Decoding Semantic Ambiguity in Large Language Models: Aligning Human Behavioral Responses with GPT-2's Internal Representations

Agustín Gianolini (0009-0003-7798-1134, agusgianolini@gmail.com)[1],
Belén Paez (0000-0001-5252-4840, bpaez2@gmail.com)[1],
Facundo Totaro (0009-0003-0475-6892 , facutotaro@gmail.com)[1],
Julieta Laurino (0000-0001-9132-2854, julilaurino@gmail.com)[2],
Fermín Travi (0009-0004-0833-3333, fermintravi@gmail.com)[1,3],
Diego Fernández Slezak (0000-0001-6348-1559, dfslezak@dc.uba.ar)[1,3],
Laura Kaczer (0000-0001-7969-9640, laurakaczer@gmail.com)[2],
Juan E. Kamienkowski (0000-0002-5725-6539, juank@dc.uba.ar)[1,3,4],
Bruno Bianchi (0000-0001-5252-4840, bbianchi@dc.uba.ar), and[1,3,4]

[1] Laboratorio de Inteligencia Artificial Aplicada, Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.
[2] Departamento de Fisiología, Biología Molecular y Celular, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.
[3] Instituto de Ciencias de la Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.
[4] Maestría en Explotación de Datos y Descubrimiento del Conocimiento, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.

Abstract. Large Language Models (LLMs), such as GPT-2, exhibit human-like text processing, yet their internal mechanisms for resolving semantic ambiguity remain opaque, similar to the "black-box" of human cognition. This study investigates how LLMs disambiguate concrete nouns by comparing their semantic biases to human behavioral responses. A corpus of sentences containing ambiguous words (e.g., "note") paired with biasing contexts (e.g., short paragraphs related to "music" and "education") was created. Human participants identified their perceived meanings of ambiguous words in these contexts, establishing a behavioral ground truth (i.e. human bias). The computer bias was measured via cosine distances between meanings' static embeddings and ambiguous words' contextualized embeddings. To improve the computer bias metric, two technical steps were implemented: (1) the model was fine-tuned to obtain a word-based tokenization, and (2) each ambiguous word's meaning was defined using word lists. Results revealed a non-linear dynamic in the GPT-2 computer bias and an additive effect of both improvements analyzed in the present work. Additionally, we found that the correlation between human bias and computer bias, measured layer-by-layer, topped at the middle layers. This result is in line with previous findings in human-model alignment research. This suggests shared computational principles between human cognition and LLM processing

for resolving ambiguity. The study advances interpretability research by linking model-internal representations to human behavioral benchmarks, offering insights into both artificial and biological language systems.

Keywords: LLMs, disambiguation, neurolinguistics

1 Introduction

Generative Artificial Intelligence (AI) language models, i.e. models that process and generate natural language, have become the focal point of technological advancements in recent years. These models have increasingly integrated into the daily lives of people worldwide, reshaping how we interact with technology. The rapid rise in their adoption can be attributed to the development of transformer-based architectures, which leverage the attention mechanism to efficiently process information. The remarkable performance of these models has sparked debates about their underlying mechanisms, particularly regarding their resemblance to the human brain's language processing (Abdou, 2022; Zhou et al., 2024). Do generative AI models process natural language in a way that mirrors the human brain?

Natural language is inherently complex, and for centuries, scientists have studied how it is processed in the brain, analyzing its various aspects. For the present study, we focus specifically on the process of semantic disambiguation. This process is crucial for language comprehension, as semantic ambiguity is a pervasive feature across all languages. In other words, many languages include words that share the same form but differ in meaning. Humans are generally skilled at disambiguating the meaning of a word within a given context, provided the context itself is not ambiguous. For instance, it is straightforward to infer that the word "bank" in the sentence "I went to the bank to withdraw money" refers to a financial institution, rather than "the side of a river, canal, etc. and the land near it"[1].

Studies investigating the processing of ambiguous words in humans – using diverse methodologies and behavioral measures – demonstrate that when such words are presented in contexts congruent with one of their meanings, they are processed more rapidly (Albrecht and O'Brien, 1993; Carter and Hoffman, 2024; de Groot, 1985; Hess et al., 1995; Laurino and Kaczer, 2024; Schustack et al., 1987; Schwanenflugel et al., 1988; Vu et al., 2000). These findings, alongside broader literature on lexical access, support the hypothesis that words and their meanings are encoded in the brain through distributed, multidimensional lexico-semantic representations (Rodd, 2020). Upon encountering an unambiguous word, its representation is activated, facilitating comprehension. However, when a word lacks a single unambiguous meaning, no stable lexico-semantic representation can be selectively engaged.

Building on this framework, Rodd proposes that ambiguous words initially activate an unstable intermediate representation—a blend of all possible mean-

[1] From Oxford Learner's Dictionaries: oxfordlearnersdictionaries.com

ings. This transient activation dynamically adjusts based on contextual cues, ultimately stabilizing into the contextually appropriate meaning. Nevertheless, critical questions remain unresolved regarding the precise neural and computational mechanisms enabling this disambiguation process.

In the case of AI models, the ability to accurately disambiguate the meaning of a word based on the preceding context is one of the key advancements introduced by the transformer architecture. Prior to the development of these models, the best architectures capable of effectively contextualizing words were recurrent neural networks (RNNs). These networks process one word at a time, maintaining an internal state that carries information from previous words. However, before the introduction of transformers, even the most advanced RNN-based models that topped benchmarks, such as AWD-LSTM, struggled with capturing long-term dependencies in text (Popov, 2018).

