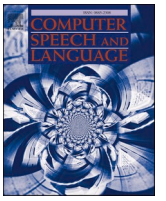
Computer Speech and Language 89 (2025) 101700

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl



Demystifying large language models in second language development research

Yan Cong

School of Languages and Cultures, Purdue University, West Lafayette, IN 47907, USA

ARTICLE INFO

Keywords:
Large language models
Natural language processing
Automatic essay scoring
L2 writing development
L2 interlanguage
Bilingualism

ABSTRACT

Evaluating students' textual response is a common and critical task in language research and education practice. However, manual assessment can be tedious and may lack consistency, posing challenges for both scientific discovery and frontline teaching. Leveraging state-of-the-art large language models (LLMs), we aim to define and operationalize LLM-Surprisal, a numeric representation of the interplay between lexical diversity and syntactic complexity, and to empirically and theoretically demonstrate its relevance for automatic writing assessment and Chinese L2 (second language) learners' English writing development. We developed an LLM-based natural language processing pipeline that can automatically compute text Surprisal scores. By comparing Surprisal metrics with the widely used classic indices in L2 studies, we extended the usage of computational metrics in Chinese learners' L2 English writing. Our analyses suggested that LLM-Surprisals can distinguish L2 from L1 (first language) writing, index L2 development stages, and predict scores provided by human professionals. This indicated that the Surprisal dimension may manifest itself as critical aspects in L2 development. The relative advantages and disadvantages of these approaches were discussed in depth. We concluded that LLMs are promising tools that can enhance L2 research. Our showcase paves the way for more nuanced approaches to computationally assessing and understanding L2 development. Our pipelines and findings will inspire language teachers, learners, and researchers to operationalize LLMs in an innovative and accessible manner.

1. Introduction

Writing is one of the most important and widely studied topics in language learning. The effective evaluation of L2 (second language) writing informs not only learners, but also teachers and researchers, on where to improve. However, manually scoring essays can be time-consuming and can lack consistency. The introduction and growing importance of the transformer large language models (LLMs), such as ChatGPT, has led to major advancements in natural language processing (NLP). Studies have shown that there are shared computational principles for language processing in humans and LLMs (Goldstein et al., 2022). These models are also known for their ability to generate coherent and contextually relevant text, frequently exceeding earlier NLP benchmarks. How they can enhance L2 research is still a mystery. Despite LLMs' broad popularity in various domains, such as law, medicine, and education (Bommasani et al., 2021), the extent to which these LLMs can be used to better our understanding of L2 writing development is a relatively untapped question. In this study, we intend to address this inquiry by utilizing LLMs in automatic essay assessment and indexing L2 development stages. With a focus on measuring and understanding Chinese L2 learners' English writing proficiency, our investigation leverages

E-mail address: cong4@purdue.edu.

https://doi.org/10.1016/j.csl.2024.101700
Received 29 December 2023; Received in revised form 11 July 2024; Accepted 24 July 2024
Available online 26 July 2024

state-of-the-art computational techniques in L2 research. Through constructing, optimizing, and evaluating a novel LLM-based NLP pipeline, this study bridges the gap between some aspects of deep learning—namely, the features contained in LLMs, the development of L2 interlanguage systems, and the automatic assessment of learner proficiency.

We introduce LLM-computed Surprisal, which mathematically represents the negative log-probability of a word sequence given previous contexts as calculated by LLMs. Conceptually, LLM-Surprisal indicates the "surprisingness" and predictability of a word sequence given previous words (Hale, 2001; Levy, 2008; Willems et al., 2015; Tunstall et al., 2022; Wolf et al., 2020; Misra, 2022). We propose that low Surprisals numerically represent highly predictable and, hence, unsurprising, natural, and fluent text. LLMs' Surprisals are exemplified below (Misra, 2022; Xiang & Kuperberg, 2015):

(1) The keys to the cabinet are on the table. [GPT2 Surprisal 38.88].
(2) The keys to the cabinet is on the table. [GPT2 Surprisal 42.76].
(3) Olivia bought a German shepherd. The dog was docile and friendly. However, it bit her hand. [GPTNeo Surprisal 6.38].
(4) Olivia bought a German shepherd. The dog was unpredictable and violent. However, it bit her hand. [GPTNeo Surprisal 9.77].

The two sentences in (1,2) differ only in subject–verb number agreement, where (1) observes the agreement and, thus, (1) is less "surprising" in LLMs' calculation, whereas (2) violates such grammatical agreement, hence it is more surprising. Reflected in LLMs' Surprisals, (2) gave rise to higher GPT2 Surprisal than (1). Surprisal not only captures syntactic grammaticality, as shown in examples (1,2), but also characterizes semantic plausibility. Both (3) and (4) are grammatically correct. But compared to example (3), (4) is less plausible and more "surprising" in LLMs' calculation, because there is no contrastive relationship between the dog being violent and the dog biting her hand. "However" is inappropriate in (4), relative to (3). Although both (3) and (4) are syntactically grammatical, (3) is semantically more plausible than (4). Translated into Surprisals, GPTNeo assigned a higher score to (4). Overall, these examples suggest that Surprisal is a composite index incorporating more than one aspect of natural language.

We propose that Surprisals as shown in examples (1-4) can help demystify LLMs in L2 development research. Many insightful works on LLMs highlight on serving engineering purposes, such as automatic essay scoring and computer-assisted learning and teaching. There is a critical need to go beyond that, to systematically investigate different kinds of LLMs' roles in capturing L2 development and how they relate to L2 research, both empirically and theoretically. We argue that LLMs can enhance L2 research in quantitatively refining theories and comparing ideas to strong alternatives. As an approach based on probability, gradient methods, and neutral networks, LLMs are not only useful for downstream software development, fulfilling engineering tasks—more importantly, they can also derive linguistically meaningful measures, which L2 researchers can tailor for answering specified L2 research inquiries, improving our grasp of interlanguage development and bilingualism.

Our investigation consists of four components: (i) establishing the LLM-Surprisals comparison baseline (native/first language speakers (L1) versus L2), (ii) examining the sensitivity of LLM-Surprisals metrics to L2 subgroups with different proficiency levels, (iii) analyzing the effectiveness of LLM-Surprisals in predicting L2 learners' overall and writing proficiency, and (iv) decomposing LLM-Surprisals from L2 lexical and syntactic perspectives. We carefully selected and automatically compiled two (written) American English datasets from publicly available corpora, and we developed and implemented an LLM-based NLP pipeline to process and analyze the datasets.

Our questions and investigation have significance. Neural models such as LLMs often perform better than the classic feature-based NLP approaches; however, they are obscure and not easy to interpret. This lack of interpretability becomes a significant issue in high-stakes situations, such as L2 writing studies, making LLMs not reliable in scientific discovery. Traditional methods, although do not necessarily suffer from opacity or interpretability issues, are often labor-intensive and inconsistent. Our investigation addresses these gaps by leveraging recent LLMs to provide more efficient, reliable, and consistent L2 writing indices. Crucially, our examination is conducted through robust and thorough comparisons with the existing methods, metrics, and frameworks, with an attempt to improving LLMs' transparency based on our knowledge of the classic L2 indices.

Our findings show that relative to the classic, widely used NLP indices, LLMs are informative of L2 inquiries and innovative in L2 proficiency assessment: LLM-derived Surprisals metrics can effectively detect L2 writing, differentiate learners' adjacent proficiency levels, and advance the classic indices in predicting the overall and writing proficiency test scores provided by human professionals. Further, our results suggest that LLM-Surprisals can capture the interface of lexical semantic diversity and syntactic complexity in L2 interlanguage development. We hope that our LLM-based approach can inspire L2 researchers to build their own research programs using LLM-derived metrics. LLMs are not merely useful in developing software utilities or making downstream L2 learning tasks more marketable; we propose that LLMs are sufficiently precise, formal, and accessible to be implemented by L2 researchers to answer L2 development and bilingualism research questions.

2. Background

2.1. LLM-computed Surprisals

LLMs-computed Surprisal score is widely used in computational psycholinguistics, in fact, it is arguably one of the most interpretable scores in human cognition contexts (Futrell et al., 2019; van Schijndel & Linzen, 2018; Wilcox et al., 2018; Michaelov et al., 2024; Michaelov and Bergen, 2022, 2023; Ryu and Lewis, 2021; Cong et al., 2023). There are studies investigating the extent to which LLM-Surprisals are sensitive to linguistic phenomena that have been shown to influence human sentence processing. For instance, Misra et al. (2020) aimed at reproducing human semantic priming effects using BERT word predictions: They suggested that BERT

predicted a word with lower Surprisal values, when the context included a related than an unrelated word. Michaelov and Bergen (2022) studied LLMs and the collateral facilitation effect. Facilitation refers to the scenario that anomalous words in a sentence are comprehended with more ease by humans because the context has semantically related words. They compared the Surprisals computed by a number of LLMs; their findings suggested that LLM-Surprisals showed sensitivity to the differences between conditions observed in human behaviors. Michaelov et al. (2024) computed LLM-Surprisals to examine the effect of context in reducing the N400[1] amplitude for word incongruency, using Dutch stimuli from Nieuwland and Van Berkum (2006). In a similar study, Ryu and Lewis (2021) suggested that the Surprisal values computed by GPT2 predicted the facilitatory effects of interference in ungrammatical sentences.

There are mixed findings in Surprisals-related studies. Shain (2024) used naturalistic reading datasets, self-paced reading, maze tasks, and eye tracking in their investigation of Surprisal theory. They suggested that GPT2 strongly correlates with human measures of language processing, both behaviorally and neurologically. Interestingly and perhaps unexpectedly, their findings indicated that larger LLMs do not necessarily lead to a better fit to reading times. In fact, in terms of aligning Surprisals with human reading times, GPT2-small outperforms GPT3-davcin-002 and other larger LLMs with instruction tuning, hypothetically because instruction tuning (such as reinforcement learning with human feedback) is likely to contaminate next-word predictability in LLMs. Shain et al. (2024) provided reading time evidence that GPT2-small best estimates human subjective Surprisals, showing a predictive fit to human reading times. They concluded that compared to GPT2-small, *n*-gram seems too constrained, whereas GPT3 is too powerful. Further, Huang et al. (2024) examined the hypothesis that a word's Surprisal computed by LLMs can be linearly mapped onto human reading times and processing difficulty; they concluded that LLM (LSTM and GPT2-small)-computed Surprisals failed to account for syntactic processing difficulty. Bearing in mind the ongoing debate and discussions, for the computation of LLM-Surprisals in this study, we plan to include multiple GPT- and non-GPT-type LLMs with different sizes in tracking and indexing L2 interlanguage development and assessment. The goal is to systematically examine how LLMs with different scales and architectures influence Surprisals' efficacy in an L2 setting.

LLM-Surprisals are also used in acoustic and phonetic studies. Kakouros et al. (2023) studied Surprisals as a feature to aid speech synthesis prosody; their findings indicate that Surprisals can help a speech synthesizer, with a minimal improvement compared to the baseline. In this study, in order to delimitate the scope, we focused on analyzing Surprisals in text, and our calculation of Surprisals was aggregated over textual paragraphs. It is worth mentioning that Surprisals calculation can be operationalized using other units, such as phones, utterances, words, phrases, clauses, sentences, and so on. For example, Surprisal of a word has been found to correlate with behavioral metrics of real-time processing difficulty, such as reading time (Smith and Levy, 2013), and neurolinguistic metrics, such as the N400 (Frank et al., 2015). We focused on paragraphs because LLMs showed promising sensitivity to long text (Tunstall et al., 2022), and with paragraph, it is computationally meaningful and feasible for us to operationalize as a unified measurement unit, across different indices and LLMs.

2.2. Classic indices in L2 writing development

L2 development in writing production is multidimensional and multifaceted. A lot of work has been devoted to establishing effective metrics that can characterize and predict L2 learners' developmental trajectories (Egbert, 2017; Kyle and Crossley, 2017; Lu, 2017; Zhang and Lu, 2022; Ouyang et al., 2022; Lu, 2011, 2012; Yang et al., 2015). Computational approaches to L2 writing development have been focused on aspects such as complexity, accuracy, and fluency. They characterize various production units, ranging from words and phrases to larger syntactic units such as clauses and sentences (Crossley et al., 2015, 2018; Kyle and Crossley, 2015; Polio, 2001; Narcy-Combes, 2003; Ortega, 2003, 2012).

Pertaining to our work, Lu (2010) introduced the L2 Syntactic Complexity Analyzer (L2SCA), which utilizes 14 traditional metrics for modeling and measuring linguistic fluency and complexity, such as length of T-units (the minimally terminable unit, e.g., a sentence), degree of phrasal complexity, and so on. Kyle and Crossley (2017, 2018) developed the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC). TAASSC includes 190 usage-based metrics of syntactic sophistication, 31 fine-grained metrics of clausal complexity, and 132 metrics of phrasal complexity. L2SCA and TAASSC are among the most widely used classic L2 writing development analysis tools. Further, lexical diversity metrics such as contextual distinctiveness and lexical semantic diversity (LSD) have been shown to effectively capture L2 development in lexical proficiency (Berger et al., 2017; Hoffman et al., 2012; Kyle et al., 2018). A word's LSD refers to the degree to which different contexts are associated with a given word in its meanings. Words with a high LSD score tend to appear in many contexts; and they are not contextually distinctive. Berger et al. (2017) found that LSD plays a role in the L2 development of lexical proficiency: More proficient speakers showed higher LSD, suggesting that they use semantically diverse words, words that occur in a broad range of semantic contexts.

