Yet another approximation of human semantic judgments using LLMs... but with quantized local models on novel data

Andrea Amelio Ravelli *
Università di Bologna

Marianna Marcella Bolognesi **
Università di Bologna

This study investigates the automatic generation of semantic norms on word specificity using various quantized open-source local Large Language Models (LLMs), including a comparison with a proprietary model (i.e. GPT-4). Word specificity norms on English are still not public, thus they are not included in the training datasets of all tested models. This offers a novel contribution by assessing LLMs ability to generalize beyond pre-trained knowledge. Our findings reveal that smaller, local quantized models such as Llama3, Phi3, and Mistral underperform in generating human-like judgments of word specificity, while a larger model such as Mixtral, even if slightly less accurate than GPT-4, represents a viable alternative to proprietary models if adequate computational resources are accessible. These findings open up new perspectives for research on linguistic features and on the scalablility of semantic norms without relying on proprietary models.

1. Introduction

As in many papers published over the past few years (and probably several included in this Special Issue) the Introduction of this contribution may start by describing the revolution in Natural Language Processing (NLP) triggered by the introduction of Transformer models with Attention by Google (Vaswani et al. 2017), which opened the doors towards Generative AI. The Transformer model uses a self-attention mechanism to weigh the importance of different words in an input sentence when encoding a particular word, and this mechanism enables models to capture long-range syntactic, semantic, or discursive dependencies more efficiently than previous models based on Recurrent Neural Networks (RNNs) or Long Short-Term Memory Networks (LSTMs). One of the most popular implementations of this recent architecture is BERT (Devlin et al. 2019), which in few months set new benchmarks in many NLP tasks. Nevertheless, with the constant growth in terms of model parameters and training data, as well as with the transition towards decoder-only architectures in much recent Large Language Models (LLMs), even the state-of-the-art achieved by BERT and similar has been not only surpassed by a margin, but also relegated to a baseline in numerous subsequent

* ABSTRACTION Research Group abstractionproject.eu
Department of Modern Languages, Literatures and Cultures – University of Bologna. Via Cartoleria 5, 40124 Bologna (BO) Italy.
E-mail: andreaamelio.ravelli@unibo.it (corresponding)
** ABSTRACTION Research Group abstractionproject.eu
Department of Modern Languages, Literatures and Cultures – University of Bologna. Via Cartoleria 5, 40124 Bologna (BO) Italy.
E-mail: m.bolognesi@unibo.it

studies (Rogers, Kovaleva, and Rumshisky 2020; Zhang et al. 2020; Zhao, Strube, and Eger 2023).

1.1 The bigger, the better

For some years, the AI industry has developed new models by adhering to "the bigger, the better" paradigm, (Kaplan et al. 2020). Architectures have grown at a steady pace for a good period, and it seemed that we were eventually going to hit the threshold of *gazillions* of parameters. As we can see, ELMo (Peters et al. 2018), one of the earliest models sensitive to context and word order (i.e., it interprets words and produces their representations by considering the specific context in which they appear, rather than static representation computed on the average of all contexts of appearance), had an LSTM network with approximately 94 million parameters when it was introduced by the Allen Institute for AI in 2018. In the same year, OpenAI launched the initial version GPT (Radford et al. 2018), which broke the ceiling of 100 million parameters. However, it was with the arrival of BERT (Devlin et al. 2019) in 2019 that we had a significant increase in model size, which grew threefold (BERT-large was released with 340 millions of parameters). Since then, we have witnessed an explosion of models, and an exponential and continuous growth, with OpenAI, in the same year, surpassing the limit of one billion parameters, with its GPT-2 (Solaiman et al. 2019) (1.5 billion parameters). Google's T5-11b (Raffel et al. 2020), also introduced in 2019, vaunted 11 billion parameters, but it is with the release of OpenAI's GPT-3-davinci(Brown et al. 2020) that we can find a significant leap, with 175 billion parameters. This rapid expansion continued into 2022 with Nvidia and Microsoft's MT-NLG (Smith et al. 2022), which featured an unprecedented 530 billion parameters. If we compare ELMo to MT-NLG, the latter is more than 5 thousands times bigger than the first. We have witnessed an explosion of models, and an exponential and continuous growth, to a point that billions of parameters was the *new black* of the new generation of LLMs, and the jury is still out on whether elements of general and true intelligence can be attributed to such sophisticated architectures.[1]

The push for larger models slowed once it was discovered that the size of the training data has a more significant impact on the performance of LLMs than the complexity of their architecture. After the empirical demonstration that more data (and larger context windows) lead to better performances on some of the most popular tasks and benchmarks (Hoffmann et al. 2022), with *few billions* parameters models outclassing way bigger ones, we witnessed a shift of paradigm, where *the bigger* refers no more to the model architecture, but to the size of the training set. *Data lakes* define precisely this new concept: large-scale repositories that replace *datasets* and indicate the extensiveness of new collections of data. The metaphor behind that name does illustrate well the depth of knowledge these systems are exposed to. The entire internet is no longer sufficient, to the extent that training data are now augmented with synthetic examples produced

1 There are often comparisons made between artificial and human intelligence, and a good portion of the lexicon used in AI results in borrowings from biology and neuroscience. It is possible that the strong overlap between the two dimensions has led AI researchers to infer that, at some point, artificial intelligence emerged from the complexity of architectures, given that the human intelligence arose from the growing complexity and inter-connectivity of biological structures. While the AI industry keeps on pushing for the interpretation of this new generation of *speaking machines* as *thinking machines*, there is much less sensationalism among academia researchers about *stochastic parrots* (Bender et al. 2021).

by other LLMs (Kumar, Choudhary, and Cho 2020; Yoo et al. 2021; Sahu et al. 2022; Li et al. 2023).

1.2 Large models, large resources

When it comes to large models, whether in terms of architecture or training datasets, the primary challenge is computational capability. In other words, the challenge is represented by the computational costs a research organization can afford. The training costs of state-of-the-art AI models have reached unprecedented levels. The Stanford Institute for Human-Centered Artificial Intelligence in its AI Index Report for 2024 (Maslej et al. 2024) estimated that OpenAI's invested $78 million worth of compute to train GPT-4 (Achiam et al. 2023), while Google's Gemini Ultra (Anil et al. 2023) required $191 millions. Given the high demanding investments required, private investors and companies play a key role in the development of these new technologies, while universities and publicly funded research centers can only test these models capability and challenge them with new tasks. A notable example of this scenario is the fate of EVALITA,[2] the Italian evaluation campaign for NLP systems, which is probably facing the conclusion (Basile 2022). Established in 2007 and conducted across 8 editions until 2023, EVALITA was designed as a comprehensive suite of tasks addressing various aspects and phenomena of language (primarily Italian), allowing participants to benchmark their systems and methodologies. In its place is now CALAMITA, a new evaluation campaign format specifically designed for LLMs. Instead of proposing new models to address specific tasks, participants now propose new tasks for a single LLM (or a small set of LLMs) to solve.

