Modeling Chinese L2 Writing Development: The LLM-Surprisal Perspective

Jingying Hu[1], **Yan Cong**[1,2]
[1] Department of Linguistics, Purdue University
[2] School of languages and cultures, Purdue University
{hu880, cong4}@purdue.edu

Abstract

LLM-surprisal is a computational measure of how unexpected a word or character is given the preceding context, as estimated by large language models (LLMs). This study investigated the effectiveness of LLM-surprisal in modeling second language (L2) writing development, focusing on Chinese L2 writing as a case to test its cross-linguistical generalizability. We selected three types of LLMs with different pretraining settings: a multilingual model trained on various languages, a Chinese-general model trained on both Simplified and Traditional Chinese, and a Traditional-Chinese-specific model. This comparison allowed us to explore how model architecture and training data affect LLM-surprisal estimates of learners' essays written in Traditional Chinese, which in turn influence the modeling of L2 proficiency and development. We also correlated LLM-surprisals with 16 classic linguistic complexity indices (e.g., character sophistication, lexical diversity, syntactic complexity, and discourse coherence) to evaluate its interpretability and validity as a measure of L2 writing assessment. Our findings demonstrate the potential of LLM-surprisal as a robust, interpretable, cross-linguistically applicable metric for automatic writing assessment and contribute to bridging computational and linguistic approaches in understanding and modeling L2 writing development. All analysis scripts are available at https://github.com/JingyingHu/ChineseL2Writing-Surprisals.

1 Introduction

The rapid development of large language models (LLMs) has opened new avenues for modeling second language acquisition (SLA) and quantifying interlanguage systems. Among these, LLM-derived surprisal (hereafter LLM-surprisal), an information-theoretic measure, has shown strong potential for quantifying linguistic unpredictability across different contexts.

LLM-surprisal has been widely used in psycholinguistics studies to model human language comprehension (Wilcox et al., 2023; Huber et al., 2024). Recent research has highlighted its potential in modeling second language (L2) writing development. For example, Cong (2025) found that LLM-surprisal is potentially linked to L2 writing naturalness and can effectively capture lexical diversity and syntactic complexity in English L2 writing. As such, LLM-surprisal shows promise as a holistic metric for evaluating English L2 writing proficiency.

Despite these findings, the cross-linguistic generalizability of LLM-surprisal remains underexplored, particularly in typologically distant languages such as Chinese. The linguistic complexities that LLM-surprisal captures in Chinese L2 writing may differ from those observed in Cong's (2025) studies on English L2 writing. Therefore, a closer investigation of these differences is crucial not only for validating the cross-linguistic applicability of LLM-surprisal but also for understanding what specific linguistic features LLM-surprisal measures in the Chinese L2 writing context.

LLMs have recently demonstrated impressive language understanding and generation abilities, but their performance can vary across model architecture, scale, and training data. Notably, most mainstream LLMs are trained predominantly on English or other high-resource languages, raising concerns about their efficacy in low-resource settings or typologically diverse language contexts. Among these, Traditional Chinese texts remain particularly underrepresented due to their non-Latin script and limited presence in large-scale training corpora. This study also examined how different types of LLM (multilingual, Chinese-

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 172–183

general, and Traditional-Chinese-specific language models) process Traditional Chinese written texts, contributing to broader discussions on multilingual LLM performance in low-resource settings.

To summarize, the present study investigates the potential of LLM-derived surprisal as a robust and cross-linguistically applicable metric for L2 writing assessment, addressing the following research questions:

(1) **Cross-linguistic efficacy of LLM-surprisal in L2 writing assessment**
Can LLM-surprisal differentiate proficiency levels in Chinese L2 writing, thereby supporting its validity as a cross-linguistic metric for L2 writing evaluation?

(2) **The efficacy of multilingual LLMs in low-resource language settings**
If so, how do three types of LLMs, which vary in the scale of their training data on Traditional Chinese, differ in their ability to evaluate Chinese L2 writing?

(3) **LLM-surprisal's interpretability in the Chinese L2 context**
What aspects of linguistic complexity are captured by LLM-surprisal in Chinese L2 writing, and how do they differ from those captured in English L2 writing assessment?

For **RQ1**, we hypothesize that LLM-surprisal can differentiate different proficiency levels in Chinese L2 writing. That is, advanced-level essays tend to exhibit lower LLM-surprisal scores than beginner-level ones, as higher proficiency is associated with more natural and predictable language production.

For **RQ2**, among three LLMs examined, we hypothesize that the LLM pre-trained on Traditional Chinese-specific data will outperform both multilingual and general Chinese LLMs in modeling Chinese L2 writing development, due to its language-specific optimizations.

For **RQ3**, unlike classic complexity indices, which focus on specific aspects of language, we hypothesize that LLM-surprisal can capture the multidimensional nature of linguistic complexity in Chinese L2 writing. Building on prior work in English L2 research (Cong, 2025; Tang, 2024), we hypothesize that LLM-surprisal also captures lexical and syntactic complexity in the context of Chinese L2 writing assessment. Moreover, it may further capture character-level and discourse-level features, given the typological differences between Chinese and English.

The significance of this study lies in both its theoretical and practical contributions. By validating the effectiveness of LLM-surprisal in Chinese L2 writing, this study not only introduces a new potential quantitative metric for the automated writing assessment system for Chinese but also provides empirical evidence supporting the cross-linguistic applicability of surprisal as a universal and robust metric for L2 writing assessment. Additionally, by analyzing what linguistic complexity LLM-surprisal specifically measures in Chinese L2 writing, this study further improves the interpretability of LLM-surprisal in modeling L2 acquisition. Practically, the study provides insights into how LLMs can be applied in SLA research, particularly in selecting models for low-resource languages like Traditional Chinese.

