Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on Human-Centred AI

Probing Cultural Awareness in LLMs: A Case Study of Cross-Culture Aesthetic Stylistics

Jiashuo Wang[1], **Fenggang Yu**[1], **Jian Wang**[1], **Chak Tou Leong**[1], **Xiaoyu Shen**[2], **Chunpu Xu**[1], **Jiawen Duan**[3], **Wenjie Li**[1], **Johan F. Hoorn**[1,4,5,6]
[1] Department of Computing, Hong Kong Polytechnic University
[2] Institute of Digital Twin, Eastern Institute of Technology, Ningbo
[3] Department of Language Science and Technology, Hong Kong Polytechnic University
[4] School of Design, Hong Kong Polytechnic University
[5] Research Institute for Quantum Technology, Hong Kong Polytechnic University
[6] Department of Communication Science, Vrije Universiteit Amsterdam

Abstract

Large Language Models (LLMs) are increasingly deployed in diverse cultural contexts, yet their ability to master aesthetic stylistics, i.e., the strategic use of language to evoke cultural resonance, remains underexplored. We curate **C^{4}STYLI**, a benchmark of highly stylized translated movie titles and advertising slogans from Hong Kong and the Chinese Mainland, to evaluate LLMs via the lens of behavioral recognition and productive competence. Extensive evaluations show that LLMs differ from humans in stylistic recognition, and this recognition ability varies across text domains. In addition, stylistic recognition and generation performance in LLMs are not consistently aligned. To further examine whether LLMs genuinely capture stylistic information in stylistic recognition, we conduct structural ablation with logistic regression probes. We find that, in the Hong Kong setting, stylistic recognition in LLMs relies primarily on surface-level linguistic information rather than stylistic structure. This suggests limited sensitivity to Hong Kong-specific stylistic structure. Our code and data are available at https://github.com/wangjs9/C4STYLI.

1 Introduction

Recent advances in Large Language Models (LLMs) have demonstrated remarkable capabilities in tasks such as question answering [Kim *et al.*, 2024], reasoning [Zhang *et al.*, 2024], and dialogue [Wang *et al.*, 2024b]. Consequently, studying their cultural awareness, the ability to recognize and appropriately respond to cultural nuances [Pawar *et al.*, 2025], has become critical for developing trustworthy, culturally aligned applications across diverse populations.

Culture encompasses various dimensions, ranging from tangible artifacts to intangible values. Among these, **aesthetic stylistics** stands as one of the most sophisticated yet under-explored facets in LLM research. It represents the strategic choice made among various grammatically correct and semantically equivalent ways of expressing the same content. [Hickey, 2014]. Unlike explicit facts or linguistic rules (Figure 1), we conceptualize aesthetic stylistics as *an implicit way that determines how to tailor the expression of specific content* to attract the reader's attention and provoke resonance without altering the core semantics [Riffaterre, 1978].

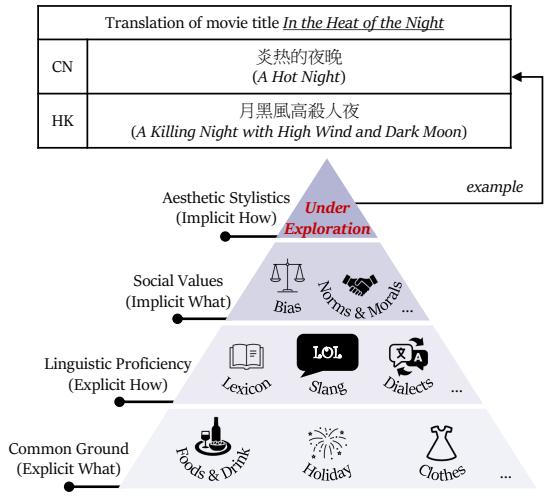


Figure 1: Hierarchy of cultural awareness in LLMs.

This work investigates cultural awareness in LLMs through aesthetic stylistics. We introduce **C^{4}STYLI** (*C*ross-*C*ulture *C*hinese-*C*hinese **Styli**stics), a curated benchmark including translated movie titles and advertising slogans in the Chinese Mainland (CN) and Hong Kong (HK). This dataset offers two primary advantages. (1) **Minimal Linguistic Interference**: By utilizing the mutual convertibility between Simplified and Traditional Chinese, we minimize surface-level linguistic variance, allowing for a more focused analysis of cultural aesthetic stylistics. (2) **High Stylistic Salience**: Unlike neutral prose, these texts are purposely engineered to maximize resonance and attention, providing a high-signal environment for observing aesthetic stylistics.

With this dataset, we analyze current popular LLMs' capacities in terms of their behavioral recognition, productive

C⁴Styli-T	*C⁴Styli-S*
Original Title: Sleepless in Seattle **Title (CN):** 西雅图夜未眠 (Sleepless Night in Seattle) **Title (HK):** 緣份的天空 (The Sky of Destiny) **Year:** 1993 **Plot Summary:** Sam Baldwin, a recently widowed Chicago architect, moves to Seattle with his eight-year-old son Jonah, to start a new life. A year later on Christmas Eve, Jonah calls in to a nationally syndicated radio talk show and persuades a reluctant Sam to go on the air to talk about how much he misses his wife, Maggie, and how he knew she was the one for him when he first took her hand ...	**Brand:** 公益廣告 (Public Service Announcement) **Product:** 交通安全宣導，不酒駕 (Road Safety Campaign; No Drink Driving) **Slogan:** 喝得醉醉 撞得碎碎 (Drink Drunk, Crash to Pieces) **Year:** 2024 **Region:** HK **Brand:** 公益广告 (Public Service Announcement) **Product:** 交通安全宣传，不酒驾 (Road Safety Campaign; No Drink Driving) **Slogan:** 喝酒不开车 开车不喝酒 (If you drink, don't drive. If you drive, don't drink.) **Year:** 2011 **Region:** CN

Figure 2: An instance from C⁴STYLI-T (left) and two instances from C⁴STYLI-S (right). English translations are provided for illustration.

competence, and internal representation. Our analysis yields three primary findings: **(1)** LLMs show clear differences from humans in recognizing cultural stylistics, and their performance varies substantially across different text domains. **(2)** We observe a decoupling between recognition and production: models that excel at identifying styles do not necessarily mirror that proficiency in generating them. **(3)** Through structural ablation, we reveal that LLMs' awareness of HK culture remains superficial. While models can "match" HK styles through sparse, high-intensity lexical anchors, they lack the integrated structural understanding observed in their representation of CN styles.