Unfortunately, in the private sector —especially with significant private investments— we cannot expect adherence to the principles of Fairness, Accountability, Transparency, and Robustness (FAIR) of research (Wilkinson et al. 2016). Despite its non-profit status, OpenAI develops and distributes proprietary models. Although the ChatGPT web interface is available for free,[3] access to its API requires a paid subscription after an initial trial period. The source code for their models, like that of many other leading LLMs, has not been released, rendering them as *black boxes*. While neural networks are inherently somewhat opaque, having public access to code and parameter weights allows for probing the model's competencies across various aspects of language, such as with probing tasks aimed at investigating what happens during the processing by observing the representations from each layer of the architecture (Alain and Bengio 2017; Conneau et al. 2018; Miaschi et al. 2022).

We were still analyzing the impacts of the *BERTology* era (Michel, Levy, and Neubig 2019) when the *GPTology* era (Ong 2024) emerged. A search for *"[anything] + (chat)GPT"* on Google Scholar or other search engines reveals numerous studies using OpenAI's models for various tasks across diverse research fields. In some applications, especially those involving sensitive data like medical information, using closed models hosted on third-party servers raises significant concerns about data security (Wu, Duan, and Ni 2024). Consequently, open-source alternatives such as LLaMa (Touvron et al. 2023) and Mistral (Jiang et al. 2023) present advantages. These models are accessible and adaptable to various scenarios, and their absence of licensing fees or restrictive usage policies makes them viable for many users, including smaller research organizations.

2 `https://www.evalita.it`

3 Like many popular *free services*, users pay through their data, which becomes the currency.

The already mentioned size and computational cost of these models is still a significant barrier to their deployment on-premises, either on private servers or personal computers, limiting the possibility of manipulating these models for individual researchers or students. To overcome this, various techniques can help reducing the size of large models, such as model pruning (LeCun, Denker, and Solla 1989), low-rank factorization (Srebro and Jaakkola 2003) or adaptation (Hu et al. 2022), knowledge distillation (Hinton, Vinyals, and Dean 2015), or quantization (Gray and Neuhoff 1998).

Model quantization consists in converting the weights and activations within a language model from a high-precision data representation (i.e. floating point) to a lower-precision representation (i.e. integer). Essentially, this process transforms data types that can store extensive information into ones with less storage capacity, allowing for the optimization of memory usage during inference. Quantized models have gained significant popularity due to the development and dissemination of simple-to-use open-source software libraries for inference, such as llama-cpp,[4] ollama,[5] or lmstudio.[6] These software enable anyone to load a model locally, without the need of an internet connection and, more importantly, on consumer hardware such as common laptops. The quantization process entails a trade-off between computational efficiency and accuracy (Yao et al. 2023), but some studies (Lee et al. 2023; Liu et al. 2024; Ramesh et al. 2023) show that quantized models performance are comparable to their original counterparts on many popular benchmarks.

The question remains open as to whether open, local and quantized models compare favorably with larger, proprietary LLMs in real use cases scenarios. To ensure a fair comparison, it is crucial that the models are tested under neutral conditions, namely: on a task for which the human benchmark used for evaluating the models' performance is safely novel, not previously "seen" by any of the contenders. For this study, we have provided a novel dataset consisting of human-generated semantic judgments of word specificity.

The goal of this investigation is to determine to which extend local quantized models can be used as a viable strategy to approximate human judgments in a typical psycholinguisitcs norming study. We test some quantized open LLMs to reproduce an annotation campain to collect values of word specificity, using as gold standard a norming study (Ravelli, Bolognesi, and Caselli 2024) that is definitively not included in any LLM's training set.[7] We faithfully reproduce the data collection protocol used with humans, and specifically we employ the best-worst scaling methodology. To the best of our knowledge, our work is the first to use this methodology with local and quantized open LLMs for simulating psycholinguistic norming studies.

To summarize our contribution:

1. we exploit some quantized local LLMs to replicate an unpublished (thus not included in any training data) psycholinguistic norming study on word specificity, by means of best-worst scaling methodology;
2. we evaluate the resulting specificity scores by correlating them to those obtained with humans, and those obtained with GPT-4

4 https://github.com/ggerganov/llama.cpp
5 https://ollama.com
6 https://lmstudio.ai
7 At the time of writing, the paper is in the final round of reviews before publication.

3. we discuss on the viability of using quantized local LLMs to generate psycholinguistic data

2. Related Works

2.1 Human ratings replication as testing field for (Large) Language Models

In NLP, many tasks have the objective of generating predictions along a continuous range of values (e.g. 0-1 range with floating values) and verify them with regressions against human-generated ratings, which are often derived from psycholinguistic or behavioral studies. The most common methodology to collect human judgments is to ask participants to vote stimuli with Likert scales (Likert 1932). Although it is an easy annotation methodology to implement, it is not easy the same from the point of view of the rater, which has to approximate their choice, leading to many possible biases, such as ordina-cardinal conflation (Ethayarajh and Jurafsky 2022), interval preference (Kiritchenko and Mohammad 2017), or central tendency bias (Guilford 1954). Best-Worst Scaling (BWS) (Louviere and Woodworth 1991; Louviere, Flynn, and Marley 2015), on the contrary, collect judgements by presenting participants with a set of items and asking them to identify those which maximise or minimize the variable of interest. This approach requires participants to make explicit comparisons between stimuli, resulting in an easier and more reliable task (Kiritchenko and Mohammad 2017), thanks to the *context*[8] in support of the choice. Recently, this methodology has been applied in exploiting LLMs to model emotion intensity, showing that the BWS methodology is more reliable than Likert scales also with artificial agents (Bagdon et al. 2024).

2.2 Large Language Models as a mean to expand semantic norms

Recent studies in cognitive science, cognitive psychology, linguistics and related disciplines report experiments involving the approximation of language-based judgments provided by humans, using LLMs (Aher, Arriaga, and Kalai 2023; Argyle et al. 2023; Törnberg 2023; Zhu et al. 2023; Gilardi, Alizadeh, and Kubli 2023; Trott 2024). These studies explore various levels of language processing, leveraging the extensive capabilities of LLMs to annotate and analyze linguistic data, and their overall good correlation with human benchmarks. The proliferation of such experiments underscores the growing reliance on LLMs as tools for psycholinguistic research and their potential to scale up existing psycholinguistic resources with unprecedented speed and good accuracy, as argued recently (Trott 2024).

Among human-generated datasets that have recently attracted the attention of researchers interested in scaling up resources with LLMs, there are datasets of so-called *semantic norms*. These datasets typically consist of evaluations of word meanings through the collection of human judgments about one or more semantic dimensions (e.g., concreteness, familiarity, imageability, age of acquisition, valence, etc). Collecting such data is often time-consuming and labor-intensive, as it requires extensive participation from human subjects and careful consideration of the experimental design and instructions they are presented with. Furthermore, these datasets of semantic norms are

8 The best and the worst options are chosen among a group, and not judged in isolation. In this terms we can consider the set of stimuli as the context in which the two selected items configure as the best and the worst.

often limited in size. LLMs have recently been used to assist researchers in generating and expanding semantic judgments more efficiently.

