AI Language Models: An Opportunity to Enhance Language Learning

Yan Cong

School of Languages and Cultures, Purdue University, West Lafayette, IN 47907, USA; cong4@purdue.edu

Abstract: AI language models are increasingly transforming language research in various ways. How can language educators and researchers respond to the challenge posed by these AI models? Specifically, how can we embrace this technology to inform and enhance second language learning and teaching? In order to quantitatively characterize and index second language writing, the current work proposes the use of similarities derived from contextualized meaning representations in AI language models. The computational analysis in this work is hypothesis-driven. The current work predicts how similarities should be distributed in a second language learning setting. The results suggest that similarity metrics are informative of writing proficiency assessment and interlanguage development. Statistically significant effects were found across multiple AI models. Most of the metrics could distinguish language learners’ proficiency levels. Significant correlations were also found between similarity metrics and learners’ writing test scores provided by human experts in the domain. However, not all such effects were strong or interpretable. Several results could not be consistently explained under the proposed second language learning hypotheses. Overall, the current investigation indicates that with careful configuration and systematic metrics design, AI language models can be promising tools in advancing language education.

Keywords: large language models; natural language processing; second language writing; automatic writing assessment; cosine similarity scores; distributional semantics; interlanguage development

1. Introduction

Rapid progress in large language models (LLMs) has enabled the deployment of many downstream natural language processing (NLP) applications. How can LLMs enhance education and language learning, particularly second language (L2) writing assessment? Despite LLMs’ broad popularity in several domains outside of computer science [1], to what extent these LLMs can be used to better our understanding of L2 writing development is a relatively unanswered question. On the application side, industry practitioners as well as researchers have been implementing LLMs such as BERT [2] and T5 [3] in L2 writing proficiency measurement. Systematic incorporation of larger and more recent LLMs in L2 writing research appears to be relatively underexplored. This study hopes to bridge the gap by utilizing multiple different LLMs for enhancing automatic essay scoring and improving our grasp of L2 writing development.

1.1. Scope and Overviews

This article intends to quantify L2 writing proficiency and index its development using LLM-derived similarity metrics. The current work presents the potential utilities and challenges of LLMs’ application in assessing L2 writing development. The current work attempts to fill in the research gap through a case study using different LLM-derived similarity metrics to benchmark writing proficiency in Chinese L2 learners studying English. There are three major components in the examination: (i) establishing the LLM similarities comparison baseline (native speakers/first language (L1) versus L2), (ii) investigating the sensitivity of LLM similarity metrics to L2 sub-groups with different proficiency levels,
and (iii) quantifying the effectiveness of LLM similarities in indexing L2 learners’ writing and overall proficiency. The current work carefully selected and compiled two written American English corpora (L1 and L2) from publicly available corpora. The results suggest that with cautious interpretations, LLM-derived similarity metrics are informative of L2 text processing and proficiency measurements. The hope is that the LLM-based approach can inspire L2 researchers to build their own research programs using LLMs.

The scope and focus of the current investigation concern second language learning. In second language acquisition and learning (SLA), a “second language” (L2) refers to any language learned after a person’s native language (L1). Studying SLA is crucial for understanding how individuals acquire and use additional languages, which has significant implications for educational practices, cognitive development, and social integration. Research in SLA helps educators develop effective teaching strategies, supports multilingualism and (linguistic) diversity, and enhances communication in increasingly globalized societies. Additionally, insights from SLA can inform policies and practices that facilitate language learning and proficiency in diverse populations. These motivate the current work to delimitate the examination of writings produced by L2 speakers.

1.2. Research Questions and Hypotheses

The current work aims to address three questions in this paper: (1) How effectively can LLM-based similarity metrics separate out L1 from L2 learners’ writing? (2) How effectively can similarity metrics distinguish proficiency levels for L2 learners’ essays? (3) What is the relationship between similarity metrics and L2 learners’ overall proficiency and writing test scores?

The overall hypothesis is that if LLM-derived similarities are effective and interpretable, such metrics should be able to tease apart L1 from L2 text, distinguish L2 learners’ proficiency levels based on L2 essays, and significantly correlate with L2 learners’ writing and overall proficiency scores. It is worth highlighting that the current work proposes LLM similarity as a composite score of text connectedness and repetition, which can serve as a proxy to the measurement of L2 writing cohesion. The hope is that English as a second language instructors and L2 researchers can utilize these LLM-derived similarities as an intermediate and assistive metric. Such metrics can (in-)directly inform educators how to help students improve writing coherence. The current work acknowledges that LLMs’ pre-training and architecture can be opaque and “black-boxy”, and hence they might not be “interpreted” in general human cognition aspects. This study defines “interpretability” as the extent to which LLM generated results can be explained by L2 learning hypotheses. In other words, an indicator of LLM-derived similarities showing interpretability is that such metrics suggest a pattern that can be translated into and accounted for by L2 learning hypotheses, in the sense that the pattern aligns with L2 writing hypotheses about text connectedness and repetition.

Concretely, the current work computed cosine similarity, the cosine angle of two LLM-derived numeric vectors in a multi-dimensional space. A cosine similarity close to 1 means two vectors are closely related semantically and syntactically. The current work views LLM-derived similarity as a measure of two aspects of discourse: (i) word and sentence connectedness and cohesion and (ii) syntactic and semantic repetitions. The current work proposes that LLM similarity in an L2 development context is a relative concept, and it can be interpreted as a spectrum of cohesion and repetition. High similarities indicate high text connectedness; hence, the writing is likely to be coherent. On the other hand, overly high similarities can be associated with large numbers of repetitions and with low variation in syntactic–semantic construction usage. Low similarities imply low repetition and overlapping in lexical and syntactic usage. However, extremely low similarities are likely indicators of incoherence and lacking connectedness. The current work examined similarities within one sentence (word-level metrics) and beyond one sentence (sentence-level metrics). The purpose is to characterize L2 learners’ writing using different measurement units.
The sub-hypothesis about LLM-derived cosine similarity is as follows: with different comparison baselines, the current work expects to find distinct LLM similarity patterns. First, when comparing L1 and L2 essays, it is predicted that there will be a lower similarity minimum but higher maximum in L2 produced text. Relative to L1 text, L2 speakers may repeatedly use a small cluster of words, due to their smaller vocabulary. Meanwhile, L2 learners’ interlanguage system may show less flexibility in syntax construction usage, leading to less-coherent and connected text. Therefore, assuming that L2 text contains more incoherence and repetition than L1 text, translated into similarity, the current work expects to find the L2 similarity minimum to be lower than that for L1 (more incoherence) and the L2 similarity maximum to be higher than that for L1 (more repetition). Additionally, the current work predicts higher variance in L2 text than in L1 text. This is because L2 production tends to show less stability and higher individual differences [4–6].