Related to our work, other lexical indices such as type–token ratio and *n*-gram association strength have proven to be valid in benchmarking L2 development and characterizing L2 proficiency (Bulté and Roothooft, 2020; De Clercq, 2015; De Clercq and Housen, 2016; Kettunen, 2014; Kim et al., 2017; Kyle and Crossley, 2017; Treffers-Daller et al., 2016). For L2 development at the phrase level, Bestgen and Granger (2014) quantified the structure, meaning, and use of "word combination" in L2 English writing; they attempted to use a longitudinal study to validate whether learners' phraseological competence develops over time, and to examine how phraseological competence indices relate to human raters' assessments of writing quality. Their results indicated that phraseology is a key aspect of L2 writing development. Relatedly, Paquot (2017) intended to define and circumscribe phraseological complexity. By

[1] N400 is a negative-going waveform that peaks around 400 milliseconds post-stimulus onset. It is part of the normal brain response to language stimuli (Luck, 2012).

comparing measures of phraseological complexity with the earlier indices of syntactic and lexical complexity, Paquot (2017) examined the extent to which phraseological complexity indices can be used to describe L2 performance. They found that pointwise mutual information (MI)-based measures are especially useful to characterize L2 writing development across proficiency levels. Going beyond the phrasal level, Wilson et al. (2017) examined L2 writing development at three levels: word choice, sentence syntax, and discourse cohesion; they suggested that a subset of nine Coh-Metrix measures were able to model each level. Relatedly, Shin and Gierl (2021) made comparisons of L2 writing indices' efficacy between two systems: a support vector machine model with Coh-Metrix features, and CNNs (convolution neural networks). Their results indicated that CNNs outperform the Coh-Metrix model, in the sense that they showed better alignment with human raters. Lan et al. (2019) used metrics such as grammatical complexity to understand and benchmark L2 writing development. NLP utility *Stanza* was used to tag part of speech (POS) and grammatical features.

Previous computational studies targeting Chinese learners' L2 English development provide insights from various perspectives. Using college-level L2 writing from the *Written English Corpus of Chinese Learners* (Wen et al., 2005), Lu (2010) developed a computational system that can take an essay and output complexity indices. The results suggested that such indices can effectively differentiate proficiency levels, hence indexing L2 development trajectories. From an information-theoretic perspective, Wang et al. (2022) investigated the extent to which Kolmogorov complexity metrics of morphological complexity can tease apart L2 learners' proficiency levels and developmental stages; they used argumentative writings produced by Chinese L1 English L2 learners to examine Kolmogorov complexity metrics' ability to index proficiency levels. Wang et al. (2022) compared the new Kolmogorov metrics with the classic NLP ones calculated by L2SCA and TAASSC. The results suggested that such metrics are distinct from the classic metrics, and they can distinguish proficiency levels. Crucially, the findings indicated that, as L2 learners' interlanguage system develops and proficiency improves, their writings' complexity increases.

Despite a large amount of NLP indices proposed for (Chinese L1 English L2) writing development, there are some aspects that have not yet been fully characterized: First, LLM-Surprisal as an index of L2 development in writing: How LLM-Surprisal differs from the widely used measures such as complexity, and how Surprisals can describe L2 performance and trace L2 development; Second, systematic examination of LLMs: Compared to the previous NLP systems, LLMs are known for being holistic, multi-purpose, and adaptable (Bommasani et al., 2021), but there is still relatively little discussion as to exactly how LLMs with different architectures and different pre-training scales can better our understanding of L2 writing development. Here, we hope to fill these research gaps by examining Surprisals as calculated by different LLMs in L2 English writing, and to further relate Surprisals to L2 writing development.

2.3. LLM-related L2 writing assessment

The current study on L2 is conducted against the background of Automatic Essay Scoring (AES). AES systems, such as *e-rater* developed by ETS (Educational Testing Service, https://www.ets.org/), primarily relied on NLP techniques and statistical models to assess writing quality (Attali & Burstein, 2006). AES systems can outperform traditional methods in evaluating the linguistic complexity and coherence of L2 essays (Perelman, 2020). Aside from identifying specific language issues common among L2 learners, such as grammatical errors and inappropriate word usage, AES can provide feedback for improvement (Y.-J. Lee, 2020). Chen and Pan (2022) conducted a comparative study of AES and instructors' feedback, using Chinese college students' English writing. The results suggested that *Aim Writing*, a writing facilitator developed by Microsoft Research Asia, needs to be used in conjunction with instructor's feedback in writing, in order to support all students' needs. Liu and Kunnan (2016) conducted a case study on AES software "WriteToLearn" and its implementation among Chinese undergraduate English majors. The results indicated that this utility is more consistent but also more stringent than manual grading, and it failed to reliably detect writing errors. Similarly, Liu et al. (2016) analyzed teacher comments on 327 English major students' writing and conducted AES-based automatic feedback classification. The results implied that the proposed approach is feasible, and system-generated feedback can be useful. Related studies and reviews have been conducted in Weigle (2013).

Recent studies in AES show a significant shift towards LLMs. For example, Xiao et al. (2024) investigated the efficacy of GPT4 and fine-tuned GPT3.5 in AES, suggesting that LLMs can not only automate the assessment but also enhance human graders' performance. Mizumoto and Eguchi (2023) used prompt engineering in automated writing evaluation. Specifically, OpenAI's text-davinci-003 was used to grade English essays. The results indicated that integrating GPT LLMs could provide consistency and efficiency for essay grading. Ludwig et al. (2021) compared LLMs with a logistic regression model based on bag-of-words (BOW) in AES. Their dataset contained 2088 emails in German. Lee et al. (2024) used GPT3.5 and GPT4 for automatic scoring; they utilized several prompt engineering strategies, in combination with zero-shot or few-shot learning with chain-of-thought reasoning. Ormerod et al. (2021) used transformer-based LLMs for AES. Their results refute the paradigm in NLP that bigger LLMs lead to better assessment accuracy: They achieved decent results using LLMs with fewer parameters than most pre-trained LLMs.

Despite the advanced capabilities, LLM-based AES systems are not without challenges. One significant concern is the lack of transparency in LLM decision-making processes, which poses difficulties for educators who seek to understand and interpret the scores provided by these systems (Kumar & Boulanger, 2021; Kumar et al., 2023). Also, reliance on automated systems must be balanced with human oversight to ensure that the feedback is contextualized and pedagogically sound. As such, the role of educators remains essential in guiding and supporting L2 learners in their writing development. Recent work by Schneider et al. (2023) emphasized the importance of teacher-AES system collaboration to maximize the educational impact.

Further, a few more challenges and gaps have been identified in Ramesh and Sanampudi (2022). They conducted a systematic literature review on the AES, summarizing that most studies focus on multiple-choice questions, but few focus on content-based essay evaluation. For content-based L2 writing studies, Crossley and Holmes (2022) contrasted transformer LLMs' performance with that of more traditional NLP approaches. They adopted LLMs' semantic embeddings to model the receptive vocabulary of L2 learners in a

writing task. The results indicated that LLMs' embeddings showed better performance than the static model word2vec in making predictions about L2 learners' vocabulary scores. Additionally, for content grading purposes, Bexte et al. (2022) introduced an architecture that uses a BERT-based approach to learn a similarity model. Relatedly, Takano and Ichikawa (2022) showed a BERT-embedding-based model for the generation of short-answer questions. For the existing content-based LLM-inspired AES studies, as far as our knowledge goes, there is no systematic investigation on different LLMs-computed Surprisals as interpretable L2 indices in AES.

We hope to bridge these gaps through an interpretable content-related AES measure: LLM-Surprisal, modeling the process of how L2 writing incrementally develops and improves. LLM-based AES systems represent a significant advancement in the field of AES, offering enhanced capabilities for assessing L2 learners' writing. While these systems show great promise, addressing issues of transparency, and appropriate integration into L2 and general educational research and practices is essential for maximizing their potential benefits. Here, to improve LLM-computed Surprisals' interpretability in L2 research and teaching, we carried out our analysis based on (1) thorough comparisons with the classic indices in L2 writing development; (2) systematic examinations of different kinds of LLMs with various pre-training mechanisms; (3) carefully interpretation of Surprisals from the aspects of content-related constructs, including lexical diversity and syntactic complexity.

2.4. Current work

The current work investigates questions from the aspects of computation and L2 learning theory. Computationally, as a utility, how effective are LLMs in an L2 context? We provided three experiments to approach the question: LLM-Surprisals detect L2 writing, index L2 proficiency levels, and predict L2 proficiency scores given by human professionals. Further, for differences in LLMs, we predict that scale matters, and larger-scale LLMs are expected to give better results. Decoder LLMs (GPT-type) should outperform encoder models (BERT and T5), because Surprisals are derived from next-word prediction, which aligns better with the decoder LLMs' pre-training paradigm (Tunstall et al., 2022). Theoretically, from the perspective of L2 learning, what does Surprisal characterize in the context of L2 development? Drawing significant insights from previous studies (Michaelov and Bergen, 2022; Berger et al., 2017; Michaelov et al., 2024; Bulté and Housen, 2012, 2014; Dahl, 2004; Paquot, 2017; Rezaii et al., 2023), we propose that, in an L2 context, the Surprisal of a text is a multifaceted measure—a gradient value that involves both lexical diversity and syntactic complexity. LLM-Surprisals can be decomposed as a numerical representation of the interplay between those two factors. In general, we hypothesize that LLM-Surprisals are higher in relatively low-proficiency L2 writing, and Surprisals decrease as their writing improves. In particular, there are three primary sub-hypotheses: First, we expect to see high lexical semantic diversity in L1 and in proficient L2 learners, as their vocabulary mastery improves. Second, we expect to see high syntactic complexity in L1 and L2 with higher proficiency, because their grammatical knowledge improves. Third, we propose that higher lexical semantic diversity scores and higher syntactic complexity scores together lead to a decrease in LLM-Surprisals, hence indicating less-surprising writings and a more developed L2 interlanguage system.

There are critical motivations for us to analyze LLM-derived Surprisals, aside from their recent introduction within the context of Artificial intelligence (AI) development. The classic indices, such as clausal complexity and sentence length, have been examined and validated in various ways. They might be arguably more transparent and informative of L2 interlanguage development than values calculated by LLMs, which seem "blackboxy" and opaque. We propose that Surprisals can supplement and advance the existing indices in two aspects. First, Surprisal captures and processes dynamic context. The context window length of a recent LLM called *Llama2* is 4096 tokens; this is approximately equivalent to six pages of English words, suggesting that recent LLMs can memorize up to four thousand tokens during text generation and other inferencing tasks (Touvron et al., 2023). This increase in knowledge base and contextual understanding enables greater complexity and a more fluent exchange of natural language. Surprisal derived from such contextually "aware" LLMs leads us to believe that it deserves attention to capture long-text L2 writing. Second, Surprisal is a composite and holistic index. Surprisal changes can occur across disfluencies in different language components (syntax; semantics) with different measurement units (lexical; sentential; paragraph). In contrast, for many of the existing feature-based NLP indices, each index characterizes one aspect of L2 proficiency. For example, the clausal complexity measures capture clausal syntax but do not necessarily reflect its semantic cohesion. It can accurately count how many embedded clauses there are in a sentence, but such a number is barely informative of the sentence's semantic plausibility, lexical proficiency, or overall text surprisingness and naturalness. To comprehensively and holistically examine L2 development, hundreds of related or unrelated indices need to be calculated and interpreted separately. This might cause inconsistency, inefficiency, and multicollinearity, for both utility operation and scientific measurement. The advantage of employing Surprisal is that it synthesizes multiple isolated indices into a single measure, which is a uniquely and naturally human-cognition-inspired index (c.f., Section 2.1). We maintain that these characteristics provide flexibility for Surprisals in complementing and enhancing the classic indices.

3. Methods

3.1. Data

3.1.1. L2 writing data

We used the publicly available *University of Pittsburgh English Language Institute Corpus* (PELIC) (Naismith et al., 2022), a large learner corpus, including both written and spoken texts. Data collection was conducted in an *English for Academic Purposes* (EAP) context over seven years in the University of Pittsburgh's Intensive English Program. Texts were produced by English learners, who has

a broad range of linguistic backgrounds and various levels of proficiency. Proficiency levels in PELIC range from level 2 (pre--intermediate), approxinately corresponding to the *Common European Framework of Reference for Languages* (CEFR, 2001) A2/B1, to level 5 (advanced), which is approximately equal to CEFR B2+/C1. Level 3 (intermediate) is approximately equal to B1, and level 4 (upper intermediate) to B1+/B2. These levels are provided by PELIC in the level identifier (level_id), which indicates the speaker's proficiency level at which the text was produced.

We selected L2 English learners whose L1 is Chinese to study Chinese L1, English L2 learners' interlanguage system. Chinese is among the top five most common L1s in PELIC (Naismith et al., 2022). Moreover, data processing and compilation indicated that this group has the most complete and detailed documented demographic information. We focused on "allow_text" tasks, which allow students to write an answer instead of choosing a word from a word bank. To maintain consistency in text content and length, and to reduce confounding factors caused by different tasks or question types, only the "paragraph_writing" question type was selected. This includes questions such as "*In five sentences or less, give instructions on how to make tea.*" or "*Describe a person that you know well. What does he or she look like? Write one paragraph, between 7 and 12 sentences. You can use some of the words on page 71 of "Refining Composition Skills.*", and so on.