In a recent study (Trott 2024), GPT-4 has been exploited to gather various types of semantic norms for English words, and compared them against human "gold standard" judgments stored in published datasets of norms. Across each dataset, it was found that GPT-4's judgments positively correlated with human judgments, sometimes even matching or surpassing the average inter-annotator agreement observed among human participants. However, many of the datasets used in this study, as well as many datasets used to investigate other types of linguistic annotations in previous experiments, have been published in prior research before 2022. It cannot be excluded, therefore, that LLMs like GPT-4 have "seen" such datasets in their training set. This raises concerns about the extent to which the models' performance and correlation to a specific dataset of human judgments may be influenced by prior exposure to the dataset itself. If the data were part of the training corpus or otherwise accessible to the models, their ability to generate accurate responses and high correlation with human judgment might be skewed by prior knowledge. This familiarity with the data could lead to inflated performance metrics, as the models may be leveraging pre-existing information rather than demonstrating genuine inferential capabilities. This phenomenon, also known as *data leakage*, or contamination and overlap between training and test set, is addressed in the study and partially contrasted, but as the author suggests, "future work should aim to address this issue by continuing to evaluate an LLM's performance on norms that are unlikely to have been observed in its training set" (Trott 2024). The issue of bias arising from asking participants to perform tasks they have previously encountered is a crucial consideration. When individuals are aware of the tasks or have seen similar tasks before, their responses may be influenced by this prior knowledge, potentially introducing biases into the results. Similarly, LLMs exposed to familiar data might demonstrate a form of "bias" where their outputs reflect the patterns they have already learned rather than providing novel insights. Understanding and addressing these biases is essential for ensuring the reliability and validity of both human and model-based annotations, particularly when interpreting findings from experiments involving repeated or familiar tasks.

Based on these considerations, as well as those provided in the Introduction, the present study reports a series of experiments aimed to investigate and discuss the difference in performance between open vs proprietary LLMs in their ability to approximate semantic norms of word specificity, a variable that is becoming increasingly popular in cognitive sciences and related disciplines (Bolognesi, Burgers, and Caselli 2020; Bolognesi and Caselli 2023; Ravelli, Bolognesi, and Caselli 2024), and for which there is still a substantial lack of resources, thus the opportunity to exploit LLMs would be highly desirable to overcome time and costs of annotation campaigns.

3. Specificity ratings

Categorical specificity refers to the degree of inclusiveness with which a word denotes a specific category or concept (Bolognesi, Burgers, and Caselli 2020; Bolognesi and Caselli 2023). Specificity is a a variable involved in the phenomenon of conceptual abstraction, which defines the level of generality or specificity at which a concept is expressed. Higher levels of abstraction represent more general, less detailed categories (e.g., "animal"), while lower levels of abstraction represent more specific, detailed categories (e.g., "Husky").

Specificity shall not be confused with Concreteness, which refers to the extent to which a concept refers to something that can be experienced through sensory channels. Concrete concepts can be experienced directly through the senses (e.g., "apple"), while abstract concepts are more intangible (e.g., "justice"). While there is a positive correlation between concreteness and specificity (Bolognesi et al., 2023), where more concrete concepts tend to be more specific, and less concrete concepts tend to be less specific and therefore more general, the two variables are theoretically distinct. However, while the scientific literature provides various resources of semantic judgments of word concreteness (Brysbaert, Warriner, and Kuperman 2014; Montefinese et al. 2014; Guasch, Ferré, and Fraga 2016; Soares et al. 2017; Yao et al. 2017; Bonin, Méot, and Bugaiska 2018), with recent interests also in the influence of the linguistic context over the perception of the variable (Gregori et al. 2020; Montefinese et al. 2023), for Specificity the resources are quite limited.

One reason for the scarcity of resources regarding semantic judgments of categorical specificity is the difficulty human subjects face when evaluating specificity using classic Likert scales, especially when words are presented in isolation. Specificity, by its nature, is a relational attribute. A word denoting a conceptual category can be hardly judged for specificity in absolute terms. A word is instead perceived to be more specific than another, or less specific than another. For instance, "dog" is more specific than "animal", and less specific than "Husky". Consequently, experimental paradigms that incorporate this relational characteristic are necessary. These paradigms, however, are more demanding in terms of the effort required and the number of stimuli and participants needed to effectively carry out the tasks, because each word, to be properly evaluated, must be compared to various other words.

Previous results show that Specificity is an important variable in psycholinguistic research, that holds the key to explaining various processing advantages, some of which have been previously attributed solely to concreteness. For instance, variations in word Specificity have been shown to have an impact in lexical and semantic decision latencies during language processing (Bolognesi and Caselli 2023; Ravelli, Bolognesi, and Caselli 2024) as well as in word-picture categorization tasks (van Hoef, Connell, and Lynott 2023). Morever, Specificity has been shown to explain a portion of the variance in contextual availability above and beyond concreteness (Rambelli and Bolognesi 2023, 2024). To further investigate the role of categorical specificity in cognitive processing, more extensive resources for specificity ratings are urgently needed. Because the Best-Worst Scaling (BWS) method is time-consuming when involving human participants, a promising direction that we hereby investigate is the possibility to augment existing resources and datasets using large language models (LLMs) in a responsible and sustainable manner for researchers. This is accomplished by using open, quantized models, and comparing their performance to a human gold standard (i.e., existing dataset of BWS spec ratings in English) as well as to a reference model as machine gold standard (i.e., GPT-4).

4. Generating specificity norms with LLMs

The experiment reported here addresses the ability of LLMs to generate linguistic norms of categorical specificity for a set of English words, replicating the recently completed data collection with human annotators. The dataset of words is the ANEW (Affective Norms for English Words) (Bradley and Lang 1999), which consists of 1,034 emotionally charged words—words with a high potential to evoke emotions (positive or negative) in those who read them. Despite its small size, this dataset is quite popular in the field

of psycholinguistics, and many norms have been collected for the words featured in this dataset.

In preparing our task, we aimed to maintain consistency with the experimental settings used to collect human judgments with the same method, on the same dataset. We therefore employed the same sets of 4-word tuples, created using the method described in (Kiritchenko and Mohammad 2017). A stimuli set consisting of 2,068 four-word combinations was generated and distributed across 53 annotation lists, each containing 40 4-word tuples organized based on their part of speech. For each local LLM tested and for each annotation list, we ran the experiment 12 times to simulate annotations from multiple individuals, and we aggregated the results for comparison with the human gold standard.

We ran the task using Ollama[9] as local LLMs server for quantized models and Langchain[10] as the framework for invoking models and performing inference. This combination of tools allowed us to efficiently design the workflow in a simple way, and to exploit different models (local and hosted) without the need to code ad-hoc scripts for each of them.

