frontiers in Language Sciences

TYPE Original Research
PUBLISHED 23 February 2026
DOI 10.3389/flang.2026.1721326



OPEN ACCESS

EDITED BY
Seana Coulson,
University of California, San Diego, CA, United States

REVIEWED BY
Yifei He,
University of Marburg, Germany
Álvaro Cabana,
Universidad de la República, Uruguay

*CORRESPONDENCE
Maryam Meghdadi
maryam.meghdadi@helmholtz-munich.de

RECEIVED 09 October 2025
REVISED 07 January 2026
ACCEPTED 23 January 2026
PUBLISHED 23 February 2026

CITATION
Meghdadi M, Duff J and Demberg V (2026) Integrating language model embeddings into the ACT-R cognitive modeling framework.
*Front. Lang. Sci.* 5:1721326.
doi: 10.3389/flang.2026.1721326



# Integrating language model embeddings into the ACT-R cognitive modeling framework

Maryam Meghdadi[1,2]*, John Duff[3] and Vera Demberg[1,4]

[1]Department of Language Science and Technology, Saarland University, Saarbrücken, Germany, [2]Institute for Human-Centered AI, Helmholtz Computational Health Center, Munich, Germany, [3]Department of Linguistics, UCLA, Los Angeles, CA, United States, [4]Department of Computer Science, Saarland University, Saarbrücken, Germany

In 2025, psycholinguistic research has the benefit of large, high-quality datasets of human behavior, and massively-scalable metrics for variables of interest like frequency and association. This means we have more data than ever before to shed light on classic language processing phenomena like associative priming. But in order to build and test rigorous theories against this data, we also need computational modeling tools that can simulate cognitive mechanisms and generate quantitative predictions at the same scale. In this paper, we assemble one such case, adapting the ACT-R cognitive modeling framework to make use of association metrics derived from language model embeddings, in service of a scalable model of associative priming in the Lexical Decision Task. ACT-R implements a model of memory retrieval that can use itemwise predictors like frequency and association to predict task response times (RTs), via interpretable and meaningfully-parameterized components like spreading activation. But currently, ACT-R's spreading activation calculations rely on manually-coded similarity scores, which are labor-intensive and prone to inaccuracies, particularly for large vocabularies. In this study, we replace these hand-coded associations with cosine similarity scores derived from Word2Vec and BERT embeddings, thereby improving both scalability and predictive accuracy while retaining ACT-R's interpretability. We compare various versions of our model against observed human RTs from the Semantic Priming Project dataset, observing impressive item-wise prediction accuracy, and achieving the strongest alignment with a model where spreading activation is penalized via a scalable approximation of the classic "fan effect." These findings provide a proof of concept for integrating embedding-based representations into algorithmic-level models of language processing. More than an insight into models of priming, we see this as a first step toward scalable and specific models of more complex phenomena.

KEYWORDS
**ACT-R, associative priming, cognitive modeling, distributional semantics, language models, psycholinguistics**

## 1 Introduction

### 1.1 Associative priming effects in lexical decision

In the Lexical Decision Task (LDT), participants are shown orthographic strings and asked to judge whether they are true words of a given language. It is thought that positive responses require accessing a matching lexical item from the mental lexicon; given this assumption, the task has a long history as a probe for the structure of the mental lexicon.

For instance, it is well-known that correct LDT response times are faster for words which occur with a higher frequency in everyday speech and print (e.g., Whaley, 1978; Balota et al., 2004). Patterns like these have supported models of lexical access as memory retrieval, with words like any other item in memory, whose resting activation levels increase with greater use.

We focus here on how response times in the LDT also change as a function of their context. In particular, response times to a target stimulus exhibit associative priming; that is, target processing is facilitated when preceded by a related stimulus (Meyer and Schvaneveldt, 1971; Traxler and Tooley, 2012). For instance, exposure to the prime "snow" might reduce the response time to a subsequent target "ski," when compared to an alternative prime which is less closely related.

Work in cognitive psychology has studied the phenomenon of associative priming extensively, and many hypotheses have emerged as potential explanations (Hutchison, 2003; Lerner et al., 2012). At long latencies between the prime and the target, there is some evidence that increased target activation is a strategic process, as participants anticipate potential targets (Neely, 1977). Nevertheless, for priming effects at short intervals (e.g., under 300ms), the most prominent hypotheses agree that primes raise the activation of associated targets by some automatic consequence of relatedness in memory. Specific models differ in the structure of memory they assume: models which treat items as atomic nodes in memory have posited automatic spreading activation flowing from active nodes along connections weighted by relatedness (Anderson, 1983; Collins and Loftus, 1975; McNamara and Altarriba, 1988); other models, which treat items as patterns of distributed activations within a neural network, can capture relatedness as simple overlap between representations (Plaut, 1995; Lerner et al., 2012). Researchers within each class of model also debate whether these weighted connections or representational overlaps emerge from shared semantic features, or mere statistical association between items in exposure (see lengthy discussion in Hutchison, 2003).

Consider a spreading activation model where connections are graded by association, like that of Anderson (1983), as it would explain the case of "snow" and "ski." When encountering the orthographic form "snow," a corresponding word will be recognized and retrieved from memory. Once it is focused in attention, "snow" will then spread activation to other items in memory as a function of their distance of association, including a relatively high amount of activation to "ski." Consequently, if "ski" is presented shortly afterwards, it will benefit from this temporarily-boosted activation, yielding faster response times.

These models all agree that effects of associative priming should be gradient along a scale of prime-target relatedness. In the past two decades, regression analyses on large-scale datasets have offered strong support for this gradience: controlling for the influence of other known variables like frequency and orthographic complexity, as various measures of relatedness increase, target response times in the LDT decrease (Hutchison et al., 2008; Günther et al., 2016; Mandera et al., 2017). These kinds of analyses have become especially noteworthy in the past ten years, as work in natural language processing identifies more and more powerful methods for modeling associative strengths.

But as our tools for modeling association have jumped forward, we observe that there has been little work incorporating these large-scale measures into computational cognitive models that specify the links between our models of memory and predicted performance in the LDT. Without this type of model, we miss an exciting chance to bring our new abilities to study priming at scale to bear on the fine-grained questions at the heart of this literature, like the particular dynamics of activation across the lexicon.

Part of the gap we identify may be due to a disconnect between the modeling tools which have been successful for treating fine-grained architectural questions, and the scale of the data which is available for modeling. Cognitive psychology has developed computational architectures like Adaptive Control of Thought–Rational (ACT-R; Anderson and Schooler, 1991; Anderson and Lebiere, 1998; Anderson et al., 2004) which are specialized to extract fine-grained response time predictions from hypotheses of cognitive organization and procedure. ACT-R in particular provides a detailed model of spreading activation, elaborated from the original model in Anderson (1983), and embedded within its treatment of memory retrieval and activation. Nevertheless, in its standard form, the spreading activation relies on overlap between hand-coded attributes, and cannot be easily extended to predict item-specific priming data.

To address this gap, we present here an elementary model of associative priming in lexical decision which enriches ACT-R with high-quality, scalable measures of target-prime relatedness derived from language model embeddings. As we will show, the resulting model is able to generate item-specific predictions for priming effects at a large scale, for a variety of user-specified configurations. As such, we can compare how various hypothetical mechanisms impact the quality of fit to a large dataset (the Semantic Priming Project; Hutchison et al., 2013), capturing the possible contributions of length, frequency, association, associative density, and their interactions. The best version of our model reconstructs impressively human-like patterns along all these dimensions. We intend this model as a proof of concept, highlighting how frameworks like ACT-R, once adapted appropriately, remain useful tools for specific and explainable computational cognitive modeling at the interface between large-scale predictors and large-scale human data.

## 1.2 Scalable measures of association

Researchers interested in measuring relatedness, either as a product of semantic similarity or association, have often turned to the methods of distributional semantics. These methods adopt Harris's (1954) famous distributional hypothesis, which posits that words with similar meanings tend to occur in similar contexts. In their original form, these approaches were count-based models, which map each word to the number of co-occurrences of words in particular contexts in a matrix and apply transformation functions on the output matrix (Pennington et al., 2014; Turney and Pantel, 2010; Landauer and Dumais, 1997; Lund et al., 1996). Measures of similarity could then be extracted from the resulting vectors, most typically *cosine similarity* (CoSim), the cosine between two

vectors, with values approaching one for highly similar word pairs and nearing zero for unrelated pairs. Such measures, extracted from early distributional models like LSA (Landauer and Dumais, 1997) and HAL (Lund et al., 1996), have then been tested for their value as predictors of priming, either to prove their validity as measures of semantic meaning, or as a way to evaluate theories of priming itself. This expected relationship is often observed (Arambel and Chiarello, 2006; Chwilla and Kolk, 2002; Günther et al., 2016), although not always (Hutchison et al., 2008). A second wave of this research has adopted a more modern variant of distributional semantic vectors, assembled not directly through text statistics, but instead through machine learning on prediction tasks, as in Word2Vec (Mikolov et al., 2013a,b). Word2Vec embeddings are fit to maximize a simple neural network's correct prediction of a target word given a window of context words (Continuous Bag of Words–CBOW) or context words given a target word (skip-gram). While these models were not originally designed for psychological tasks, some researchers have considered them to be more cognitively plausible than count-based models (Mandera et al., 2017; Günther et al., 2019), with reference to the hypothesis that human language-users may also learn lexical representations through context-based prediction. Mandera et al. (2017) demonstrate that CoSim over CBOW Word2Vec embeddings performs remarkably well as a predictor of item-level response times for the 200ms stimulus onset asynchrony LDT data from the large Semantic Priming Project (SPP) dataset (Hutchison et al., 2013).