In this study, the text input for the LLM-based NLP pipeline was a *paragraph*, a long-text preprocessed response produced by a participant to a "paragraph_writing" question. All the computations were aggregated over a paragraph. Taking paragraphs as text inputs ensures that the text length is controlled and matched across L1 and L2 corpora. Also, LLMs can handle long-text paragraphs well (c.f., Section 2). Sentence boundaries were automatically detected based on the punctuation ".", "!", "?", and "...". Direct quotes were excluded from each paragraph. Minimal cleaning was conducted to remove non-characteristic, non-numeric symbols, and redundant spaces, and to keep the raw content. Overall, a total of 297 samples produced by 113 Chinese L2 learners of English were selected and preprocessed. For each proficiency level, we randomly sampled 99 paragraphs matched in demographics and text length, using the *MatchIt* function in R (Ho et al., 2011). There were significantly fewer data from level 2 learners (7 samples in total). According to Naismith et al. (2022), this is because the English Language Institute did not regularly offer level 2 when they collected the data. Therefore, we excluded level 2 learners' writings from our analyses.

Each of the 297 writing samples has its corresponding test scores, provided by PELIC. Regarding specific tests, we selected "Writing_Sample", which includes in-house writing test scores on a scale of 1-6, with 6 representing the most proficient writing. The writing was assessed by at least two certified human raters, provided by the *Examination for the Certificate of Proficiency in English* (ECPE). We also included one of the *Michigan Test of English Language Proficiency* scores "MTELP_Conv_Score" as a measure of L2 learners' overall combined proficiency. To sum up, we considered level_id, Writing_Sample, and MTELP_Conv_Score as the response variables, and level_id and Writing_Sample as the gold standard variables, since they are assessed and provided by human professionals. Table S1 in the Appendix details the basic descriptive statistics for Writing_Sample and MTELP_Conv_Score.

3.1.2. L1 writing data

We argue that it is necessary to establish an interpretation baseline, examining LLMs' performance in basic detection tasks. Thus, we included an L1 corpus from the *Michigan Corpus of Upper-level Student Papers* (MICUSP), which was developed at the English Language Institute of the University of Michigan. MICUSP is a historically important collection of language data for linguistic analysis and materials development. By including L1 writing data, the hope is to justify the extent to which LLMs' Surprisal metrics can distinguish L1 from L2 writings. Building on top of that, we can further investigate how to operationalize LLMs in understanding L2 development stages and indexing subgroups among L2 learners.

We randomly selected ten essays produced by ten final-year undergraduate students who are native speakers of English and whose major is English. Ten essays were selected because, after slicing, sampling, and matching, ten essays can give us a matched set of paragraphs samples. A *paragraph* in the L1 writing data is a long text that can stand alone as a response or a response component to a writing assignment. We sliced each essay into short paragraphs, so that the text length (total number of words) aligned with the L2 dataset. Same cleaning and matching procedures used in the L2 dataset were applied to the L1 data. In total, preprocessing ten essays gave us 99 matched samples. Table 1 details the text length information for the L1 and L2 datasets used in this study.

As illustrated in Table 1, after sampling and matching, there were 99 paragraphs from each group included for analyses. The selected L2 paragraphs were drawn from 297 samples produced by L2 learners in a paragraph-writing task. The selected L1 paragraphs were drawn from ten essays produced by L1 speakers.

It is worth noting that we chose MICUSP as our L1 dataset because MICUSP is relatively suitable in terms of content, size, and writers' diverse academic background, making it comparable to L2 speakers' writing. MICUSP offers immediate and free access to a set of 830 top-quality papers (totaling around 2.6 million English words) authored by University of Michigan graduate and senior undergraduate students across 16 different disciplines spanning major academic fields (Römer and Swales, 2010). MICUSP illustrates the writing style of proficient writers (Hardy and Römer, 2013) and has proven to be a valuable resource of genuine language that educators and researchers can utilize to create language-based activities and materials (Cobb and Boulton, 2015). There are other publicly available corpora, including the *Corpus of Contemporary American English* (COCA), the *Michigan Corpus of Academic Spoken*

Table 1
Descriptive statistics of the text length for L1 and L2 datasets. *N* stands for the number of paragraphs.

	Mean	Min	Max	Std	N
L2 dataset (PELIC)	141	56	299	58.13	99
L1 dataset (MICUSP)	170	55	321	63.01	99

English (MICASE), and the *British National Corpus* (BNC). However, they do not fit in with our research paradigm due to the considerations of text length, content, and genre. For instance, the Wiki sample from COCA could be an appropriate L1 corpus, since it is written American English and is sufficiently short. But its content can be too technical to match the L2 PELIC corpus, which might affect LLMs' performance (Tunstall et al., 2022). COCA has a movies corpus that is publicly accessible, yet it is too interactive and lengthy: one passage can contain over 1,000 words. MICASE could be a good match in terms of content, but it is all spoken English. BNC contains short written English. Given the difference between the British and American varieties of English, and the fact that all PELIC L2 learners learn English in a U.S. institution, we did not choose BNC.

3.1.3. Validation data

We additionally validated the L1-L2 writing detection task with a separate dataset. We are aware that despite MICUSP and PELIC sharing similarities and the lack of a more appropriate L1 corpus, MICUSP and PELIC are not directly comparable, since they are two separate corpora built and maintained by two institutions. Therefore, we validated our findings using an additional corpus *CROW* (Corpus and Repository of Writing, Staples and Dilger, 2018), which includes both L1 and L2 English essays. CROW is an English corpus with 17,839 samples of college writing from both L1 and L2 participants, containing 17,823,912 words. CROW also provides detailed demographic information about the speakers. Texts have been collected longitudinally from the fall semester of 2009 through the spring semester of 2021. CROW allows filtering by country, gender, assignment type, date, and TOEFL (Test of English as a Foreign Language) scores. The following is a count of files by type of assignment: argumentative paper, 1429; reflection, 609; rhetorical analysis, 330. Count of files by draft: final, 1690; first, 677.

For validation purposes, we took final drafts only, and we extracted all three types of assignments to balance potential writing style and content differences. The same cleaning used in the L1 and L2 datasets was applied to the validation CROW dataset, and we additionally cleaned references and bibliographies. We randomly sampled 37 L1 paragraphs and 37 L2 counterparts whose L1 is Chinese, which is the maximal amount of L1 Chinese samples we can obtain from CROW, after applying the same matching procedures used in the other two datasets. Note that this CROW dataset was only used for validation. For all the other analyses, we still used PELIC as the L2 dataset and MICUSP as the L1 dataset.

3.2. Indices

3.2.1. New indices

To operationalize the LLM-derived Surprisals metrics, we used *minicons* (Misra, 2022), an open-source utility that provides a standard API (application programming interface) for behavioral analyses of LLMs. We used LLMs to tokenize the text. A "token" can be viewed as a subword; for instance, a text sequence "walked" would be tokenized by an LLM into two tokens: "walk" and "ed". Word-level Surprisals are obtained by taking the average of subword tokens' Surprisals, as described by Misra (2022). Mathematically, Surprisal is defined as the negative logarithm of the probability of a word sequence given its preceding context as calculated by LLMs. Formally, following Misra (2022) and Misra et al. (2020), the Surprisal of the target word w_t in the context $w_{1\ldots t\text{-}1}$ was computed as (1).

$$Surprisal(w_t) = -logP(w_t|w_{1\ldots t-1}) \tag{1}$$

In formula (1), $P(w_t|w_{1\ldots t-1})$ is the conditional probability of the word w_t given the context $w_{1\ldots t-1}$. In LLMs' pre-training, they learn the statistical relationships between words and their contexts, allowing them to estimate $P(w_t|w_{1\ldots t-1})$ (Tunstall et al., 2022). This pre-training mechanism is built up on bigrams or trigrams models, where $P(w_t|w_{1\ldots t-1})$ is calculated based on the frequencies of w_t occurring after the preceding one or two words (Jurafsky & Martin, 2024). Transformer LLMs can handle longer and more complex contexts than *N*-gram models (Jurafsky & Martin, 2024). Such LLMs are pre-trained on large datasets to predict the probability of the next word in a sequence, effectively learning $P(w_t|w_{1\ldots t-1})$.

Derived from formula (1), we computed median Surprisal of the paragraph, which is the summation of each word's Surprisal in the paragraph, normalized by the total number of tokens in the paragraph. We computed medians instead of means in order to account for outliers' impact on the Surprisals metrics. Word features such as frequency, difficulty, and predictability can give rise to exceptionally high Surprisals. Studies have shown that there is a longitudinal decrease in L2 writing in the use of high-frequency word collocations that are less typical in L1 writing (Bestgen and Granger, 2014). These collocations in L2, made up of vocabulary that is less frequently used in L1 writing, can lead to outstanding Surprisal scores. This motivated us to calculate median instead of mean Surprisals. Table S2 in the Appendix details the descriptive statistics for each Surprisal metric derived from LLMs, for both the L1 and L2 datasets.

3.2.2. LLMs selection

In this study, we compared encoder LLMs (e.g., BERT-large-uncased; Devlin et al., 2018) and decoder LLMs (e.g., GPT2; Brown et al., 2020). Encoder LLMs have only the encoder part of the transformer. They are pre-trained in a paradigm called "masked language modeling", which is a pre-training task where the model sees corrupted texts that are generated by covering tokens randomly, and the model predicts the original text (Tunstall et al., 2022). Encoder LLMs are also called bidirectional LLMs, since the prediction is based on both the left- and right-hand sides of the masked token. On the other hand, decoder LLMs only have the decoder transformer and use "causal language modeling", a pre-training task where the model reads the text in sequential order and needs to predict the next word (Tunstall et al., 2022). Decoder LLMs are also called unidirectional LLMs, since the prediction is based on only the left-hand side of the current token. Compared with bidirectional LLMs, the pre-training task in unidirectional LLMs aligns better with the next-word Surprisal operation (Tunstall et al., 2022; Michaelov et al., 2024). We additionally included T5-large (Raffel et al., 2020), which

integrates both the encoder and the decoder parts of the transformer architecture. This selection exhausted all three major types of transformer LLMs that are currently publicly available and widely popular (Tunstall et al., 2022; Wolf et al., 2020).

To examine how scaling would influence LLMs' capacity, we included three variants of decoder–transformer LLMs of different sizes: GPT2, with 124 million parameters (Radford et al., 2019); DistilGPT2, with 82 million parameters (Sanh et al., 2019), trained as a student network with the supervision of GPT2; and GPTNeo, with 1.3 billion parameters (Black et al., 2022; Gao et al., 2020), which is an open-source LLM developed by EleutherAI. It is similar in scale to the smaller models in the GPT-3 family, such as the smallest GPT-3 model (which has 125 million parameters) and GPT-3 medium (which has 350 million parameters). Given the rapidly evolving landscape of LLMs, we included *text-davinci-003*, with 175 billion parameters (OpenAI, 2023). It should be noted that this was one of the largest and latest unidirectional LLMs that allowed us to compute log-probabilities as of the time when we wrote this manuscript. As to the other types of LLMs, BERT-large-uncased has 336 million parameters, and T5-large has 770 million parameters.

A systematic comparative analysis with various LLMs could provide a broader perspective on the capabilities and limitations of LLMs in an L2 context. We acknowledge that it does not seem feasible to constantly exhaust the latest and the largest LLMs that are out there in the market. We chose this set of LLMs because it represents the major LLM *architectures*. We suggest that gaining a thorough comprehension of LLM structures would empower L2 researchers to select the most suitable LLM initially and expand or upscale it later if needed. By outlining the linguistic nuances of LLMs via an easily accessible NLP pipeline, we aim to clarify the utilization of LLMs in L2 research, improving LLMs' interpretability and transparency.

3.2.3. Classic indices

We examined how much the new LLM-based indices can advance the classic NLP indices in L2 writing. Such advancement was quantified by comparing LLM metrics with the classic NLP indices. It is worth noting that LLM-Surprisals themselves are not necessarily directly comparable to the classic indices, such as mean length of T-units, phrasal complexity, and so on. The new and the classic indices capture different aspects of L2 development, conceptually and mathematically. However, both kinds of indices are potentially useful in automatic essay assessment, and both can inform us on L2 interlanguage development. Unlike the well-studied classic indices such as mean length of T-unit, the new LLM indices lack transparency; hence, we are unsure what LLM-Surprisals are characterizing and how they are related to or distinct from the classic indices. In other words, we are comparing LLM metrics' *efficacy* with that of the classic indices in separating L1 and L2 and indexing L2 proficiency levels. As a byproduct, based on what we know about L2 writing (e.g., complexity, sophistication, fluency), such comparison of efficacy between the well-known and the less-known indices will distinguish what LLM-Surprisals are measuring in an L2 learning context.

Complexity and fluency are widely studied L2 variables that showed efficacy in indexing L2 writing (Bulté and Housen, 2012, 2014; Housen and Kuiken, 2009). Hence, we selected and computed 15 such metrics, which were validated and proved effective in predicting L2 writing quality (Kyle and Crossley, 2018; Zhang and Lu, 2022; Wang et al., 2022). This collection includes six classic syntactic complexity and production fluency metrics from L2SCA (Lu, 2010), along with nine fine-grained clausal and phrasal complexity metrics from TAASSC (version 1.3.8; Kyle, 2016). Concretely, for L2SCA, we adopted all three metrics concerning text length as measures of linguistic productivity and fluency, in addition to the complexity indices, including the amount of subordination, the amount of coordination, and the degree of phrasal sophistication. Previous studies have shown the efficacy of these L2SCA indices in indexing L2 development (Norris and Ortega, 2009; Wang et al., 2022; Ouyang et al., 2022). For TAASSC, we selected seven phrasal and two clausal metrics. This selection was based on previous studies by Wang et al. (2022) as well as Ouyang et al. (2022), which validated that those phrasal and clausal indices are effective in indexing L2 development stages.

