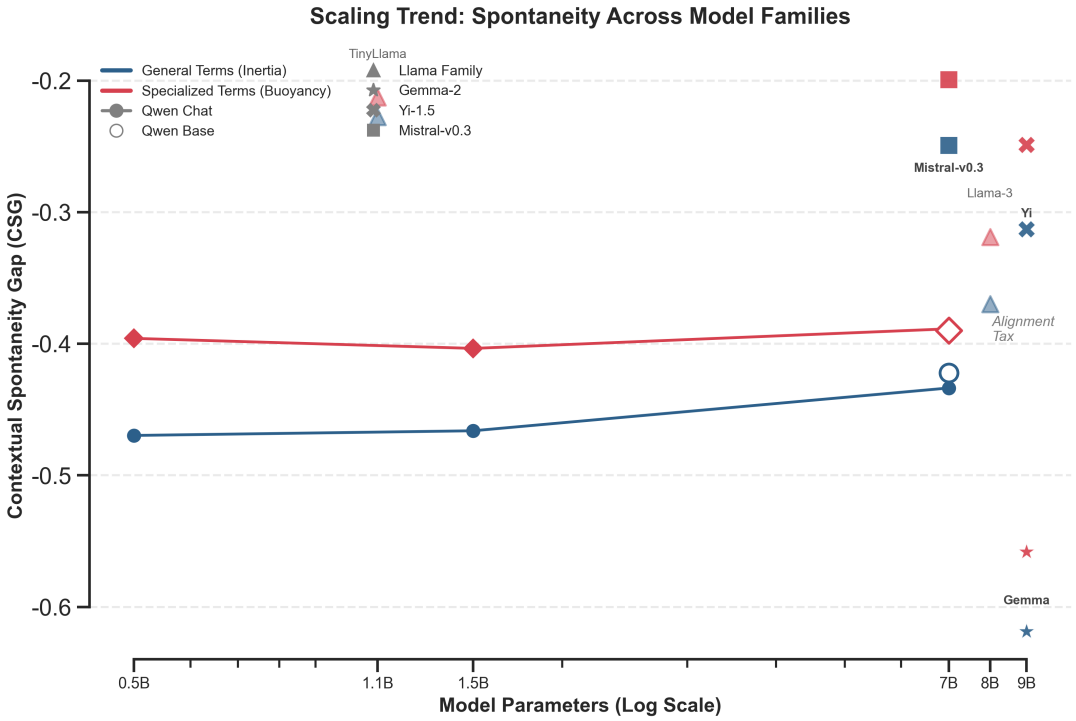


Figure 7: **Scaling Trend of Spontaneity Across Model Families.** Solid lines represent General Terms (Blue, Inertia) vs. Specialized Terms (Red, **Activation Preference**). While Contextual Inertia keeps all values negative, *Specialized Terms* consistently show a higher CSG (Lift) across scales. Note the "Alignment Tax" where Instruct models often drop below Base models.

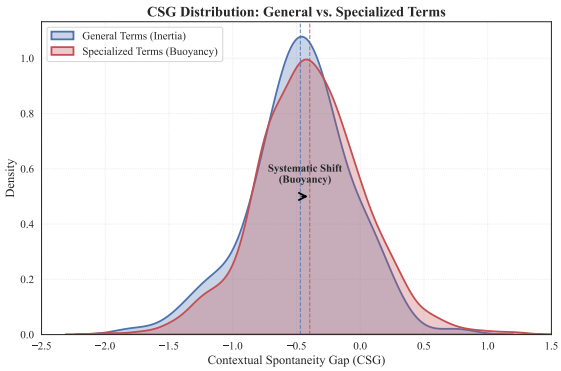


Figure 8: **Systematic Shift in CSG Distribution.** Specialized terms (Red) exhibit a consistent rightward shift compared to General terms (Blue), statistically confirming the **intrinsic activation preference** effect.

insufficient to completely reverse the strong inhibition of the Chinese context (i.e., failing to make CSG > 0), they objectively and significantly weaken the impact of contextual inertia. This suggests that the model's handling of code-mixing is not a simple binary decision (Switch or Not), but a refined gradient adjustment within the latent probability space based on term information content.

F Qualitative Analysis & Case Studies

To provide concrete intuition into the model behaviors and our data format, we present raw samples from our evaluation pipeline, complemented by a visual illustration of the decision process.

F.1 Naturalness: Expert Response Samples

Table 3 showcases the responses generated by *DeepSeek-V3* (one of the top-performing models in our benchmark) under the expert simulation prompt. The questions cover complex concepts (e.g., BERT, CoT, NER), requiring the model to correctly interweave specific English terminology within Chinese syntax.

