Universiteit Leiden
The Netherlands

Bachelor Data Science and Artificial Intelligence

Explaining semantic priming effects using predictors derived from Large Language Models

Qiu van Leeuwen

Supervisors:
Tom Heyman & Evert van Nieuwenburg

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl 01/07/2025

Abstract

Humans are quicker to process a word (like “cat”) if they have previously seen a word that is related (like “dog”). This decrease in reaction time compared to unrelated words is known as semantic priming. Strongly related words result in stronger semantic priming effects, whereas weakly related words result in weaker semantic priming effects. The current thesis explored whether modern Large Language Models (LLMs) can predict semantic priming in human lexical processing. Using the already existing English subset of the Semantic Priming Across Many Languages (SPAML) dataset of Buchanan et al., the similarity scores were generated with OpenAI’s gpt-4o-mini model and its embeddings with text-embedding-3-small [BCC$^+$25]. These LLM scores assessed word pair relatedness based on either common association (e.g., doctor-nurse) or shared semantic features (e.g., deer-pony). To assess predictive power, three multiple regression models were used to take multicollinearity into account, including one model combining all six predictors, and two separate models with both three similar predictors (Difference Score and Only Related Score). These models revealed that LLM-derived predictors explained a significant portion of the human priming effect. However, a large amount of variance remained unexplained, suggesting that while LLMs can mimic certain human linguistic behaviors, their current representation of word meaning does not fully align with the complex cognitive processes underlying semantic priming. It may also be that the information reflecting this alignment is present, but has not been fully captured by our current predictive measures.

Contents

1 Introduction

1.1 Semantic Priming

A fundamental characteristic of understanding human language is the influence of the prior context on subsequent word processing. This is evident when exposure to one word unconsciously prepares us to process a related word more efficiently. The psychological finding, where the meaning of one stimulus influences the processing of another, is known as semantic priming. It is the decrease in reaction time for target words that are semantically related to a previously presented stimulus (i.e., cue), compared to an unrelated stimulus [CGA+23]. People would react faster to the word "cat" if it is presented shortly after the related word "dog" compared to an unrelated word, for example "bus". The semantic priming effect says something about semantic memory or the structure of people's mental lexicon, which can be seen as the dictionary of the human brain. Related concepts have a stronger or closer connection with each other in our mental lexicon, whereas unrelated concepts have weaker connections.

To understand the nature of these semantic relationships that give rise to priming, researchers also focus on human performance in psycholinguistic tasks with computational approaches. The paper of Mandera et al. discussed two types of distributional semantic models (DSMs), known as count and predict models [MKB17]. These models are based on the idea that words with similar meaning are used in a similar context. Count models, such as Latent Semantic Analysis (LSA), build word representations based on counting word co-occurrences in text, whereas prediction models, such as Word2Vec, use neural networks to predict words in a given context, learning vector representations (embeddings) that encode semantic meaning [MKB17]. The way these semantic relationships are encoded and expressed can vary significantly across languages, influenced by differences in script, syllables, morphology, semantics, and broader cultural contexts [BCC+25]. The (associative) link between words is often shaped by cultural norms or daily experiences. For instance, words that seem closely related in the Dutch language and culture, such as "cheese" and "snack" might share a strong association, because cheese is commonly consumed as a snack or appetizer in the Netherlands. However, this particular association may not be as obvious in other languages or cultures where cheese is not a typical snack. Such cross-linguistic variations highlight the complexity of studying semantic representation.

1.1.1 Activation Theory

The activation theory of semantic processing proposed by Quillian (1968) is an influential theory of how semantic priming developed [CL75]. His model conceptualized semantic memory as an interconnected network of nodes, where each node represents a concept or word. The activation spreading from two or more concepts continued until an intersection was found. The intersection is the point where different concepts in the semantic network converge, providing the basis for establishing a relationship between two or more concepts. The interconnected nodes are linked by associative pathways, where the strength of these pathways reflects the degree of relatedness between the concepts. A mechanism named "attention mechanism" allowed a model to dynamically weigh the importance of different words when processing information or generating an output. The weights are expressed in the length of the pathways between two or more nodes in a network. The shorter a line, the greater the relatedness. A node can have multiple connections, enabling them to be both a cue and a target concept. When a cue (e.g., "red") is encountered, its corresponding node

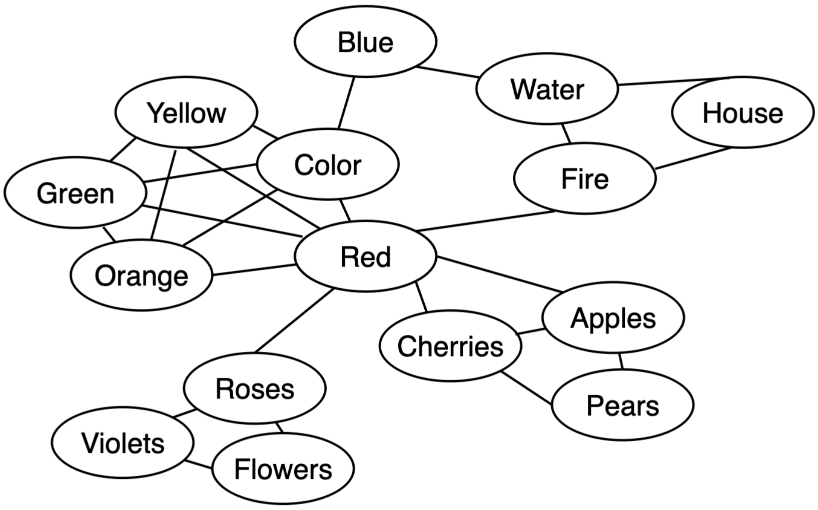

Figure 1: A schematic representation of concept relatedness in a stereotypical fragment of human memory adapted from [CL75]. A shorter line represents greater relatedness.

in the semantic network becomes activated. This activation can spread to other connected nodes using the associative pathways. Nodes representing closely related concepts (e.g., “color”, “fire”, “cherries”) receive more activation than unrelated concepts. For example in Figure 1, the line from red-cherries is shorter and therefore, stronger than red-fire, because fire can also be other colors.

If the presented target word (e.g., “cherries”) corresponds to a node that has already received some pre-activation from the cue, its processing threshold is reached more quickly. This means that there is less additional input or processing time needed to identify the target. The resulting shorter reaction time for related words forms the priming effect. Stronger associations lead to more activation spread and thus stronger priming.

1.2 Large Language Models (LLMs)

The recent introduction of Large Language Models (LLMs) is a game changer in the field of artificial intelligence (AI). LLMs are generative mathematical models of the statistical distribution of tokens of human-generated text. Such models are trained on large text datasets, enabling them to learn and understand complex patterns in human languages [Sha24].

Modern LLMs can be seen as sophisticated evolutions of the earlier mentioned predict models in distributional semantics discussed by Mandera et al. [MKB17]. A key component in many modern LLMs is already discussed in subsection 1.1.1: the attention mechanism [CL75]. By weighing the importance of different words in the input when processing information, an LLM is able to capture long-range dependencies and more context-sensitive meanings generating the best possible response. LLMs are capable of learning linguistic aspects, such as grammar, showing noticeable abilities in capturing semantic relationships. This is done by the core engine of LLMs, also known as the neural networks or artificial neural networks (ANNs). The foundational concept of ANNs took inspiration from the structure and function of biological neural networks in the human brain. Early models, such as the McCulloch-Pitts neuron (Shown in Figure 2), provided a formal representation of an artificial neuron [MP43]. Their research questioned how the human brain could produce complex patterns through connected brain cells, or neurons. One of their main findings was the comparison of neurons with a binary threshold to Boolean logic (logical operations e.g., True, False, AND, OR). The McCulloch-Pitts neuron consisted of interconnected processing units (analogous to neurons) that transformed inputs through weighted connections (analogous to synapses) and activation functions, which determined whether a unit was activated. McCulloch and Pitts (1968) demonstrated the theoretical power of such networks, showing that they could compute any computable function and implement fundamental logical operations like AND, OR, and NOT through simple configurations [RN16].

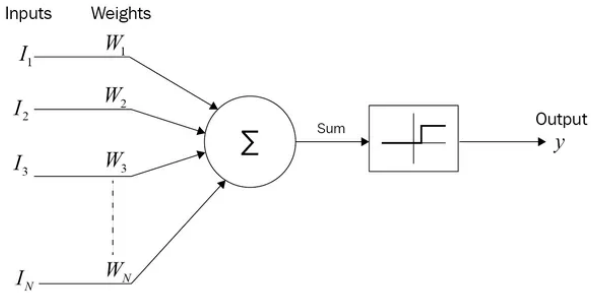

Figure 2: The McCulloch-Pitts neuron representing an artificial neuron.

LLMs try to imitate the human brain by using the three different types of layers, called the input layer, the hidden layers and the output layer (shown in Figure 3). The input layer receives information and data, which are numerical representations of words called vectors. Note that tokens are the linguistic units, while vectors are the mathematical representations of these units. Each token is integrated into a vector in the LLM processing pipeline. Next, the hidden layers perform the core computations. They capture complex patterns by adjusting their internal weights during the training process based on the errors on the training data. LLMs have a lot of hidden layers making them deep neural networks. This ensures that they generate the desired final output in the last output layer.

Many LLMs are trained to predict the next word in a sequence given the preceding context. Through this process, LLMs learn statistical patterns, grammatical structures, and representations of word meaning and their relationships (as shown with the attention mechanism). Words are internally represented as high-dimensional vectors, or embeddings, where words with similar meanings or used in similar contexts are closer to each other

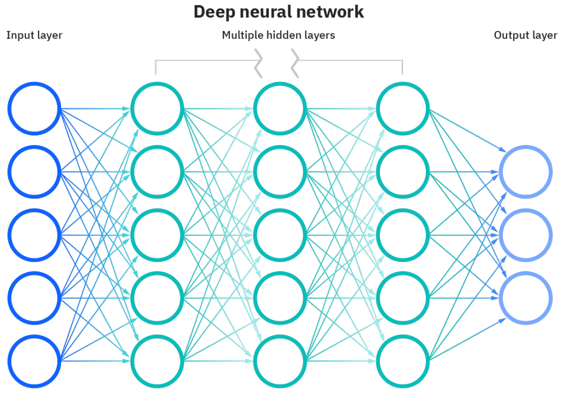

Figure 3: The three layers in Artificial Neural Networks (ANN) [But22].

in the vector space. To determine the degree of semantic relatedness between words as represented by their vectors, the cosine similarity is often calculated, which is a metric measures of the cosine angle between the two vectors. Vectors that are closer together (have a smaller angle) will have a higher cosine similarity, suggesting a stronger semantic relationship.