Our contributions are as follows: **(1) Benchmark:** We introduce **C⁴STYLI**, which mitigates surface linguistic interference via script normalization to focus on implicit aesthetic stylistics. **(2) Behavioral Gap:** We find LLM–human differences, domain variation, and a decoupling between stylistic recognition and production. **(3) Shallow Encoding of HK Stylistics:** We find that LLMs encode HK style via superficial lexical shortcuts instead of integrated structures.

2 Related Work

Cultural Awareness in LLMs. Culture is an expansive and multifaceted construct, necessitating a granular approach to its study within LLMs. To systematically examine cultural awareness, we organize its various dimensions into a hierarchical pyramid structure (Figure 1), reflecting a progression from explicit manifestations to implicit nuances. At the base of our framework lie the foundational layers, representing the explicit "what" and "how" of culture. *Common Ground* encompasses tangible cultural artifacts and factual concepts, such as traditional clothing [Zhou *et al.*, 2025b] and regional cuisine [Zhou *et al.*, 2025a]. *Linguistic Proficiency* focuses on the formal communicative conventions and regional codes, including specialized lexicons [Cao *et al.*, 2024], slang [Sun and Xu, 2022], and regional dialects [Faisal and Anastasopoulos, 2025]. The higher echelons of the pyramid transition into the implicit realm. *Social Values* represent the implicit "what", governing how language adheres to societal morals and ethics to ensure appropriateness within cultural norms. Prior work in this layer has focused on identifying cultural biases [Nangia *et al.*, 2020], aligning models with regional norms and morals [Li *et al.*, 2024], and evaluating cross-cultural ethical reasoning [Mohammadi *et al.*, 2025]. Finally, at the apex sits *Aesthetic Stylistics*, which we define as the implicit "how." This sophisticated layer dictates the artistic treatment of language required to achieve pragmatic effects and stimulate profound resonance within an audience.

Cantonese NLP. Cantonese (Yue), despite being used by a large and geographically dispersed speech community, remains comparatively underrepresented in current NLP research. Even for classical NLP tasks, such as sentiment classification [Xiang *et al.*, 2019], rumor detection [Chen *et al.*, 2020], and dialogue generation [Wang *et al.*, 2024a], the number of Cantonese-focused studies remains limited. Unlike simplified Chinese (Zh) and English (En), for which a wide range of mature processing tools are available (e.g., cntext, NLTK, SpaCy), Cantonese lacks comparable tool support, with only a small number of dedicated resources such as PyCantonese [Lee *et al.*, 2022]. This scarcity of linguistic infrastructure further constrains large-scale empirical research on Cantonese. In the era of LLMs, Cantonese-specific models are still rare. At present, several models released by SenseTime constitute the few publicly available non-commercial LLMs with explicit Cantonese support. Correspondingly, evaluation benchmarks for Cantonese remain limited. Earlier resources were primarily constructed from television programs for machine translation [Lee, 2011] or for linguistic annotation [Leung and Law, 2001]. More recently, a small number of benchmarks, such as Yue-GSM8K, Yue-MMLU, and Yue-ARC, have been proposed to evaluate the reasoning abilities of both Cantonese-specific and general-purpose LLMs [Jiang *et al.*, 2025].

3 Dataset: C⁴STYLI

3.1 Overview

C⁴STYLI consists of two subcomponents. **C⁴STYLI-T** contains Chinese movie titles in HK and CN, translated from English source titles. **C⁴STYLI-S** comprises advertisement slogans in HK and CN. Figure 2 shows instances. The dataset focuses on texts in domains that can provide classic instances for studying language style [Wales, 2014]. Accordingly, C⁴STYLI serves as a diagnostic testbed for analyzing the extent to which cultural stylistic preferences in HK and CN are distinguishable at the level of highly stylized short texts.

Subsplit	# Instances	# HK : # CN	Avg. Len	Year Range
C^{4}STYLI-T	1,530	1,530 : 1,530	5.10	1907-2027
C^{4}STYLI-S	2,837	1,976 : 861	47.50	1953-2025

Table 1: Statistics of the C^{4}STYLI dataset.

Title	**CN**	人(Man), 总动员(Mobilization), 美国(America), 世界(World), 特工(Secret Service), 游戏(Game), 爱(Love), 疯狂(Crazy), 之战(War), 传奇(Legend), 电影(Movie), 王牌(Ace), 杀手(Killer), 最后(Last)
	HK	電影(Movie), 人(Man), 奇兵(Wonder Soldiers), 世紀(Century), 王(King), 新(New), 傳奇(Legend), 戰士(Warrior), 特攻(Special Attack), 神奇(Magical), 追擊(Chase), 俠(Hero), 星球大戰(Star Wars), 無敵(Invincible)
Slogan	**CN**	中国(China), 更(More), 健康(Health), 科技(Technology), 世界(World), 生活(Life), 爱(Love), 妈妈(Mom), 核心(Core), 营养(Nutrition), 吃(Eat), 传承(Heritage), 典范(Model), 幸福(Happiness), 掌握(Master), 品质(Quality), 人生(Life), 看(See), 美好(Beautiful)
	HK	全新(Brand New), 俾(Give), 香港(Hong Kong), 會(Will), 更(More), 想(Want), 健康(Health), 人(People), 新(New), 快(Fast), 仲有(Also/Still Have), 令(Make), 食(Eat), 最(Most), 買(Buy), 嘢(Thing), 使(Use/Need), 揀(Choose), 再(Again), 好似(Like/Seem), 送(Give/Deliver), 真係(Really), 系列(Series)

Figure 3: Comparison of the most frequent words used across different text domains and regions.