2 Related Work

2.1 LLM-surprisal and multilingual LLMs

Mathematically, LLM-surprisal is defined as the negative log-probability of a word given its preceding context as computed by LLMs (Misra, 2022). LLM-surprisal has shown a strong correlation with **human language comprehension**, with higher LLM-surprisal indicating greater processing difficulty. Behavioral studies found that the higher LLM-surprisal predicts longer reading times, as cognitive load increases when processing less predictable input (Goodkind & Bicknell, 2018; Rethi, 2021). Neurocognitive further supports this relationship: words with higher surprisal elicit larger N400 amplitudes or increased P600 responses, both of which are neural markers of processing difficulty (Aurnhammer et al., 2021; Li et al., 2024).

The application of LLM-surprisal has also been extended to evaluate **human language production.** Recent studies suggest that LLM-surprisal has merged as a promising metric for assessing both writing quality and language proficiency among English L2 learners. Tang (2024) analyzed essays written by English L2 learners and found that as proficiency increases, learners convey more informative content while maintaining lower levels of unpredictability in their writing, as measured by entropy and LLM-surprisal respectively. Cong's (2025) study also confirmed LLM-surprisal's predictive power in tracking English L2 writing development, showing that it numerically represents the interplay between

syntactic complexity and lexical diversity in English L2 interlanguage development.

However, the robustness of LLM-surprisal as a metric for assessing Chinese L2 writing quality has not been sufficiently investigated. Furthermore, the typological difference between English and Chinese raises critical questions about whether LLM-surprisal captures comparable dimensions of linguistic complexity in Chinese L2 contexts. This dual gap highlights the need to examine both LLM-surprisal's cross-linguistic validity and its capacity to capture language-specific features in non-English settings.

The choice of LLM is important, as the effectiveness of LLM-surprisal is contingent upon the underlying language model's performance. Higher quality language models can produce more accurate surprisal estimates, which in turn better predict human behavior (Hao et al., 2020; Oh, 2023). A key consideration in LLM selection is whether to use a multilingual or monolingual model, yet previous studies have reported mixed findings. While some studies suggest that English-centric multilingual LLMs perform robustly across languages (Nguyen et al., 2023; Joshi et al., 2024; Kargaran et al., 2024), Xu et al. (2023) found that multilingual LLMs rely on translation-like behavior for cross-linguistic generalization, which may introduce biases in language-specific tasks. Moreover, multilingual LLMs tend to perform significantly better on high-resource languages, particularly those using Latin scripts, but struggle with low-resource languages and complex linguistic structures (Alam et al., 2024; Shu et al., 2024).

Based on these findings, and given that Traditional Chinese is a low-resource language in LLM training, it remains unclear how multilingual, Chinese-general, and Traditional Chinese-specific LLM differ in their ability to capture Traditional Chinese linguistic complexity or to provide more reliable surprisal estimates. Addressing these gaps is critical for understanding the applicability of LLM-surprisal in assessing Chinese L2 writing.

2.2 Classic linguistics indices in assessing Chinese L2 writing development

Previous studies on Chinese L2 writing assessment primarily focus on syntactic and lexical complexity indices. Early Chinese L2 studies adapted T-unit analysis from English, but Jin (2007) found it ineffective for distinguishing proficiency levels due to Chinese's topic-prominent structure. As an alternative, Jin (2007) proposed the Terminal Topic-Comment Unit (TTCU), which was later validated as a more effective measure (Jiang, 2013; Yu, 2021). Recent research has shifted from large-grained to more fine-grained syntactic analysis. At the level of phraseological complexity, Lu & Wu (2022) identified noun-phrase complexity as a stronger predictor of L2 Chinese writing quality, while Hu et al. (2022) highlighted the importance of word-combination-based measures. Hao et al. (2024) found that fine-grained syntactic indices more effectively predicted Chinese L2 writing quality than large-grained ones.

Lexical complexity indices have also been widely used in Chinese L2 writing evaluation. For example, Wang (2017) found that lexical errors, the number of unique word types, and the use of high-frequency words were effective indicators of the writing performance of Chinese learners.

It is worth noting that *Chinese Proficiency Grading Standards for International Chinese Language Education* (2021) (hereafter referred to as the *Grade Standard*), which defines the characters, vocabulary, and syntactic structures that Chinese learners at each proficiency level are expected to master, provides an effective tool for measuring Chinese L2 writing complexity. For example, Wang et al. (2022) used advanced-level vocabulary and grammar items from the *Grade Standard* to assess lexical and grammatical sophistication in Chinese L2 writing. They found that the use of advanced-level vocabulary and grammatical structure was strongly correlated with learner proficiency.

Despite these findings, few studies have focused on lexical semantic diversity and its role in tracking Chinese L2 writing development. Different from lexical diversity measured by TTR (the ratio of unique word types of total words), lexical semantic diversity is a computationally derived measure of the variability in a word's meaning across different contexts (Hoffman et al., 2012). A word with a high semantic diversity value indicates that it appears across more varied, semantically distinct contexts. Berger et al. (2017) found that advanced learners of English have greater lexical semantic diversity values in their language production, suggesting they can use words across many semantic diverse contexts.

Taken together, various linguistics complexity indices were used to characterize Chinese L2

learners' writing development, which provides a strong foundation for testing the reliability and validity of the new indices. Among the classic indices, the role of *Grade Standard* and lexical semantic diversity needs to be further investigated. Moreover, while much attention has been given to lexical, phrasal, and sentence-level complexity in Chinese L2 writing, relatively little is known about whether discourse-level features can effectively distinguish different proficiency levels. LLM-surprisal, which captures both the local and global unpredictability and naturalness based on prior context information, holds the potential to fill this gap by evaluating L2 quality at the discourse level or textual level that spans across a larger context.