1.3 Related Work

As mentioned before, semantic priming is the decrease in reaction time for a target word that is semantically related to a previously presented stimulus word, compared to an unrelated stimulus word. The nature of what exactly makes a semantically related pair has been a central topic of investigation in psycholinguistics. In a meta-analysis of 26 studies, Lucas demonstrated that automatic semantic priming can indeed occur even without pre-existing word associations [Luc00]. It means that priming is not solely dependent on how frequently words co-occur or how readily one word brings another to mind (strong associations, e.g., doctor-nurse). Instead, purely meaning-based connections, such as those based on shared semantic features (e.g., deer-pony), category membership, or functional relationships, are sufficient to prepare target word processing. This finding challenged models that viewed associative links as the primary or sole driver of automatic priming. The meta-analysis showed that even when controlling for association values, ensuring that semantically related pairs (e.g., deer-pony) had little or no associative strength, a significant priming effect remained. For example, Lucas highlights the early work of Fischler (1977), who specifically constructed cue-target pairs that were semantically related, but had low word association values. The effects of semantic priming in Fischler's study were substantial (d = 0.79), but smaller than for pairs that were both semantically and associatively related (d = 1.17) [Luc00]. This is highly relevant for the current thesis, as it shows the importance of being able to capture the relationships purely semantic (without association) to accurately predict human priming behavior.

Furthermore, she investigated "associative boost" where word pairs that are both semantically and associatively related (e.g., doctor-nurse) tend to cause stronger priming effects than pairs that are only semantically related (e.g., deer-pony). Her analysis confirmed the existence of associative boost. It suggests that although purely semantic relationships are sufficient to cause priming, adding an associative link between words significantly enhances the effect (on average doubling the effect size). This distinction is crucial for understanding the mechanisms underlying semantic priming and for interpreting data from priming experiments. For LLMs, the question of whether they treat internally associated words differently from semantically related words has arisen.

To answer the above question about LLMs, robust, large-scale datasets on human priming are needed. One such comprehensive dataset has been provided by a recent research about semantic priming conducted by Buchanan et al. (2025) "Measuring the Semantic Priming Effect Across Many Languages" (SPAML) [BCC$^+$25]. They conducted a study across 19 different languages collecting data from 25.163 adult participants (18+). It involved a continuous lexical decision task (LDT) in which adult participants had to indicate as quickly as possible whether a certain letter string forms an existing word or a non-existing word (nonword). Some of the consecutively presented stimuli were critical word pairs, where the first one of the sequence is called the cue and the second is called the target.

This experiment consisted of approximately 800 trials with the combination word-word,

word-nonword, nonword-word and nonword-nonword. The interval between the cue and target words was approximately 500 ms. In addition to existing words and nonwords, there was also a distinction between target words that were preceded by semantically related or unrelated cue words. It is important to note that these underlying pairings were not evident to the participants, as they were presented with a continuous stream of individual letter strings, making a lexical decision for each one. All word pairs are eventually categorized into related, unrelated or nonword.

The latter research will form the basis for the current thesis for several reasons. Firstly, the dataset is large and utilizes 1.000 unique target words. Previous studies had too few observations to reliably predict the priming effect. The study by Buchanan et al. provides a robust and reliable estimate of the priming effect using a large dataset, which minimizes noise and enables more stable calculations of the dependent variable for our predictive models [BCC+25]. Secondly, the SPAML project was specifically designed to measure item-level semantic priming effects, which is essential for the current thesis. Item-level semantic priming means that the priming effect is examined, measured, and analyzed for each individual cue-target pair rather than an average priming effect across all word pairs [HPL+24]. For example, consider the pairs cat-dog and nurse-doctor. An analysis focusing on the average effect would combine the priming scores of both pairs to report a single mean priming effect. In contrast, an item-level analysis would treat each pair individually, retaining separate priming scores for cat-dog and nurse-doctor. Item-level semantic priming allows researchers to investigate why some pairs produce stronger priming than others by correlating these individual priming scores with item-specific properties. In addition, the explicit clear categorization, where each target is paired with both a designated related cue and a designated unrelated cue, is suitable for testing the predictive power of LLMs. The LLM assessment of these predefined related and unrelated pairs can directly correlate with observed human priming magnitudes for those same targets.

1.4 Large Language Models predicting human processing

The introduction of LLMs was not only interesting in the field of AI, but also in human cognition, especially in language processing. As Niu et al. mentioned, LLMs have become a point of interest for cognitive scientists seeking to unravel the mysteries of human cognition [NLB+24]. This interest stems from the remarkable capabilities LLMs demonstrate in mimicking human-like performance on various cognitive tasks, including aspects of language understanding, reasoning, and even sensory judgments. Niu et al. highlighted that the relationship is bidirectional: cognitive science provides foundational knowledge that shapes the design and improvement of LLMs, while the capabilities of these models offer new perspectives that challenge and reshape our current understanding of human cognition [NLB+24]. This raises the question: What kind of human behavior can be explained and/or predicted by LLMs?

Earlier research has demonstrated that computational models, including earlier forms of distributional semantics and more recent LLMs, can predict human similarity judgments with accuracy. For instance, a study by De Deyne showed that vector space models can capture the structure of human free association norms [DD24]. Similarly, other work by Gerz et al. has highlighted the ability of LLM-derived embeddings to align with human ratings of word similarity or relatedness [GVH+16]. These studies raise a further, critical question: To what extent can such models also predict more subtle effects/behavioral measures, in particular semantic priming?

It is also important to acknowledge the differences in how LLMs and humans acquire language. Humans typically learn language with far less explicit textual data than LLMs are trained on, and human language is deeply embedded in a multimodal environment. We have access to different types of input with our senses, such as visual, auditory and social interaction, which are (for now) absent in standard LLMs. It is noteworthy that standard LLMs are evolving towards multimodel LLMs, which can process information from various modalities as humans can with their senses, such as noise, images and videos. However, the interaction of advanced multimodal systems still differ significantly from human learning. For many standard LLM applications, and particularly for the text-based similarity judgments used in this research, the input remains unimodal. Furthermore, the statistical patterns that an LLM extracts are language-specific. An LLM that is trained primarily in the English language will develop representations that reflect English linguistic and cultural patterns. The English patterns may differ significantly from the patterns that an LLM would learn, for example, in Dutch. This highlights the importance of taking into account the training data and the linguistic context when interpreting LLM-derived measures.

However, when considering the question of what kind of human behavior LLMs can explain or predict, the nature of the prompt we humans provide becomes a critical determinant. Similarity in word relationship can be judged in different ways, such as rhyme, number of common letters or context. The distinction between association-based and feature-based similarity, previously detailed in Section 1.3, is important for the current approach to capture relationship strength. For clarity, association-based means that a word pair is related if they co-occur in similar situations or derive their meaning from the context of the situation or sentence. Examples are cold-hot, spider-web or needle-haystack. Relatedness in terms of feature-based is defined by shared semantic features, such as deer-pony, cat-dog or guitar-violin. Both types of relationships can lead to priming, however their results might differ depending on the specific task parameters. The difference between association-based and feature-based are especially important when working with LLMs. For instance, word embeddings capture a blend of associative and feature-based information due to their training on co-occurrence statistics [DD24]. In contrast, asking a specific question to an LLM allows for a more targeted assessment of similarity. For example, the words “cat” and “table” are nothing alike and will have a low chance of being used in the same context. However, we can still expect that cat-table would have a higher similarity score when using a feature-based prompt compared to an association-based prompt. This is because they both have four legs and, therefore have a shared semantic feature.

Given the previous considerations, the current thesis provides several key expectations. Firstly, it is expected that LLM-derived similarity measures obtained through direct prompting, will offer significant predictive power for human semantic priming effects. This can potentially surpass the general similarity captured by embeddings. Secondly, we hypothesize that our LLM predictors designed to capture association-based relatedness might show a stronger overall relationship with the observed priming data compared to the purely feature-based predictors. The reasoning comes from the paper of Lucas in which associative links provide a significant boost in human priming strength over purely semantic links [Luc00]. Therefore, if our LLM successfully captures this cognitive disparity between the two semantic relationships, we expect its association-based scores to be a more powerful predictor of the priming effect. Furthermore, given that the experimental priming effect is defined as the difference in reaction times (RT) between related and unrelated words, we hypothesize that LLM-derived predictors calculated as the difference score will be more accurate than predictors based solely on the similarity of the related pair [CGA+23]. In addition, we take some statistical issues into account when combining all the different predictors together in one analysis, which is also known as multicollinearity. It may lead to challenges in disentangling their unique contributions and could potentially influence the results for some predictors.

Therefore, this paper will try to answer the research question: "To what extent can similarity predictions derived from Large Language Models explain semantic priming effects in human lexical processing?"

1.5 Thesis overview

The current thesis contains an experiment using multiple predictors derived from an LLM to explain semantic priming effects in human lexical processing. Section 2 details the methodology that is divided into multiple different subsections, which contains a description of the experimental dataset, including the priming effect, by Buchanan et al. [BCC+25]. In addition, it gives a conceptual overview of how the LLM-based predictors were generated with different prompts and an in-depth explanation of the data processing steps in RStudio. The calculations needed to create the final set of predictors used in this study will also be presented. Following, Section 3 presents the results of the statistical analyses, which includes the correlation analyses and the results of three distinct multiple linear regression analyses to address potential multicollinearity and better isolate conceptual contributions. Afterwards, Section 4 will be the general discussion section. The results of the experiment is interpreted, any limitations are discussed and recommendations for future research is given. Finally, the thesis ends with a conclusion in Section 5 about the results of the experiment to answer the main research question.

2 Methodology

The experimental data for this study was obtained from the original research by Buchanan et al. [BCC+25]. While they collected data across 19 different languages, the current thesis focuses exclusively on the English dataset derived from SPAML. The English subset consists of 1.000 unique English target words and for each of these target words the SPAML project also included 1.000 unique English cue words. The data for the English subset was derived from the analysis of 5.964 adult participants. It is important to note that a specific word from a set could serve as a cue in one pair and as a target in another, allowing for a varied set of linguistic stimuli. An important aspect of the current research is the pairing strategy with both a semantically related and unrelated cue for each target word. A related cue had a meaningful connection to the target, whereas a unrelated cue word was chosen to be a control condition without any connection to the target. This is how the in total 2.000 cue-target pairs were formed and labeled as related or unrelated pairs to perform the lexical decision task for the English SPAML project. The initially selection and matching of these related and unrelated cues to targets was performed using word embeddings. Specifically, the cosine similarity between word embeddings was used. Cosine similarity measures the cosine of the angle θ between two word vectors, A and B, in the vector space. A cosine similarity value closer to 1 indicates a smaller angle and thus higher similarity, while a value closer to 0 indicates greater dissimilarity. A value closer to

-1 would indicate opposite meanings (this is less common for standard word embeddings). The formula looks as follows:

$$\text{similarity}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

In the formula, the dot product of vectors $A \cdot B$ is calculated as $\sum_{i=1}^{n} A_i B_i$, where A_i and B_i are the i-th components of vectors A and B. $\|A\|$ (and similarly $\|B\|$) denotes the Euclidean norm or magnitude (length) of vector A, calculated as $\sqrt{\sum_{i=1}^{n} A_i^2}$. This was used to identify comparable items across the various languages studied by Buchanan et al. [BCC+25].