We used an *off-the-shelf* approach, in the sense that we do not tune any parameters (e.g., temperature, top_k, top_p, etc.) of the tested models,[11] nor prepared them to the task with a fine-tuning. In the spirit of a genuine emulation of human judgement ratings with the average speaker (i.e., not specifically trained to the task), we decided to keep to the minimum any possible *influence* on the models. The objective was to simulate an annotation campaign with crowd-sourced raters, who typically lack specific knowledge in the fields of linguistics or psychology, rather than to emulate the performance of expert annotators. The code to generate specificity ratings with LLMs is available in this OSF repository: `https://osf.io/j8xb5/`.

4.1 Methodology

The same instructions given to human annotators were used in this experiment. We provided these instructions as a `system prompt`, introducing each request. Using chains from Langchain, we configured our requests as `system prompt + user request | model`, where the `user request` is the 4-word tuple stimulus.

The task could be classified as a *two-shots learning* task, even if we did not explicitly chain examples along with answers in the prompt template. Within the instructions for humans there were two examples, and they have been marked for the models using the tags `[EXAMPLE]` and `[/EXAMPLE]`, at the beginning and end of each, and they are followed by the correct answers tagged with `[ANSWER]` and `[/ANSWER]`, making it clearer to the models which parts of the text corresponded to examples and answers during the parsing of the system prompt. In this way, we did not alter the information and their order in the instructions with respect to the version used with people. The only modifications to the original instructions were the addition of two sentences at the end, specifying that the model should respond exclusively by providing the indexes of the two words, among the four presented as stimuli, selected as best and worst, separated by a comma (e.g., "1, 3" or "4, 1"). We chose to query the models for the numeric indices

9 `https://ollama.com`

10 `https://www.langchain.com/langchain`

11 We experimented with some basic parameter tuning, but the results were that some models were replying always with the same choice, limiting the possibility to observe the variability that naturally occurs when dealing with humans.

associated with the words in the tuples rather than the words themselves due to an unusual behavior observed during the experiment setup phase. Indeed, sometimes the response might include a graphically comparable word instead of an actual tuple word in certain instances, causing an error in the step of aggregating all replies to compute the best-worst scaling results: e.g., the word *dove* were often re-written as *Dover*. This could be probably due to the default settings for temperature (0.8) and top_p (0.9) in the Langchain implementation of Ollama. To address this issue, we assigned a number in the range of 1–4 to each word in a tuple and instructed the model to respond using these numbers. While some models tended to ignore the instruction not to generate additional text, they still provided the two indices separated by a comma, often followed by an explanation of the decision on a new line. Interestingly, in some responses, we observed that the models began to recreate the task with new combinations of candidate words. To prevent the generation of unwanted text, we set a limit of 48 tokens. We then easily converted the indices back to words using simple text post-processing. Due to the censorship of some models, not all tuples containing "sensitive" words (e.g., *rape*) were annotated, and the output contained only a circumstantial phrase by the model; we opted to disregard such cases without a valid response.

Finally, we assessed the performance by correlating the specificity values obtained with the annotations generated by models with those produced by humans.

4.2 Models

In this section, the models used for the experiment are described. Table 1 summarizes some information about the models. Local quantized models have been downloaded from the Ollama website; we opted for *instruct* versions due to their better focus on the task execution.

Table 1
Features of the models used in this experiment. Asterisc indicates unofficial information.

Model	Size	Release	Arch.	Params	Context	Quantization
llama3:instruct	4.7 GB	18/04/2024	llama	8.0B	8,192	Q4_0
phi3:instruct	2.4 GB	23/04/2024	phi3	3.8B	4,096	Q4_K_M
mistral:instruct	4.1 GB	27/09/2023	llama	7B	32,768	Q4_0
mixtral:instruct	26 GB	11/12/2023	llama	47B	32,768	Q4_0
GPT-4	-	14/03/2023	gpt	1.76T*	8,192	-

4.2.1 LLama 3

LLama 3 (Dubey et al. 2024) is the latest version of the LLama family of LLMs by Meta. It has been released in 8 billion and 70 billion parameter configurations, both as pre-trained and instruction-tuned versions. According to the technical report, these models expand their functionalities by intruducing advanced reasoning capabilities. LLama3 models show state-of-the-art performance on a variety of industry benchmarks, outperforming many of the available open-source chat models.

4.2.2 Phi 3

Phi 3 (Abdin et al. 2024) is a family of lightweight open models developed by Microsoft, with state-of-the-art performance with respect to bigger models. The *mini* model

features 3.8 billions of parameters, and has been extensively trained on a mixture of synthetic and filtered publicly accessible websites information, with the goal of develop high quality reasoning properties. Upon evaluation against benchmarks assessing common sense, language comprehension, mathematics, coding, long context, and logical reasoning, the Phi 3 mini model demonstrated competitive performance among models up to 13 billion parameters.

4.2.3 Mistral and Mixtral

Mistral (Jiang et al. 2023) is the flagship model of Mistral AI, and it demonstrates that high performances can be obtained without the need to sacrifice efficiency. The 7B-sized model scores better than best 13B models on all the benchmark tested by the developers, and also than some 34B models from competitors in mathematics and code generation. Mistral architecture employs two variation of the attention mechanisms: grouped-query attention (GQA) and sliding window attention (SWA). On hone and, The implementation of GQA accelerates inference speed and reduces decoding memory consumption; on the other, SWA effectively manages longer sequences at decreased computational expense.

Mixtral (Jiang et al. 2024) is part of the Mistral family of models, but it implements a sparse mixture-of-experts (MoE) network. Each layer in Mixtral features a feedforward block that selects from a set of 8 distinct sets of parameters (i.e. 8 Mistral 7B models). For every token, a router network determines two *expertis* out of the set to process the token and aggregates their outputs additively. This approach expands the model's parameter set while controlling cost and latency by exploiting only a fraction of the total parameters per token. Mixtral features 46.7 billion total parameters but employs about 12.9 billion parameters for each token. As a result, Mixtral retains input processing and output generation speeds, as well as costs, equivalent to those of a 13B parameter model.[12]

4.2.4 GPT 4

GPT 4 (Achiam et al. 2023) is a state-of-the-art language model developed by OpenAI, representing a significant advancement over its GPT predecessors. One such innovation is the supposed[13] use of a MoE architecture, where different components of the model can specialize in various subtasks or domains, and the reply is the result of the reasoning of multiple models. For our experiment we employed the GPT-4 *base* model (i.e., the model accessible by calling *gpt-4*: gpt-4-0613) with updated training as of September 2021 and a context window size of 8k tokens.

5. Results

For both human and LLM-generated annotations, we applied the BWS methodology to compute Specificity scores for the target words, by adapting the python script from (Basile and Cagnazzo 2021). We computed Specificity scores on both the full set of

12 Due to its memory footprint, this model has not been run on the testing consumer computer. It was possible to test it effectively during the experiment design phase, but for an efficient execution in terms of time (and due to the necessity of having the laptop available), we chose to perform the annotation on a dedicated CPU-only computing machine with larger memory.

13 No specific technical details regarding the model and its architecture have been disclosed or confirmed. Rumors suggest a multi-agent architecture with an estimated 8x220 billion parameters.