Furthermore, previous studies have emphasized the need for assessment metrics that are sensitive to Chinese-specific linguistic properties. Unlike English, a subject-prominent language, Chinese is a topic-prominent and pro-drop language, allowing subject omission in the discourse (Li and Thompson, 1976; Liu, 2010). Chinese also has a logographic writing system, where each character represents a morpheme or meaning unit, in contrast to English's alphabetic system (Wang, 2015). Additionally, Chinese lacks rich inflectional morphology found in English and instead relies on aspect markers and contextual cues (Klein et al., 2000). These typological differences not only set Chinese apart from English but also shape how Chinese L2 learners implicitly organize their writings across lexical, syntactic, and discourse levels.

Therefore, this study applied LLM-surprisal to Chinese L2 writing to examine its predictive power in assessing writing proficiency and its ability to capture Chinese-specific typological features. Additionally, we examined the interpretation of LLM-surprisal in the Chinese L2 writing context, and how this may differ from its established interpretations in English L2 assessment.

3 Method

3.1 Dataset

We used the publicly available TOCFL Learner dataset[1] (Lee et al., 2018), which collected written essays from the standardized Test of Chinese as a Foreign Language. This dataset includes 2,837 essays written by learners from 46 different L1 backgrounds, covering proficiency levels A2 to C1, as defined by the CEFR framework. Although each essay was originally scored on a 0-5-point scale by at least two Chinese teachers, only essays that scored above 3, which is indicative of sufficient proficiency to meet the passing grade, were included in this dataset.

In the present study, we selected 65 essays from each CEFR level (A2, B1, B2, C1) to ensure balanced comparisons across proficiency groups. These essays were also carefully matched based on their scores and the learners' L1 backgrounds (see Appendix A for details). Given that LLM-surprisal can be influenced by text length, we also explicitly controlled for essay length in the experiment, with each essay containing approximately 200 Chinese characters. After applying these controls, we compiled a balanced dataset of 260 Traditional Chinese essays for subsequent analysis.

3.2 LLM-surprisals calculation

LLMs-surprisals were calculated as shown below in (1) (Misra, 2022; Cong, 2025).

$$surprisal(w_t) = -logP\ (w_t|(w_{1\ldots t-1}) \quad (1)$$

In order to answer whether LLM-surprisal can effectively distinguish different proficiency levels in Chinese L2 writing, we calculated mean LLM-surprisal scores for each essay. Specifically, we first computed character-wise surprisal within each essay, and the surprisal scores of all characters were summed and then divided by the essay length (total number of characters). We hypothesize that low surprisal, as an indicator of low unpredictability, is associated with advanced learner's writing, given that as proficiency increases, proficient learners tend to produce natural writings in their L2.

Three transformer-based language models were selected to calculate the LLM-surprisal scores:

1) *bigscience/bloom-7b1* (Le Scao et al., 2023), a large-scale multilingual model trained on 45 natural languages with 7.07 billion parameters;

2) *hfl/chinese-llama-2-7b* (Cui et al., 2023), a pre-trained transformer model trained on both simplified and traditional Chinese language with 7 billion parameters;

3) *Taiwan-LLM-7B-v2.1-chat* (Lin and Chen, 2023), an LLM exclusively tailored for Traditional Chinese with 7 billion parameters, with an emphasis on linguistic norms specific to Taiwan.

[1] https://github.com/NYCU-NLP/TOCFL

These selections enable us to compare how model architecture and training data of LLM affect their ability to model (Traditional) Chinese L2 writing, especially in the low-resource language setting. All selected LLMs are publicly available on HuggingFace (https://huggingface.co/). We utilized minicons (Misra, 2022) to conduct a systematic evaluation of different LLMs' behavior.

3.3 Classic Chinese complexity indices

To tease apart what aspects of L2 Chinese the LLM-indices can characterize, and to examine the reliability and validity of the LLM-surprisal, we selected 16 well-established Chinese complexity indices, including character, lexicon, syntax, clause, discourse coherence, and text length indices. These indices have been validated in previous Chinese L2 writing research and have also been incorporated into different linguistic complexity calculation tools (Cui et al., 2022; Sung et al., 2016). We used L2C-Rater (Wang & Hu, 2021) and CTAP for Chinese (Cui et al., 2022) to calculate these complexity indices. Additionally, we correlated these indices with LLM-surprisal to better understand what aspects of linguistic complexity are captured by LLM-surprisal in Chinese L2 essays.

We first calculated **the cohesive complexity**: first personal pronouns per token. As a pro-drop and topic-prominent language, Chinese allows the omission of subject pronouns when they can be inferred from context (Li & Thompson, 1989), a phenomenon known as zero anaphora. At the discourse level, this feature results in fewer overt first-person pronouns compared to non-pro-drop languages, which rely on pronoun retention to maintain coherence. Therefore, the appropriate use of first-person pronouns, particularly the management of pronoun dropping and retention, serves as a crucial indicator of learners' grammatical knowledge and their mastery of Chinese-specific discourse conventions. Moreover, it also allows us to examine whether LLM-surprisal is sensitive to discourse-level cohesive complexity.

Given the Chinese L2 context, we also included the **character, lexical, and syntactical sophistication indices based on the *Grade Standard***, which provides a standardized metric for assessing the difficulty of Chinese characters, words, and grammatical structure for Chinese L2 learners. Higher levels indicate greater complexity or difficulty. For each essay, we calculated the average levels of characters, average levels of words, and average levels of grammar. Advanced Chinese learners are expected to produce writings with higher average levels in all three dimensions.

In addition to lexical sophistication, we included **lexical semantic diversity,** following Cong (2025). We used the semantic diversity norms established by Chang & Lee (2018), which provide a semantic diversity value for each Chinese character. By mapping each character in the essays to its corresponding value using a dictionary-based approach, we calculated the mean lexical semantic diversity for each essay. Higher lexical semantic diversity is expected to be observed in advanced learners' essays due to their ability to utilize diverse contextual words.