3.2 Collection and Annotation

C^{4}STYLI-T. Movie titles were collected from TMDb (The Movie Database) and Wikipedia. Using a web-scraping framework, we extracted the titles and plot summaries in English, along with their officially released Chinese titles for both HK and CN. Instances were manually filtered if they met any of the following criteria: (1) the Chinese title matches the original English title, (2) the plot summary is missing, or (3) the Chinese titles in CN and HK are identical (ignoring simplified vs. traditional script). Each instance includes the English title, the movie release year, the plot summary in English, and translated titles in HK and CN.

C^{4}STYLI-S. Advertising slogans were collected from company websites, social media platforms, Bilibili, and YouTube. For content from company websites and social media, we manually extracted the brand, product, year, region, and slogan information. For Bilibili and YouTube, we first identified seed videos containing the keywords (经典广告 or 經典廣告) in their titles and then collected their related videos, including recommended videos and videos in the same default playlist. We retained only videos whose titles contained the brand name. Slogans were primarily extracted from available subtitles; if subtitles were missing or incomplete, we used qwen3-asr-flash, a speech-to-text system supporting both Mandarin and Cantonese, to generate them. All slogans were manually verified to ensure quality. Each instance consists of the brand name, product information, the release year, the region, and the slogan.

Table 1 summarizes the overall statistics of C^{4}STYLI, including the instance number, average length of target texts (i.e., translated movie titles and slogans), and temporal coverage. It primarily reflects the differences across text domains.

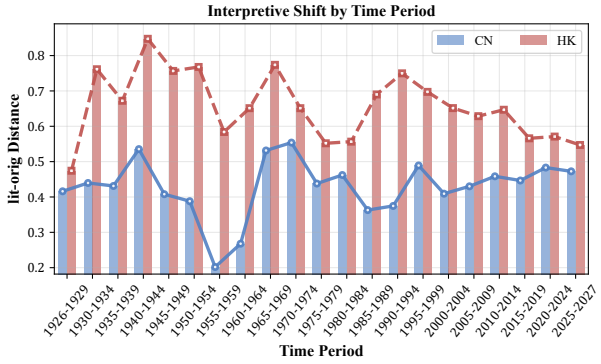
3.3 Linguistic Features

To characterize regional linguistic variation, we analyze differences in language use and lexical preferences between CN and HK. For translated movie titles, we examine their stylistic properties by comparing them with their original English counterparts. Specifically, we quantify source–target divergence by computing $1 - \cos(\mathbf{e}_{bt}, \mathbf{e}_{src})$, where $\mathbf{e}_{bt}$ denotes the embedding of the back-translated Chinese title and $\mathbf{e}_{src}$ denotes the embedding of the original English title, both represented in English embedding space. Our results, HK: 0.617 vs. CN: 0.447, show that **titles released in HK exhibit a larger source–target divergence than those in CN, indicating a greater degree of deviation from the source titles**. A top frequency-based lexical analysis further reveals systematic stylistic differences between the two regions, as shown in Figure 3. CN titles tend to preserve stronger formal equivalence to the source text, frequently relying on literal renderings of high-frequency content nouns (e.g., "Man," "American," "World"). In contrast, HK titles more often employ expressive reformulations by introducing affectively salient modifiers. For example, the film *King Arthur* is rendered in HK as 王者無敵 (*lit.* King Invincible), where the modifier 無敵 (Invincible) enhances emotional salience beyond the source title. **These observations point to distinct stylistic preferences rather than rigid translation conventions across cultures**. To further examine affective differences, we measure the change in sentiment expression between a translated title and its original English version, defined as: $|\text{senti}_{\text{bt}}| - |\text{senti}_{\text{src}}|$, where senti denotes the sentiment polarity score computed using the VADER analyzer. We find that HK titles exhibit a larger increase in absolute sentiment strength than CN titles (HK: 0.111 vs. CN: 0.052), suggesting that **HK translations more frequently amplify affective expression relative to the source text than their CN counterparts**.

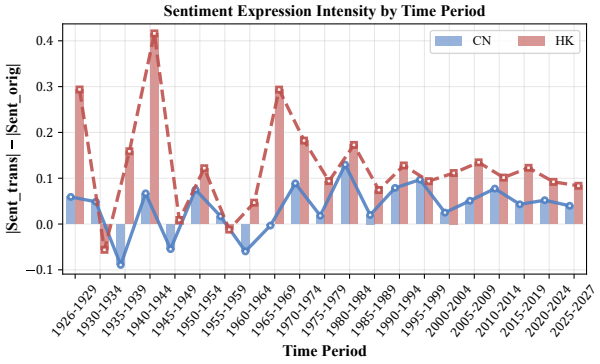
For advertising slogans, we first examine the prevalence of English usage. Specifically, we compute the proportion of English tokens in slogans from each region. English accounts for 12.36% of tokens in HK slogans, compared to only 1.47% in CN slogans, reflecting **HK's longstanding bilingual sociolinguistic context**. We further analyze syntactic and lexical stylistic differences by computing the modifier ratio, defined as (# adjectives + # adverbs) / # nouns, for slogans in both regions. The modifier ratio is substantially higher in HK (51.79%) than in CN (34.67%), indicating systematic differences in stylistic composition. Specifically, **CN slogans tend to emphasize substantive attributes through noun-heavy constructions, whereas HK slogans more frequently rely on predicate-heavy formulations**. These tendencies are also reflected in the most frequent lexical items shown in Figure 3. CN slogans often foreground nouns such as 科技 (Technology) and 传承 (Heritage), contributing to a more formal and descriptive register. In contrast, HK slogans favor adjectives and adverbs (e.g., 快 [Fast], 真係 [Really]) to achieve greater conversational immediacy.