Moreover, to address the L2 theoretical question of what Surprisal is capturing, we additionally included a lexical diversity index: the LSD that was proposed and examined by Hoffman et al. (2012) and Berger et al. (2017), using a latent semantic analysis technique (LSA; Landauer and Dumais, 1997). Lexical diversity is associated with a wide variety of measures, for example, word frequency: high-frequency words are likely to show up in diverse contexts (Kyle et al., 2018). Crucially, studies showed that lexical semantic diversity is mostly influenced by the meaning of a word, rather than its frequency (Johns, 2022; Caldwell-Harris, 2021). They examined how semantic uniqueness, word frequency, and contextual uniqueness could forecast human accuracy in identifying spoken words amidst different levels of noise. Results suggested that *semantic* uniqueness had a stronger impact on response times compared to the other factors. This motivated us to focus on lexical *semantic* diversity rather than other lexical diversity features such as frequency when decomposing LLM-Surprisals. In operation, a word's LSD was extracted, and we aggregated each word's LSD over the whole paragraph by taking the mean.

For the L2SCA and TAASSC indices, the automatic text analysis software took a paragraph at a time and output a score for the given paragraph. LSDs were normalized by paragraph length, and any sequence in the paragraph that is not represented in the LSD index database was not counted toward paragraph length, for example, extremely low-frequency words or misspellings. To sum up, all the indices (classic and new) in the current work were calculated and aggregated over a paragraph, to ensure valid comparisons of efficacy across different kinds of L2 writing indices.

4. Analysis and results

Statistical analyses were conducted in R (R Core Team, 2023). Shapiro–Wilk tests on all the independent variables showed p-values smaller than 0.05. An examination of the histogram plots additionally suggested that the data distributions violate the normality assumption for parametric tests. Therefore, we transformed and standardized our data through centering (subtracting the mean) and scaling (dividing by the standard deviation). The standardization was conducted using the *scale*() function in R (R Core Team, 2023). This is also to control for mean values across different indices, and to ensure equal weighting for accurate comparisons. Moreover, we

adopted non-parametric statistical tests throughout this paper. The alpha level in this paper is 0.05.

4.1. Detecting L2 writing

Before examining LLMs' efficacy in differentiating proficiency levels and predicting gold-standard test scores provided by human professional raters, it is critical to investigate these new metrics' performance in basic and fundamental tasks, such as distinguishing L2 from L1 essays. Therefore, first, to establish a baseline, we analyzed L1 and L2 datasets and quantified how effectively LLMs can differentiate L2 from L1 writing. We conducted Welch two-sample *t*-tests to determine whether the difference in the means of Surprisal metrics between two groups was significant. Effect sizes (Cohen's *d*) and significance (*p*-value) are summarized in Table 2.

The findings reveal that some LLMs managed to detect L2 writing. GPTNeo-, T5-, and BERT-large-unbased-derived median Surprisals showed statistically significant effects. Interestingly, BERT and GPTNeo showed negative effects, whereas T5-large showed *positive* effects, suggesting that writings produced by L1 speakers are associated with high T5 Surprisal scores, whereas those produced by L2 speakers are associated with low T5 Surprisals. The inverse pattern was found when using BERT and GPTNeo.

In order to validate whether the difference in Surprisal metrics between the two groups remains significant using a separate dataset[2], we conducted the same LLM-Surprisal calculations and Welch two-sample *t*-tests on the validation CROW dataset. Effect sizes (Cohen's *d*) and significance (*p*-value) are summarized in Table 3.

The results suggested that our main findings are reproduced. With a separate validation dataset, aside from GPTNeo and BERT-large-unbased, DistilGPT2- and GPT2-derived median Surprisals showed statistically significant effects. This indicated the feasibility and reliability of LLMs in detecting L2 writings. We also replicated the directionality of different LLMs: BERT, GPT2, DistilGPT2, and GPTNeo showed negative effects, whereas T5-large gave the inverse again. This further justified that, unlike all the other LLMs, writings produced by L1 speakers are associated with high Surprisal scores when using T5, across different datasets. This makes T5 a distinct and the least interpretable LLM for detecting L2 writing.

4.2. Indexing L2 proficiency levels

How effective are LLM-computed Surprisals in tracing L2 interlanguage development and indexing L2 learners' adjacent proficiency levels, relative to the classic NLP indices? Table 4 provides the results of Kruskal–Wallis tests on the new Surprisals indices and the classic NLP indices. The classic indices include six traditional and nine fine-grained syntactic complexity indices.

We found that the classic indices from L2SCA showed generally stronger effects than the fine-grained ones from TAASSC and the new LLM indices, suggesting that the classic measures of productivity, fluency, and complexity remain robustly informative in indexing L2 development stages. Among all the LLM-Surprisals indices, T5 led to the strongest effect, even stronger than some of the classic text-length indices from L2SCA. In contrast, BERT and the GPT-type LLMs, including text-davinci-003, showed weaker effects. Overall, we found significant effects in both LLM-based metrics and the classic NLP indices. As a follow-up analysis, to pinpoint the specific proficiency levels at which the indices become informative, for all of the adjacent proficiency levels, we conducted pairwise Wilcoxon comparisons on the Surprisals and the classic indices that showed main effects in the Kruskal–Wallis tests. The results are visualized in Fig. 1.

Fig. 1 suggested that all of the classic indices from L2SCA can distinguish whether a writing was produced by level 4 or level 5 L2 learners, and only the mean length of sentence, mean length of T-unit, and dependent clauses per clause were able to distinguish level 3 and level 4 L2 writings. This implies that there are noticeable improvements in fluency and productivity from the upper-intermediate to the advanced level. As for the classic indices from TAASSC, all but *dependents per nominal subject* showed significant differences between level 4 and level 5 L2 writings. This index also turned out to be the only one that can distinguish level 3 and 4 L2 writings. We can thus infer that the complexity increase is more notable from the upper-intermediate to the advanced level than from the intermediate to the upper-intermediate level. Put otherwise, our findings suggested that fluency, productivity, and complexity improvement become more indexable and visible when learners' proficiency arrives at the advanced level (level 5).

For the new LLM indices, we found that T5 was the only LLM that showed statistical differences for all of the pairwise comparisons. BERT managed to distinguish level 4 and level 5 L2 writings, and DistilGPT2 was able to benchmark L3 and L4 L2 writings, while text-davinci-003 did not show significance for groupwise comparisons. It is also worth noting that T5 showed an increase of Surprisal throughout the course of L2 writing development, whereas BERT and DistilGPT2 demonstrated the inverse. This echoes the findings about T5 in L2 writing detection (Section 4.1).

Overall, our findings indicated that as L2 learners' proficiency levels improve, there are increases in productivity and fluency as illustrated in the text length indices, an increase in phrasal complexity, and a decrease in LLM-Surprisals, at least for BERT and DistilGPT2. We can infer that, as L2 learners' interlanguage system becomes more advanced, their writing becomes more and more natural, less surprising, hence decreasing LLMs' Surprisals. This decrease in Surprisals is possibly related to the increase in linguistic production fluency and phrasal complexity in their writing.

[2] Note: We did not include text-davinci-003, because as of the time when we were revising this manuscript (May 2024), the LLMs hosted by OpenAI no longer exposed token-wise log-probabilities for a predicted token.

Table 2
LLM-Surprisals indices for differentiating L1 and L2 English writings.

LLMs	Cohen's *d*	Sig (*p*-value)
BERT	-0.99	0.000
GPT2	-0.275	0.059
DistilGPT2	-0.233	0.109
GPTNeo	-0.33	0.024
text-davinci-003	0.068	0.634
T5	1.23	0.000

Table 3
Comparison of L1 and L2 English writing's LLM-Surprisals, with matched samples from a separate corpus (*CROW*).

LLMs	Cohen's *d*	Sig (*p*-value)
BERT	-1.283	0.000
GPT2	-1.429	0.000
DistilGPT2	-1.237	0.000
GPTNeo	-1.602	0.000
T5	0.105	0.653

Table 4
Efficacy comparisons between the classic and the LLM-Surprisal indices in indexing L2 proficiency levels.

	Index	$\chi^2(2)$	Sig (*p*-value)	Effect size (*eta2 [H]*)
Classic indices: text length from L2SCA	Mean length of sentence (MLS)	46.122	0.000	0.15
	Mean length of T-unit (MLT)	70.71	0.000	0.234
	Mean length of clause (MLC)	8.545	0.014	0.022
	Dependent clauses per clause (DC/C)	64.294	0.000	0.212
	Coordinate phrases per clause (CP/C)	2.023	0.364	0.000
	Complex nominals per clause (CN/C)	18.17	0.000	0.055
Fine-grained Classic indices: phrasal complexity from TAASSC	Dependents per object of the preposition	22.942	0.000	0.071
	Prepositions per object of the preposition	14.687	0.000	0.043
	Adjectival modifiers per object of the preposition	16.229	0.000	0.048
	Average number of dependents per direct object	3.634	0.163	0.005
	Standard deviation of the number of dependents per direct object	0.695	0.707	-0.004
	Standard deviation of the number of dependents per nominal subject	8.72	0.013	0.023
	Adjectival modifiers per nominal subject	0.281	0.869	-0.006
Fine-grained classic indices: clausal complexity from TAASSC	Number of nominal subjects per clause	15.905	0.000	0.047
	Number of adverbial modifiers per clause	1.264	0.532	-0.003
LLM-derived Surprisals indices	BERT	16.822	0.000	0.05
	GPT2	4.975	0.083	0.01
	DistilGPT2	6.736	0.034	0.016
	GPTNeo	4.304	0.116	0.008
	text-davinci-003	6.081	0.048	0.014
	T5	51.733	0.000	0.169

4.3. Predicting L2 proficiency scores

To examine how much LLM-derived metrics can advance the classic measures, we constructed elastic net models using *scikit-learn* (Pedregosa et al., 2011) for the L2 dataset. We used both the classic and the new indices to predict the two gold-standard scores provided or validated by human professionals: L2 overall proficiency (MTELP_Conv_Score) and writing proficiency (Writing_Sample). We chose elastic net models because we had a relatively small dataset and only a handful of features, and these features are different but related. Elastic net regression is suitable and can handle multicollinearity. Since our sample size was small, this was intended to be an exploratory analysis, and we did not implement fine-grained hyperparameter tuning. The default setting was used: an equal balance of 0.5 was used for "l1_ratio", and a full weighting of 1.0 was used for "alpha". We evaluated the elastic net model on the L2 dataset using repeated 10-fold cross-validation, with three repeats. We reported the average mean absolute error (MAE) on the test dataset in Table 5. The resulting MAE value entails an average measure of how far the elastic net model's predictions are from the actual target values in the test set. Standard deviation is given inside the parenthesis.

Our findings indicated that, when predicting L2 writing proficiency, there is no noticeable difference in terms of keeping or removing the new LLM indices. For all three models with different feature combinations (both LLMs and classic indices; LLMs only; and

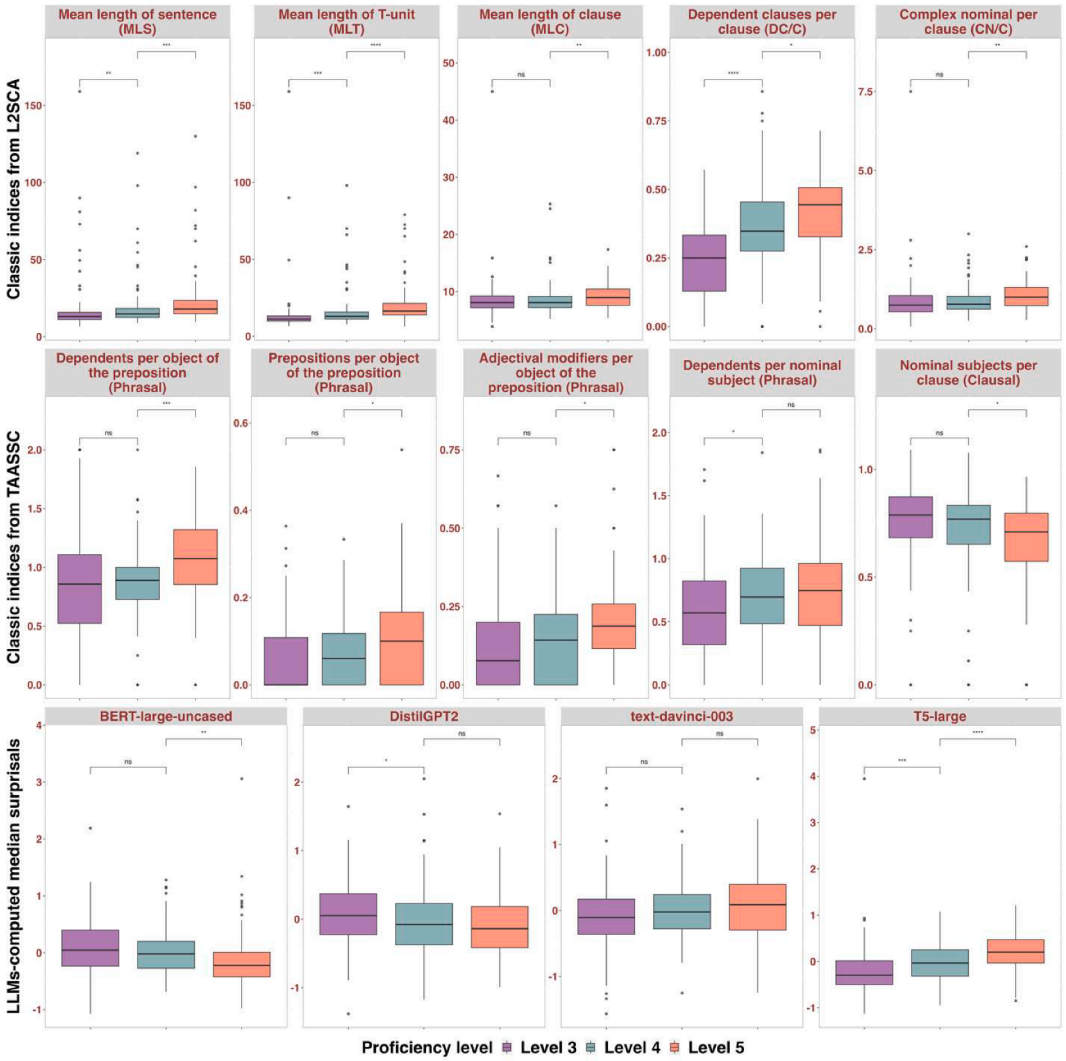


Fig. 1. Paired comparisons across learner proficiency levels: LLMs indices and selected classic indices for indexing L2 development stages. Significance notation: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; *ns:* $p > 0.05$.

