PLOS ONE

RESEARCH ARTICLE

Establishing vocabulary tests as a benchmark for evaluating large language models

Gonzalo Martínez[1], Javier Conde[2], Elena Merino-Gómez[3], Beatriz Bermúdez-Margaretto[4], José Alberto Hernández[1], Pedro Reviriego[2]*, Marc Brysbaert[5]

1 Departamento de Ingeniería Telemática, Universidad Carlos III de Madrid, Leganés, Spain, **2** ETSI de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain, **3** Escuela de Ingenierías Industriales, Universidad de Valladolid, Valladolid, Spain, **4** Departamento de Psicología Básica, Psicobiología y Metodología de las CC. del Compto, Universidad de Salamanca, Salamanca, Spain, **5** Department of Experimental Psychology, Ghent University, Ghent, Belgium

* pedro.reviriego@upm.es

Citation: Martínez G, Conde J, Merino-Gómez E, Bermúdez-Margaretto B, Hernández JA, Reviriego P, et al. (2024) Establishing vocabulary tests as a benchmark for evaluating large language models. PLoS ONE 19(12): e0308259. https://doi.org/10.1371/journal.pone.0308259

Editor: Jessie S. Barrot, National University Philippines, PHILIPPINES

Received: June 26, 2024

Accepted: October 4, 2024

Published: December 12, 2024

Data Availability Statement: All data is available on a Github repository: https://github.com/WordsGPT/LLM_Vocabulary_Evaluation.

Funding: This work was partially supported by the project CyberTutor: Asistente educativo personalizado basado en Grandes Modelos de Lenguaje (LLM), funded by "Primeros Proyectos" call from ETSIT, UPM; by the FUN4DATE (PID2022-136684OB-C22) and ENTRUDIT (TED2021-130118B-I00 projects funded by the Spanish Agencia Estatal de Investigación (AEI); by

Abstract

Vocabulary tests, once a cornerstone of language modeling evaluation, have been largely overlooked in the current landscape of Large Language Models (LLMs) like Llama 2, Mistral, and GPT. While most LLM evaluation benchmarks focus on specific tasks or domain-specific knowledge, they often neglect the fundamental linguistic aspects of language understanding. In this paper, we advocate for the revival of vocabulary tests as a valuable tool for assessing LLM performance. We evaluate seven LLMs using two vocabulary test formats across two languages and uncover surprising gaps in their lexical knowledge. These findings shed light on the intricacies of LLM word representations, their learning mechanisms, and performance variations across models and languages. Moreover, the ability to automatically generate and perform vocabulary tests offers new opportunities to expand the approach and provide a more complete picture of LLMs' language skills.

Introduction

In a seminal paper Landauer and Dumais [1] presented latent semantic analysis (LSA) as a new theory of knowledge representation. Meaning was inferred from local co-occurrences of words in representative text. The main idea was that one can learn the meaning of an unfamiliar word "X" from the words that frequently occur with "X." Results with the LSA model suggested that English vocabulary could be acquired in that way at a rate comparable to schoolchildren, without prior linguistic or perceptual knowledge.

Landauer and Dumais not only proposed the theory but also presented a mathematical model that, starting from the co-occurrences of words in the texts, constructed a matrix that was then mapped to a space of reduced dimensions. In this space of a few hundred dimensions, each word was represented by a point and the distance between the points represented the distance in meaning. The points were defined as semantic vectors and had about 300 dimensions. Words with similar meanings had semantic vectors that were close to each other.

Landauer and Dumais evaluated their LSA model with a vocabulary test. They used 80 items from the synonymy section of the Test of English as a Foreign Language (TOEFL). In this task, target words were presented with four response alternatives from which the correct

the Chips Act Joint Undertaking project SMARTY (Grant no. 101140087) and by the OpenAI API Research Access Program. The funders had not played in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

one had to be chosen. Landauer and Dumais found that the semantic vector of the correct answer was closest to that of the target word in 51 items (64% correct). This was comparable to US college candidates from non-English-speaking countries who scored 65% correct on average [1].

Although the scores of the LSA model were not perfect (they were at the same level as speakers of English as a second language), they were much better than what was available at the time. Jarmasz and Szpakowicz reported that other algorithms for word meaning scored barely above the chance level (25%) on the TOEFL test [2]. In subsequent years, authors tried to improve the performance of LSA-type models by optimizing the algorithm and training materials. Performance on the TOEFL test (as it became known) gradually increased until Bullinaria and Levy reported 100% correct performance for a model tweaked to the test [3, 4]. Landauer and Dumais argued that their LSA model could be viewed as a simple three-layer neural network, an argument that was confirmed when real connectionist networks were developed that quickly outperformed the co-occurrence statistics used in LSA [5–7].

The work of Landauer and Dumais showed how vocabulary can be learned from a large corpus of text using only local relationships between words. This breakthrough was a fundamental step in understanding how language can be acquired and how computer systems can be implemented that can learn the meaning of words and text. These concepts and models can be seen as forerunners of today's Large Language Models (LLMs). Since vocabulary tests were used to evaluate the predecessors of LLMs, a research question which we try to answer in this paper naturally arises: are vocabulary tests still relevant to evaluate LLMs?

Large language models design and evaluation

Over the past decade, artificial intelligence has made impressive progress in language modeling (and other areas). The improvements were possible due to the availability of huge text datasets, more powerful computing and storage, and architectures that can implement very large models efficiently (such as transformers). First, the introduction of the transformer [8] and then the development of popular models based on it, such as BERT [9] or T5 [10], paved the way for the development of LLMs with many billions of parameters, such as GPT4 [11]. Transformers are complex neural networks with many interconnected layers and novel mechanisms such as attentional focus that allow them to learn complex relationships, for example, between words in text. By increasing the size of transformers and training datasets, unprecedented performance has been achieved in many language processing tasks, but more importantly, they have been integrated into products such as ChatGPT, Gemini, or Bing, reaching hundreds of millions of users [12]. The main features of these LLMs are the huge size of their training datasets and number of model parameters, their ability to learn different languages (including programming languages) and perform a wide range of tasks such as generating text in genres, translating, answering questions and summarizing [13]. LLMs operate on units called tokens, which in some cases correspond to words but can also be sequences of a few letters from the training dataset. Texts are decomposed into tokens as input, and the model generates output tokens that are assembled into words and sentences. LLMs are models that predict the next token, and tokens are mapped into words. This process can be applied recursively to construct complex sentences or even longer texts or to build more powerful tools based on LLMs. A particularly attractive application of LLMs is the development of intelligent chatbots, such as the well-known ChatGPT used by hundreds of millions of people worldwide. Chatbots are models that use a base pre-trained LLM to specialize in answering questions. To do this, the model has been partially retrained (finetuned) with a dataset of question-answer pairs, allowing the neural network to learn how to process questions and respond to them.