In contrast, transformer-based models do not process words one at a time but instead analyze the entire text as a whole (right-masked in the case of Causal Language Models). This architecture is built upon the attention mechanism, which adjusts the vector representations assigned to words (embeddings) by combining them with the vectors of neighboring words. The weights used for this combination —essentially, the weighted average of the vectors— are learned by the models during pretraining (Radford et al., 2018; Vaswani et al., 2017). Despite having a precise understanding of the mathematical operations underlying transformer-based models and full access to the learned weights and model activations in response to a stimulus, we still do not fully understand how the process of contextualization is performed. In other words, it is interesting to explore, just as is done in the field of neurolinguistics, the internal processes of semantic disambiguation in language models. In the present work, we will stimulate one of these models (GPT-2) with stimuli designed to deepen the analysis of this process.

The methodologies for studying semantic disambiguation mechanisms in humans and language models are fundamentally distinct. While language models allow full, explicit access to their internal activation states, human cognition permits only indirect measurement via behavioral responses, EEG, or fMRI. Nevertheless, these constraints raise a critical question: whether the mechanisms underlying semantic disambiguation are shared between these systems – despite their architectural and observational disparities – remains an open empirical frontier.

In recent years, advancements in Artificial Intelligence models, particularly in the field of Natural Language Processing (NLP), have driven the emergence of a new area of research that bridges cognitive science and artificial intelligence (namely CogniAI). This emerging field aims to understand the functioning of both the human brain and NLP models, which, despite their effectiveness, remain opaque, leaving us without a detailed understanding of the reasons behind their performance.

A pioneering study in this field mapped fMRI BOLD signals to semantic word embeddings during natural speech listening using voxel-wise alignment—a linear

regression of voxel activity onto word embedding dimensions (Huth et al., 2016). This work not only revealed organized semantic maps in the human brain but also established a framework for subsequent research linking linguistic processes to computational language models.

Building on this foundational work, numerous studies have expanded this approach to investigate diverse linguistic processes (Caucheteux and King, 2022; Caucheteux et al., 2023; Défossez et al., 2023; Reddy and Wehbe, 2021). To our knowledge, however, no prior work has systematically analyzed semantic ambiguity using methodologies bridging cognitive neuroscience and artificial intelligence. The present study addresses this gap by examining behavioral-level human semantic disambiguation (without neuroimaging), thereby establishing an initial framework for future research in this direction.

Beyond the previously mentioned challenges, such as understanding the internal mechanisms of the brain and language models, a current technical challenge lies in the representation of words. As noted earlier, we do not know precisely how our brain represents concepts: whether we store individual representations for every possible word or rely on a compositional representation system. For example, it is possible that we store the different morphemes that make up words separately and access the meaning of a concept based on the joint activation of these elements. Despite this debate, there is a consensus in the CogniAI field to conduct analyses at the word level—that is, to analyze brain activation in response to observing a complete word.

In contrast, we have a precise understanding of the representations used by language models. State-of-the-art models employ a process called tokenization, in which words are converted into tokens that do not necessarily carry inherent meaning. This happens because the tokenization process is learned statistically from the corpus used during the model's pretraining phase. The most widely used tokenizer today is known as Byte Pair Encoding (BPE). The BPE tokenizer is a sub-word tokenization method that iteratively merges the most frequent pairs of characters or sub-words in a text corpus to reduce vocabulary size and handle out-of-vocabulary words more effectively.

Subsequently, the model learns an embedding for each token, which becomes its representation in the pretrained model. Consequently, when analyzing a text with a language model to compare it with human processing, we do not have direct access to a vector representation of our words of interest but instead work with representations at the token level. In a previous study (Vaidya et al., 2023), researchers modified the tokenizer of a small model, enabling them to analyze activations at the word level.

In this work, we investigate the alignment between human cognition and a language model in semantic disambiguation. On one hand, we leverage behavioral measurements of human bias in a semantic disambiguation task. On the other, we develop a computationally analogous metric for language models based on cosine similarity. Our primary objective is to identify parallels between the internal processing of the tested language model and human behavioral responses. To achieve this, we optimized the model architecture (by revising its tokeniza-

tion scheme) and refined the computational bias metric. Our results reveal that intermediate layers of the model exhibited the strongest alignment with human responses, suggesting shared mechanisms in semantic resolution.