Table 5
Evaluation of elastic net regression in predicting L2 learners' writing (variable *Writing_Sample*) and overall proficiency (variable *MTELP_Conv_Score*).

	LLMs and classic indices	LLM indices	Classic indices
Predicting writing proficiency	0.605 (0.087)	0.604 (0.087)	0.605 (0.087)
Predicting overall proficiency	11.168 (1.337)	11.358 (1.254)	11.501 (1.427)

classic indices only), elastic net models gave almost the same MAE (0.604–0.605). When predicting L2 learners' overall proficiency, we found that elastic net models with only the classic indices gave rise to the highest MAE, i.e., the worst performance. Adding LLM-Surprisals on top of the classic features led to the best elastic net model (MAE = 11.168). LLM indices in conjunction with the classic features contributed to the most effective model for predicting overall proficiency. This led us to argue that LLM indices have the potential to advance the classic indices in benchmarking L2 interlanguage development. For the best-configured elastic net model (LLMs and classic indices predicting overall proficiency), we found that, with respect to model coefficients, the top three most informative features in the model's predictions were BERT (coefficient = -1.866), T5 (coefficient = 1.657), and the classic syntax complexity index *dependents per object of the preposition* (coefficient = 1.448). This implied that LLM indices weigh more than the classic ones in predictive models.

4.4. Decomposing L2 Surprisals

We have shown that computationally, as a utility, LLM-computed Surprisals can detect L2 writing, index L2 development stages, and they can enhance the existing indices. Conceptually, what does LLM-Surprisal capture in an L2 setting? How do we decompose and interpret Surprisals in L2 research? We propose two approaches to improve LLMs' interpretability and transparency in an L2 context: First, we created linear mixed-effects models for both the new and the classic indices, modeling the trajectory of how these indices manifest themselves throughout L2 development. Second, we conducted error analyses, qualitatively analyzing the successes and failures in LLMs.

As illustrated in Section 2.4, we hypothetically interpreted LLM-Surprisals as a numeric representation of the interplay between syntactic complexity and lexical diversity in L2 interlanguage development. For syntactic complexity, we selected the L2SCA-computed measure DC/C (dependent clauses per clause). This selection was made due to the following considerations: First, in our investigation so far, this index showed significance across multiple analyses. Second, DC/C is informative of the subordination amount, which is critical to syntactic complexity, because it introduces additional hierarchical layers within sentences, increasing the depth of the syntactic tree and cognitive load required for processing. Each subordinate clause adds new dependency relations and nuanced

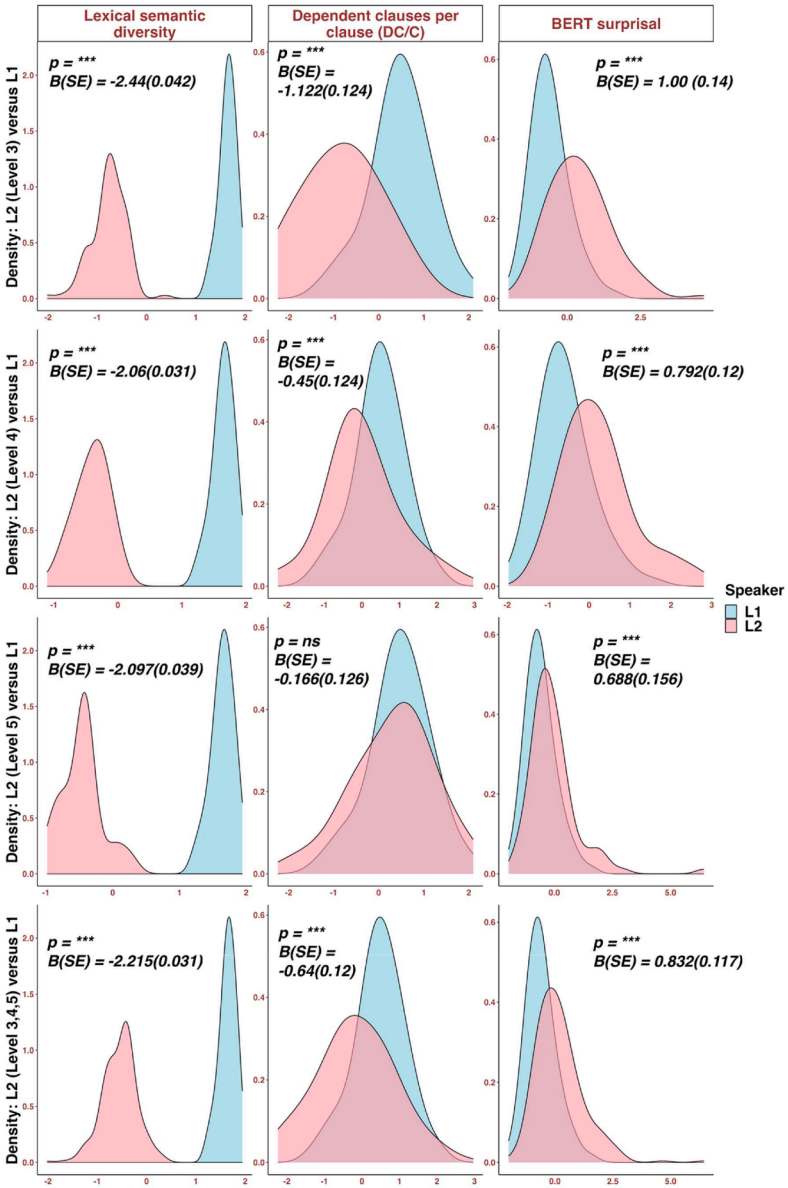


Fig. 2. Density plot of lexical diversity, syntactic complexity, and Surprisals, in L1 and L2 with various proficiency levels. Significance notation: $*p < 0.05$; $**p < 0.01$; $***p < 0.001$; *ns:* $p > 0.05$.

relationships such as causality or temporality, making the sentence structure more advanced cognitively. Third, subordination also tends to lengthen sentences and contribute to more complex structures, affecting readability and comprehension. It is a commonly used, recommended measure of productive complexity (Vercellotti, 2019). For lexical diversity, we calculated the lexical semantic diversity index, as motivated and discussed in Section 2.

We constructed three linear mixed-effects models for each individual proficiency level, as well as for the combined levels. The three models differ in response variables: (a) a model with lexical semantic diversity as the response variable, (b) a model with syntactic complexity (DC/C: dependent clauses per clause) as the response variable, and (c) a model with Surprisal as the response variable. For all the linear mixed-effects models, we used L2 status (is or is not L2 speaker) and number of words as fixed effects, with participant ID as a random effect. For illustration purposes, we chose BERT Surprisals as a representative of LLM-Surprisals, because in our investigation so far, BERT showed significance across all analyses, and it also gave the most decisive features in the predictive model (c.f., elastic net regression in Table 5). We visualized a density plot in Fig. 2, with the essential statistics coefficient (*B*), standard error (*SE*), and *p*-value printed on the top row.

As illustrated in Fig. 2, for lexical semantic diversity, there was statistically significant differences in L1 and L2 with level 3, 4, 5 proficiencies. The effect of lexical diversity also manifested when all three levels are combined. Compared to L2 (regardless of proficiency levels), a higher density of L1 showed high lexical semantic diversity. From level 3 to 4 L2 learners, this difference in lexical semantic diversity between L1 and L2 became noticeably smaller (*B* changed from -2.44 to -2.06). These findings support previous results in Berger et al. (2017), indicating that proficient speakers showed higher lexical semantic diversity. There was a slight increase of such difference from level 4 to level 5 L2 learners (*B* changed from -2.06 to -2.097). This implied the sophisticated and potentially non-linear nature of L2 interlanguage development. It is progressive with fluctuations. Moreover, our findings suggested that for lexical diversity, the distribution of L1 is more concentrated than L2. For syntactic complexity, a significantly higher syntactic complexity (DC/C dependent clauses per clause) was found in L1 than in L2 with level 3 and 4 proficiencies, and this difference disappeared when L2 learners' proficiency improves to level 5. From level 3 to 4, the difference in syntactic complexity between L1 and L2 became smaller, suggesting that L2 learners' interlanguage systems become more and more sophisticated and developed. When combining all three levels, we found significantly lower syntactic complexity in L2 than in L1. Our finding is generally in line with previous work by Wang et al. (2022). We additionally found that the distribution of syntactic complexity in L1 is clustered, whereas it is more spread-out in L2, suggesting more variance in the L2 group. As to BERT Surprisals, compared to L1, a significantly larger density of L2 showed higher Surprisals, across all L2 proficiency levels. As L2 learners' proficiency improves, this difference in Surprisals between L2 and L1 became smaller, indicating that L2 learners' interlanguage systems become more advanced. When combining all three levels, significantly higher Surprisal was found in L2 than in L1. We again found more variance in L2 groups' Surprisal distribution than in L1. Overall, our findings were as predicted. LLM-Surprisals can be interpreted as a collective representation of lexical semantic diversity and syntactic complexity. Our findings reveal that the underlying reason why L2 text's Surprisal decreases as proficiency increases is likely associated with lexical diversity and syntactic complexity.

In order to further demystify LLMs in L2 research, we additionally provided concrete examples in Table 6 for a qualitative error analysis. BERT assigned a Surprisal of 2.67 to an L1 essay, which was expectedly lower than that of level 3 and 4 L2s' writing. However, BERT unexpectedly assigned a higher Surprisal to an L1's than to a level 5 L2's essay, suggesting that the level 5 L2's essay was considered by BERT to be less surprising and more natural than the L1 essay. We speculated that this unexpected pattern of BERT

Table 6
Concrete examples exerpts from L1 and L2 writings, and the corresponding LLM-computed Surprisal scores. Note: *xxx* refers to words that can lead to identifiable information; readers may refer to the source PELIC dataset for full length text [https://github.com/ELI-Data-Mining-Group/PELIC-dataset].

Corpus	Text excerpt (de-identified)	Median Surprisals
L1	*Right from the start of the semester, my life at the campus has been both educational and challenging. Fortunately, my passionate desire and self-empowerment to pursue my education to the end, coupled with the unbowed backing of my support system, kept me pushing. At the end of it all, I have grown both as a person and as a professional in the last few months. The ongoing coronavirus outbreak has also been a major setback for my education during the final semester. It is now common knowledge that this virus has indiscriminately devastated people of all cultures and nationalities.*	BERT 2.67; GPT2 3.87; DistilGPT2 4.36; GPTNeo 3.41; T5 11.71
L2—level 3	*You know three of us in our family like to use computers, but we just have two, one desktop and one laptop. During weekends or vacation days, we are often lack of one computer. So we decided to buy a new computer, but we argued for a long time about buying a laptop or desktop. You know each one has advantages and disadvantages. First, desktop usually has large space in hard disk and memory cards, but it isn't convenient for moving. Second, desktop often runs faster than laptop. If you use computer edit video or make three-dimensional objects, desktop is preponderant. However, you can't use it or go online anywhere like laptop.*	BERT 4.17; GPT2 4.54; DistilGPT2 4.96; GPTNeo 3.97; T5 11.04
L2—level 4	*Failing a test is not easy successful, but if you want to do this, there are many method can reach this goal. For example, you needn't to do your homework and don't go to class on time. When you have a test, don't write anything on your test paper if you know the answer. On the other hand, you needn't review your class everyday and don't highlight any important vocabulary. sentences on your book.*	BERT 2.96; GPT2 4.6; DistilGPT2 4.86; GPTNeo 4.08; T5 11.3
L2—level 5	*our car was carefully checked in case something unexpected would happen. So our car was expected to be in good condition for a long distance trip. Generally we were supposed to be ready for any weather condition, but we did nothing because it was the first time we had such a long drive. We were so happy that we forgot to take a look at the weather condition along the route. When we were driving on the express way very close to xxx and immediately after we were told to take care because of lake side effect, our wind shield glass was covered a thick layer of snow in a very short time and wind shield wiper was blocked by quickly-frozen snow. So anything could not be seen through our car window.*	BERT 1.84; GPT2 3.88; DistilGPT2 4.19; GPTNeo 3.64; T5 11.8

Surprisal was caused by relatively low-frequency words such as "self-empowerment" and "unbowed". The same pattern was found in DistilGPT2. For GPT2, the L1 essay was assigned a lower Surprisal than all the L2 essays. However, GPT2 unexpectedly assigned a higher Surprisal to a level 4 L2 essay than to a level 3 L2 essay. It is likely that the overuse and misuse of negation in the level 4 L2 essay confused GPT2, leading to high Surprisals. GPTNeo assigned a lower Surprisal to L1 than to all the L2 essays. Unexpectedly, GPTNeo assigned *higher* Surprisals as L2 proficiency increased from level 3 to level 4. This is likely an indicator that L2 interlanguage system is complex, and its development might ne nonlinear. An alternative explanation is that GPTNeo did not fully capture L2 learners' interlanguage evolvement. T5 appeared to give the least interpretable pattern, assigning the highest Surprisal to level 5 L2 essays, whereas it assigned the lowest Surprisal to level 3 L2 essays, with L1 essays Surprisals in between. Although T5's pattern seems uninterpretable, it is consistent, and it echoes findings in the previous experiments.