For **clausal complexity**, we calculated four important phrases per simple clause: coordinate phrases, noun phrases, prepositional phrases, and verb phrases. Additionally, syntactical indices such as mean dependency distance and the height of the highest parse tree were also included to measure **syntactical complexity**. Moreover, we included **text length indices** such as mean length of sentences, clauses, and T-units, as well as the number of clauses per sentence, and number of T-units per sentence to measure sentence production complexity. We expect that advanced Chinese learners show higher values for these clausal, syntactical, and text length indices in their essays.

4 Results

Statistical analyses were conducted in R (R Core Team, 2023). The results of the Shapiro–Wilk test and Levene's test on all dependent variables suggested that the data violates the assumptions (normality and homogeneity of variance) for parametric tests. Therefore, we used non-parametric statistical tests throughout the paper. The alpha level is 0.05.

4.1 LLM-surprisal's efficiency in modeling Chinese L2 writing development

In order to answer whether LLM-surprisal can differentiate proficiency levels in Chinese L2 writing, we conducted three separate Kruskal-Wallis tests to examine the differences in LLM-surprisal across different proficiency levels. Effect sizes for the statistical tests are reported in Table 1.

The result showed significant differences in LLM-surprisal scores across proficiency levels for all three LLMs (Bloom: $\chi2=33.39$, $p<.000$,

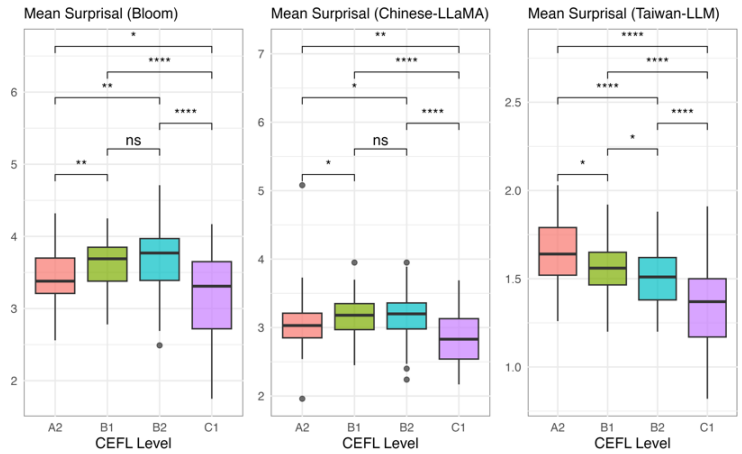

Figure 1: Paired comparisons across four proficiency levels (A2–C1). Significance notation: *p < 0.05; **p < 0.01; ****p < 0.0001; ns: p > 0.05.

η2=0.12; Chinese-LLaMA: χ2=32.4, p<.000, η2=0.12; Taiwan-LLM: χ2=62.81, p<.000, η2=0.23), indicating that LLM-surprisal is effective in distinguishing between L2 proficiency levels. Among the three LLMs, Taiwan-LLM demonstrated the largest effect size, suggesting its greater sensitivity to proficiency differences.

As a post-hoc analysis, to identify the specific proficiency levels at which the LLM-surprisal indices become informative, we conducted Mann-Whitney U tests on LLM-surprisal scores between adjacent proficiency levels. The results are visualized in Figure 1. Detailed descriptive statistics can be found in Appendix B.

LLM-surprisal scores calculated by Bloom and Chinese-LLaMA showed a similar trend across different proficiency levels. That is, as proficiency levels increase from A2 to B2 level, the mean LLM-surprisal scores slightly increase. However, no statistically significant difference was found in LLM-surprisal scores between the B1 and B2 levels (p > .05). Notably, the C1 level showed significantly lower LLM-surprisal scores than the other proficiency levels. In summary, LLM-surprisal scores calculated by Bloom and Chinese-LLaMA exhibited less distinct separation between adjacent levels, but both confirmed that advanced Chinese learners produce essays with the lowest LLM-surprisal score.

Taiwan-LLM showed the most consistent LLM-surprisal trends across proficiency levels, with surprisal scores decreasing significantly as proficiency increased. All pairwise comparisons were statistically significant (p < .05). These results support our hypothesis. That is, Taiwan-LLM can capture the surprisal scores difference across all proficiency levels. In other words, Taiwan-LLM is more sensitive to the subtle variations in learner writing at different proficiency stages.

In summary, these findings confirm that LLM-surprisal effectively differentiates proficiency levels in Chinese L2 writing, supporting its cross-linguistic applicability despite typological differences between Chinese and English. Across three LLMs, essays written by advanced Chinese learners (C1) consistently exhibited the lowest surprisal scores. On the other hand, Taiwan-LLM outperformed both Bloom and Chinese-LLaMA, given that it shows the largest effect size in the Kruskal-Wallis tests and demonstrated a more distinct separation between adjacent proficiency levels.

4.2 Interpreting LLM-surprisal in Chinese L2 writing context

To further validate the effectiveness of LLM-surprisal and identify which aspects of linguistic complexity it captures in Chinese L2 writing, we conducted a correlation analysis between LLM-surprisal scores and 16 classic complexity indices.

Table 1 provides the results of Kruskal–Wallis tests on three LLM-surprisal indices and 16 classic linguistic complexity indices. We found that most classic indices showed generally stronger effects than the new LLM-surprisal indices, suggesting that the classic complexity measures at the levels of characters, lexicon, phrases, coherence, syntax, and text length remain robustly informative in indexing Chinese L2 writing development.

Figure 2 shows a heatmap visualization of Spearman's rank correlations between LLM-surprisal scores and 16 classic complexity indices.

