Universiteit Leiden

Predicting Human Semantic Priming Using Multilingual Large Language Models

Nina Rojewska (s3908666)

Supervisors:
Tom Heyman & Evert van Nieuwenburg

Bachelor Thesis– Informatica

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 1, 2026

Abstract

Semantic priming is the psycholinguistic phenomenon where word recognition is faster when it is preceded by a semantically related word (e.g. related *cat-dog*, and unrelated *table-dog*). This concept provides an important insight into the organisation of the human mental lexicon and is widely used to evaluate semantic representation models. This thesis tests whether semantic representations derived from large language models (LLMs) can predict the size of item-level semantic priming effects in English and Polish using data from the Semantic Priming Across Many Languages (SPAML) project. Three types of LLM-based measures were evaluated: cosine similarity from subtitle-trained fastText embeddings, cosine similarity from two OpenAI embedding models (text-embedding-3-small and text-embedding-3-large), and GPT-4o-mini ratings of association strength and feature overlap. Their predictive value was assessed using nested linear regression models with 10-fold cross-validation while controlling for word frequency and length.

The results show that cosine similarity predictors consistently predicted item-level semantic priming in both languages, with fastText being the strongest single predictor, followed by the OpenAI large embedding. Both significantly predicted priming in both languages and performed better in English than in Polish, even after applying a correction for attenuation. This suggests that the differences in prediction accuracy may be the result of greater morphological complexity in Polish and different training data. GPT ratings of association strength and feature overlap did not sufficiently predict priming in either language. Therefore, these findings support the use of cosine similarity as a predictor of human semantic priming, while showing that prompt-based semantic ratings do not capture behavioural processes in the priming effect.

Contents

1 Introduction

1.1 The Situation

Large Language Models (LLMs) have become increasingly strong at capturing semantic patterns in human language, which raises the question of whether their internal representations reflect how people actually understand and process words. Models such as GPT are trained on massive text data and learn to process language by predicting semantic patterns, without being explicitly taught what words mean. The challenge is therefore not in showing that a model performs well on standard benchmarks like text classification, but in showing that it captures semantic knowledge in the same way human behaviors do.

Semantic priming is a standard and commonly used measurement for this problem. The effect shows that when a person sees a related word before a target word, they respond to the target word faster than if the preceding word was unrelated. This reaction is quick and automatic, and it reflects the structure of the human mental lexicon - how words and their meanings are connected in human memory. The size of the facilitation that happens varies depending on the word pairs, and understanding what causes these differences is still an open question.

However, if an embedding model can capture human semantic knowledge, then it should be able to predict the size of priming in different word pairs using its similarity score. While the majority of previous research in this area studies if related pairs produce more priming than unrelated ones on average, this thesis focuses on LLM-derived similarity measures and whether they are able to predict the priming effect size for specific word pairs. This is a stricter approach as a model must capture small differences in semantic relatedness rather than just general related/unrelated distinction.

Another important question is whether any of these patterns are shared across languages. Since most NLP is done in English, the question if models trained on mostly English data perform equally well in other languages is understudied. Therefore, this thesis compares English and Polish, two morphologically distinct languages. Because Polish words change forms based on the grammatical context, which might influence the accuracy of embedding models, this comparison would present useful insights into how embedding models generalise across languages.

For this purpose, three types of measures are evaluated: cosine similarity from a fastText model trained on subtitles, cosine similarity from two OpenAI embedding models of different sizes, and two semantic similarity ratings generated by GPT-4o-mini, measuring association strength and feature overlap. All of these are used to predict item-level semantic priming effects from the SPAML dataset (Buchanan et al., 2026), which collected data from participants for 1000 word pairs in both English and Polish.

1.2 Thesis Overview

This paper covers the background knowledge required to understand the problem, as well as a full review of the methods and findings of the current study. Section 2 provides background on the main concepts: semantic priming, the lexical decision task and automatic and controlled processes, distributional semantics (including fastText embeddings and cosine similarity), and semantic simi-

larity norms. Related work on large-scale priming databases, distributional models as predictors of psycholinguistic data, use of LLMs as cognitive models and cross-linguistic considerations will be reviewed in Section 3. Further, after stating the research questions (Section 4), the full methodology of this project, including the SPAML dataset, computation of cosine similarity, GPT ratings, and statistical analysis pipeline (with attenuation correction and cross-linguistic comparison on correlations and regression analysis), will be explained in Section 5. Finally, the reported results will be analysed and interpreted in Section 6 based on the correlation analysis, regression results, and assumption checks, and Sections 7 and 8 will discuss the findings and limitations, and draw final conclusions.

2 Background

2.1 Semantic Priming

Semantic priming is a psycholinguistic phenomenon where the target word (e.g. *cat*) is recognised faster when it is preceded by a semantically related prime word (e.g. *dog*) compared to when it is preceded by an unrelated word (e.g. *table*). The effect is defined as a faster response to related prime-target pairs, reflecting facilitated linguistic processing (Buchanan et al., 2026; Hutchison et al., 2008). The phenomenon, which was first presented by Meyer and Schvaneveldt (1971), is often used to investigate cognitive processing, word recognition, and the structure and organisation of semantic memory (Buchanan et al., 2026).

2.1.1 Lexical Decision Tasks

The most common method for measuring semantic priming is the *lexical decision task* (LDT), where participants see a letter string and have to decide whether it is a real word or not. Responses tend to be faster when the target is preceded by a related prime than an unrelated one (Lucas, 2000). A similar paradigm is the *naming task* (pronunciation), where participants have to say the word aloud. While both tasks can capture priming, they engage different mechanisms: naming reflects only forward-looking processes (prospective) like spreading activation and expectancy generation, whereas lexical decision additionally involves backward-looking process (retrospective) of semantic matching (Hutchison et al., 2013) all of which are discussed in more detail in the following sections. A type of lexical decision task that reduces expectancy is the *continuous decision task* (CLD), where participants judge every presented item (cue and target) for lexicality (word or nonword), which makes the prime-target relationship less obvious. Because of that, CLD is often used to test the automaticity of semantic priming. Supporting this idea, results of a meta-analysis by Lucas (2000) showed that semantic priming effects were comparable across different tasks, including LD and CLD, suggesting that a large part of semantic priming can reflect automatic processes.

2.1.2 Automatic vs. Controlled Processes

The distinction between *automatic* and *controlled* (attentional) processes has been formally defined by Posner and Snyder (1975) based on three criteria. Automaticity is a fast processing of stimulus without awareness, attention and intention. In semantic priming research, the most common way of separating these two types of processing is the *stimulus onset asynchrony* (SOA) between prime and target. Priming effects observed at short SOAs (typically 250-300 ms or less) reflect automatic processing, where priming at longer SOAs additionally involve controlled strategies like conscious expectancy and retrospective semantic matching (Hutchison et al., 2013; Lucas, 2000; Neely, 1991). Automatic priming is thought to reflect the stable organisation of the mental lexicon, while controlled processes are more influenced by individual and task differences (Heyman et al., 2016). Within this framework, Neely and Keefe (1989) distinguish three main processes where priming can arise: *expectancy generation, semantic matching* and *spreading activation*, each of which is discussed in more detail in the following sections.

Expectancy Generation

According to Neely (1991), *expectancy generation* is a slower and controlled process in which participants consciously generate a set of potential target words based on the prime words, e.g. after seeing a prime *cat*, participants expect to see *dog*, *pet*, *meow*, *milk* etc. If the actual target is among the expected candidates, word recognition is facilitated. The probability of a target appearing in the set of expected candidates is related to the association strength between prime and target, as words with stronger associations are more likely to be generated as expectations. Due to the time necessary for the generation, the process requires longer SOAs in order to have an effect and is influenced by individual differences in mental capacity and conditions. Moreover, expectancy generation is one of the prospective (forward-looking) processes, as it is triggered by the prime, and can be used in both lexical decision and naming tasks.

Semantic Matching

Semantic matching is another controlled, but this time retrospective (backward-looking) process described by Neely (1991). Opposite to expectancy generation, it is triggered after processing the target word, not the prime, when the participants check semantic relatedness between the target and the preceding prime. Based on the decision, the target is more likely considered a word if the pair is semantically related, and a nonword if it is unrelated. As a controlled process, semantic matching usually occurs at longer SOAs because participants need time to process the prime before making a decision about the prime-target relation. It is mostly used in the lexical decision tasks, as backtracking to the prime in the naming task would provide no additional information on target pronunciation. Specifically, the continuous lexical decision task is used to reduce or eliminate these controlled strategies, as both prime and target require a response, which makes the prime-target coupling less evident (Buchanan et al., 2026).

Spreading Activation

Unlike expectancy generation and semantic matching, *spreading activation* is an automatic process as it operates quickly, unconsciously, and its facilitation decreases over time. Because it does not require conscious attention, spreading activation operates on short SOAs, and it is considered a prospective process as it is triggered by the prime. As a forward-looking process, it is suitable for both lexical decision and naming tasks. This process is rooted in the fundamental theory proposed by Collins and Loftus (1975) - *spreading activation mechanism*. They claim that semantic memory is organised as a network of concept nodes connected by links of different strength (where shorter hence stronger links represent greater semantic relatedness), and processing of a concept automatically activates the network spreading it in parallel across paths. The strength of this spread decreases with distance and decays over time, meaning that concepts sharing more properties receive stronger cumulative activation than distant ones (Collins and Loftus, 1975). When multiple sources activate the same node simultaneously, their activation sums up until it reaches an activation threshold, which allows the model to capture complex semantic relations (Collins and Loftus, 1975). This approach is especially important for predicting semantic priming as the processed prime (e.g. *cat*) activates its node, which spreads to related nodes (e.g. *dog, pet*). When the target word *dog* is presented shortly after, it requires less processing to reach the recognition threshold, as it was already partially activated by the prime word *cat* (Heyman et al., 2016). A

related idea is represented by the feature overlap models, where word meaning is represented as a pattern of activation over semantic features. This means that because related words share many features, when a prime appears, it activates features overlapping with the target, which produces facilitation through feature similarity (see more in Section 2.3).

While spreading activation explains how a prime facilitates target recognition, it does not explain where the connections come from and how to measure the strength of the semantic relation between words without relying on human judgments. An alternative approach based on the distributional representations that addresses it directly is introduced in the next section.

2.2 Distributional Semantics and Word Embeddings

2.2.1 Distributional Hypothesis

Distributional semantics offers such an alternative, relying on the *distributional hypothesis* which states that the meaning of a word can be determined from the words that commonly occur in similar semantic contexts (Lenci, 2018; Turney and Pantel, 2010). To capture this, co-occurrence patterns between words are extracted from large text corpora and mapped to numerical vectors in high-dimensional semantic space. This means that words which appear in similar contexts receive similar vector values and are assumed to have similar meanings, which has been shown to successfully predict semantic priming at the item-level (Mandera et al., 2017).

The early distributional models used count-based methods, where they built large frequency matrices from the corpus and applied dimensional reduction to extract the semantic patterns. The most influential in this area was *Latent Semantic Analysis* (LSA) which used *Singular Value Decomposition* (SVD) to reduce the original high-dimensional word document matrix to a smaller semantic space of around 300 dimensions (Turney and Pantel, 2010; Buchanan et al., 2019; Hutchison et al., 2008). The high performance of LSA on synonym tests was comparable to human results and became widely used in psycholinguistic research.

The modern prediction-based models introduced a different approach. Instead of relying on co-occurrence frequencies and matrix factorisation, they use a neural network to learn vector representations by predicting a word from its context (*CBOW*) or predicting context from the target word (*skip-gram*) (Mandera et al., 2017; Mikolov et al., 2013). Baroni et al. (2014) show that prediction-based models generally outperform count-based models in semantic tasks, as even when properly optimized, count-based models can reach results only comparable to the prediction-based ones (Mandera et al., 2017).

2.2.2 fastText and subs2vec

Building on the prediction-based apprach, *fastText embeddings* (Bojanowski et al., 2017) extended the skip-gram models by representing words through their character n-grams rather than treating each word as a single unit. This allows a single word e.g. *cats* to be stored not by a single vector, but by the sum of partially contributing vectors of the word (such as *ca, cat, ats*, and *ts*) along the full form.

This approach has two advantages important for the cross-linguistic comparison. First, it handles morphologically rich languages, where the inflected forms (e.g. in Polish *cat* can be *kot, kotu,*

kota etc.) receive related vector representations, even when they are not equally represented in the training corpus. Second, fastText can create vectors for out-of-vocabulary words that were missing from the training by combining the vectors of their character n-grams (Bojanowski et al., 2017). This makes fastText especially suitable for Polish, which has more complex word inflection than English.

The fastText embeddings used in the SPAML study (Buchanan et al., 2026) come from the *subs2vec project* (Van Paridon and Thompson, 2021), where the word vectors were trained on the *OpenSubtitles corpus* in 55 languages. The OpenSubtitles is a large multilingual collection of movie and television subtitles from the opensubtitles.org website, and due to the breadth of conversational language patterns it captures, it has become one of the most widely used corpora in NLP and psycholinguistic research (Lison and Tiedemann, 2016). Research has shown that subtitle-based embeddings predict lexical and psycholinguistic measures better than embeddings trained on formal written text, because they reflect everyday spoken language more closely (Mandera et al., 2017; Van Paridon and Thompson, 2021). For this reason, the SPAML project used subs2vec fastText embeddings to match prime-target pairs across languages.

2.2.3 Cosine Similarity

Once words are represented as vectors, their semantic relatedness can be measured. The standard method for this is calculating their *cosine similarity*. The cosine of the angle between two vectors in the high-dimensional space ranges from 1 for vectors pointing in the same direction (highly related words), to 0 for orthogonal vectors (semantically unrelated words), and to -1 for opposite directions (Turney and Pantel, 2010):

$$\textit{Cosine Similarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{1}$$

The greatest advantage of this metric is that it is insensitive to vector length and depends only on their directions. This means that the two high-frequency words, such as *the* and *an* (long vectors), will not be considered semantically related as they are pointing in different directions (Turney and Pantel, 2010). This makes cosine a suitable measure of pure semantic similarity, independent of word frequency and other non-semantic factors.

In the context of semantic priming, one hypothesis is that cosine similarity may serve as a proxy for the structure that spreading activation operates on. Word pairs with higher cosine similarity are assumed to produce stronger pre-activation of the target word, and therefore larger priming effect. Findings of Mandera et al. (2017) confirmed that adding cosine similarity from distributional models improves the prediction of reaction times in both lexical decision and naming tasks. Altogether, these findings suggest that distributional similarity can partially reflect human priming, which provides evidence for using distributional embeddings to predict item-level effect.

