The Flywheel of Invisible Dialect Loss: A New Mission for NLP

Anonymous ACL submission

Abstract

Large language models (LLMs) are becoming everyday conversational interfaces. This raise an urgent issue for dialects and minority varieties with weakening intergenerational transmission: will LLMs support preservation, or accelerate their disappearance? Using Sinitic varieties as a testbed, we argue that current LLM ecosystems instantiate a *flywheel of invisible dialect loss*: Mandarin-centric defaults shift dialect interaction toward Mandarin; model-generated pseudo-dialect can re-enter future corpora as synthetic contamination; and benchmark erasure leaves these regressions unmeasured. We ground this mechanism in two empirical probes of Qwen 2.5, GLM-4, and DeepSeek-V3 on authentic Wikipedia text in Mandarin, Cantonese, Wu, and Hakka. Cantonese and Hakka show consistent tokenizer fragmentation relative to Mandarin, while Wu shows near-parity—an apparent fairness success that instead reflects advanced Mandarinization, where standard disparity metrics can no longer detect the variety as distinct. In zero-shot continuation, models frequently shift dialect inputs into Mandarin, and Hakka is sometimes mapped laterally into Cantonese. We propose a three-pillar research agenda: dialect-aware evaluation, provenance-governed data, and dialect-preserving model design. The central lesson is that dialect loss in LLM ecosystems is not only due to data scarcity, but also a self-destructive feedback loop.

1 Introduction

LLM-powered chatbots are rapidly becoming a default interface for written and spoken interaction. Users increasingly ask them to explain, translate, tutor, summarize, and converse; children and language learners may encounter them as everyday literacy partners. For dialects and minority varieties with weakening intergenerational transmission—a category spanning thousands of varieties from te reo Māori (Hinton, 2013) to Yiddish to Sinitic varieties spoken by hundreds of millions (Joshi et al., 2020)—this shift creates a critical issue for NLP: do current LLMs help preserve linguistic diversity, or do they normalize the dominant variety as the only practical language of computation?

This issue is especially urgent for Sinitic varieties. Chinese LLM development operates at large scale, yet most Chinese-language systems implicitly treat Standard Mandarin as "Chinese." The Sinitic family includes Mandarin and several major non-Mandarin varieties, including Wu, Cantonese, Min, Hakka, Gan, and Xiang. Many have large nominal speaker communities but declining childhood use and limited written infrastructure. Cantonese benefits from Jyutping, active online communities, and recent benchmark construction (Jiang et al., 2025a; Cheng et al., 2025; Jiang et al., 2025b), yet LLMs still often answer Cantonese input in Mandarin. Wu, including Shanghainese and related varieties, has a large speaker base but a written register that is often heavily Mandarinized. Hakka remains still less visible in mainstream Chinese NLP resources. These cases make Sinitic varieties a useful testbed for a broader problem: dialects can be erased not only because they are absent from data, but because model behavior, data pipelines, and benchmarks reinforce one another.

We call this mechanism the *flywheel of invisible dialect loss* with three turns. First, Mandarin-centric defaults fragment, correct, translate, or replace dialectal interaction. Second, degraded model outputs may re-enter web corpora as plausible pseudo-dialect, especially when dialects share Chinese characters with Mandarin. Third, standard benchmarks do not measure dialect preservation, so these regressions remain invisible to model developers and leaderboard users. The result is a feedback loop: dialectal capability degrades, degraded output contaminates future data, and the absence of evaluation allows the process to con-

tinue unpenalized.

Several components of this loop have been studied separately. Prior work documents Mandarin-centric behavior in Chinese LLMs (Wang et al., 2025), recursive degradation from model-generated data (Shumailov et al., 2024), and dialect gaps in evaluation (Jiang et al., 2025a; Cheng et al., 2025). Our contribution is to connect them as a single LLM-era mechanism for dialect loss. This differs from global taxonomies of language invisibility, such as the "Invisible Giants" category of Khanna and Li (2025), because our focus is not a static state of under-representation but a self-reinforcing process. It also extends earlier concerns about self-reinforcing NLP availability gaps (Hovy and Spruit, 2016) by analyzing recursive synthetic contamination and script-overlap ambiguity. Finally, Wang et al. (2025) show that Chinese-developed LLMs are not necessarily more inclusive of China's languages than Western systems, but they exclude Han dialects because of unstandardized writing. We address precisely this difficult case and fill the gap in the literature.