F.2 Spontaneity: Logits & CSG Analysis

Table 4 displays the raw scoring data for *Qwen-7B* on Minimal Contrastive Pairs (MCP).

- *Score A:* English term log-likelihood $\mathcal{S}(t_{en})$.
- *Score B:* Chinese equivalent log-likelihood $\mathcal{S}(t_{zh})$.
- *CSG:* The calculated gap $(\mathcal{S}_{en} - \mathcal{S}_{zh})$.

We observe that General Terms (e.g., "zero") often exhibit a larger negative gap (Stronger Inertia) compared to Specialized Terms (e.g., "visual

ID	*Expert Query (Prompt)*	*Model Response (DeepSeek-V3)*
3	请解释BERT 预训练阶段MLM 任务的原理。	BERT 的*MLM* 任务在预训练时随机遮蔽输入序列中的部分*token*，模型通过双向上下文预测这些被遮蔽的原始*token*，从而学习语言的深层表示。
28	在复杂推理任务中，CoT 提示策略的核心优势是什么?	*CoT* 的核心优势是通过中间推理步骤引导模型分解复杂问题，模拟人类逐步思考过程，从而显著提升逻辑推理和数学计算等任务的准确性。
29	在IE 流水线中，NER 的准确率对后续Relation Extraction 有什么影响?	*NER* 的准确率直接影响关系抽取的性能，错误实体识别会传递至关系分类，导致召回率和精确度下降。

Table 3: **Real-world Generation Samples.** Responses from DeepSeek-V3 demonstrate the "Terminology Layering Strategy," where acronyms (MLM, CoT, NER) are retained in English while the matrix language remains Chinese. Data sourced from raw evaluation logs.

feature"), consistent with our **intrinsic activation preference** hypothesis.

F.3 Raw Evaluation Data Format

Figure 9 presents the raw JSON data format utilized in our automated evaluation pipeline. As illustrated, each entry records the specific prompt (expert query) and the model's corresponding generated response. This structured format facilitates the subsequent calculation of the Expert Deviation Penalty (EDP) by allowing automated scripts to parse and analyze the interweaving of English terminology within the Chinese matrix language.

Term Pair (t_{en} / t_{zh})	**Type**	*Score A* (S_{en})	*Score B* (S_{zh})	*CSG*
zero / 零	General	-3.48	-3.01	-0.47
visual feature / 视觉特征	Specialized	-3.29	-2.92	*-0.37*
vocab / 词表	Specialized	-3.38	-2.77	-0.61
ADALINE / 自适应线性神经元	Specialized	-2.75	-2.33	-0.42

Table 4: **Raw Spontaneity Scores (Qwen-7B).** Comparing Score A and Score B reveals the model's latent preference. Note that while all CSG values are negative (due to context), Specialized terms like "visual feature" show a relatively higher CSG compared to general terms like "zero", illustrating the "Lift" effect.


```
{
    "id": 1,
    "term": "Unknown",
    "prompt": "请从技术角度定义一下 AI 的核心目标及其在现代社会的应用。",
    "response": "AI's core objective is to develop intelligent systems that can perform tasks that typically require human intelligence, such as learning, reasoning, problem-solving, and perception. Its applications in modern society range from speech recognition and natural language processing in virtual assistants, to image recognition in self-driving cars, to personalized recommendations in e-commerce platforms."
  },
  {
    "id": 2,
    "term": "Unknown",
    "prompt": "相比于目前的 Narrow AI，实现 AGI 的主要技术瓶颈在哪里？",
    "response": "主要技术瓶颈在于实现人类水平的泛化能力和理解能力，以及实现自我学习和反思能力。"
  },
  {
    "id": 3,
    "term": "Unknown",
    "prompt": "请解释 BERT 预训练阶段 MLM 任务的原理。",
    "response": "BERT (Bidirectional Encoder Representations from Transformers) 的预训练阶段中，Masked Language Modeling (MLM) 任务旨在学习单词在句子中的语境，通过随机掩盖文本中的一些单词，然后让模型预测掩盖单词的原始单词。这个过程帮助 BERT 学习语言模型，并捕捉到句子中单词之间的关系。"
  },
```


Figure 9: **Raw Output Samples from the Evaluation Pipeline.** The JSON snippet displays structured prompt-response pairs used in our naturalness assessment. It demonstrates the specific inputs (e.g., queries about AGI and BERT) and the raw generated text before EDP scoring.