2.3 Semantic Similarity Norms

While cosine similarity captures relatadness from text, semantic similarity can also be measured from human-generated norms. In free association tasks, participants are presented with a cue word and asked to name the first word that comes to their mind. The proportion of people

giving a certain answer defines the *forward association strength* (FAS) from cue to target, and *backward association strength* (BAS) for the reversed proportion (Hutchison et al., 2008). The norms originally established by the University of South Florida (USF) contained association data for more than 72000 word pairs based on the 5019 cue words (Nelson et al., 2004). These were more recently expanded by the *Small World of Words project* (De Deyne et al., 2019) to over 12000 cue words. These advances have consistently shown that association strength can be used to predict semantic priming (Mandera et al., 2017; Hutchison et al., 2008).

A complementary approach uses semantic feature norms, where participants are asked to list the properties or features that describe a concept, such as *animal, pet, tail, meow* for the word *cat*. This way, semantic similarity between two words is measured as a feature overlap, so the number of features they share (Buchanan et al., 2019). Unlike association norms, which measure whether two words are related, feature norms capture why they are related by identifying their common properties. Research shows that association norms and feature norms capture partially different aspects of semantic priming, as they overlap only modestly at approximately 37%, which confirms differences in the measured semantic relations and their ability to predict semantic priming (Buchanan et al., 2019).

One of the main questions in the semantic priming literature relates to what drives priming more - association strength or feature overlap. Lucas (2000) meta-analysis using purely semantically related prime-target pairs while controlling for association, showed that semantic priming can occur automatically and without association at different SOAs and task types. This suggested that semantic relationships are part of the mental lexicon and are automatically accessed during word recognition. Lucas additionally identified "an associative boost" where adding associative relationship to an existing semantic one doubled the effect size, suggesting that both mechanisms contribute to the priming independently. Hutchison (2003), on the other hand, took a more critical approach by questioning the association purity of the stimulus pairs in Lucas (2000). After more rigorous control for association, he observed noticeably weaker relationships, specifically among the items sharing a category (e.g. *horse-deer*). While acknowledging, consistent with Lucas (2000), that some feature overlap may produce priming independently from association, Hutchison concluded that association remains a more robust predictor across tasks and stimulus type. Overall, neither of these semantic similarity measures is sufficient for priming alone. While association strength is a reliable predictor of priming, feature overlap contributes independently, capturing additional information, especially when association strength is weak. This distinction motivates the use of both types of relatedness measures in addition to the cosine similarity measures introduced in Section 2.2.3.

Beyond the implicit distributional patterns captured by cosine similarity, LLMs can also be used to derive semantic similarity estimates by prompting them for explicit judgments about word pairs. This offers a practical advantage over human-generated norms, which require large-scale datasets from thousands of participants (such as the USF and Small World of Words discussed above), whereas prompt-based ratings can be obtained directly from the model. Another advantage of using prompt-based ratings compared to embedding-based is the possibility to access distinct types of semantic relatedness, such as association strength and feature overlap, allowing each mechanism to be tested independently. De Deyne (2024) supported this by showing that GPT-4 can capture human-like semantic similarity judgments with high accuracy, which motivates using this approach

to address the debate about which measurement drives priming more.

Having introduced the theoretical foundations of semantic priming and its measures, the following section reviews the empirical work that applies these approaches.

3 Related Work

3.1 Large-Scale Priming Databases and Item-Level Analysis

Traditional semantic priming studies relied on factorial design, which compared overall reaction time distributions between related and unrelated pairs, typically by averaging across all items within a condition. This approach required manual selection and matching of stimuli which introduced a bias, as well as treating continuous variables like association strength and word frequency like discrete categories, which could obscure the underlying relations. These limitations, recognised by Hutchison et al. (2008) (see Section 3.2), motivated them to address this issue in the *Semantic Priming Project* (SPP) (Hutchison et al., 2013), a large-scale behavioural database that provided reliable priming estimates for individual word pairs. The SPP used four prime types: first-associate related, first-associate unrelated, other-associate related, other-associate unrelated for 1661 target words in both LDT and naming task at short (200 ms) and long (1200ms) SOAs. This allowed researchers to use multiple regression analysis to predict item-level priming from continuous measures of prime-target relatedness, instead of relying on categorical comparisons.

Building on that, Heyman et al. (2016) directly correlated item-level priming effects across tasks. The findings from calculating a separate priming effect for each target word showed significant correlations at the short SOA, with strong Bayesian evidence for shared automatic mechanisms. In contrast, the long SOA correlations were much weaker, failing to reach significance and showing data more consistent with the null hypothesis, which suggests that strategic processes add individual differences to priming effects. Even after restricting analysis to forward associates only, the same pattern held, which supports fast automatic processes like spreading activation instead of retrospective strategies.

While these findings confirm that item-level effects can capture automatic semantic activation, the SPP project was limited to a single set of English word pairs and faced item-level reliability concerns, with reported low split-half reliability estimates (Buchanan et al., 2026). Similarly, Hutchison et al. (2008) noticed that small sample sizes per item introduced measurement noise, although item-level prediction was still possible with a smaller dataset. Therefore, Buchanan et al. (2026) developed the *Semantic Priming Across Many Languages* (SPAML) project to test whether relationships hold across languages and at a greater scale. By collecting item-level estimates from over 25000 participants across 19 languages in a CLD task, SPAML enabled cross-linguistic comparison, while addressing the reliability concerns that limited earlier work.

3.2 Distributional Models as Predictors of Psycholinguistic Data

Beyond building datasets, research has a long history of asking which measures actually predict priming. The first important study on predicting item-level semantic priming was conducted by Hutchison et al. (2008), where they used multiple regression analysis to predict priming effects for 300 prime-target pairs in LD and naming tasks at both short and long SOAs. The analysis used lexical properties of the prime and target, such as word frequency and length, as a baseline, and three correlation measures: forward association strength (FAS), backward association strength (BAS) and similarity based on Latent Semantic Analysis (LSA). The results showed that FAS was predicting priming across both tasks and SOAs, BAS predicted priming at both tasks for longer

SOAs, but LSA similarity failed to predict priming in any condition. This led to the conclusion that distributional models based on contextual similarity have limited abilities to capture behavioural priming processes, which was later addressed by Mandera et al. (2017) based on the data from the SPP (Hutchison et al., 2013).

In this research, a larger dataset operating on 1661 targets compared count and prediction-based models, including LSA, CBOW, skip-gram, and fastText. As reported, adding distributional cosine similarity as a predictor significantly improved the variance explained in LDT (Mandera et al., 2017). Additionally, distributional models performed equally or better than human-generated association and feature norms, which led to the conclusion that distributional semantics can successfully explain semantic priming data.

3.3 LLMs as Cognitive Models

While distributional models such as fastText and word2vec have been successful in predicting psycholinguistic data, they use single-vector word representations and do not account for changes in meaning based on contexts. Recent work has shifted to LLMs to study whether contextualised representations can better capture how people process word meaning.

Cassani et al. (2023) compared contextualised (*BERT*) embeddings with static (*word2vec*) embeddings on both explicit semantic similarity judgments and implicit semantic priming. They found that while average representation of contextualised embeddings correlated stronger with explicit similarity ratings (above .5), outperforming static embeddings (around .3), the static models provided a better fit for implicit priming effects. This suggests that automatic semantic activation may rely on simpler, context-free representation, while conscious similarity judgments benefit more from understanding the context.

De Deyne (2024) extended this comparison to GPT-4 and showed its wide capabilities in capturing both closely related concepts and weakly related pairs in triads. By using the semantic triad task, the model had to identify which two of the three presented words were most similar to each other, across different levels of semantic relatedness. GPT-4 captured human similarity judgments well by recognising weak relations with an .84 correlation for remote triads, outperforming earlier word embedding models. Importantly, GPT-4 performed equally well for weakly and strongly related pairs, suggesting that detailed semantic distinctions are captured well in more recent, larger language models, unlike older, distributional models, which struggle with weak semantic relations. This finding supports the idea that LLMs can be used as suitable models of human semantic organisation.

3.4 Cross-Linguistic Considerations

Like most of the work discussed so far, research on semantic priming has been conducted mostly in English and Western populations, making it hard to recognise whether the observed patterns are universal or language-specific. Psychology often refers to this as the *WEIRD (Western, Educated, Industrialised, Rich, Democratic)* problem, and to address it, comparable baseline norms must be used. For English, the *SUBTLEX-US* database (Brysbaert and New, 2009) provides reliable word frequency trained on the subtitles corpora, and for Polish, the equivalent *SUBTLEX-PL* (Mandera

et al., 2015) offers the same. Using these norms ensures comparable control for frequency as a baseline predictor in both languages, as word frequency can predict item-level priming effects beyond semantic relatedness (Hutchison et al., 2008).

The SPAML project (Buchanan et al., 2026) also addressed the WEIRD problem by collecting data across 19 languages using a controlled set of translated stimuli to allow cross-linguistic analysis of priming patterns in different languages. This project used computational word embeddings to select the stimulus set based on semantic similarity across languages instead of translating a single-language set, which ensured that stimuli capture comparable semantic relations in each language.

4 Current study

Building on this work, the present study used the SPAML dataset (Buchanan et al., 2026) to test whether LLM-derived semantic representations can predict item-level priming effects in English and Polish. Three types of predictors were evaluated: cosine similarity from fastText embeddings trained on subtitle data, used as the baseline given the established success in predicting priming (Mandera et al., 2017; Van Paridon and Thompson, 2021); cosine similarity from two OpenAI embedding models of different sizes, to test whether more modern models trained on larger corpora outperform the simpler fastText; and prompt-based GPT-4o-mini ratings of association strength and feature overlap, motivated by the theoretical debate on which of the two mechanism drives priming more (Lucas, 2000; Hutchison, 2003) and the evidence that LLMs can capture human-like semantic judgments (De Deyne, 2024). Word frequency and word length were included as control variables, as both have been shown to independently predict item-level priming effects (Hutchison et al., 2008). Based on this, four research questions were established.

RQ1. Can distributional embeddings (fastText, text-embedding-3-small and text-embedding-3-large) predict item-level semantic priming effects in English and Polish?

RQ2. How do OpenAI embedding models perform compared to fastText cosine similarity in predicting semantic priming?

RQ3. Do prompt-based GPT-4o-mini ratings of association strength and feature overlap provide additional predictive information beyond distributional cosine similarity?

RQ4. Are the results consistent across English and Polish, and do they share cross-linguistic patterns in predicting priming effects?

5 Methods

5.1 Data

5.1.1 SPAML Dataset

This study used data from the *Semantic Priming Across Many Languages* (SPAML) project (Buchanan et al., 2026), a large-scale multilingual psycholinguistic dataset designed to investigate semantic priming effects across multiple languages. The original study collected data from 25163 participants across 19 languages and was developed in English using 1000 target words, then translated into additional languages, including Polish, to enable cross-linguistic comparison.

Participants completed a continuous lexical decision task (CLD) in which both the prime and the target were presented sequentially, and participants decided whether each item was a real word or not. Each target appeared in two conditions: following a semantically related prime (e.g. *cat-dog*) or an unrelated prime (e.g. *table-dog*), with each participant seeing the target in only one condition. The English dataset included 5122 participants, and the Polish dataset included 1188 participants in the analysis. Semantic facilitation effects reported were measured through differences in reaction times between these conditions.

5.1.2 English and Polish Datasets

English and Polish datasets were processed separately while following the same procedures and analysis pipeline. Separate analyses were necessary because each language relied on independent lexical information, embedding vocabularies, and frequency norms.

Specifically, the main variables - cosine similarity and word frequency - differed across languages because they were derived from separate English and Polish embedding models and lexical databases. Processing the datasets independently ensured that all semantic representations and control variables remained consistent within each language before cross-linguistic comparison.

5.1.3 Filtering

Raw reaction time data were obtained from the SPAML dataset for both English and Polish. The trials were filtered using three quality flags labeled *keep_target, keep_participant, and keep_participant_answer* in the dataset, established by Buchanan et al. (2026) as part of the original SPAML data quality procedure. *Keep_target* flagged individual trials that should be excluded because of incorrect responses, timeouts or reaction times below 160ms. *Keep_participant* checked whether participants met the criteria such as age and completing at least 100 trials within their session (a single sitting in which the participant completed the task) with accuracy over 80% or higher. Accuracy here was calculated as the number of correct responses divided by the total number of trials. *Keep_participant_answered* followed the same criteria, but the calculated accuracy used only the trials to which the participant responded, excluding timeouts for an increased participant sample. While the original SPAML analyses apply these filters separately, depending on the specific analysis, the current study required all three conditions to be satisfied at the same time for a trial to be retained.

This approach in the filtering procedure was adopted to improve data quality and consistency

across all analyses. The resulting dataset provided a reliable foundation for analysis, ensuring that the computed priming effects were based only on observations meeting both participant and item quality requirements.

5.1.4 Priming Effect Computation

Priming effects were computed using *Z-scored reaction times* (RT) from the SPAML dataset. While processing the original data, raw RTs for each participant were standardised within their own session to control for individual differences in overall response speed. For each target word, the priming effect was computed by averaging the z-scored RTs across participants separately for the related and unrelated conditions, and taking the difference between the condition means:

$$Priming\ Effect = zRT_{\text{unrelated}} - zRT_{\text{related}} \tag{2}$$

Therefore, positive values indicate semantic facilitation, reflecting faster recognition when targets were preceded by semantically related primes compared to unrelated primes (Buchanan et al., 2026).

Because item-level priming predictions are averaged across multiple participants, they contain measurement error. This is addressed by the *split-half reliability* with $r\_yy = .719$ for English and $r\_yy = .524$ for Polish (Buchanan et al., 2026) and used to correct for attenuation in the reported correlations (see more in Sections 6.1.5 and 6.2.4).

5.1.5 Subset Matching

The original SPAML study was designed on 1000 unique English target words, which were later translated into 18 additional languages. To enable cross-linguistic analysis and direct comparison between English and Polish semantic priming effects, the *pl_matched.csv* file provided as part of the SPAML materials was used to map translation pairs between target words across languages.

The prime-trial datasets contain multiple observations for each target word, reflecting responses collected from multiple participants. After aggregating the reaction times into item-level priming effects, the matching file was used to map English and Polish target words based on their direct translations. Aligning both datasets to the same item order ensured that cross-linguistic analyses were performed on directly corresponding word pairs across languages, resulting in a final sample of n = 1000 triads (related and unrelated primes, target word) per language.

5.1.6 Control Variables

Word frequency and word length were used as control variables due to their established relationship with item-level priming effects. Low-frequency targets are supposed to show greater priming than high-frequency targets, as words that are slower to recognise benefit more from activation triggered by the prime (Hutchison et al., 2008). Similarly, longer words with slower response times tend to produce larger priming effects. By controlling for both variables, we ensure that the observed effects reflect semantic relatedness instead of differences in lexical accessibility. Word length was calculated as the number of characters in each target word. Frequency was obtained from the

SUBTLEX-US database for English (Brysbaert and New, 2009) and SUBTLEX-PL for Polish (Mandera et al., 2015). Both resources are subtitle-based frequency norms providing logarithmic frequency estimates.