This paper makes three contributions. First, we formulate the flywheel of invisible dialect loss as a feedback mechanism linking Mandarin-centric defaults, recursive pseudo-dialect contamination, and benchmark erasure. Second, we introduce two lightweight empirical probes—tokenizer fragmentation and first-turn dialect preservation—and apply them to three widely used Chinese LLMs across Mandarin, Cantonese, Wu, and Hakka Wikipedia text. These probes reveal an intra-Sinitic asymmetry: Cantonese and Hakka remain visibly under-supported, while Wu becomes difficult to detect precisely because its written form has already been Mandarinized. Third, we propose a counter-agenda for NLP research organized around dialect-aware evaluation, provenance-governed data, and dialect-preserving model design.

2 The Flywheel of Invisible Loss

The flywheel of invisible loss consists of three mutually reinforcing components: (1) Mandarin-centric deployment defaults degrade dialect interaction, (2) degraded outputs re-enter future corpora as synthetic contamination, and (3) benchmark erasure makes the regression difficult to observe. We describe each component below, using Wu as the central case and Cantonese and Hakka as contrasts, then propose a falsifiable prediction.

2.1 Default behavior: fragmentation and Mandarin shift

Commonly used LLM tokenizers and pretraining corpora encode Mandarin as standard Chinese. Dialect texts may be incorrectly segmented, silently translated, or treated as material to be explained rather than continued. Petrov et al. (2023) show that tokenizer design introduces cross-linguistic unfairness before inference: semantically equivalent inputs can require different numbers of tokens, changing cost, latency, and effective context length.

Analysis 1: tokenizer fragmentation. Following Petrov et al. (2023), we measure tokenizer fragmentation on Wikipedia entries in four Sinitic varieties: Mandarin (`zh`), Cantonese (`zh-yue`), Wu (`wuu`), and Hakka (`hak`). The corpus contains introductory passages from parallel articles on seven topics, with one missing Hakka entry, yielding n=27 passages. We compare three LLMs commonly used in Chinese conversations—Qwen 2.5 (Qwen Team, 2024), GLM-4 (Team GLM, 2024), and DeepSeek-V3 (DeepSeek-AI, 2024)—and report tokens per CJK Unified Ideograph, normalized by each model's Mandarin baseline (Figure 1(a)).

Cantonese and Hakka require consistently more tokens than Mandarin, with mean normalized ratios of $1.19\times$ and $1.18\times$ across the three models. Wu, by contrast, shows near-parity with Mandarin, with a mean ratio of $1.03\times$. This is the *Wu paradox*: low measured disparity is not necessarily evidence of fairness. In our corpus, Wu Wikipedia text often follows Mandarin grammatical patterns with sporadic Wu particles. The text is sociolinguistically marked as Wu, but the usual tokenization disparity signal largely disappears. Standard tokenizer-fairness probes can therefore fail exactly when a dialect has already been absorbed into a Mandarinized written register.

Analysis 2: first-turn dialect preservation. Tokenizer parity does not imply interactive preservation. We therefore test whether models continue in the input variety when given a dialect passage with no instruction to translate or standardize. We submit the same n=27 passages as single-turn prompts to Qwen 2.5-7B-Instruct, GLM-4.6, and DeepSeek-V3, yielding 81 responses, including 60 responses to non-Mandarin dialect prompts. Each response is annotated for whether it preserves the input variety, silently shifts to Mandarin, translates or explains the variety, mixes varieties, or substitutes another

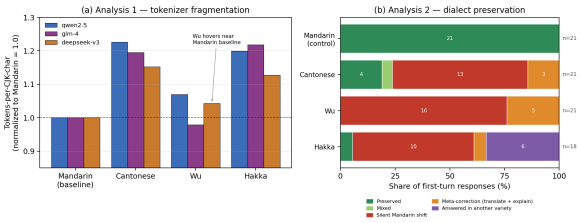

Figure 1: Empirical probes on Wikipedia entries. **(a) Analysis 1:** Cantonese and Hakka show higher tokenizer fragmentation than Mandarin, while Wu shows near-parity. **(b) Analysis 2:** Wu inputs show the strongest Mandarin shift; Hakka inputs are sometimes answered in Cantonese.

non-Mandarin variety. Full protocol details and representative outputs appear in Appendix A–B.