Second, when comparing L2 essays produced by speakers with different proficiency levels, the current work expects lower averaged similarities (e.g., median or mean) as learners’ proficiency increases. The current work hypothesizes that such decrease in averaged similarity in an L2 text would be attributed to vocabulary expansion and grammar improvement. An increase in lexical diversity can lead to a decrease in similarity median. This is because a change in lexical items will be reflected in embedding representations of the text. A pair of words or sentences with similar vocabularies gives rise to higher similarity than the same pair with different vocabularies. In other words, learners have more vocabulary to use as their proficiency improves; hence, similarities within and beyond sentences decrease. Similarly, as learners’ interlanguage system develops, they will most likely show more flexibility and diversity in using various syntactic constructions. As a consequence, L2 writing improvement implies that words and sentences in an essay become less “similar”, and averaged similarities may decrease. Put otherwise, a pair of sequences with different syntax should lead to lower similarities than the same pair with similar syntax.

Third, it is predicted that the magnitude and directionality of correlations between LLM metrics and proficiency scores will depend on the LLMs’ architecture. It is likely that certain LLMs will show greater sensitivity to L2 writing patterns than other LLMs, and that different LLMs will capture different aspects of writing proficiency, leading to different directionalities in the correlation.

2. Background

2.1. Computational Approach to L2 Writing Proficiency

L2 development in writing is multidimensional and multifaceted [6–13]. Measurements of L2 writing proficiency have focused on complexity, accuracy, and fluency, among other indices [5,14–16]. For example, [9] developed the L2 Syntactic Complexity Analyzer (L2SCA), including metrics such as degree of phrasal complexity, length of production unit, and so on. Other metrics, including type-token ratio and \( n \)-grams, have been validated in characterizing L2 proficiency writing [8,17–22].

Despite a large amount of insightful work quantifying L2 essay proficiency computationally, there are more and more empirical studies targeting recent and open-source LLMs such as Llama2 [23]. Such LLMs can generate fluent text and answer questions, and they excel in various downstream tasks, ranging from translation and summarization to text completion. Recent studies have attempted to incorporate LLMs for language education purposes. They focus on personalized and adaptive feedback, automated writing assessment, and generation of teaching content [24]. Studies have shown that LLM is a promising tool that can enhance students’ learning experience and improve educators’ work efficiency [24].

2.2. Contextualized Meaning Representations in LLMs

Research on meaning representations in computational linguistics is rooted in distributional semantics, the advancement of which has completely transformed the field of
natural language processing and understanding in the last decade. One of the key driving forces behind this remarkable transformation is the movement from static to contextualized vectors produced by LLMs. The main advantage of this development is that for contextualized LLMs, each word token in a specified (sentence) context receives a unique vector representation. The so-called “contextualized embeddings” are realized via multi-layer networks, where each word vector is learned in a way that the word’s distinct context activates different states, giving rise to distinct contextualized vector representations [25]. Recent LLMs for obtaining contextual embeddings are composed of a stack of transformer network layers [25]. Moreover, thanks to the attention algorithms of LLMs, instead of learning huge amounts of text, LLMs can learn to select which words are most relevant to the current word. Consequently, related aspects of the context structures become encoded in the output vector embeddings. This enables the contextualized meaning representations to capture multiple linguistic features [26], especially the context-sensitive aspects of word meaning [27]. Such capability addresses the issue of lack of context sensitivity in the static models, which ineffectively conflate all the word senses in one single vector [28].

Put another way, unlike static models, which have one singular vector per word token, the newer LLMs are sensitive to context, and can deliver a different vector for a word for every context in which that word appears. There is much evidence that these representations are modulated in intuitive and realistic ways by those contexts [29]. To summarize, LLMs learn word representations that are dramatically different from the ones we can obtain from earlier static models. Such advancement makes a difference in improving LLMs’ performance in text processing and understanding.