3.4 Temporal Characteristics

Having characterized aggregate stylistic differences between cultures, we next examine whether these features remain temporally stable within each culture. Figure 4 illustrates the temporal variation of source–target divergence and affective change in translated movie titles across cultures. Across all time periods, titles released in HK consistently exhibit larger

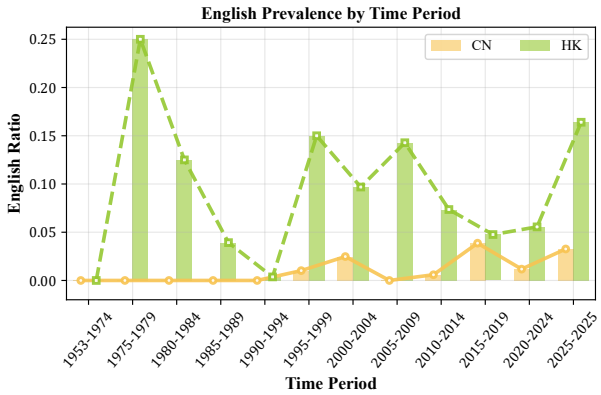


(a) Semantic divergence between translated titles, measured after back-translation into English, and original titles.

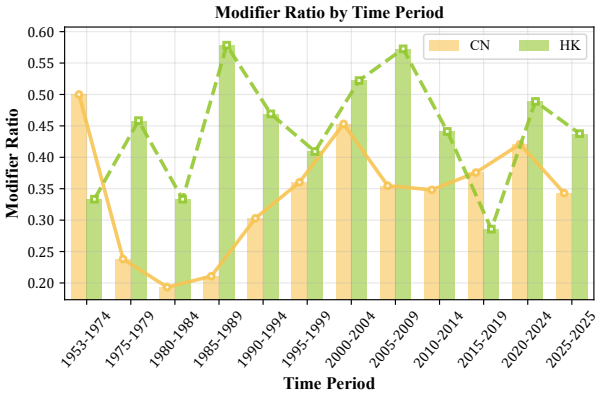


(b) Divergence in affective content between translated titles and original titles, quantified as the difference in sentiment scores.

Figure 4: Temporal variation in source–target divergence of C^{4}STYLI-T in CN and HK.



(a) Prevalence rate of English tokens.



(b) Modifier-to-noun ratio, defined as $\frac{(\#\text{adjectives}+\#\text{adverbs})}{\#\text{nouns}}$.

Figure 5: Temporal variation of stylistic features in C^{4}STYLI-S from CN and HK.

semantic divergence from their original English counterparts than those released in CN. A similar temporal pattern is observed for affective divergence. As shown in Figure 4(b), HK titles generally demonstrate a larger increase in absolute sentiment strength relative to the original English titles, whereas CN titles exhibit smaller and more conservative changes. Notably, although this cultural contrast remains robust throughout the time span, the magnitude of the gap shows signs of narrowing in more recent periods.

Figure 5 presents the temporal variation of stylistic features in advertising slogans, focusing on English usage and syntactic composition. Across all time periods, HK slogans exhibit substantially higher English token ratios than those from CN. While English prevalence in HK slogans fluctuates over time, the overall separation between the two cultures remains evident throughout the entire temporal span. In terms of syntactic composition, the modifier ratio also displays consistent cultural differences over time.

From the results, *the cultural differences are temporally ordered rather than merely average-based*: across all observed time periods, the maximum value observed for CN remains below the minimum value observed for HK. In other words, these differences characterize distributional tendencies at the population level, rather than providing sufficient cues for unambiguous culture identification at the level of individual texts or for explicit cultural attribution.

4 Overall Setup

Assessed LLMs. For behavioral recognition and productive competence, we evaluate the cultural awareness of 18 LLMs, spanning six major model families: (1) OpenAI series, (2) Meta-Llama series, (3) Google Gemma series, (4) DeepSeek series, (5) Qwen series, and (6) SenseNova series. These models are categorized according to their relatively stronger languages, i.e., En, Zh, and Yue.

Dataset Splitting. We split the dataset into three subsets. For C^{4}STYLI-T, we randomly selected 1010 instances for training and 520 for testing. For C^{4}STYLI-S, we randomly selected 2,287 instances for training and 550 for testing.

5 Can LLMs Discriminate Cultural Stylistics?

Method. To investigate whether LLMs can identify culturally grounded stylistic distinctions, we formulate a classification task that evaluates whether LLMs can discriminate cultural stylistics by predicting the cultural origin (HK vs. CN)