Figure 1(b) shows that dialect preservation is rare. All 21 Wu responses fail to preserve Wu: 16 silently shift to Mandarin and 5 translate or explain the input in Mandarin. Cantonese is somewhat more robust, but still shifts to Mandarin in 13 responses and is preserved in only 4. Hakka shows a different failure mode: in addition to 10 Mandarin shifts, 6 responses answer Hakka input in Cantonese, defaulting to a higher-resource non-Mandarin neighbor. These failures matter because users experience them as conversational defaults. A model that repeatedly answers Wu, Cantonese, or Hakka in Mandarin teaches the user that the dialect is not a viable medium for interaction.

2.2 Synthetic-data contamination

Recursive training on model-generated data can degrade model quality over time (Shumailov et al., 2024). For Sinitic varieties, this risk takes a particularly severe form. Unlike many dialect settings in which orthography sharply separates varieties, written Sinitic varieties share a large character inventory with Mandarin. This script overlap makes pseudo-dialect difficult to detect: Mandarin syntax with scattered dialect particles can look plausible, especially to automatic language-identification pipelines and non-expert curators.

This issue is exacerbated as pseudo-dialect is used in future training data. Cantonese LLM failures often involve Mandarinized syntax with Cantonese lexical substitutions (Jiang et al., 2025a; Liu et al., 2025). Similar outputs for Wu or Hakka may be scraped, deduplicated, and relabeled as dialect text. The Wu near-parity observed above is consistent with this danger: if online Wu is increasingly written in Mandarinized form, the boundary that tokenizers or classifiers might detect begins to dissolve. The variety remains named, but its computational representation becomes increasingly Mandarin-like.

Existing dialect-LLM resources vary in provenance discipline. Cantonese efforts, including Cantonese-LLM, YueTung/YueData, and related resources (Jiang et al., 2025b), increasingly incorporate human-curated corpora. Wu and Hakka resources remain more dependent on web scrapes and weakly marked machine-generated content. Because broad web pipelines such as Common Crawl's Chinese language identification are not designed to reliably distinguish Wu, Hakka, and Mandarin, dialect-labeled subsets may contain substantial Mandarin contamination. The harm is therefore recursive: today's degraded dialect outputs can become tomorrow's training examples.

2.3 Benchmark erasure

The third turn of the flywheel is evaluation. Major Chinese-language benchmarks and evaluation platforms, including C-Eval (Huang et al., 2023), CMMLU (Li et al., 2024), AGIEval (Zhong et al., 2024), OpenCompass (OpenCompass Contributors, 2023), CMRC (Cui et al., 2019), and CLUE (Xu et al., 2020), do not treat dialect preservation as a first-class evaluation axis. Speech evaluation is similarly dominated by Mandarin ASR and text-to-speech systems. Cantonese has only recently begun to move beyond this regime through HKCanto-Eval (Cheng et al., 2025), YueTung/YueData (Jiang et al., 2025b), and Cantonese-specific LLM evaluation (Jiang et al., 2025a). Zhang et al. (2025) cover several non-Han minority languages, including Tibetan, Uyghur, Kazakh, and Mongolian, but explicitly bracket Han dialects. For Wu and Hakka, the evaluation landscape remains thinner: prior Wu work is largely confined to narrow speech-engineering settings (Gao et al., 2025), while Hakka has ASR resources (Chen et al., 2025) but no general LLM-text benchmark.

This asymmetry is central to the flywheel. Without metrics, dialect loss is invisible. Without visibility, developers face little incentive to correct Mandarin-centric defaults. Benchmark erasure feeds back into model development: when dialect preservation is not measured, it is not optimized.

Falsifiable prediction. The flywheel predicts that Sinitic varieties without benchmark coverage will exhibit widening gaps across successive

model generations, whereas varieties that cross a critical benchmark threshold will begin to narrow those gaps. The 2024–2026 divergence between Cantonese, which has gained multiple evaluation efforts, and Wu/Hakka, which remain under-benchmarked, provides a concrete setting for testing this prediction in the near future.