Table 2
Spearman r correlation with human annotations.

MODEL	Single run				All runs (12 per list)
	Mean	Min	Max	Std	
Llama 3	0.099	0.076	0.131	0.016	0.120
Phi 3	0.334	0.316	0.352	0.011	0.369
Mistral	0.321	0.304	0.342	0.011	0.371
Mixtral	0.627	0.620	0.633	0.003	0.636
GPT 4	0.772	-	-	-	0.772

stimuli lists and individual lists: the former was used for per-model evaluation, while the latter was used for per-annotator evaluation in the ablation study described in 5.1.

Table 2 reports the results of the experiments in terms of Spearman r correlation with the human annotation. Data in Figure 1 illustrate the results in two groups of bars: *single run* bars show the average correlation of one round of annotation from a model, while the bars of *multiple runs (12)* report the correlation of the aggregation of all the 12 rounds of annotation with each model, with the exclusion of GPT4, which has been exploited for only one round of annotation.[14]

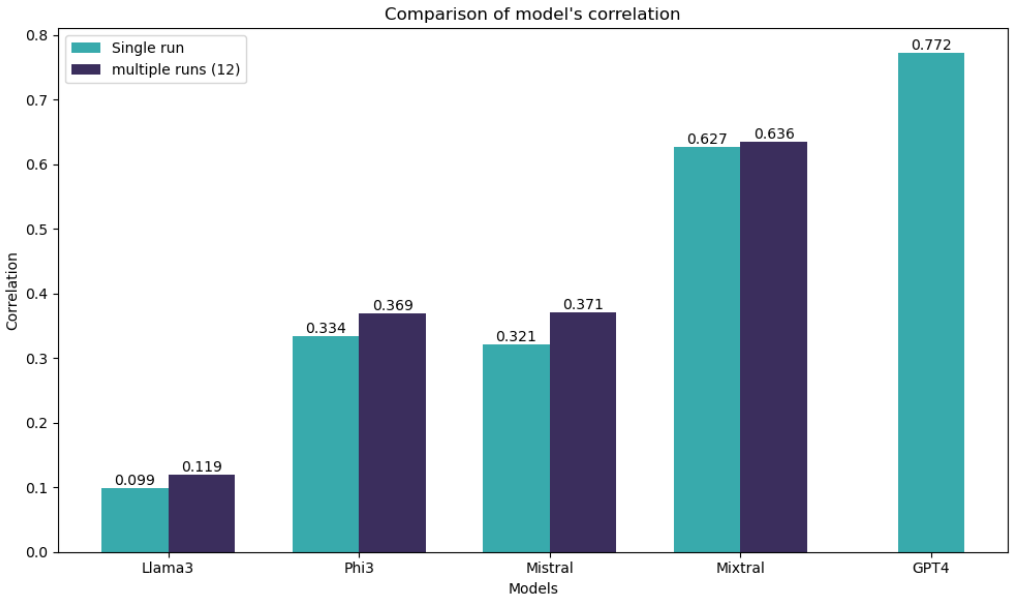

Figure 1
Correlation with human data of Specificity scores obtained with LLMs. Lighter bars: average over 12 runs (except GPT-4); Darker bars: sum of all the 12 runs.

14 As stated in Section 3 we have included GPT-4 in this experiment as a *machine gold standard*, given that it is the model of choice for these kind of psycholinguistics explorations with LLMs. The cost of the annotations with the openAI APIs (53 lists of 40 stimuli for one run, thus a total of 2,068 API calls) was of $36.40.

Average correlation per single annotator

If we focus on the performances of only one round of annotation, we can see that the Specificity ratings obtained with the Llama 3 model have little to no correlation with those derived from human annotation, with an average Spearman r over 12 runs of 0.099. Phi 3 and Mistral show comparable performance, with an average Spearman r of 0.334 and 0.321 respectively. Mixtral records the highest correlation among local quantized models, scoring about the double of the other models, with an average of 0.627. GPT 4 obtains the highest score, correlating with human derived Specificity with a Spearman r of .0.772.

Correlation of all annotators

If we aggregate the data from all 12 runs of the local quantized models, similar to the procedure used with human annotators, we observe a general trend of higher correlation with human derived data across all the models, but not a significant increase compared to the average of the single run correlations. The model that benefit more from the aggregation of multiple runs is Llama 3, whose performance shows an increase of about the 20% with respect to the average correlation of a single run, but still not sufficient to show a considerable correlation (Spearman r: 0.119). The correlation of Phi 3 increases of about the 10%, scoring a Spearman r of 0.369, and we can see a similar result with Mistral, which get a correlation of 15% higher (0.371) than the average of a single run. Mixtral, instead, do not show a significant increase with the aggregation of multiple runs, with a gain of a bit more than the 1% (0.636).

5.1 Annotation consistency

As explained in Section 4.1, during our experiments we gave the models wide freedom in generation in order to prioritize creativity over precision, trying to avoid to obtain always the same reply (i.e., the model selects the same best and worst out of a 4-word tuple at every run) and to better mimic the results from multiple human annotators. An ablation study was carried out to evaluate if this strategy lead to a consistent data variability. For each stimuli list, we calculated the BWS with the data from every separate run and with the data from the remaining eleven runs. We then analyzed the correlation of the two Specificity scales resulting from the computation of these two BWS scores. GPT 4 model has been run only once, so we have no data to perform an ablation. Moreover, an ablation analysis was conducted also on the data collected with humans to evaluate inter-annotator consistency among them and compare the results with LLMs.

Table 3
Ablation of each run against the others, per list.

MODEL	% Significant	Mean	Std	Min	Max	Median
Llama 3	99.371	0.736	0.059	0.464	0.916	0.742
Phi 3	100	0.771	0.059	0.488	0.914	0.773
Mistral	99.843	0.737	0.085	0.338	0.923	0.749
Mixtral	100	0.954	0.029	0.761	0.991	0.963
HUMAN	94.369	0.587	0.132	0.199	0.869	0.613

Table 3 reports the results of the ablation conducted on all the annotation runs, over every annotation lists, per model, plus the ablation on the human collected data. The *Significance* column indicates the percentage of significant correlations obtained against the total number of raters. As we can see, all the models display strong consistency in their decisions across various runs, despite no generation constraint (i.e., no parameter tuning). On average, there is an inter-annotator correlation of 0.8, with Mixtral exhibiting a peak of over 0.9. Notably, this extremely high inter-rater ablation correlation on Mixtral results confirms that combining more annotation runs from this model does not yield substantial enhancements. Judgement consistency among human raters was found to be lower, with a mean correlation below 0.6, as opposed to the higher correlations and consistency observed among the models.

Given that one of the objectives of this study is to verify to which extend a LLM can be exploited in typical psycholinguistics rating tasks, the high *internal* consistency of these models do not reply to our question. Thus, we execute another ablation study but considering, for every annotation run in each annotation list, the resulting Specificity BWS from an agent against all the human annotators' judgments collected for the same list of stimuli.