2 Methods

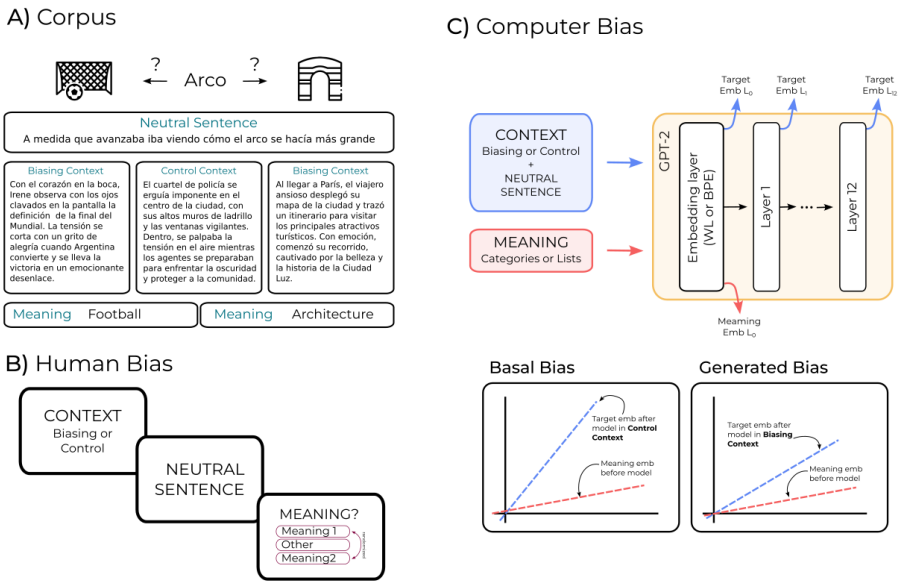

Fig. 1: **Methodological overview of the study. (A) Corpus:** Example of a stimulus from the corpus using the ambiguous Spanish word "arco" ("goal" vs. "arch"), including a neutral sentence and three contexts: two biasing contexts related to distinct meanings ("football" and "architecture"), and one control context. **(B) Behavioral procedure:** in an web platform, participants read a neutral sentence preceded by a context (biasing or control) and choose the perceived meaning of the ambiguous word among three options. **(C) Computational bias estimation:** GPT-2 (with the original BPE tokenizer or the fine-tuned word-level one) receives the combined input (context + neutral sentence), and target word embeddings are extracted from each layer. These embeddings are compared with meaning embeddings (defined using one word or via extension) extracted from the Embedding layer (Layer 0) using cosine distance. Basal Bias is computed using the control context, and Generated Bias using the biasing contexts.

2.1 Stimuli

Neutral sentences We compiled a selection of 48 ambiguous words in Spanish, all of which are concrete nouns with at least two distinct meanings (e.g., the word "palm" can refer to both "the palm of a hand" and "a type of tree"). These target words were sourced from the "Small World of Words" database and were carefully selected by Dr. Laura Kaczer and Lic. Julieta Laurino for a word association experiment (Laurino and Kaczer, 2024). The manual selection process was specifically tailored for the purposes of the referenced experiment, taking into account multiple linguistic and associative variables derived from the database responses. This meticulous curation underscores that the stimuli were not chosen automatically or at random, but rather with experimental constraints in mind. Importantly, this is not a purely computational study: the design includes experiments involving human participants, which imposes practical limits on the number of stimuli that can be feasibly used.

For each target word, a neutral sentence was constructed (Figure 1A), i.e., a sentence in which the meaning of the target word cannot be disambiguated. For example, "Enzo movió la palma muy rápido para evitar que hiciera ruido" ("Enzo moved the palm very quickly to prevent it from making noise").

Biasing contexts To bias the interpretation of the ambiguous words, two short paragraphs were created for each target word. These paragraphs were designed without including either the target word itself or any words directly associated with the intended meaning. For example, for the target word "palm" contexts were: (1) about the hand for the meaning "body", and (2) about Caribbean beaches for the meaning "vegetable".

A third context was constructed for each target word. This context was designed to reflect a meaning unrelated to the target word (i.e. neutral context). For example, for the target word "palm", the neutral context is related to "politics," a domain completely detached from any of the word's meanings.

Control sentences and contexts To include a control condition, we selected 18 unambiguous words (e.g., "robot"). For these words, we designed a sentence (similar to the neutral sentence of the ambiguous words) and one paragraph associated with their single unambiguous meaning (e.g., "family"). This control set provides a baseline for comparison with the polysemous stimuli. Also, this will be used as catch trials in the behavioral experiments. That is, will serve as correctness control for the subjects.

Word meaning Up to now, the meanings of ambiguous words presented in this work were based on categories. For example, for the word "palm", the meanings presented were "body" and "plant". These meanings are useful for measuring biases in humans, who can likely understand which interpretation each of these categories refers to. However, in language models, this way of defining meanings may not be the most appropriate.

In the present study, we explore the possibility of defining each meaning of ambiguous words using a list of associated words. This approach is called *definition by extension*. For example, the list of words *leaf, root, soil, flowers, trunk, woods* can be used to express the meaning of the word "tree".

For each meaning of the target words, we selected a list of 8 words that we considered could describe the target word in the context of that meaning. Since this selection was handpicked, validation was necessary. For this purpose, an online experiment was conducted. In each trial of this experiment, participants were shown a target word in a non-neutral context along with the list of 8 words related to that context. They were instructed to select the 3 words they believed were most strongly related to the target word.

2.2 Bias metrics

The main analysis of the present study is based on measuring the change in the semantic bias given each of the previously introduced texts. That is, for each target word, measure how the representation or interpretation of its meaning changes when presented after a neutral context against when it is presented after a biasing context. Based on this, we defined 2 metrics: *Basal Bias* and *Generated Bias*. The former one relates to the bias (model and human) towards a meaning given that no clue was presented to disambiguate that meaning. The latter one refers to the bias to a meaning when presenting the word in a context that relates to that particular meaning. For each target word, we will analyze the base and the generated bias for each meaning.