5.1.7 Missing Data

English dataset In the English dataset, one target word (*hiv*) was absent from SUBTLEX-US, resulting in a missing value for the word frequency control variable. No cosine similarity values were missing, indicating full vocabulary coverage for the English word pairs. The final English dataset contained n = 999 valid rows.

Polish dataset In the Polish dataset, 64 items (6.4% of the dataset) had missing cosine similarity values. The gaps occurred when at least one of the three words within the item - the target word, the related cue, or the unrelated cue - had no embedding value.

The main cause was multi-word expressions: 23 target words and 40 cue words (related or unrelated) were phrases (e.g, *klatka piersiowa - chest*, *piłka nożna - football*, *wschód słońca -sunrise*). Single-token embedding models cannot represent multi-word expressions because phrases are not encoded as units in the vocabulary. The same missing cosine values were already present in the original SPAML Polish dataset, confirming that the issue originated in the source study stimuli rather than the present implementation. Furthermore, a subset of 24 Polish target words (all multi-word expressions) was missing from the SUBTLEX-PL frequency norms, as subtitle-based frequency databases generally do not include phrase entries.

A smaller number of missing cases involved single words absent from the embedding vocabulary due to misspelling e.g. *grzebietowy* instead of *grzbietowy - dorsal* or rare, archaic forms such as *pierzchanie*. However, most of these cases overlapped with items already affected by the presence of the phrase.

All rows containing missing values (cosine similarity and/or word frequency) were excluded from analyses, resulting in a final Polish sample of n = 936 valid items.

Starting from 999 valid English items and 936 valid Polish items, the intersection of complete cases across both matched datasets resulted in n = 935 items used in all subsequent analyses.

5.1.8 Preprocessing

During preprocessing, lowercasing and whitespace trimming were applied to all textual columns before joining the datasets and frequency. These normalisation steps were necessary as inconsistent formatting could result in failed matches.

The preprocessing revealed that a small number of Polish cue words in the SPAML trial data contained trailing punctuation characters. These formatting errors were likely introduced during initial translation and resulted in missing values. To address this issue, trailing punctuation was removed from all target and cue word columns before analysis.

5.2 Semantic Representations

5.2.1 Cosine Similarity Effect Computation

Cosine similarity measures the semantic relatedness between words based on vector representations generated by embedding models. Each word is represented as a numerical vector in a multidimensional semantic space. Cosine similarity is computed as the cosine of the angle between two word vectors (1), producing values between -1 and 1. Higher cosine similarity values reflect stronger semantic relatedness between words, which was previously associated with stronger semantic priming effects (Mandera et al., 2017).

For each target word, cosine similarity values are calculated separately for the related and unrelated word pairs. The cosine similarity effect is the difference between the two conditions, meaning higher values indicate a larger semantic similarity difference between related and unrelated word pairs.

$$Cosine\ Effect = Cosine_{\text{related}} - Cosine_{\text{unrelated}} \tag{3}$$

In this study, three separate sets of cosine similarity values are used as predictors, including fastText embeddings from the subs2vec project and two OpenAI embedding models of different sizes.

fastText embeddings. The *fastText* cosine similarity values were obtained directly from the SPAML dataset (*en_words.csv* and *pl_words.csv*). The original SPAML study uses fastText word vectors from the subs2vec project (Van Paridon and Thompson, 2021), trained on the OpenSubtitles corpus. In the stimulus construction of SPAML, related prime-target pairs were selected based on cosine similarity, while unrelated pairs were created through random word shuffling to minimise semantic similarity.

OpenAI embeddings. Word vectors were additionally collected via the OpenAI embeddings API. Initially, *text-embedding-3-small* (with 1536 dimensions in word vectors) was used as the OpenAI model. To investigate whether any observed difference in performance reflected limitations specific to that model or the OpenAI embeddings in general, the analysis was extended to include *text-embedding-3-large* (3072 dimensional vectors) as well (OpenAI, 2024). For each prime-target pair and both models, cosine similarity was computed separately for the related and unrelated conditions, and the cosine similarity effect was defined as the difference between these two values, following the same procedure as applied to fastText embeddings.

5.2.2 GPT Semantic Relatedness Ratings

Ratings extraction

Additional semantic ratings were collected using GPT-4o-mini through the OpenAI API. The model was prompted to evaluate each prime-target pair in two categories: *association strength* and *feature overlap* (see Figure 1). Previous research has shown that large language models can capture human-like semantic similarity judgments comparable to human semantic association patterns (De Deyne, 2024). Association measured how strongly one word brings another to mind, while feature overlap measured the number of shared semantic properties or characteristics. The

ratings were collected for related and unrelated pairs in each category on a 0-100 scale using temperature = 0 for increased output consistency. Each rating extraction was performed twice on both language datasets to assess rating stability. This resulted in 1000 word pairs x 2 GPT rating types (association strength/ feature overlap) x 2 conditions (related/ unrelated) x 2 runs = 8000 API calls per language.

Association Strength

System: You are an assistant that rates word association strength on a 0–100 scale.

User: Rate how strongly the word `[word1]` brings the word `[word2]` to mind, from 0 (no association) to 100 (very strong). Return only a single integer.

Feature Overlap

System: You are an assistant that rates semantic feature overlap between words on a 0–100 scale.

User: Rate how many features the words `[word1]` and `[word2]` share, from 0 (none) to 100 (identical). Return only a single integer.

Figure 1: Prompts used to collect GPT-4o-mini association strength (top) and feature overlap (bottom) ratings.

Variable construction

Semantic relatedness effect scores were computed as the difference between the unrelated condition rating and the related condition rating. Separate effect scores were calculated for association strength and feature overlap:

$$\textit{Association Effect} = \textit{Association}_{\text{related}} - \textit{Association}_{\text{unrelated}} \tag{4}$$

$$\textit{Feature Effect} = \textit{Feature}_{\text{related}} - \textit{Feature}_{\text{unrelated}} \tag{5}$$

The original GPT ratings were collected on a 0-100 scale and later rescaled to a 0–1 range before regression analysis by dividing all values by 100. The initial 0-100 scale was used to improve consistent numerical ratings while maintaining more detailed semantic distinctions between word pairs. Rescaling was applied to improve the interpretability of regression coefficients, as coefficients in the original scale would represent the effect of a one-point increase on a 100-point scale, resulting in very small estimates. After conversion to a 0-1 range, each regression coefficient represented the effect in the full range of the semantic measure, allowing more meaningful comparisons between predictors.

Stability was assessed by computing Pearson correlations separately for related and unrelated pair ratings across two independent runs and prompt types, before computing difference scores. High

correlations were observed across both languages and rating types, confirming that the model produced stable outputs at temperature=0. Based on these results, ratings from both runs were averaged before computing the final semantic effects and used in all analyses (see Section 6.1.1).

5.3 Statistical Analysis

All statistical analyses were conducted on the same matched subset of 935 items in both languages. This number was determined mainly by the Polish dataset, where 64 items contained missing cosine similarity values because of the multi-word expressions or missing vocabulary, and one additional item from the English dataset with a missing frequency value. As a result, the same corresponding 935 word pairs were selected in the Polish and English datasets, ensuring direct comparability of complete cases across the languages.

5.3.1 Correlation Analysis

Pearson correlations were computed as a first step to explore the reliability and validity of all semantic predictors and the relationships between them, before the main regression analysis (described in Section 5.3.2) were conducted. All analyses were performed separately for English and Polish, except for the cross-linguistic comparison, which directly compared the priming effects across the same matched rows in the two datasets.

Four sets of correlational analyses were conducted:

1. **GPT stability:** As mentioned in Section 5.2.2, correlation between run 1 and 2 ratings was computed for all GPT measures (association-related, association-unrelated, feature-related, feature-unrelated) to assess the reliability of GPT outputs. Assuming high stability across the runs, all subsequent analyses were performed on averaged difference scores. A similar stability check was not necessary for the cosine similarity embedding measures, as they are calculated from fixed vector values, so they produce identical values across runs. Reliability of priming effects was not estimated in this study, as Buchanan et al. (2026) reports them for both languages (see section 6.1.5).

2. **Convergent validity:** Correlations were calculated within and across measure families: between the three cosine similarity embedding models (fastext, OpenAI small, OpenAI large), between GPT-derived scores (association effect and feature effect), and between all embedding-based and GPT-based measures. Altogether, these correlations assessed how different semantic measures capture similar semantic relationships.

3. **Predictive validity:** Evaluated relatedness of each semantic measure (both cosine similarity and GPT-derived values) to semantic priming effects before the regression analysis. This helped determine whether higher semantic similarity was generally associated with stronger semantic priming effects.

4. **Cross-linguistic correlation:** The relationship between item-level priming effects across languages was assessed by directly correlating matched English and Polish priming effects for the same 935 word pairs. This analysis examined whether priming effect patterns are shared across languages, providing context to interpret the subsequent regression analyses.

5.3.2 Regression Analysis

The purpose of performing a set of linear regression analyses was to determine whether semantic similarity measures derived from LLMs could independently predict the human semantic priming effect. The dependent variable was the item-level priming effect, and the analysis was conducted separately on English and Polish data (n=935) using $\alpha = .05$ as the significance threshold.[1]

The process consisted of two steps:

Step 1: Cosine predictors. The predictive contribution of cosine similarity was evaluated by comparing three cosine similarity models with a baseline model containing only the control variables - word frequency and word length (Model 0). Model 1 used the fastText cosine effect from the SPAML dataset, Model 2 used cosine similarity values from text-embedding-3-small (OAI small), and Model 3 used the text-embedding-3-large (OAI large). Model 4 combined all three cosine predictors to evaluate whether the different embedding models explain overlapping or independent variance in semantic priming.

Step 2: GPT measures. Model 5 evaluated the GPT-derived association and feature effects without any cosine predictors. Model 6 combined fastText cosine similarity with GPT measures, while Model 7 combined the best-performing OpenAI embedding predictor with GPT measures. By comparing Model 6 and Model 7 against the cosine-only baselines (Model 1 and Model 3), the additional predictive contribution of GPT ratings outside the cosine similarity were captured.

5.3.3 Model Comparison

Model performance was evaluated using three measures:

- R^2: the proportion of variance in priming effects explained by the model
- Adjusted R^2: R^2 corrected for the number of predictors, penalising for unnecessary model complexity
- CV-R^2: estimated using 10-fold cross-validation to evaluate the model's performance on the unseen data. In 10-fold cross-validation, the dataset is randomly divided into 10 equal folds, then the model is trained on 9 folds and tested on the remaining one. This is repeated until each fold is used as the test set once, and the average R^2 from the 10 test folds is reported, as it provides a less biased estimate of out-of-sample performance than the in-sample R^2.

To determine the additional contribution of semantic predictors beyond the control variables, ΔR^2 was calculated as the difference between the full model and the baseline control model:

$$\Delta R^2 = R^2_{\text{full}} - R^2_{\text{baseline}} \tag{6}$$

[1]Because a large number of correlation and regression analysis were performed, the overall probability of Type I error across this study is greater than normal $\alpha = .05$. A stricter correction such as Bonferroni correction could have been applied, but given the exploratory nature of this study, it was not selected. Therefore, results that are close to the .05 significance threshold should be carefully interpreted.

The statistical significance of each ΔR^2 was tested using incremental F-tests with nested model comparison in R. Additionally, cross-linguistic comparison on separately fitted regression models was performed.

5.3.4 Assumption Checks

To validate the models, five assumptions were checked to ensure that the linear models were appropriate for the data:

- **Linearity:** scatterplots of each predictor against the priming effect were inspected with a regression line to detect any systematic nonlinear patterns.
- **Independence:** each row in the dataset was a unique target word, and the priming effects were aggregated across participants before the analysis. Item-level independence was therefore confirmed.
- **Multicollinearity:** *Variance Inflation factors* (VIFs) were calculated and compared against the standard threshold 10 to measure whether predictors were too highly correlated to be included in one model. Expected similarity between two OpenAI embedding models caused by similar architecture and training data justified selecting only better-performing OpenAI predictor for the combined model with GPT measures (Model 7).
- **Normality of residuals:** assessed by visual inspection of Q-Q plots and residual histograms for all models. Formal statistical tests were not used, as normality tests like Shapiro-Wilk are known to be overly sensitive with large samples, flagging trivial deviations as significant, which made visual inspection a more suitable approach.
- **Homoscedasticity:** observation of residual-vs-fitted plots in all models. Any systematic deviations from a horizontal line were checked using the *locally estimated smoothing* (LOESS) curves, to notice unequal residual variance.

5.3.5 Software and Implementation

All statistical analyses were performed in R. Pearson correlations were computed using the *cor.test()* function, while regression analyses used *lm()*. For model comparison, *anova()* was used to obtain incremental F-tests, and cross-validated R^2 values were computed using 10-fold cross-validation from the *caret package*. A fixed random seed (*set.seed(18)*) ensured reproducibility. GPT values were rescaled from the original 0-100 to a 0-1 scale before the analysis to improve the comparability of regression coefficients. The embedding API calls and GPT rating collection were implemented using the *httr2* and *openai* R packages. The full analysis code can be found at https://github.com/ninaroj/bachelor_project_26-semantic_priming_llms.git.

6 Results

6.1 Correlation Analysis

6.1.1 GPT Stability

To evaluate the stability of the GPT-4o-mini semantic ratings, Pearson correlations were calculated between run 1 and run 2 for each rating type (association and feature, related and unrelated) and language (English and Polish). As shown in Table 1, the correlations had high values in both languages ($r = .867 - .986$), indicating that the model produced consistent ratings across multiple calls at temperature=0. The lowest correlation was observed for unrelated feature ratings in Polish ($r = .867$) and the highest value was reached for related feature ratings in English ($r = .986$). Based on these results, ratings from the two runs were averaged to create the final association and feature scores used in the following analysis.

Table 1: GPT stability: Pearson correlations between run 1 and run 2 for each rating type, condition, and language. Computed on 0–100 scale at temperature = 0 before difference scores.

Rating type (condition)	English r	Polish r
Association (related)	.975	.972
Association (unrelated)	.971	.956
Feature (related)	.986	.975
Feature (unrelated)	.892	.867

6.1.2 Convergent Validity

While evaluating effect scores (related minus unrelated) for all measures, the strongest correlation among all three embeddings was observed between the two OpenAI models ($r = .777$ in English, $r = .656$ in Polish). FastText showed weaker correlations with other embeddings ($r = .326$ in English, $r = .247$ in Polish) for small and ($r = .416$ in English, $r = .310$ in Polish) for large OpenAI models. This suggests that the two OpenAI models capture semantic representations in a more similar way than with the fastText model, due to the differences in the training corpus and architecture. The two GPT measures (association and feature ratings) were strongly correlated [2] with each other in both languages ($r = .582$ in English, $r = .589$ in Polish), which aligned with the expectation that they capture related but distinct aspects of semantic similarity.