	Model	C^{4}STYLI-T				C^{4}STYLI-S			
		Accuracy	Precision	Recall (macro / HK / CN)	F1	Accuracy	Precision	Recall (macro / HK / CN)	F1
En	o4-mini	**0.701**	0.701	**0.701** / 0.681 / 0.720	**0.701**	**0.788**	**0.836**	0.726 / 0.479 / **0.973**	0.740
	GPT-4.1	0.685	**0.705**	0.685 / 0.529 / 0.841	0.677	0.819	0.828	**0.797 / 0.675** / 0.920	**0.806**
	Llama-3-8B-Instruct	0.535	0.546	0.540 / 0.729 / 0.351	0.520	0.687	0.748	0.611 / 0.256 / 0.966	0.590
	Llama-3-70B-Instruct	0.624	0.637	0.627 / **0.765** / 0.490	0.618	0.699	0.688	0.675 / 0.548 / 0.802	0.678
	gemma-3-4b-it	0.558	0.571	0.552 / 0.293 / 0.810	0.522	0.691	0.685	0.691 / 0.693 / 0.690	0.686
	gemma-3-12b-it	0.547	0.713	0.537 / 0.083 / **0.991**	0.422	0.746	0.782	0.701 / 0.459 / 0.943	0.705
	gemma-3-27b-it	0.604	0.626	0.608 / 0.800 / 0.416	0.591	0.779	0.797	0.746 / 0.567 / 0.925	0.754
Zh	DeepSeek-V3.2	0.700	0.719	0.700 / 0.553 / **0.846**	0.693	0.824	**0.866**	0.775 / 0.571 / 0.979	0.793
	DeepSeek-V3.2 (Thinking)	0.745	0.756	0.745 / 0.847 / 0.642	0.742	0.754	0.764	0.721 / 0.542 / 0.900	0.727
	Qwen2.5-7B-Instruct	0.598	0.618	0.593 / 0.368 / 0.819	0.574	0.761	0.760	0.737 / 0.608 / 0.865	0.743
	Qwen2.5-14B-Instruct	0.648	0.651	0.650 / 0.705 / 0.594	0.648	0.804	0.809	0.783 / 0.663 / 0.902	0.790
	Qwen2.5-32B-Instruct	0.686	0.692	0.688 / 0.769 / 0.606	0.685	**0.831**	0.829	**0.819 / 0.751** / 0.887	**0.823**
	Qwen2.5-72B-Instruct	**0.799**	**0.814**	**0.799 / 0.910** / 0.687	**0.796**	0.810	0.820	0.785 / 0.651 / **0.919**	0.794
	Qwen3-8B	0.661	0.673	0.663 / 0.785 / 0.542	0.657	0.707	0.699	0.682 / 0.546 / 0.818	0.686
	Qwen3-14B	0.685	0.686	0.685 / 0.712 / 0.659	0.685	0.733	0.737	0.702 / 0.527 / 0.876	0.707
	Qwen3-32B	0.680	0.685	0.680 / 0.767 / 0.593	0.677	0.706	0.700	0.678 / 0.522 / 0.833	0.681
Yue	SenseChat-5-Cantonese	0.623	0.631	0.623 / **0.752** / 0.494	0.616	0.586	**0.775**	0.582 / 0.164 / **1.000**	0.495
	SenseNova-V6-5-Turbo	**0.652**	**0.654**	**0.651** / 0.592 / **0.710**	**0.650**	**0.762**	0.761	**0.715 / 0.528** / 0.901	**0.725**
	Human Baseline	**0.897**	**0.897**	**0.897 / 0.901 / 0.893**	**0.897**	**0.933**	**0.930**	**0.935 / 0.948 / 0.922**	**0.932**

Table 2: LLM performance in cultural stylistic identification across C^{4}STYLI-T and C^{4}STYLI-S.

of a given text, conditioned on relevant contextual information. Specifically, for translated movie titles, we provide the original title and plot summary as context; for advertising slogans, we supply the brand name, product information, and year of release. We examine model behavior under different prompting conditions by varying (i) the language variety used in the prompt (Cantonese vs. Mandarin) and (ii) the presence or absence of in-context exemplars.

Evaluation. The performance is evaluated by standard classification metrics, including **accuracy**, **precision**, **recall**, and **F1 score**. All metrics are reported in their macro-averaged form to account for potential class imbalance. In addition to macro recall, we separately report recall scores by treating HK and CN as the positive class, respectively, to assess potential asymmetries in model sensitivity across cultures.

LLM Overall Performance. Table 2 summarizes LLM performance on the cultural stylistic identification task. Each model is evaluated using four prompt variants, with the reported results averaged across variants. A human baseline is established by asking two human participants to identify the cultural style class of each text; inter-annotator agreement reaches a Krippendorf's α of 0.613, indicating substantial agreement. Overall, **LLM performance remains substantially below the human baseline, revealing a persistent gap in cultural stylistic recognition**. Beyond this overall performance difference, the human–model gap is further evidenced along two additional dimensions. First, **LLMs exhibit strong sensitivity to text type**. Advertising slogans are consistently easier to identify than translated movie titles, as reflected by higher scores across all evaluation metrics. We attribute this discrepancy primarily to the fact that movie titles are substantially shorter than advertising slogans, offering fewer tokens and less contextual redundancy, which makes cultural style inference more challenging for LLMs. Advertising slogans, by comparison, provide richer contextual information that facilitates more reliable identification of cultural cues. Human participants, however, can reliably recognize cultural stylistic cues even from very short texts. Second, **LLMs display systematic, text-type-dependent biases in label assignment**. For movie titles, LLMs tend to misclassify CN-labeled instances as HK, whereas the opposite pattern is observed for advertising slogans. This asymmetry is reflected in recall scores: recall is higher when HK is the primary label in the movie title domain, but this pattern reverses in the slogan domain. Together, these results indicate that LLMs' cultural stylistic biases are not uniform, but vary substantially with the textual genre and informational structure. Regarding training language exposure, **models primarily trained on Chinese corpora generally outperform those trained mainly on English**, suggesting that access to Chinese-language data is beneficial for recognizing Chinese cultural styles. However, this advantage does not straightforwardly extend to Cantonese. Models trained on more Cantonese data (i.e., SenseNova models) do not consistently yield superior performance on cultural stylistic identification. This limitation may stem from several factors, including the relatively limited scale of available Cantonese corpora, the divergence between written and spoken Cantonese, and a higher level of annotation noise.