3. Materials and Methods
3.1. Text Preprocessing

The current work used the publicly available University of Pittsburgh English Language Institute Corpus (PELIC) [30] as the L2 data in the investigation. The pipeline of pre-processing and essay selection is illustrated in Figure 1. PELIC is a large L2 English learner corpus of written and spoken texts. It contains a broad range of proficiency levels, ranging from level 2, which is approximately equal to the Common European Framework (CEFR) A2, to level 5, which is equal to CEFR B2+/C1. The two intermediate levels, level 3 and level 4, correspond to CEFR B1 and B1+/B2, respectively. The current work focused on learners whose L1 was Chinese. The English Language Institute of PELIC did not regularly offer level 2 when they collected the data [30]; therefore, there were significantly fewer level 2 writing data (level 2 N = 7; level 3 N = 292; level 4 N = 478; level 5 N = 110). For this reason, the current work excluded level 2 learners. To keep consistency and reduce potential confounding variables caused by tasks and question types, only “allow_text” tasks and “paragraph_writing” question types were selected. These tasks and questions allow students to write an answer instead of choosing a word from a word list. Basic pre-processing was conducted to remove extra space, non-characteristic symbols, and citations and references. Sentence boundaries were automatically detected, based on punctuation (“.” “!” “...” “?”). The current work additionally curated a list of proper nouns that may contain “.”, where the period does not necessarily indicate a sentence boundary, for example, “Dr.” and “U.S.”. Further processing was conducted to ensure that sentence boundaries were meaningful. Since the dataset has been de-identified by PELIC and only “paragraph_writing” question types were chosen, there was not much identifiable personal information or information that could be used to trace personal identify in the text. The current work conducted an additional round of de-identification to extract and replace potential names and places, to ensure that privacy issues were handled appropriately.
After the screening, the current work collected 887 samples produced by 156 Chinese L2 learners of English. PELIC provided all 887 writing samples with the corresponding proficiency tests scores. Specifically, for the test scores, the current work selected “Writing_Sample”, a variable indicating in-house writing test scores with a scale of 1–6, for which at least two certified human raters assessed each essay. A higher “Writing_sample” number indicates higher writing proficiency. The current work also included the Michigan Test of English Language Proficiency (MTELP) score as the L2 learners’ overall proficiency scores. Particularly, the current work focused on the “MTELP_Conv_Score” for the total combined score. Descriptive statistics for the MTELP_Conv_Score and Writing_Sample are provided in the Supplementary Table S1.

To establish an interpretation baseline, the current work included an L1 corpus from the Michigan Corpus of Upper-level Student Papers (MICUSP). The MICUSP is a collection of language data developed for linguistic analysis purposes at the English Language Institute at the University of Michigan. The purpose of involving a separate L1 corpus was to quantify the extent to which LLM similarity metrics can tease apart L1 from L2 text. Once that baseline task is checked, one can further examine how to operationalize LLMs in understanding L2 text development stages and index L2 learners’ sub-groups’ writing. With that goal in mind, the current work randomly sampled ten essays produced by ten final-year undergraduate students. All ten students were native speakers of English. Their major was English. Similar cleaning to Figure 1 was conducted to remove non-speech symbols and redundant spaces. To align with the L1 dataset in terms of text length, each essay was sliced into short paragraphs. After the pre-processing, the current work collected 99 matched samples. Table 1 provides the text length information for the selected L1 and L2 corpora. The text length was measured by the number of words in the text.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>134.564</td>
<td>51</td>
<td>299</td>
<td>58.802</td>
</tr>
<tr>
<td>L2</td>
<td>162.647</td>
<td>54</td>
<td>298</td>
<td>60.693</td>
</tr>
</tbody>
</table>

As shown in Table 1, this matching of L1 and L2 samples in text length enabled a comparative analysis of text length and writing proficiency (text connectedness and repetition), facilitating the operationalization of LLMs in understanding L2 text development and indexing L2 learners' writing stages.

Figure 1. A flowchart illustrating text processing steps for the L2 dataset.
3.2. Technical Details

To obtain the LLMs’ contextualized embedding vectors, the current work used mini-cons [31], an open-source utility that provides a standard API for behavioral analyses of LLMs. The current work computed the cosine similarities of embedding vectors. Specifically, LLMs were employed to compute cosine similarity by embedding texts into high-dimensional vector spaces, which capture their semantic meanings. In evaluating the text connectedness and repetition of L2 learners’ English essays, LLMs convert words and sentences into vector representations. Cosine similarity can then be calculated between vectors of text chunks to measure their semantic closeness [32]. Higher cosine similarity scores indicate better coherence, reflecting how smoothly ideas transition between sentences or words [32]. Lower similarity, on the other hand, suggests higher lexical and syntactic diversity. Extremely high or low cosine similarity suggests repetition and incoherence (c.f. Section 1.2). This approach leverages the LLMs’ deep understanding of language nuances to assess the logical flow, connectivity, and repetition within the essays.

More formally, first, Equation (1) was used to calculate cosine similarity of vectors, then Equation (2) was used to compute the normalized cosine similarity. Cosine similarity quantifies how similar two vectors are, regardless of their size, as illustrated in Formula (1). A cosine similarity of $-1$ indicates two strongly opposite vectors, 0 means independent (orthogonal) vectors, and a similarity of 1 refers to (positive co-linear) identical vectors. Intermediate values are generally used to assess the degree of similarity.

$$\text{CosineSimilarity}(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|}$$  \hspace{1cm} (1)

$$\text{NormalizedCosineSimilarity}(x, y) = \frac{\text{CosineSimilarity}(x, y) - (-1)}{1 - (-1)}$$  \hspace{1cm} (2)

In Formula (1), $x \cdot y$ is the dot product of the vectors $x$ and $y$. For different metrics, the two vectors represent different linguistic units. For similarity measurement within a sentence, the vectors represent words, whereas for those beyond a sentence, the vectors represent sentences. $\|x\|$ and $\|y\|$ refer to the length of the two vectors $x$ and $y$. $\|x\| \cdot \|y\|$ is the cross product of the two vectors. The cosine similarity is robust and widely used [32]. Even when the two similar vectors are far apart by other distance measures because of their size, they could still have a smaller cosine angle between them. In general, the smaller the cosine angle, the higher the similarity of the two vectors. When plotted on a multi-dimensional space, the cosine similarity captures the orientation (the angle) of the two vectors, and not the magnitude.

Formula (2) is the normalized version of (1). The $\text{CosineSimilarity}(x, y)$ value is in the range $[-1, 1]$. The current work normalized $\text{CosineSimilarity}(x, y)$ using the minimum and the maximum values of cosine similarity throughout the paper, so that one could make direct comparisons across LLMs and across corpora. The current work focused on cosine similarity, as opposed to other techniques such as Hamming distance and Jaccard similarity, because the cosine similarity is one of the most used in natural language processing [32]. The hope is to systematically explore and compare different similarity techniques for future research.