Human bias | Behavioral experiment: To address the human biases, we conducted an online behavioral experiment (Figure 1B). This experiment was performed on the pcibex platform[2]. After entering the platform, participants, recruited via social media, were instructed to carefully read the texts that were going to be presented and answer some comprehension multiple-choice questions. The experiment was semi-self-paced, with forced pauses planned to avoid users skipping texts before reading them. The procedure was as follows:

1. A context (neutral o biasing) was presented,
2. After 10 seconds a "continue" button appear at the bottom of the context,
3. The context disappear and the neutral sentences appears,
4. After 5 seconds a comprehension question and 3 options (Meaning 1, Meaning 2, and Other) were presented. The option "Other" was always presented in the middle, and the 2 meanings were randomly shuffled between top and bottom positions,
5. After completing 15 trials, participants were invited to continue with a new block or to exit the website.

Results of this experiment were used to measure the human Base and Generated Biases:

[2] farm.pcibex.net

- **Human Basal Bias:** For a given word-meaning, the human Basal Bias is defined as the proportion of participants that choose the given meaning after reading the neutral sentence contextualized by the neutral context,
- **Human Generated Bias:** For a given word-meaning, the human Generated Bias is defined as the proportion of participants that choose the given meaning after reading the neutral sentence contextualized by the corresponding biasing context.

Computational bias | GPT-2 experiment: To address the computer biases, we conducted an experiment inputting the previously presented texts to a Language Model (Figure 1C). The procedure was as follows:

1. A context-sentence pair was input to a model (see Model section below),
2. The embedding of the target word was extracted from each layer of the model,
3. In the case of multi-token words, all the embeddings were extracted and then averaged together.

Additionally, a meaning embedding was obtained from the embedding layer of the model. For meanings defined as lists of words, the meaning embedding was calculated by averaging the embeddings of all five candidate words, weighted according to the proportion of participants who selected each word in the validation experiment described above.

Results of this procedure were used to measure the Computer Base and Generated biases:

- **Computer Basal Bias:** For a given word-meaning, the computer Basal Bias is defined as the cosine distance between the meaning embedding and the target-word embedding when contextualized with the neutral context
- **Computer Generated Bias:** For a given word-meaning, the computer Generated Bias is defined as the cosine distance between the meaning embedding and the target-word embedding when contextualized with the corresponding biasing context

2.3 Models

For this research, we utilized an open-source version of the GPT-2 model available on HuggingFace[3]. This specific model was pre-trained exclusively on a Spanish language corpus, providing us with a monolingual model. We chose a monolingual architecture to ensure that the embeddings would not be influenced by cross-linguistic interference, as can occur in multilingual models. Additionally, we deliberately selected a lightweight version of GPT-2 (12 layers, 512 embedding dimensions) that could be fine-tuned locally using the hardware available in our laboratory (an NVIDIA Titan RTX with 24GB of VRAM). This setup

[3] https://huggingface.co/DeepESP/GPT-2-spanish

enabled us to retrain the model and modify its tokenizer as needed, which was essential for the objectives of our study. Initially, this approach also allowed us to minimize the previously mentioned challenge of utilizing sub-word tokens when making comparisons with human language processing. In future work, we aim to extend this study by incorporating more modern language models.

As mentioned, despite the advantages of using a monolingual model, we found that several words of interest in the present study are represented by multiple tokens. For this reason, we modified the model's vocabulary. To accomplish this, we followed the methodology described by Vaidya et al. (Vaidya et al., 2023), which consisted of:

1. Obtaining a new text corpus exclusively in Spanish; for this purpose, the Large Spanish Corpus[4] dataset was used.
2. Tokenizing the selected corpus at the word level (i.e., separating by spaces), thus obtaining a new vocabulary. This process involved vocabulary generation, parsing, cleaning, and determining the final word count. To reduce vocabulary size, words containing question marks (¿?) and exclamation marks (¡!) were filtered out. Uppercase letters were also removed.
3. Obtaining static embeddings for each word in the new vocabulary. For this, pre-trained embeddings from the first layer of GPT-2 were used. For words composed of multiple tokens, the embedding was obtained by averaging all the vectors that constitute them.
4. Replacing the first layer of static embeddings in the original model with a new layer containing the obtained vectors. Since the new vocabulary size differed from the original size, it was necessary to resize this layer before inserting the new embeddings.
5. Retraining the complete model with the new corpus on the text generation task. Similar to GPT-2 training, at each step t, the model predicts the next word. Cross-entropy was then used as the loss function to backpropagate all weights, including those of the new embedding layer. Due to the large size of the dataset and to avoid lengthy training times, we decided to use 15% of the dataset for training. Of this fraction, 90% was used for training and 10% for validation. Based on previous work (Vaidya et al., 2023), the model was trained for 10 epochs.

3 Results

3.1 Tokenizer change

First, to analyze the outcome of the GPT-2 model tokenizer change, we conducted a subjective analysis of its performance. We selected incomplete text fragments from different sources and input them into both the original model and the word-tokenized model. The models were then used to complete these fragments based on their predictions. Since the models return probabilities for

[4] https://huggingface.co/datasets/josecannete/large_spanish_corpus

the next word, we randomly sampled from words accumulating 95% of the probability for generation (top-p sampling). This approach prevented the models from entering cycles where they repeatedly generate the same word and allowed for more varied token generation.