Table 4
Ablation of each annotation run over single lists from LLMs against annotations from humans

MODEL	Total	Significant	Mean	Std	Min	Max	Median
Llama 3	636	78 (12.264%)	0.173	0.142	-0.327	0.323	0.205
Phi 3	636	304 (47.799%)	0.264	0.075	0.162	0.514	0.246
Mistral	636	288 (45.283%)	0.254	0.066	0.163	0.502	0.243
Mixtral	636	574 (90.252%)	0.425	0.098	0.188	0.681	0.423
GPT 4	53	53 (100%)	0.596	0.098	0.420	0.789	0.615
HUMAN	515	486 (94.369%)	0.587	0.132	0.199	0.869	0.613

Table 4 reports the summary of the ablation of single LLM raters against human raters. The table shows the total number of annotations, the count of significant annotations (i.e., the number of annotations with $p < 0.05$) and their percentage, and typical statistical measures (mean, std, min, max, and median) for each model. What we see here confirms the trend observed in the general results. The synthetic annotators based on Llama 3 exhibit significantly different behavior compared to human annotators, with an average cross-ablation correlation of 0.173. Notably, about the 12% of the Llama 3 raters showed a significant correlation, thus very limited. Annotators based on Phi 3 and Mistral produced again similar outcomes, obtaining a correlation score with human counterparts of 0.264 and 0.254, respectively, but still with not enough annotators with a significant correlation (less than the 50% for both models). Once again, the correlation scores of Mixtral and GPT 4 represent a significant leap forward with respect to the results of the other tested models, confirming that the MoE architecture leads to better performances. Mixtral achieved an average cross-ablation correlation of 0.425, and approximately 90% of synthetic annotators based on this model yielded statistically significant results. Synthetic annotators based on GPT 4 all produced statistically significant correlation values, regardless of the *one-shot* annotation (i.e. only one rater per list and not 12 as per the local LLMs tested), scoring an average correlation of 0.596 when compared to human.

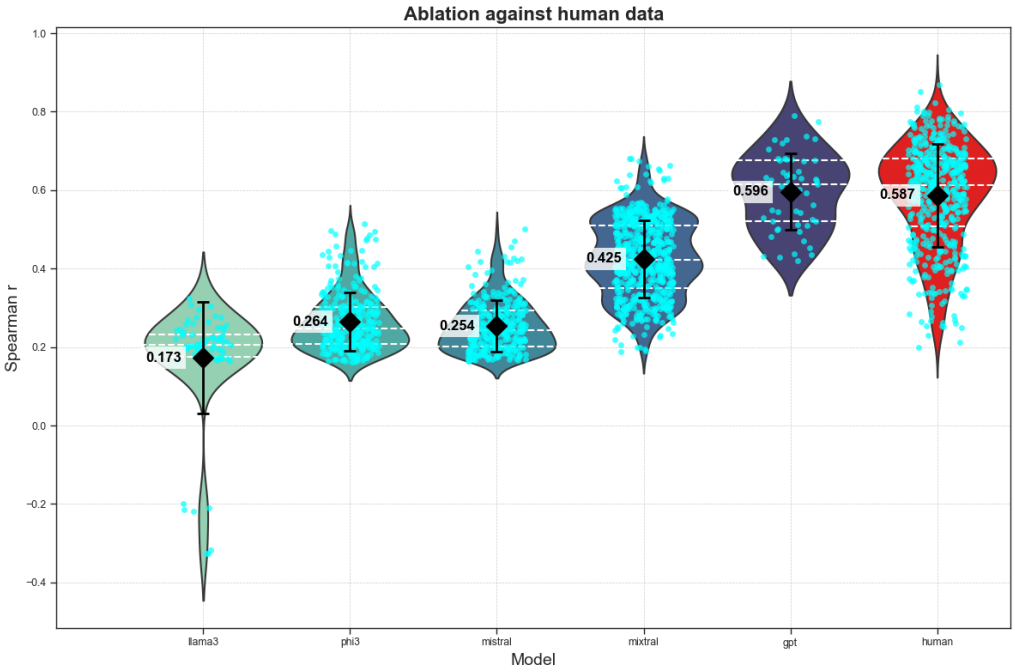

Figure 2
Distribution of statistically significant correlations between LLMs and human raters, plus the inter-rater correlations from human raters for comparison (red violin on the right).

GPT-4 cross-ablation correlation surpasses, slightly, the average ablation correlation scored by human annotators, suggesting that synthetic annotations collected with this model may be indistinguishable from human annotations. This behaviour is particularly evident if we take a look at Figure 2. The violin plot shape of GPT-4 and, consequently, the distribution of individual correlation values for this model, is remarkably similar, and almost indistinguishable, to that of human annotators. Nevertheless, by applying typical psycholinguistic data filtering methods to human correlations, such as the outlier boundaries defined as $mean \pm 1.5 * std$, it yields a lower bound of 0.389. Using this as threshold, we realize that all the synthetic annotations generated by GPT 4 (min: 0.42) and more than half of those from Mixtral (median: 0.434) could potentially pass unnoticed among the data collected with human raters.

6. Discussion

This investigation allowed us to assess the capabilities of off-the-shelf local quantized LLMs within a common psycholinguistic research scenario, i.e. the collection of human judgements on a semantic feature of the lexicon, namely Specificity. The objective was to evaluate the potential applications and limitations of these models in such context, as well as their ability to generate annotations comparable to those produced by human labelers.

Overall, we reported that Llama3, Phi3 and Mistral models displayed low correlations with the annotations provided by human raters. Conversely, the results obtained with Mixtral, which is a MoE model, are quite promising.

In particular:

- Through the analysis of the average correlation per single annotator we found that LLama3 is not able to tackle this task, with a very low correlation with human data (Spearman r: 0.099); Phi3 and Mistral perform better than Llama3, but still with a mild correlation with our human gold standard (0.334 and 0.321, respectively); MoE models such as Mixtral and GPT-4 are able to make BWS choices that better reflect human behaviour, with Mixtral scoring a correlation of 0.627 Spearman r and GPT-4 (used as *machine gold standard*) obtaining a correlation of 0.772 with just one single run.
- Through the analysis of the correlation of all annotators we found that running the experiment multiple times, thus emulating the typical data collection with many raters and averaging the results, benefits overall all the models, with an average increase in terms of correlation with human data of about the 15% for Llama3, Phi3 and Mistral, while it is not effective the same with Mixtral, which shows only a 1% increase in performance, at the cost of x12 the time needed for executing all the runs in this experiment.
- Through the analysis of the annotation consistency with ablation, we found that all the models show less variability than humans, with higher inter-annotator correlation with respect to human raters, despite the relative freedom expected from the configuration of the inference parameters.
- Through the comparison of models' annotations and the human ones with cross-ablation (i.e. one synthetic annotator vs. all human annotators on the same annotation list), we found that all the synthetic annotations generated by GPT-4 and more than half of those generated by Mixtral could easily blend in the group of human annotators. While it is not feasible to use GPT-4 to cheat on crowdsourcing platform due to the cost that may probably surpass the earnings from those platforms, it would be easier to exploit Mixtral.

