Jeremiah Henderson

Second Language Vocabulary Assessment

Use this space for legal copy.
State whether the information is confidential and how it's intended to be used.

Introduction

- “Exploiting the English Vocabulary Profile for L2 word-level vocabulary assessment with LLMs”

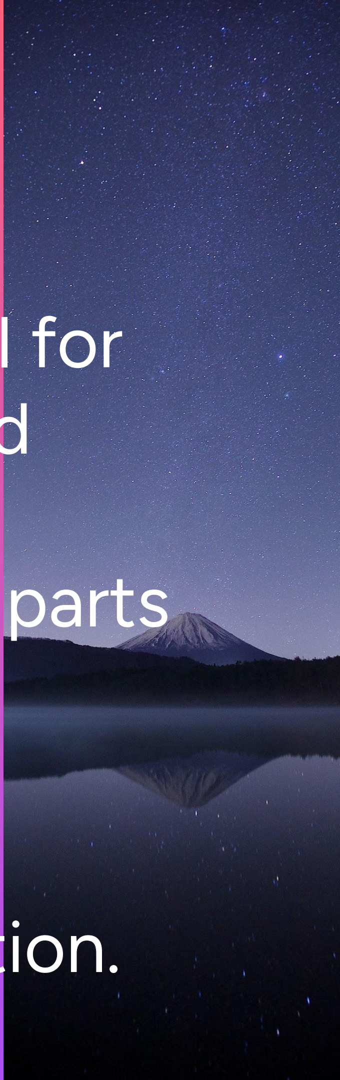
nill, Mark J. F.

Introduction

Importance: This helps us improve how we measure language proficiency not just with grammar but with what words someone uses and in what context. It gives us a true measure of the depth of someone's vocabulary too.

Background

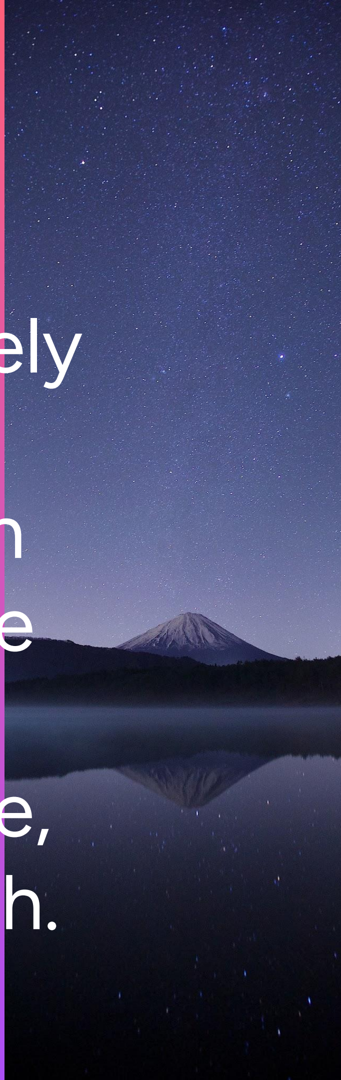
- The authors state that vocabulary is essential for language proficiency and communication and traditional automated systems only assess vocabulary in context independent ways like parts of speech rules or frequency counts.
  These systems miss things like polysemy (different meanings for words and phrases), multi-word expressions and contextual variation.



Research Question

Can Large Language Models (LLMs) accurately assess second language learners vocabulary proficiency at the word level in context when combined with the English Vocabulary Profile (EVP)?

The authors aim to create a context-sensitive, replicable and accurate assessment approach.



Early Studies

- Some early approaches where done in the 1980s. One such study for example is the study run by Richard Anderson and Peter Freebody where they aimed to assess specific aspects of vocabulary use and knowledge. Another was the study run by John Read in 2013 where they shifted from knowledge of grammatical and lexical elements to performance in real-world-like tasks.

Key Concepts

- During the different studies two major dimensions of vocabulary knowledge came to be. One being Lexical diversity which is how many different words learners use and Lexical sophistication which is how advanced or rare a word is. This idea was formulated in studies by Melissa Berk, Shilo Drake and Kurtis Foster where they focused on the depth of lexical knowledge and presence of relatively rare or uncommon words in different writing samples.

Key Concepts

- English Vocabulary Profile (EVP) - A public reference that maps words, phrases and idioms to the CEFR levels of proficiency and gives learner and dictionary examples. This was created by Annette Capel in 2015. Its grounded in tons of research using the Cambridge Learner Corpus which has a growing collection of exam scripts written by learners worldwide.