	Index	χ2 (3)	Sig	Eta2
LLM-surprisal scores	Bloom surprisal	33.392	0.000	0.119
	Chinese-LLaMA surprisal	32.397	0.000	0.115
	Taiwan-LLM surprisal	62.808	0.000	0.234
Classic index: cohesive complexity	First Personal Pronouns per Token	94.124	0.000	0.356
Classic index: character	Average Character Levels	159.049	0.000	0.610
Classic indices: lexicon	Lexical Semantic Diversity	98.308	0.000	0.372
	Average Word Levels	165.828	0.000	0.636
Classic indices: clausal complexity	Coordinate Phrases per Simple Clause	70.698	0.000	0.264
	Noun Phrases per Simple Clause	77.881	0.000	0.293
	Prepositional Phrases per Simple Clause	29.662	0.000	0.104
	Verb Phrases per Simple Clause	30.388	0.000	0.107
Classic indices: syntactic complexity	Mean Dependency Distance	69.672	0.000	0.260
	The Height of the Highest Parse Tree	46.281	0.000	0.169
	Average Grammatical Levels	9.317	0.025	0.025
Classic indices: text length	Mean Length of Sentences	112.998	0.000	0.430
	Mean Length of Clauses	92.557	0.000	0.350
	Mean Length of T-Units	132.487	0.000	0.506
	Number of Clauses per Sentence	59.361	0.000	0.220
	Number of T-Units per Sentence	20.253	0.000	0.067

Table 1: Efficacy comparisons between the classic and the LLM-Surprisal indices in modeling Chinese L2 writing proficiency and development.

At **the discourse coherence level**, all LLM-surprisal scores were positively correlated with first personal pronouns per token, indicating that essays with lower surprisal scores tend to have fewer first personal pronouns per token, that is, less first personal pronouns repetition in the essay.

At **the character level**, all LLM-surprisals were strongly negatively correlated with average character levels, indicating that essays with lower mean surprisal scores had higher average character levels. Taiwan-LLM surprisal showed the strongest correlation coefficient, which means Taiwan-LLM is more effective at capturing character complexity than the other two LLMs.

At **the lexicon level**, all LLM-surprisals were strongly negatively correlated with average word levels and lexical semantic diversity, indicating that essays with higher mean surprisal scores had higher average word levels and higher lexical semantic diversity. Taiwan-LLM also showed the highest correlation coefficient here.

Notably, only Taiwan-LLM surprisal scores correlated with **clausal, syntactic, and sentence complexity indices**. First, Taiwan-LLM surprisal scores were strongly negatively correlated with prepositional phrases per simple clause, noun phrases per simple clause, and coordinate phrases per simple clause. That indicated that essays with lower mean surprisal scores had more complex phrases per clause. Second, Taiwan-LLM surprisal scores were strongly negatively correlated with mean dependency distance and the height of the highest parse tree, indicating that essays with lower surprisal scores exhibited higher syntactic complexity. Third, Taiwan-LLM surprisal scores also strongly negatively correlated with the mean length of sentence, clause, and T-units, as well as number of clauses per sentence.

We also found strong positive correlations within LLMs-surprisal scores. That is not surprising, since they are all transformer-based decoder models and share the core architecture. Besides that, Taiwan-LLM showed a higher correlation coefficient with Chinese-LLaMa than with Bloom. The stronger correlation may be attributed to the overlapping Traditional Chinese training data within these two Chinese LLMs.

Taken together, LLM-surprisal can capture linguistic complexity at the levels of coherence, characters, lexicon, phrases, syntax, and text length in Chinese L2 writing.

5 Conclusions and Discussions

In this study, we attempted to answer whether LLM-surprisal can serve as an effective and interpretable metric for L2 writing assessment across languages, and whether multilingual LLMs can effectively handle low-resource languages, such as traditional Chinese text.

Consistent with our hypothesis, we found that LLM-surprisal can effectively differentiate essays written by Chinese L2 learners across different proficiency levels. Specifically, advanced Chinese L2 learners exhibit lower surprisal values in their essays compared to less proficient learners. In line with Cong's (2025) work on English L2 writing, the present study demonstrates that LLM-surprisal is also effective in a typologically distinct language, such as Chinese, thereby further supporting its cross-linguistic robustness as a measure of