3 A Three-Pillar Counter-Agenda

Breaking the flywheel requires intervention at each turn: evaluation must expose the failure, data governance must prevent recursive contamination, and model design must make dialect preservation an explicit objective.

3.1 Pillar I: dialect-aware evaluation

We propose a benchmark specification organized around four axes. (1) **tokenization fidelity** should report per-variety tokens per character relative to a Mandarin baseline, following Analysis 1 and the cross-linguistic calibration of Petrov et al. (2023). (2) **generation faithfulness** should test whether a model produces genuine dialectal output rather than Mandarin syntax with dialectal vocabulary substitution. This requires authentic reference corpora and should be distinguished from translation evaluation, even when translation resources such as FLORES-200 Cantonese subsets are available (Liu et al., 2025). (3) **correction and refusal rate** should measure how often dialect prompts are rewritten, translated, refused, or handled as metalinguistic objects. Lin et al. (2025) provide a related template for AAVE; Analysis 2 gives a minimal Sinitic instantiation. (4) **contamination detection** should test whether a held-out classifier can distinguish authentic vernacular from LLM-generated pseudo-dialect. This axis is difficult, but it directly targets the recursive fuel of the flywheel. Together, these axes define a dialect-preservation leaderboard rather than a single task.

3.2 Pillar II: provenance-governed data

The second pillar is a corpus manifest standard for Sinitic varieties, building on Data Statements (Bender and Friedman, 2018) and Datasheets (Gebru et al., 2021). At minimum, such manifests should record source, speaker-consent status, dialect-classification confidence, machine-generation status, and attribution chain. This specialization is necessary because script-overlapping varieties make contamination unusually subtle: an unflagged Mandarinized pseudo-dialect sample can be nearly indistinguishable from authentic vernacular in standard pipelines. Recording classifier provenance and uncertainty would give downstream curators a principled basis for filtering recursively contaminated data before training begins. It would also make negative evidence explicit. A corpus builder should be able to say not only that a passage is labeled Wu or Hakka, but who labeled it, whether the label reflects community authorship or automatic classification, whether the text was machine-generated, and how much Mandarinized structure it contains. Such manifests would not solve scarcity, but they would prevent scarcity from being addressed by silently importing contaminated data.

3.3 Pillar III: dialect-preserving model design

The third pillar concerns model design. Promising approaches include dialect-aware tokenization with variety-tagged subword units; contrastive Mandarin–dialect objectives during fine-tuning; variety routing in mixture-of-experts architectures, inspired by the language-routing mechanisms of NLLB (NLLB Team et al., 2022); and small, variety-specific "dialect tutor" models that can be incorporated into larger systems. None is likely to break the flywheel alone. Together, however, they make dialect preservation a first-class modeling objective rather than an accidental byproduct of multilingual scale. The goal is not to freeze dialects into a single standardized form; it is to prevent model defaults from collapsing living variation into the dominant written standard. For that reason, model improvements should be evaluated jointly with community-authored data and dialect-specific user studies, not only with automatic scores.

4 Conclusion

Dialect loss in LLM ecosystems is due to not only under-representation, but also a self-destructive feedback loop. In Sinitic varieties, Mandarin-centric defaults, pseudo-dialect contamination, and benchmark erasure reinforce one another, making dialect loss both accelerated and difficult to detect. Our tokenizer and first-turn preservation probes show how this mechanism appears across Cantonese, Wu, and Hakka, with Wu illustrating the especially dangerous case where low measured disparity may signal prior erasure rather than fairness. Breaking this flywheel requires evaluation, data governance, and model design to treat dialect preservation as a first-class NLP objective.

Limitations

Sinitic variety selection. We chose Wu mainly because its large nominal speaker base makes the invisibility of decline empirically demonstrable, rather than because it is the most endangered Sinitic variety. Pinghua, Hui, and smaller Min and Wu subvarieties may be more critical and remain unaddressed here. Future work should test whether the flywheel framework generalizes across these varieties.

Short-paper scope. Pillar III's model-design directions are an agenda, not a validated system. Each direction has analogues in low-resource language modeling, but none is demonstrated here as a stand-alone intervention for Sinitic dialect preservation.