Key Concepts

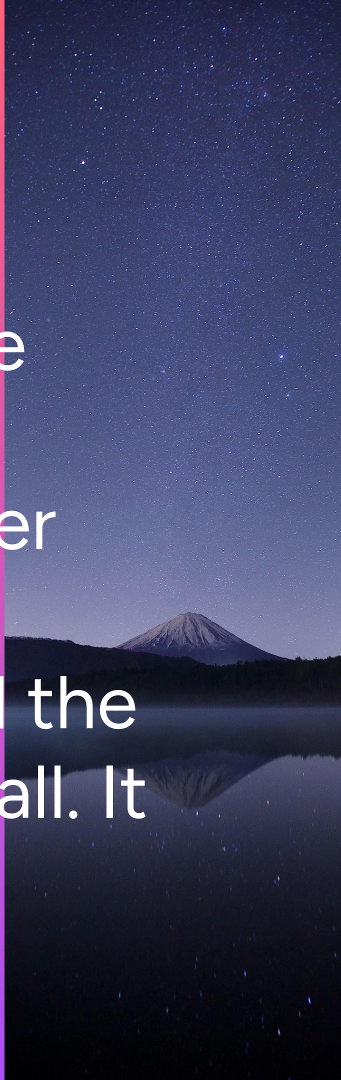
- The CEFR(Common European Framework of Reference for Languages) is a standardized guideline for language ability and breaks it down into levels from A1 to C2. A1 being the lowest and C2 being the highest level. A1 is the most basic user of the language and C2 is a proficient user of the language.

Methodology

- In the first experiment, LLMs were asked to pick the correct meaning of polysemous words from examples given by the EVP. They tested models like: GPT-4o, 4o mini, Llama 3.1 and Qwen 2.5. They found that GPT-4o performed the best with an accuracy of about 84%.

Methodology

- In the second experiment, they tested if the LLMs could accurately predict the CEFR proficiency levels of each word in the learner sentences based on EVP entries.
  They found that Qwen 2.5 outperformed all the other models and had a 87% accuracy overall. It was also found that PoS based systems struggled with ambiguous words.

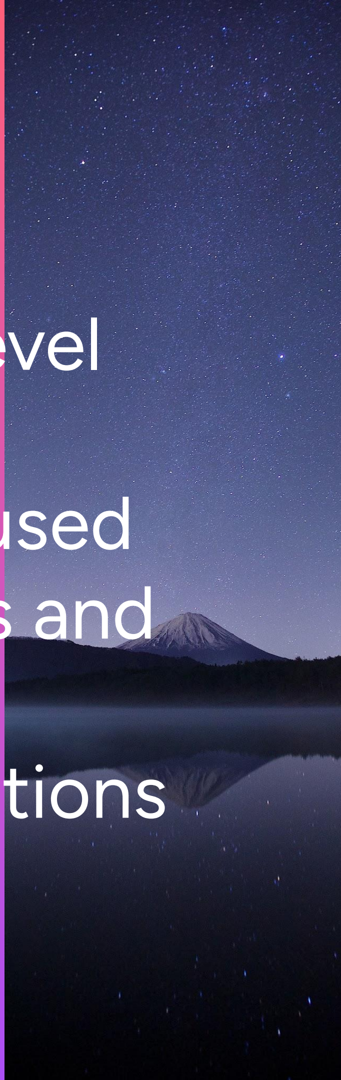


Methodology

- In the third experiment, the predicted word levels were used to estimate the proficiency level of essays. The results from the LLMs were very similar to human evaluations. So when a human thought the essay was advanced the LLMs results also showed the essay to be advanced. It was also found that vocabulary sophistication was the best predictive analytic feature for overall proficiency scores.

Methodology

In the last experiment, they tested the EVPs level consistency. They used two very common multi-meaning words- “work” and “like”. They used the LLM to look at thousands of learner essays and predict what CEFR level each use of the two words were based on context. The LLM predictions confirmed EVP level consistency across proficiency levels.



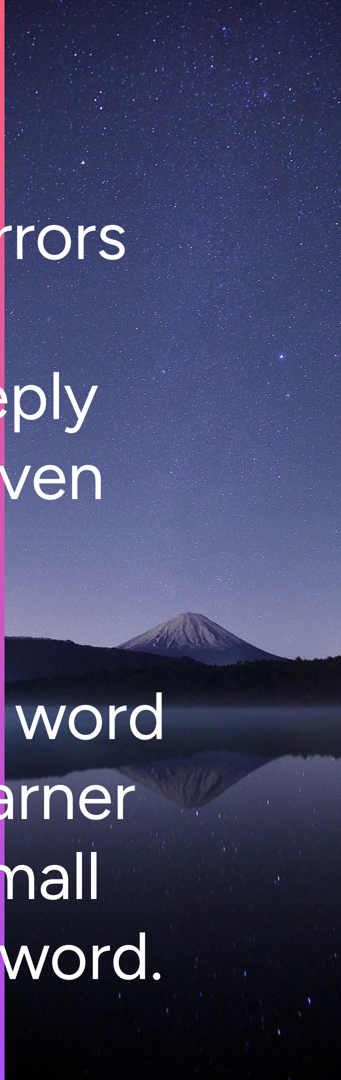
Main Findings

- LLMs effectively handle semantic ambiguity and contextual meaning better than the PoS-based or rule-based models
  Qwen 2.5 achieved the best balance of performance and efficiency.
  Word-level predictions correlate with overall essay proficiency and vocabulary/ phrase scores.
  The most optimal approach might be a hybrid one with LLMs for ambiguous words and PoS for simples ones.