Buchanan et al. utilized the embedding-based similarity approach not only for constructing the English stimuli, but also as a method to identify and create comparable stimulus items across the diverse languages included in their broader cross-linguistic study [BCC+25]. For their stimulus selection and initial similarity calculations, they relied on pre-trained fastText word embeddings, which are mainly derived from the OpenSubtitles corpus via the subs2vec project (by Paridon and Thompson (2021) as cited by Buchanan et al.). This ensured a degree of consistency in the relational strength of related pairs across different linguistic contexts. It is important to note that while the SPAML stimuli were selected based on similarity scores from these fastText embeddings, the current thesis uses embeddings for constructing its own cosine similarity predictors generated by OpenAI's text-embedding-3-small model [Ope25]. The focus of this thesis is therefore on evaluating the predictive power of text-embedding-3-small based similarities (and prompted gpt-4o-mini judgments) for the human priming effects observed with the SPAML stimuli. The current thesis will only use the English data where 1.000 related and 1.000 unrelated unique pairs form the basis for deriving the experimental priming effect.

2.1 Data Processing

As mentioned, the data for the thesis comes from the study by Buchanan et al. and was provided as a CSV file named en_SPAML.csv [BCC+25]. For the data processing and analyses in this study, R version 4.3.3 was used using RStudio. The entire analysis process is divided into multiple R scripts, which can be executed in the correct order using a master script, **Master.R**, to provide a structured and reproducible approach. The complete R code utilized for this research is available through the link in the Appendix or click here.

The first script **CodeSPAML.R** initiates the data handling process. Firstly, it reads the CSV file named "en_SPAML.csv", which contains the experimental data from Buchanan et al. [BCC+25]. The contents are saved into a data frame called en_SPAML. Then it reads another CSV file named "en_words.csv", which contains information about the word pairs, including their type as discussed in Section 1.3 (related, unrelated, nonword) and a pre-calculated cosine similarity score between the cue and target words of each pair. These two data frames, en_SPAML and the data from "en_words.csv" serve as essential inputs for the next script **API.R**.

2.2 Prompting Methodology

The subsequent script, **API.R**, is responsible for generating the new LLM-derived similarity scores used as predictors. As mentioned in Section 2, the original SPAML study utilized fastText-based similarities for stimulus construction. However, the current thesis will be calculating the cosine similarity scores again. A key reason for this decision was to ensure consistency across the LLM-derived measures, as our second type of predictor also utilizes an OpenAI model (see below). By employing an embedding model from the same provider as the generative model used for prompted judgments, we aim for a more reliable comparison. To enable RStudio to interact with OpenAI's services, an API key was configured within the R environment. **API.R** generates two runs for two primary types of predictors to investigate the relationship between LLM-derived semantic similarity and human priming effects:

1. Cosine Similarity from Embeddings: The generated predictor is based on word embeddings generated by OpenAI's text-embedding-3-small model. The similarity between word pairs is then calculated using the cosine similarity formula, as previously detailed.

2. LLM Similarity Scores via Prompting: For this predictor, similarity ratings were obtained directly from an LLM using OpenAI's API with the gpt-4o-mini model.

To ensure that only relevant word pairs were processed for the LLM-based judgments, the **CodeSPAML.R** script is sourced at the beginning of **API.R**. This step filters out nonword pairs from the en_SPAML data, as the focus of this investigation is exclusively on existing words, categorized as either related or unrelated pairs. Two distinct prompt types were utilized inspired by the paper of De Deyne, one focusing on association-based relatedness and the other on feature-based relatedness [DD24]:

- **Association-Based Prompt:** "In this study we want to investigate the degree to which English words can be considered related. We will present you with two words: ', cue_word, ' and ', target_word, '. Words are related if they co-occur in similar situations and derive their meaning from the context in which the word occurs. Your task is to rate the relatedness of a word pair based on how often these two words are used together in everyday language. Use a numerical rating from 0 – 1 with 3 decimals. A rating of 0 means the pair has no possible relation/association in everyday language. A rating of 1 means that the pair has the highest possible degree of relation/association. Evaluate relatedness solely in regard to the co-occurrence patterns and not on meaning of a word. For example, the word pair cold - hot should have a high relatedness of 0.887, since they co-occur in similar situations. However, frog - square should have a low relatedness of 0.112, since they are not used in similar situations. Only return a numeric value, nothing else."

- **Feature-Based Prompt:** "In this study we want to investigate the degree to which English words can be considered related. We will present you with two words: ', cue_word, ' and ', target_word, '. Words must have shared semantic features of words to be considered related. The relationship is based on shared properties, attributes and categories (e.g., 'has four legs', 'is a vehicle', 'is edible'). Explicitly ignore word associations and co-occurrence in language (e.g., do not rate 'hot' and 'cold' high because they are often used in daily language). Focus solely on overlapping features. Use a numerical rating from 0 – 1 with 3 decimals. A rating of 0 means the pair has no possible relation in semantic features. A rating of 1 means that the pair has the highest possible degree of relation in semantic features. For example, the word pair cat - dog should have a high relatedness (eg., 0.887), since their features are similar (animal, pet, furry, has four legs). However, cat - bus should have a low relatedness (eg., 0.112), since they share very few features. Only return a numeric value, nothing else.”

For the prompt-based similarity, the gpt-4o-mini model was instructed to provide a similarity score calculated for each word pair individually. The individual processing approach was chosen to ensure that each judgment was made independently, minimizing potential contextual carry-over effects that might arise if multiple pairs were presented simultaneously. In constructing the input for the LLM, the specific cue and target word for the current evaluated pair are inserted into predefined placeholders within the prompt template, such as “, cue_word, ” and “, target_word, ”. It was asked to give a numerical rating with 3 decimals for each word pair on a scale from 0 (no possible relation) to 1 (highest possible degree of relation). To enhance clarity and consistency of the API responses, the prompts include examples and explicitly requested only a numeric value as output.

As mentioned earlier, the R script **API.R** runs two times for each distinct prompt. Each of the two distinct prompt types (association-based and feature-based) was submitted twice for every word pair, resulting in two separate similarity ratings per prompt type. An example of the previous mentioned variability over two runs is shown in Table 1 with both association-based and feature-based prompt for the first five pairs. The reasoning behind the repeated measures was to capture and reduce minor variability by averaging the two ratings for each prompt-pair combination in a later stage. These runs are saved into separate CSV files containing the four columns “pair”, “type”, “cosine_similarity” and “prompt_similarity”. These separate files are saved under the names: “results_association1.csv”, “results_association2.csv”, “results_feature-based1.csv”, and “results_feature-based2.csv”.

	Cosine Similarity		Association-Based		Feature-Based	
Pair	Run 1	Run 2	Run 1	Run 2	Run 1	Run 2
paradise-father	0.253*	0.253*	0.200	0.200	0.100	0.100
spring-autumn	0.463*	0.464*	0.742	0.822	0.800	0.800
fidelity-devotion	0.559*	0.559*	0.872	0.874	0.850	0.850
bridge-tunnel	0.450*	0.450*	0.764	0.786	0.552	0.600
excellent-band	0.238*	0.238*	0.125	0.245	0.143	0.112

Table 1: Variability of LLM-derived similarity scores across two runs for selected word pairs, showing cosine similarity, association-based and feature-based.

**Note:* While the cosine similarity scores between repeated runs were generally almost identical, minor variations could be observed at a high level of decimal precision.*

Unlike the prompt-based similarity scores, the cosine similarity derived from word embeddings (text-embedding-3-small) is based on a standard calculation that is independent of the specific prompt type used. Therefore, the cosine similarity value for any given word pair remains constant, regardless of whether the pair is evaluated using an association-based or a feature-based prompt. Although all four originally generated CSV files contained the cosine_similarity column, we consider it redundant in the feature-based files. It will be identical to that in the association-based files for the same word pairs. As a result, for subsequent analyses involving individual feature-based predictors, this particular cosine similarity column from the feature-based runs will be ignored. Their identical scatter plots can be seen in Appendix Figure 9 for more clarity and as proof.

2.3 Averaging Results and Correlation Calculation

As mentioned, for each condition (association-based and feature-based), the two result files were combined. From the combined data, the average "prompt_similarity" and average "cosine_similarity" were calculated for each unique word pair and type. Note that the cosine_similarity column in the feature-based files will be ignored (explained in Section 2.2). The averaged results were then saved into two new CSV files: "average_results_association.csv" and "average_results_feature-based.csv". These files contain the same four columns as earlier before the merging. Finally, after averaging the data the Pearson correlation coefficients were calculated between the "mean_cosine_similarity" and the "mean_prompt_similarity" for all pairs, only related pairs, and only unrelated pairs. In addition, the correlation between run 1 and 2 for both prompts are also calculated to ensure reliability. This was done separately for the association-based and feature-based conditions in the R scripts **AssociationCorrelation.R** and **FeatureBasedCorrelation.R**.

2.4 Predictor Variables

After establishing the averaged similarity scores for both association-based and feature-based prompts, the next step involves transforming these into a set of distinct predictor variables derived from the LLM. This was done in the **Priming.R** script creating two new CSV files named "predictors_priming_association.csv" and "predictors_priming_feature_based.csv". For each target word in the dataset we generated predictors that reflect two main types of information:

1. **Only_Related**: The related predictors quantify the similarity score only between the target word and its semantically related cue, either through embeddings or prompted LLM responses. It represents the most direct and intuitive test of the hypothesis that a stronger semantic link leads to a stronger priming effect. By including the simple measurement of the Only Related score, we can really prove that the Difference Score approach offers superior predictive power (as expected). Unrelated word pairs are ignored for this specific predictor, representing the direct semantic proximity captured by embeddings.

2. **DifferenceScores**: The difference predictors aim to more closely mirror the experimental calculation of the priming effect itself. For each target word, we calculated the difference between the LLM's similarity rating for its related cue-target pair and its unrelated cue-target pair.

   $$DifferenceScore = related - unrelated$$

   For example, if we have the target word "dog", which is paired with the related cue "cat" and the unrelated cue "table", the difference score would be calculated as:

   $$DifferenceScore = Similarity(dog, cat) - Similarity(dog, table)$$

A larger positive difference score indicates that the LLM perceives a much stronger relationship for the related pair compared to the unrelated control. This contrastive approach was chosen because it mirrors the calculation of the priming effect itself, which is also defined as the difference (in reaction times) between a related and an unrelated condition (further discussed in Section 2.5). By constructing the predictor this way, we hypothesize that it more accurately captures the variance in the model.