Table A.1 shows three examples of texts generated by both models. It is important to note two relevant points: (1) unlike state-of-the-art LLMs, the GPT-2 model is a base Language Model. That is, it is not trained to function in a chat format; (2) as a 2019 model, its performance falls considerably short of the capabilities observed in current state-of-the-art models. In the first input, the Fine-Tuned Model demonstrates an ability to complete text similar (or even better) than the Base Model. In this example, the Base Model loses the original thread and fails to form sentences with internal coherence. However, the Fine-Tuned Model maintains considerable coherence, forming grammatically correct sentences. This indicates that the retraining did not substantially impair the model's functionality.

Conversely, examining the second and third examples in Table A.1, it becomes evident that there are cases where the model performs worse (significantly worse in Example 3). A notable observation in these examples is the repeated appearance of the ⟨unk⟩ token in the Fine-Tuned Model's predictions, which is used when the model processes words not found in its vocabulary. This is expected, as this model, with its vocabulary consisting exclusively of complete words, cannot generate new words. In contrast, the Base Model's vocabulary includes sub-word tokens, which can be concatenated to form new words. Thus, when the input contains words unrecognized by the Fine-Tuned Model, the ⟨unk⟩ token leads to poor predictions. This does not pose a significant problem for the remainder of our work because: (1) we do not aim to generate text; and (2) our corpus does not contain a large number of words unrecognized by the Retrained Model .

3.2 Meaning Definition by Extension

To define each meaning of ambiguous words as precisely as possible, we constructed lists that define meaning by extension. That is, for each meaning of each ambiguous word, we created lists of 8 words related to it. However, these lists were compiled manually according to our judgment. To validate these lists, we conducted an online experiment in which participants selected 3 out of the 8 words as the closest to the corresponding meaning. With 512 participants and an average of 95.54 responses for each word-meaning pair, we determined that the lists to be used from this point forward would contain 5 words each. This decision was based on the observation that for certain word-meaning pairs, words ranked 6th through 8th received very few selections from participants (Figure A.2). In other words, these words, which were arbitrarily chosen by the researchers, did not reflect the general population's understanding of the particular meaning of that word.

This experiment not only allowed us to validate and filter the word lists but also enabled us to weight each word in the lists based on the number of times it was selected by participants. Thus, when obtaining the embedding for a

meaning, it could be calculated as a weighted average of the embeddings of each word.

3.3 Measurement of the Biases

Human Bias: The results of the online behavioral experiment demonstrate that, in general terms, the generated corpus functions correctly. That is, when comparing the Basal Bias with the Generated Bias for each word-meaning pair, it is evident that in the vast majority of cases, the Generated Bias exceeds the Basal Bias (Figure 2A). This confirms that the biasing contexts successfully induce the interpretation of the desired meaning for the ambiguous words. It is worth noting that, despite the general trend, there are cases where the generated bias is very close to the baseline bias (being slightly higher or lower), although there is no case where the Generated Bias is substantially lower than the Basal Bias. This demonstrates that within our corpus of ambiguous words, some words (and particularly certain meanings of these words) proved difficult to bias.

Computer Bias: Conversely, when analyzing the biases in the final layer of the original GPT-2 model, it shows a much more subtle effect. In this analysis, the distance between the target word and the categorical meaning, a bias metric, was used. In this case, there are many more instances where the Generated Bias has a value similar to the Basal Bias (points closer to the diagonal).

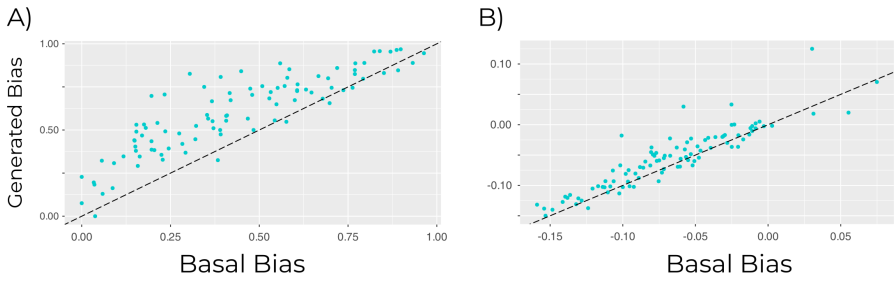

Fig. 2: **Human and Computer Bias:** Basal and Generated Bias for (A) Humans and (B) Computer. Human biases were obtained from a behavioral experiment. Computer biases were obtained from a cosine similarity metric. Dashed line indicates the identity function.

Based on these preliminary, slightly positive results, we decided to optimize the bias measurement through three approaches: (1) layer-by-layer analysis of the model; (2) definition of meanings through word lists; and (3) retraining the model to incorporate a word-level tokenizer. The objectives of these new analyses are varied. With (1), we aim to deepen our understanding of the dynamics models follow when biasing embeddings; with (2), we seek to understand the best way to

measure embedding bias; and with (3), we want to analyze whether word-level embeddings are superior to sub-word token embeddings.