The capabilities of LLM based chatbots make them attractive for use in L2 teaching [14, 15]. For example, they can be used to interact with students by engaging in conversations or generating questions to practice different skills and correct the mistakes of the students. They can also assist students when writing. However, as with any new technology there are also drawbacks and limitations. Chatbots can be used to complete assignments students should do and write essays for them. Chatbots can also instill a false confidence in students, reducing the incentive to learn the language. LLMs may further produce invalid and incorrect information that is hard to detect as it is well written. Regardless of these limitations LLMs are poised to change the way languages are learned in the near future.

As mentioned before, LLMs are trained to predict the next token in a sentence [16]. This is done by taking existing texts, removing tokens from them and using these tokens as criteria for the outcome to be predicted by the model. This approach makes it possible to use huge datasets containing almost every text of interest on the Internet. The reason why models work with tokens is that words in all languages can be generated with just a list of tens or hundreds of thousands of tokens. Otherwise, if they worked with all the words from every language, it would make the models too heavy and slow. In contrast, if models operated with single characters, many predictions would be needed to construct long texts. The tokenization strategy (design of the token dictionary) is very relevant when designing an LLM. The most common practice is to assign tokens to the most frequent words. Likewise, many models have predominantly been trained in English, tokenizers have also been designed to optimize generation in this language. This means that generating the same text in two different languages may take more time, cost, or produce responses with lower quality.

Interestingly, predicting the next token assumes that languages and words can be learned from the words that appear in texts together with the target word. This is exactly the assumption Landauer and Dumais made in their LSA model. Therefore, LLMs can, in a way, be seen as descendants of LSA. The prediction of words based on surrounding words is not the only link between LLMs and LSA. The use of a point in a multidimensional space to represent word meanings, introduced by LSA, is also used by LLMs to map inputs to so-called "embeddings." These embeddings are vectors of values that correspond to a point in the space where the meaning is located, so that, similar to semantic vectors in LSA, points that are close together correspond to similar meanings. In the case of LLMs, embeddings are usually larger than the 300 dimensions used in LSA, with several thousand dimensions.

To evaluate LLM performance, several benchmarks have been proposed [11]. In most cases, the test evaluates how well LLMs answer questions on almost any topic [17, 18], or are able to perform reasoning based on a given text [19]. For example, there are benchmarks with thousands of mathematical problems covering almost every discipline in mathematics [20]. There are also frameworks that can be easily extended so that new tasks or tests can be added at will (see for example https://github.com/openai/evals). Those expanded benchmarks focus on quantifying how well LLMs perform on different knowledge tasks. Fig 1 shows the evolution of models over the years along with their sizes expressed in billions of parameters, since the introduction of the transformers architecture in 2017 [21]. In November 2022, the trend of LLMs emerged, coinciding with OpenAI's launch of ChatGPT.

Large language models produce language, which is likely to create a feedback loop

LLMs do not just answer questions or solve problems. At the same time, they produce language output. They create text when they provide answers or translations. In fact, they are already being used to help write entire novels and textbooks. So, people and future LLMs will

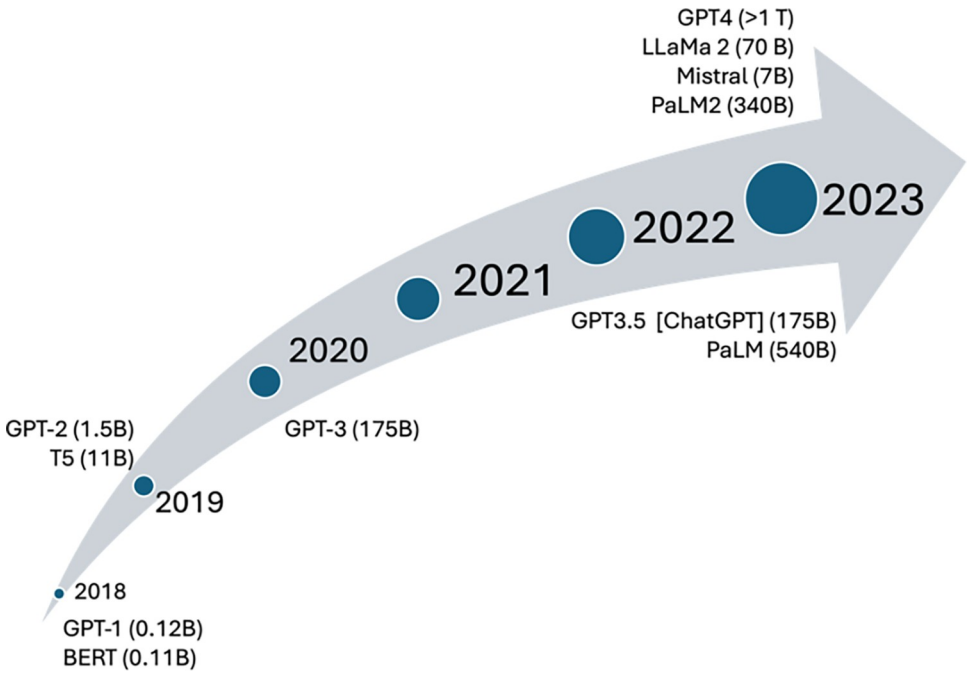

Fig 1. Evolution of LLMs. Released models and model sizes over the years. The sizes of commercial models are not official.

https://doi.org/10.1371/journal.pone.0308259.g001

be increasingly exposed to LLMs' output, creating a feedback loop [22]. As a simple example, take a word that is not produced by LLMs because a simpler synonym exists. This word will appear less and less in the language to which future people and LLMs are exposed until it becomes extinct.

Some research is beginning to appear on these subtler, as well as more fundamental, linguistic aspects of the use of LLMs, but it is still very limited compared to the large number of articles evaluating the performance of LLMs on various knowledge tasks. There are some studies that have focused on the linguistic characteristics of the text generated by LLMs. A comparison of the linguistic characteristics of humans and LLMs is presented in [23], which focuses on the analysis of news generated by an open-source LLM. There are also studies that focus on phonological [24] and lexical [25] aspects of LLMs. Finally, the effect of LLMs in academic writing has also been recently studied showing how some words are becoming more popular, probably due to the use of generative AI tools by the authors when writing papers [26].

In this article, we look at how LLMs perform on vocabulary tests, what this tells us about how LLMs learn language, and try to answer the question of whether vocabulary tests can be a useful addition for evaluating LLMs.

Vocabulary tests and large language models

As discussed at the beginning of this article, the TOEFL test was an important tool for evaluating Landauer and Dumais' LSA model and its later extensions. Vocabulary tests are also widely used to assess language proficiency in humans [27]. Surprisingly, current LLMs are no longer tested on vocabulary tests, probably because everyone assumes they will be error-free. Given that LSA-type models were already achieving flawless performance in early 2010 and that current LLMs are vastly superior in design and the amount of training material, it seems a waste of time to test them on something as simple as the TOEFL test, which only taps into knowledge

known to undergraduates with English as a second language. Indeed, this is what the present authors expected when they used the TOEFL test as the starting point of a study that would use more taxing vocabulary tests.