In summary, a close scrutiny of concrete examples indicated that none of the LLMs fully capture L2 interlanguage development. It also indirectly demonstrated that L2 interlanguage development is incremental, multi-dimensional, with occasional regressions, although the general trend is towards improvement. There are possible solutions to further improve LLMs' accountability: For example, addressing the effect of low-frequency vocabulary and misuse of negations, among other possible triggers of high Surprisals. For future research, we plan to compute the maximum, minimum, and standard deviation of Surprisals. This will provide a more elaborate picture of the distribution of Surprisals and its interplay with potentially edge cases in linguistics.

5. Discussion

In this study, we attempt to address how accurate and effective state-of-the-art LLMs can be in facilitating L2 writing assessment and informing L2 development research. In an L2 writing context, we explored and showcased LLMs' gradient representations in Surprisals, architectural assumptions and their implications, and broad accessibility relative to the classic NLP indices. Our findings reveal that LLMs can be reliable utilities, and they are most effective when used in conjunction with the classic metrics. We hope that our pipeline and findings provide inspirations for future researchers that LLMs can be innovative with respect to L2 research inquiry.

5.1. LLM-Surprisals as a proxy to L2 naturalness

We speculate that LLM-derived Surprisal can be further interpreted as a proxy measure of L2 writing *naturalness*. Farghal (1992) highlighted that naturalness involves cohesion and fluency, which make L2 writing appear seamless and "native-like". Farghal (1992) also noted that, while grammar matters for evaluating L2 writing, the natural flow of language is equally critical for high proficiency L2 writings. This study emphasizes the need for L2 writing to not only be grammatically well-formed but also to sound natural and authentic to native L1 speakers. Similarly, Silva and Matsuda (2010) discussed how naturalness is important for ensuring that L2 learners can effectively communicate with (L1) speakers in real-world contexts, beyond the confines of the classroom. This aligns with the view that L2 writing should be assessed not only for its structural correctness but also for its ability to convey meaning naturally and fluently. Relatedly, the LLMs that we used in the current work are mostly pre-trained in "native" and authentic English data, which are dominantly produced by L1 English speakers. Conceptually, LLMs consider a text to be surprising and assign higher Surprisals based on what they are pre-trained on, namely, contextualized predictions in a native and natural L1 text (Tunstall et al., 2022). From this perspective, LLM-Surprisal is very reminiscent of the naturalness concept in L2 writing assessments, both referring to how closely L2 writing mimics the linguistic patterns of native L1 speakers.

LLM-Surprisals and the L2 naturalness index share a common characteristic: both are composite indices of L2 development. Similar to Surprisals, naturalness not only characterizes grammatical well-formedness (Sinclair, 1984); it is a holistic measure that includes several factors, such as grammatical accuracy, idiomatic usage, and overall fluency and cohesion (Kobayashi and Rinnert, 1992; Polio, 1997; Silva and Matsuda, 2010). According to Polio (1997) and Kobayashi and Rinnert (1992), the importance of naturalness lies in its significant impact on the overall perceived proficiency of L2 learners, influencing both evaluators' judgments and the learners' communicative effectiveness. Additionally, di Gennaro (2006) argues for a clear distinction between L2 writing ability and other forms of L2 knowledge, suggesting that a comprehensive model of L2 writing proficiency should illustrate how it interacts with overall L2 proficiency. This highlights the multifaceted nature of naturalness, integrating it as a core component of L2 writing assessment. To sum up, both naturalness and LLM-Surprisals emphasize that effective evaluation must consider both grammatical accuracy and the natural use of language to provide a holistic measure of L2 proficiency.

Although LLM-Surprisal aligns with L2 naturalness in certain aspects, with the current evidence, we can only *speculate* that Surprisals are numerical proxies to approximate writing constructs such as naturalness. The definition of naturalness from internationally recognized proficiency standards diverges from the LLM-derived Surprisal in many ways. In the context of L2 learning, naturalness is often related to the ability to use the language fluently, spontaneously, and appropriately in a variety of real-life situations. This concept is discussed in frameworks such as CEFR (2001) and the American Council on the Teaching of Foreign Languages (ACTFL, 2012) Proficiency Guidelines. According to the CEFR, naturalness is linked to the descriptors for spoken interaction and production across different proficiency levels. At higher levels (C1 and C2), learners are expected to communicate with a high degree of fluency and spontaneity, producing language that sounds natural and is appropriate to the context. The ACTFL Proficiency Guidelines also emphasize naturalness in language use, particularly at the Advanced and Superior levels—they can handle a variety of communicative tasks with naturalness and ease (ACTFL, 2012).

In both frameworks, the notion of naturalness involves not only grammatical accuracy and vocabulary usage but also the ability to engage in authentic communication that feels natural and spontaneous to native speakers. On the other hand, here we decomposed LLM-Surprisals as a numeric representation of the interplay between lexical diversity and syntactic complexity, with no intended

quantification on how much Surprisals can represent a L2 speaker's engagement capability, and how authentic it is to L1 speakers. Although both concepts capture multiple dimensions of the L2 interlanguage, and there appears to be intersections of what they characterize, they are not at all equivalent to each other. More in-depth investigation is needed to measure engagement in authentic communication and its representation in LLMs, and to understand to what extent LLM-Surprisals can capture the naturalness construct.

5.2. LLM-Surprisals and other L2 measures

We hypothesized and justified LLM-Surprisal as a composite index of lexical diversity and syntactic complexity, and we further speculated the Surprisal index to be potentially connected to naturalness. However, there are likely associations between Surprisals and other L2 indices. First, it is possible that LLM-Surprisals manifest themselves in *semantic anomaly*. Low-proficiency L2 learners may have a less intuitive grasp of English semantics compared to native or proficient L2 speakers. We can infer that their semantics might deviate more from what is expected according to an LLM, which is pre-trained in (L1) English data. Thus, semantic anomaly in L2 essays might potentially give rise to higher Surprisal scores for some LLMs.

Second, LLM-Surprisal scores are related to but different from *acceptability*. In the field of NLP, LLM-Surprisals have been proposed as a measure of acceptability (Misra, 2022). We maintain that in an L2 setting, there appears to be an overlap between Surprisal and acceptability, but Surprisal may go above and beyond, and it is not equivalent to acceptability. Syntax and syntax–semantics interface factors are decisive in acceptability judgements (Sprouse, 2007). Our findings indicate that Surprisal scores can also be attributed to lexical factors, aside from syntax. How much lexical semantic diversity influences acceptability judgements is open to discussion, and it is out of the scope of the current work.

Third, it is also possible that LLMs correlate with *topic*-level measures. For example, certain LLMs are effective at quantifying topic complexity (Raffel et al., 2020). We speculate that L1 writing consists of more sophisticated topic structures, whereas L2 writing, depending on the learners' proficiency level, might involve different degrees of topic complexity. For future research, we plan to examine whether LLM-derived Surprisal metrics reflect topic complexity. We also acknowledge that Surprisal is likely to increase in contexts involving creative and specialized topics or content, or cultural descriptions. This could potentially complicate the operation and interpretation of LLM-Surprisals in differentiating between L1 and L2 speakers or predicting L2 proficiency.

5.3. LLMs' usage in L2 studies

Three of our findings about LLMs were not precisely as predicted: First, LLM-derived metrics are not always more robust and effective than the classic indices. We speculate that this is likely due to most LLMs' lack of long-text *reasoning*. Studies have indicated that LLMs are currently not very successful in pragmatic and contextual reasoning, although they can *process* long text (Barrett et al., 2018; Collins et al., 2022; Lake and Murphy, 2023; Mahowald et al., 2023). It is also important to note that L2 writing proficiency involves multiple aspects, ranging from grammar, vocabulary, and coherence to creativity. LLM-derived metrics might not capture all dimensions of writing proficiency. We hope that, as a starting point for demystifying LLMs' usage in L2 research, LLM-computed metrics can provide *some* insights into the complexity, fluency, and predictability of L2 text. Even though Surprisal scores alone may not be sufficient to determine writing quality, they provide clues as to the relative unexpectedness of certain language usage. For future research, we will consider integrating additional LLM-generated metrics (e.g., similarity, perplexity) into human experts' evaluations, in order to obtain a more comprehensive understanding and assessment of L2 writing proficiency.

Second, it was found that LLMs involving an encoder (e.g., BERT-large-uncased) did not always underperform decoder GPT-type LLMs, contrary to our predictions. In the tasks of detecting L2 writing and predicting human professionals' ratings, encoder LLMs sometimes showed stronger effects than decoder ones. BERT's bidirectional architecture allows it to take into account both the previous and future contexts. This property likely enables better *comprehension* of the overall context, resulting in more accurate predictions of word probabilities. Since encoder LLMs' pre-training tasks involve not only masked language modeling, but also next-sentence prediction (Devlin et al., 2018), we speculate that these pre-training tasks contribute to encoder LLMs' ability to capture writing proficiency indices effectively. In contrast, GPT-type LLMs are effective at text *production* due to their unidirectional nature. In other words, our findings can be potentially explained by the architecture difference that encoder LLMs are more proficient at comprehension, whereas decoder LLMs are more proficient at production. Hence, we propose that—at least for separating L1 from L2 writing and for predicting human professionals' ratings of L2 essays—compared to decoder LLMs, encoder LLMs would be more effective at comprehending the overall text, leading to a generally more sensible L2 writing proficiency metric.

Third, we did not find larger LLMs to always outperform the smaller ones, which is not exactly as predicted but is consistent with some previous studies (Ormerod et al. 2021). GPT2 was found to outperform larger LLMs in aligning with human reading times (Shain, 2019, 2024; Shain et al., 2024). Our approach and findings highlight the necessity for a more refined approach to evaluating the effectiveness of LLMs in L2 research (Gebru et al., 2022; Shain et al., 2024; Shwartz-Ziv and Tishby, 2017). Smaller LLMs, because of their reduced architectural complexity, seem more likely to exhibit better performance and adaptability in domain-specific datasets (Henderson et al., 2018), such as L2 learners' writing.

There are disadvantages and limitations of the proposed LLM approaches. For example, GPT-type LLMs, especially the latest GPT models hosted under the OpenAI API, lack stability in deriving probability-based indices (Hu et al., 2023; Chen et al., 2013, 2024). Although we reported standard error in our statistics, there are still uncertainties as to the generalizability and reproducibility of our findings. We plan to publish all of our pipeline code and datasets for future researchers and practitioners to explore. Moreover, we speculate that disfluencies such as spelling errors may affect LLMs' capabilities. Pre-trained LLMs are likely to be brittle when dealing with disfluent text. For now, we only conducted basic pre-processing. We hope to conduct different levels of pre-processing and

examine how it may influence LLMs' competence in automatic writing assessment. Further, according to Ramesh and Sanampudi (2022), for the content-based features extracted and studied for automatic assessment, no models have been created or validated for whether a student's response is *relevant* to the given question prompt, few studies have considered cohesion and coherence, no thorough explanation about *consistency* in writing has been provided, and no machine-learning-based ontological approach has been conducted (e.g., similarity; knowledge graphs). Although LLM-Surprisals might seem related to content-based evaluation such as naturalness and lexical cohesion, we acknowledge that more evidence is needed to connect Surprisal with other content analysis factors, such as consistency and relevance. We also plan to combine Surprisals with similarity and knowledge graphs in future studies to build predictive models.

Another potential limitation of using LLMs in L2 studies is that employing closed-source LLMs such as GPT3 poses scientific obstacles. The absence of publicly available information about the specific models under examination—including details such as their parameter counts, training datasets, and corpus sizes, among other crucial aspects—creates significant challenges. During the course of this project, OpenAI ceased support for the GPT model in obtaining token-wise probabilities, highlighting another issue in working with closed-source models: the difficulty in ensuring long-term replicability and the dependency on non-transparent LLMs. Although we managed to reproduce and validate our findings by using an open-source LLM with similar architectures, there seems no assurance of indefinite access to certain LLMs. This highlights the uncertainty nature of relying on such models for scientific research.