In total, there are six distinct predictors, which are divided into two embedding-based cosine similarity predictors derived from text-embedding-3-small and four prompt-based similarity predictors derived from gpt-4o-mini (two for association-based and two for feature-based). This results in the following six key predictor variables that were prepared for the multiple regression analysis. Prefixes "A." denote association-based, and "Fb." denote feature-based:

1. RelatedCos: Cosine similarity for the related word pairs.
2. DifferenceCos: Difference score using cosine similarity for word pairs.
3. A.RelatedPrompt: Prompt-based similarity (association-based) for the related pair.
4. A.DifferencePrompt: Difference score using prompt-based similarity (association-based).
5. Fb.RelatedPrompt: Prompt-based similarity (feature-based) for the related pair.
6. Fb.DifferencePrompt: Difference score using prompt-based similarity (feature-based).

These six predictors were then merged into a single CSV file "PrimingPredictors.csv" by R script **Predictors.R**.

2.5 Priming Effect

The dependent variable for the current thesis is the observed semantic priming effect, derived from the reaction time (RT) data collected by Buchanan et al. only for the English word pairs [BCC+25]. To account for individual differences in overall response speed and to normalize the distribution of reaction times (RTs), the raw RTs were Z-transformed [BCC+25]. This transformation was performed for each participant individually. For each unique target word, the mean Z-transformed RT (zRT) was then calculated separately for the two conditions related and unrelated:

1. zRTMean_related: the mean zRT when the target word was preceded by its semantically related cue.
2. zRTMean_unrelated: the mean zRT when the same target word was preceded by its semantically unrelated cue.

The final item-level priming effect for each target word was then computed as the difference between these two mean zRTs:

$$\text{zRT\_Priming\_Effect} = \text{zRTMean\_unrelated - zRTMean\_related}$$

A positive value for zRT_Priming_Effect indicates that the average responses to target words were faster when they were preceded by a related cue compared to an unrelated cue.

The zRT_Priming_Effect score, calculated per target word, serves as the primary outcome variable that the LLM predictors aim to explain. This observed priming effect was merged with the LLM-derived predictors for each target word into a final dataset, as detailed in Section 2.4 in the "PrimingPredictors.csv" file, which forms the direct input for the multiple regression analyses detailed in the subsequent sections.

2.6 Multiple Regression Analysis

To assess the predictive power of the LLM-derived scores, a series of Multiple Regression Analyses (MRA) were performed using the R script **MRA.R** on the "PrimingPredictors.csv" dataset. The dependent variable for all analyses is the zRT_Priming_Effect.

The primary analysis involved a single multiple linear regression model specified to predict the zRT_Priming_Effect using all six predictors simultaneously named in section 2.4. The general form of this regression equation is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_6 X_6 + \epsilon$$

Here Y is the zRT_Priming_Effect, $X_1, \ldots, X_6$ are the six predictor variables listed above, β_0 is the intercept, $\beta_1, \ldots, \beta_6$ are the regression coefficients for each predictor, and ϵ is the error term. This was implemented using a standard RStudio function called lm(), which is used to fit linear models to data frames in the R Language [R C23]. The formula is constructed as follows:

```
zRT_Priming_Effect ~ RelatedCos + DifferenceCos +
                      A.RelatedPrompt + A.DifferencePrompt +
                      Fb.RelatedPrompt + Fb.DifferencePrompt
```

The output of the multiple regression analysis, including the estimated coefficients (β) for each predictor, their standard errors (Std. Error), t-values, and corresponding p-values, is also presented in a new CSV file named "multiple_regression_table.csv". Additionally, the overall model fit statistics, such as R-squared (R^2), Multiple R-squared, and the F-statistic for the model were obtained to assess the proportion of variance in the zRT_Priming_Effect explained by the set of predictors [R C23].

Given the overlap between predictors (e.g., between a Difference Score and its corresponding Only Related score), we fitted two additional regression models to avoid issues with multicollinearity. Multicollinearity can increase standard errors and make it difficult to distinguish the unique contributions of individual predictors in a combined model. This topic will be discussed in more detail in Section 3.4.1. The first additional model included only the three Difference Score predictors, while the second included only the three Only Related Score predictors. By grouping similar predictor types, it became possible to more clearly assess the unique contribution of each predictor group to the priming effect, without the influence of the other group.

3 Results

This section presents the results of the statistical analysis conducted to investigate the relationship between the similarity measures derived from the LLM and the semantic priming effects (zRT_Priming_Effect). First, the results of the correlation analyses between the cosine_similarity and prompt_similarity for both prompts are detailed. Additionally, the descriptive statistics for the primary outcome variable and the generated predictors will be presented. The next subsection will show the results of the correlation analyses between the LLM predictors and the priming effect, followed by the outcomes of the three multiple linear regression analyses.

3.1 Internal Correlations of LLM-Derived Measures

The consistency between the two data collection runs for prompt-based similarity, and the relationship between averaged prompt_similarity and averaged cosine_similarity, were examined for both the association-based and feature-based prompt conditions (note that the cosine_similarity for both distinct prompts is identical). This was done in the earlier mentioned R scripts **AssociationCorrelation.R** and **FeatureBasedCorrelation.R**.

To assess the reliability of the similarity scores generated by the LLM, Pearson correlations were calculated between the outcomes of the first and second run for each measure. For the prompt-based similarity scores, the correlation between Run 1 and Run 2 was very high for the association-based prompt (r = .99). A similarly high correlation was found for the feature-based prompt (r = .99). The embedding-based cosine similarity scores also demonstrated consistency across the two runs, with a correlation of r = .99. These high correlations between runs indicate a strong consistency in the LLM's responses across repeated evaluations with the same prompt. The findings show the reliability of the measures for further analysis after averaging. The almost perfect correlation for cosine similarity confirms its near-deterministic nature when derived from the same embeddings.

The overall correlation between the averaged mean_cosine_similarity and the averaged mean_prompt_similarity for association-based measures was $r = .85$. When considering only related pairs, the correlation was $r = .28$, and for unrelated pairs, it was $r = .31$.

The overall correlation between the averaged mean_cosine_similarity and the averaged mean_prompt_similarity for feature-based measures was $r = .82$. For related pairs, this correlation was $r = .20$, and for unrelated pairs, it was $r = .25$. For more clarity, the Pearson correlation coefficients (r) that are discussed above are presented in Table 2.

Correlation Type	**Association-Based**	**Feature-Based**
Correlations based on Mean Scores		
Overall	.85	.82
Only Related Pairs	.28	.20
Only Unrelated Pairs	.31	.25
Correlations between Runs		
prompt_similarity	.99	.98
cosine_similarity	.99	.99

Table 2: Pearson Correlations (r) between Cosine Similarity and Prompt Similarity Measures.

3.2 Descriptive Statistics

The experimental semantic priming effect (zRT_Priming_Effect) was calculated as the difference between mean Z-transformed reaction times to unrelated and related targets.

This showed a mean of 0.12 (SD = 0.15, min = -0.66, max = 0.94) across the 1.000 unique target words in the data. The descriptive statistics for the six predictors are presented in the Table 3.

Predictor	**Observations**	**Mean**	**SD**	**Min**	**Max**
RelatedCos	1000	0.54	0.11	0.16	0.90
DifferenceCos	1000	0.29	0.13	-0.16	0.70
A.RelatedPrompt	1000	0.79	0.13	0.05	1.00
A.DifferencePrompt	1000	0.65	0.16	-0.11	0.95
Fb.RelatedPrompt	1000	0.74	0.21	0.00	1.00
Fb.DifferencePrompt	1000	0.66	0.21	-0.16	1.00

Table 3: Descriptive Statistics for LLM-Derived Predictor Variables

As shown in Table 3, the mean score for RelatedCos was 0.54 (SD = 0.11), with scores for related pairs ranging from 0.16 to 0.90. The corresponding DifferenceCos (M = 0.29, SD = 0.13) suggests that related pairs generally had a higher cosine similarity than their unrelated counterparts for the same target, as expected.

For the prompt-based predictors derived from the LLM, the Only Related scores (A.RelatedPrompt: M = 0.79, SD = 0.13 and Fb.RelatedPrompt: M = 0.74, SD = 0.21) were generally high, suggesting that the gpt-4o-mini model, when prompted, tended to rate the "related pairs" as indeed quite related. These results align with the experimental design. The standard deviations, particularly for Fb.RelatedPrompt (SD = 0.21), indicated a reasonable degree of variability in these ratings, implying that the LLM perceived varying strengths of relatedness across different related pairs. The Fb.RelatedPrompt showed a slightly lower mean, but a larger standard deviation compared to A.RelatedPrompt. This suggests that the feature-based prompt might have led to more diverse or more strict ratings. The feature-based prompt explicitly instructed the model to focus solely on overlapping features and to explicitly ignore word associations and co-occurrence, whereas the association-based prompt allowed for ratings based on broader contextual co-occurrence. Overall, the results showed that there is agreement between the experimental definition of relatedness and the assessment of the LLM.

The Difference Score predictors (A.DifferencePrompt: M = 0.65, SD = 0.19 and Fb.DifferencePrompt: M = 0.66, SD = 0.21) also had positive means, indicating that the LLM rated the related cue as more similar to the target than the unrelated cue. An important finding from Table 3 is that the minimum values for all Difference Score predictors were negative (DifferenceCos: Min = -0.16, A.DifferencePrompt: Min = -0.11 and Fb.DifferencePrompt: Min = -0.16). A negative difference score means that the LLM rated the unrelated pair as more similar to the target than its designated related cue. These occurrences highlight the potential differences between the relatedness computed by the LLM and the assumed relatedness in the experimental data. It can also be explained by the noise in the similarity measures.