Data, artifacts, and ethics. Probe data are public Wikipedia introductions (CC BY-SA 4.0), used only for measurement; we redistribute no derived corpus, and the encyclopedic text contains no targeted personal data. The studied models (Qwen 2.5, GLM-4, DeepSeek-V3) are used for analysis under their public release terms, consistent with their intended research use. One risk of this work is that a coarse corpus signal — such as Wu near-parity — could be misread as a claim about speaker competence rather than written register; we frame it as the latter throughout.

Generative AI usage. AI tools (Claude Code and ChatGPT) were used for paper editing and coding to generate the probes and figures. All scientific claims, analyses, and final draft were verified by the authors who are responsible for this entire work.

References

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Chih-Hsi Chen, Pei-Jun Liao, Chia-Hua Wu, Pang-Cheng Wu, and Hsin-Min Wang. 2025. The AS-SLAM system for Formosa speech recognition challenge 2025. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pages 504–511, National Taiwan University, Taipei City, Taiwan. Association for Computational Linguistics.

Tsz Chung Cheng, Chung Shing Cheng, Chaak-ming Lau, Eugene Lam, Wong Chun Yat, Hoi On Yu, and Cheuk Hei Chong. 2025. HKCanto-eval: A benchmark for evaluating Cantonese language understanding and cultural comprehension in LLMs. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 1–11, Vienna, Austria. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5883–5889, Hong Kong, China. Association for Computational Linguistics.

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

Zhipeng Gao, Akihiro Tamura, and Tsuneo Kato. 2025. Wenzhou dialect speech to Mandarin text conversion. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 36–43, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Leanne Hinton, editor. 2013. *Bringing Our Languages Home: Language Revitalization for Families*. Heyday Books, Berkeley, CA.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-Eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*, volume 36.

Jiyue Jiang, Pengan Chen, Liheng Chen, Sheng Wang, Qinghang Bao, Lingpeng Kong, Yu Li, and Chuan Wu. 2025a. How well do LLMs handle Cantonese? benchmarking Cantonese capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics.

Jiyue Jiang, Alfred Kar Yin Truong, Yanyu Chen, Qinghang Bao, Sheng Wang, Pengan Chen, Jiuming Wang, Lingpeng Kong, Yu Li, and Chuan Wu. 2025b. Developing and utilizing a large-scale Cantonese dataset for multi-tasking in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1924–1944, Suzhou, China. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.

Saurabh Khanna and Xinxu Li. 2025. Invisible languages of the LLM universe. *Preprint*, arXiv:2510.11557.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael Wooldridge, Janet B. Pierrehumbert, and Furu Wei. 2025. Assessing dialect fairness and robustness of large language models in reasoning tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Hannah Liu, Junghyun Min, Ethan Yue Heng Cheung, Shou-Yi Hung, Syed Mekael Wasti, Runtong Liang, Shiyao Qian, Shizhao Zheng, Elsie Chan, Ka Ieng Charlotte Lo, Wing Yu Yip, Richard Tzong-Han Tsai, and En-Shiun Annie Lee. 2025. SiniticMTError: A machine translation dataset with error annotations for sinitic languages. *Preprint*, arXiv:2509.20557.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

OpenCompass Contributors. 2023. OpenCompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems*, volume 36.

Qwen Team. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759.

Team GLM. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Andrea Wen-Yi Wang, Unso Eun Seo Jo, and David Mimno. 2025. Do chinese models speak chinese languages? *Preprint*, arXiv:2504.00289.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, and 13 others. 2020. CLUE: A Chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Chen Zhang, Mingxu Tao, Zhiyuan Liao, and Yansong Feng. 2025. MiLiC-eval: Benchmarking multilingual LLMs for China's minority languages. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11105–11120, Vienna, Austria. Association for Computational Linguistics.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

A Detailed Evaluation Setup

This appendix documents the full protocol for Analysis 2 (dialect preservation). Analysis 1 (tokenizer fragmentation) uses the same corpus and is described in §2.1.

Corpus and varieties. Both analyses draw on introductory paragraphs of Wikipedia articles across four Sinitic varieties: Mandarin (zh), Cantonese (zh-yue), Wu (wuu), and Hakka (hak). The corpus contains seven topics (Shanghai, Harbin, the Chinese language, Beijing, the Hakka people, Guangzhou, Chengdu), plus variety-specific topics where needed. The Hakka "Chinese language" article was absent at retrieval, giving n=27 source passages. Passages are community-authored vernacular rather than machine translations, which is essential: the Wu passages in particular reveal the Mandarinized written register central to our argument.