In recent years, predictive language modeling has advanced beyond the basic architecture used in Word2Vec, providing more advanced methods for encoding lexical meaning, and with them, a new alternative for calculating relatedness. While the vector-based methods above provide static, type-level embeddings, transformer-based large language models (LLMs) leverage an attention mechanism to produce contextualized embeddings for every token in a string (token-level and dynamic). Previous work, such as Vulić et al. (2020b), has demonstrated that LLMs also capture lexical semantic knowledge. In this work, the authors use pre-trained LLMs with type-level inputs in two distinct ways:

i) Decontextualized or isolated embeddings (ISO): a target word is input in isolation (with or without special tokens),[1] as illustrated in the left panel of Figure 1. The resulting ("ISO") embeddings are decontextualized and capture the meaning of words in isolation.
ii) Contextualized or average-over-context (AOC) embeddings: a target word is input within a sentence, as shown in the right panel of Figure 1. The LLM outputs an embedding vector for each token in the sentence. The target word embeddings are then extracted from these vectors, and their average across multiple sentences is computed to produce a single embedding for the target word. These ("AOC") embeddings represent typical word meaning across diverse sentential contexts.

1 Although it may seem counter-intuitive to include special tokens here at all, they may in principle be relevant for model representations of meaning in isolated word contexts, and previous work in related tasks (Vulić et al., 2020b) has demonstrated they may be beneficial.

Previous work has applied these type-level embedding extraction methods to BERT models, and has found impressive performance in alignment between CoSim and human similarity ratings (Vulić et al., 2020a; Bommasani et al., 2020). Cassani et al. (2023) demonstrates that BERT-based CoSims of both types are also predictive of response times in the Semantic Priming Project database, although they do not match the predictive power of Word2Vec CoSims.

In sum, vector-based models of lexical meaning have been fruitful as a source for quantitative measures of association, fueling regression-based analyses which have validated the presence of gradient priming effects. Nevertheless, this result offers only limited insight for theories of lexical meaning in cognitive science—all prominent theories considered in the literature would expect that (some sense of) association should serve as a continuous predictor of priming effects. There is more that can be learned here, if we look more closely. These scalable metrics of association, in conjunction with large datasets like the Semantic Priming Project, have the potential to allow us to compare specific theories of associative priming. In order to use such metrics for this purpose, it is necessary to construct computational cognitive models which can interface with these metrics. To this end, we shift our focus to models of lexical access in ACT-R, and how they can be adapted to interface with lexical embeddings.

## 1.3 Modeling lexical decision performance in ACT-R

The ACT-R framework is a cognitive architecture that integrates basic cognitive mechanisms based on empirical findings, allowing for predictions about more complex cognitive phenomena, including language comprehension (Anderson and Schooler, 1991; Anderson and Lebiere, 1998; Anderson et al., 2004).

ACT-R conceptualizes the human mind as a system of modules and their associated buffers. Modules represent mental functions, and each module is associated with a buffer for processing input and output information.

The main modules in ACT-R represent *declarative* and *procedural* memories in humans. Declarative memory stores factual information about the world, such as the fact that Berlin is the capital of Germany. In ACT-R, this knowledge is represented in *chunks* of attribute-value pairs in declarative memory. For example, the word "snow" may be stored as: [Form: snow, Meaning:⟦snow⟧, Domain: winter, Number: singular]. The associated buffer, the retrieval buffer, is responsible for retrieving chunks from the declarative memory module based on their content, and keeping them temporarily accessible to other processes. Procedural memory, on the other hand, deals with implicit knowledge such as how to perform tasks (e.g., driving). ACT-R models these processes using production rules (if-then statements), and the goal buffer helps regulate which rules fire at which state of the whole procedure. Elsewhere, visual buffers are responsible for encoding representations of information in the visual field. Typically, models also employ an imaginal buffer, which maintains longer-lasting

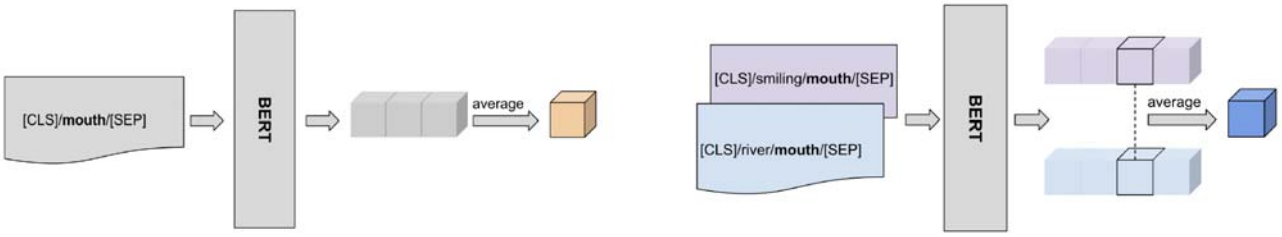


FIGURE 1
**(Left)** ISO embeddings of the target word "mouth," **(Right)** AOC embeddings of the target word "mouth."

internal representations of the contextual information which is relevant to the current cognitive process.

We can use cognitive frameworks like ACT-R to simulate a set of component procedures which we assume lie within a participant's LDT performance (cf. van Rijn and Anderson, 2003):

- Locating the orthographic stimulus on the screen and visually encoding it,
- Attempting to retrieve a matching target word from declarative memory, where chunk activations are influenced by the recent processing of the prime word,
- Executing a motor command to press the appropriate key (e.g., "Y" or "N"), based on whether a matching word was successfully retrieved.

ACT-R models these steps serially, with each component requiring a specific amount of time. The latency of each step would then contribute to the total response time (RT) to the target word. The framework already implements a version of spreading activation, which would affect total RT by decreasing the latency of target retrieval in the second step. In the next section, we break down the details of this spreading activation mechanism, before going on to explain how we might adapt it to our present needs.

### 1.3.1 Modeling chunk activation in ACT-R

In ACT-R, the chunks in the imaginal buffer store contextual information (here the prime word), and spread activation to semantically-related chunks in the declarative memory (like other words in the lexicon). The amount of spreading activation depends on the strength of this association. The activation of a chunk is determined by its past usage and influences both the likelihood of successful retrieval and the time required for retrieval (Anderson and Schooler, 1991; Anderson and Lebiere, 1998). This activation is composed of two main components: base-level activation, reflecting the frequency and recency of the chunk's use, and spreading activation, which depends on the context in which the chunk has been used. More specifically, ACT-R formulates the activation of each chunk as below (Anderson and Lebiere, 1998; Lewis and Vasishth, 2005):

$$A_i = B_i + \sum_{j=1}^{C} \sum_{x=1}^{X_{ij}} W_j^x S_{ji}^x + \epsilon \tag{1}$$

where,

- $B_i$ is the base-level activation. The more frequent or recent a chunk is, the higher its $B_i$.
- $S_{ji}^x$ is the strength of association between the value of attribute $x$ of chunk $i$ and attribute $x$ of some chunk $j$, across all shared attributes of $i$ and $j$ ($X_{ij}$) and across all $C$ chunks in attention. Here, we will only ever treat one chunk in attention, the item in the imaginal buffer.
- $W_j^x$ is the weight of the spreading activation from attribute $x$ of source chunk $j$. By default, this is even across all attributes of all source chunks.
- $\epsilon$ is instantaneous noise.

The strength of association $S_{ji}^x$ is formulated as:

$$S_{ji}^x = \begin{cases} 0, & \text{if attribute } x \text{ does not match between chunk } j \text{ and chunk } i, \\ S - \log(\text{fan}_j^x), & \text{otherwise.} \end{cases} \tag{2}$$

where,

- $S$ is the maximum association strength, a user-defined constant.
- $\text{fan}_j^x$ is the total number of chunks in the declarative memory which match $j$'s value for attribute $x$, including chunk $j$ itself.

As shown in the formula, the number of associated chunks ($\text{fan}_j^x$) penalizes the strength of association. In other words, when more chunks in memory share a given attribute-value pair, the activation associated with that match spreads across a larger number of chunks, reducing the effect of spreading activation on any one chunk. This phenomenon is known as the *fan effect*.