5.4. Broader impacts and practical applications

How can our findings be applied in practical educational settings? Our investigations can inform and advance automatic assessment (Ramesh and Sanampudi, 2022; Chen and Pan, 2022). We detailed what metrics are being considered and how they differ from classic proficiency assessment methods. Furthermore, we explored the underlying mechanisms of LLMs in differentiating L2 proficiency levels and L1 and L2 writing in general, thereby adding depth to our research and demystifying LLMs in practical applications. By specifying which classic NLP indices are being compared, we hope to explain how this comparison contributed to the understanding of LLMs' efficacy in language assessment, aside from the LLMs' understanding of language complexity and development. We additionally provide recommendations on the use of different LLMs in different L2 tasks. In summary, integrating LLMs into education practice is timely and relevant given the rapid advancements in NLP technologies. We hope that the proposed pipeline demonstrates an initial framework for using LLMs in benchmarking L2 development and general language assessment, showcasing a potential path for combining computational linguistics with language learning.

Our investigation supports and advances previous studies targeting Chinese learners of English as a foreign language. In agreement with Chen and Pan (2022), our findings suggest that new tools give the best performance when used together with the classic existing tools. Our study also supports previous findings by Liu and Kunnan (2016), in that we showcase the *feasibility* of such an LLM-based automated approach for grading Chinese English learners' essays and understanding their interlanguage development. For future research, we hope to incorporate feedback into our LLM-based pipeline (Weigle, 2013; Liu et al., 2016).

Further, our study has general educational implications. First, language educators and researchers may benefit from incorporating LLM metrics into the battery of linguistic indices to track and understand L2 writing development. Second, viewing LLM-Surprisals as a proxy measurement of naturalness is also relevant to foreign- and second-language teaching. For example, teaching natural text in L2 education is important because it exposes learners to authentic and context-rich language use, which helps develop their overall language proficiency and communicative competence. Natural text reflects the complexities of real-world communication, including idiomatic expressions, cultural nuances, and text patterns. This would enable learners to understand and produce language that is more natural in various social and cultural contexts (Olshtain and Celce-Murcia, 2016). Third, our pipeline will greatly help front-line teachers. Manual L2 writing assessment can be labor-intensive and tends to generate low inter-rater reliabilities. Text-based evaluation is especially associated with time and cost difficulties (Paquot, 2017), making the implementation of holistic computational metrics particularly relevant and urgent. Our NLP pipeline is efficient, scalable, and reliable; we validated it with multiple large-scale L2 corpora. According to Narcy-Combes (2003), we should highlight the need for large-scale L2 corpora, because they enable detailed analysis of acquisition sequences. Filling this research gap is an important priority. We provided interpretability methods that can improve transparency in current AES systems (Kumar & Boulanger, 2021; Kumar et al., 2023), advocating for human interpretable indices and AES-teacher collaboration (Schneider et al., 2023). We hope that the validated LLM-based NLP pipeline can be utilized by teaching practitioners. The practical applications of our method and findings can lead to more personalized and effective language learning experiences, benefiting both educators and learners.

6. Conclusions

AI systems such as LLMs are transformative. However, there has been relatively little study as to what an accessible and interpretable LLM-based pipeline can bring. This study attempts to bridge that gap. We developed and evaluated such a pipeline, showcasing that, as a utility, pre-trained LLMs have the potential to generate meaningful language indices without computationally intensive domain-specific training. Such indices are predicative of standardized test scores, and they can decently detect and index L2 writings. Crucially, LLM indices show research merit. They provide a scalable paradigm for L2 researchers to track and understand L2 interlanguage systems. Our interpretability analyses suggested that, as L2 learners' interlanguage system evolves, their writing's syntactic complexity increases noticeably, and their lexical diversity also changes, possibly giving rise to more developed and natural writings, hence the lower LLM-Surprisals that we observed on the surface. This implied that, as a holistic index derived from holistic LLMs, Surprisals can characterize L2 writing's development more thoroughly, revealing underlying patterns that may not seem

immediately salient with other classic measurements. We hope that our investigation will lead to more nuanced questions on pinpointing AI models' role in understanding language learning and building educational applications.

CRediT authorship contribution statement

Yan Cong: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The second language data were drawn from PELIC https://github.com/ELI-Data-Mining-Group/PELIC-dataset. Corpus citation: Juffs, A., Han, N-R., & Naismith, B. (2020). The University of Pittsburgh English Language Corpus (PELIC). https://doi.org/10.5281/zenodo.4577423. The first language data were drawn from MICUSP https://elicorpora.info/main. Corpus citation: Michigan Corpus of Upper-level Student Papers (MICUSP). (2009). Ann Arbor, MI: The Regents of the University of Michigan. The validation data were drawn from CROW. Corpus citation: Staples, S., & Dilger, B. (2018-). Corpus and repository of writing (CROW) [Learner corpus articulated with repository]. Available at https://crow.corporaproject.org. The script for the analysis in this paper is available online: https://doi.org/10.17605/OSF.IO/EYHPQ.

Acknowledgments

We would like to thank Phillip Wolff, Emmanuele Chersoni, Yu-Yin Hsu, Sunny X. Tang, Jiyeon Lee, and Arianna N. LaCroix for their inspiration and fruitful discussions. Special thanks to Elaine J. Francis and Bradley Dilger for introducing and granting access to the CROW corpus. Thank you all the anonymous reviewers and editor for their constructive comments. This research was funded by the School of Languages and Cultures, Purdue University. All errors remain mine.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.csl.2024.101700.

References

ACTFL, J, 2012. ACTFL Proficiency Guidelines 2012. ACTFL, White Plains, New York.
Attali, Y., Burstein, J., 2006. Automated essay scoring with e-rater® V. 2. J. Technol. Learn. Assess. 4 (3).
Barrett, D., Hill, F., Santoro, A., Morcos, A., Lillicrap, T., 2018. Measuring abstract reasoning in neural networks. In: *Proceedings of the* International Conference on Machine Learning, pp. 511–520.
Berger, C.M., Crossley, S.A., Kyle, K., 2017. Using novel word context measures to predict human ratings of lexical proficiency. J. Educ. Techno Soc. 20 (2), 201–212.
Bestgen, Y., Granger, S., 2014. Quantifying the development of phraseological competence in L2 English writing: an automated approach. J. Second. Lang. Writ. 26, 28–41. https://doi.org/10.1016/j.jslw.2014.09.004.
Bexte, M., Horbach, A., Zesch, T., 2022. Similarity-based content scoring-how to make S-BERT keep up with BERT. In: Proceedings of the 17th Workshop on Innovative Use of Nlp for Building Educational Applications (Bea 2022), pp. 118–123.
Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U.S., Purohit, S., Reynolds, L., Tow, J., Wang, B., Weinbach, S., 2022. GPT-NeoX-20B: an open-source autoregressive language model. In: Proceedings of the BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models. https://doi.org/10.18653/v1/2022.bigscience-1.9.
Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., & others. (2021). On the opportunities and risks of foundation models. *ArXiv Preprint ArXiv:2108.07258*.
Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., others, 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.
Bulté, B., Housen, A., 2012. Defining and operationalising L2 complexity. Language Learning & Language Teaching. John Benjamins Publishing Company, pp. 21–46. https://doi.org/10.1075/llt.32.02bul.
Bulté, B., Housen, A., 2014. Conceptualizing and measuring short-term changes in L2 writing complexity. J. Second. Lang. Writ. 26, 42–65. https://doi.org/10.1016/j.jslw.2014.09.005.
Bulté, B., Roothooft, H., 2020. Investigating the interrelationship between rated L2 proficiency and linguistic complexity in L2 speech. System 91, 102246. https://doi.org/10.1016/j.system.2020.102246.
Caldwell-Harris, C.L., 2021. Frequency effects in reading are powerful–But is contextual diversity the more important variable? Lang. Linguist. Compass. 15 (12), e12444.
Chen, H., Pan, J., 2022. Computer or human: a comparative study of automated evaluation scoring and instructors' feedback on Chinese college students' English writing. Asian-Pac. J. Second and Foreign Lang. Educ. 7 (1), 34.
Chen, H., Xu, J., He, B., 2013. Automated essay scoring by capturing relative writing quality. Comput. J. 57 (9), 1318–1330. https://doi.org/10.1093/comjnl/bxt117.

Chen, J., Lin, H., Han, X., Sun, L., 2024. Benchmarking large language models in retrieval-augmented generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, 38, pp. 17754–17762.
Cobb, T., & Boulton, A. (2015). Classroom applications of corpus analysis. *Cambridge Handbook of Corpus Linguistics*, 478–497.
Collins, K. M., Wong, C., Feng, J., Wei, M., & Tenenbaum, J. B. (2022). Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. *ArXiv Preprint ArXiv:2205.05718*.
Cong, Y., Chersoni, E., Hsu, Y.Y., Blache, P., 2023. Investigating the Effect of Discourse Connectives on Transformer Surprisal: Language Models Understand Connectives; Even So They Are Surprised. In: 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP.
Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, 2001. Common European Framework of Reference For Languages: Learning, Teaching, Assessment. Cambridge University Press.
Crossley, S., Holmes, L., 2022. Assessing receptive vocabulary using stateoftheart natural language processing techniques. J. Second Lang. Stud. 6 (1), 1–28. https://doi.org/10.1075/jsls.22006.cro.
Crossley, S.A., Salsbury, T., Mcnamara, D.S., 2015. Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. Appl. Linguist. 36 (5), 570–590.
Dahl, Ö., 2004. The growth and maintenance of linguistic complexity. Studies in Language Companion Series. John Benjamins Publishing Company. https://doi.org/10.1075/slcs.71.
De Clercq, B., 2015. The development of lexical complexity in second language acquisition: a cross-linguistic study of L2 French and English. EUROSLA Yearbook 15, 69–94. https://doi.org/10.1075/eurosla.15.03dec.
De Clercq, B., Housen, A., 2016. The development of morphological complexity: a cross-linguistic study of L2 French and English. Second. Lang. Res. 35 (1), 71–97. https://doi.org/10.1177/0267658316674506.
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv: 1810.04805*.
di Gennaro, K. (2006). *Second language writing ability: Towards a complete construct definition*.
Egbert, J., 2017. Corpus linguistics and language testing: navigating uncharted waters. Lang. Test. 34 (4), 555–564. https://doi.org/10.1177/0265532217713045.
Farghal, M., 1992. Naturalness and the notion of cohesion in EFL writing classes. IRAL 30 (1), 45–50.
Frank, S.L., Otten, L.J., Galli, G., Vigliocco, G., 2015. The ERP response to the amount of information conveyed by words in sentences. Brain Lang. 140, 1–11.
Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., Levy, R., 2019. Neural language models as psycholinguistic subjects: representations of syntactic state. In: Proceedings of the Conference of the North. https://doi.org/10.18653/v1/n19-1004.
Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., & others. (2020). The pile: an 800gb dataset of diverse text for language modeling. *ArXiv Preprint ArXiv:2101.00027*.
Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé, H., Crawford, K., 2022. Excerpt from datasheets for datasets *. Ethics of Data and Analytics. Auerbach Publications, pp. 148–156. https://doi.org/10.1201/9781003278290-23.
Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S.A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., Hasson, U., 2022. Shared computational principles for language processing in humans and deep language models. Nat. Neurosci. 25 (3), 369–380. https://doi.org/10.1038/s41593-022-01026-4.
Hale, J., 2001. A probabilistic earley parser as a psycholinguistic model. In: Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies 2001 - NAACL '01. https://doi.org/10.3115/1073336.1073357.
Hardy, J.A., Römer, U., 2013. Revealing disciplinary variation in student writing: a multi-dimensional analysis of the Michigan Corpus of Upper-level Student Papers (MICUSP). Corpora 8 (2), 183–207.
Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., Meger, D., 2018. Deep reinforcement learning that matters. In: Proceedings of the AAAI Conference on Artificial Intelligence, 32. https://doi.org/10.1609/aaai.v32i1.11694.
Ho, D.E., Imai, K., King, G., Stuart, E.A., 2011. MatchIt: nonparametric preprocessing for parametric causal inference. J. Stat. Softw. 42 (8), 1–28. https://doi.org/10.18637/jss.v042.i08.
Hoffman, P., Lambon Ralph, M.A., Rogers, T.T., 2012. Semantic diversity: a measure of semantic ambiguity based on variability in the contextual usage of words. Behav. Res. Methods 45 (3), 718–730. https://doi.org/10.3758/s13428-012-0278-x.
Housen, A., Kuiken, F., 2009. Complexity, accuracy, and fluency in second language acquisition. Appl. Linguist. 30 (4), 461–473.
Hu, J.-M., Liu, F.-C., Chu, C.-M., Chang, Y.-T., 2023. Health care trainees' and professionals' perceptions of ChatGPT in improving medical knowledge training: rapid survey study. J. Med. Internet. Res. 25, e49385.
Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., Linzen, T., 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. J. Mem. Lang. 137, 104510.
Johns, B.T., 2022. Accounting for item-level variance in recognition memory: Comparing word frequency andcontextual diversity. Mem. Cogn. 50, 1013–1032.
Jurafsky, D., Martin, J.H., 2024. Speech and Language Processing 3rd Edition Draft. Draft.
Kakouros, S., Šimko, J., Vainio, M., Suni, A., 2023. Investigating the utility of surprisal from large language models for speech synthesis prosody. In: Proceedings of the *12th ISCA Speech Synthesis Workshop (SSW2023)*. https://doi.org/10.21437/ssw.2023-20.
Kettunen, K., 2014. Can type-token ratio be used to show morphological complexity of languages? J. Quant. Linguist. 21 (3), 223–245. https://doi.org/10.1080/09296174.2014.911506.
Kim, M., Crossley, S.A., Kyle, K., 2017. Lexical sophistication as a multidimensional phenomenon: relations to second language lexical proficiency, development, and writing quality. Mod. Lang. J. 102 (1), 120–141. https://doi.org/10.1111/modl.12447.
Kobayashi, H., Rinnert, C., 1992. Effects of first language on second language writing: translation versus direct composition. Lang. Learn. 42 (2), 183–209.
Kumar, V.S., Boulanger, D., 2021. Automated essay scoring and the deep learning black box: how are rubric scores determined? Int. J. Artif. Intell. Educ. 31, 538–584.
Kumar, Y., Parekh, S., Singh, S., Li, J.J., Shah, R.R., Chen, C., 2023. Automatic essay scoring systems are both overstable and oversensitive: explaining why and proposing defenses. Dialogue & Discourse 14 (1), 1–33.
Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication*.
Kyle, K., Crossley, S.A., 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. Tesol Q. 49 (4), 757–786.
Kyle, K., Crossley, S., 2017. Assessing syntactic sophistication in L2 writing: a usage-based approach. Lang. Test. 34 (4), 513–535. https://doi.org/10.1177/0265532217712554.
Kyle, K., Crossley, S.A., 2018. Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. Mod. Lang. J. 102 (2), 333–349. https://doi.org/10.1111/modl.12468.
Kyle, K., Crossley, S., Berger, C., 2018. The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. Behav. Res. Methods 50, 1030–1046.
Lake, B.M., Murphy, G.L., 2023. Word meaning in minds and machines. Psychol. Rev. 130 (2), 401.
Lan, G., Liu, Q., Staples, S., 2019. Grammatical complexity:'what does it mean'and 'so what'for L2 writing classrooms? J. Second. Lang. Writ. 46, 100673.
Landauer, T.K., Dumais, S.T., 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. 104 (2), 211.
Lee, G.-G., Latif, E., Wu, X., Liu, N., Zhai, X., 2024. Applying large language models and chain-of-thought for automatic scoring. Comput. Educ. Artif. Intell., 100213
Lee, Y.-J., 2020. The long-term effect of automated writing evaluation feedback on writing development. Engl. Teach. 75 (1), 67–92.
Levy, R., 2008. Expectation-based syntactic comprehension. Cognition 106 (3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006.
Liu, M., Li, Y., Xu, W., Liu, L., 2016. Automated essay feedback generation and its impact on revision. IEEe Trans. Learn. Technol. 10 (4), 502–513.
Liu, S., Kunnan, A.J., 2016. Investigating the application of automated writing evaluation to chinese undergraduate english majors: a case study of" WriteToLearn". Calico J. 33 (1), 71–91.