We incorporated GPT-4 as a comparison model within our experimentation due to its widespread usage, both for its state-of-the-art performance across various tasks and the ease of access for many researchers through the *chatGPT* web interface, but we did a *one-shot* annotation with this model, because its performance were not the focus of our experiment. Nevertheless, we can observe that GPT-4 within our experiment obtain results comparable and in line with the findings from previous works on approximation of human semantic judgements (Trott 2024).

It is not surprising that GPT-4 performs the best in this task, but its result allows us to better evaluate the result obtained with Mixtral, since both are MoE models, making them comparable architectures. We can observe a performance advantage of approximately 15% of GPT-4 over Mixtral, regardless of whether we consider the average, sum, worst or best scenario across the 12 runs of the latter model. Performance-wise, Mixtral cannot be considered the model of choice to conduct such data collection, or to augment it. However, it is important to highlight that research projects may have specific requirements or constraints, such as time, costs or accessibility to high

performance computing equipment. If the latter is not a problem, Mixtral could still be a valuable option, especially when privacy and data security is a concern.

7. Conclusions

In this study, we conducted a series of tests aimed at automatically generating semantic norms related to the semantic variable of word specificity, leveraging and comparing the performance of various LLMs. A key innovation of our work lies in the fact that, differently to other studies replicating human data collection through LLMs, the task and the data we used were not included in the training datasets of any of the LLMs used. This is due to the fact that, at the time of the models' training, no publicly available resources or datasets with human ratings was addressing word specificity as a target semantic feature.[15] As a result, the ability of these models to infer specificity presents a novel contribution to the field, highlighting the models' potential to generalize beyond their training data and opening up new avenues for exploring linguistic variables in computational settings and possibly scale up efficiently and reliably the collection of semantic norms.

We illustrated how smaller local quantized models such as Llama3, Phi3, and Mistral do not yet deliver performance levels that would qualify them as robust solutions for advanced tasks such as the generation of human-like semantic judgments on novel variables, such as word specificity. Their utility in relation to these tasks may be limited to preliminary experimentation or piloting on personal computing systems. However, models like Mixtral, along with other similar alternatives, present a feasible option that may be considered as an alternative to GPT-4, albeit with a slight reduction in overall performance. If access to dedicated computational infrastructure is available, it would reduce the time investment for the execution of MoE models.

8. Limitations

It is crucial to emphasize that the experiments reported here were conducted using off-the-shelf versions of the tested models, indicating potential for further enhancement through fine-tuning or optimization techniques, such as more efficient prompt engineering. While fine-tuning is also an option with GPT-4, this approach significantly raises the financial costs of the experiments. In contrast, with locally hosted models, the primary constraint shifts from financial expense to time consumption, making them an attractive option when resources are more constrained but flexibility in tuning is available.

Many limitations may be addressed to this work, especially due to the nature of LLMs and their unpredictability in the generation, and to the complexity of the evaluation of their output. Moreover, minimal changes in the prompt can have a serious impact on the results. In the experiments we described herein, we opted for the most *ecological* way of presenting the task, as if it was conducted with human subjects, without changing the instruction given to them. We already had the Gold-Standard against which verify the performances of the models, thus we choose to measure the agreement, again, in order to evaluate models just as human subjects in these kind of

15 It is important to emphasize that, given the lack of comprehensive documentation regarding the training data of currently available models —whether closed- or open-source— it cannot be categorically stated that data related to the specificity of words are not included at all. However, it is unequivocally certain that this specific dataset of word specificity for English, developed through best-worst scaling, has not been part of any training process.

judgment collections. We decided to not limit the *freedom* of the models to let them *change idea* at each run just as different persons may do. We are aware that the BWS methodology may not be the most efficient way to elicit reasoning about specificity in Language Models. As an example, the comparison of simple copular constructions like *"X is a kind of Y"*[16] may be more effective in verifying a vertical and taxonomic relation such as the specificity of the concept evoked by a word. Still, the BWS methodology is a convenient way to collect relational features, especially when those are not already available. While it would be easy to assess if *lion* is more or less specific than *animal*, it is not the same if we compare two words referring to concepts which are not in a direct taxonomic relation, for which it is necessary an effort of abstraction. With the BWS we can try to force on this effort, and to overcome the lack of taxonomic relations by randomly combining target words in 4-word tuples, and by presenting a word in multiple contexts of 4 words.

Funding

Funded by the European Union (GRANT AGREEMENT: ERC-2021-STG-101039777). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Author contributions

This article is the result of the collaboration between the two authors. In particular, AAR and MB conceived the study, AAR designed the experimental setup. AAR collected the data and run all the experiments and analyses. Both author wrote the article, reviewed and approved the final version of the manuscript before submission. For the specific concerns of the Italian academic attribution system, AAR is responsible for sections 1,2,4,5, MB is responsible for section 3 and 7. Section 6 was written together by the two authors.

References

Abdin, Marah, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aher, Gati V., Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *ICML'23: Proceedings of the 40th International Conference on Machine Learning*, pages 337–371, Honolulu, Hawaii, USA, July. PMLR.

Alain, Guillaume and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Workshop Track Proceedings*, Toulon, France, April 24th-26th.

16 An example of task based on these constructions is the PRETENS @ SEMEVAL 2022 (Zamparelli et al. 2022), in which the objective is to evaluate the acceptability of sentences such as *the lion is a kind of animal* and *the animal is a kind of lion*, in multiple languages.

Anil, Rohan, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 1.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Bagdon, Christopher, Prathamesh Karmalkar, Harsha Gurulingappa, and Roman Klinger. 2024. "you are an expert annotator": Automatic best–worst-scaling annotations for emotion intensity modeling. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7924–7936, Mexico City, Mexico, June 16th - 21st. Association for Computational Linguistics.

Basile, Valerio. 2022. Is EVALITA done? On the impact of prompting on the Italian NLP evaluation campaign. In Debora Nozza, Lucia C. Passaro, and Marco Polignano, editors, *Proceedings of the 6th Workshop on Natural Language for Artificial Intelligence, NL4AI 2022*, volume 3287, pages 127–140, Udine, Italy, November 28th - December 2nd.

Basile, Valerio and Christian Cagnazzo. 2021. Litescale: A lightweight tool for best-worst scaling annotation. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 121–127, Held Online, September 1st - 7th. INCOMA Ltd.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, Virtual Event, Canada, March 3rd - 10th. Association for Computing Machinery.

Bolognesi, Marianna, Christian Burgers, and Tommaso Caselli. 2020. On abstraction: decoupling conceptual concreteness and categorical specificity. *Cognitive Processing*, 21(3):365–381.

Bolognesi, Marianna Marcella and Tommaso Caselli. 2023. Specificity ratings for italian data. *Behavior Research Methods*, 55(7):3531–3548.

Bonin, Patrick, Alain Méot, and Aurélia Bugaiska. 2018. Concreteness norms for 1,659 french words: Relationships with other psycholinguistic variables and word recognition times. *Behavior research methods*, 50:2366–2387.