3.3. Experimental Design

To investigate how interpretable and effective LLMs are in facilitating L2 research and benchmarking L2 writing development, the current work analyzed several variables. First, the current work used LLM-derived similarities as the predictor variables. Contextualized embeddings were derived from three different types of transformer LLMs: bidirectional BERT-large-uncased with 336 million parameters [2], encoder–decoder T5-large with 770 million parameters [3], and unidirectional Llama2-7B with 7 billion parameters [23]. The selection of these LLMs was motivated by the intention to systematically examine pre-trained LLMs’ efficacy in automatic L2 essay assessment and in indexing L2 development.

The calculation of similarity statistics is adapted from [33], which surveyed, identified, and examined NLP coherence measures in previous studies. These metrics are scalable, meaning they are easier to apply to large-scale datasets. Each similarity metric
was normalized in order to make comparisons across different LLMs and different datasets. Following [33], in addition to the mean, the current work chose the median, Q5 (5% quantile), Q95 (95% quantile), and interquartile range (IQR) of similarity metrics, since they are robust to outliers and can represent the variance of the distribution. All the independent and dependent variables are summarized below:

(a) Independent variables in the format of LLM_measurement unit_statistics. For example, bert_mv5_q5 refers to a BERT-large-uncased-derived metric, which calculates the average similarity of each pair of words in a 5- or 10-word moving window (mv5/10) for the whole sample, then takes the Q5 for the sample-aggregated mv5/10:

- **LLMs**: BERT-large-uncased; T5-large; Llama2.
- **Measurement unit**:
  - **Within a sentence**:
    - **mv5/10**: average similarity of each pair of words in a 5-word or 10-word moving window.
    - **k1:10**: pairwise word-to-word similarity at k inter-word similarity, with k ranging from 1 to 10.
  - **Beyond a sentence**:
    - **foc**: first-order coherence, referring to the cosine similarity of adjacent sentence pairs.
    - **soc**: second-order coherence, referring to the cosine similarity of sentence pairs with one intervening sentence.
- **Statistics**: Mean; Median; Q5; Q95; IQR.

(b) Dependent variables, validated and provided by the L1 and L2 corpora:

- **MTELP_Conv_Score**: MTELP total combined proficiency score.
- **level_id**: indicates the level of the speaker.
- **Writing_Sample**: in-house writing test score (scale of 1–6; a bigger score is associated with higher writing proficiency).

In addition to the choices motivated and inspired by [33], the current work selected these word- and sentence-level metrics because they have psychological and educational relevance. In L2 writing assessments, using metrics derived from LLMs, such as BERT, T5, and Llama2, combined with statistical measures (mean, median, quantiles, and interquartile range) and specific metrics within and beyond sentences provides a multi-faceted approach to evaluate language proficiency and index the interlanguage system’s development. Metrics such as the moving window average similarity (mv5/10) and pairwise word-to-word similarity (k1:10) assess the cohesion (text connectedness and repetition) within a sentence, highlighting how consistently a learner uses their L2 knowledge. First-order coherence (foc) and second-order coherence (soc) measure how well sentences connect logically and flow in a broader context. These measures are psychologically and educationally relevant, also because they provide detailed insights into a learner’s ability to produce text that is not only grammatically correct but also contextually and semantically coherent. Additionally, these metrics can track and index L2 development stages. This holistic assessment can inform targeted instructional strategies and address specific areas of difficulty in language learning.

4. Results

4.1. LLM Similarities Detect L2 and L1 Writing

To address how effectively and reliably LLM-derived similarities can separate out L2 and L1 writing, the current work constructed logistic regression models for each LLM-derived metric, with the similarity metrics as predictors predicting whether an essay is written by an L1 or L2 speaker. The current work plotted estimates of the logistic regression results for each LLM’s word- and sentence-level metrics in Figure 2 and Figures S1–S8 in the Supplementary Materials. The blue neutral line illustrates the vertical intercept, indicating no effect. The figures demonstrate the estimated values, and the asterisks represent the
significant level of the p-values. The horizontal line with brackets represents a 95% confidence interval. For illustration purposes and space considerations, only some of the figures are displayed, with the rest of the results presented in the Supplementary Materials.

Figure 2. BERT-large-uncased mv5/10 metrics’ estimates of predicting whether a text is produced by an L2 learner. Significance notation: * p < 0.05; *** p < 0.001. The red line indicates that the standardized beta coefficients values are positive; and the orange line indicates that the values are negative.

4.1.1. Word-Level Metrics

Regarding the word-level metric mv5/10 in BERT, Figure 2 shows the coefficient Q5 predicted change (in log-odds of being L2) when BERT mv10 cosine similarity Q5 was increased. For example, the mv10_q5 coefficient was 0.78, meaning that the odds of an essay being characterized as L2 were predicted to be 2.18 times higher than it being characterized as L1 for essays with high mv10_q5 ($e^{0.78}$ = 2.18). This finding was not as predicted. However, it was as predicted that mv5_q95 was significantly higher in L2 than in L1, suggesting that an essay with higher similarity Q95 (hence more repetition) was more likely to be captured as L2 rather than L1. mv5_q95 showed a wider 95% CI than mv10_q5, suggesting that the latter is a more reproducible metric. The main pattern that emerged was that a given text with higher BERT mv5_q95 and mv10_q5 was more likely to be produced by an L2 learner than by an L1 learner. Given that predictors were standardized, one can interpret their values as relative effect sizes: the mv10_q5 effect was stronger than mv5_q95.

Regarding pairwise word similarity with an inter-word distance of k in BERT-large-uncased (Figure S1), it was found that a given text with significantly higher k1_q5, k2_q5, k3_q5, k2_iqr, k2_q95, and k7_q95 was more likely to be produced by an L2 learner. On the other hand, an essay with lower k3_q95 and k10_q5 was more likely to be produced by an L2 learner. The results suggest that the k1_q5 effect was the strongest. Overall, half of the results were as predicted, for which a lower Q5 was found in L2, and higher IQR and Q95 were found in L1.