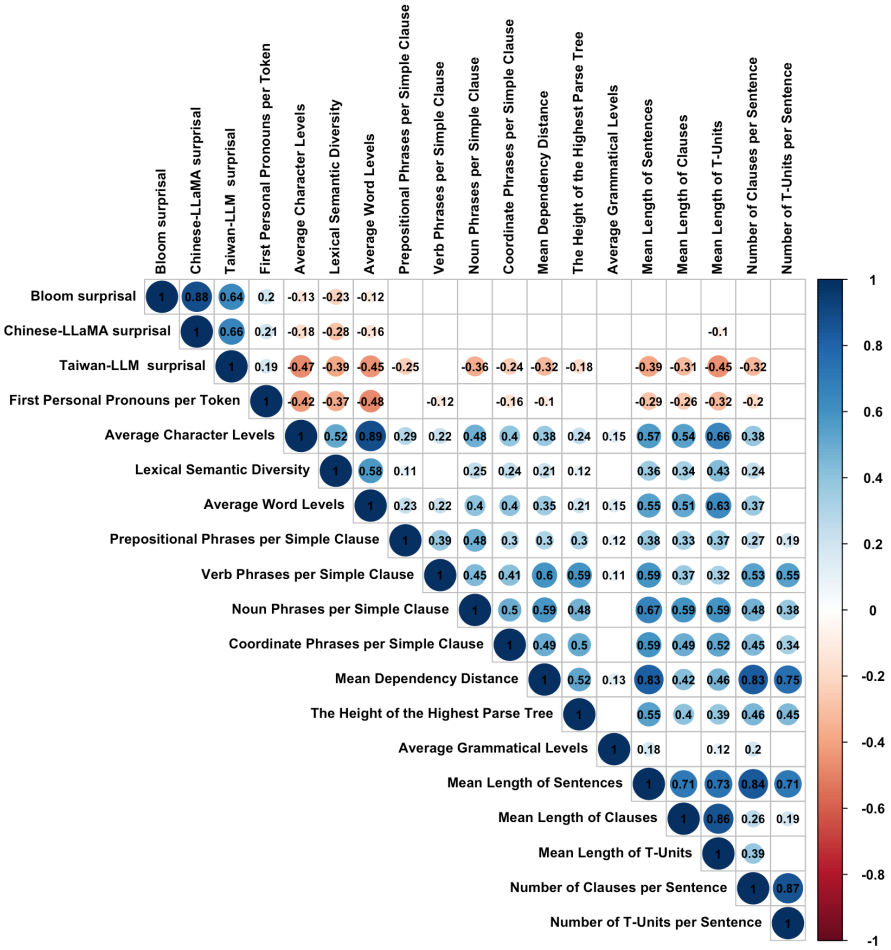

Figure 2: Correlations coefficients heatmap of LLMs-surprisals and 16 selected classic linguistic complexity indices. Darker colors indicate stronger correlations. Insignificant cells are left blank.

linguistic proficiency. These findings align with previous studies that demonstrated the cross-linguistic universality of surprisal effects in naturalistic reading (Wilcox et al., 2023; Xu et al., 2023). Extending this line of work, the present study broadens the application of LLM-surprisal by applying it to modeling language production, particularly within the contexts of automated L2 writing assessment and L2 acquisition modeling. In sum, this study highlights the potential of LLM-surprisal as a universal and effective metric for modeling human language behavior across both receptive (comprehension) and productive (writing) modalities.

Could the cross-linguistic efficacy of LLM-surprisal lie in its ability to capture multiple dimensions of linguistic complexity, including language-specific features? L2 learners' writing development is complex and encompasses multiple facets of language complexity. Cong's (2025) study on English L2 writing showed that LLM-surprisal functions as an integrated measure, capturing both lexical diversity and syntactic complexity. We speculate that LLM-surprisal may serve as a proxy for evaluating both linguistic complexity and the naturalness of learners' essays in the Chinese L2 writing context. Our correlation analysis indicated that LLM-surprisal computed by Taiwan-LLM significantly showed a significant correlation with a wide range of linguistic complexity indices (see Figure 2), suggesting its capacity to model language complexity at the character, lexical, syntactic, clausal, sentential, and discourse levels. In other words, while LLM-surprisal can capture similar linguistic complexities in both English and Chinese L2 contexts, it also uniquely showed sensitivity to certain characteristics specific to Chinese, such as character complexity and cohesive complexity. For example, we found that essays with lower surprisal scores exhibit reduced usage of first-person pronouns, a characteristic of pro-drop languages such as Chinese where subject pronouns can be omitted when implied by the context. Therefore, this study suggests that LLM-

surprisal can capture the appropriate use of first-person pronouns following Chinese-specific discourse coherence conventions.

With the rapid development of LLMs, it has become increasingly important to understand how multilingual LLMs and monolingual LLMs differ in their performance across tasks. Previous studies showed mixed findings on whether multilingual or monolingual language models perform better (Goyal et al., 2020; Rönnqvist et al., 2019; Kargaran et al., 2024). In the present study, we utilized three different LLMs (Bloom, Chinese-LLaMa, Taiwan-LLM) to calculate the mean surprisal scores of each essay written in Traditional Chinese. The three LLMs feature different architectural designs and were trained on progressively larger Traditional Chinese data, allowing us to further investigate LLM's performance in low-resource languages. Among the tested models, Taiwan-LLM exhibited the best performance, characterized by the largest effect sizes, clear distinctions between different proficiency levels, and strong correlations with multiple classic language complexity indices. In contrast with Chinese-general LLM, Taiwan-LLM is trained on Traditional Chinese data with diverse textual sources, and can better capture linguistic features in the Chinese learners' essays written in Traditional Chinese. In short, LLM selection is indeed crucial for low-resource languages, as the performance of these models heavily depends on the availability and quality of training data specific to such languages. Our findings highlight that monolingual LLMs outperformed multilingual LLMs in the low-resource language setting.

Why does the Taiwan-LLM outperform the other two models? The strong performance of the Taiwan-LLM may initially raise concerns about potential data overlap or overfitting, especially given its pretraining on traditional Chinese texts. However, we argue that the observed outstanding performance cannot be fully attributed to such data familiarity. Our correlation analysis provides evidence that the LLM-surprisal estimates reflect more than mere memorization of surface patterns. Specifically, Taiwan-LLM surprisal scores showed significant correlations with a broad range of linguistic complexity indices, including lexical diversity, syntactic depth, and discourse coherence. These correlations suggest that the model captures meaningful structural and functional aspects of language that are relevant to L2 proficiency, rather than simply reproducing patterns from potentially familiar training data. In this sense, Taiwan-LLM's strongest performance likely reflects a genuine sensitivity to various linguistic complexity indices of proficient writing, thereby reinforcing the interpretability and potential utility of LLM-surprisal in L2 assessment contexts.

This study has important implications in different aspects. First, this study introduces, validates, and demystifies LLM-surprisal as a novel and robust tool for analyzing linguistic complexity in Chinese L2 writing. Given its powerful ability to capture Chinese-specific features, this study expands our methodological toolkit for automatic Chinese L2 essay scoring or writing assessment. This study also advances computational approaches to modeling L2 acquisition and human language behavior. We not only demonstrate LLM-surprisal's cross-linguistic utility in modeling language production but also provide insights into the role of LLM architecture and training data in modeling linguistic complexity in the low-resource language setting.