Error Study. We conduct an error analysis to better understand the failure modes of LLMs, which reveals three major sources of error at different levels. **(1) Inherent ambiguity in the data.** A subset of errors stems from cases where cultural stylistics are genuinely difficult to distinguish, even for human annotators. Specifically, 17 instances in C^{4}STYLI-T, primarily from films released around 2025, exhibit highly similar CN and HK titles. For example, the CN title of *Mickey 17* is "编号17 (Number 17)," while the HK title is "米奇17號 (*Mickey 17*)." Such translations offer minimal stylistic cues, rendering reliable attribution inherently challenging. A similar pattern appears in 30 instances from C^{4}STYLI-S advertising slogans. For example, the 2001 HK slogan of Ausupreme, "守護健康，澳至尊 (*lit. Safeguarding Your*

Health, Ausupreme)," is stylistically compatible with both CN and HK conventions. These cases, therefore, reflect an upper bound imposed by data ambiguity rather than model-specific limitations. **(2) LLM heuristic-based reasoning in movie title classification.** Beyond data ambiguity, LLMs exhibit systematic failures in classifying movie titles due to stereotypical and oversimplified heuristics. Models tend to reduce stylistic judgment to a single literal–adaptive axis, implicitly associating more literal renderings with CN and freer adaptations with HK. As a result, titles that adopt adaptive yet concise translation strategies are often misclassified. For instance, given the HK title of *Ready Player One*, "挑戰者1號 (*lit. Challenger 1*)," the model incorrectly classified it as CN, despite the title reflecting a narrative-driven translation strategy characteristic of HK conventions. Moreover, such reasoning is highly fragile: even changing the script from Simplified to Traditional Chinese can reverse the model's judgment. For the HK title of *Alpha* ("馴狼紀 [*lit. Taming the Wolf*]"), the same model interpreted the title as a creative HK-style adaptation under a Simplified Chinese prompt, but as a literal translation under a Traditional Chinese prompt. **(3) LLMs' limited understanding of HK-style humor and rhetoric in advertising slogans.** A distinct failure pattern is observed in advertising slogans, where LLMs struggle to recognize culturally grounded humor and rhetorical strategies specific to HK. Although models often state that HK slogans emphasize interaction, wordplay, and playfulness, this knowledge remains largely superficial. A representative failure is the 2025 HK advertisement for the Value Partners Asian Income Fund: "進退之間，如何攻守兼利？ (*lit. Between advancing and retreating, how can one achieve both growth and protection?*)." The slogan employs metaphorical and subtly humorous rhetoric, drawing on strategic and military imagery to frame personal finance decisions, a recurring pattern in HK advertising discourse. Nevertheless, all evaluated LLMs misclassified such cases as CN, indicating that their judgments are driven by a stereotype that CN advertising favors abstract and grand narratives, instead of sensitivity to culturally specific rhetorical strategies.

6 Can LLMs Produce Cultural Stylistics?

Method. Productive capability is another crucial aspect of assessing LLMs' cultural awareness. Thus, we formulate a conditional text generation task in which LLMs are instructed to produce texts tailored to a target region. For movie titles, the LLM is provided with the original English title and a plot summary in English, and is instructed to generate a Chinese title that reflects the stylistic conventions of either HK or CN. For advertising slogans, the LLM receives the brand name, product information, and release year, and is asked to generate a slogan appropriate for the specified region.

Evaluation. For **automatic evaluation** of the cultural stylistics of LLM generations, we train lightweight **classifiers** on frozen hidden representations extracted from Qwen3-8B. We concatenate the hidden states from the last 3rd and 4th layers and apply mean pooling over tokens to obtain fixed-length embeddings. Separate MLP classifiers are trained for movie titles and advertising slogans. The classifier achieves an accuracy of 82.13% on the movie title generation task and 89.73% on the advertising slogan generation task. We use these classifiers to determine the *proportion of generated texts that align with the target cultural style*. In addition, we conducted a **human evaluation** on outputs generated by *GPT-4.1*, *Qwen3-32B-Instruct*, and *DeepSeek-V3.2 (thinking)*. For each instance, annotators were presented with the input context alongside the model-generated text. They were asked to judge whether the text more closely reflected HK culture, CN culture, or whether the style was not apparent. Each instance was independently annotated by two annotators, one from HK and one from CN. In cases of disagreement, a third annotator was consulted to determine the final label. For each model and each domain, we randomly sampled 240 instances.

	Model	C^{4}Styli-T		C^{4}Styli-S	
		HK (%)	CN (%)	HK (%)	CN (%)
En	o4-mini	82.9	**81.7**	55.5	58.9
	GPT-4.1	**93.6**	73.2	59.6	61.0
	Llama-3-8B-Instruct	74.3	36.5	44.8	67.5
	Llama-3-70B-Instruct	78.6	44.1	43.2	**69.1**
	gemma-3-4b-it	82.8	21.5	63.4	61.7
	gemma-3-12b-it	79.6	72.2	58.6	60.2
	gemma-3-27b-it	81.3	68.9	**70.2**	62.1
Zh	DeepSeek-V3.2	**90.4**	77.6	**64.9**	63.0
	DeepSeek-V3.2 (Thinking)	88.6	**81.4**	53.6	**64.3**
	Qwen2.5-7B-Instruct	66.5	63.3	56.3	60.6
	Qwen2.5-14B-Instruct	85.8	54.7	46.5	64.1
	Qwen2.5-32B-Instruct	84.9	76.1	48.9	64.1
	Qwen2.5-72B-Instruct	84.4	79.3	52.5	61.0
	Qwen3-8B	75.0	72.1	42.8	**63.3**
	Qwen3-14B	76.3	78.9	48.2	61.4
	Qwen3-32B	83.1	69.2	49.0	60.9
Yue	SenseChat-5-Cantonese	70.2	**57.6**	46.8	60.0
	SenseNova-V6-5-Turbo	**83.1**	57.1	**50.2**	**62.1**

Table 3: LLM performance in cultural stylistic generation, automatically evaluated by classifiers.

