Large Language Model Tokens are Psychologically Salient

David A. Haslett
haslett@ust.hk
Division of Social Science,
Hong Kong University of Science and Technology

Antoni B. Chan
abchan@cityu.edu.hk
Department of Computer Science,
City University of Hong Kong

Janet H. Hsiao
jhhsiao@ust.hk
Division of Social Science and Department of Computer Science & Engineering, Hong Kong University of Science and Technology

Abstract

Large language models segment words into chunks called tokens, using compression algorithms that ignore semantics. We investigated whether tokenization corrupts representations of word meanings in 17 languages. We found that GPT-4o and Llama 3 inflate the similarity of words that share tokens. However, tokens turned out to be good predictors of orthographic priming, such that people recognize a target word faster after reading a prime that ends with the same token. This boost in priming far exceeds what other overlapping strings of letters explain, which suggests that tokenization selectively identifies functional subword units. The pattern extends to the production of word associates in English: Tokens capture phonologically motivated associations, while other strings of letters do not. So, tokenization does influence semantic representations, but because tokens correspond to psychologically salient orthographic and/or phonological constituents, they may endow large language models with human-like language networks and facilitate alignment with human word processing.

Keywords: large language models; tokenization; conceptual alignment; semantic priming; subword processing

Introduction

Every day, millions of people use ChatGPT and other technologies powered by large language models (LLMs). As people delegate decisions to LLMs, conceptual misalignment—when LLM representations differ from human representations—will become increasingly consequential. In this study, we investigate one potential source of misalignment with human meanings: the seemingly unnatural way that LLMs segment words into tokens.

Does tokenization affect semantic representations?

Distributional semantic models, such as Latent Semantic Analysis (Landauer & Dumais, 1997) and word2vec (Mikolov et al., 2013), represent word meanings with embeddings, which situate words in semantic space. The embeddings are derived from patterns in text, based on the assumption that words which occur in similar contexts have related meanings (Harris, 1954; Firth, 1957). The cosine similarity of embeddings correlates with semantic priming effects: People recognize a target word faster after reading a semantically related prime word (Meyer & Schvaneveldt, 1971), and priming effects vary as a function of the proximity of primes to targets in distributional semantic space (Gatti et al., 2023; Jones et al., 2006; Lund & Burgess, 1996; Mandera et al., 2017). Günther et al. (2016) conclude that these models offer a "cognitively plausible measure of meaning similarities."

LLMs are sometimes called "contextualized" distributional semantic models (e.g., Ethayarajh, 2019), because whereas "static" models such as word2vec learn a single, stable embedding for each word, LLMs tailor embeddings to contexts (Devlin et al., 2018; Peters et al., 2018; Radford et al., 2019). For example, the representation of *bark* will differ in contexts that refer to dogs versus trees. The similarity of contextualized LLM embeddings predicts semantic priming and other word processing effects, but not as well as static embeddings do, at least in older LLMs such as BERT (Cassani et al., 2023; Lenci et al., 2022).

This deviation from human semantic representations might stem, in part, from tokenization. LLMs use compression algorithms to segment text into tokens for efficient encoding (see Mielke et al., 2021), and LLM embeddings represent tokens, not words. In standard tokenization methods, such as byte-pair encoding (Gage, 1994; Sennrich et al., 2016), tokens are strings of bytes that occur often in training data. English is predominant on the internet and in the largest datasets, so common English words tend to be single tokens in state-of-the-art LLMs, while most other words are segmented into subword tokens. For example, *interdisciplinary* is represented by GPT-4o as a single token, whereas the rare word *intradisciplinary* is segmented into three tokens (*_intr*, *ad*, *isciplinary*), and the French translation *interdisciplinaire* is segmented into four tokens (*_inter*, *disc*, *ipl*, *inaire*). (The underscores indicate whitespace incorporated into a word's first token.)

People decompose words to morphemes, i.e., meaningful, productive constituents such as "pre" and "ing" (Marslen-Wilson et al., 1994; Niswander et al., 2000; Rastle et al., 2004; Taft & Forster, 1975). This invites comparisons to tokenization, but as *intradisciplinary* illustrates, tokens do not reliably correspond to morphemes (Bostrom & Durrett, 2020; Church, 2020; Hofmann et al., 2021). Several studies report that tokens which deviate from morphemes hinder performance on downstream tasks, such as classifying titles and pairing words with definitions (e.g., Batsuren et al., 2024; Haslett, 2025; Hofmann et al., 2022). However, other studies report that LLMs perform equally well regardless of whether tokens correspond to morphemes (e.g., Arnett, Rivière et al., 2024; Gutiérrez et al., 2023). Gutierrez-Vasques et al. (2023) found that byte-pair encoding identifies patterns which correspond to linguistic typology, and they conclude that

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

tokens "might not correspond to usual morphological analyses, but they are structural elements." Haslett and Cai (2025) further found that tokens capture non-morphological patterns in form and meaning, such as etymological relationships, and so convey semantic information that morphemes miss. **It is an open question whether the impact of tokenization is negative, negligible, or positive**.