For Llama2 mv5/10 metrics (Figure S3), the results indicated that only mv5_q95 showed significance, with lower mv5_q95 in L2 than in L1 essays. This suggests that when measuring similarity locally (in a 5-word window within a sentence), L2 essays gave a lower Q95 than L1, which was not as predicted. For Llama2 k1:10 metrics (Figure S4), findings showed that higher k8_q95 and k7_iqr were more likely to be detected in L2 than in L1 essays, which was as predicted. Surprisingly, lower k3_q95, k2_q95, and k1_q95 were...
more likely to be predictors of L2 than L1 writings. This also indicates that, in a smaller local window \((k < 5)\), lower similarity extreme values (e.g., Q95) are more likely to be found in L2 than in L1.

As for T5-large metrics generated within a sentence, it was found that an essay with higher \(mv10_q5\) and \(mv5_iqr\) was more likely to be produced by an L2 learner, whereas a given text with lower \(mv5_q95\) was more likely to be produced by an L2 learner (Figure S6). The current work interprets this as the following: an essay with high Q5 and variation (IQR) in similarities in a given text window (5 or 10 words long) is more likely to be produced by an L2 learner than by an L1 speaker. In contrast, an essay with a lower Q95 in a 5-word window is more likely to be produced by an L2 learner than by an L1 speaker, which is not in line with the prediction. Among all the metrics, \(mv5_q95\) had the strongest effect.

Regarding T5-large k1:10 metrics (Figure S7), the findings showed that a given text with higher \(k10_q95\) but lower \(k4_q95\), \(k2_q95\), \(k8_q5\), \(k1_q95\), and \(k3_q95\) was more likely to be produced by an L2 learner than by an L1 speaker. This result is partially as predicted. The findings revealed that \(k3_q95\) showed the strongest effect and a narrow 95% CI.

4.1.2. Sentence-Level Metrics

For the metrics that go beyond a sentence, the findings in BERT suggest that essays with higher \(soc_q95\) but lower \(soc_q5\) were more likely to be produced by an L2 learner (Figure S2). This finding was as expected. It indicates that, when using BERT-large-uncased to derive contextualized embedding vectors, extremely low sentence-vector similarity is associated with L2-generated text. The interpretation is that there is low sentence-relatedness in L2 writing, compared to L1 writing. This is likely due to L2 learners’ vocabulary limitations. Ref. [34] suggested that variations in a vocabulary’s frequency and part of speech features can have a strong influence on the performance of contextualized embeddings. Meanwhile, extremely high similarities were found in L2 text, suggesting more repetition in L2 than in L1. On the other hand, it was also found that L2 text showed lower \(soc_iqr\), which was not predicted.

For Llama2 sentential metrics (Figure S5), it was found that lower \(soc_iqr\) and \(foc_q5\) were more likely to be detected in L2 writing than in L1. Lower Q5, which was predicted, can be attributed to more incoherence or less text connectedness in L2 writing, compared to essays produced by L1 speakers. Additionally, \(soc_iqr\) gave a narrower 95% CI and, hence, higher reproducibility than \(foc_q5\).

Regarding T5-large sentence similarity or pairwise sentence similarity with an intervening sentence (Figure S8), the finding showed that T5 sentential metrics were the most interpretable and informative among all the LLM metrics in teasing apart L1 and L2 text. It was found that, as predicted, when using T5-large to generate contextualized embedding vectors, an essay with higher \(soc_q95\) but lower \(foc_iqr\), \(soc_iqr\), \(foc_q5\), and \(soc_q5\) was more likely to be produced by an L2 learner than by an L1 speaker.

4.1.3. Regression Model Comparison

The current work reports the AUC value (the area under the ROC (receiver operating characteristic) curve) for each logistic regression model in Table 2. The baseline would be AUC = 0.5. The AUC of the \(bert_mv510\) model was 0.91, suggesting that the model had the explanatory power to predict whether an essay was produced by an L2 learner. The AUC of the regression model using T5 \(mv5/10\) metrics gave the lowest performance, with an AUC of 0.66, indicating above-chance explanatory power. The best model was the regression using BERT k1:10 metrics (AUC = 0.95), indicating strong explanatory power, although not all the BERT k1:10 metrics were interpretable and supportive of the hypothesis.
Table 2. AUC of logistic regression models classifying L1 and L2.

<table>
<thead>
<tr>
<th></th>
<th>BERT</th>
<th>Llama2</th>
<th>T5</th>
</tr>
</thead>
<tbody>
<tr>
<td>mv5/10</td>
<td>0.91</td>
<td>0.79</td>
<td>0.66</td>
</tr>
<tr>
<td>k1:10</td>
<td>0.95</td>
<td>0.89</td>
<td>0.92</td>
</tr>
<tr>
<td>foc/soc</td>
<td>0.7</td>
<td>0.72</td>
<td>0.73</td>
</tr>
</tbody>
</table>

With respect to specific word and sentential metrics, it was found that k1:10 appeared to be more effective than mv5/10 or foc/soc in detecting L2 and L1 essays. This is probably because there were, overall, more k1:10 predictor variables (10 variables) than mv5/10 or foc/soc predictors (two variables each), potentially improving the models’ performance.

With respect to LLMs, the results showed that BERT gave a better overall performance than Llama2 and T5, when predicting whether an essay was produced by L2 or L1 speakers.

To summarize, it was found that similarity metrics within a sentence showed mixed results for all the LLMs. Although significant effects were found in several measurements, not all of them showed patterns as predicted. For similarity metrics that go beyond one sentence, T5’s results were precisely as predicted, showing the highest interpretability. Overall, real world analysis gave a more sophisticated picture than what one would predict theoretically. The findings suggest that all LLM metrics showed some significance in detecting L2 writing from L1, and their interpretability varied. The current work concludes that LLMs, especially T5-large foc/soc, are potentially effective in establishing the interpretation baseline and identifying L2 text features. The hope is that the findings can shed light on LLM metric selections for L2 text assessment purposes.