Consider the associative priming example between the "snow" and "ski" chunks in Figure 2. These chunks share the value of their domain attribute. As a result, there will be non-zero spreading activation from "snow" to "ski" for the domain attribute, $S_{\text{snow, ski}}^{\text{domain}} \neq 0$. The maximal association strength for a matching attribute is $S$, but this will be penalized based on the size of the fan of the domain value "winter." If the chunk "ski" is the only other chunk in declarative memory with domain "winter," $\text{fan}_{\text{snow}}^{\text{domain}}$ will be 2 (for "ski," and "snow" itself). If there is just one additional chunk with

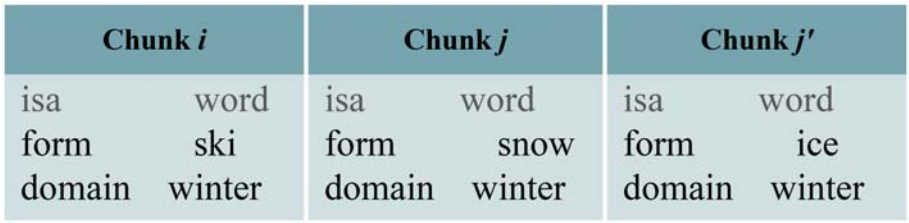


FIGURE 2
An example of target chunk (*i*) and existing chunks in declarative memory (*j* and *j*′). In ACT-R's notation, *isa* shows the type of the chunk's content.

domain "winter" in declarative memory, such as *j*′, then $\text{fan}_{\text{snow}}^{\text{domain}}$ would be 3, and so on.

Equations 1, 2 capture how frequency, recency, and associations of a chunk can affect its activation in memory. For instance, a word with higher frequency has a higher base-level activation, resulting in a higher total activation. Similarly, a word which shares an attribute with a word in the focus of attention will have a higher $S_{ji}^{x}$, leading to increased activation.

As modeled by Anderson and Schooler (1991), there is an inverse exponential relationship between a chunk's activation and its retrieval time. ACT-R computes the retrieval time for chunk *i* as follows:

$$T_i = Fe^{-fA_i} \tag{3}$$

where *F* is the latency factor, and *f* is the latency exponent, two free parameters available for model adjustment. In this formulation, as activation increases, retrieval time decreases exponentially; conversely, as activation decreases, retrieval time exponentially rises. In particular, chunks receiving higher levels of spreading activation will be retrieved faster due to their greater total activation.

In this way, ACT-R has the potential to model the associative priming effect, provided that the total weighted similarity of all attributes of the prime and the target chunk ($S_{ji}$) effectively represents the similarity between prime and target words. However, with this implementation of chunks and spreading activation in ACT-R, representing gradient similarity between chunks requires adding a large quantity of hand-coded, discrete features for every chunk. This is neither realistic nor efficiently scalable. In contrast, the flexible similarity metrics described in the last section are a more promising candidate for automatically generating values for $S_{ji}$. In the next section, we introduce an adaptation of ACT-R's spreading activation which can make use of this alternative definition of similarity.

# 2 Materials and methods

We model the associative priming effect using ACT-R considering different components in the modeling process. First, we define three baseline models based on simple components contributing to reading time and to baseline activation. These will serve as comparison points to help isolate the portions of the dataset which must be explained by a spreading activation component. In the rest of the section, we will introduce our method to improve the spreading activation in ACT-R, and the three target models which incorporate this form of spreading activation.

## 2.1 Establishing the baselines

We will investigate the following baselines:

- Baseline B1: is an ACT-R model which considers only the visual processing difficulty of the target in a lexical decision task, neglecting the components of the chunk activation, i.e. all the chunks have the same activation time (the default value). This difficulty is computed based on the Eye Movements and Movement of Attention (EMMA) model (Salvucci, 2001).
- Baseline B2: is an ACT-R model that only considers the frequency of target words affecting the base-level activation in a lexical decision task, setting the EMMA and spreading activation components to their default values.
- Baseline B3: is an ACT-R model that combines the EMMA and frequency components, setting the spreading activation to its default value.

### 2.1.1 How do we compute visual processing difficulty using EMMA?

Following the EMMA model (Salvucci, 2001), as used in Reichle et al. (1999) and Dotlačil (2018), a shift of attention to a visual object triggers two processes: (1) an immediate attempt to encode the object as an internal representation, and (2) a corresponding eye movement. The time required for visual encoding is given by Equation 4:

$$T_{\text{enc}} = KDe^{kd} \tag{4}$$

Here, *K* and *k* are free parameters, *D* represents the reading difficulty of the object, and *d* is the distance (in degrees of visual angle) between the current focal point and the target object.

In our models, we set *D* to the length of the target word, and use the default values of $K = 0.01$ and $k = 1$. Given a small arbitrary $d = 0.32^{\circ}$ (all target words were presented in the same location on the simulated screen, slightly displaced from center), our $T_{\text{enc}}$ amounts to about 14ms per character.

### 2.1.2 How do we compute baseline activation as a function of frequency?

We obtain the frequency of each target word from the `wordfreq` database (Speer, 2022) and group them into 100 linearly spaced bins, such that each word is assigned to one frequency band.[2] To estimate the base activation associated with these frequencies, we follow the approach of Brasoveanu and Dotlačil (2020), who simulate a simple lexical decision task. In their setup, they use average word frequencies across 16 frequency bands to estimate the number of times a 15-year-old speaker has been exposed to words in each band.

Following Brasoveanu and Dotlačil (2020), we adopt the average number of words a 15-year-old child has been exposed to (≈112.5 million words) as an estimate of total language exposure. This value is not intended to reflect absolute word frequency effects, but rather to provide a relative scaling of exposure across frequency bands. To simulate this exposure, we compute the total number of seconds in the 15-year lifespan and use the mean frequency vector of the target words to generate a schedule of linearly spaced word presentations (or "rehearsals") across frequency bands. These schedules then serve as the input to standard ACT-R equations for baseline activation $B_i$ as a function of previous exposure.

## 2.2 Integrating LMs into ACT-R

As previously mentioned, ACT-R does not inherently model similarities between words, requiring manual encoding of these similarities to simulate the associative priming effect. To address this limitation and improve the efficiency of such models, we integrate language model embeddings into the ACT-R framework. To ensure the embeddings effectively capture human-like patterns in the data, we evaluate the human-likeness of several high-performing language models to identify the most suitable one for our ACT-R model of the associative priming effect. In this section, we first introduce the language models selected for experimentation and then explain how we integrate these models into the ACT-R framework.

### 2.2.1 Language model selection

We select the following language models (LMs) to build on one well-established measure of association, compare the value of an alternative measure from a widely-used LLM, and probe the applicability of a more recent state-of-the-art LLM:

i) Word2vec (Mikolov et al., 2013a): this static embedding model establishes a starting point grounded in previous psycholinguistic investigations of vector-based measures of similarity. We use the embeddings from Mandera et al. (2017), which have been validated on psycholinguistic tasks. These 300-dimensional word vectors were trained on a concatenation of the UkWaC and SUBTLEX corpora, using a context window of 5 words to the left and right of the target word.

ii) BERT-base-uncased (Devlin, 2018): with 12 layers of transformers containing bidirectional encoders, and a vocabulary size of 30,522 768-dimensional embeddings, BERT is a more complex source for vector representations of meaning. Recent work has established BERT embeddings as an effective measure of lexical association standing near or beyond Word2Vec (Cassani et al., 2023; Vulić et al., 2020b; Bommasani et al., 2020), and including them here allows us to continue these ongoing comparisons. Following Timkey and van Schijndel (2021), we normalize each BERT layer across the corpus to prevent a few dimensions with disproportionately high variance from dominating, which can reduce the representational quality of other dimensions when measured using cosine similarity. See Appendix A.3 for more details.

iii) State-of-the-art LLMs: in Appendix A, we describe the performance of other state-of-the-art LLMs, such as Gemma-2B (Gemma Team et al., 2024), and LLM2Vec (BehnamGhader et al., 2024). We do not describe them further in the main body of the paper, as they did not outperform BERT in our experiments.

For the more complicated models (ii and iii), there are many options for extracting embeddings, including various methods of producing type-level embeddings (see above), but also the choice of which layer to extract embeddings from, and whether special tokens marking the beginning and end of an input string should be included in the embedding contexts (see Vulić et al., 2020a and Bommasani et al., 2020 for more detail). Before examining the performance of our ACT-R model, we established the likely best-performing options for each model through preliminary correlations between cosine similarity and response times, as described in Section 3.2.