Vocabulary tests traditionally consist of multiple-choice questions that require participants to choose the correct answer from a number of alternatives [28]. Indeed, this is the format of the TOEFL test. The difficulty of the test then depends on the difficulty of the target words and the number and difficulty of the answer alternatives. By varying these, it is possible to take more demanding vocabulary tests than the TOEFL to see what level various LLMs achieve.

Lemhöfer and Broersma introduced another format for vocabulary tests that may be of particular interest for LLM testing [29]. They presented participants with a list of letter strings and asked them to indicate which words they knew. To prevent participants from selecting all items without knowing them, legal letter strings were added that do not exist as words in English (such as "plound" or "ternace"; these are so-called non-words or pseudowords). Performance was estimated based on both word and non-word performance, so that a participant who selected all items would receive a score of zero. The Yes/No format gained momentum when Lemhöfer and Broersma published an English language proficiency test for use in psycholinguistic research, which they called LexTALE [30]. The Yes/No format is interesting because the test taker must refrain from choosing the non-words. LLMs are known to tend to present nonexistent information based on word co-occurrences (so-called hallucinations), as discussed in [16]. So, we thought it would be interesting to see how LLMs would perform on tests with non-words.

Considering the evaluation of LLMs, vocabulary tests have a number of features that are of interest:

- First, vocabulary tests are pure language tests with stimuli that are not embedded in an informative context. Thus, vocabulary tests could potentially be used to evaluate LLMs' knowledge of languages. This is an important point since LLMs should perform equally well for all the languages they claim to support (see [31], for evidence that this unfortunately is often not the case yet).
- Second, if a model makes errors, these errors can be analyzed to study how LLMs learn a language and to evaluate whether cognitive theories of language acquisition apply to LLMs. For example, early acquired words are often well remembered by humans even though they almost never occur in everyday language use [32]. Vocabulary tests allow fine-grained evaluation of LLM, which facilitates analysis and understanding of the mechanisms and algorithms underlying the models.
- Third, how do LLMs interpret non-words and are they able to distinguish them from valid words? Do they follow the same mechanisms as humans (discussed in [33])? How does the tokenization that LLMs use affect word learning? To what extent does performance depend on the specific question posed to the LLM?
- A fourth important feature of vocabulary tests is that they can be automated, both for test generation and for test execution. This is especially true for the Yes/No format. There are software tools that can be used to generate non-words in different languages [34] and even workflows to generate entire test suites [35]. This makes it possible to generate large-scale vocabulary tests. Similarly, the execution of those tests in different LLMs can be automated, allowing the evaluation of multiple LLMs in multiple languages, without practical limitations on the number of words tested [36].

These characteristics make vocabulary tests potentially interesting for evaluating LLMs. The main question is: are vocabulary tests still relevant to evaluate LLMs? To try to answer it

we pose ourselves three more specific research questions to get a better understanding of the potential of vocabulary tests for LLM evaluation:

- Research Question 1 (RQ1): What is the performance of current LLMs on existing vocabulary tests?
- Research Question 2 (RQ2): Are vocabulary tests capable of discriminating LLM performance?
- Research Question 3 (RQ3): Which type of vocabulary tests are more relevant for LLM evaluation?

In the following we present an initial assessment of the use of existing vocabulary tests for evaluating LLMs so that we can answer those questions and decide whether or not it is a fruitful approach.

Materials and methods

To assess the usefulness of vocabulary tests in LLM evaluation, we conducted several vocabulary tests on different LLMs. To have a representative sample of current LLMs, we selected two company-owned, commercial LLM tools: ChatGPT (based on GPT3.5 and GPT4 see https://openai.com/blog/chatgpt) and Bard (based on PaLM 2, see https://ai.google/static/documents/google-about-bard.pdf, now replaced by Gemini), together with two open source LLMs: Llama 2 [37] and Mistral [38].

ChatGPT, developed by OpenAI, is the most popular LLM-based chatbot today and probably the one that has shown the best performance across a variety of tasks. Two versions of ChatGPT (with different numbers of parameters) were tested. Bard was developed by Google and is intended to compete with ChatGPT, so both are good examples of commercial chatbots. Parameters and source code for these LLMs are not available, which makes them less interesting for research purposes because they can be modified overnight without researchers being able to verify what was done. Still, because of their massive use by the public, it is worthwhile to determine their performance at some point. Llama 2, developed by Meta, is probably the best known open-source LLM right now. Another open-source LLM with good performance is Mistral, developed by a startup of the same name. We tested three versions of Llama 2, of different sizes, to see to what extent performance improves as network complexity increases. The seven models considered in our evaluation are summarized in Table 1. They range from relatively small models to the largest models that were publicly accessible at the time of conducting the experiments.

- Mistral-7B. LLM developed by Mistral AI and released in September 2023. It has 7.3 billion parameters and is based on a transformer architecture. It stands out for its efficiency,

Table 1. LLMs considered in the experiments.

LLM	Company	Type	Parameters
Mistral-7B	Mistral	Opensource	7.3 billion
Llama 2-7B	Meta	Opensource	7 billion
Llama 2-13B	Meta	Opensource	13 billion
Llama 2-70B	Meta	Opensource	70 billion
PaLM 2 (Bard)	Google	Commercial	> 340 billion (non-official)
GPT-3.5-turbo (ChatGPT)	OpenAI	Commercial	175 billion
GPT-4 (ChatGPT)	OpenAI	Commercial	> 1 trillion (non-official)

https://doi.org/10.1371/journal.pone.0308259.t001

outperforming larger models like Llama2 at the time of its release. It is an open-weight model, but the dataset used for its training is not public [38].

- Llama 2 (7/13/70B). Developed by Meta AI and released in July 2023. It is available in versions of 7B, 13B, and 70B parameters, utilizing a transformer architecture with techniques such as RMSNorm pre-normalization, SwiGLU activation function, and rotary positional embeddings [37]. Unlike other big tech companies, Meta has chosen to offer Llama 2 as open source. It has been trained in over 20 languages, although most of the data is in English (89.7%), while other languages, such as Spanish, represent only 0.13% of the training data.
- PaLM 2. Developed by Google AI and presented in May 2023, it is a lighter version compared to its predecessor (PaLM) while offering better capabilities. The training dataset includes more data in other languages to enhance multilingual capabilities compared with PaLM. Neither the training dataset nor the model weights have been released [39].
- GPT3.5. Commercial model developed by OpenAI and a version of GPT-3. It is estimated to have a size equal to GPT-3 (175 billion parameters), which was primarily trained in English (93%). GPT-3.5 is the model on which the first version of ChatGPT was based (November 2022) [40].
- GPT4: Latest version of the GPT family released in March 2023 [41]. The model adds image processing capabilities and improves its performance compared to its predecessor. Various sources indicate that it contains more than 1 trillion of parameters within a Mixture of Experts (MoE) architecture [42], which allows it to dynamically activate different subsets of its neural network depending on the input.