4.2. LLM Similarities Index L2 Proficiency Levels

To address LLM metrics’ effectiveness in distinguishing L2 learners’ proficiency levels (level 3 versus 4, level 4 versus 5, and level 3 versus 5), multiple Welch t-tests with Bonferroni corrections were conducted. The results are grouped by metrics (Figures 3 and 4).

![Figure 3. LLM similarities mv5/10 index L2 proficiency levels. Significance notation: ** p < 0.01; *** p < 0.001; **** p < 0.0001.](image-url)
4.2.1. Word-Level Metrics

For the mv5/10 metrics using word as a measurement unit in a five- or ten-word window (Figure 3), the findings showed significant effects in all LLMs except for T5 mv5, suggesting the informativeness of this metric in certain LLMs. All LLMs showed a decrease in similarity from level 3 to 5. When using Llama2, there was a significant increase in similarity from level 3 to 4.

For the k1:10 metrics using word as a measurement unit in a word-pair format (Figure 4), a significant comparison effect was found in all LLMs except for t5 k3. Specifically, when using BERT and T5, there was a clear pattern, showing that similarities decrease...
from lower to higher levels. When Llama2 was used, the pattern mostly held, and an increase in similarity was found from level 3 to 4. This indicates that for local measurements within a sentence, as L2 learners’ writing evolves, there are progressively lower similarities. This decrease in similarity is probably attributed to the expansion of their vocabulary.

4.2.2. Sentence-Level Metrics

For the foc/soc metrics using sentence as a measurement unit (Figure 5), it was found that there is a significant increase in foc sentence similarity from level 3 to 5 using Llama2. Further, the findings suggest a significant decrease in foc and soc from level 3 to 4 and from level 3 to 5 when using T5. This indicates that when using Llama2, with causal language modeling, L2 learners’ writings show higher sentence similarities as their proficiency improves. In contrast, the inverse was found when using T5, with a different framework such as text-to-text transfer transformer.

![Figure 5. LLM similarities foc/soc index L2 proficiency levels. Significance notation: * p < 0.05; ** p < 0.01; **** p < 0.0001.](image)

In summary, the findings suggest that LLMs are generally effective in distinguishing L2 learners’ proficiency levels, based on similarity metrics derived from word pairs and sentence pairs. Statistical significances were found in multiple LLMs metrics. However, it is also worth pointing out that the results are a mixture of findings that (mis-)align with the predictions. As L2 learners’ proficiency level develops and improves, text-vector similarities are not always increased or decreased. The current work takes this to mean that non-linearity and multi-dimensionality are characteristics of L2 text development. The statistical modeling provides recommendations on which similarity metrics are effective and which are insensitive in capturing L2 development. The hope is that these systemic examinations will inspire L2 text researchers in selecting the appropriate metrics.

4.3. LLM Similarities Correlate with Overall Scores and Writing Scores

To study the relationship between LLM-derived similarity metrics and L2 learners’ writing and overall proficiency scores, Pearson correlation analyses were conducted. For illustration purposes and space consideration, the current work reports part of the Pearson correlations Rho in Table 3, with significance asterisks in superscripts. To reduce the Type I error probability, a lower significance level (alpha = 0.035) was set. The complete report can be found in the Supplementary Table S2.
Table 3. Pearson correlation coefficients of LLM similarities median and IQR and L2 learners’ proficiency test scores. Significance notation: * $p < 0.035$; ** $p < 0.01$; *** $p < 0.001$.

<table>
<thead>
<tr>
<th></th>
<th>BERT Overall</th>
<th>Writing</th>
<th>Llama Overall</th>
<th>Writing</th>
<th>T5 Overall</th>
<th>Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>foc_iqr</td>
<td>0.03</td>
<td>−0.01</td>
<td>foc_iqr</td>
<td>0.02</td>
<td>foc_iqr</td>
<td>0.03</td>
</tr>
<tr>
<td>soc_iqr</td>
<td>−0.05</td>
<td>0.06 **</td>
<td>soc_iqr</td>
<td>−0.02</td>
<td>soc_iqr</td>
<td>−0.02</td>
</tr>
<tr>
<td>soc_median</td>
<td>−0.02</td>
<td>0.05</td>
<td>soc_median</td>
<td>0.01</td>
<td>soc_median</td>
<td>−0.09 ***</td>
</tr>
<tr>
<td>mv10_iqr</td>
<td>−0.03</td>
<td>−0.01</td>
<td>mv10_iqr</td>
<td>0.03</td>
<td>mv10_iqr</td>
<td>−0.01</td>
</tr>
<tr>
<td>mv10_median</td>
<td>−0.15 ***</td>
<td>−0.05</td>
<td>mv10_median</td>
<td>0.07 **</td>
<td>mv10_median</td>
<td>−0.11 ***</td>
</tr>
<tr>
<td>mv5_iqr</td>
<td>0.01</td>
<td>0.03</td>
<td>mv5_iqr</td>
<td>0.03</td>
<td>mv5_iqr</td>
<td>0.04</td>
</tr>
<tr>
<td>mv5_median</td>
<td>−0.13 ***</td>
<td>−0.04</td>
<td>mv5_median</td>
<td>0.09 ***</td>
<td>mv5_median</td>
<td>−0.05 *</td>
</tr>
</tbody>
</table>

For the relationship between BERT and writing sample scores, positive correlations in foc_median were found. For BERT and the overall proficiency scores, negative correlations in mv10_median and mv5_median were found. Similar patterns were found when using T5. Different directionalities were found when using Llama2, where positive correlations were identified in almost all the metrics, except for soc_iqr.