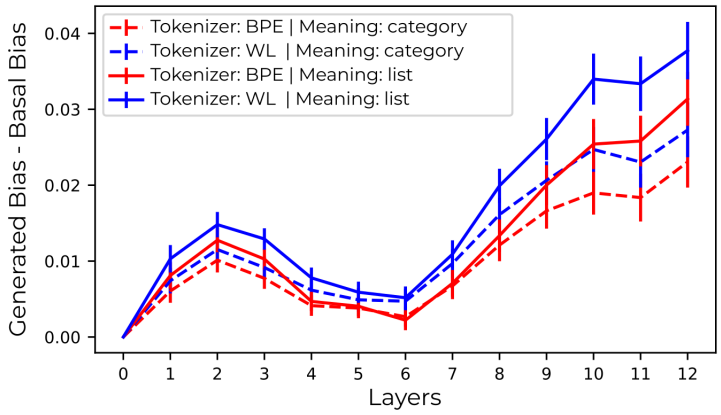

Fig. 3: **Optimization analysis:** difference between generated and basal bias for each layer of the GPT-2 tested models, varying tokenizer (BPE: byte-pair-encoding, WL: word level) and meaning definition (category: 1 word defining the semantic category, list: 5-word list curated by humans)

The results of these new analyses reveal significant changes in the outcomes (Figure 3). First, the layer-by-layer analysis demonstrates a non-linear dynamic in the bias induced in vector representations of words. The bias is not induced uniformly across layers. Instead, the difference between Generated Bias and Basal Bias exhibits an initial increase, followed by a decrease that reaches its minimum at layer 6, after which it grows monotonically until the final processing stage. Secondly, defining meanings through word lists (continuous lines) enabled a more robust measurement of bias, yielding higher values, particularly in the model's final layers. This represents a crucial improvement, as the original metric used in our preliminary analysis (red-dashed line) proved to be the least sensitive measure overall. Finally, implementing a model with a word-level tokenizer (blue lines) enhances the bias difference throughout the entire model. Notably, the effects of changing the tokenizer and redefining meaning measurement are cumulative. Consequently, applying both modifications simultaneously (blue-continuous line) produces the most pronounced bias measurement.

3.4 Human-Computer similarity

Finally, to assess the similarity between human and model biases, we analyzed the Pearson correlation between these metrics. Specifically, we compared the bias difference (Generated Bias - Basal Bias) observed in the human behavioral experiment with the bias computed from the fine-tuned model, using word

meanings defined by extension (blue continuous line in Figure 3). The results of this analysis indicate the highest similarity between the two metrics at layer 5 of the analyzed GPT-2 model (Figure 4). This finding is consistent with previous studies reporting that similarities between human cognition and language models tend to emerge in the middle layers across different types of analyses (Caucheteux and King, 2022; Cheng and Antonello, 2024; Zhou et al., 2024).

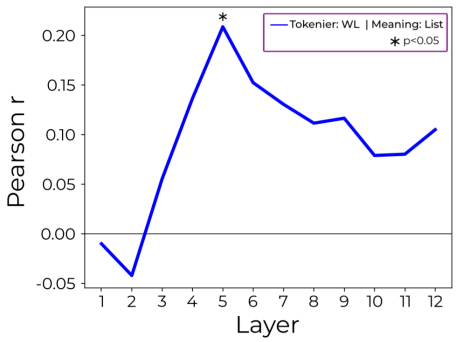

Fig. 4: **Layer-by-Layer correlation with human bias:** Person r statistic between human and computer biases difference, for the best-performing model (Tokenizer: WL — Meaning: list).

4 Discussion

In the present study, we analyzed the semantic disambiguation of nouns in Spanish by both humans and language models. Specifically, within the broader line of research in which this work is framed, we are interested in examining the similarities between the processes that allow humans and language models to disambiguate the meanings of such words.

Current hypotheses regarding the neural mechanisms underlying this process resemble our understanding of the internal workings of language models. In neuroscience, theories have been proposed involving distributed representations that resemble the word embeddings used in language models to encode meaning. These distributed representations are believed to be differentially activated depending on the context in which an ambiguous word appears, and this differential activation is thought to allow the interpretation of its meaning (Rodd, 2020). This is, once again, reminiscent of the contextualization mechanism in language models, where attention layers within the transformer architecture dynamically modulate word embeddings based on context (Vaswani et al., 2017).

For our analysis, we selected 48 ambiguous Spanish nouns, each with at least two concrete noun senses. For each target word, we designed a neutral sentence (in which the meaning cannot be unambiguously resolved), two biasing contexts

(each promoting one of the two target meanings), and one additional neutral context. Using these stimuli, we conducted a behavioral experiment to measure human bias, as well as a computational experiment to assess bias in a pre-trained GPT-2 model in Spanish.

Preliminary results showed that, while the stimuli successfully induced measurable bias in human participants, the model did not display a comparable pattern. This led us to refine our methodology to improve measurement sensitivity. Two key modifications were introduced: (1) the definitions of the target word meanings were revised, shifting from categorical definitions to extension-based definitions (i.e., sets of words validated by human annotators); and (2) the pre-trained model was adapted to use a word-level tokenizer. In addition, the analysis was conducted across all layers of the model.

These updated analyses revealed significant improvements in bias measurement, particularly when both modifications were applied simultaneously. This highlights the importance of a robust definition of meaning embeddings. Initially, the meaning embedding for each sense was based on a single word representing its semantic category. In contrast, the revised approach used the average of multiple embeddings, each corresponding to a different word within a semantically validated set, resulting in a more precise representation of the intended meaning.

Moreover, using a model with a word-level tokenizer improved the quality of word representations. In the original model with a byte-pair encoding (BPE) tokenizer, many target words were split into multiple tokens, requiring additional processing (e.g., averaging) to represent a single word. This introduced additional sources of noise in the measurements. With word-level tokenization, we were able to avoid this issue and perform measurements directly at the word level.