The tests were run automatically using the Application Programming Interfaces (APIs) of the LLM-based chatbots to create the questions in each test and then produce an excel file with all the answers, as described in [36]. The only exception was Bard (now Gemini), whose API was not accessible in Spain at the time of the evaluation, for Bard the user interface was used for testing. Automation is interesting for running tests at scale, since we evaluated seven LLMs on tests with dozens of questions each. In addition, the use of the API allowed control over LLM parameters, such as temperature, which adjust the variability of answers. For certain models, such as Llama 2, the responses included extra text along with the selected answer (A/B/C/D). When feasible, this text was automatically processed to isolate the answer and compare it with the correct one to generate the evaluation metrics. In cases where this was not possible, a manual analysis was performed to classify the response as correct or incorrect. During evaluation, LLMs were not given context information and default parameters were used except temperature, which was set to zero if it was controllable to produce deterministic responses. The prompts used to interrogate the chatbots were simple and similar to those used in the human tests (see below). The performance of LLMs can be improved by providing context or more sophisticated prompts that force the LLMs to solve the questions step by step using a chain of thought [16]. However, our goal was to understand how LLMs perform in vocabulary tests when presented with the same questions as humans and not to modify the questions or provide additional information to improve the LLMs' answers.

We use several representative vocabulary tests, both with multiple choice and yes/no questions, in our evaluation. In every case, our use of these tests adheres to the terms and conditions set forth by the sources. For example, for some of the tests, the questions cannot be made publicly available and thus are not included in the dataset that contains the results of the different LLMs for the different tests.

Table 2. Vocabulary tests considered in the experiment.

Test	Type	Language	Number of items
TOEFL	Multiple choice	English	80
StuVoc-EN	Multiple choice	English	150
StuVoc-SP	Multiple choice	Spanish	80
LexTALE-EN	Yes/No	English	60
LexTALE-SP	Yes/No	Spanish	90

https://doi.org/10.1371/journal.pone.0308259.t002

The details of the multiple-choice and yes/no questions tests are summarized in Table 2. The first test was the TOEFL introduced by Landauer and Dumais [1]. It contains 80 target words with four alternatives to choose from (these are also single words). The difficulty of the items varies, but the level is adapted to non-English-speaking students who want to study at English-speaking universities. The test cannot be freely shared due to copyright restrictions, but researchers can request access to the stimulus material, if use is strictly limited to research purposes. We were kindly granted access to the stimuli by the LSA research group at the University of Colorado.

The second vocabulary test was the StuVoc test, published by Vermeiren et al. [43]. This test contains three subtests with 50 validated English items each (thus 150 items in total). Items consist of target words in short neutral sentences along with four response alternatives. Unlike the TOEFL, the alternatives can include short descriptions of words. The first two subtests are difficult enough for English-speaking university students. The third subtest is easier and better suited for second-language speakers with high proficiency [27]. The level of the last subtest is similar to the TOEFL, while the first two tests are more demanding.

The third vocabulary test was the Spanish adaptation of StuVoc. Bermúdez-Margaretto & Brysbaert translated 146 of the English StuVoc items into Spanish and validated them on a group of adult native Spanish speakers [39]. They selected the 80 best items (good distribution of difficulty levels, good correlation between item performance and overall test performance, and a clear transition from unknown to known based on item response theory analysis). The remaining 66 items were considered less interesting for various reasons.

The fourth and fifth tests were yes/no tests. For English, we used LexTALE, proposed by Lemhöfer and Broersma [29]. The test contains 40 English words and 20 non-words. Since the test is aimed at advanced second language speakers, the level is comparable to the TOEFL. Native speakers typically score more than 90% correct on the test (scores obtained after subtracting the % yes answers to non-words).

Finally, we tested the models on a Spanish Yes/No test published by [40]. This test is more comprehensive and more difficult than the English LexTALE because the authors wanted the test to be usable by both native Spanish speakers and second-language speakers [41]. The test contains 60 Spanish words and 30 non-words. We presented all vocabulary tests to all LLMs. For the multiple-choice tests, we asked: "Answer the option which is the meaning of the following word "squelch": a. suppress b. revive c. acquire d. dispute. Please, first just answer the letter of the option and below your explanation.". For the Yes/No tests, we asked: "Please answer "Yes" or "No" to the following question: Is X an existing word in English (Spanish)?". The prompts were the same, in English for all tests to avoid prompts from causing differences in the results We have selected these five tests, which include two languages and two types of evaluations (multiple-choice and yes/no for words/pseudowords), because they allow to assess different aspects of LLMs, such as their performance degradation when using a language other than English and their ability to recognize pseudowords as invalid (which is a task they struggle with). Table 3 contains samples of all the test except for TOEFL as it is not public.

Table 3. Samples of vocabulary tests considered in the experiments.

Test	Example of question	Correct answer
StuVoc-EN	Answer the option which is the meaning of the following word "ablution": a. did all her duties as a minister b. washed herself to get ready c. played her set piece of music d. did her exercises to stay healthy This is an example sentence: She performed her ablutions. Please, first just answer the letter of the option and below your explanation.	b
StuVoc-SP	Answer the option which is the meaning of the following word "ablución": a. tocar su pieza musical b. hizo sus ejercicios para mantenerse saludable c. cumplió con todos sus deberes como ministra d. se lavó para prepararse This is an example sentence: Ella realizó sus abluciones. Please, first just answer the letter of the option and below your explanation.	d
LexTALE-EN	Please respond with "Yes" or "No" to the following question: Is "mensible" an existing word in English?	no
LexTALE-SP	Please respond with "Yes" or "No" to the following question: Is "terzo" an existing word in Spanish?	no

https://doi.org/10.1371/journal.pone.0308259.t003

Results

The aim of our study is to determine whether vocabulary tests are still relevant for evaluating LLMs. First, we will analyze the performance of models using different types of vocabulary tests and in different languages (RQ1). Next, we will examine whether these tests can differentiate the quality of the models (RQ2). Finally, we will investigate which vocabulary tests are most suitable for LLM evaluation (RQ3). The results obtained from the multiple-choice tests are summarized in Table 4.

As expected, most models performed well on the TOEFL test, and the larger Llama 2 models outperformed the basic 7B version. Still, the performance was not flawless. Moreover, the models differed in the items they got wrong, suggesting that suboptimal performance was not due to one or two weak items. The three commercial models failed on the item "fashion," where they chose the option "rage" instead of "manner" (which Llama 2-7B and Mistral-7B did get right). For the item "figure," GPT3.5, Bard and Mistral-7B chose the option "express" instead of "solve." GPT4 and Mistral-7B obtained the best scores. However, the differences with PaLM 2 (Bard), Llama 2-70B and GPT3.5 were small. Unfortunately, no further information can be given for the TOEFL test as the items are copyrighted. For the other tests, the results for each model and item are available in a public GitHub repository (available at https://github.com/WordsGPT/LLM_Vocabulary_Evaluation). This dataset contains for each item on each test the response given by each of the LLMs in machine readable files. The links to the LLM models and versions used are also provided.

Despite the fact that the StuVoc-EN is more demanding for human speakers than the TOEFL, LLMs' performance on this test was generally higher than on the TOEFL. One reason

Table 4. Performance of LLMs on the multiple-choice tests (percentage of correct answers).