To sum up, significant correlations were found between LLM text similarity metrics and L2 learners’ overall proficiency and their writing test scores, suggesting that LLM similarity metrics are potentially effective in characterizing L2 learners’ writing and overall proficiencies. The results mostly suggest that correlation direction depends on the LLM and the specific metrics. It depends on LLMs and specific metrics in terms of whether more-proficient L2 learners’ texts are likely to be associated with low similarity scores. Most of the effect sizes were weak, even though many text similarity metrics were found to be statistically significant.

5. Conclusions and Discussion

This investigation used LLMs to generate contextualized word- and sentence-embedding vectors. Normalized cosine similarities were computed for word/sentence vector pairs to quantify L2 text proficiency. The results showed that LLM similarity metrics are potentially useful in indexing L2 writing development. Several significant effects were found across LLMs, but not all of these were as predicted; hence, not all LLM-generated results were interpretable under the current hypotheses. The current work made recommendations as to which metrics specifically would be appropriate and generalizable. The LLM-based NLP pipeline is not only relevant and timely with respect to the L2-studies community, but also informative and innovative with respect to general text processing. The hope is that the systematic analyses and evaluation of LLM-derived similarities can inspire researchers to develop a cutting-edge text analytics methodology.

5.1. Interpreting LLM Similarity Scores in an L2 Setting

The comparison of L1 and L2 writing gave a complicated picture. Relative to L1 writing, L2 writing did not always behave as hypothesized. The current work predicted a higher maximum, lower minimum, and higher IQR in L2 writing (c.f. Section 1.2). Only T5 sentence-level metrics showed interpretable results that were in line with the expectation. It is inferred that several factors could contribute to this result. First, although native English speakers (i.e., L1 speakers) typically have a more extensive and intuitive grasp of the language, leading to higher diversity and lower repetition and similarity in their writing samples, their familiarity with idiomatic expressions, colloquial expressions, and linguistic conventions may result in more standardized, connected, and cohesive texts, leading to higher similarity scores. Further, the educational context and writing instruction provided to native English speakers may emphasize coherence and adherence
to standard language norms. This emphasis on standardization and conformity could result in more consistent and similar writing styles among native learners compared to L2 learners, who may receive instruction tailored to the individual L2 learner groups’ language acquisition and proficiency development. Overall, these factors may have complicated the experiments and the findings. It is likely that more fine-grained experiments and larger datasets can further tease apart the interplay between text connectedness and repetition, hence demystifying LLM similarity scores in L2 writing analyses.

Regarding comparisons within L2 learners’ writings, the results show that, relative to low-proficiency L2 learners’ writing, high-proficiency learners’ writing showed lower similarity scores in almost all the LLM metrics, although such decrease was not always consistent across L2 proficiency levels. This is partially as predicted (c.f. Section 1.2). The current work considers several factors in interpreting and explaining these results. High-proficiency L2 learners might demonstrate a more extensive vocabulary and varied sentence structures. This linguistic diversity and mastery of more vocabulary and syntax can lead to lower similarity scores. Although task selections were controlled and focused on the paragraph writing task, the current work did not examine specific prompt effects. It is possible that the similarity effects are related to different prompt questions under which writing samples were produced. High-proficiency L2 learners may have been tasked with proportionally more open-ended prompts, encouraging individual expression and contributing to variability in texts, hence lower similarity scores.

Why are there differences among the three LLMs in these L2 writing assessment tasks? The current work speculates the following: the differences among BERT, Llama2, and T5 arise from variations in their architectures, training datasets, and training objectives. BERT, a bidirectional transformer, is pre-trained on BooksCorpus and English Wikipedia, using a masked language model objective [2]. T5, based on a text-to-text framework, is trained on the Colossal Clean Crawled Corpus (C4) and handles various NLP tasks as text generation problems [3]. Llama2, a fine-tuned transformer, utilizes diverse internet-based data, including recent content, and balances understanding and the generation of tasks through specific training adjustments [23]. These distinctions result in unique performance variations across L2 writing assessment tasks. It is likely that the text-to-text framework is particularly suitable for L2 text detection at the sentence level, enabling T5 to have a more sensitive LLM in L1- and L2-writing differentiation. On the other hand, masked language modeling seems more sensitive for word-level measurements, enabling BERT to have a more appropriate LLM in writing assessment within a sentence (c.f. Table 2).

5.2. LLM Implications in Language Learning and Teaching

The application of LLMs for assessing L2 text proficiency has significant implications for both learning and teaching. By utilizing LLMs to generate contextualized word- and sentence-embedding vectors and computing normalized cosine similarities, the study offers a novel approach to evaluating L2 proficiency levels.

For learners, the findings suggest the potential for personalized feedback and tailored language instruction. LLM-based metrics can provide learners with detailed insights into their language development, highlighting areas of strength and areas requiring improvement. By leveraging these metrics, educators can design targeted interventions to address specific linguistic challenges and support more effective language acquisition strategies.

Further, the recommendations for appropriate and generalizable metrics offer valuable guidance for language instructors. By incorporating LLM-based assessments into their pedagogical practices, educators can implement evidence-based teaching methodologies that align with the principles of communicative language teaching. This integration of technology-enhanced assessment tools can foster a more dynamic and interactive learning environment, promoting learner engagement and motivation.

Additionally, the relevance of the LLM-based NLP pipeline extends beyond individual language learners and classrooms. It has implications for curriculum development, assessment design, and the broader discourse surrounding language education policy.
and practice. By demonstrating the utility of advanced text-analytics methodologies in evaluating L2 proficiency, the study contributes to the ongoing dialogue on best practices in language teaching and assessment.

5.3. AI Tool Usage in Education

The implementation of AI models in language learning has several educational implications. First, LLMs can facilitate personalized and adaptive learning experiences by analyzing vast numbers of textual data to provide tailored feedback and recommendations to learners. Through their ability to generate contextualized word and sentence embeddings, LLMs can identify patterns in language usage, pinpoint areas of difficulty for individual learners, and offer targeted interventions to support their language acquisition journey. Second, LLMs can empower educators with powerful tools for assessment and evaluation. By leveraging LLM-derived similarity metrics, educators can more accurately gauge students’ language proficiency levels, track their progress longitudinally, and identify domains for improvement. This not only enhances the effectiveness of assessment practices but also enables educators to design more-targeted and impactful instructional strategies. Further, the integration of LLMs into educational technology platforms can facilitate the development of innovative learning resources and environments. LLMs can power intelligent tutoring systems, virtual language assistants, and automated grading systems, providing learners with immersive and interactive learning experiences that are responsive to their individual needs and preferences.