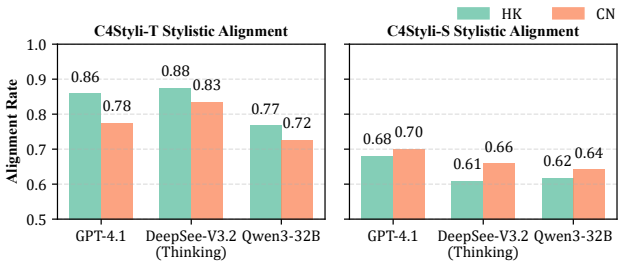


Figure 6: LLM performance in cultural stylistic generation, evaluated by human annotators. Krippendorff's α is 0.776, indicating a substantial agreement.

Overall Performance. Table 3 and Figure 6 presents LLM performance on cultural stylistic generation for translated movie titles and advertising slogans. Here, HK and CN scores indicate how well the generated text conforms to the target HK or CN cultural style, respectively. Across all models, **generation quality is consistently higher for translated movie titles than for advertising slogans**. For titles, many models achieve accuracies above 80% when generating HK-style text, whereas slogan generation remains substantially more challenging, with most accuracies ranging between 45% and 65%. This pattern contrasts with the mod-

Movie Context	Movie Title: Materialists Plot: A young, ambitious New York City matchmaker finds herself torn between the perfect match and her imperfect ex. Year: 2025
DeepSeek-V3.2 (Thinking)	CN-Style: 物质主义者 (Materialists) HK-Style: 紐約愛情選擇題 (A New York Love Dilemma)
Slogan Context	Brand Name: 中国人寿 (China Life) Product: 保险 (Insurance)
DeepSeek-V3.2 (Thinking)	CN-Style: 中国人寿，伴您一生平安。 (China Life, safeguarding your peace for a lifetime.) HK-Style: 中國人壽，守護香港每一家。 (China Life, protecting every family in Hong Kong.)

Table 4: LLM-generated texts in different cultural styles.

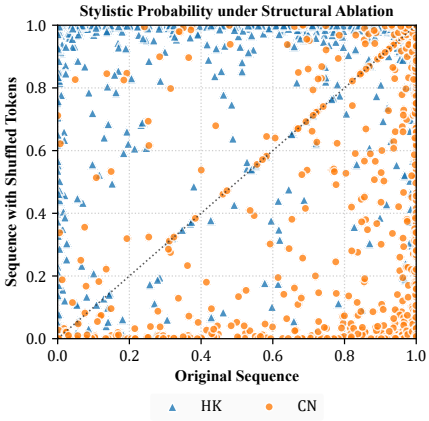


Figure 7: Stylistic Probabilities: Original Text vs. Sequence with Shuffled Tokens. The dotted line is $y = x$ for reference.

els' discrimination performance. This phenomenon is intuitive: translated movie titles are typically very short, which makes cultural distinctions harder to recognize but easier to generate, whereas advertising slogans tend to be longer and stylistically richer, making them easier to discriminate but more challenging to generate. **A notable asymmetry between HK and CN stylistic generation emerges across domains.** For translated movie titles, models tend to achieve higher accuracy when generating HK-style text than CN-style text, whereas for advertising slogans, CN-style generation often attains higher accuracy. This pattern indicates that LLMs' generation biases with respect to cultural stylistics are domain-dependent. **The effect of the model's primary training language on cultural stylistic generation remains inconclusive.** Performance differences are largely model-specific rather than language-specific, and comparable accuracies can be observed across English-, Chinese-, and Cantonese-oriented models.

Case Study. We present representative examples generated by DeepSeek-V3.2 (Thinking) in Table 4. The results demonstrate that the model effectively captures the essence of CN-style literalism for movie titles. In contrast, for HK-style adaptations, the model provides localized titles that align more closely with the underlying narrative or plot. Regarding slogans, while the model successfully reflects the concise and abstract nature of the CN-style, it struggles to replicate the distinctive humor and witty charm characteristic of Hong Kong's creative style, indicating a persistent gap in capturing nuanced cultural flair. Furthermore, a notable limitation is the model's tendency to rely on a superficial and somewhat tautological approach to localization. It frequently resorts to the explicit inclusion of the word "香港 (Hong Kong)" as a lexical crutch to signal regional identity, rather than naturally embedding cultural or linguistic nuances. This rigid dependency on explicit geographic markers highlights a significant deficit in LLMs' genuine understanding of cultural stylistics.

7 Do LLMs Internally Encode Cultural Stylistics?

Task. To discern whether LLMs leverage integrated structural understanding or merely rely on superficial lexical anchors, we evaluate representational depth by disrupting syntactic structures. We employ a Logistic Regression (LR) probe, trained on the frozen hidden states of Qwen3-8B to classify cultural styles (CN vs. HK). For each sample in the val set, we utilize the LR probe to obtain the probability of the ground-truth label. By comparing these scores across original sequences and their shuffled variants (which preserve token frequency but eliminate syntax), we quantify the marginal contribution of structural information to cultural identity. This examines whether the model's stylistic discernment is rooted in high-level syntax or isolated token features.

Results. Figure 7 illustrates the classification probabilities of sequences before and after token shuffling for each instance. The structural ablation results yield a critical finding: **the LLM fails to achieve a genuine understanding of HK culture, particularly regarding its aesthetic stylistics**. For HK-style texts (blue triangles), the probabilities exhibit a paradoxical increase or remain heavily clustered near $y = 1.0$ following the disruption of syntactic structure. This suggests that the model's recognition of HK culture is primarily driven by superficial lexical anchors rather than holistic structural understanding. Conversely, for CN-style texts (orange circles), the probabilities undergo a catastrophic drop upon shuffling, with many samples shifting from $x = 1.0$ to $y \approx 0.0$. This sensitivity to structural integrity demonstrates that the identification of CN-style identity relies significantly on integrated syntactic coherence.