Limitations

- Learner errors like spelling errors and grammar errors could reduce lemmatization accuracy.
- Multi-word expressions and idioms were not deeply analyzed so there is room to make the method even smarter.

  When the researchers labeled each word in a sentence, they only marked what CEFR level the word belonged to but not the specific meaning the learner intended for the word. This sometimes causes small mismatches when guessing the difficulty of the word.



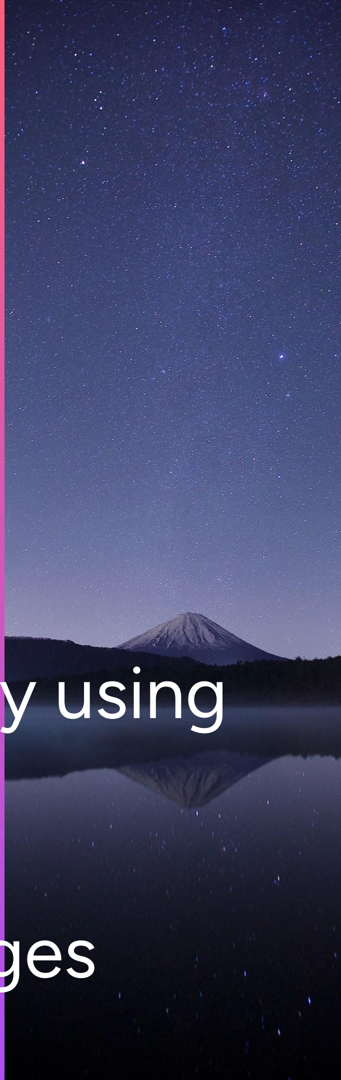
Quiz: Question 1

What is the main goal of the study?
A) To teach English using LLMs

B) To assess grammar errors in learner writing

C) To evaluate word-level vocabulary proficiency using LLMs and the English Vocabulary Profile

D) To translate English essays into other languages



Quiz: Question 2

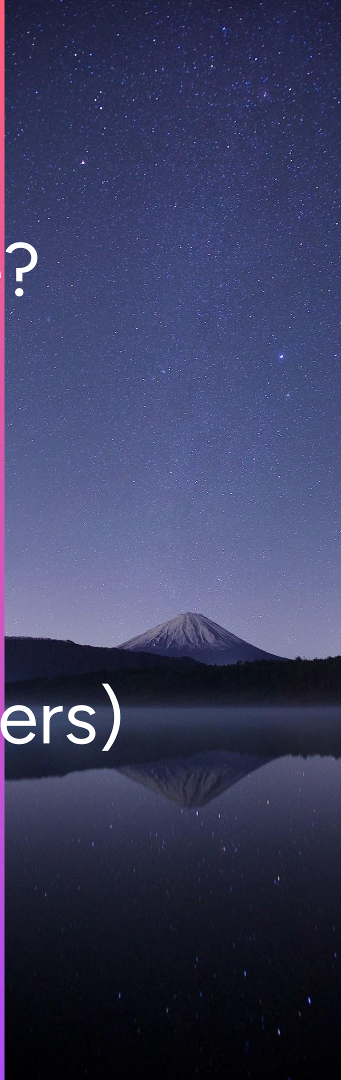
What type of data did the researchers analyze?

A) Spoken interviews

B) Short stories written by native speakers

C) Essays written by English learners (L2 learners)

D) Tweets and social media posts



Quiz: Question 3

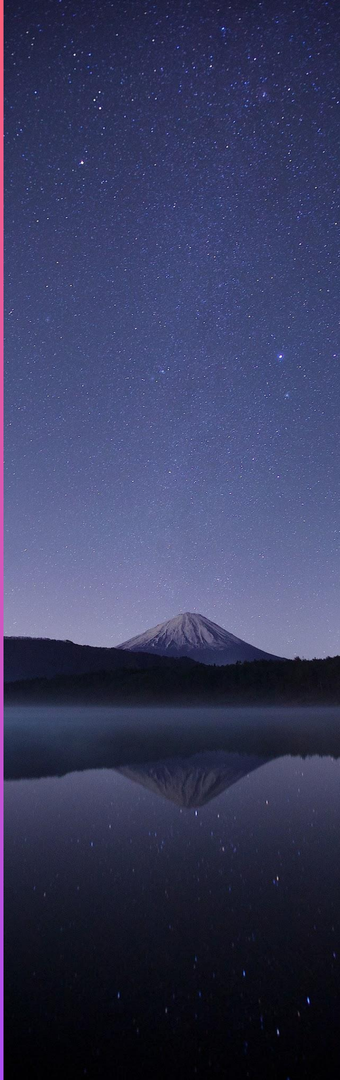
What were the two major dimensions of vocabulary knowledge found in the studies?

A) Lexical Sophistication

B) Lexical Accuracy

C) Lexical Density

D) Lexical Diversity



Quiz: Question 4

In the first experiment, what model had performed the best?

A) 40 mini

B) GPT-4o

C) Llama 3.1

D) Qwen 2.5