### 2.2.2 Proposed spreading activation

After identifying the optimal embeddings for the associative priming effect, we integrate these embeddings into our ACT-R model. We follow previous work (Mandera et al., 2017, among many others), and use the cosine between two embeddings as a measure of their similarity. Accordingly, we can replace Equation 2 with Equation 5, so that the strength of association ($S_{ji}$) is modulated by the cosine similarity (CoSim) between the embeddings of the prime ($V_i$) and target ($V_j$) words, higher for more similar prime-target pairs.

Note that we map all negative cosine values to 0, ensuring that $S_{ji}$ lies within the range $[0, S]$. If the prime and target words are not semantically related (or are negatively related), the term $\max\{0, \cos(V_i, V_j)\}$ evaluates to 0, leading to $S_{ji} = 0$. Conversely,

2 Since computing activations for each individual word frequency is computationally expensive, we aggregate words into frequency bands to approximate frequency effects more efficiently. We chose 100 bins as it met our precision goals for a proof of concept.

if the prime and target words are highly related, their cosine can reach 1, resulting in $S_{ji} = S$.[3]

$$S_{ji} = S * \max\{0, \cos(V_i, V_j)\} \quad (5)$$

To incorporate the potential for a fan effect, we extend Equation 5 to Equation 6, emulating $\log(\text{fan}_j)$ (used in Equation 2) with the average cosine similarity between the prime word embedding ($V_j$) and its top $n$ (a free parameter) associated neighbors present in declarative memory. This derives smaller values of $S_{ji}$ from primes $j$ which have many near neighbors.

$$S_{ji} = S * \max\{0, \cos(V_i, V_j) - \frac{1}{n-1} \sum_{k \neq i} cos(V_k, V_j)\} \quad (6)$$

While Smith and Vasishth (2020) suggest this way of approximating fan effects from cosine similarity data, to our knowledge our model is the first to implement it.

| ACT-R Model | EMMA | Frequency | CoSim | Fan |
|---|---|---|---|---|
| B1 | ✓ | | | |
| B2 | | ✓ | | |
| B3 | ✓ | ✓ | | |
| M1 | | | ✓ | |
| M2 | ✓ | ✓ | ✓ | |
| M3 | ✓ | ✓ | ✓ | ✓ |

FIGURE 3
Baseline and target models.

## 2.3 Target models

In our model experiments, we will investigate the predictions of three target ACT-R models, comparing against each other and against the baselines mentioned above.

- Target M1: only considers simple spreading activation based on CoSim in Equation 5 This model allows us to investigate the effectiveness of CoSim on its own.
- Target M2: considers a combination of visual processing, baseline activation, and simple spreading activation (Equation 5). This model is a fair representation of ACT-R's standard strengths, as it integrates the baseline components with CoSim, all within a specified model of the task.
- Target M3: considers a combination of visual processing, baseline activation, and spreading activation subject to a fan effect (Equation 6). This model allows us to compare the evidence for fan effects as a component of the spreading activation mechanism.

These target models allow us to present a few key comparisons of interest. First, comparing between target models M1 and M2 will allow us to see the value of predictions coming from several integrated components within an ACT-R model, as compared to a model which merely aims to capture a target effect.

More interestingly, comparing models M2 and M3 allows us to isolate the contributions of fan effect penalties within this system of spreading activation. Fan effects have an important status in the broader literature, as they provide a mechanism for spreading activation theories to predict cases of asymmetric priming, such that a prime "termite" facilitates target "wood" (one of its few highly-associated items) more than a target "wood" facilitates "termite" (one of many highly-associated items among a large fan). Although some experiments probing these conditions have produced such findings (e.g. Hutchison, 2002), in many cases researchers have observed 'backwards' priming effects ("wood" priming "termite") of the same size as the canonical 'forwards' effects, especially in lexical decision; see Hutchison (2003) for lengthy discussion. This data might suggest a fundamental flaw of the spreading activation approach. Nevertheless, to date, work on asymmetrical priming has focused on hand-selected items; there has been no investigation of associative density as a component of itemwise response time predictions in a large corpus like the Semantic Priming Project. It remains possible that stimulus selection or a focus on conditional differences in response times may be arbitrarily concealing a role for fan effects. Our target model M3 thus represents a first attempt to measure the evidence for asymmetrical priming at scale.

We will test each of the target models with both Word2Vec and (the best layer of) normalized BERT embeddings. For an overview of the models see Figure 3.

# 3 Results

In this section, we first introduce the dataset we use for measuring priming effects in the Lexical Decision Task (LDT). Next, in order to select the language model embeddings that would be most suitable for modeling this task, we conduct an initial comparison of the simple correlations between CoSim and LDT response times across a variety of embedding types (this first step was done independently of our ACT-R models). Finally, we examine the performance of our baseline and target ACT-R models, equipped with the best embeddings from our initial comparison,

3 We opted for this approach because in our experiments, most cosine similarity values were positive, with only a few small negative values very close to zero. While an alternative method would have been to scale all cosine values to the range $[0, 1]$, we opted for the maximum operation for its higher flexibility. Scaling requires precomputing all cosine values to determine the scaling factors, whereas using the maximum allows for on-the-fly computation, resulting in a more adaptable ACT-R model.

as a case study in how procedural models can bridge the gap from large-scale statistical patterns toward mechanistic insight into language processing. The code for both steps is publicly available at https://github.com/maryam97/LMs_in_ACT-R.

## 3.1 Dataset

We assess the priming effect using lexical decision data from the Semantic Priming Project (SPP) dataset (Hutchison et al., 2013). The SPP comprises 6,641[4] prime-target pairs collected from 512 native English speakers (mean age = 21.14 years; mean education = 13.68 years) across four universities in different regions of the United States. Pairs were presented to participants to compare various experimental conditions, including relatedness (related vs. unrelated), stimulus onset asynchrony (SOA; short vs. long), prime type (first/most common associate vs. other associate), and target lexicality (word vs. non-word).

We analyze here only correct responses to trials where the target was a word. In the main text, we further focus on the subset of data where prime words preceded the targets at a short SOA of 200 ms (SPP-Short). Each prime-target pair in this subset received between 2 and 33 observations. Previous work agrees that this short SOA is more likely to isolate the automatic portions of priming (Neely, 1977; Hutchison, 2003). Cassani et al. (2023) compare BERT cosine similarities as predictors for SPP-Short and also additional 1,200 ms SOA data from the Semantic Priming Project (SPP-Long). They observe that the former is much better captured. The same is true for our models in a supplementary analysis comparing SOAs, see Appendix A.

## 3.2 Language model selection

In this first experiment, we evaluate the performance of CoSim from BERT and Word2Vec embeddings as simple predictors for the SPP-Short dataset. This step enables us to identify the most suitable embeddings for simulating the priming effect within the ACT-R framework in the subsequent experiment.

### 3.2.1 Implementation

We extract Word2Vec embeddings, as well as normalized ISO and AOC embeddings from each individual layer of BERT, with and without including special tokens, for all prime and target words in the SPP-Short dataset.[5] For the AOC embeddings, we take the average over 50 contextual texts for each word sourced from the FineWeb corpus (Penedo et al., 2024).

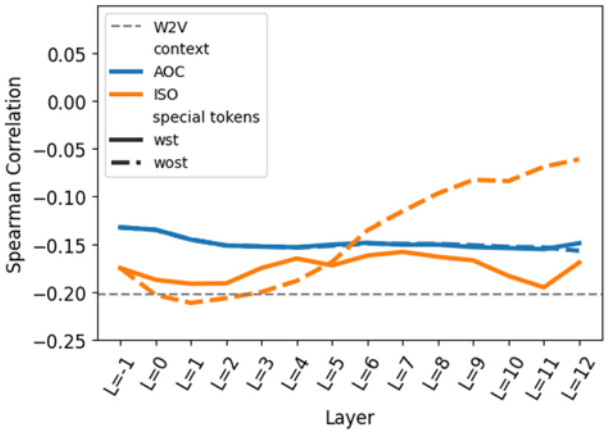


FIGURE 4
Spearman correlation results for Word2Vec and all layers of BERT (normalized) on the SPP-Short dataset. "wst" and "wost" indicate embeddings with and without special tokens, respectively.

### 3.2.2 Evaluation

We evaluate the performance of embeddings from Word2Vec and each BERT source on the SPP-Short dataset by calculating Spearman rank correlation between the cosine similarities of prime–target embeddings and RTs.[6] Higher similarity between prime–target pairs is expected to correspond to shorter RTs. Therefore, more similar word pairs should exhibit stronger negative correlations, while less similar pairs should show weaker correlations. Given this expected negative relationship between RTs and cosine similarities, the LM with the strongest negative correlation would demonstrate the best performance. We compare these correlations across BERT and Word2Vec embeddings.