Prompt construction. Each passage is submitted verbatim as a single user turn, with no system prompt and no instruction—a zero-shot continuation setting. This isolates the model's default behavior: absent any instruction to translate or standardize, does it continue in the variety it was

given? Prompts are built by `build_prompts.py`. Mandarin passages serve as controls, where "continue in the input variety" and "answer in Mandarin" coincide.

Models and access. We query three Chinese-trained models: Qwen 2.5-7B-Instruct, GLM-4.6, and DeepSeek-V3 (Qwen Team, 2024; Team GLM, 2024; DeepSeek-AI, 2024). Decoding used `max_tokens`=200 and `temperature`=0.2.

Annotation rubric. Each of the n=81 responses received one of five labels:

- **preserved** — dominant morphosyntax and function words match the input variety rather than Standard Mandarin;
- **mixed** — substantial code-mixing without a single dominant variety (counted as not preserved);
- **mandarin_shift** — the model silently answers in Standard Mandarin, with no acknowledgement of the input variety;
- **meta_correction** — the model recognizes the variety but treats it as material to be translated into Mandarin, often with English explanation, rather than continued;
- **other** — the model answers in a different non-Mandarin variety; in every observed case, Cantonese substitutes for Hakka.

B Representative Responses by Model and Variety

The excerpts below (input prompt truncated; model output truncated and stripped of Markdown) illustrate the preservation outcomes from Appendix A. They show that the failure is not monolithic: silent shift, knowing translation, and lateral substitution are qualitatively distinct, and the same prompt can elicit opposite behavior from different models.

Control (preserved) — Mandarin, GLM-4.6. *Prompt:* 上海市，简称沪，别称申，是中华人民共和国直辖市..... *Response:* 这段关于上海市的介绍非常准确且精炼.....关于"申"与春申君：上海的母亲河黄浦江..... *Insight:* with Mandarin input, the model naturally continues in Mandarin; "preserved" and "Mandarin" coincide, anchoring the control condition.

Rare dialect preservation — Cantonese, GLM-4.6. *Prompt:* 上海中一城，亦一埠。而家中人民共和直市..... *Response:* 呀，你得好。上海唔止中最大都市，亦全球重要金融中心同埋航中心。位於江入海口..... *Insight:* the dominant non-Mandarin function words (, , , 同埋) are sustained—one of just five preserved dialect responses, all from GLM-4.6 on Cantonese.

Silent Mandarin shift — Cantonese, DeepSeek-V3. *Prompt:* (identical Cantonese passage as above) *Response:* 上海是中国的经济、金融、贸易和航运中心，拥有丰富的历史和现代都市风貌。以下是关于上海的一些关键信息..... *Insight:* the same prompt yields Standard Mandarin (是, 的) with no hint that the input was Cantonese—the most common failure mode and the one most invisible to users.

Meta-correction — Wu, GLM-4.6. *Prompt:* 上海（上海），「」，「申」.....个直市之一。地理方位勒拉中部..... *Response:* 您提供了一段用上海话（吴语）写的关于上海的介绍。以下是这段话的普通话和英语翻译..... *Insight:* the model correctly identifies Wu, then translates it into Mandarin and English and glosses its particles. Recognition is not preservation; the variety is handled as an object of study, not as a language to speak.

Lateral substitution (Hakka → Cantonese) — Hakka, GLM-4.6. *Prompt:* 客家，族一支，其基本特徵使用客家。主要分在、江西、福建..... *Response:* ，你得好。客家族一支好重要民系，有著特言、文化同埋俗.....世界各地都有足..... *Insight:* faced with Hakka, the model answers in Cantonese (, 同埋,), defaulting to the higher-resource non-Mandarin neighbor.

Total silent shift in the small model — Hakka, Qwen-7B. *Prompt:* 上海，中直市之一.....目前中展最亟最城市之一..... *Response:* 上海确实是中国的一个直辖市，位于中国东部的长江三角洲地区，面向东海，处于长江入海口南岸..... *Insight:* Qwen-7B shifted to Mandarin on all of its dialect prompts; the Hakka markers are dropped without comment. Where GLM-4.6 sometimes recognizes the variety, the smaller model erases silently.