All embedding models correlated positively with both GPT measures in English and Polish (all $p < .001$), suggesting that the different approaches capture some common semantic information. OpenAI large showed the strongest moderate correlation with GPT measures ($r = .317$ and .363 in English, $r = .279$ and .401 in Polish) followed by OpenAI small and fastText. In all cases, feature overlap was slightly more correlated to the embedding measures than association strength. These results suggest that embedding-based and GPT measures reflect a shared semantic convergence, while capturing unique information. In general, OpenAI embeddings showed stronger correlations with GPT measures than fastText did. All results can be found in Table 2.

[2]Following Cohen (1988) conventions, correlations of .10, .30, and .50 are considered weak, moderate, and strong, respectively. These thresholds are used as approximate guidelines throughout the results.

Table 2: Convergent validity: Pearson correlations between all semantic measures (three embedding-based, two GPT-based), for English and Polish. All correlations were computed on effect scores (related - unrelated).

	English		Polish	
Measure pair	r	p	r	p
fastText - OAI small	.326	$< .001$	.247	$< .001$
fastText - OAI large	.416	$< .001$	.310	$< .001$
OAI small - OAI large	.777	$< .001$	.656	$< .001$
Association - Feature	.582	$< .001$	.589	$< .001$
fastText - Association	.164	$< .001$	.157	$< .001$
fastText - Feature	.238	$< .001$	.226	$< .001$
OAI small - Association	.283	$< .001$	.198	$< .001$
OAI small - Feature	.325	$< .001$	.345	$< .001$
OAI large - Association	.317	$< .001$	.279	$< .001$
OAI large - Feature	.363	$< .001$	.401	$< .001$

6.1.3 Predictive Validity

As presented in Table 3, individual correlation of each embedding model with priming effect resulted in significant values for English (fastText: $r = .286$, OAI large: $r = .216$, OAI small: $r = .162$, all $p < .001$), with fastText showing the strongest relation, followed by OAI large and OAI small. In Polish, on the other hand, the correlations were weaker. As in English, the strongest association with the priming effect was observed in fastText ($r = .127$, $p < .001$), which was the only predictor reaching significance, as OAI large ($r = .060$, $p = .069$) and OAI small ($r = .023$, $p = .483$) failed to reach significance. These results suggest that embedding-based measures of semantic similarity are generally related to the human priming effect, but this relationship is stronger and more consistent in English, where all three embeddings were significant, than in Polish, where only fastText reliably predicted priming.

In general, GPT ratings did not meaningfully predict priming in either language. While feature overlap showed a small, statistically significant relationship in English ($r = .076$, $p < .05$) but not in Polish ($r = .025$, $p = .446$), association strength failed to predict priming in both cases ($r = .055, p = .090$ in English, $r = -.002, p = .952$ in Polish). These findings suggest that although GPT ratings capture meaningful semantic relations between words, they do not consistently predict item-level semantic priming effects.

6.1.4 Cross-Linguistic Comparison

Based on the correlations observed in Sections 6.1.1-6.1.3, the item-level priming effects themselves were only weakly correlated across the two languages ($r = .228$, $p < .001$), indicating that the same word pairs do not produce equivalent priming in English and Polish. In addition, fastText, being the strongest predictor in both languages, showed more than twice as large values in English ($r = .286$) as in Polish ($r = .127$). OpenAI embeddings performed similarly (OAI large: $r = .216$

in English, $r = .060$ in Polish; OAI small: $r = .162$ in English, $r = .023$ in Polish), but here the difference was even more noticeable, as the OpenAI small failed to reach significance in Polish. While the GPT ratings were generally weak in both languages, they again reached higher values for English (association: $r = .055$ in English, $r = -.002$ in Polish; feature: $r = .076$ in English, $r = .025$ in Polish). The disproportion between semantic priming effects across English and Polish is visually presented by Figure 2. This pattern was also visible in convergent validity, where the reported correlations were consistently lower in Polish as well, whereas GPT stability showed comparably high results in both languages. The English advantage possibly also reflects a lower reliability of the Polish priming estimates, which more strongly affected the tested correlations. Therefore, the level to which the observed differences are influenced by the reliability rather than true differences in captured priming effects is examined through attenuation correction in the following section.

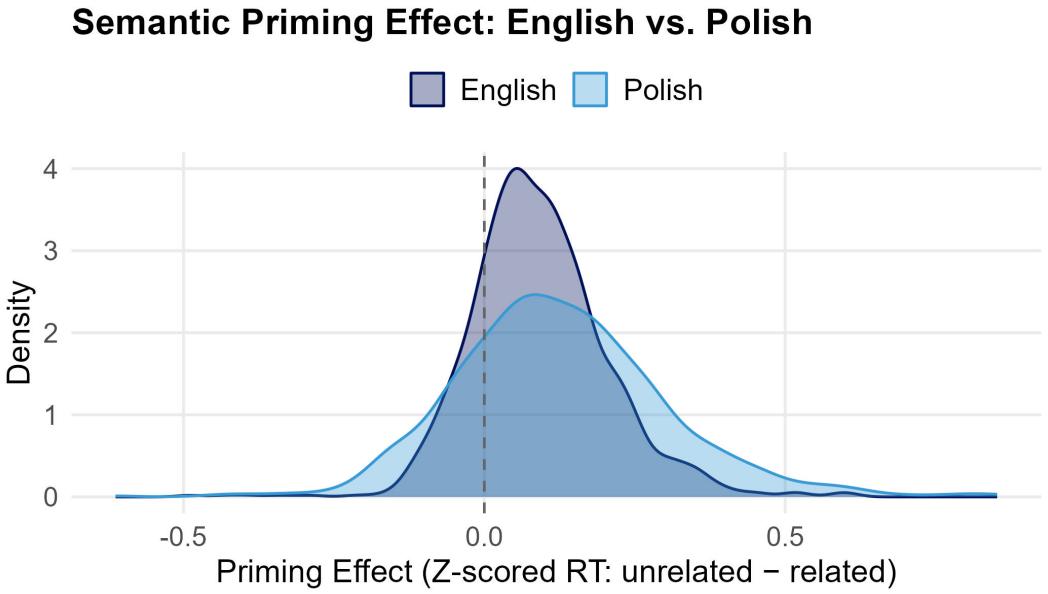

Figure 2: Distribution of the item-level priming effect (z-scored RT: unrelated - related) for English and Polish. Values above zero (dashed line) indicate semantic facilitation.

6.1.5 Attenuation Correction

Because observed correlations are limited by the reliability of the priming measures, the true predictive validity correlations might be stronger than those observed. To account for this, the correlations were corrected using $r_{corr} = r/\sqrt{r_{yy}}$, where r_{yy} is the *split-half reliability* estimate reported by SPAML($r_{yy} = .719$ for English, $r_{yy} = .524$ for Polish). Attenuation correction was applied only to the priming measures and supported the earlier observations (see results in Table 3). FastText remained the strongest predictor in both languages ($r_{corr} = .337$ in English, $r_{corr} = .175$ in Polish), followed by OpenAI large ($r_{corr} = .255$ in English, $r_{corr-pl} = .082$ in Polish) and OpenAI small ($r_{corr} = .191$ in English, $r_{corr} = .032$ in Polish). Although the gap between English and Polish became smaller, it did not disappear, as English correlations remained nearly twice as large as those in Polish even after adjusting for the difference in priming reliability. This suggests that

the weaker predictive validity in Polish cannot be explained only by lower reliability and that other factors, like differences in the stimuli, translation, or the quality of the embedding representations, also likely contribute.

Table 3: Predictive validity: Pearson correlations of each semantic measure with the item-level priming effect, for English and Polish. r = observed correlation; r_{corr} = correlation corrected for attenuation using SPAML split-half reliabilities (English $r_{yy} = .719$, Polish $r_{yy} = .524$)

	English			Polish		
Predictor	r	r_{corr}	p	r	r_{corr}	p
fastText cosine effect	.286	.337	$< .001$	.127	.175	$< .001$
OAI small cosine effect	.162	.191	$< .001$	.023	.032	.483
OAI large cosine effect	.216	.255	$< .001$	.060	.082	.068
Association effect	.055	.065	.090	$-.002$	$-.003$	.952
Feature effect	.076	.090	.015	.025	.034	.446

6.2 Regression Results

6.2.1 Step 1: Cosine Predictors

Controls (Model 0). The baseline model, containing only control variables (word frequency and word length), explains a small but significant amount of variance in priming effect in both English ($R^2 = .042$, $R^2_{adj} = .040$, $R^2_{cv} = .038$) and Polish ($R^2 = .034$, $R^2_{adj} = .032$, $R^2_{cv} = .028$). Both word frequency (English: $b = -.040$, $p < .001$; Polish: $b = -.050$, $p < .001$) and word length (English: $b = -.005$, $p < .05$; Polish: $b = -.007$, $p < .05$) were significant negative predictors in both languages, which aligns with the expectation that less frequent words show higher priming effect, because they benefit more from semantic facilitation (Hutchison et al., 2008). Full regression results are reported in Table 5.

fastText cosine similarity (Model 1). Adding fastText cosine similarity predictor led to the largest single contribution improvement over the baseline model in both languages. In English, the model explains 11.4% of variance ($R^2 = .114$, $R^2_{adj} = .111$, $R^2_{cv} = .108$), showing significant increase over the control model ($\Delta R^2 = .072$, $p < .001$) (see Table 4). In Polish, the improvement was smaller but still significant ($\Delta R^2 = .014$, $p < .001$). FastText cosine similarity was a significant positive predictor of priming in both languages ($b = .251$, $p < .001$ in English; $b = .149$, $p < .001$ in Polish).

OpenAI cosine similarity (Models 2 and 3). The OAI small (Model 2) significantly predicted priming in English ($\Delta R^2 = .029$, $p < .001$), but not in Polish ($\Delta R^2 = .003$, $p = .111$). Differently, OAI large predictor (Model 3) showed significant improvement over controls in both languages ($\Delta R^2 = .044$, $p < .001$ in English; $\Delta R^2 = .004$, $p < .05$ in Polish), although with a much smaller effect size in Polish. In both languages, OAI large outperformed OAI small justifying its selection for the combined regression model in Step 2.

All cosine predictors (Model 4). When all three cosine similarity predictors were entered together, the combined model reached the best overall performance (English: $R^2 = .125$, $R^2_{adj} = .121$, $R^2_{cv} = .115$; Polish: $R^2 = .048$, $R^2_{adj} = .043$, $R^2_{cv} = .035$). In English, both fastText ($b = .204$, $p < .001$) and OAI large ($b = .108$, $p < .05$) made significant independent contributions to predicting priming, while OAI small did not ($b = .016$, $p = .717$). In Polish, only fastText remained significant ($b = .137$, $p < .05$), and neither OpenAI predictor contributed to additional variance (OAI large: $b = .032$, $p = .546$; OAI small: $b = .008$, $p = .879$). These results suggest that while in English fastText and OAI large capture semantic information together, in Polish most of the predictive information comes from fastText alone.

6.2.2 Step 2: GPT Ratings

GPT measures only (Model 5). When GPT ratings (association strength and feature overlap) were added to the control variables, no significant improvement in prediction of priming effects was observed in either language ($\Delta R^2 = .006$, $R^2_{adj} = .004$, $R^2_{cv} = .001$, $p = .056$ in English; $\Delta R^2 = .002$, $R^2_{adj} = .000$, $R^2_{cv} = -.005$, $p = .379$ in Polish). Although feature overlap showed a weak, approaching significance regression coefficient in English ($b = .032$, $p = .092$), neither GPT predictor reached significance threshold after controlling for word frequency and length.

GPT ratings and cosine predictors (Models 6 and 7). Adding the GPT measures to the fastText (Model 6) and the OAI large (Model 7) did not improve the prediction beyond the previous cosine similarity measures. In both languages the increase was similarly negligible and non-significant. None of the GPT predictors were significant in any of the combined models and additionally, the cross-validated R^2 values for Models 6 and 7 slightly decreased in the presence of GPT ratings compared to the corresponding models with cosine similarity only (English M1 vs M6: .108 → .103, M3 vs M7: .080 → .0.76; Polish M1 vs M6: .040 → .034, M3 vs M7: .028 → .023) (visually presented in Figure 3). These results show that GPT association strength and feature overlap ratings did not explain more variance in item-level priming beyond what is already captured by cosine similarity predictors.

Table 4: Model comparisons reporting change in fit (ΔR^2, ΔR^2_{adj}, (ΔR^2, ΔR^2_{adj}, ΔR^2_{cv}) with incremental F-tests, for English and Polish. Each row compares a model against the baseline.

		English					Polish				
Comparison	df	ΔR^2	ΔR^2_{adj}	ΔR^2_{cv}	F	p	ΔR^2	ΔR^2_{adj}	ΔR^2_{cv}	F	p
M1 vs M0 (fastText)	$(1, 931)$	.072	.071	.070	75.25	$< .001$	.014	.013	.011	13.53	$< .001$
M2 vs M0 (OAI small)	$(1, 931)$	.029	.028	.027	29.42	$< .001$	.003	.002	.000	2.54	.111
M3 vs M0 (OAI large)	$(1, 931)$	.044	.043	.042	44.54	$< .001$	.004	.003	.000	3.98	.046
M4 vs M0 (all cosine)	$(3, 929)$	.084	.081	.078	29.55	$< .001$	.015	.012	.007	4.78	.003
M5 vs M0 (GPT only)	$(2, 930)$	.006	.004	.001	2.89	.056	.002	.000	$-.005$	0.97	.379
M6 vs M0 (fastText, GPT)	$(3, 929)$	.072	.069	.065	25.10	$< .001$	.015	.012	.006	4.75	.003
M6 vs M1 (GPT added)	$(2, 929)$	.000	$-.002$	$-.005$	0.10	.909	.001	$-.001$	$-.005$	0.37	.691
M7 vs M0 (OAI, GPT)	$(3, 929)$	.044	.041	.039	14.90	$< .001$	.005	.002	$-.006$	1.60	.187
M7 vs M3 (GPT added)	$(2, 929)$	.000	$-.002$	$-.003$	0.12	.890	.001	$-.001$	$-.006$	0.42	.660

Table 5: Full regression results for all models (M0–M7) in English and Polish (n = 935 each).