3.3 Correlation Analyses for all six Predictors

To examine the relationships between the LLM-derived predictors and the zRT_Priming_Effect, the Pearson correlation coefficients (r) were calculated. These correlations are visualized in the six individual scatter plots for each predictor. Starting with the two cosine predictors in Figure 4 below:

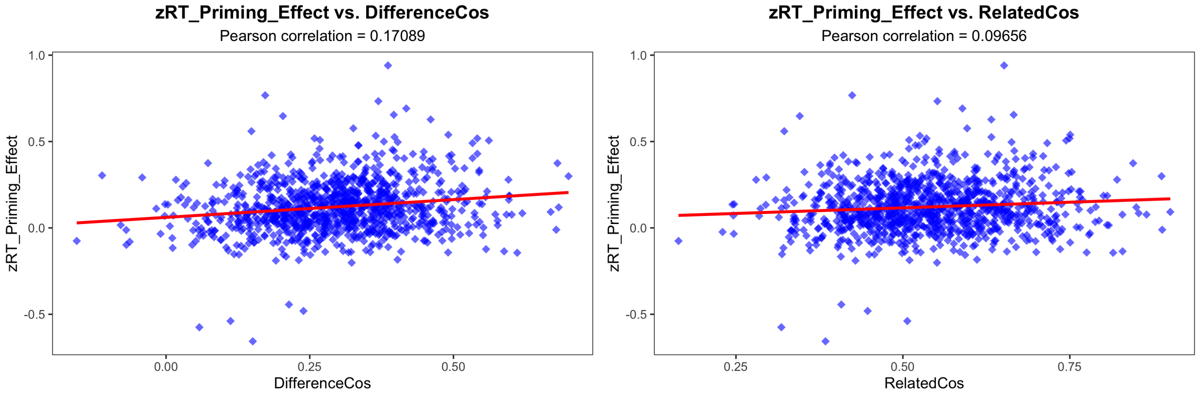

Figure 4: Two scatter plots showing the relationship between the zRT_Priming_Effect and two cosine similarity-based predictors. The plots visually demonstrate that the DifferenceCos measure has a notably stronger positive correlation with the priming effect. Each point represents a word pair, and the red line indicates the linear regression fit.

The DifferenceCos predictor, representing the difference in embedding-based cosine similarity between related and unrelated cue-target pairings, showed a positive relationship with the zRT_Priming_Effect. The regression line indicated that a larger difference in cosine similarity tends to be associated with a stronger priming effect. For the RelatedCos predictor, which reflects the direct cosine similarity of only the related cue-target pair, the correlation with the priming effect was quite lower compared to the previous predictor showing a weaker relationship. Comparing these two embedding-based measures, the DifferenceCos predictor exhibited a somewhat stronger linear relationship with the experimental priming effect than the RelatedCos predictor.

Now we continue with the two association-based predictors in Figure 5. The predictor A.DifferencePrompt demonstrated a comparable Pearson correlation to DifferenceCos. The corresponding scatter plot visually confirms this positive association. The A.RelatedPrompt predictor showed a weaker correlation with the priming effect, showing a less clearly positive trend. Within the association-based measures, similar to the cosine predictors, the difference score (A.DifferencePrompt) showed a stronger correlation with priming compared to its counterparts based solely on the related pair (A.RelatedPrompt).

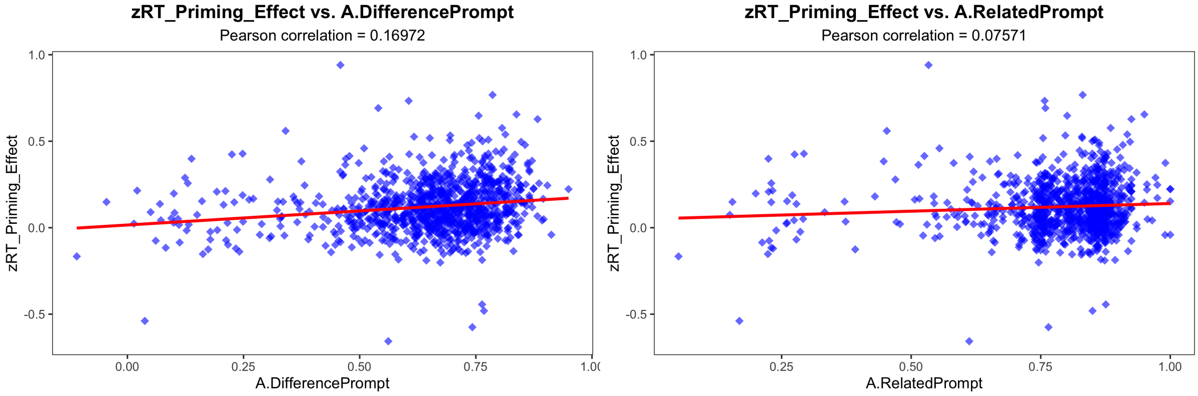

Figure 5: Two scatter plots showing the relationship between the zRT_Priming_Effect and two association-based predictors. The plots clearly illustrate that the A.DifferencePrompt is a stronger predictor of the priming effect. Each point represents a word pair, and the red line indicates the linear regression fit.

Lastly, the two feature-based predictors in Figure 6 show with both scatter plots a low Pearson correlation. Again, the Fb.DifferencePrompt predictor showed a stronger positive relation compared to Fb.RelatedPrompt. However, they both have a very weak positive trend. The Fb.RelatedPrompt predictor showed the weakest correlation of all predictors, confirming a nearly straight line.

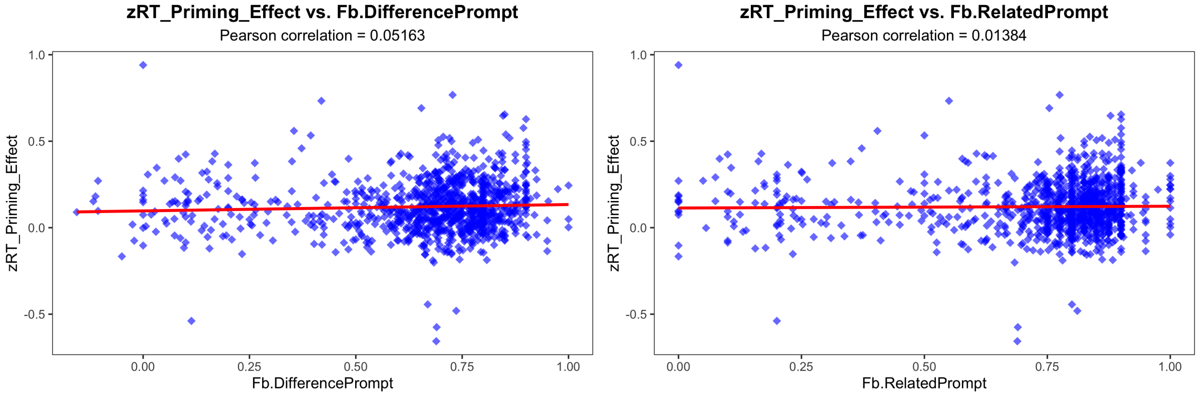

Figure 6: Two scatter plots showing the relationship between the zRT_Priming_Effect and two feature-based predictors. Both plots demonstrate that the feature-based judgments have a straight line, showing almost no ability to predict priming effect. Each point represents a word pair, and the red line indicates the linear regression fit.

In summary, these correlations suggest that all LLM-derived measures show a positive linear relationship with human semantic priming. The Difference Score calculation tends to have a slightly stronger correlations than the Only Related scores across both cosine and prompt-based measures. Furthermore, predictors based on cosine similarity and association-based judgments appear to capture more variance in the priming effect than those based on feature-based judgments.

3.4 Multiple Linear Regression Analysis

To assess the combined predictive power of all six LLM-derived predictors and their contributions to explaining the zRT_Priming_Effect, a multiple linear regression model was fitted. The model included all six predictors: RelatedCos, DifferenceCos, A.RelatedPrompt, A.DifferencePrompt, Fb.RelatedPrompt, and Fb.DifferencePrompt. The results of this regression analysis are summarized in Table 4:

Predictor	Estimate	Std. Error	t value	Pr($> \|t\|$)
(Intercept)	0.11	0.04	2.70	0.00699 **
RelatedCos	-0.19	0.09	-2.18	0.02987 *
DifferenceCos	0.32	0.08	4.04	5.70e-05 ***
A.RelatedPrompt	-0.18	0.10	-1.75	0.08107 .
A.DifferencePrompt	0.34	0.08	4.17	3.31e-05 ***
Fb.RelatedPrompt	0.09	0.11	0.79	0.43283
Fb.DifferencePrompt	-0.19	0.11	-1.73	0.08390 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.15 on 993 degrees of freedom
Multiple R-squared: 0.06926, Adjusted R-squared: 0.06363
F-statistic: 12.31 on 6 and 993 df, p-value: 2.098e-13

Table 4: Multiple Linear Regression Results Predicting zRT_Priming_Effect.

Examining the overall statistical significance of the multiple linear regression model, the F-statistic serves as a key indicator. The analysis showed a F-statistic value of 12.31, which represents the ratio of the variance explained by the model to the unexplained variance (error). A higher F-value generally suggests a better model fit. Looking at both degrees of freedom, the first degree of freedom (6) is the number of predictors in the model. The second degree of freedom (993) is the number of observations minus the number of estimated parameters (number of observations - number of predictors - 1 for the intercept). The p-value associated with this F-statistic is extremely small (p = 2.098e-13). Therefore, we reject the null hypothesis that all regression coefficients (excluding the intercept) are equal to zero. This indicates that the overall regression model is statistically significant, ($F(6, 993) = 12.31, p < .001$), which suggests that the set of six LLM-derived predictors, when considered together, collectively explains a portion of the variance in the zRT_Priming_Effect.

Multiple R-squared ($R^2 = 0.06926 \approx 7.0\%$) is a crucial measure of the explanatory power of the model. It means that about 7.0% of the total variation in the zRT_Priming_Effect between the different target words can be explained or predicted by the combination of the six predictors. Additionally, it also means that there is still 93% of the variance in the priming effect that is not explained by the current set of predictors. The Adjusted R-squared is an adjusted version of R^2 that takes into account the number of predictors in the model. R^2 will always remain the same or increase if there are more predictors added. The Adjusted R^2 (6.4%) is slightly lower than the Multiple R^2 (7.0%), indicating that not many unnecessary predictors were added.

Looking at the contributions of the individual predictors within the multiple regression model, several patterns emerge (see Table 4). The A.DifferencePrompt score, reflecting the associative difference between related and unrelated pairs, is a statistically significant positive predictor of the zRT_Priming_Effect ($\beta = 0.34, SE = 0.08, t(993) = 4.17, p < .001$). This indicates that a larger perceived associative difference by the LLM is uniquely associated with a stronger priming effect, even when controlling for other predictors. Similarly, the DifferenceCos predictor, based on the difference in embedding cosine similarity, also emerged as a significant positive predictor ($\beta = 0.32, SE = 0.08, t(993) = 4.04, p < .001$). This suggests that a greater difference in cosine similarity between related and unrelated pairs also uniquely predicts a larger priming effect.

Interestingly, the RelatedCos predictor showed a significant negative relationship with the priming effect ($\beta = -0.19, SE = 0.09, t(993) = -2.18, p = .030$). It implies that, when controlling for the other variables in the model, a higher cosine similarity for only the related pair is associated with a weaker priming effect. This finding could indicate complex interactions or suppressor effects, possibly due to a strong linear relationship with the DifferenceCos measure. This is also known as multicollinearity, which will be further discussed in Section 3.4.1.