Finally, the layer-by-layer analysis revealed a non-linear dynamic in bias induction. That is, bias increased in the early layers, decreased around layer 6, and then increased again in later layers. This suggests that the operations performed by successive attention layers are not monotonic or unidirectional. On the contrary, the function of each layer may vary throughout the network.

This final result also informs our concluding analysis, where we compared human and computational biases using Pearson correlation. This comparison revealed a finding consistent with prior literature: the highest similarity between human and model biases occurs in the intermediate layers of the model. While the specific pattern of this similarity may vary across studies—depending on the comparison techniques and the nature of the data used—our findings align with those reported in the field. In our case, we compared behavioral data with internal model activations. In contrast, other studies compare model activations with brain activity measured via functional MRI, EEG, or MEG (Caucheteux and King, 2022; Deniz et al., 2023; Huth et al., 2016). These types of neural data also enable other forms of similarity analyses, such as alignment, representational similarity analysis (RSA) (Abnar et al., 2019), and centered kernel alignment (CKA) (Cheng and Antonello, 2024), among others. Notably, previous studies have employed a range of language models, and all tend to report a similar pattern of peak similarity in intermediate layers. However, there is still a lack

of studies involving more modern architectures, likely due to the substantial computational cost associated with accessing internal activations locally. This is a necessary step for such analyses and presents a technical limitation. In future work, we aim to address this challenge and expand the investigation to include more recent language models.

Nevertheless, despite the differences in methodology, our behavioral-based analysis of model-human similarity converges with previous findings in the field, reinforcing the relevance of using behavioral evidence to explore parallels between neural and computational representations.

References

Abdou, M. (2022). Connecting neural response measurements & computational models of language: A non-comprehensive guide. https://arxiv.org/abs/2203.05300

Abnar, S., Beinborn, L., Choenni, R., & Zuidema, W. (2019). Blackbox meets blackbox: Representational similarity and stability analysis of neural language models and brains. *arXiv preprint arXiv:1906.01539*.

Albrecht, J. E., & O'Brien, E. J. (1993). Updating a mental model: Maintaining both local and global coherence. *Journal of experimental psychology: Learning, memory, and cognition*, *19*(5), 1061.

Carter, G.-A., & Hoffman, P. (2024). Discourse coherence modulates use of predictive processing during sentence comprehension. *Cognition*, *242*, 105637.

Caucheteux, C., Gramfort, A., & King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, *7*(3), 430–441.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications biology*, *5*(1), 134.

Cheng, E., & Antonello, R. J. (2024). Evidence from fmri supports a two-phase abstraction process in language models. *arXiv preprint arXiv:2409.05771*.

Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., & King, J.-R. (2023). Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, *5*(10), 1097–1107.

de Groot, A. M. (1985). Word-context effects in word naming and lexical decision. *The Quarterly Journal of Experimental Psychology Section A*, *37*(2), 281–297.

Deniz, F., Tseng, C., Wehbe, L., la Tour, T. D., & Gallant, J. L. (2023). Semantic representations during language comprehension are affected by context. *Journal of Neuroscience*, *43*(17), 3144–3158.

Hess, D. J., Foss, D. J., & Carroll, P. (1995). Effects of global and local context on lexical processing during language comprehension. *Journal of Experimental Psychology: General*, *124*(1), 62.

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Laurino, J., & Kaczer, L. (2024). Pupil size tracks the effects of global context and semantic ambiguity on word-meaning processing. *psyarxiv*. https://osf.io/preprints/psyarxiv/7dt8j_v1

Popov, A. (2018). Neural network models for word sense disambiguation: An overview. *Cybernetics and information technologies*, *18*(1), 139–151.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. .

Reddy, A. J., & Wehbe, L. (2021). Can fmri reveal the representation of syntactic structure in the brain? *Advances in neural information processing systems*, *34*, 9843–9856.

Rodd, J. M. (2020). Settling into semantic space: An ambiguity-focused account of word-meaning access. *Perspectives on Psychological Science*, *15*(2), 411–427.

Schustack, M. W., Ehrlich, S. F., & Rayner, K. (1987). Local and global sources of contextual facilitation in reading. *Journal of Memory and language*, *26*(3), 322–340.

Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of memory and language*, *27*(5), 499–520.

Vaidya, A. R., Turek, J., & Huth, A. G. (2023). Humans and language models diverge when predicting repeating text. *arXiv preprint arXiv:2310.06408*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vu, H., Kellas, G., Metcalf, K., & Herman, R. (2000). The influence of global discourse on lexical ambiguity resolution. *Memory & cognition*, *28*, 236–252.

Zhou, Y., Liu, E., Neubig, G., Tarr, M., & Wehbe, L. (2024). Divergences between language models and human brains. *Advances in Neural Information Processing Systems*, *37*, 137999–138031.