Test	Llama 2-7B	Llama 2-13B	Llama 2-70B	Mistral-7B	GPT3.5	GPT4	PaLM 2
TOEFL	76.3%	93.9%	96.3%	98.8%	96.3%	98.8%	97.5%
StuVoc-EN	91.3%	96.0%	94.7%	99.3%	97.3%	100%	98.7%
StuVoc-SP	73.8%	76.4%	61.3%	68.8%	93.8%	95.0%	96.3%

https://doi.org/10.1371/journal.pone.0308259.t004

could be the availability of a short, neutral context sentence (given information about the part-of-speech). Another reason could be that multi-word descriptions describe the meaning of the target words better than one-word synonyms. Again, the best scores were obtained by GPT4 and Mistral-7B, closely followed by PaLM 2 and GPT3.5. The Llama 2 models had lower scores, and Llama 2-13B performed better than Llama 2-70B. The item "Let's not pussyfoot around" was censored by PaLM 2 and Llama 2-7B. What is further striking is that no item was censored by all LLMs. Every item was scored correctly by at least 4 of the 7 LLMs. Three LLMs had only two items wrong: "The boy shuddered." where they selected "almost fell" instead of "shook" and for "She parried the comments" where the three failing models selected different answers.

There was a significant performance drop for StuVoc-SP compared to StuVoc-EN, even though the items were direct translations. The drop was especially large for the open-source models (Llama 2 and Mistral), which fell below 80%. For the commercial models, the decline was smaller, but still noticeable. Bard (PaLM 2) achieved the highest score in this test.

The distribution of the failures among the words is shown in Fig 2. It can be observed that for TOEFL there are words that are failed by five out of seven models and for StuVoc-SP even for six out of seven models. Instead for StuVoc-EN the worst case are three failures for a word. Given that TOEFL and StuVoc-EN have similar failing percentages, and both are low, the results suggest that TOEFL has some specific questions that are harder for LLMs while that is not the case for StuVoc-EN. For StuVoc-SP, the failure rates are much higher, and therefore, a few questions failing on five or six models can be due to the large number of failures. In the case of Llama2, the best model for StuVoc is the intermediate sized one (13B), not the largest as one might expect. The StuVoc-SP translation stands out, where 70B gets 61.3% of the questions correct, while 13B gets 76.4% and 7B gets 73.8%. Comparing the Llama2 and StuVoc-SP versions, the 70B model makes 38.7% of its errors on its own (12 questions), while in 13B this percentage is much lower (5.3%, 1 question). For example, the word ablución is correctly identified by 13B and 7B, but incorrectly by 70B which identifies it as exercise. These results are aligned with the poorer performance of the models in languages other than English, where more powerful versions in English do not imply they are better in Spanish.

After the analysis of the results of the multiple-choice tests we can go back to our research questions:

- RQ1 (What is the performance of current LLMs on existing vocabulary tests?): LLMs achieve good performance on the vocabulary tests but no model got all the answers correctly except for GPT4 on StuVoc-EN. The performance is significantly worse in Spanish than in English.

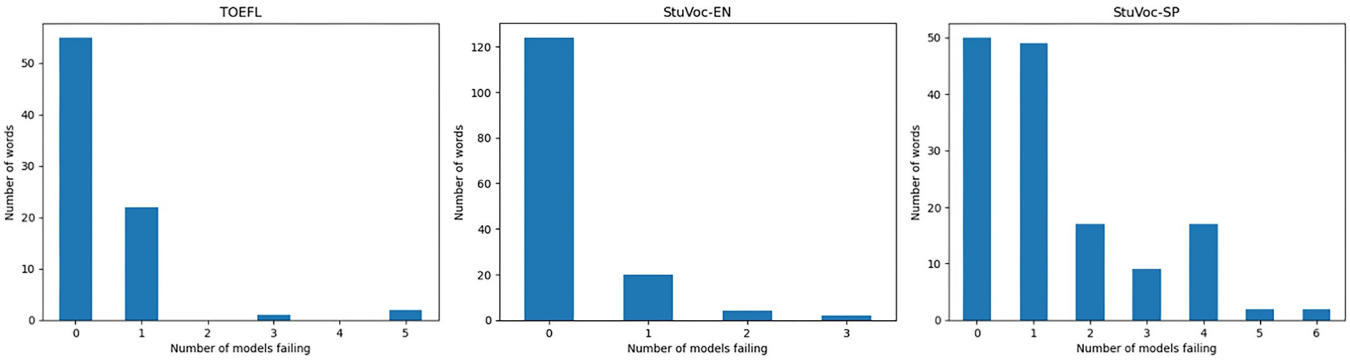

Fig 2. Distribution of the failures per word in TOEFL (left), StuVoc-EN (center) and StuVoc-SP (right).

https://doi.org/10.1371/journal.pone.0308259.g002

Table 5. Performance of October 2023 LLMs on the Yes/No tests (percentage of correct answers).

Test	Type	Llama 2-7B	Llama 2-13B	Llama 2-70B	Mistral-7B	GPT3.5	GPT4	PaLM 2
LexTALE-EN	Words	92.5%	87.5%	100%	100%	100%	100%	100%
LexTALE-EN	Non-words	100%	95%	85%	95%	95%	100%	85%
LexTALE-SP	Words	96.7%	96.7%	100%	100%	96.7%	100%	100%
LexTALE-SP	Non-words	7.1%	60.7%	14.3%	0%	67.9%	82.1%	46.4%

https://doi.org/10.1371/journal.pone.0308259.t005

- RQ2 (Are vocabulary tests capable of discriminating LLM performance?): the tests are able to discriminate LLM performance and can be used to compare LLMs. For example, they show the limitations of open models in Spanish.

The results for the Yes/No tests are presented in Table 5 and are reported independently for words and non-words to better understand the results. Performance on LexTALE-EN was quite good and tended to be better for words than for non-words. Remember that this is a fairly easy test designed for second language speakers. Interestingly, for Llama 2 we see better performance on words as the model gets larger and at the same time worse performance on non-words. ChatGPT4 (GPT4) was flawless on all items, but Google Bard (PaLM 2) in its current version was poor (with Llama 2-70B) in hallucinating the existence of non-words.

Analyzing the data from LexTALE-SP, we saw that two of the non-words (vegada, capillo) are existing words in Spanish because they appear in some dictionaries. Therefore, we excluded these two items from the analyses. Performance was good for the words (especially considering that some were more difficult than the LexTALE-EN words). When further asked about the meaning of the words, most models gave the English translation. However, performance was very poor for the Spanish non-words, where the models not only gave "yes" answers, but when asked, readily gave meanings and translations for letter combinations that do not exist in Spanish. This was especially true for Llama 2 and Mistral. Mistral was unable to identify even one non-word and performed like a boastful test taker, claiming to know all the "words" but in reality scoring zero points. Performance was best for GPT4, but even this model presented interpretations for 18% of the non-words. Performance was also poor for Bard, where interpretations were given for more than half of the non-words.