		English						Polish					
Model	Predictor	b	t	p	R^2	R^2_{adj}	R^2_{cv}	b	t	p	R^2	R^2_{adj}	R^2_{cv}
M0	Frequency	−.040	−6.39	< .001	.042	.040	.038	−.050	−5.57	< .001	.034	.032	.028
	Length	−.005	−2.82	.005				−.007	−2.75	.006			
M1	Frequency	−.035	−5.77	< .001				−.048	−5.40	< .001			
	Length	−.005	−2.82	.005	.114	.111	.108	−.006	−2.72	.007	.047	.044	.040
	fastText	.251	8.68	< .001				.149	3.68	< .001			
M2	Frequency	−.041	−6.70	< .001				−.051	−5.72	< .001			
	Length	−.007	−3.58	< .001	.071	.068	.064	−.007	−3.02	.003	.036	.033	.028
	OAI small	.153	5.42	< .001				.061	1.60	.111			
M3	Frequency	−.039	−6.27	< .001				−.049	−5.54	< .001			
	Length	−.006	−3.46	< .001	.086	.083	.080	−.007	−3.04	.002	.038	.035	.028
	OAI large	.215	6.67	< .001				.078	2.00	.046			
M4	Frequency	−.035	−5.81	< .001				−.048	−5.34	< .001			
	Length	−.006	−3.20	.001				−.007	−2.82	.005			
	fastText	.204	6.48	< .001	.125	.121	.115	.137	3.19	.002	.048	.043	.035
	OAI small	.016	0.36	.717				.008	0.15	.879			
	OAI large	.108	2.08	.038				.032	0.61	.546			
M5	Frequency	−.040	−6.39	< .001				−.050	−5.61	< .001			
	Length	−.005	−2.93	.004	.048	.044	.039	−.007	−2.93	.004	.036	.032	.024
	Assoc.	.008	0.41	.684				−.020	−0.65	.513			
	Feature	.032	1.69	.092				.042	1.38	.168			
M6	Frequency	−.035	−5.77	< .001				−.048	−5.40	< .001			
	Length	−.005	−2.83	.005				−.007	−2.78	.006			
	fastText	.248	8.31	< .001	.114	.109	.103	.146	3.51	< .001	.048	.043	.034
	Assoc.	.003	0.15	.885				−.023	−0.76	.446			
	Feature	.005	0.26	.797				.023	0.76	.446			
M7	Frequency	−.039	−6.26	< .001				−.049	−5.54	< .001			
	Length	−.006	−3.48	< .001				−.008	−3.11	.002			
	OAI large	.217	6.22	< .001	.086	.081	.076	.072	1.69	.091	.039	.033	.023
	Assoc.	−.009	−0.48	.629				−.023	−0.77	.444			
	Feature	.005	0.28	.782				.026	0.83	.406			

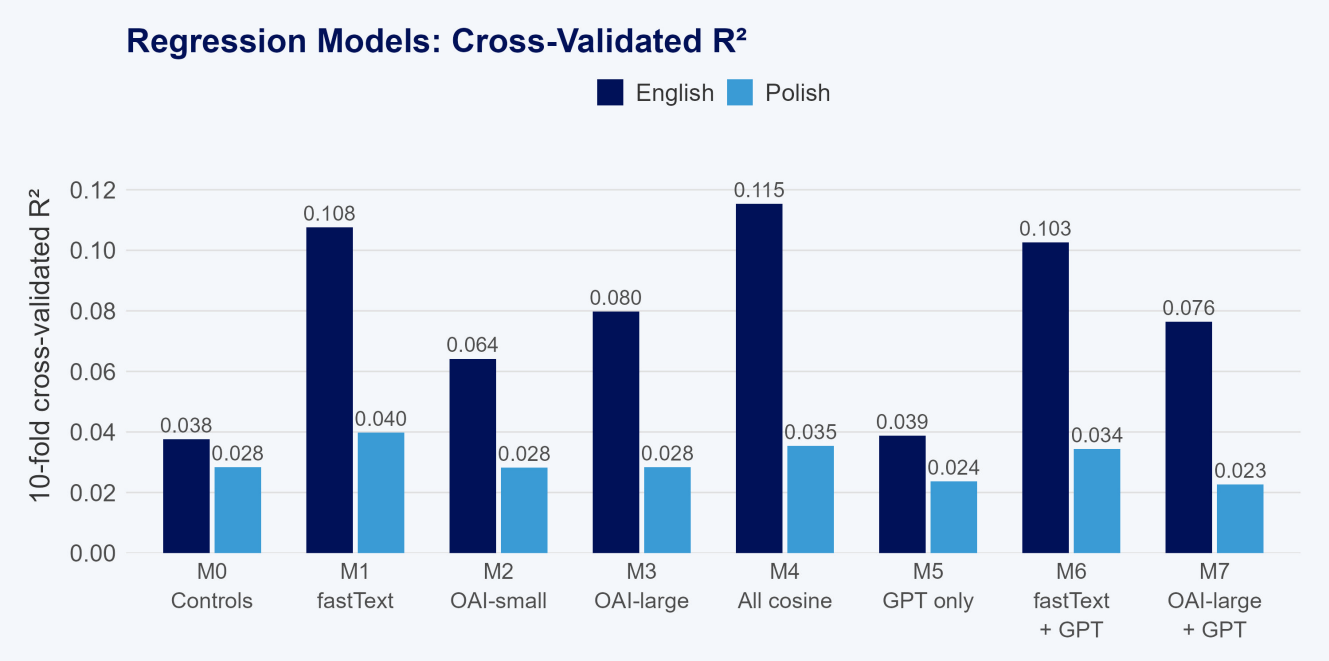

Figure 3: 10-fold cross-validated R^2 for all regression models (M0–M7) in English and Polish.

6.2.3 Cross-Linguistic Comparison

Across all regression analyses, the English models consistently explained more variance than the Polish models (see Figure 3), and OAI small failed to significantly predict priming in Polish. These confirm that the relationship between LLM semantic similarity measures and human priming effect is stronger and more consistent in English.

Analysing the cosine similarity distribution (see Table 6) revealed differences between the two languages and helped explain the weaker results for Polish. For fastText, related word pairs had slightly lower cosine similarity in Polish ($M = .498$) than in English ($M = .553$), while unrelated pairs were very similar across languages (English: $M = .109$, Polish: $M = .100$). As a result, the cosine effect, calculated as a difference between related and unrelated pairs, was smaller for Polish.

A difference appeared in OAI large embeddings as well, where the related pairs showed similar cosine similarity in both languages (English: $M = .533$, Polish: $M = .561$), but unrelated pairs were noticeably higher in Polish ($M = .322$) than in English ($M = .238$). This reduced the cosine effect in Polish not because of lower semantic similarity of related pairs, but because of higher similarity of unrelated pairs, as visually presented in Figure 4.

Table 6: Cros-linguistic statistics (M, SD, Mdn) of related and unrelated cosine similarity for each embedding model, by language.

		English			Polish		
Model	Condition	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*
fastText	Related	.553	.113	.551	.498	.133	.502
	Unrelated	.109	.031	.115	.100	.034	.106
OAI small	Related	.541	.113	.537	.451	.138	.435
	Unrelated	.247	.063	.240	.266	.063	.266
OAI large	Related	.533	.100	.527	.561	.138	.572
	Unrelated	.238	.050	.233	.322	.076	.336

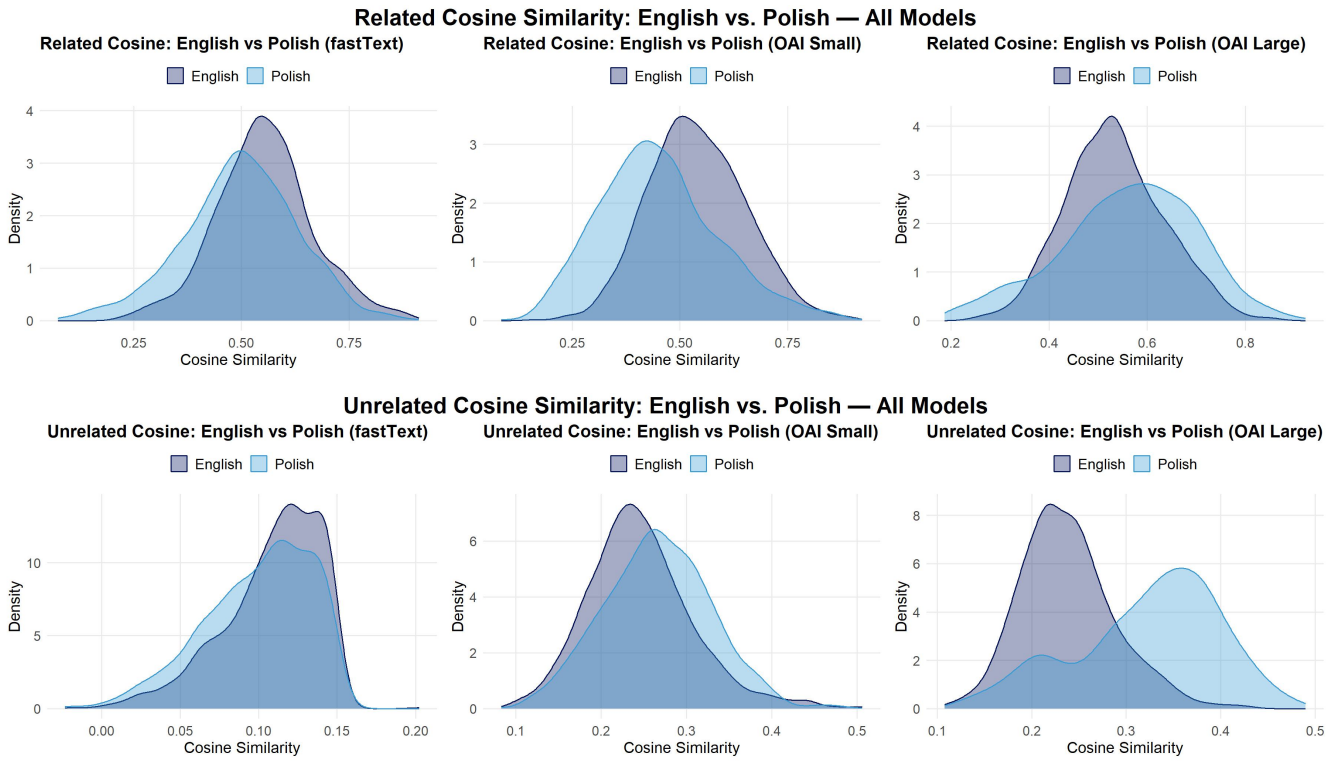

Figure 4: Distribution of cosine similarity for related (top row) and unrelated (bottom row) word pairs across the three embedding models (fastText, OAI small, OAI large), comparing English and Polish.

6.2.4 Attenuation Correction

In order to remove the noise in priming effect measurements and the differences in reliability of languages, the observed predictive correlations were corrected for attenuation using split-half reliability estimates reported by Buchanan et al. (2026) (English: $r_{yy} = .719$; Polish: $r_{yy} = .524$).

Table 7: Attenuation-corrected model comparisons for English and Polish (observed -> corrected). Correction used SPAML split-half reliabilities (English $r_{yy} = .719$; Polish $r_{yy} = .524$).

	English			Polish		
Comparison	ΔR^2	ΔR^2_{adj}	ΔR^2_{cv}	ΔR^2	ΔR^2_{adj}	ΔR^2_{cv}
Step 1: Cosine models vs. controls (M0)						
M1 vs M0	$.072 \rightarrow .100$	$.071 \rightarrow .099$	$.070 \rightarrow .097$	$.014 \rightarrow .026$	$.013 \rightarrow .026$	$.011 \rightarrow .022$
M2 vs M0	$.029 \rightarrow .041$	$.028 \rightarrow .040$	$.027 \rightarrow .037$	$.003 \rightarrow .005$	$.002 \rightarrow .004$	$-.000 \rightarrow -.000$
M3 vs M0	$.044 \rightarrow .061$	$.043 \rightarrow .060$	$.042 \rightarrow .059$	$.004 \rightarrow .008$	$.003 \rightarrow .007$	$.000 \rightarrow .000$
M4 vs M0	$.084 \rightarrow .116$	$.081 \rightarrow .114$	$.078 \rightarrow .108$	$.015 \rightarrow .028$	$.012 \rightarrow .025$	$.007 \rightarrow .013$
Step 2: GPT models						
M5 vs M0	$.006 \rightarrow .008$	$.004 \rightarrow .006$	$.001 \rightarrow .002$	$.002 \rightarrow .004$	$-.000 \rightarrow .002$	$-.005 \rightarrow -.009$
M6 vs M0	$.072 \rightarrow .100$	$.069 \rightarrow .097$	$.065 \rightarrow .090$	$.015 \rightarrow .028$	$.012 \rightarrow .025$	$.006 \rightarrow .012$
M6 vs M1	$.000 \rightarrow .000$	$-.002 \rightarrow -.002$	$-.005 \rightarrow -.007$	$.001 \rightarrow .001$	$-.001 \rightarrow -.001$	$-.005 \rightarrow -.010$
M7 vs M0	$.044 \rightarrow .061$	$.041 \rightarrow .058$	$.039 \rightarrow .054$	$.005 \rightarrow .010$	$.002 \rightarrow .007$	$-.006 \rightarrow -.011$
M7 vs M3	$.000 \rightarrow .000$	$-.002 \rightarrow -.002$	$-.003 \rightarrow -.005$	$.001 \rightarrow .002$	$-.001 \rightarrow -.000$	$-.006 \rightarrow -.011$

As shown in Table 7, after applying correction the variance explained by the best-performing cosine model (fastText) increased from .072 to about .100 in English and from .014 to .026 in Polish. The GPT only model on the other hand showed little improvement in either language. Even after correcting for reliability, the English models continued to outperform Polish models noticeably - the corrected variance explained by fastText was .100 for English and only .026 for Polish. This suggests that the weaker performance in Polish cannot be explained only by lower reliability of the dataset, as the other language-related factors are also likely contributing to the difference.

6.3 Assumption Checks

Regression assumptions were checked for all seven models in both languages. Linearity was assessed using visual inspection of scatterplots for each predictor against the priming effect, and no systematic deviations from linearity were observed. Each row represented a unique target word with a priming effect averaged across participants, confirming item-level independence. Multicollinearity was assessed using Variance Inflation Factors (VIF) for all models. Because only Models 4, 6 and 7 had multiple predictors beyond the control variables, they were reported in detail in Table 8, as if multicollinearity is not a problem in these cases, it is unlikely to be an issue in the simpler models. All VIF values were below the standard threshold of 10, meaning that collinearity was not a concern.