6 Limitations

This study provides new insights into LLM-surprisal as a cross-linguistic metric for L2 writing assessment. However, several limitations should be acknowledged. First, the number of essays per proficiency level is limited. The writing genres were also not well-controlled. This is attributed to the inherent design of the Test of Chinese as a Foreign Language (TOCFL). This standardized test assigns different genres to different proficiency levels, for example, practical messages and picture-based storytelling at A2, functional writing and letters at B1–B2, and argumentative or report-style essays at C1. Genre variability may introduce differences in rhetorical structure, topical content, and linguistic features, potentially confounding the relationship between LLM-surprisal and proficiency, as different genres have distinct lexical and syntactic characteristics. Although our results showed the robust effectiveness of LLM-surprisal in evaluating L2 essay proficiency with a broad spectrum of genres, the nature of the dataset and the variability of writing tasks in this study restrict our ability to isolate genre-specific effects. Future research should investigate the impact of genre by analyzing essays from a single genre across multiple proficiency levels and using larger, more balanced datasets

We maintain that our correlation analysis provides an approach to unpack LLM-surprisal, improving LLMs' interpretability and transparency in L2 modeling. While the Taiwan-LLM demonstrated particularly strong performance, questions about the potential overlap between its training data and the learner essays remain outside the scope of our current investigation. Future studies should evaluate model performance on out-of-domain writing samples and systematically investigate how different pretraining corpora influence surprisal estimates. In addition, fine-tuning multilingual models on controlled datasets may help disentangle the effects of language exposure, model architecture, and data familiarity in surprisal-based assessments.

Another limitation lies in our exclusive focus on written text, which leaves open questions about how LLM-surprisal operates in spoken or multimodal L2 contexts. LLM-surprisal can also be measured at phoneme or utterance level. Previous studies have shown that disfluencies tend to occur before high-surprisal and syntactically complex elements (Dammalapati et al., 2021), and words with higher surprisal are associated with longer articulation durations (Lazic et al., 2025). Future studies should explore the applicability of surprisal in L2 spoken data and the effectiveness of LLM-surprisal on automatic phonetic evaluation of L2 speech.

Acknowledgments

Special thanks to Elaine J. Francis for her kind support. We appreciate anonymous reviewers' constructive and helpful comments. This project is supported by the School of Interdisciplinary Studies, Purdue University.

References

Alam, F., Chowdhury, S. A., Boughorbel, S., & Hasanain, M. (2024, March). LLMs for low-resource languages in multilingual, multimodal, and dialectal settings. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, 27–33.

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., & Crocker, M. (2021). Retrieval (N400) and integration (P600) in expectation-based comprehension. *PLOS ONE, 16*(9), e0257430. https://doi.org/10.1371/journal.pone.0257430

Berger, C. M., Crossley, S. A., & Kyle, K. (2017). Using novel word context measures to predict human ratings of lexical proficiency. *Journal of Educational Technology & Society*, *20*(2), 201-212.

Chang, Y. N., & Lee, C. Y. (2018). Semantic ambiguity effects on traditional Chinese character naming: A corpus-based approach. *Behavior Research Methods, 50*(6), 2292–2304.

Cong, Y. (2025). Demystifying large language models in second language development research. *Computer Speech & Language, 89*, 101700.

Cui, Y., Zhu, J., Yang, L., Fang, X., Chen, X., Wang, Y., & Yang, E. (2022, June). CTAP for Chinese: A linguistic complexity feature automatic calculation platform. *Proceedings of the 13th Language Resources and Evaluation Conference*, 5525–5538.

Cui, Y., Yang, Z., & Yao, X. (2023). Efficient and effective text encoding for Chinese LLAMA and Alpaca. *arXiv preprint arXiv:2304.08177.*

Dammalapati, S., Rajkumar, R., Ranjan, S., & Agarwal, S. (2021, February). Effects of duration, locality, and surprisal in speech disfluency prediction in English spontaneous speech. In *Proceedings of the Society for Computation in Linguistics 2021* (pp. 91-101).

Goyal, N., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.747

Goodkind, A., & Bicknell, K. (2018, January). Predictive power of word surprisal for reading times is a linear function of language model quality. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, 10–18.

Hao, Y., Mendelsohn, S., Sterneck, R., Martinez, R., & Frank, R. (2020). Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. *arXiv preprint arXiv:2009.03954.*

Hao, Y., Wang, X., Bin, S., Yang, Q., & Liu, H. (2024). How syntactic complexity indices predict Chinese L2 writing quality: An analysis of unified dependency syntactically-annotated corpus. *Assessing Writing*, 61, 100847.

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods, 45*(3), 718–730.

Hu, R., Wu, J., & Lu, X. (2022). Word-combination-based measures of phraseological diversity, sophistication, and complexity and their relationship to second language Chinese proficiency

and writing quality. *Language Learning, 72*(4), 1128–1169.

Huber, E., Sauppe, S., Isasi-Isasmendi, A., Bornkessel-Schlesewsky, I., Merlo, P., & Bickel, B. (2024). Surprisal from language models can predict ERPs in processing predicate-argument structures only if enriched by an Agent Preference principle. *Neurobiology of Language, 5*(1), 167–200.

Jiang, W. (2013). Measurements of development in L2 written production: The case of L2 Chinese. *Applied Linguistics, 34*(1), 1–24.

Jin, H. G. (2007). Syntactic maturity in second language writings: A case of Chinese as a foreign language (CFL). *Journal of the Chinese Language Teachers Association, 42*(1), 27.