Lu, X., 2010. Automatic analysis of syntactic complexity in second language writing. Int. J. Corpus Linguistics 15 (4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu.
Lu, X., 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. TESOL Quart. 45 (1), 36–62. https://doi.org/10.5054/tq.2011.240859.
Lu, X., 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. Mod. Lang. J. 96 (2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x.
Lu, X., 2017. Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. Lang. Test. 34 (4), 493–511. https://doi.org/10.1177/0265532217710675.
Luck, S. J. (2012). *Event-related potentials.*
Ludwig, S., Mayer, C., Hansen, C., Eilers, K., Brandt, S., 2021. Automated essay scoring using transformer models. Psych. 3 (4), 897–915.
Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: a cognitive perspective. *ArXiv Preprint ArXiv:2301.06627*.
Michaelov, J.A., Bardolph, M.D., Van Petten, C.K., Bergen, B.K., Coulson, S., 2024. Strong prediction: language model surprisal explains multiple N400 effects. Neurobiol. Lang. 5 (1), 107–135. https://doi.org/10.1162/nol_a_00105.
Michaelov, J., Bergen, B., 2022. Collateral facilitation in humans and language models. In: Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL). https://doi.org/10.18653/v1/2022.conll-1.2.
Michaelov, J., & Bergen, B. (2023). Rarely a problem? Language models exhibit inverse scaling in their predictions following few-type quantifiers. *Findings of the Association for Computational Linguistics: ACL 2023*. 10.18653/v1/2023.findings-acl.891.
Misra, K. (2022). Minicons: enabling flexible behavioral and representational analyses of transformer language models. *ArXiv Preprint ArXiv:2203.13112*.
Misra, K., Ettinger, A., & Rayz, J. (2020). Exploring BERT's sensitivity to lexical cues using tests from semantic priming. *Findings of the Association for Computational Linguistics: EMNLP 2020*. 10.18653/v1/2020.findings-emnlp.415.
Mizumoto, A., Eguchi, M., 2023. Exploring the potential of using an AI language model for automated essay scoring. Res. Methods Appl. Linguistics 2 (2), 100050.
Naismith, B., Han, N.-R., Juffs, A., 2022. The University of Pittsburgh English Language Institute Corpus (PELIC). Int. J. Learn. Corpus Res. 8 (1), 121–138. https://doi.org/10.1075/ijlcr.21002.nai.
Narcy-Combes, J.-P., 2003. Rod Ellis, Task-based Language Learning and Teaching. Oxford University Press, pp. 87–88. https://doi.org/10.4000/apliut.3696, 2003. *Les Cahiers de l'APLIUT, Vol. XXII N° 3*.
Nieuwland, M.S., Van Berkum, J.J.A., 2006. When peanuts fall in love: N400 evidence for the power of discourse. J. Cogn. Neurosci. 18 (7), 1098–1111.
Norris, J.M., Ortega, L., 2009. Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. Appl. linguist. 30 (4), 555–578.
Olshtain, E., Celce-Murcia, M., 2016. Teaching language skills from a discourse perspective. Handbook of Research in Second Language Teaching and Learning. Routledge, pp. 144–158. https://doi.org/10.4324/9781315716893-11.
Ormerod, C. M., Malhotra, A., & Jafari, A. (2021). Automated essay scoring using efficient transformer-based language models. *ArXiv Preprint ArXiv:2102.13136*.
Ortega, L., 2003. Syntactic complexity measures and their relationship to L2 proficiency: a research synthesis of college-level l2 writing. Appl. Linguist. 24 (4), 492–518. https://doi.org/10.1093/applin/24.4.492.
Ortega, L, 2012. Interlanguage complexity: a construct in search of theoretical renewal. Linguistic Complexity. DE GRUYTER, pp. 127–155. https://doi.org/10.1515/9783110229226.127.
Ouyang, J., Jiang, J., Liu, H., 2022. Dependency distance measures in assessing L2 writing proficiency. Assess. Writ. 51, 100603 https://doi.org/10.1016/j.asw.2021.100603.
Paquot, M., 2017. The phraseological dimension in interlanguage complexity research. Second. Lang. Res. 35 (1), 121–145. https://doi.org/10.1177/0267658317694221.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.
Perelman, L., 2020. The BABEL generator and e-rater: 21st century writing constructs and automated essay scoring (AES). J. Writ. Assess. 13 (1), 1–10.
Polio, C.G., 1997. Measures of linguistic accuracy in second language writing research. Lang. Learn. 47 (1), 101–143.
Polio, C.G, 2001. Second Language development in writing: measures of fluency, accuracy, and complexity. Kate Wolfe-Quintero, Shunji Inagaki, and Hae-Young Kim. Honolulu: University of Hawai'i Press, 1998. Pp. viii + 187. 20.00 paper. Stud. Second. Lang. Acquis. 23 (3), 423–425. https://doi.org/10.1017/s0272263101263050.
R Core Team. (2023). *R: a language and environment for statistical computing*. https://www.R-project.org/.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., others, 2019. Language models are unsupervised multitask learners. OpenAI Blog 1 (8), 9.
Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21 (140), 1–67.
Ramesh, D., Sanampudi, S.K., 2022. An automated essay scoring systems: a systematic literature review. Artif. Intell. Rev. 55 (3), 2495–2527.
Rezaii, N., Michaelov, J., Josephy-Hernandez, S., Ren, B., Hochberg, D., Quimby, M., Dickerson, B.C., 2023. Measuring sentence information via Surprisal: theoretical and clinical implications in nonfluent aphasia. Ann. Neurol. 94 (4), 647–657.
Römer, U., Swales, J.M., 2010. The Michigan corpus of upper-level student papers (MICUSP). J. Engl. Acad. Purp. 9 (3), 249.
Ryu, S.H., & Lewis, R.L. (2021). Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. *arXiv preprint arXiv:2104.12874*.
Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv Preprint ArXiv:1910.01108*.
Schneider, J., Schenk, B., Niklaus, C., & Vlachos, M. (2023). Towards llm-based autograding for short textual answers. *ArXiv Preprint ArXiv:2309.11508*.
Shain, C., 2019. A large-scale study of the effects of word frequency and predictability in naturalistic reading. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4086–4094.
Shain, C., 2024. Word frequency and predictability dissociate in naturalistic reading. Open Mind 8, 177–201.
Shain, C., Meister, C., Pimentel, T., Cotterell, R., Levy, R., 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. Proc. Natl. Acad. Sci. 121 (10), e2307876121.
Shin, J., Gierl, M.J., 2021. More efficient processes for creating automated essay scoring frameworks: a demonstration of two algorithms. Lang. Test. 38 (2), 247–272.
Shwartz-Ziv, R., & Tishby, N. (2017). Opening the black box of deep neural networks via information. *ArXiv Preprint ArXiv:1703.00810*.
Silva, T., Matsuda, P.K., 2010. Practicing Theory in Second Language Writing. Parlor Press LLC.
Sinclair, J. (1984). Naturalness in Language. In: Aarts, J., Meijs, W. (Eds.), Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research. Rodopi, Amsterdam , pp. 203–210.
Smith, N.J., Levy, R., 2013. The effect of word predictability on reading time is logarithmic. Cognition 128 (3), 302–319.
Sprouse, J., 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. Biolinguistics. (Nicos) 1, 123–134.
Staples, S., & Dilger, B. (2018). *Corpus and repository of writing [Learner corpus articulated with repository]*.
Takano, S., Ichikawa, O., 2022. Automatic scoring of short answers using justification cues estimated by BERT. In: Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022. https://doi.org/10.18653/v1/2022.bea-1.2.
Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., & others. (2023). Llama 2: open foundation and fine-tuned chat models. *ArXiv Preprint ArXiv:2307.09288*.
Treffers-Daller, J., Parslow, P., Williams, S., 2016. Back to basics: how measures of lexical diversity can help discriminate between CEFR levels. Appl. Linguistics. https://doi.org/10.1093/applin/amw009 amw009.
Tunstall, L., Von Werra, L., Wolf, T., 2022. Natural Language Processing With Transformers. O'Reilly Media, Inc.

van Schijndel, M., Linzen, T., 2018. A neural model of adaptation in reading. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/d18-1499.

Vercellotti, M.Lou, 2019. Finding variation: assessing the development of syntactic complexity in ESL speech. Int. J. Appl. Linguistics 29 (2), 233–247.

Wang, G., Wang, H., Wang, L., 2022. Kolmogorov complexity metrics in assessing L2 proficiency: an information-theoretic approach. Front. Psychol. 13 https://doi.org/10.3389/fpsyg.2022.1024147.

Weigle, S.C., 2013. English as a second language writing and automated essay evaluation. Handbook of Automated Essay Evaluation. Routledge, pp. 36–54.

Wen, Q., Wang, L., Liang, M., 2005. Spoken and Written English Corpus of Chinese Learners. Foreign Language Teaching and Research Press.

Wilcox, E., Levy, R., Morita, T., Futrell, R., 2018. What do RNN language models learn about filler–gap dependencies?. In: Proceedings of the EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP https://doi.org/10.18653/v1/w18-5423.

Willems, R.M., Frank, S.L., Nijhof, A.D., Hagoort, P., van den Bosch, A, 2015. Prediction during natural language comprehension. Cereb. Cortex 26 (6), 2506–2516. https://doi.org/10.1093/cercor/bhv075.

Wilson, J., Roscoe, R., Ahmed, Y., 2017. Automated formative writing assessment using a levels of language framework. Assess. Writ. 34, 16–36.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A., 2020. Transformers: state-of-the-art natural language processing. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

Xiang, M., Kuperberg, G., 2015. Reversing expectations during discourse comprehension. Lang. Cogn. Neurosci. 30 (6), 648–672.

Xiao, C., Ma, W., Xu, S. X., Zhang, K., Wang, Y., & Fu, Q. (2024). From automation to augmentation: large language models elevating essay scoring landscape. *ArXiv Preprint ArXiv:2401.06431*.

Yang, W., Lu, X., Weigle, S.C., 2015. Different topics, different discourse: relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. J. Second. Lang. Writ. 28, 53–67. https://doi.org/10.1016/j.jslw.2015.02.002.

Zhang, X., Lu, X., 2022. Revisiting the predictive power of traditional vs. fine-grained syntactic complexity indices for L2 writing quality: the case of two genres. Assess. Writ. 51, 100597 https://doi.org/10.1016/j.asw.2021.100597.