Normality of residuals was tested using Q-Q plots (Figure 5) and residual histograms for all models (Figure 6). Residuals were normally distributed across the central line, with minor deviations in the tails, which given the large sample size are unlikely to affect the validity of statistical tests.

To measure homoscedasticity, residuals/fitted values plots were used (Figure 7). In both languages, the locally estimated scatterplot smoothing (loess) lines were generally flat, suggesting that the variance of the residuals stayed relatively consistent across predicted values and homoscedasticity was not a concern. Visual inspection of the plots identified a small number of outlying observations in both datasets, but there was not clear theoretical or methodological justification to remove them from the analysis.

Table 8: Variance inflation factors (VIF) for the multi-predictor models (M4, M6, M7) in English and Polish. Values < 10 indicate no multicollinearity concern.

Model	Predictor	English VIF	Polish VIF
M4	Frequency	1.32	1.13
	Length	1.32	1.15
	fastText	1.22	1.12
	OAI small	2.57	1.83
	OAI large	2.75	1.86
M6	Frequency	1.30	1.10
	Length	1.30	1.12
	fastText	1.07	1.06
	Association	1.52	1.54
	Feature	1.57	1.61
M7	Frequency	1.29	1.10
	Length	1.30	1.13
	OAI large	1.19	1.22
	Association	1.54	1.54
	Feature	1.71	1.71

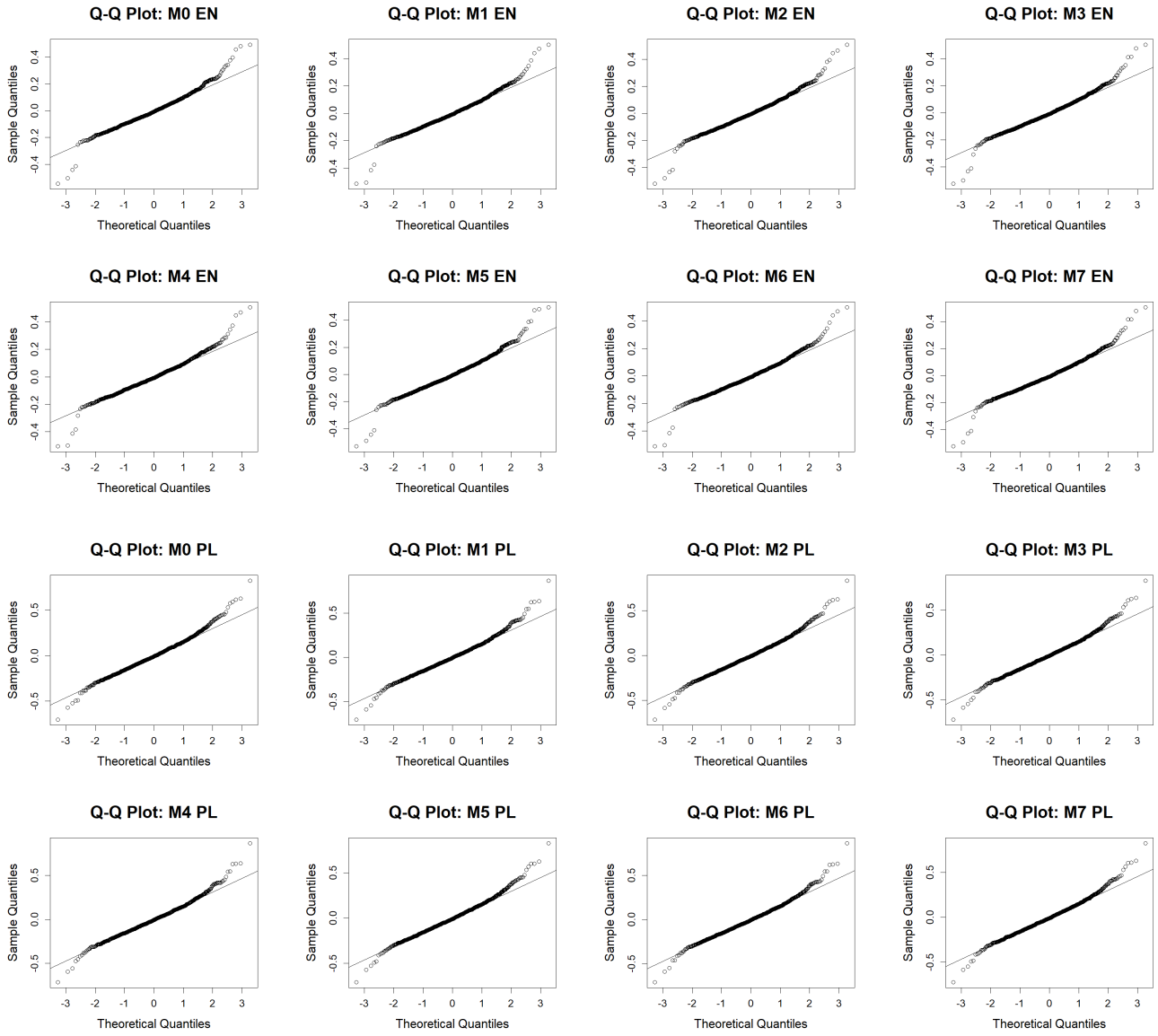

Figure 5: Q-Q plots of residuals for all models (M0–M7), English (top two rows) and Polish (bottom two rows). Each panel plots sample quantiles of the residuals (y-axis) against theoretical quantiles (x-axis). Points following the diagonal indicate approximately normal residuals.

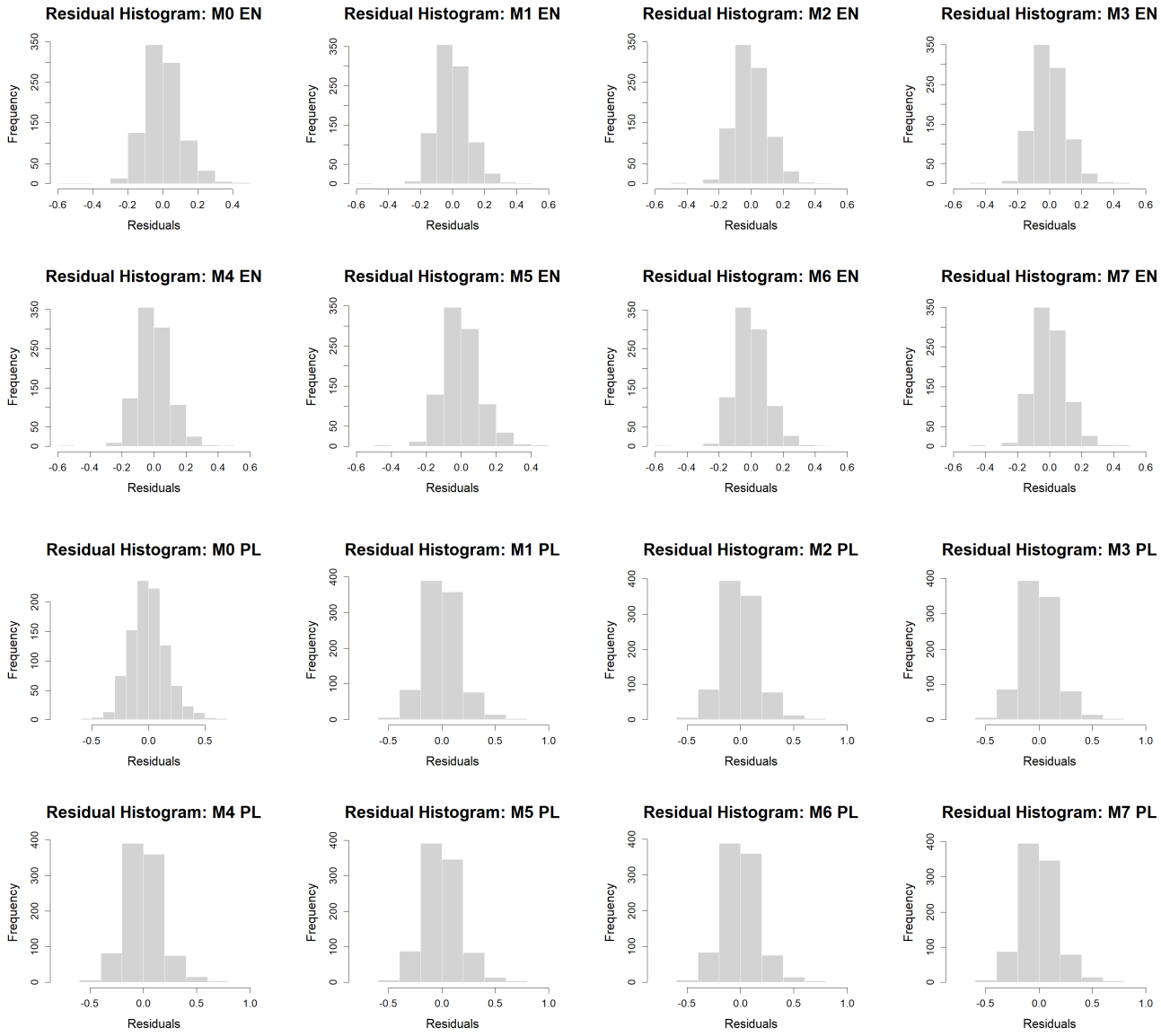

Figure 6: Residual histograms for all models (M0–M7), English (top two rows) and Polish (bottom two rows). Each panel plots frequency (y-axis) against residual value (x-axis) to assess the normality of residuals.

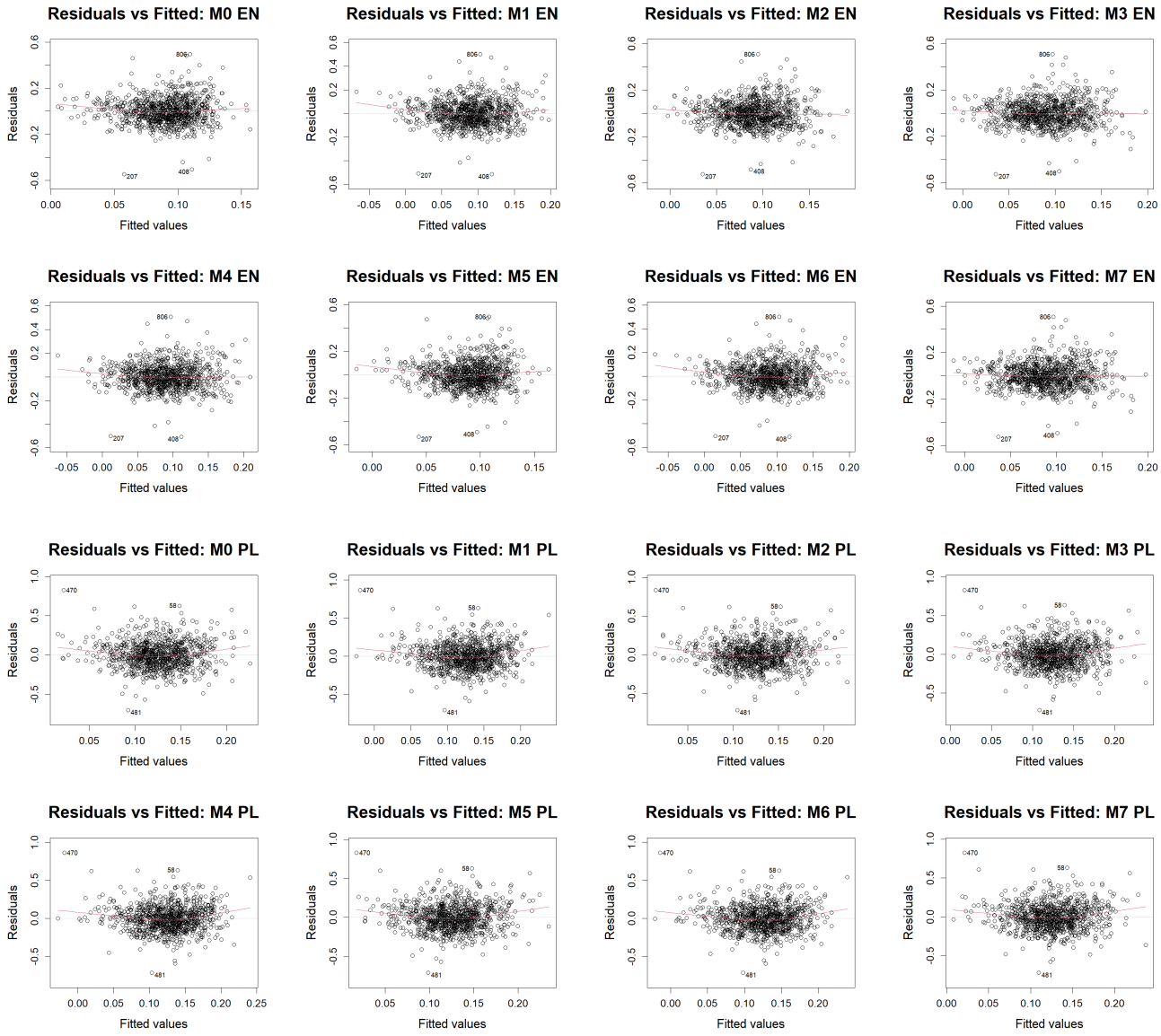

Figure 7: Residuals versus fitted values for all models (M0–M7), English (top two rows) and Polish (bottom two rows). Each panel plots residuals (y-axis) against fitted values (x-axis) to assess homoscedasticity, indicated by approximately flat LOESS curves.

7 Discussion

This study investigated whether semantic representations derived from Large Language Models can predict item-level semantic priming magnitude in English and Polish. Three types of LLM-derived measurements were tested: cosine similarity from fastText model trained on subtitle corpus, cosine similarity from two OpenAI embedding models of different sizes, and GPT-4o-mini ratings of association strength and feature overlap. The results showed that cosine similarity consistently predicted semantic priming across both languages, with fastText providing the strongest predictions, while prompt-based GPT ratings did not independently contribute to priming predictions beyond cosine similarity. The correlation between LLM-based measurements and human priming effects were consistently stronger in English than in Polish, also after applying correction for attenuation. Each of these findings is discussed in the following sections.

7.1 Cosine Similarity as a Predictor of Priming

The main finding of this study is that cosine similarity values derived from distributional embedding models significantly predict item-level semantic priming in both English and Polish. This supports the hypothesis that LLMs, which are trained only on large text without any grounding in the human behavioural context, can partially capture the semantic structure of the mental lexicon reflected in priming effects. The results are consistent with distributional semantic theory, which states that word meanings are largely determined by the contexts in which they appear (Lenci, 2018).