Supplementary Material

Table A.1: Examples of completitions using the Base Model and the Fine-tuned model (word-level tokenizer). Three examples are shown where (1) both models behave correctly, (2) Fine-tuned model is better, and (3) Base Model is better

Input	Base Model	Fine-Tuned Model
A diferencia de otros países donde los créditos hipotecarios son una herramienta fundamental para facilitar el acceso a la vivienda	y que cualquier otro país se ocupa de protegernos de las enfermedades y enfermedades que afectan a un mismo. Los ciudadanos a menudo no se ocupan de los medios de transporte de la vivienda ni de la comida; ni de la comida ni de los alimentos pueden permitirse vivir en un mismo lugar. Desde luego, yo mismo me habría sentido beneficiado si hubiera tratado de ser un simple ciudadano de una ciudad a otro. En mi opinión, las circunstancias son semejantes a las de un país que se está muriendo	o a las personas mediante el uso de los seguros y los de su uso a los mercados en general o a las empresas en general. Como hemos visto en capítulos anteriores, el derecho de propiedad es un proceso económico basado en el empleo y la producción de bienes propios y no en la compra de bienes propios mediante la compra del activos. De este modo la posesión de bienes de los países del mundo se da como un medio de garantizar los acceso y acceso al país a través de los pagos de los individuos mediante
El presidente del Gobierno de España Pedro Sánchez aseguró este lunes que tras un periodo de reflexión sobre su futuro	En lo que respecta a la prensa, el presidente del Consejo de Seguridad Nacional Juan Carlos Menem, cuando habló sobre su inminente futuro, dijo: “No tiene más remedio que esperar”. El almirante Juan Carlos, que había pasado la entrevista con su secretario general, manifestó que, pese a la crisis reinante, la historia no lo impresionó ni conmocionó a las decenas de miles de personas que se habían reunido en el país y los que le acompañaban en la reunión, ya que, en realidad, la mayoría de	profesional, habría realizado los trámites necesarios para el tratamiento del tratamiento del tratamiento de la enfermedad mental a la ⟨unk⟩ lo siguiente que hizo es asegurar que no habrá ni un ápice de modificación en su decisión de permitir que el tratamiento de la enfermedad mental pueda ser usado con eficacia en la consulta y tratamiento del tratamiento en la vida de la enferma y de su futuro profesional y económico en la enfermedad mental de la misma manera y por ⟨unk⟩ y lo siguiente que hizo la misma señora del ⟨unk⟩ a la mujer que le rece ⟨unk⟩
El Deportivo Feniz sigue a paso firme en el Torneo de Primera División Femenino que organiza y fiscaliza la prestigiosa Liga Municipal	[2], y de nuevo en la Federación de Fútbol se establece una reglamentación de cada uno de los elementos de la Federación. Las partidas de Liga de Liga son muy importantes y las de Liga de Liga de Liga de Liga del Mundo y Liga de Liga de Liga. Las primeras dos partidas han de ser muy importantes para las ligas locales de las Naciones Unidas. Son muchas las actividades que han realizado desde los tiempos de la Federación. La Federación de Fútbol y Liga del Mundo ha desarrollado	de ⟨unk⟩ ⟨unk⟩ en el que se celebran los ⟨unk⟩ el ⟨unk⟩ con más de tres mil ⟨unk⟩ ⟨unk⟩ y se celebran las ⟨unk⟩ la ⟨unk⟩ ⟨unk⟩ y la ⟨unk⟩ no se ⟨unk⟩ ⟨unk⟩ y no hay ⟨unk⟩ ⟨unk⟩ si hay ⟨unk⟩ no hay ⟨unk⟩ ⟨unk⟩ ⟨unk⟩ y en todas las ⟨unk⟩ el ⟨unk⟩ no hay ⟨unk⟩ ⟨unk⟩ no hay ⟨unk⟩ no hay ⟨unk⟩ que exista tanto ⟨unk⟩ ⟨unk⟩ ⟨unk⟩ ⟨unk⟩ ni ⟨unk⟩ ni ⟨unk⟩ ⟨unk⟩ se ⟨unk⟩ no hay ⟨unk⟩ ⟨unk⟩ no hay ⟨unk⟩ que existan para ⟨unk⟩ ⟨unk⟩ ⟨unk⟩ aunque exista por ⟨unk⟩ y que se haga ⟨unk⟩

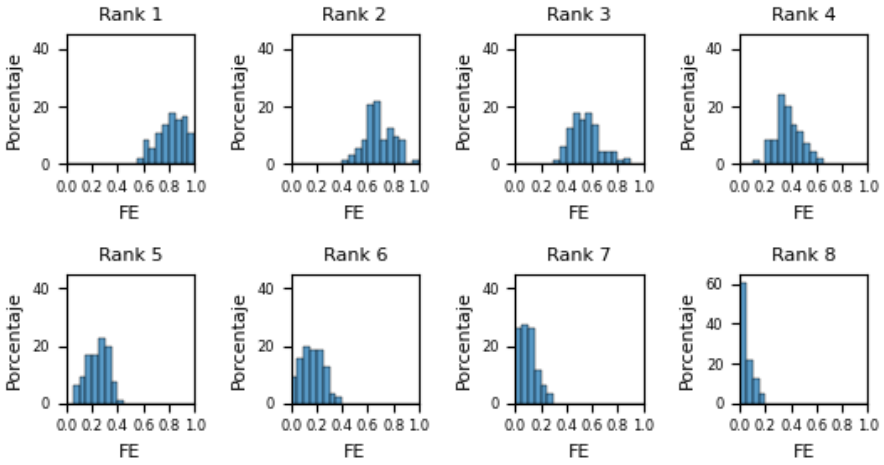

Fig. A.2: **Frequency of Election (FE) Distribution:** each histogram represents the distribution of frequency of election of the 8 proposed items for each meaning-word pair. The frequency of election is the proportion of participants that choose a given item. Since each participant choose 3 words, frequencies of a single meaning-word pair sums up to 3.