### 3.2.3 Findings

Figure 4 presents the Spearman correlations for normalized BERT embeddings on the SPP-Short dataset across different layers,[7] compared to Word2Vec embeddings. The results reveal that BERT ISO embeddings without special tokens perform best in the early to middle layers, achieving the highest correlation at layer $L = 1$ ($\rho = -0.21$). Word2Vec embeddings follow closely, with a correlation of $\rho = -0.20$. Based on these findings, we select Word2Vec and normalized BERT-L1 ISO embeddings without special tokens for our ACT-R experiments. See Appendix A for further discussion and comparisons, and Appendix B for supplemental comparisons among these models on an explicit semantic similarity task.

4 We use 6,631 pairs for ISO embeddings, as 10 pairs were not included in the Word2Vec embeddings provided by Mandera et al. (2017). For AOC embeddings, we use 5,521 pairs after removing low-context words. Excluding the same pairs for ISO embeddings did not affect their performance; hence, we report ISO results with 6,631 pairs.

5 Both ISO and AOC embeddings were extracted using an adapted version of https://github.com/MilaNLProc/psycho-embeddings.

6 It is worth noting that Cassani et al. (2023) report Pearson's correlation, which assumes a linear relationship between RTs and cosine similarities—a condition that may not necessarily hold in this context.

7 In this notation, $L = -1$ corresponds to the static embedding layer before positional encoding, $L = 0$ corresponds to the static embedding layer after positional encoding, and $L = 1$ corresponds to the output after the first layer of transformers.

## 3.3 ACT-R models

Using the selected LMs, we construct our ACT-R models, and test their performance on the SPP-Short dataset. In the following, we describe the setup, dataset, and implementation for an ACT-R model with all the components (i.e. the M3 model).

Finally, we report the results for all the baseline and target models.

### 3.3.1 Implementation

We define our ACT-R models using the pyactr package in Python (Brasoveanu and Dotlačil, 2020). We also adopt those authors' fitted values for the latency factor and latency exposure parameters in Equation 3 ($F = 0.379287$, $f = 0.363791$), and an additional decay parameter used in the calculation of baseline activation $B_i$ from frequency (0.153496). They obtain these values from Bayesian model-fitting of the frequency curve in (unprimed) lexical decision. To set the final free parameter, maximum association strength $S$, we run a grid-search over $[1.0 \ldots 10.0]$ for each model variant, and report the parameterization with the highest performance. To simplify the setup, we set the instantaneous noise to $\epsilon = 0.0$, ensuring consistent results across runs, as we are mainly interested in the model's best prediction for an average response to each stimulus. While this eliminates variability for the current study, future experiments may incorporate positive noise to examine predictions as distributions rather than points, a la Balota et al. (2008).

As described above, baseline activations (where included) are calculated based on the mean frequency of the target word's corresponding frequency band from wordfreq (Speer, 2022). Fan estimates for Equation 6 (where included) come from the average BERT/Word2Vec CoSim between the prime word and its 20 nearest neighbors in the Mandera et al. (2017) Word2Vec space, extracted using those authors' online tools.[8]

In this ACT-R model, first, we initialize declarative memory to include chunks for the prime and target words. Then, we simulate participants' behavior during their response to the target word, assuming the prime word has already been processed and is currently being maintained in the imaginal buffer, controlling spreading activation. Upon presentation of the target word string on the screen, the model follows defined production rules to (i) visually find and attend to the string, (ii) encode the string in an intermediate buffer in memory, (iii) attempt to retrieve a corresponding target word from declarative memory, and (iv) initiate a motor command to press the appropriate key depending on whether retrieval is successful. The total response time is measured from the moment the model begins reading the target word on the screen until it successfully presses a key. Except for the effect of length on visual encoding, variation in this latency depends primarily on the retrieval time, as discussed in Section 1.3.1, which is modulated by baseline and spreading activation (see Equations 1, 3). For the current experiment, each observation was simulated in isolation, with no effects of context other than those coming from the relevant prime.

### 3.3.2 Evaluation

We evaluate the models' predictions by comparing the predicted RTs with the observed response times in the SPP-Short dataset using two measures. Our primary measure, Spearman rank correlation, evaluates the degree to which each model reproduces the observed rankings of item-wise difficulty. While this is a liberal measure of model fit, we use it here to abstract away from deflated performance due to (1) any potential non-linear divergence between prediction and observation, particularly at extreme values,[9] and (2) free parameters which were not tuned for maximal performance of these models on this dataset. As a more stringent, secondary measure, we also present Root Mean Squared Error (RMSE), to evaluate the degree to which our model reproduces the absolute RTs for each item. We note that this measure will be especially sensitive to the un-tuned free parameters which mediate between model states and predicted RTs.

### 3.3.3 Findings

Table 1 presents the Spearman correlation with 95% CI,[10] and RMSE of the ACT-R baseline and target models relative to the observed RTs from the SPP-Short dataset. In Table 2, we use Williams' T2 statistic to compare two correlations onto a shared dependent variable (Dunn and Clark, 1971). We additionally compare our ACT-R models with a less constrained alternative, *lreg*, a linear regression that predicts RTs using log-frequency, CoSim (from Word2Vec), and target word length. See Appendix D for details on this regression.

#### 3.3.3.1 Baseline models (B1–B3)

Among the ACT-R baseline models, B3 achieves the highest correlation ($\rho = 0.48$), significantly outperforming both EMMA (B1; $\rho = 0.30$) and the frequency-only model (B2; $\rho = 0.41$), $p < 0.01$. These results suggest that within this model schema, effects of target word length (on encoding time) and frequency (on retrieval time, via base-level activation) are valuable basic components of appropriate RT predictions.

Examining prediction errors using RMSE, we find that B2, the frequency model, achieves the lowest RMSE (≈70 ms), diagnosing the closest absolute predictions. In fact, this holds across all ACT-R baseline and target models. Models that exclude frequency exhibit substantially larger errors, while other models that include frequency among other predictors demonstrate slightly larger errors. This pattern is not surprising, as the free parameters which scale the relationship between activation and response time were fit in a model where frequency was the only determinant factor of RT (Brasoveanu and Dotlačil, 2020).

8 Available at https://www.pawelmandera.com/snaut-en/. BERT-based nearest neighbor selection would be more appropriate for the BERT-based model, but we judged this to be an acceptable shortcut.

9 Nevertheless, see Appendix C for very similar results with Pearson correlations.

10 We compute this interval by bootstrap sampling using 3,000 bootstraps for 6,631 data points.

TABLE 1 Measures of fit for baseline and target models, including Spearman correlation with bootstrapped 95% confidence intervals, and mean squared error.

| Model | Embedding | $S$ | $\rho$ | 95% CI | RMSE |
|---|---|---|---|---|---|
| lreg | – | – | 0.51 | (0.49–0.52) | 65.16 |
| B1 | – | – | 0.30 | (0.28–0.32) | 282.18 |
| B2 | – | – | 0.42 | (0.39–0.44) | 69.91 |
| B3 | – | – | 0.48 | (0.46–0.50) | 72.86 |
| M1a | W2V | 4.0 | 0.20 | (0.17–0.22) | 199.88 |
| M1b | BERT-L1 | 5.0 | 0.21 | (0.18–0.23) | 207.21 |
| M2a | W2V | 4.0 | 0.50 | (0.49–0.52) | 79.81 |
| M2b | BERT-L1 | 5.0 | 0.49 | (0.47–0.51) | 79.37 |
| M3a | W2V | 4.0 | 0.51 | (0.49–0.53) | 75.97 |
| M3b | BERT-L1 | 5.0 | 0.50 | (0.48–0.51) | 76.52 |

Embedding type and maximum association strength value ($S$) are listed where relevant.

#### 3.3.3.2 Target models (M1–M3)

Among the target models, M1, which incorporates only the CoSim term in the spreading activation (Equation 5), yields a correlation of $\rho = 0.20$ with the Word2Vec-based model (M1a) and $\rho = 0.21$ with the BERT-L1 model (M1b).[11] By contrast, M2, which integrates M1 with B3, achieves significantly higher correlations of $\rho = 0.50$ with Word2Vec (M2a) and $\rho = 0.49$ with BERT-L1 (M2b), demonstrating the added predictive value which comes from incorporating CoSim together with other known predictors of performance. As shown in Table 2, M2a significantly outperforms both B3 and M1a, $p < 0.01$.

M3, which extends M2 with the fan effect term (Equation 6), yields the highest correlations overall: $\rho = 0.51$ with Word2Vec (M3a) and $\rho = 0.50$ with BERT-L1 (M3b). This improvement suggests that the complete dataset was best accounted for by an asymmetric spreading activation mechanism. Notably, M3a significantly outperforms M2a, $p = 0.02$, making it our best-performing model in our study.

Turning again to RMSE, we again note that models excluding frequency (M1a, M1b) produce the highest errors. The RMSEs for M2 and M3 approach that of B2 and decrease as Spearman correlation improves, indicating that these models better capture the appropriate order, and depart only minimally from absolute accuracy. Among the target models, M3a achieves the lowest RMSE (≈76 ms).