The distribution of the failures among the words, non-words and all test items is shown in Fig 3. It can be observed that failures are concentrated on non-words. As for the multiple-choice tests, the English version does not show strong correlation among model failures. For the Spanish test, the failure rates for non-words are so high that again having a few questions failing on most or all models is expected even with no correlation among failures. In this case, Mistral7B is not capable of detecting non-words in LexTALE-SP. However, other models like GPT-3.5, which contains only 0.8% Spanish in its training dataset, can. Mistral was possibly trained with a similar percentage of Spanish (the information is not public). Nevertheless, the size of the training dataset is not the only factor that determines multilingual capabilities. Other aspects, such as the model's architecture or the tokenization strategy, can also affect the model's performance.

As with the multiple-choice tests, after the analysis of the results of the yes/no tests we can go back to our research questions:

- RQ1: LLMs also achieve good performance on the yes/no tests but again except for GPT4 on Lex-TALE-EN scores are also below 100%. Results are worse for nonwords and for Spanish.
- RQ2: again, the tests are able to discriminate LLM performance and can be used to compare LLMs. For example, they show the inability of Mistral-7B in recognizing nonwords in Spanish.

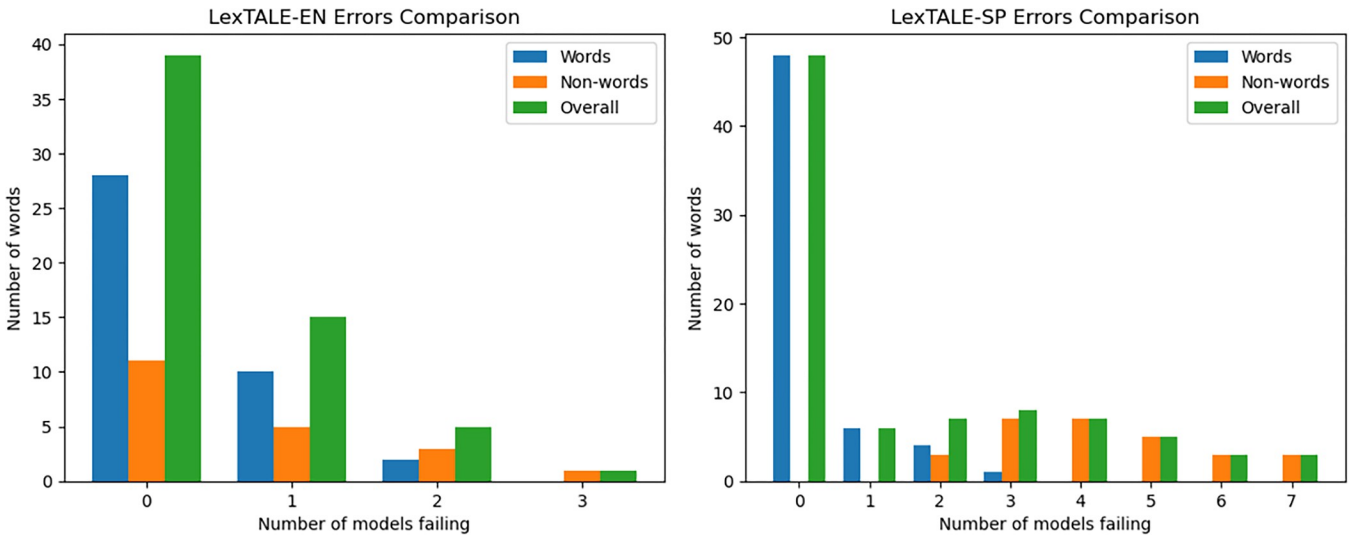

Fig 3. Distribution of the failures per word in LexTale-EN (left) and LextTale-SP (right).

https://doi.org/10.1371/journal.pone.0308259.g003

Finally, having all the results we can also go back to our last research question:

- RQ3 (Which type of vocabulary tests are more relevant for LLM evaluation?): both tests, multiple-choice and yes/no can be used to evaluate and compare LLMs. However, yes/no tests are more interesting as LLMs have a tendency to identify nonwords as words.

The Spanish results suggest that current LLMs are likely to perform less well in languages other than English, for which the training materials were most extensive. To better understand this effect, the information on the training datasets for GPT3 (see https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv), Palm [44] and Llama 2 [37] is summarized in Table 6. For the other models, the information on the training dataset is not publicly available as it is considered a key element of the design, and it is kept confidential for commercial reasons (see https://huggingface.co/mistralai/Mistral-7B-v0.1/discussions/8).

What is first noticeable is the small percentage of training materials in Spanish, even though Spanish is one of the most widely spoken languages in the world (and present in the USA). To some extent, it is surprising that performance on StuVoc-SP and for some LLMs on LexTALE-SP was so good, given the limited amount of Spanish training material these models received. At the same time, the results clearly show that the use of LLMs for languages other than English is oversold.

Discussion

In the early days of machine language models, developers tested the quality of their models with a vocabulary test (specifically, the TOEFL test introduced by Landauer & Dumais [1]. At present this is no longer done, possibly because developers assume that current models are error-free given the improvements in design and amount of training over the past decade. In this paper we have analyzed the performance of LLMs on several vocabulary tests to understand if they are still relevant as benchmarks for LLMs.

When we tested the performance of available LLMs, however, we saw that most models did not achieve 100% scores on the various vocabulary tests (RQ1) (the exception was GPT4 on StuVoc-EN and LexTALE-EN). Performance was especially poor in Spanish and in Yes/No tests. Therefore, the tests can still be used to discriminate LLM performance (RQ2).

Table 6. Percentage of the training dataset in English and Spanish for several LLMs.

Model	English	Spanish
GPT3	92.64%	0.77%
PaLM 2	77.98%	2.11%
Llama 2	89.70%	0.13%

https://doi.org/10.1371/journal.pone.0308259.t006

The origin of the poor performance in Spanish is not difficult to find. Given the small percentage of training material the models received in Spanish (Table 6), it is to some extent surprising that the models still performed so well (RQ1). At the same time, our findings suggest that errors in Spanish are more likely than in English and that the quality of Spanish answers will be linguistically worse than the quality of English answers (RQ2). Given that Spanish is one of the most widely spoken languages in the world (and is present in the U.S.), it is to be expected that performance will decline even further for languages with less training material. Our findings indicate that training datasets should be more balanced to avoid bias against languages other than English, and they demonstrate the usefulness of vocabulary tests to perform comparisons of language proficiency in LLMs.

The high number of non-words accepted as meaningful by LLMs (again, particularly in Spanish) is also important and worthy of further investigation. Needless to say, language models that hallucinate meaning when there is none are poor assistants. One possible cause could be that some models do not work with words as input and output units, but with tokens of a certain length, regardless of word boundaries. In such models, words have no special status and non-words similar to existing words (as good pseudowords should be) can activate lexical output. Even in humans, there is evidence that some non-words activate word-related meanings [32]. Further factors contributing to hallucinations of non-words are likely to be cross-language contamination (a non-word in one language may be a word in another language) and spelling errors in training materials. Therefore, again vocabulary tests are relevant to understand how LLMs interpret words.