Two other prompt-based predictors showed trends towards statistical significance. The A.RelatedPrompt had a negative coefficient that approached significance levels ($\beta = -0.18, SE = 0.10, t(993) = -1.75, p = .081$). This negative estimate implies that, when controlling for all other predictors in the model, for every 1.0 unit increase in the A.RelatedPrompt score, the zRT_Priming_Effect is expected to decrease by 0.175 units. This suggests that a higher perceived associative relatedness of the cue-target pair alone might be associated with a weaker priming effect, which is possibly due to multicollinearity issues that will be discussed in Section 3.4.1. Similarly, the Fb.DifferencePrompt predictor also showed a trend towards significance with a negative coefficient ($\beta = -0.19, SE = 0.11, t(993) = -1.73, p = .084$), indicating that a larger feature-based difference perceived by the LLM might be associated with a weaker priming effect. This finding for a difference score, especially a feature-based one, could be influenced by the complex interplay of predictors within the model.

The Fb.RelatedPrompt predictor ($\beta = 0.09, SE = 0.11, t(993) = 0.79, p = .433$) was not found to be statistically significant in the primary multiple regression model, since $p > .05$. It means that once we already accounted for the effects of the other five predictors, Fb.RelatedPrompt did not provide any additional, unique information to help explain the differences in zRT_Priming_Effect.

An examination of the model's residuals provided insights into the model's fit (see Figure 7). The residuals ranged from a minimum of -0.75 to a maximum of 0.77, with a median of -0.01, which is very close to zero. The first quartile (1Q) was -0.09 and the third quartile (3Q) was 0.08. Overall, the distribution suggested that the residuals were approximately symmetrically distributed around zero.

3.4.1 Multicollinearity

Multicollinearity arises when there is a strong linear relationship between two or more explanatory variables in a regression model. This phenomenon is generally considered unfavorable in regression analyses as it can lead to less precise or unstable estimates of the individual regression coefficients. When predictors are highly correlated, they provide redundant information, making it difficult for the model to disentangle their unique contributions to the explained variance in the dependent variable [Pau06].

In the primary multiple regression analysis, the simultaneous inclusion of both Only Related scores and Difference Scores (for both cosine similarity and prompt-

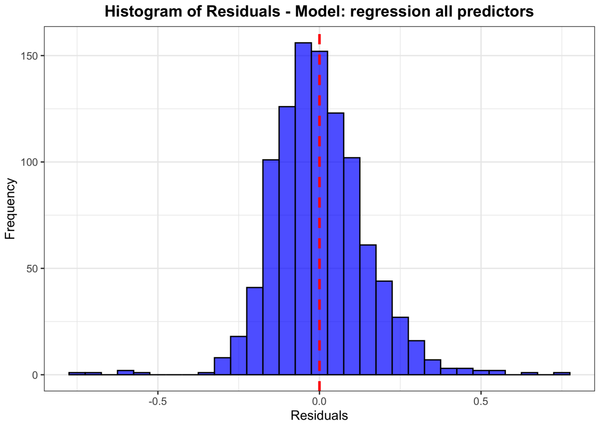

Figure 7: Histogram of the residuals from the multiple linear regression all six predictors model predicting the zRT_Priming_Effect. The data follows a roughly normal bell-shaped distribution and is centered at zero, showing a well-fitted linear regression model.

based similarity) raises consideration of multicollinearity. The Difference Scores are mathematically constructed using its corresponding Only Related score (formula: DifferenceScore = RelatedScore - UnrelatedScore). The Only Related predictors (RelatedCos, A.RelatedPrompt and Fb.RelatedPrompt) and the Difference Score predictors (DifferenceCos, A.DifferencePrompt and Fb.DifferencePrompt) are very similar, because the Difference Score depends directly on what the Only Related predictors do. If the scores of the Only Related predictors are high, the predictors of the Difference Score will generally also be high.

A notable finding in the primary multiple regression analysis was the statistically significant negative coefficient for RelatedCos ($\beta = -0.19$), despite its positive correlation with the zRT_Priming_Effect. A higher direct cosine similarity of the related pair associated with a weaker priming effect can potentially be explained by the presence of multicollinearity. In a multiple regression model, the unique contribution of each predictor is assessed. The DifferenceCos predictor, which exhibits a strong positive and significant effect ($\beta = 0.32$), represents the contrast in cosine similarity between the related and unrelated conditions. This measure appears to effectively capture the positive prediction arising from the cosine similarity contrasts.

The DifferenceCos predictor captures the primary positive relationship with the priming effect. Once the model accounts for this dominant predictor, it then assesses the remaining contribution of RelatedCos. This is where a suppressor effect, a statistical byproduct of multicollinearity, becomes evident [Bec12]. A high RelatedCos score that is not accompanied by a correspondingly high DifferenceCos score is quite unusual (because DifferenceCos = RelatedCos - UnrelatedCos). This scenario can only occur when the similarity of the unrelated control pair (UnrelatedCos) is also unusually high. Since high similarity to an unrelated word is associated with a weaker priming effect, the unique information that RelatedCos adds becomes negatively correlated with priming. Therefore, the negative coefficient highlights that DifferenceCos is the more comprehensive and dominant predictor of the similarity contrast relevant to priming. It does not necessarily imply that higher RelatedCos in isolation hinders priming.

3.4.2 Two Separate Multiple Linear Regression Analyses

To more clearly demonstrate the predictive utility of the different types of LLM-derived measures and to explore the multicollinearity concerns highlighted in the combined six-predictor model, two separate multiple linear regression models were fitted. As shown in Table 5, the first model included only the three Difference Score predictors (DifferenceCos, A.DifferencePrompt and Fb.DifferencePrompt). The second model, detailed in Table 6, included only the three Only Related score predictors (RelatedCos, A.RelatedPrompt and Fb.RelatedPrompt). Both models aimed to predict the zRT_Priming_Effect.

Predictor	**Estimate**	**Std. Error**	**t value**	**Pr(> $\|t\|$)**
(Intercept)	-0.01	0.02	-0.27	0.791
DifferenceCos	0.16	0.04	4.17	3.29e-05 ***
A.DifferencePrompt	0.26	0.05	5.58	3.04e-08 ***
Fb.DifferencePrompt	-0.13	0.03	-3.93	8.99e-05 ***

Residual standard error: 0.1471 on 996 degrees of freedom
Multiple R-squared: 0.05872, Adjusted R-squared: 0.05588
F-statistic: 20.71 on 3 and 996 DF, p-value: 5.069e-13
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 5: Separate Multiple Linear Regression Results for Difference Score Predictors Predicting zRT_Priming_Effect.

The regression model containing only the three Difference Score predictors was statistically significant overall ($F(3, 996) = 20.71, p < .001$)(Table 5). The model explained approximately 5.6% of the variance in the zRT_Priming_Effect (Adjusted $R^2 = 0.06$). Both A.DifferencePrompt ($\beta = 0.26, p < .001$) and DifferenceCos ($\beta = 0.16, p < .001$) remained highly significant positive predictors. This confirms that the contrast in both prompt-based associative relatedness and embedding-based cosine similarity uniquely contribute to predicting the priming effect. Fb.DifferencePrompt also emerged as a statistically significant predictor in this model, but with a negative coefficient ($\beta = -0.13, p < .001$) implying that a larger feature-based difference is associated with a weaker priming effect when only difference scores are in the model.

Predictor	**Estimate**	**Std. Error**	**t value**	**Pr(> $\|t\|$)**
(Intercept)	-0.01	0.03	-0.22	0.82940
RelatedCos	0.11	0.04	2.44	0.01505 *
A.RelatedPrompt	0.17	0.06	2.77	0.00567 **
Fb.RelatedPrompt	-0.08	0.04	-2.26	0.02430 *

Residual standard error: 0.1504 on 996 degrees of freedom
Multiple R-squared: 0.01694, Adjusted R-squared: 0.01398
F-statistic: 5.72 on 3 and 996 DF, p-value: 0.0006989
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 6: Separate Multiple Linear Regression Results for Only Related Score Predictors Predicting zRT_Priming_Effect.

The second separate regression model containing only the three Only Related score predictors was also statistically significant overall ($F(3, 996) = 5.72, p < .001$). However, it explained a considerably smaller portion of the variance in the zRT_Priming_Effect (Adjusted $R^2 = 0.01398 \approx 1.4\%$). The individual coefficients of RelatedCos was now a statistically significant positive predictor ($\beta = 0.11, p = .015$), while in the analysis with all six predictors it was negative. A.RelatedPrompt also became a statistically significant positive predictor ($\beta = 0.17, p = .006$). In contrast, Fb.RelatedPrompt showed a statistically significant negative coefficient ($\beta = -0.08, p = .024$), indicating that higher direct feature-based similarity is associated with a weaker priming effect when considered alongside other direct relatedness measures.

An inspection of the residuals for both models was conducted to assess the assumption of normally distributed errors in the multiple regression analysis. The residuals for the Difference model (Min = -0.73, 1Q = -0.10, Median = -0.01, 3Q = 0.08, Max = 0.77) were approximately symmetrically distributed around zero. The distribution of the numerical data of the residuals is visually represented in the left histogram in Figure 8. The residuals for the Only Related model (Min = -0.74, 1Q = -0.10, Median = -0.01, 3Q = 0.08, Max = 0.79) were also reasonably symmetrically distributed. This distribution is also demonstrated in the right histogram in Figure 8. Both indicated that the residuals were approximately bell-shaped and centered close to zero. However, the residual distribution for the Difference Score predictors model appeared slightly more peaked around the mean compared to that of the Only Related predictors model. Despite a few outliers in both tails, the overall shape provides reasonable support for the assumption of normally distributed errors for this model.

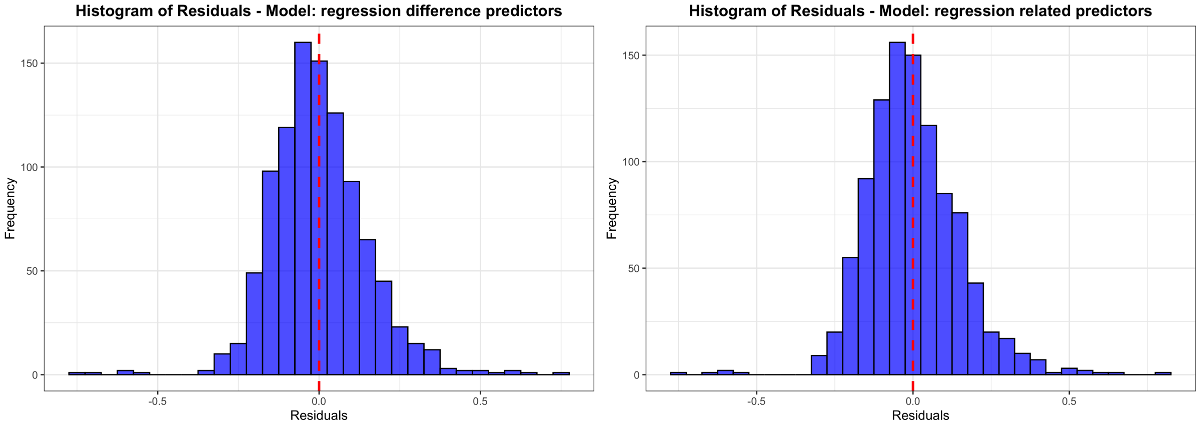

Figure 8: Histograms of the residuals for two separate multiple linear regression models predicting the zRT_Priming_Effect. On the left the residual distribution for the model with Difference Score predictors and on the right the residual distribution for the model with Only Related score predictors. Both distributions are centered close to zero (indicated by the dashed red mean line) and show a bell shape.