#### 3.3.3.3 Linear regression

Finally, the linear regression represents a best-case scenario in which all parameters are directly optimized to fit the data. Unsurprisingly, this model achieves the lowest RMSE (≈65 ms), but it does not improve on the Spearman correlation of our best target model ($\rho = 0.51$). Despite the fact that our ACT-R models are operating with mostly static parameters, and limited, pre-conceived relationships between predictors and RT, we therefore conclude that our best target model captures a high degree of the variation which can be explained by these three predictors.

11 These results are similar to the CoSim correlations reported in Cassani et al. (2023); see Appendix A for more details.

TABLE 2 Selected Williams' T2 statistics comparing Spearman correlations in Table 1.

| Comparison | $T2$ | $p$ |
|---|---|---|
| B1 vs. B3 | −10.90 | <0.01 |
| B2 vs. B3 | −21.83 | <0.01 |
| M2a vs. B3 | 7.29 | <0.01 |
| M2a vs. M1a | 23.79 | <0.01 |
| M3a vs. M2a | 2.39 | 0.02 |

#### 3.3.3.4 Further investigations

Figure 5 visualizes 2,000 randomly sampled prime–target pairs, plotting observed RTs ($x$-axis) against predicted RTs ($y$-axis) from the Word2Vec-based M3a (left) and BERT-based M3b (right). The colormap encodes CoSim values for each pair. The majority of observed RTs and ACT-R predictions cluster between 550–900 ms, with the models tending to underestimate true RTs, and failing to capture extreme outliers. Interestingly, most of these outliers have near-zero CoSim values, suggesting that CoSim is not a decisive factor for these cases. This points to the existence of additional latent factors influencing RTs, which we will revisit in the Discussion (Section 3.3.4).

To further examine the behavior of our model, Figure 6 compares average observed RTs (left) with average RTs predicted by M3a (right), across four frequency bands (from low to high) and four levels of CoSim values. The figure illustrates that our model successfully replicates a notable human RT pattern: for low-frequency words, both humans and ACT-R exhibit substantially faster RTs as the CoSim between prime and target words increases, whereas for high-frequency words, CoSim has a comparatively smaller effect (Becker, 1979). This is a prime example of a complex aspect of the data which requires a specified model to explain. In the case of the ACT-R model, this scaling occurs because frequency and spreading activation contribute additively toward total activation levels, which have an inverse exponential relationship to retrieval latencies (Equation 3). The presence of one activation source (e.g. high baseline activation due to frequency) brings activation to a higher level, where unit changes in activation produce smaller latency changes, so that variation in another activation

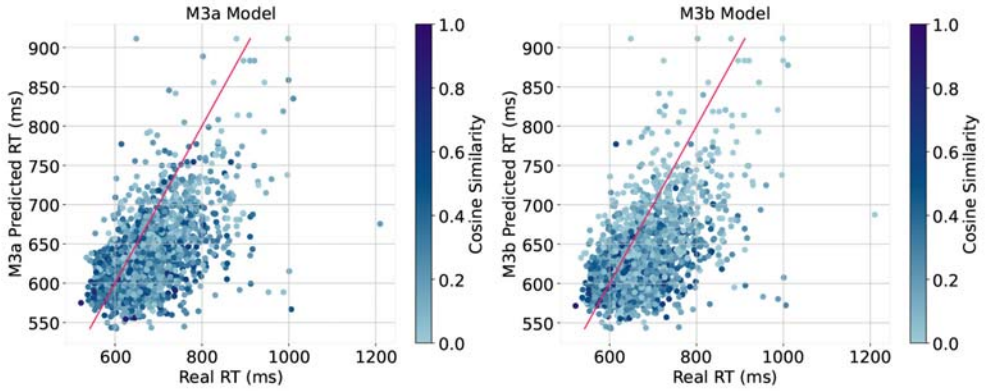


FIGURE 5
Observed and M3-Predicted RTs using Word2Vec [**(left)**; M3a] and BERT-L1 [**(right)**; M3b]. The colormap indicates the CoSim value for each pair. The pink line shows the unit slope ($y = x$).

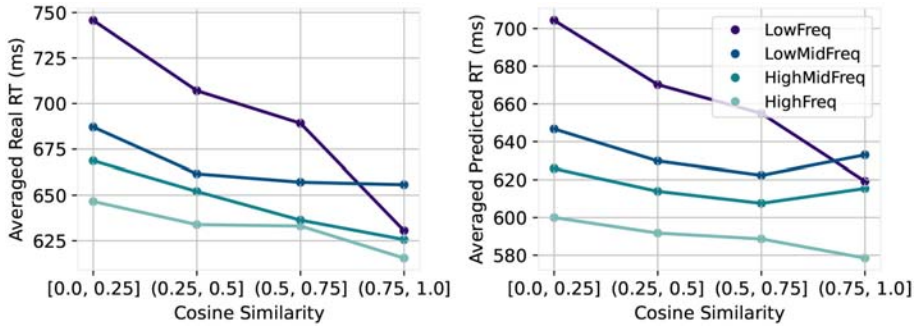


FIGURE 6
Interaction between frequency and CoSim on averaged real RTs **(left)** and averaged predicted RTs using M3a **(right)**.

source (e.g. spreading activation) will have a smaller effect. ACT-R also correctly predicts that this interaction should not hold between either of these sources and visual complexity (Figure 7, see Appendix D); in the model, visual complexity determines a separate latency component which combines additively with the latency of memory retrieval.

### 3.3.4 Remarks

To get a better sense of how CoSim-based spreading activation affects the predictions of our model, we zoom in here on the behavior of particular items. Figure 8 illustrates how the addition of Word2Vec CoSims in M3a changes some of the predictions of B3 across 1,000 random samples from the SPP-Short dataset. Arrows pointing toward the $y = x$ line indicate cases where CoSim improves the predictions relative to B3, whereas arrows pointing away from the line indicate cases where CoSim worsens them. While in many instances predictions remain unchanged, several notable shifts can be observed, indicated by black dotted arrows in the figure. For example, the point indicated on the left corresponds to the prime-target pair (*employee–employer*) with a CoSim of 0.70, a prime fan of 0.41, and a target frequency in band q30. Similarly, the point indicated on the right corresponds to (*nerd–geek*) with a CoSim of 0.54, a prime fan of 0.37, and a target frequency in band q7. Both pairs have high CoSim values, and we observe that CoSim affects the predicted RTs from Word2Vec more than expected. One possible explanation is that the fan term does not penalize spreading activation strongly enough in these cases, suggesting that more refined measures of fan could further improve predictions.

To gain deeper insights, we next examine the largest prediction shifts across the entire SPP-Short dataset. Table 3 reports the worst (top) and best (bottom) shifts when moving from B3 to M3a.

Notably, two prime-target pairs appear in both sets of shifts. The first set is *(zit–pimple)* and its reversed version *(pimple–zit)*. Both words have very low frequency (q0 band) and identical CoSim values as well as very close fans. Human RTs, however, differ substantially: 800.23 ms for the first case and 713.11 ms for the second. Including CoSim reduces the predicted RTs for both, as expected, but this adjustment improves the prediction for *zit–pimple* while worsening it for *pimple–zit*. One possible explanation lies in the slightly smaller fan value for *pimple* (by 0.02), which results in weaker penalization of its spreading activation.

A closer inspection reveals small yet meaningful differences in the actual frequencies of the two words (*pimple* = 0.67, *zit* = 0.25), even though both were assigned the q0 average frequency (1.37) in our model. Due to the negative exponential relationship between activation and RT (Equation 3), this subtle frequency difference may have had the potential to capture this RT discrepancy more

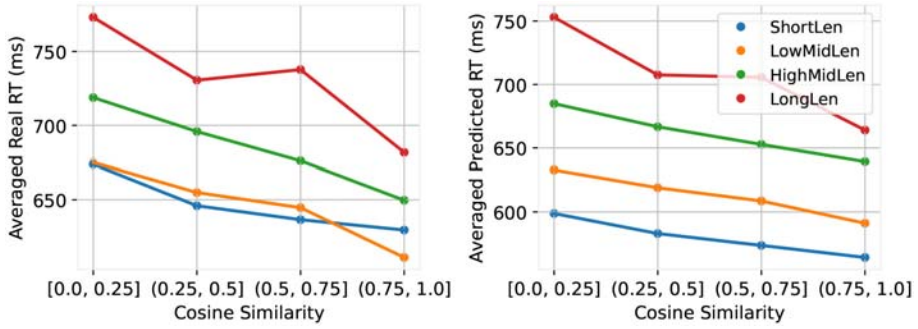


FIGURE 7
Interaction between Target word length and CoSim on averaged real RTs **(left)** and averaged predicted RTs using M3a **(right)**.

accurately, had we treated frequency at a finer grain. Additionally, orthographic factors may play a role: *zit* may receive especially slow responses due to the low typicality of the character *z* in English orthography; we have not treated any orthographic or phonological effects here, although their influence has been a fixture in previous work, see Andrews (1997) for a review.