However, it is important to acknowledge the challenges and limitations associated with the integration of LLMs in education, such as concerns regarding data privacy, algorithmic bias, and the digital divide. Additionally, while LLMs offer powerful capabilities for analyzing and understanding language, they should be complemented with pedagogical expertise and human-centered approaches to education.

In summary, LLM and AI technologies have the potential to revolutionize education by enhancing language learning and teaching practices, facilitating personalized learning experiences, and enabling more accurate assessment and evaluation. By harnessing the power of LLMs, educators can unlock new opportunities for improving learning outcomes and fostering linguistic diversity and inclusion in educational settings.

5.4. Limitations and Future Directions

There are a few findings that are worth further discussion. As expected, inconsistencies were found in similarity directionalities within and across LLMs. Furthermore, for the correlational models, most of the effect sizes were small. In future research, the plan is to include static models, such as LSA or GloVe, to make comprehensive comparisons. It is likely that the alleged superiority of contextualized LLMs over the static models is not as real as has been claimed [34]. Static models have been shown to surpass BERT representations in most tasks that isolate contexts. Systematic evaluation of contextualized embeddings produced by LLMs in [34] has indicated that contextualized embeddings generated by BERT mostly underperform. Additionally, the current work acknowledges that the analysis is limited to the use of pre-trained LLMs for sentence-pair similarity calculation, while fine-tuning is left for future work. Once these types of embedding experiments become more standard and streamlined, the current work maintains that it will be necessary to take a fine-tuning approach in the assessment of L2 writing. Fine-tuning on a curated L2 corpus would lead to a more holistic and effective LLM in benchmarking L2 writing.

The current work is aware that the training sets for LLMs are continually updated, and that better and more-powerful LLMs continue to emerge. However, the current work maintains that the proposed approach and the findings of this study are generalizable to new LLMs with similar architectures. The hope is that the experimental design can provide recommendations for which type of LLMs and which kinds of metrics to focus on in future research. This is one of the reasons why the current work chose three LLMs with different
architectures. Although there are constantly new LLMs coming out, the major types of architectures are not infinite, and the three types of LLM the current work selected in this study cover the three major types of transformer architectures. Therefore, it is suggested that the approach has the potential to generalize to more recent LLMs.

The plan is to comprehensively compare static and contextualized metrics’ performances in indexing and quantifying L2 text development. This will also inform us as to which metrics and which LLMs are robust and accessible to society for text and discourse. Despite the claimed impressive performance of contextualized embeddings over static embeddings [1,35], those findings are not reproducible, and static embeddings can still be useful, considering the simplicity of static models and the much greater amount of computation power that is needed to train contextual LLMs [34]. The hope is to justify these hypotheses and develop accessible text analytics for future research.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/informatics11030049/s1, Figure S1: BERT-large-uncased k1:10 metrics’ estimates predicting whether a text is produced by an L2 learner. Significance notation: * p < 0.05; ** p < 0.01; *** p < 0.001; **** p < 0.0001; Figure S2. BERT-large-uncased foc/soc metrics’ estimates predicting whether a text is produced by an L2 learner. Significance notation: * p < 0.05; ** p < 0.01; *** p < 0.001; **** p < 0.0001; Figure S3. Llama2 mv5/10 metrics’ estimates predicting whether a text is produced by an L2 learner. Significance notation: * p < 0.05; ** p < 0.01; *** p < 0.001; **** p < 0.0001; Figure S4. Llama2 k1:10 metrics’ estimates predicting whether a text is produced by an L2 learner. Significance notation: * p < 0.05; ** p < 0.01; *** p < 0.001; **** p < 0.0001; Figure S5. Llama2 foc/soc metrics’ estimates predicting whether a text is produced by an L2 learner. Significance notation: * p < 0.05; ** p < 0.01; *** p < 0.001; **** p < 0.0001; Figure S6. T5-large mv5/10 metrics’ estimates predicting whether a text is produced by an L2 learner. Significance notation: * p < 0.05; ** p < 0.01; *** p < 0.001; **** p < 0.0001; Figure S7. T5-large-uncased k1:10 metrics’ estimates predicting whether a text is produced by an L2 learner. Significance notation: * p < 0.05; ** p < 0.01; *** p < 0.001; **** p < 0.0001; Figure S8. T5-large mv5/10 metrics’ estimates predicting whether a text is produced by an L2 learner. Significance notation: * p < 0.05; ** p < 0.01; *** p < 0.001; **** p < 0.0001; Table S1. Dependent variable descriptive statistics; Table S2. Pearson correlation coefficients between LLM metrics and writing and overall proficiency scores. Significance notation: * p < 0.035; ** p < 0.01; *** p < 0.001; **** p < 0.0001.

**Funding:** This research was funded by the College of Liberal Arts, School of Languages and Cultures, Purdue University, Faculty Start-Up fund, grant number F10072818.05.001.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.


**Acknowledgments:** The author acknowledges Cameron Pilla and Emily Tumacder for their help with compiling the datasets and optimizing the NLP pipeline. The author thanks Sunny X. Tang, Phillip Wolff, and Emmanuelle Chersoni for their inspiration and fruitful discussions. All errors remain mine.

**Conflicts of Interest:** The author declares no conflicts of interest.

**References**


17. Bulté, B.; Roothoot, H. Investigating the interrelationship between rated L2 proficiency and linguistic complexity in L2 speech. *System* 2020, 91, 102246. [CrossRef]


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.