These separate analyses provided crucial insights when compared to the initial model that included all six predictors (Adjusted $R^2 = 6.4\%$, $F(6, 993) = 12.31, p < .001$).

First, the combined model of six predictors explained the most variance (Adjusted $R^2 = 6.4\%$). The model with only Difference Score predictors explained a substantial portion of this (Adjusted $R^2 = 5.6\%$), while the model with only Only Related predictors explained considerably less (Adjusted $R^2 = 1.4\%$). This suggests that the separate

Difference Score analysis captures more of the predictive information relevant to priming than the Only Related scores alone.

Another notable effect of fitting separate models is seen in the coefficients for RelatedCos and A.RelatedPrompt. In the combined model, both RelatedCos and A.RelatedPrompt had a significant negative coefficient. However, when included in the Only Related scores model (Table 6), both RelatedCos and A.RelatedPrompt became significant positive predictors, aligning with theoretical expectations and their positive correlations with priming. This strongly supports the expectation in Section 3.4 that their negative coefficients in the combined model were indeed statistical byproducts of multicollinearity, primarily due to their shared variance with the more dominant DifferenceCos and A.DifferencePrompt predictors.

A third notable difference was present in the feature-based predictors. Fb.DifferencePrompt was significantly negative in the Difference Score model, and Fb.RelatedPrompt was significantly negative in the Only Related model. In the combined six predictor model, when controlling for other similarity metrics, Fb.DifferencePrompt also trended negative, while Fb.RelatedPrompt was non-significant. The consistent negative association for feature-based measures, even when isolated with similar predictors, suggests that the way gpt-4o-mini judge feature-based similarity does not align positively with priming in this dataset.

In summary, conducting separate regression analyses confirms the superior predictive utility of Difference Score measures, particularly A.DifferencePrompt and DifferenceCos. It also clarifies that the counterintuitive negative coefficients for RelatedCos and A.RelatedPrompt in the combined model were likely due to multicollinearity. The consistent negative coefficients from the feature-based predictors across all model specifications suggests a potential misalignment between the LLM's feature-based judgments and the priming in the dataset.

4 Discussion

This research aimed to investigate to what extent similarity predictions derived from Large Language Models (LLMs) can explain semantic priming effects in human lexical processing. The primary measure of human semantic priming was the zRT_Priming_Effect, derived from the English subset of the Semantic Priming Across Many Languages (SPAML) dataset by Buchanan et al. [BCC+25]. The analytical approach involved generating six distinct predictor variables based on cosine similarity from OpenAI's text-embedding-3-small embeddings and API prompt-based similarity from gpt-4o-mini [Ope25]. These predictors were designed to capture both association-based and feature-based relatedness, distributed as either Only Related scores (direct similarity of the related cue-target pair) or Difference Scores (the contrast in similarity between a target's related and unrelated cues).

4.1 Summary and Interpretation

The primary multiple linear regression analysis, combining all six LLM-derived predictors, revealed that the overall model was statistically significant ($F(6, 993) = 12.31, p < .001$). This main finding indicates that the set of LLM-derived predictors can account for a meaningful portion of the variance in the semantic priming effect. Specifically, the model explained approximately 7.0% of the variance in the zRT_Priming_Effect (Adjusted

$R^2 = 0.064$). While these LLM scores only explained a small fraction of human priming, the model's overall significance supports the hypothesis that computational measures from LLMs capture aspects of semantic relatedness relevant to human lexical processing. This aligns with the work of Niu et al. suggesting that LLMs encode human-like knowledge about word relationships [NLB+24].

The analysis highlighted the dominant role of difference scores as predictors. The associative difference score of direct prompting, A.DifferencePrompt, emerged as the most robust unique predictor ($\beta = 0.34, p < .001$). Similarly, the DifferenceCos score, based on cosine similarity, made a strong unique positive contribution ($\beta = 0.32, p < .001$). This suggests that a predictor performs better when measuring the difference between a related and an unrelated pair, than to simply ask how related two words are. The superiority of the Difference Score predictors is further confirmed by our separate analysis (see Section 3.4.2), where the model with only the three Difference Score predictors accounted for the vast majority (Adjusted $R^2 = 5.6\%$) of the variance explained by the full model. The strength of the associative measure aligns well with theories like The Spreading Activation Theory [CL75].

Interestingly, the direct relatedness predictors exhibited more complex effects. The RelatedCos predictor showed a significant negative coefficient ($\beta = -0.19, p = .030$), and the prompt-based A.RelatedPrompt showed a similar negative trend ($p = .081$). As discussed in Section 3.4.1, these negative relationships are likely statistical byproducts of multicollinearity, acting as suppressor variables. This is confirmed by the separate regression analysis, where the same predictors became significant positive predictors once the influence of the Difference Score variables was removed.

Finally, the feature-based predictors were the least effective in the models. The Fb.DifferencePrompt score showed only a weak negative trend ($p = .084$), while the Fb.RelatedPrompt score was clearly not significant ($p = .433$). This indicates that judgments based on semantic features, as prompted to gpt-4o-mini, added little unique predictive value for the associative priming effect in this dataset.

4.2 Comparison of Predictor Types

4.2.1 Difference Scores vs. Only Related Scores

The findings show that the Difference Scores are more powerful and interpretable predictors of priming than Only Related scores. It was most evident in the separate regression analyses. The model containing only the three Difference Scores explained a substantial portion of the variance in priming (Adjusted $R^2 = 5.6\%$), whereas the model with the Only Related scores explained very little in comparison (Adjusted $R^2 = 1.4\%$). This demonstrates that the predictive information relevant to priming is more effectively captured by these contrastive measures.

The same pattern was also evident in the initial model combining all six predictors. Both A.DifferencePrompt and DifferenceCos were significant positive predictors, and their correlations with the zRT_Priming_Effect were generally higher than their Only Related counterparts (e.g., $r = .17$ for A.DifferencePrompt vs. $r = .08$ for A.RelatedPrompt).

The Difference Score methodology closely mimics the experimental design of priming research as mentioned in Section 2.4. By subtracting the score of an unrelated control pair, the method helps to control for baseline effects the LLM might have toward the target word itself, isolating the unique contribution of the cue word. This results in a cleaner and more direct measure of the semantic contrast that appears to drive the priming effect.

4.2.2 Cosine Similarity vs. Prompt-based Similarity

Another notable outcome was that the judgments from the generative model gpt-4o-mini did not necessarily outperform the embedding model text-embedding-3-small or vice versa. This was most evident in the performance of DifferenceCos ($\beta = 0.32$) and A.DifferencePrompt ($\beta = 0.34$) with almost identical effect sizes. The near-identical performance of these predictors demonstrates that a well-applied difference score on static embeddings is just as powerful as a prompted judgment.

However, an alternative interpretation should be considered. Given that both models are developed by OpenAI and likely trained on overlapping data corpora, their strong similarity is perhaps not surprising. It is possible that text-embedding-3-small represents a version of the same underlying semantic knowledge base present in gpt-4o-mini. From this perspective, their similar performance reflects a shared knowledge source rather than a convergence of two different architectural approaches, highlighting that this comparison is limited to the OpenAI ecosystem.

In addition, both approaches also showed similar behavior regarding multicollinearity, because their direct relatedness scores (RelatedCos and A.RelatedPrompt) changed from negative to positive predictors in the separate analyses.

4.2.3 Feature-based vs. Association-based Predictors

The results consistently indicated stronger predictive performance for association-based predictors compared to feature-based predictors. In the main six-predictor model, the A.DifferencePrompt predictor was the best performing predictor, while the three feature-based predictors were either non-significant or showed only weak, negative trends. The disparity between the prompt-based predictors became even more clear in the separate regression analyses. Both Fb.DifferencePrompt ($\beta = -0.13, p < .001$) and Fb.RelatedPrompt ($\beta = -0.08, p = .024$) emerged as statistically significant negative predictors of priming. Several interpretations for this disparity are possible:

First, the experimental priming effects observed in the English SPAML dataset might be more driven by associative links between words than by pure semantic feature overlap. This would be consistent with the "associative boost" phenomenon discussed by Lucas where associative strength often plays a dominant role [Luc00]. If the priming effects in the SPAML dataset are similarly dominated by this associative component, it logically follows that our association-based LLM predictors would be more successful compared to the feature-based predictors. In addition, it is possible that the stimuli within the SPAML dataset, while categorized as related, might possess stronger associative links than clearly defined feature-based links.

Second, the association-based prompt provided to gpt-4o-mini might have been more effective in judgments that align with human priming mechanisms than the feature-based prompt. It is possible that an LLM find it challenging to define a feature overlap in a way that consistently reflects the cognition of humans. Feature-based relations require world knowledge and grounded concepts (such as texture or noise), which are not necessarily robustly present in a purely language-based model. The negative correlations of feature-based predictors may be a symptom of this. Therefore, the concept of co-occurrence in similar situations for association might be less challenging to capture.

Further research is needed to explore these possibilities. One approach could be to use stimuli that are specifically designed to have a more clearly contrast between associative relatedness and feature-based relatedness.

4.3 Limitations

Several limitations should be considered when interpreting the findings of this research.

First of all, the prompt-based scores are closely tied to the phrasing of our prompts and the capabilities of gpt-4o-mini. Using different prompts or switching to other larger LLMs made by different research organizations may lead to different similarity scores and different predictive outcomes. Designing prompts that show accurate and precise constructs like associative strength or feature overlap remain a complex challenge. Similarly, the cosine similarity scores are specific to the text-embedding-3-small model. Other embedding models might produce different similarity patterns.

Second, the current study is limited by the predictors chosen and the overall model performance. While our six predictors were theoretically motivated, other LLM-derived features or calculation methods might offer enhanced predictive power. Crucially, while the main six predictors model was statistically significant, it explained a relatively small portion (approximately 7%) of the variance in the priming effect. This highlights that human semantic priming is a complex cognitive phenomenon influenced by multiple factors.