A similar pattern is observed for *(inhale-exhale)* and its reversed version, where the higher frequency and orthographic typicality of *inhale* compared to *exhale* may explain the shorter observed RT for the pair *(exhale-inhale)*.

As potential directions for future research, it is clear that more precise approximations of frequency would be beneficial for model performance, particularly for low-frequency words where small estimation errors can disproportionately influence predicted RTs. Moreover, extending the model to include additional factors—such as orthographic and phonological similarity, sensorimotor effects to account for concrete vs. abstract words, and individual variability—could further improve predictions and help capture the extreme outliers (see Figure 5).

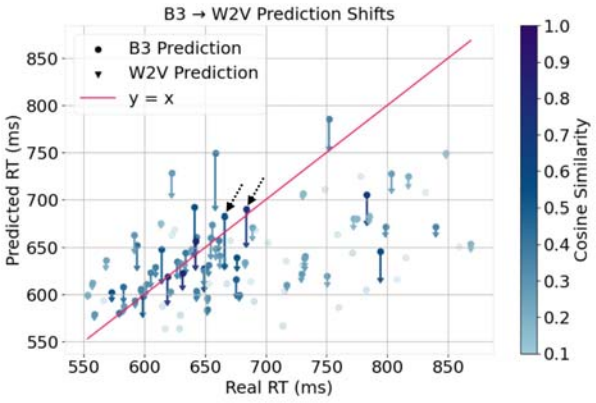


FIGURE 8
B3 to M3a prediction shifts for 1,000 random samples of the SPP-Short dataset. The colormap shows the CoSim value for each pairs. The dashed arrows point to the worst shifts within these 1,000 samples.

## 4 Discussion

In laying out one of the classic mechanistic theories of spreading activation, Anderson expresses one worry about its evaluation and application:

> "One of the problems in using this model is that one needs to specify all of the long-term memory network connected at any distance to the working memory elements to derive precise predictions about activation patterns" (Anderson, 1983, p. 266).

Indeed, we think this obstacle has stood in the way of much progress at the interface between models like Anderson's and the large scale of modern psycholinguistic data.

In this research, we worked to surmount this obstacle, by incorporating language model embeddings into ACT-R to help estimate the wide network of associations between items in a realistically-sized mental lexicon. This integration allows us to model associative priming with two advantages which have not traveled together in prior work:

1. Fixed and interpretable, thanks to ACT-R's modular structure,
2. Scalable and realistic, through replacing hand-coded discrete features with CoSim values derived from LMs.

In particular, having a scalable but interpretable model allowed us to generate large-scale itemwise predictions for two theoretically-relevant alternative models, comparing a simplified symmetric spreading activation to a more complex, asymmetric mechanism which includes penalties for associative density. Our results here suggest some unexpected support for the full, asymmetric variant, in contrast to previous evidence for symmetry from many factorial studies.

More than a model of this particular task, we present this work as a proof of concept and a first step toward enriching cognitive architectures with modern language technologies. This represents just one set of simple predictions for an elementary task, based on a conveniently-available memory model. It is our hope that future

TABLE 3 Worst (top) and best (bottom) prediction shifts from B3 to M3a.

| P | T | Obs. RT | B3 RT → M3 RT | P-T CoSim | T's Freq. | P's Fan |
|---|---|---|---|---|---|---|
| Pimple | Zit | 800.23 | 799.9 → 686.1 | 0.54 | q0 | 0.49 |
| Breathe | Exhale | 760.56 | 777.1 → 681.3 | 0.64 | q1 | 0.54 |
| Inhale | Exhale | 757.73 | 777.1 → 682.5 | 0.61 | q1 | 0.45 |
| Duck | Quack | 803.31 | 763.4 → 674.7 | 0.57 | q1 | 0.47 |
| Ash | Ashtray | 865.93 | 855.5 → 768.5 | 0.39 | q0 | 0.38 |
| Plastic | Tupperware | 868.13 | 897.2 → 814.0 | 0.39 | q0 | 0.46 |
| Introvert | Extrovert | 780.65 | 883.3 → 754.4 | 0.62 | q0 | 0.50 |
| Zit | Pimple | 713.11 | 841.6 → 728.8 | 0.54 | q0 | 0.51 |
| Freckle | Pimple | 728.23 | 841.6 → 734.8 | 0.51 | q0 | 0.51 |
| Container | Tupperware | 810.71 | 897.2 → 801.2 | 0.44 | q0 | 0.43 |
| Entertain | Amuse | 712.68 | 763.4 → 674.5 | 0.57 | q1 | 0.46 |
| Exhale | Inhale | 682.81 | 752.4 → 671.3 | 0.61 | q2 | 0.48 |

P and T refer to prime and target, respectively.

work can apply similar techniques to compare the performance of more complex models at a similar scale. For instance, van Maanen et al. (2012) present an adapted form of ACT-R's visual processing and memory architecture based on sequential sampling approaches, and an implementation of spreading activation which flows continuously, rather than originating only from the current focus of attention, closer to the original semantic networks proposal of Collins and Loftus (1975). Comparisons between our present approach and theirs could inform future theorizing about the nature of spreading activation. Likewise, a similar approach could be used to extend distributed memory models like Lerner et al. (2012), and compare their performance on the same data.

Future work can also extend our particular ACT-R model in many other directions:

- Other dimensions of similarity: while we focused on semantic similarity, other forms of similarity such as *orthographic distance* (cf. Mandera et al. 2017) could be incorporated as additional activation components.
- Cognitive individual differences: one promising extension is to parameterize spreading activation by participants' working memory capacity (WMC) (Daily et al., 2001). In this view, higher WMC would increase the relative influence of associative priming on RT, perhaps aligning model predictions with known individual variability (Yap et al., 2016).
- Trial-specific models: instead of exposing participants to a single prime, one could investigate continuous priming effects by modeling sequences of primes. This would test the dynamic behavior of the model in more naturalistic settings.
- Alternative models of the fan effect: our current M3 model estimates fan effects using cosine similarity among nearest neighbors. Future work could explore alternative operationalizations, perhaps including clustering methods for identifying meaningful neighborhoods.
- Domain-specific embeddings: following Škrjanec et al. (2023), embeddings trained on specialized corpora may capture differences in the representations and associations of words across individual differences in expertise. This could shed new light on how domain knowledge modulates cognitive processing.
- Beyond lexical decision: the Semantic Priming Project dataset also contains data from a primed naming task, sometimes argued to be a better measure for isolating automatic priming from strategic effects (see sources reviewed in Hutchison, 2003). We refrain from treating that data here largely because modeling naming latencies would require non-trivial assumptions about the timing of speech planning and articulation after successful retrieval, but future work would benefit from extending toward this salient additional task.
- Beyond priming effects: finally, building on this foundation, one could easily extend our approach to similarity-based interference tasks beyond priming, as envisioned by Smith and Vasishth (2020). Such tasks would offer a broader testbed for evaluating the robustness of integrating LLM embeddings into the ACT-R framework.

In conclusion, our work demonstrates the potential of combining advanced language technologies with cognitive architectures to bridge the gap between computational models and human cognition. By addressing the outlined directions, we aim to refine these models further, broadening their applicability and improving their cognitive plausibility.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://www.montana.edu/attmemlab/spp.html.

## Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional

requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

## Author contributions

MM: Writing – review & editing, Visualization, Validation, Methodology, Formal analysis, Writing – original draft. JD: Writing – original draft, Investigation, Supervision, Methodology, Conceptualization, Writing – review & editing. VD: Investigation, Funding acquisition, Resources, Writing – review & editing, Supervision.

## Funding

The author(s) declared that financial support was received for this work and/or its publication. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC Starting Grant "Individualized Interaction in Discourse", grant agreement No. 948878).

## Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. The authors used an AI tool only to refine the wording and phrasing of some parts of the manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note



## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/flang.2026.1721326/full#supplementary-material

## References

Anderson, J. R. (1983). A spreading activation theory of memory. *J. Verbal Learning Verbal Behav.* 22, 261–295. doi: 10.1016/S0022-5371(83)90201-3

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., Qin, Y., et al. (2004). An integrated theory of the mind. *Psychol. Rev.* 111, 1036–1060. doi: 10.1037/0033-295X.111.4.1036

Anderson, J. R., and Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Anderson, J. R., and Schooler, L. J. (1991). Reflections of the environment in memory. *Psychol. Sci.* 2, 396–408. doi: 10.1111/j.1467-9280.1991.tb00174.x

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: resolving neighborhood conflicts. *Psychon. Bull. Rev.* 4, 439–461. doi: 10.3758/BF03214334

Arambel, S. R., and Chiarello, C. (2006). Priming nouns and verbs: Differential influences of semantic and grammatical cues in the two cerebral hemispheres. *Brain Lang.* 97, 12–24. doi: 10.1016/j.bandl.2005.07.003

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., and Yap, M. J. (2004). Visual word recognition of single-syllable words. *J. Exp. Psychol. Gen.* 133, 283–316. doi: 10.1037/0096-3445.133.2.283

Balota, D. A., Yap, M. J., Cortese, M. J., and Watson, J. M. (2008). Beyond mean response latency: response time distributional analyses of semantic priming. *J. Mem. Lang.* 59, 495–523. doi: 10.1016/j.jml.2007.10.004

Becker, C. A. (1979). Semantic context and word frequency effects in visual word recognition. *J. Exp. Psychol. Hum. Percept. Perform.* 5, 252–259. doi: 10.1037//0096-1523.5.2.252

BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., and Reddy, S. (2024). LLM2vec: large language models are secretly powerful text encoders. *arXiv [preprint]*, arXiv:2404.05961.

Bommasani, R., Davis, K., and Cardie, C. (2020). "Interpreting pretrained contextualized representations via reductions to static embeddings," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, eds D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault (Association for Computational Linguistics), 4758–4781. doi: 10.18653/v1/2020.acl-main.431

Brasoveanu, A., and Dotlačil, J. (2020). *Computational Cognitive Modeling and Linguistic Theory*. Chm: Springer Nature. doi: 10.1007/978-3-030-31846-8

Cassani, G., Günther, F., Attanasio, G., Bianchi, F., and Marelli, M. (2023). Meaning modulations and stability in large language models: an analysis of BERT embeddings for psycholinguistic research. *Preprint*. doi: 10.31234/osf.io/b45ys

Chwilla, D. J., and Kolk, H. H. J. (2002). Three-step priming in lexical decision. *Mem. Cognit.* 30, 217–225. doi: 10.3758/BF03195282

Collins, A. M., and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychol. Rev.* 82, 407–428. doi: 10.1037/0033-295X.82.6.407

Daily, L. Z., Lovett, M. C., and Reder, L. M. (2001). Modeling individual differences in working memory performance: a source activation account. *Cogn. Sci.* 25, 315–353. doi: 10.1207/s15516709cog2503_1

Devlin, J. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv [preprint]*. arXiv:1810.04805.

Dotlačil, J. (2018). Building an ACT-R reader for eye-tracking corpus data. *Top. Cogn. Sci.* 10, 144–160. doi: 10.1111/tops.12315

Dunn, O. J., and Clark, V. (1971). Comparison of tests of the equality of dependent correlation coefficients. *J. Am. Stat. Assoc.* 66, 904–908. doi: 10.1080/01621459.1971.10482369

Gemma Team, Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., et al. (2024). Gemma: open models based on Gemini research and technology. *arXiv [preprint]*. arXiv:2403.08295.

Günther, F., Dudschig, C., and Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: evidence from priming studies. *Q. J. Exp. Psychol.* 69, 626–653. doi: 10.1080/17470218.2015.1038280

Günther, F., Rinaldi, L., and Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: a discussion of common misconceptions. *Perspect. Psychol. Sci.* 14, 1006–1033. doi: 10.1177/1745691619861372

Harris, Z. S. (1954). Distributional structure. *WORD* 10, 146–162. doi: 10.1080/00437956.1954.11659520

Hutchison, K. A. (2002). The effect of asymmetrical association on positive and negative semantic priming. *Mem. Cognit.* 30, 1263–1276. doi: 10.3758/BF032 13408

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychon. Bull. Rev.* 10, 785–813. doi: 10.3758/BF03196544

Hutchison, K. A., Balota, D. A., Cortese, M. J., and Watson, J. M. (2008). Predicting semantic priming at the item level. *Q. J. Exp. Psychol.* 61, 1036–1066. doi: 10.1080/17470210701438111

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., et al. (2013). The semantic priming project. *Behav. Res. Methods* 45, 1099–1114. doi: 10.3758/s13428-012-0304-z

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211–240. doi: 10.1037/0033-295X.104.2.211

Lerner, I., Bentin, S., and Shriki, O. (2012). Spreading activation in an attractor network with latching dynamics: automatic semantic priming revisited. *Cogn. Sci.* 36, 1339–1382. doi: 10.1111/cogs.12007

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog 0000_25

Lund, K., Burgess, C., AND Audet, C. (1996). Dissociating semantic and associative word relationships using high-dimensional semantic space. *Proc. CogSci.* 18, 603–608.

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. *J. Mem. Lang.* 92, 57–78. doi: 10.1016/j.jml.2016.04.001

McNamara, T. P., and Altarriba, J. (1988). Depth of spreading activation revisited: semantic mediated priming occurs in lexical decisions. *J. Mem. Lang.* 27, 545–559. doi: 10.1016/0749-596X(88)90025-3

Meyer, D. E., and Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *J. Exp. Psychol.* 90, 227–234. doi: 10.1037/h0031564

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Vol. 2* (Red Hook, NY: Curran Associates Inc.), 3111–3119.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: roles of inhibitionless spreading activation and limited-capacity attention. *J. Exp. Psychol.: Gen.* 106, 226–254. doi: 10.1037/0096-3445.106.3.226

Penedo, G., Kydlíček, H., Ben Allal, L., Lozhkov, A., Mitchell, M., Raffel, C., et al. (2024). "The FineWeb datasets: decanting the web for the finest text data at scale," in *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Red Hook, NY: Curran Associates Inc.).

Pennington, J., Socher, R., and Manning, C. D. (2014). "GLOVE: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, eds A. Moschitti, B. Pang, and W. Daelemans (Doha: Association for Computational Linguistics), 1532–1543. doi: 10.3115/v1/D14-1162

Plaut, D. C. (1995). Semantic and associative priming in a distributed attractor network. *Proc. CogSci.* 17, 37–42.

Reichle, E. D., Rayner, K., and Pollatsek, A. (1999). Eye movement control in reading: accounting for initial fixation locations and refixations within the ez reader model. *Vis. Res.* 39, 4403–4411. doi: 10.1016/S0042-6989(99)00152-2

Salvucci, D. D. (2001). An integrated model of eye movements and visual encoding. *Cogn. Syst. Res.* 1, 201–220. doi: 10.1016/S1389-0417(00)00015-2

Škrjanec, I., Broy, F. Y., and Demberg, V. (2023). Expert-adapted language models improve the fit to reading times. *Procedia Comput. Sci.* 225, 3488–3497. doi: 10.1016/j.procs.2023.10.344

Smith, G., and Vasishth, S. (2020). A principled approach to feature selection in models of sentence processing. *Cogn. Sci.* 44:e12918. doi: 10.1111/cogs.12918

Speer, R. (2022). *rspeer/wordfreq: v3.0, version v3.0.2*. Zenodo. doi: 10.5281/zenodo.7199437

Timkey, W., and van Schijndel, M. (2021). "'All bark and no bite: Rogue dimensions in transformer language models obscure representational quality," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, eds M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih (Punta Cana: Association for Computational Linguistics) 4527–4546. doi: 10.18653/v1/2021.emnlp-main.372

Traxler, M. J., and Tooley, K. M. (2012). "Lexical and syntactic priming in language comprehension," in *Psychology of Priming*, eds. N. Hsu, and Z. Schütt (Hauppauge, NY: Nova Science Publishers.), 79–100.

Turney, P. D., and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *J. Artif. Intell. Res.* 37, 141–188. doi: 10.1613/jair.2934

van Maanen, L., van Rijn, H., and Taatgen, N. (2012). RACE/A: an architectural account of the interactions between learning, task control, and retrieval dynamics. *Cogn. Sci.* 36, 62–101. doi: 10.1111/j.1551-6709.2011.01213.x

van Rijn, H., and Anderson, J. R. (2003). "Modeling lexical decision as ordinary retrieval," in *Proceedings of ICCM 5* (Bamberg), 207–212.

Vulić, I., Baker, S., Ponti, E. M., Petti, U., Leviant, I., Wing, K., et al. (2020a). Multi-simlex: a large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Comput. Linguist.* 46, 847–897. doi: 10.1162/coli_a_00391

Vulić, I., Ponti, E. M., Litschko, R., Glavaš, G., and Korhonen, A. (2020b). "Probing pretrained language models for lexical semantics," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7222–7240. doi: 10.18653/v1/2020.emnlp-main.586

Whaley, C. P. (1978). Word–nonword classification time. *J. Verbal Learning Verbal Behav.*, 17, 143–154. doi: 10.1016/S0022-5371(78)90110-X

Yap, M. J., Hutchison, K. A., and Tan, L. C. (2016). "Individual differences in semantic priming performance: insights from the semantic priming project," in *Big Data in Cognitive Science*, ed. M. N. Jones (Hove: Psychology Press), 203–226.