Another question is whether the performance of the best English models is already good enough. Much here depends on what one wants (or claims) to achieve. If average human performance is the goal, then the current best models may already be good enough. However, if models are intended to improve human performance, then further progress still seems possible, even in English.

Perhaps even more important than the results with the specific tests we used is the usefulness of vocabulary tests in general to examine the performance of LLMs. Many hypotheses in cognitive science linked to existing theories of human cognition can be related to the generation of specific words (and non-words) that can be tested in LLMs, to see if the predicted accuracy differences are obtained. This will likely lead to new theoretical developments that may be applicable to both LLMs and humans. By focusing on words rather than knowledge areas, vocabulary tests provide a fine-grained mechanism for investigating the fundamental cognitive mechanisms of language processing in LLMs (and humans). The vocabulary tests of Table 2 included words that were good for testing language proficiency in general, but words (and non-words) can also be selected to answer specific theoretical questions.

Another aspect that can be varied is the format of the vocabulary test. In this article, we have discussed the multiple-choice format and the yes/no format (RQ3). Other formats are those in which a correct definition or word must be generated (production rather than recognition). Such tests are more difficult for humans, but they can be easier for LLMs. Even within a format, a few changes can make a difference [43]. In the results section, we wondered to what extent the availability of neutral carry phrase improves performance. This can be easily

tested in LLMs and, if deemed interesting, later in humans. Similarly, we can examine the extent to which LLMs' performance on Yes/No tests depends on the questions asked and what this says about the underlying mechanisms.

Another finding from our study is that commercial models (PaLM2, GPT-3.5, and GPT-4) have obtained the best results. This aligns with benchmarks where commercial models outperform open models in most evaluations. To the best of our knowledge, vocabulary tests have not been evaluated on them, but it is no surprise that they perform better.

Finally, it is possible to automate the production of large-scale vocabulary tests (e.g., [34]). This makes it possible to generate large numbers of stimuli based on different criteria, such as frequency in texts, length, or even how letter sequences are tokenized by LLMs. In this way, vocabulary tests can be developed to evaluate LLMs at scale. This is particularly interesting because, unlike human participants, LLMs are not limited in the number of items they can process. Thousands of items can be easily tested in LLMs, allowing comprehensive evaluation of the entire vocabulary. This also solves the problem of experimenter bias in the selection of stimuli presented [45–48]. The evaluation of the results can also be automated as part of the pipeline.

In summary, the results obtained and the answers to all the RQs allow us to confirm that vocabulary tests are relevant to evaluate the performance of LLMs and complement existing benchmarks by providing a more fine-grained analysis of how LLMs interpret words.

Conclusion

The development of vocabulary tests for the evaluation of LLMs is an interesting area of research at the intersection of cognitive science, psycholinguistics, and artificial intelligence, which can provide valuable insights into both the operation of LLMs and theories of human cognition. Therefore, there is a strong case for designing such vocabulary tests to complement existing benchmarks for evaluating LLMs.

At the same time, we would like to point out that the performance of LLMs is also useful for creators of vocabulary tests for human participants. The models may not replace item selection based on psychometric analysis, but they can point to ambiguities in item construction. To avoid being overly influenced by one model, our data suggest that it is beneficial to test the performance of multiple models. For the good items we tested, most models gave the expected result, at least as long as the items were existing words. LLM tests can thus be used to avoid problems in new tests for humans. Also note that LLM testing alerted us to the fact that two of the non-words in the Spanish Yes/No test were present in some Spanish dictionaries.

This paper is just an initial step on the study of the use of vocabulary tests to evaluate LLMs and has several limitations. The first one is the number of LLMs and languages evaluated. Evaluating a larger number of LLMs and languages is needed to confirm the results and conclusions obtained in this work. A second limitation is the size of the tests that have only tens of items as they have been designed for humans. The creation of larger tests with thousands of items by relying on automated processes would enable a more comprehensive and focused evaluation and is also an interesting topic for future work. Finally, studying the links between the results of LLMs in vocabulary tests and cognitive science theories [49] is also an interesting idea to explore in future works.

Acknowledgments

We thank the LSA research group at Colorado University and in particular Peter Foltz for kindly providing the items for the TOEFL test.

Author Contributions

Conceptualization: Elena Merino-Gómez, José Alberto Hernández, Pedro Reviriego, Marc Brysbaert.

Data curation: Gonzalo Martínez, Javier Conde, Beatriz Bermúdez-Margaretto.

Investigation: Pedro Reviriego.

Methodology: Elena Merino-Gómez, Beatriz Bermúdez-Margaretto, Pedro Reviriego, Marc Brysbaert.

Resources: Javier Conde.

Software: Gonzalo Martínez, Javier Conde.

Supervision: Elena Merino-Gómez, Pedro Reviriego.

Validation: Gonzalo Martínez, Javier Conde.

Visualization: Gonzalo Martínez, Javier Conde.

Writing – original draft: Gonzalo Martínez, Javier Conde, Elena Merino-Gómez, José Alberto Hernández, Pedro Reviriego, Marc Brysbaert.

Writing – review & editing: Gonzalo Martínez, Javier Conde, Elena Merino-Gómez, Beatriz Bermúdez-Margaretto, José Alberto Hernández, Pedro Reviriego, Marc Brysbaert.

References

1. Landauer T. K., & Dumais S. T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104 (2), 211–240. (1997).
2. Jarmasz, M., and Szpakowicz, S. Roget's thesaurus and semantic similarity, In Proceedings of the *International Conference on Recent Advances in Natural Language Processing*, pp. 212–219 (2003).
3. Bullinaria J.A., and Levy J.P. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior Research Methods*, 44 (3), 890–907. (2012) https://doi.org/10.3758/s13428-011-0183-8 PMID: 22258891
4. Wiki ACL (2019, September 15). TOEFL Synonym Questions (State of the art). https://aclweb.org/aclwiki/TOEFL_Synonym_Questions_(State_of_the_art) (2019)
5. Mikolov T., Karafiát M., Burget L., Cernocký J., & Khudanpur S. Recurrent neural network based language model. *In Proceeding of Interspeech* 2, pp. 1045–1048. (2010).
6. Mikolov T., Chen K., Corrado G., & Dean J. Efficient estimation of word representations in vector space. Preprint at https://arxiv.org/abs/1301.3781 (2013).
7. Mandera P., Keuleers E., & Brysbaert M. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92, 57–78. (2017).
8. Vaswani A. et al. Attention is All you Need. *In Proceedings of the Neural Information Processing Systems* (2017).
9. Devlin J., Chang M., Lee K., and Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the North American Chapter of the Association for Computational Linguistics*. (2019)
10. Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21 (140):1–67. (2020).
11. Chang Y., Wang X., Wang J., Wu Y., Zhu K., Chen H., et al. A survey on evaluation of large language models. Preprint at https://arxiv.org/abs/2307.03109 (2023).
12. Ray P. P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154. (2023).