Furthermore, our analytical approach of fitting separate regression models to address multicollinearity has its own limitations. While this method successfully shows the positive underlying relationships of the Only Related predictors, it presents a simplified view of the cognitive process. In reality, these different types of semantic relationships likely operate simultaneously rather than in isolation. The full six-predictor model, despite its own complexities, may therefore offer a more realistic representation of how these factors interact.

As earlier noted, another limitation is that the findings are based on only the English subset of the SPAML dataset. Although this dataset is large and robust, the specific properties, such as the nature of the stimuli, participant sample, and task design, may shape the observed effects. Therefore, the generalizability of these results to other priming paradigms or datasets cannot be guaranteed. In addition, it is important to consider that, although English is the native language for all participants in the research conducted by Buchanan et al., it is not the case for the English text data on which the LLM is trained [BCC+25]. Differences may influence the priming effects observed.

Finally, an important consideration is the disparity between the training data of LLMs and the language experience of human participants. Although the human participants in the SPAML English dataset were native English speakers, whose language processing is based on the diverse input of interactive experiences through senses, the LLMs (such as gpt-4o-mini and text-embedding-3-small) are mostly trained on vast corpora of text [BCC+25]. This internet-derived text may not perfectly reflect an individual human's development of language exposure or their experience of the world. Furthermore, the human language processing integrates visual input (e.g., facial expressions, gestures, text), auditory input (e.g., spoken language, environmental sounds), and sensorimotor interactions. In contrast, LLMs are primarily text-based, lacking direct access to these non-linguistic contextual cues that shape human understanding and representation of meaning. This fundamental difference in learning environments and input could significantly influence how well their derived similarity measures align with human cognitive phenomena.

4.4 Future Research

The current findings, despite their limitations, open several promising avenues for future research.

The current study was limited to only English word pairs. A crucial next step is to assess the cross-linguistic validity of these findings by applying the same methodology to other languages available in the SPAML datasets. It would reveal whether the superiority of Difference Scores and the challenges with feature-based prompts are universal or language-specific phenomena. Furthermore, future work should move beyond isolated word pairs and investigate priming in more challenging environments, such as within full sentences or dialogues. This would test whether LLMs can understand that the meaning of a word is flexible and depends on the context, such as how the word "bank" means something different in "river bank" versus "money bank" [CGA+23].

Another interesting addition to the current research would be to conduct a comparative analysis of predictors derived from various LLM architectures, such as Google or Meta. By testing other models from different developers, we could identify which types of models are most effective at capturing the nuances of human semantic priming.

Future research should explore more sophisticated statistical methods to handle multicollinearity. Techniques such as Principal Component Analysis (PCA), can be used to combine the information from our six predictors into a new smaller set of variables called principal components [Pau06]. The components with very low importance, as these are considered to be mostly statistical noise, will be removed, resulting in a more stable model.

5 Conclusion

This bachelor thesis investigated to what extent similarity predictions derived from Large Language Models (LLMs) can explain semantic priming effects in human lexical processing. The primary experimental measure was the zRT_Priming_Effect, reflecting human reaction time differences to related versus unrelated word pairs from the English SPAML dataset [BCC+25]. Six predictors were generated using OpenAI's text-embedding-3-small for cosine similarities and gpt-4o-mini for prompt-based similarities, considering both association-based and feature-based prompts of relatedness [Ope25]. In addition, they are calculated as either Only Related scores or Difference Scores.

Our primary multiple regression model, which combined all six predictors, successfully explained a statistically significant portion of the variance in human semantic priming (Adjusted $R^2 = 6.4\%$). This confirms that the information extracted from LLMs is relevant for understanding factors that influence human lexical processing and semantic processing. Three key conclusions can be drawn from this research:

First, the method of measurement proved more critical than the model architecture. The most powerful predictors were consistently the Difference Scores, which mimic the nature of priming experiments. The prompt-based A.DifferencePrompt and the embedding-based DifferenceCos were the two strongest predictors, demonstrating almost identical predictive power. This suggests that a well-designed, contrastive measure is just as powerful as a prompted judgment.

Secondly, our separate regression analyses successfully clarified the role of multicollinearity. Predictors measuring direct relatedness (e.g., RelatedCos), showed negative effects in the main model and became significant positive predictors when analyzed in isolation. This confirms their value, but also shows they become statistically redundant when their more powerful Difference Score counterparts are present.

Third, there was a clear superiority of association-based over feature-based predictors. The feature-based measures consistently failed to positively predict priming and even showed significant negative relationships in separate analyses, indicating a fundamental misalignment between the LLM's interpretation of features and the associative processes driving this priming task.

In conclusion, this research provides evidence that similarity predictions derived from LLMs, especially those with associative relatedness as a difference score via prompting, offer meaningful and powerful insights into human semantic priming. Although these computational measures cannot fully capture the complexity of human cognition, they serve as valuable tools for psycholinguistic research. These findings contribute to a better understanding of how meaning might be processed in both humans and computational models. Future work should continue to refine these methods, exploring new prompting strategies and statistical approaches to improve computational tools for linguistic research.

References

[BCC+25] Erin Michelle Buchanan, Kelly Cuccolo, Nicholas Alvaro Coles, Tom Heyman, Aishwarya Iyer, Neil Anthony Lewis, Kim Olivia Peters, Niels van Berkel, Anna Elisabeth van't Veer, Jack Edward Taylor, et al. Measuring the semantic priming effect across many languages. 2025.

[Bec12] Jason W Beckstead. Isolating and examining sources of suppression and multicollinearity in multiple linear regression. *Multivariate Behavioral Research*, 47(2):224–246, 2012.

[But22] Betty Butterfly. Optical Neural Networks: The Future of Deep Learning?, 10 2022.

[CGA+23] Giovanni Cassani, Fritz Günther, Giuseppe Attanasio, Federico Bianchi, and Marco Marelli. Meaning modulations and stability in large language models: An analysis of bert embeddings for psycholinguistic research. 2023.

[CL75] Allan M Collins and Elizabeth F Loftus. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407, 1975.

[DD24] Simon De Deyne. Evaluating human-like similarity biases at every scale in large language models: Evidence from remote and basic-level triads. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

[GVH+16] Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. SimVerb-3500: A large-scale evaluation set of verb similarity. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182, Austin, Texas, November 2016. Association for Computational Linguistics.

[HPL+24] Tom Heyman, Ekaterina Pronizius, Savannah C. Lewis, A. Acar Oguz, Matúš Adamkovič, et al. Crowdsourcing multiverse analyses to explore the impact of different data-processing andanalysis decisions: A tutorial. *Manuscript*, 2024.

[Luc00] Margery Lucas. Semantic priming without association: A meta-analytic review. *Psychonomic bulletin & review*, 7:618–630, 2000.

[MKB17] Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57–78, 2017.

[MP43] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

[NLB+24] Qian Niu, Junyu Liu, Ziqian Bi, Pohsun Feng, Benji Peng, Keyu Chen, Ming Li, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Tianyang Wang, Yunze Wang, Silin Chen, and Ming Liu. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. *arXiv preprint arXiv:2409.02387*, 2024.

[Ope25] OpenAI. gpt-4o-mini and text-embedding-3-small models. https://openai.com/, 2025.

[Pau06] Ranjit Kumar Paul. Multicollinearity: Causes, effects and remedies. *IASRI, new Delhi*, 1(1):58–65, 2006.

[R C23] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.

[RN16] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. pearson, 2016.

[Sha24] Murray Shanahan. Talking about large language models. *Communications of the ACM*, 67(2):68–79, 2024.

Appendix

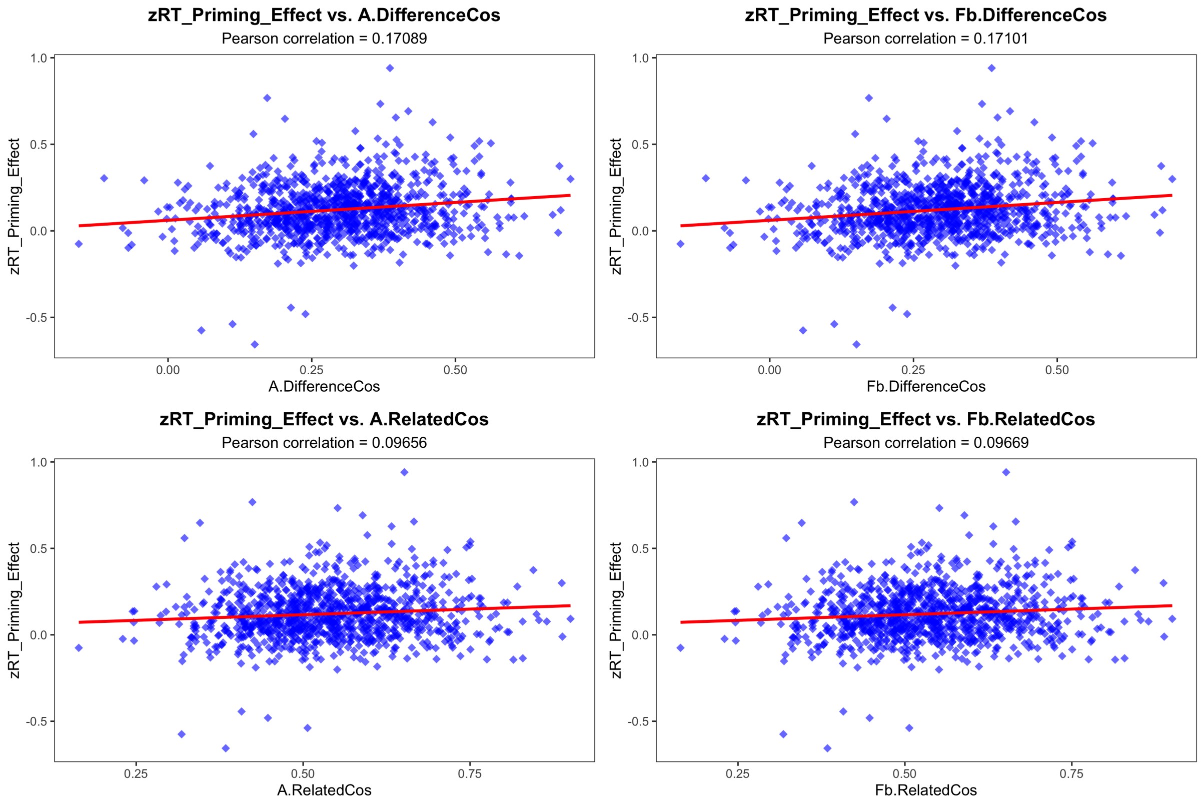

Figure 9: Four Cosine plots for both association-based and feature-based prompts for comparison.

Link to the code base: https://github.com/qiuvanleeuwen/Thesis-2025.git