Bradley, Margaret M. and Peter J. Lang. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings, Technical report C-1. Technical report, The Center for Research in Psychophysiology, Gainesville, FL.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Conneau, Alexis, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia, July 15th-20th. Association for Computational Linguistics.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2nd - 7th. Association for Computational Linguistics.

Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ethayarajh, Kawin and Dan Jurafsky. 2022. The authenticity gap in human evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6070, Abu Dhabi, United Arab Emirates, December 7th - 11th. Association for Computational Linguistics.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
Gray, Robert M. and David L. Neuhoff. 1998. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383.
Gregori, Lorenzo, Maria Montefinese, Daniele Paolo Radicioni, Andrea Amelio Ravelli, and Rossella Varvara. 2020. CONcreTEXT@EVALITA2020: The Concreteness in Context Task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online, December 17th.
Guasch, Marc, Pilar Ferré, and Isabel Fraga. 2016. Spanish norms for affective and lexico-semantic variables for 1,400 words. *Behavior Research Methods*, 48:1358–1369.
Guilford, Joy Paul. 1954. *Psychometric methods*. McGraw-Hill.
Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*, Virtual event, April 25th - 29th.
Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
Jiang, Albert Q., Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
Kiritchenko, Svetlana and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada, July 30th - August 4th. Association for Computational Linguistics.
Kumar, Varun, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In William M. Campbell, Alex Waibel, Dilek Hakkani-Tur, Timothy J. Hazen, Kevin Kilgour, Eunah Cho, Varun Kumar, and Hadrien Glaude, editors, *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China, December 4th - 7th. Association for Computational Linguistics.
LeCun, Yann, John Denker, and Sara Solla. 1989. Optimal brain damage. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann.
Lee, Janghwan, Minsoo Kim, Seungcheol Baek, Seok Hwang, Wonyong Sung, and Jungwook Choi. 2023. Enhancing computation efficiency in large language models through weight and activation quantization. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14726–14739, Singapore, December 6th - 10th. Association for Computational Linguistics.
Li, Zhuoyan, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore, December 6th - 10th. Association for Computational Linguistics.
Likert, Rensis. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22 140:55–55.
Liu, Peiyu, Zikang Liu, Ze-Feng Gao, Dawei Gao, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2024. Do emergent abilities exist in quantized large language models: An empirical study. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5174–5190, Torino, Italia, May 20th - 25th. ELRA and ICCL.

Louviere, Jordan J., Terry N. Flynn, and Anthony Alfred John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.
Louviere, Jordan J. and George G. Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. Technical report, Working paper.
Maslej, Nestor, Loredana Fattorini, Ray Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. 2024. Artificial intelligence index report 2024. *arXiv preprint arXiv:2405.19522*.
Miaschi, Alessio, Gabriele Sarti, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2022. Probing linguistic knowledge in italian neural language models across language varieties. *IJCoL - Italian Journal of Computational Linguistics*, 8(8-1).
Michel, Paul, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
Montefinese, Maria, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. The adaptation of the affective norms for english words (anew) for italian. *Behavior research methods*, 46:887–903.
Montefinese, Maria, Lorenzo Gregori, Andrea Amelio Ravelli, Rossella Varvara, and Daniele Paolo Radicioni. 2023. Concretext norms: Concreteness ratings for italian and english words in context. *Plos one*, 18(10):e0293031.
Ong, Desmond C. 2024. Gpt-ology, computational models, silicon sampling: How should we think about llms in cognitive science? *arXiv preprint arXiv:2406.09464*.
Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, USA, June 1st - 6th. Association for Computational Linguistics.
Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.
Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
Rambelli, Giulia and Marianna Bolognesi. 2023. Contextual variability depends on categorical specificity rather than conceptual concreteness: A distributional investigation on italian data. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 16–27, Nancy, France, June 20th - 20th.
Rambelli, Giulia and Marianna Bolognesi. 2024. The contextual variability of english nouns: The impact of categorical specificity beyond conceptual concreteness. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15854–15860, Torino, Italia, May 20th - 25th.
Ramesh, Krithika, Arnav Chavan, Shrey Pandit, and Sunayana Sitaram. 2023. A comparative study on the impact of model compression techniques on fairness in language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15762–15782, Toronto, Canada, July 9th - 14th. Association for Computational Linguistics.
Ravelli, Andrea Amelio, Marianna Marcella Bolognesi, and Tommaso Caselli. 2024. Specificity ratings for english data. *Cognitive Processing*.
Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.
Sahu, Gaurav, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In Bing Liu, Alexandros Papangelis, Stefan Ultes, Abhinav Rastogi, Yun-Nung Chen, Georgios Spithourakis, Elnaz Nouri, and Weiyan Shi, editors, *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland, May 27th. Association for Computational Linguistics.
Smith, Shaden, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022.

Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Soares, Ana Paula, Ana Santos Costa, João Machado, Montserrat Comesaña, and Helena Mendes Oliveira. 2017. The minho word pool: Norms for imageability, concreteness, and subjective frequency for 3,800 portuguese words. *Behavior Research Methods*, 49:1065–1081.

Solaiman, Irene, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Srebro, Nathan and Tommi Jaakkola. 2003. Weighted low-rank approximations. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 720–727, Washington, DC USA, August 21st - 24th.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Trott, Sean. 2024. Can large language models help augment english psycholinguistic datasets? *Behavior Research Methods*, 56(6):6082–6100.

Törnberg, Petter. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

van Hoef, Rens, Louise Connell, and Dermot Lynott. 2023. The effects of sensorimotor and linguistic information on the basic-level advantage. *Cognition*, 241:105606.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Wu, Xiaodong, Ran Duan, and Jianbing Ni. 2024. Unveiling security, privacy, and ethical concerns of chatgpt. *Journal of Information and Intelligence*, 2(2):102–115.

Yao, Zhao, Jia Wu, Yanyan Zhang, and Zhenhong Wang. 2017. Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 chinese words. *Behavior Research Methods*, 49(4):1374–1385, Aug.

Yao, Zhewei, Xiaoxia Wu, Cheng Li, Stephen Youn, and Yuxiong He. 2023. Zeroquant-v2: Exploring post-training quantization in llms from comprehensive study to low rank compensation. *arXiv preprint arXiv:2303.08302*.

Yoo, Kang Min, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic, November 7th - 11th. Association for Computational Linguistics.

Zamparelli, Roberto, Shammur Chowdhury, Dominique Brunato, Cristiano Chesi, Felice Dell'Orletta, Md. Arid Hasan, and Giulia Venturi. 2022. SemEval-2022 task 3: PreTENS-evaluating neural networks on presuppositional semantic knowledge. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 228–238, Seattle, United States, July 14th - 15th. Association for Computational Linguistics.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia, April 26th - 30th.

Zhao, Wei, Michael Strube, and Steffen Eger. 2023. DiscoScore: Evaluating text generation with BERT and discourse coherence. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia, May 2nd - 6th. Association for Computational Linguistics.

Zhu, Yiming, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.