13. Sallam M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, 11, 887. (2023). https://doi.org/10.3390/healthcare11060887 PMID: 36981544

14. Barrot J. S. Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, 57, 100745. (2023).

15. Barrot J. S. Google Bard as an automated written corrective feedback tool: possibilities and feedback. *TESOL Journal*, e805. (2024).

16. Zhao W. X., et al. A survey of large language models. Preprint at https://arxiv.org/abs/2303.18223 (2023).

17. Hendrycks D., Burns C., Basart S., Zou A., Mazeika M., Song D., et al. Measuring massive multitask language understanding. Preprint at https://arxiv.org/abs/2009.03300 (2020).

18. Srivastava A., Rastogi A., Rao A., Shoeb A. A. M., Abid A., Fisch A., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Preprint at https://arxiv.org/abs/2206.04615 (2022).

19. Zellers R., Holtzman A., Bisk Y., Farhadi A., & Choi Y. Hellaswag: Can a machine really finish your sentence? Preprint at https://arxiv.org/abs/1905.07830 (2019).

20. Hendrycks D., Burns C., Kadavath S., Arora A., Basart S., Tang E., et al. Measuring mathematical problem solving with the math dataset. Preprint at https://arxiv.org/abs/2103.03874 (2021).

21. Vaswani A., et al "Attention is all you need." Advances in Neural Information Processing Systems (2017).

22. Martínez G., Watson L., Reviriego P., Hernández J. A., Juarez M., & Sarkar R. Towards Understanding the Interplay of Generative Artificial Intelligence and the Internet. I*n International Workshop on Epistemic Uncertainty in Artificial Intelligence* (pp. 59–73). Cham: Springer Nature Switzerland, 2023.

23. Muñoz-Ortiz A., Gómez-Rodríguez C., & Vilares D. Contrasting linguistic patterns in human and LLM-generated text. Preprint at https://arxiv.org/abs/2308.09067 (2023).

24. Toro J. M. Emergence of a phonological bias in ChatGPT. Preprint at https://arxiv.org/abs/2305.15929 (2023).

25. Reviriego P., Conde J., Merino-Gómez E., Martínez G., and Hernández J. A. Playing with words: Comparing the vocabulary and lexical richness of ChatGPT and humans. Preprint at https://arxiv.org/abs/2308.07462 (2023).

26. Kobak D., González Márquez R., Horvát E. A., & Lause J. Delving into ChatGPT usage in academic writing through excess vocabulary. arXiv e-prints, arXiv-2406. (2024).

27. Webb S., & Nation P. How vocabulary is learned. Oxford University Press. (2017).

28. Vermeiren H., & Brysbaert M. How useful are native language tests for research with advanced second language users? *Bilingualism*: *Language and Cognition* 27 (1):204–213. (2024).

29. Meara P.M., & Buxton B. An alternative to multiple choice vocabulary tests. *Language Testing*, 4, 142–154. (1987)

30. Lemhöfer K., & Broersma M. Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, 44, 325–343. (2012). https://doi.org/10.3758/s13428-011-0146-0 PMID: 21898159

31. Petrov A., La Malfa E., Torr P. H., & Bibi A. Language Model Tokenizers Introduce Unfairness Between Languages. Preprint https://arxiv.org/abs/2305.15425 (2023).

32. Stadthagen-Gonzalez H., Bowers J. S., & Damian M. F. Age-of-acquisition effects in visual word recognition: Evidence from expert vocabularies. *Cognition*, 93 (1), B11–B26. (2004). https://doi.org/10.1016/j.cognition.2003.10.009 PMID: 15110727

33. Gatti D., Marelli M., & Rinaldi L. Out-of-vocabulary but not meaningless: Evidence for semantic-priming effects in pseudoword processing. *Journal of Experimental Psychology* 152 (3), 851–863. (2023). https://doi.org/10.1037/xge0001304 PMID: 36174173

34. Keuleers E., & Brysbaert M. Wuggy: A multilingual pseudoword generator. *Behavior research methods*, 42, 627–633. (2010). https://doi.org/10.3758/BRM.42.3.627 PMID: 20805584

35. van Rijn P., Sun Y., Lee H., Marjieh R., Sucholutsky I., Lanzarini F., et al. Around the world in 60 words: A generative vocabulary test for online research. Preprint at https://arxiv.org/abs/2302.01614 (2023).

36. Martínez G. et al. How many words does ChatGPT know? The answer is ChatWords. Preprint at https://arxiv.org/abs/2309.16777 (2023).

37. Touvron H., Martin L., Stone K., Albert P., Almahairi A., Babaei Y., et al. Llama 2: Open foundation and fine-tuned chat models. Preprint at https://arxiv.org/abs/2307.09288 (2023).

38. Jiang A. Q., Sablayrolles A., Mensch A., Bamford C., Chaplot D. S., et al. Mistral 7B. Preprint at https://arxiv.org/abs/2310.06825 (2023).

39. Rohan A., et al. "Palm 2 technical report." Preprint at https://arxiv.org/abs/2305.10403 (2023).

40. Brown T. B., et al. "Language models are few-shot learners." Preprint at https://arxiv.org/pdf/2005.14165 (2020).

41. Achiam Josh, et al. "Gpt-4 technical report." Preprint at https://arxiv.org/abs/2303.08774 (2023).

42. Liu Boan, et al. "Diversifying the mixture-of-experts representation for language models with orthogonal optimizer." Preprint at https://arxiv.org/pdf/2310.09762 (2023).

43. Vermeiren H., Vandendaele A., & Brysbaert M. Validated tests for language research with university students whose native language is English: Tests of vocabulary, general knowledge, author recognition, and reading comprehension. *Behavior Research Methods*, 55 (3), 1036–1068. (2023). https://doi.org/10.3758/s13428-022-01856-x PMID: 35578105

44. Chowdhery A., Narang S., Devlin J., Bosma M., Mishra G., Roberts A., et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24 (240), 1–113. (2023).

45. Bermúdez-Margaretto B., & Brysbaert M. How efficient is translation in language testing? Deriving valid Spanish tests from established English tests. https://doi.org/10.31234/osf.io/ypu9w (2022).

46. Izura C., Cuetos F., & Brysbaert M. Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica*, 35 (1), 49–66. (2014).

47. Ferré P., & Brysbaert M. Can Lextale-Esp discriminate between groups of highly proficient Catalan-Spanish bilinguals with different language dominances? *Behavior Research Methods*, 49, 717–723. (2017). https://doi.org/10.3758/s13428-016-0728-y PMID: 27004486

48. Forster K. I. The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition*, 28 (7), 1109–1115. (2000). https://doi.org/10.3758/bf03211812 PMID: 11185767

49. Kuperman V. Virtual experiments in megastudies: A case study of language and emotion. *Quarterly Journal of Experimental Psychology*, 68 (8), 1693–1710. (2015). https://doi.org/10.1080/17470218.2014.989865 PMID: 25406972