Joshi, R., Singla, K., Kamath, A., Kalani, R., Paul, R., Vaidya, U., ... & Long, E. (2024). Adapting multilingual LLMs to low-resource languages using continued pre-training and synthetic corpus. *arXiv preprint arXiv:2410.14815.*

Kargaran, A. H., Modarressi, A., Nikeghbal, N., Diesner, J., Yvon, F., & Schütze, H. (2024). MEXA: Multilingual evaluation of English-centric LLMs via cross-lingual alignment. *arXiv preprint arXiv:2410.05873.*

Klein, W., Li, P., & Hendriks, H. (2000). Aspect and assertion in Mandarin Chinese. *Natural Language & Linguistic Theory, 18*(4), 723–770.

Lazic, J., & Vujnovic, S. (2025). Influence of the Surprisal Power Adjustment on Spoken Word Duration in Emotional Speech in Serbian. SSRN. https://ssrn.com/abstract=5102491

Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., ... & Al-Shaibani, M. S. (2023). Bloom: A 176B-parameter open-access multilingual language model.

Lee, L. H., Tseng, Y. H., & Chang, L. P. (2018, May). Building a TOCFL learner corpus for Chinese grammatical error diagnosis. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Li, C. N., & Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar.* University of California Press.

Li, C., & Thompson, S. (1976). Subject and topic: A new typology of language. In C. Li (Ed.), *Subject and topic* (pp. 457–489). Academic Press.

Li, J., & Futrell, R. (2024). Decomposition of surprisal: Unified computational model of ERP components in language processing. *arXiv preprint arXiv:2409.06803*.

Liu, H. (2010). Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua, 120*(6), 1567–1578.

Lin, Y. T., & Chen, Y. N. (2023). Taiwan llm: Bridging the linguistic divide with a culturally aligned language model. *arXiv preprint arXiv:2311.17487*.

Lu, X., & Wu, J. (2022). Noun-phrase complexity measures in Chinese and their relationship to L2 Chinese writing quality: A comparison with topic–comment-unit-based measures. *The Modern Language Journal, 106*(1), 267–283.

Center for Language Education and Cooperation, China's Ministry of Education. (2021). *Chinese Proficiency Grading Standards for International Chinese Language Education*. Higher Education Press

Misra, K. (2022). minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112.*

Nguyen, X. P., Aljunied, S. M., Joty, S., & Bing, L. (2023). Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts. *arXiv preprint arXiv:2306.11372.*

Oh, B. (2023). Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. https://doi.org/10.18653/v1/2023.findings-emnlp.128

R Core Team. (2023). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Rathi, N. (2021, June). Dependency locality and neural surprisal as predictors of processing difficulty: Evidence from reading times. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 171–176.

Rönnqvist, S., Kanerva, J., Salakoski, T., & Ginter, F. (2019). Is multilingual BERT fluent in language generation? *arXiv preprint arXiv:1910.03806.*

Shu, P., Chen, J., Liu, Z., Wang, H., Wu, Z., Zhong, T., ... & Liu, T. (2024). Transcending language boundaries: Harnessing LLMs for low-resource language translation. *arXiv preprint arXiv:2411.11295.*

Sung, Y. T., Chang, T. H., Lin, W. C., Hsieh, K. S., & Chang, K. E. (2016). CRIE: An automated analyzer for Chinese texts. *Behavior Research Methods, 48*, 1238–1251.

Tang, Z., & van Hell, J. G. (2024). Learning to Write Rationally: How Information Is Distributed in Non-

Native Speakers' Essays. *arXiv preprint arXiv:2411.03550*.

Wang, Y. (2017). The Correlation between Lexical Richness and Writing Score of CSL Learner—the Multivariable Linear Regression Model and Equation of Writing Quality. *Applied Linguistics*, (2), 93-101

Wang, H., Cheng, Y., & Hu, X. (2022). A dynamic development study of CSL writing quality based on lexical features and grammatical patterns. *TCSOL studies*, (2), 20-31

Wang, Y., & Hu, R. (2021). A prompt-independent and interpretable automated essay scoring method for Chinese second language writing. In *Chinese Computational Linguistics* (pp. 450–470). Springer. https://doi.org/10.1007/978-3-030-84186-7_30

Hu, R., Wu, J., & Lu, X. (2022). Word-Combination-Based Measures of Phraseological Diversity, Sophistication, and Complexity and Their Relationship to Second Language Chinese Proficiency and Writing Quality. *Language Learning*, *72*(4), 1128-1169.

Wang, M., Li, C., & Lin, C. (2015). The contributions of segmental and suprasegmental information in reading Chinese characters aloud. Plos One, 10(11), e0142060. https://doi.org/10.1371/journal.pone.0142060

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics, 11*, 1451–1470.

Xu, W., Chon, J., Liu, T., & Futrell, R. (2023, December). The linearity of the effect of surprisal on reading times across languages. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 15711-15721).

Yu, Q. (2021). An organic syntactic complexity measure for the Chinese language: The TC-unit. *Applied Linguistics*, *42*(1), 60-92.

Appendix A. L1 Backgrounds Distribution of L2 Learners in the Writing Dataset

L1 of L2 Learners	Number of Essays per Level	Total number of essays	Percentage of Dataset
English	18	72	27.69%
Vietnamese	14	56	21.54%
Japanese	13	52	20.00%
Korean	9	36	13.85%
Indonesian	6	24	9.23%
French	2	8	3.08%
Hungarian	1	4	1.54%
Russian	1	4	1.54%
Swedish	1	4	1.54%

Appendix B. Summary of LLM-Surprisal Scores, Mean (SD) for Chinese L2 Essays across Four CEFR Proficiency Levels

	Bloom	Chinese-LLaMA	Taiwan-LLM
A2	3.45(0.37)	3.06 (0.39)	1.64 (0.18)
B1	3.63(0.33)	3.16 (0.28)	1.59 (0.17)
B2	3.67(0.45)	3.18 (0.35)	1.51 (0.17)
C1	3.17(0.62)	2.85 (0.38)	1.34 (0.25)