Comparison of all models in both languages resulted in fastText cosine similarity being the strongest single predictor of priming. FastText alone explained an additional 7.2% of variance in English and 1.4% in Polish building on top of the baseline model (M0 with word frequency and word length as controls). While this observation was surprising, given the simpler structure of fastText compared to OpenAI models using transformer architectures, one plausible explanation lies in the source and not size of data used for training. The fastText embeddings used in this study come from the SPAML dataset, which relied on the OpenSubtitles corpus applied to the subs2vec project (Van Paridon and Thompson, 2021). Since subtitle data reflect more closely natural, conversational language, it may capture underlying association patterns present in semantic priming better than more formal, written language, especially in the lexical decision tasks which also rely on everyday word recognition. This interpretation aligns with previous findings of Mandera et al. (2017), who showed that models trained on the subtitle data outperform different corpora, especially in predicting psycholinguistic measures.

The two OpenAI models showed significant prediction strength in English, with the text-embedding-large (3072 dimensions) consistently outperforming text-embedding-small (1536 dimensions). The strongest result for this comparison comes from Model 4, including all three cosine similarity models combined together. In this model, both fastText and OpenAI large showed significant independent contributions, while OpenAI small did not. This suggests that fastText and OpenAI large because of the differences in architecture, training data and dimensionality capture partially different aspects of semantic similarity and provide complementary information. The OAI-large may represent more detailed, structural relations, not fully captured by the sensitive for conversational associations fastText, and vice versa. Together, these two account for additional 8.4% of

variance in English, which is the best-performing model tested.

Despite this, the overall prediction magnitude remains modest, with the best-performing model (Model 4) explaining 12.5% of variance in English and only 4.8% in Polish. The majority of the item-level semantic priming therefore stays uncaptured by any of the semantic similarity measures, which is not surprising, as semantic priming is influenced by many factors. Therefore, cosine similarity reflects distributional patterns of word use, but cannot capture all aspects of semantic processing that contribute to priming.

The attenuation-corrected analysis provided a better picture of the true correlations, where the additional variance explained by the combined cosine model (Model 4) increased from 8.4% to 11.6%. This suggests that the relationship between LLM-derived semantic similarity and the human semantic processes measured by priming is stronger than the raw correlations imply.

7.2 GPT Ratings as Predictors of Priming

The second key finding is that prompt-based GPT ratings of association strength and feature overlap fail to predict item-level priming in either language, both when the GPT measures were tested alone and when added to the cosine similarity models. Although this is a null result, it provides insight into the relation between different types of semantic representations worth careful interpretation.

The low performance of GPT measures was not a result of poor quality, as the ratings showed high internal stability in both categories and languages across two runs. The models were consistent in semantic judgments, and highly correlated with all three embedding-based measures, which means they produced valid responses. This suggests that while GPT measures can show strong convergent validity with other semantic measures, they still fail to predict semantic priming.

One possible explanation lies in the nature of the semantic knowledge each type of measure captures. Semantic priming is a result of automatic spreading activation in the mental lexicon, meaning that when one word is activated, related words become more accessible and easier to recognise. While this process is fast, automatic and influenced by the language, the GPT ratings require more reflective and controlled conditions, when answering "how strongly word A brings word B to mind" or "how many features does word A and B share".

The positive correlation between GPT ratings and cosine similarity suggests that both measures capture general semantic similarity, but in different ways. Cosine similarity based on the distributional patterns reflects how often two words appear in similar contexts, while GPT ratings focus on semantic knowledge on item's category, features and broader word context. As a result, two words may have high GPT ratings because they are conceptually similar or share many features, while having low distributional co-occurance, which would result in weak cosine similarity. Consequently, words sharing high frequency in similar contexts could produce strong priming effects despite not being semantically realted.

Overall, these findings support the idea that explicit semantic judgments and implicit semantic processes are not the same thing, and while GPT ratings can capture meaningful relationships between words, they do not reflect the automatic association processes present in priming. This suggests that for predicting behavioural data related to lexical activation, distributional measures

like cosine similarity may be more appropriate than prompt-based ratings. However, future research could further study if alternative prompting approaches, different from explicit judgments, can better predict associative structures in semantic priming.

7.3 Cross-Linguistic Differences

The third major finding was that the relationship between LLM similarity measures and semantic priming was consistently weaker in Polish compared to English. When fastText explained 7.2% additional variance in English, it only reached 1.4% in Polish; OpenAI small predicted priming in English, but not in Polish. Although the cross-linguistic priming correlation was significant ($r = .228$), confirming that semantic priming effects are partially shared across languages, the magnitude was weak, which suggests that most of the item-level variance in priming is language-specific. While there are few factors that likely contributed to the weaker results in Polish, one of them is lower reliability of the Polish dataset. The split-half reliability estimate accessed from the SPAML study was lower for Polish ($r_{yy} = .524$) than for English ($r_{yy} = .719$), which means that the data contained more measurement noise. However, while attenuation correction partially addressed this issue, the corrected Polish values ($R^2 = .026$) remained significantly lower than English ($R^2 = .100$), meaning the reliability alone cannot explain the gap.

Another likely factor is the greater morphological complexity of Polish. Unlike English, Polish is a highly inflected language, where words change to different forms depending on the grammatical case, gender or number. As a result, each form may receive a different vector representation in the embedding space, potentially reducing the precision of cosine similarity estimates. This would be especially problematic if the inflected form used in the SPAML dataset was different from the most frequent form of the embedding training corpus. While a formal test would require splitting the training corpus in half and verifying the consistency of resulting vectors, convergent validity results provide indirect support of this idea (Table 2). The correlations between independently trained cosine similarity embedding models were consistently lower for Polish than for English, while the correlation between the two GPT measures was slightly higher in Polish. Since GPT ratings are based on the explicit judgments, not distributional vectors, they are less likely to be affected by morphological variations in the training data. This suggests that lower correlations in Polish may be the result of more complex word forms and not differences in semantic structure. The failure in predicting priming in Polish by smaller OAI embedding, strengthens this interpretation, as a model with more dimensionality could potentially handle morphological variations better than the smaller-dimension model.

A further consideration is the origin of the Polish SPAML dataset. The used Polish stimuli was translated from the original English dataset, and not selected based on the Polish association norms. Because direct translation does not always preserve the same association strength across languages, some translated pairs may have a weaker association in Polish than in English due to the cultural context or frequency in everyday use. This could introduce additional noise into the Polish data and weaken the relationship between semantic similarity and priming regardless of the measures' quality.

The differences in the cosine similarity distributions between the two languages further explained the weaker Polish results. For OAI large embeddings, the unrelated pairs had higher cosine simi-

larity in Polish than in English, while related pairs showed comparable values in both languages. This reduced the cosine effect in Polish not because semantically related words were less similar, but because semantically unrelated words were more similar than in English. One possible explanation is that morphologically related Polish words may appear more similar in the embedding space even when they are semantically unrelated, raising the baseline similarity of unrelated pairs.

All together, these findings suggest that the weaker performance in Polish is a result of morphological, linguistic and psycholinguistic factors. The significant and positive effect of fastText indicates that LLM-derived semantic similarity still captures meaningful aspects of semantic processing across languages. Considering that subtitles-trained model was able to predict Polish priming effect, even at a modest level, suggests that distributional semantic representations are valid for cross-linguistic analysis. At the same time, the results highlight the importance of language-specific training data and morphology-sensitive models.

7.4 Limitations

Several limitations of this study deserve acknowledgment. First, the explained variance across all models was relatively small, with the best-performing model explaining 12.5% of priming in English and 4.8% in Polish. Therefore, the majority of item-level priming variance remains unexplained. This likely shows that semantic priming is influenced by many factors including association (*doctor-nurse*), shared features (*cat-dog*), as well as differences in experimental context and participants, which may not be captured by distributional semantic models. Consistent with this, Heyman et al. (2016) found only moderate correlations between item-level priming effects across two different tasks, suggesting that a big amount of priming variance depends on the specific task and context, so it may not be captured by semantic similarity measures alone. Incorporating additional, more specific human-oriented predictors might improve predictive performance in future work.

Second, the Polish word pairs were translated from English instead of designing a dataset based on original Polish psycholinguistic norms. Because direct translation does not guarantee the same level of semantic association across languages, some word pairs may have different correlations in Polish than in English. Therefore, future cross-linguistic research should prioritise using materials developed independently for each language.

Third, 64 Polish items with missing cosine similarity values were excluded from the regression analysis. The excluded items were mainly multi-word expressions, which may produce priming effects differently from single-word items. As a result, Polish findings may not fully generalise to all types of word pairs.

Fourth, the GPT ratings were collected using only one fixed prompting command for each rating type (one prompt for association strength and one prompt for feature overlap). Different prompt instructions, word framing or rating scale could produce different values, and potentially different predictive accuracy. Although the outputs were highly consistent across repeated runs, consistency does not guarantee that ratings capture the aspects of semantic relatedness most relevant to psycholinguistic processing.

Fifth, the GPT ratings were collected using only one language model (GPT-4o-mini). Larger and more recent OpenAI models as well as models beyond the OpenAI families, may produce different

semantic similarity ratings and vary in levels of predictive accuracy. Future research should focus on expanding the study with more distinct models, to test if the absence of relation between GPT-based ratings and semantic priming is general characteristics of prompt-based ratings or if it depends on the specific model.

Sixth, the cross-language comparison was limited to English and Polish only. Although this provides meaningful insight into generalisation across languages, Polish and English represent only two specific language groups. It is therefore unknown whether the same patterns would be observed in other languages with different morphological structure, writing system, or levels of representations in LLM training data. Further research could extend the analysis to additional languages from the SPAML dataset to better understand how well LLM-derived semantic similarity measures predict human priming across a wider range of languages.

Finally, all embedding models used in this study produce static, independent from the context vector representations. Existing models like BERT or GPT can generate different representations depending on the sentence in which a word appears. Averaging them over many contexts was shown to produce more stable and realistic semantic representations (Cassani et al., 2023), so future work could study whether contextualised embeddings show stronger alignment with priming effects, especially in words whose meanings vary in different contexts.

8 Conclusion

This research shows that semantic representations derived from LLMs can predict item-level semantic priming in English and Polish. Across all tested models, fastText cosine similarity was the strongest and most consistent predictor of priming, and regardless of simpler architecture, it performed better than OpenAI large embedding, possibly due to the differences in the training data. The combined models also showed that fastText and OpenAI large both significantly contributed to the prediction, which means the models capture partially different aspects of semantic similarity. GPT-based ratings (association strength and feature overlap), on the other hand, did not reach a significant level in either language, highlighting existing differences between semantic similarity judgments and association processes present in semantic priming. The correlation results were consistently stronger in English than in Polish, and did not change after applying a correction for attenuation. This suggests that the differences in lexical morphology or direct translations may play a key role in fair cross-linguistic analysis. Overall, these findings support using cosine similarity as a valuable and scalable predictor of human semantic priming, while emphasising the importance of language-specific training data and morphology-sensitive models when extending these methods beyond English and Polish.

References

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Buchanan, E. M., Cuccolo, K., Heyman, T., van Berkel, N., Coles, N. A., Iyer, A., et al. (2026). Measuring the semantic priming effect across many languages. *Nature Human Behaviour*, 10(1):182–201.

Buchanan, E. M., Valentine, K. D., and Maxwell, N. P. (2019). English semantic feature production norms: An extended database of 4,436 concepts. *Behavior Research Methods*, 51(4):1849–1863.

Cassani, G., Bianchi, F., Attanasio, G., Marelli, M., and Günther, F. (2023). Meaning modulations and stability in large language models: An analysis of BERT embeddings for psycholinguistic research. PsyArXiv preprint.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge, New York, 2nd edition.

Collins, A. M. and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428.

De Deyne, S. (2024). 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.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., and Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, 51(3):987–1006.

Heyman, T., Hutchison, K. A., and Storms, G. (2016). Uncovering underlying processes of semantic priming by correlating item-level effects. *Psychonomic Bulletin & Review*, 23(2):540–547.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review*, 10(4):785–813.

Hutchison, K. A., Balota, D. A., Cortese, M. J., and Watson, J. M. (2008). Predicting semantic priming at the item level. *The Quarterly Journal of Experimental Psychology*, 61(7):1036–1066.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., Yap, M. J., Bengson, J. J., Niemeyer, D., and Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, 45(4):1099–1114.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, 4:151–171.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review*, 7(4):618–630.

Mandera, P., Keuleers, E., and Brysbaert, M. (2017). 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.

Mandera, P., Keuleers, E., Wodniecka, Z., and Brysbaert, M. (2015). SUBTLEX-PL: Subtitle-based word frequency estimates for Polish. *Behavior Research Methods*, 47(2):471–483.

Meyer, D. E. and Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2):227–234.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In Besner, D. and Humphreys, G. W., editors, *Basic processes in reading: Visual word recognition*, pages 264–336. Lawrence Erlbaum Associates.

Neely, J. H. and Keefe, D. E. (1989). Semantic context effects on visual word processing: A hybrid prospective-retrospective processing theory. In Bower, G. H., editor, *Psychology of Learning and Motivation*, volume 24, pages 207–248. Academic Press.

Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

OpenAI (2024). text-embedding-3-small, text-embedding-3-large. https://openai.com/index/new-embedding-models-and-api-updates/.

Posner, M. I. and Snyder, C. R. R. (1975). Attention and cognitive control. In Solso, R. L., editor, *Information processing and cognition: The Loyola symposium*, pages 55–85. Lawrence Erlbaum.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Van Paridon, J. and Thompson, B. (2021). subs2vec: Word embeddings from subtitles in 55 languages. *Behavior Research Methods*, 53:629–655.