8 Conclusion

This work studies cultural awareness in LLMs through *aesthetic stylistics*, an implicit cultural dimension governing how meaning is expressed. We first introduce C^{4}STYLI, a cross-cultural Chinese–Chinese benchmark built on highly stylized texts. With this dataset, our experiments yield three findings. First, LLMs significantly underperform humans in recognizing cultural stylistics, with strong domain sensitivity. Second, stylistic recognition and production are decoupled: accurate identification does not imply culturally aligned generation. Third, representation analyses reveal an encoding asymmetry: CN stylistics are captured in more distributed structures, whereas HK stylistics rely on sparse, high-salience lexical cues, reflecting superficial cultural awareness.

Ethical Statement

Dataset Collection. This work uses only publicly available data sources collected from open-access websites and publicly released materials. The data do not contain private, sensitive, or personally identifiable information, and are used strictly in accordance with their original usage terms and for research purposes only.

LLM Evaluation. During LLM-based evaluation, no personal or sensitive information was provided as input to any model. The prompts and test instances consist solely of publicly available or model-generated text.

Human Annotation. For human evaluation, we recruited annotators with relevant cultural and linguistic backgrounds. Participation was voluntary, and annotators were informed of the purpose of the study. The annotation tasks involved labeling model-generated content and did not expose annotators to harmful or sensitive material. No personal information about the annotators was collected or disclosed.

Use of Models. Finally, while we release trained models and code to support reproducibility, we encourage their responsible and appropriate use in accordance with relevant ethical guidelines.

Acknowledgments

We thank the anonymous area chair and anonymous reviewers for their insightful comments and valuable feedback during the review process. This study is funded by the Research Grants Council (project code: T43-518/24-N and PolyU/15213323) under the University Grants Committee, Hong Kong Special Administrative Region Government.

Contribution Statement

Jiashuo Wang and Fenggang Yu contribute equally to this work.

References

[Cao *et al.*, 2024] Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99, 2024.

[Chen *et al.*, 2020] Xinyu Chen, Liang Ke, Zhipeng Lu, Hanjian Su, and Haizhou Wang. A novel hybrid model for cantonese rumor detection on twitter. *Applied Sciences*, 10(20):7093, 2020.

[Faisal and Anastasopoulos, 2025] Fahim Faisal and Antonios Anastasopoulos. Testing the boundaries of LLMs: Dialectal and language-variety tasks. In Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jorg Tiedemann, and Marcos Zampieri, editors, *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 68–92, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[Hickey, 2014] Leo Hickey. *The Pragmatics of Style (RLE Linguistics B: Grammar)*. Routledge, 2014.

[Jiang *et al.*, 2025] Jiyue Jiang, Pengan Chen, Liheng Chen, Sheng Wang, Qinghang Bao, Lingpeng Kong, Yu Li, and Chuan Wu. How well do llms handle cantonese? benchmarking cantonese capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4464–4505, 2025.

[Kim *et al.*, 2024] Jaehyung Kim, Dongyoung Kim, and Yiming Yang. Learning to correct for qa reasoning with black-box llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8916–8937, 2024.

[Lee *et al.*, 2022] Jackson Lee, Litong Chen, Charles Lam, Chaak-ming Lau, and Tsz-Him Tsui. Pycantonese: Cantonese linguistics and nlp in python. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 6607–6611, 2022.

[Lee, 2011] John SY Lee. Toward a parallel corpus of spoken cantonese and written chinese. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1462–1466, 2011.

[Leung and Law, 2001] Man-Tak Leung and Sam-Po Law. Hkcac: the hong kong cantonese adult language corpus. *International journal of corpus linguistics*, 6(2):305–325, 2001.

[Li *et al.*, 2024] Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838, 2024.

[Mohammadi *et al.*, 2025] Hadi Mohammadi, Yasmeen FSS Meijer, Efthymia Papadopoulou, and Ayoub Bagheri. Do large language models understand morality across cultures? In *Proceedings of the 2nd LUHME Workshop*, pages 30–39, 2025.

[Nangia *et al.*, 2020] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, November 2020. Association for Computational Linguistics.

[Pawar *et al.*, 2025] Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96, 2025.

[Riffaterre, 1978] Michael Riffaterre. *Semiotics of poetry*, volume 19. Indiana University Press Bloomington, 1978.

[Sun and Xu, 2022] Zhewei Sun and Yang Xu. Tracing semantic variation in slang. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1313, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[Wales, 2014] Katie Wales. *A dictionary of stylistics*. Routledge, 2014.

[Wang *et al.*, 2024a] Hongru Wang, Wai-Chung Kwan, Min Li, Zimo Zhou, and Kam-Fai Wong. Kddres: A multi-level knowledge-driven dialogue dataset for restaurant towards customized dialogue system. *Comput. Speech Lang.*, 87(C), August 2024.

[Wang *et al.*, 2024b] Jian Wang, Chak Tou Leong, Jiashuo Wang, Dongding Lin, Wenjie Li, and Xiaoyong Wei. Instruct once, chat consistently in multiple rounds: An efficient tuning framework for dialogue. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3993–4010, 2024.

[Xiang *et al.*, 2019] Rong Xiang, Ying Jiao, and Qin Lu. Sentiment augmented attention network for cantonese restaurant review analysis. In *Proceedings of the 8th KDD Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*, pages 1–9. KDD WISDOM, 2019.

[Zhang *et al.*, 2024] Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. Llm as a mastermind: A survey of strategic reasoning with large language models. In *First Conference on Language Modeling*, 2024.

[Zhou *et al.*, 2025a] Li Zhou, Taelin Karidi, Wanlong Liu, Nicolas Garneau, Yong Cao, Wenyu Chen, Haizhou Li, and Daniel Hershcovich. Does mapo tofu contain coffee? probing LLMs for food-related cultural knowledge. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9840–9867, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[Zhou *et al.*, 2025b] Li Zhou, Lutong Yu, Dongchu Xie, Shaohuan Cheng, Wenyan Li, and Haizhou Li. Hanfubench: A multimodal benchmark on cross-temporal cultural understanding and transcreation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24627–24649, Suzhou, China, November 2025. Association for Computational Linguistics.