A Appendix

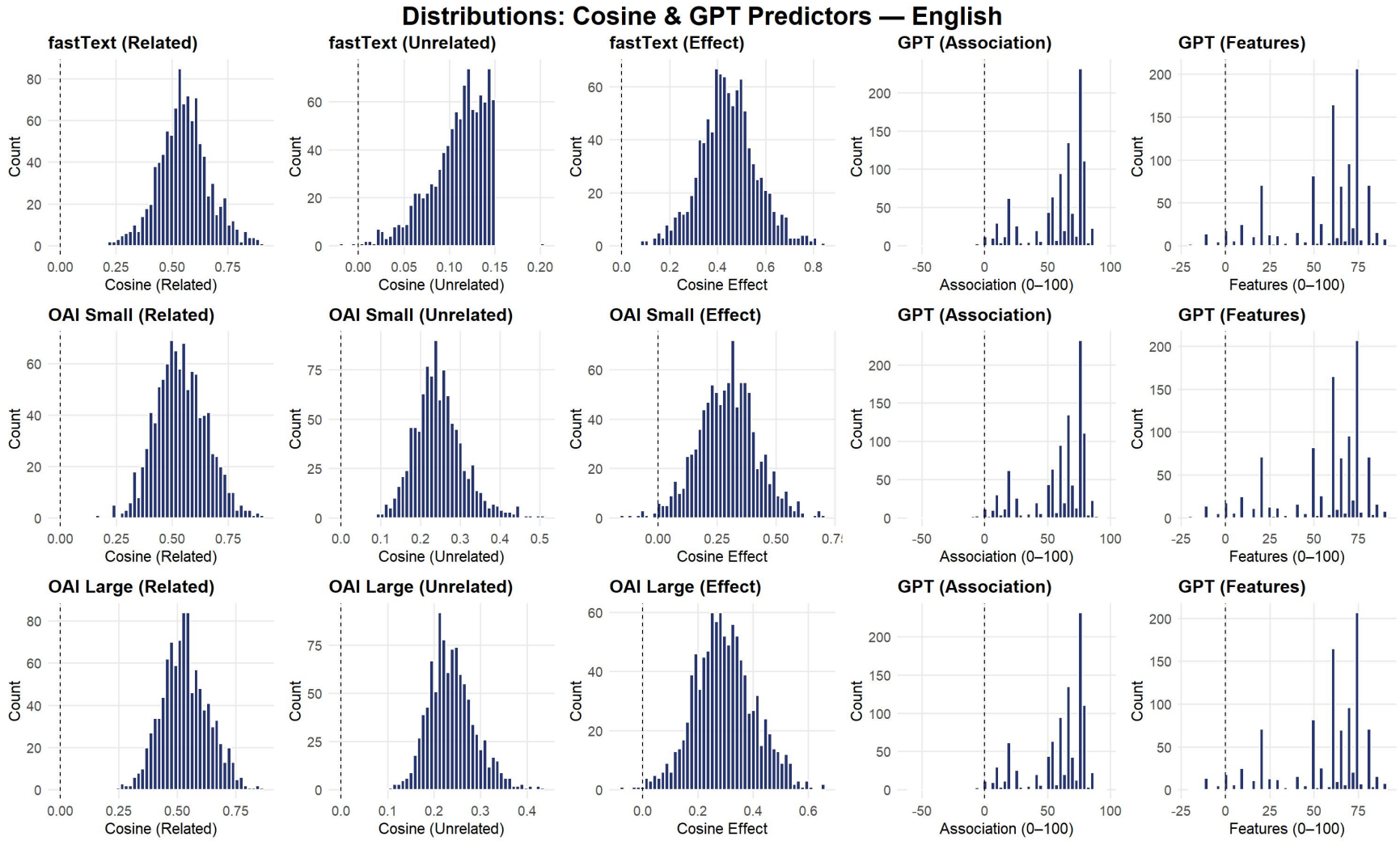

Figure 8: Distributions of the cosine and GPT predictors for English. Columns show, for each embedding model (rows: fastText, OAI small, OAI large), the related and unrelated cosine similarities, their difference (cosine effect), and the GPT association and feature effects.

B Appendix

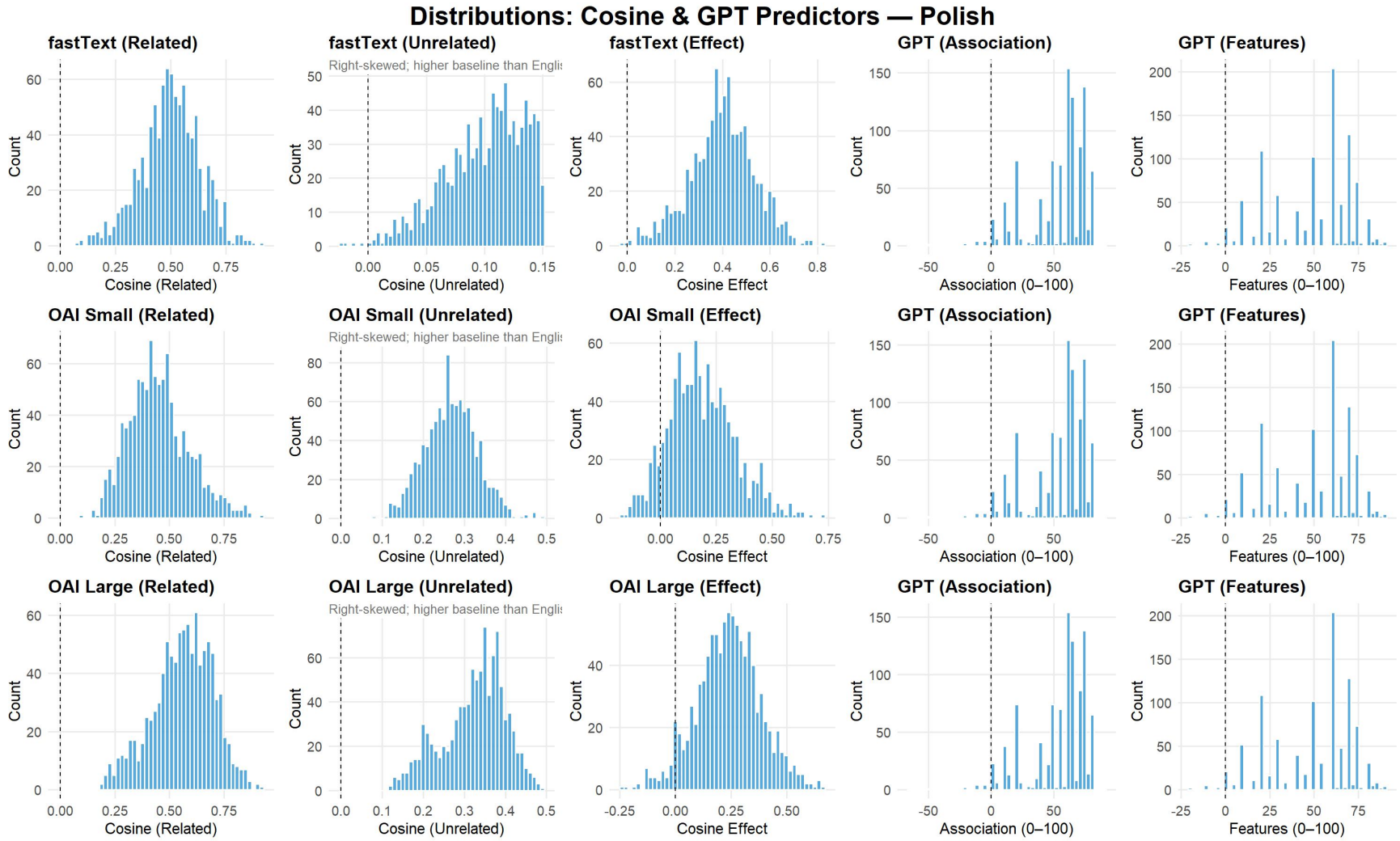

Figure 9: Distributions of the cosine and GPT predictors for Polish. Columns show, for each embedding model (rows: fastText, OAI small, OAI large), the related and unrelated cosine similarities, their difference (cosine effect), and the GPT association and feature effects.

C Appendix

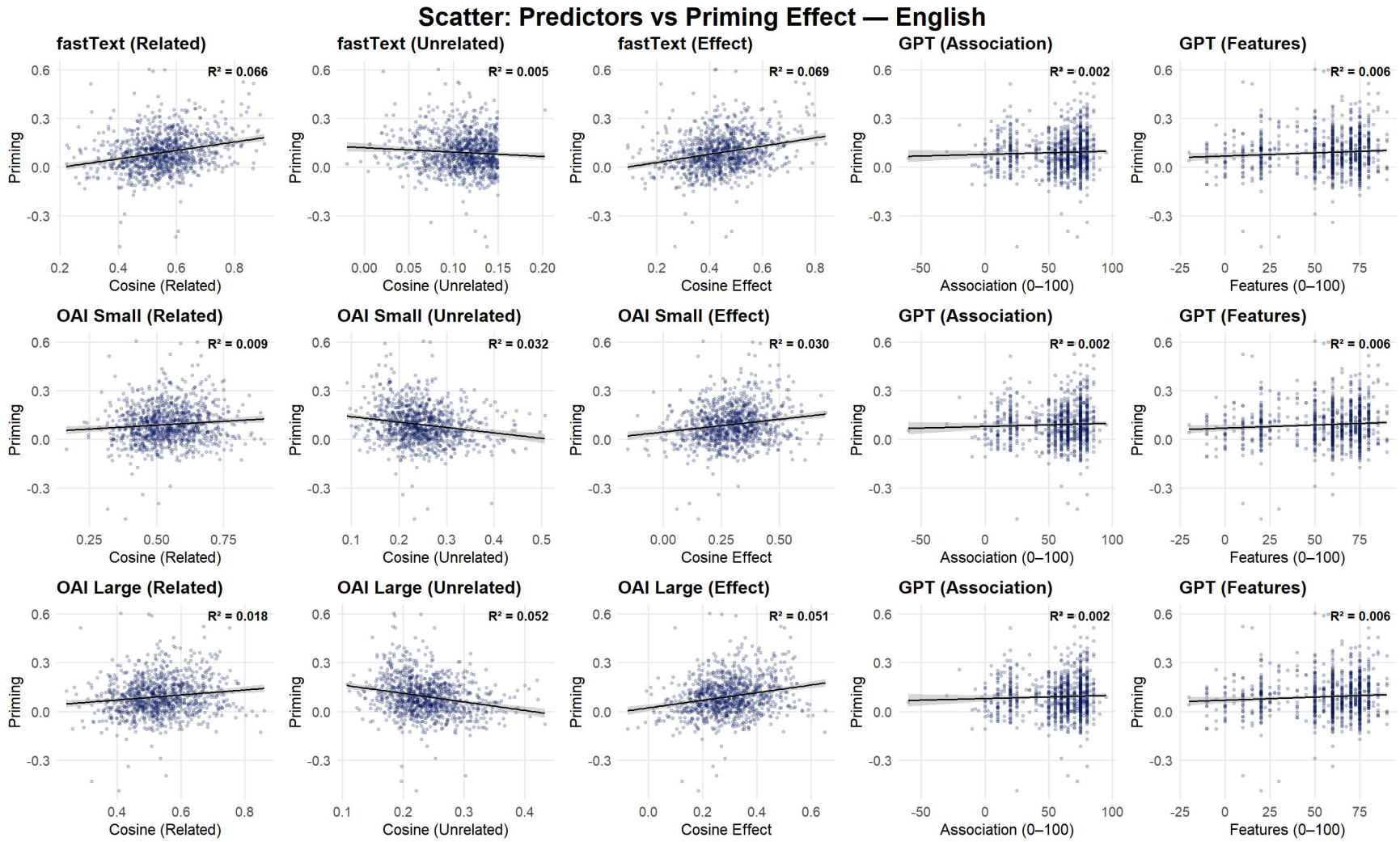

Figure 10: Scatterplots of each predictor against the item-level priming effect for English, with fitted regression lines. Panels arranged per embedding model (rows) across cosine and GPT rating effects (columns).

D Appendix

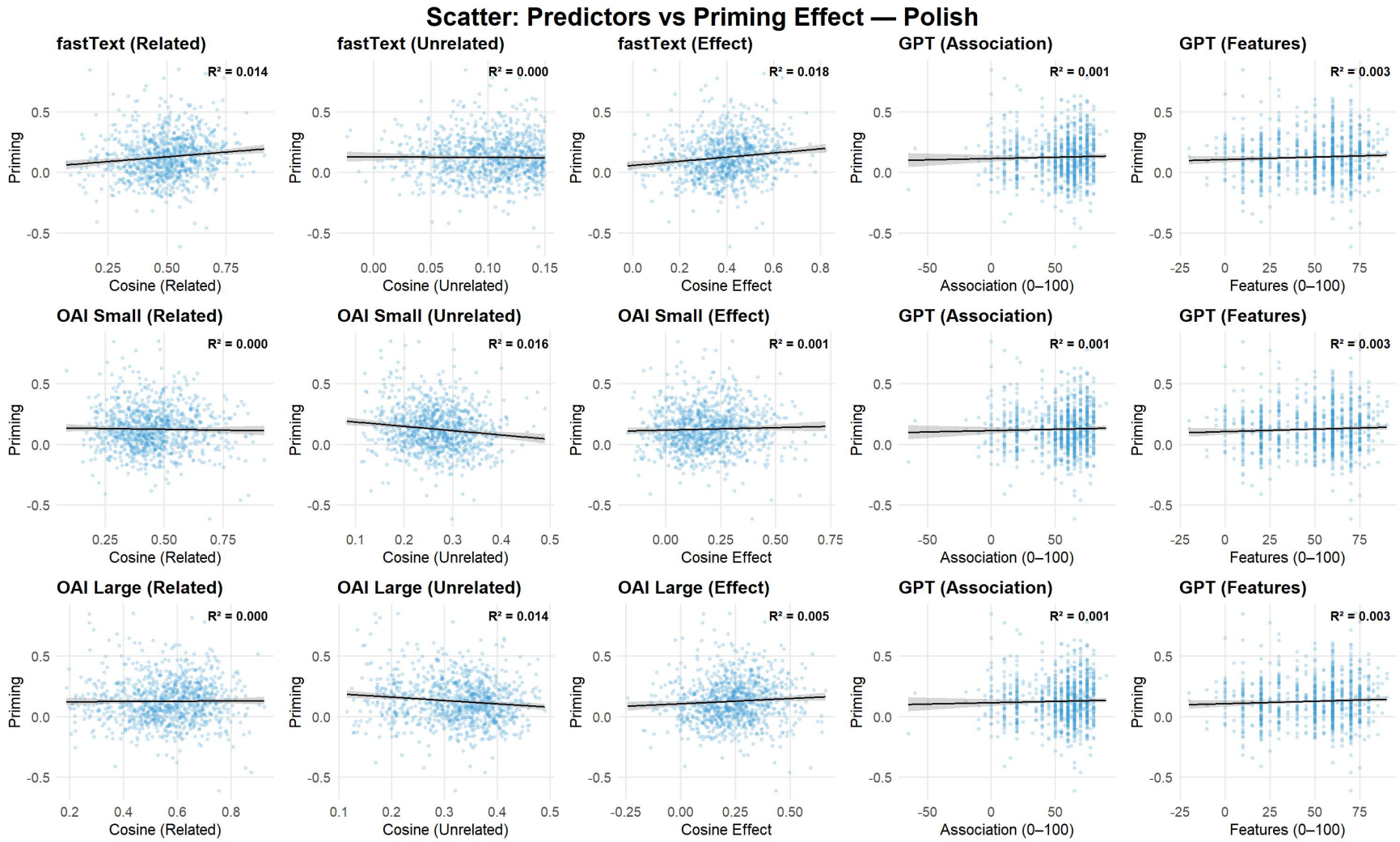

Figure 11: Scatterplots of each predictor against the item-level priming effect for Polish, with fitted regression lines. Panels arranged per embedding model (rows) across cosine and GPT rating effects (columns).