Are tokens psychologically salient?

How could a compression algorithm like byte-pair encoding, which is blind to morphology and context, split words into informative constituents? Languages evolve for the sake of efficient communication (Gibson et al., 2019), so word forms exhibit helpful regularities. Dautriche et al. (2017) found that Dutch, English, French, and German have more phonological neighbours (i.e., words that differ in only one sound) than would be expected by chance, suggesting that languages reuse strings of sounds to make articulation less effortful. There is longstanding evidence for psychologically real subsyllabic units (e.g., Bowey, 1990; Claxton, 1974; MacKay, 1972; Treiman, 1989), and tokenization should capture commonly reused chunks, which presents an opportunity to align LLMs with human word processing.

Tokenization is lossless compression because a token can be converted back into the original string of bytes, but from the perspective of LLMs, tokenization is lossy: LLMs do not have direct access to fine-grained information about the letters within a token (though see Edman et al., 2024; Itzhak and Levy, 2021; Kaushal and Mahowald, 2022). **Subsyllabic units offer an intermediate stage where tokens could correspond to psychologically real constituents**. Beinborn and Pinter (2023) report that in Dutch, English, French, and Spanish, people recognize words that are segmented into fewer tokens per letter faster and more accurately than words segmented into more tokens per letter. This is reminiscent of Treiman and Chafetz (1987), who showed that people recognize words faster when they are split into chunks that are more likely to be subsyllabic units (e.g., TW + IST as opposed to TWI + ST). However, unlike Treiman and Chafetz, Beinborn and Pinter did not manipulate the presentation of words, and correlations with tokens per letter can be confounded by word frequency (i.e., high-frequency words tend to be segmented into fewer tokens) among other lexical variables. Thus, it remains unclear whether tokenization recovers psychologically real units.

The present study

First, we tested the hypothesis that LLMs inflate the semantic similarity of words that share tokens. For example, Llama 3 and GPT-4o might relate *intradisciplinary* to *intrusiveness* because they both begin with the token *_intr*. We gathered pairs of primes and targets in 17 languages from the Semantic Priming Across Many Languages project (SPAM-L; Buchanan et al., 2021), and we measured the similarity of their representations in two state-of-the-art LLMs, Llama 3 and GPT-4o (Meta, 2024; OpenAI, 2024). The similarity of LLM representations of prime–target pairs should correlate with semantic priming effects in human participants and with representations in static distributional semantic models (e.g., Cassani et al., 2023), but we expected shared tokens to weaken those correlations (i.e., becoming less human-like), such that LLMs tend to treat unrelated words as semantically similar if they share tokens.

Although LLMs may inflate the semantic similarity of words that share tokens, causing their representations to diverge from human interpretations, these shared tokens may capture orthographic priming effects if they match well with subword units used in human word processing. For example, *intradisciplinary* might prime *intrusiveness* because they begin with the same letters. Orthographic priming depends on functional subword units—not just any string of letters will do—so the impact of shared tokens on word recognition could provide evidence that tokens are psychologically salient. **In a second analysis, we investigated whether tokens explain orthographic priming effects better than other overlapping strings of letters do**.

The above priming effects measure comprehension, so **in a third analysis, we investigated whether tokens are psychologically salient in language production**. In the word association task, people are presented with a cue word (e.g., *nocturnal*) and supply the first words that come to mind (e.g., *owl*, *night*). Crucially, when semantically related associates are hard to think of, people more often supply similar-sounding associates (Haslett & Cai, 2024; cf. Kumar et al., 2022). For example, the rare word *diurnal* sometimes elicits *urinal* (Stolz & Tiffany, 1972). If tokens skew how LLMs perceive words (i.e., because tokens obscure fine-grained information about word form), they might interfere with an LLM's ability to mimic this human behaviour. However, if tokens are psychologically salient, they should account for humans' tendency to think of similar-sounding associates better than other overlapping strings of letters do.

Analyses were conducted in Python and R (R Core Team, 2024). For data and scripts, see OSF: osf.io/p9afk.

Methods

Materials

Human semantic priming. SPAM-L comprises 1,000 target words from each of 18 languages, paired with a related prime and an unrelated prime. For example, the French target *morphine* ("morphine") was paired with the related prime *aspirine* ("aspirin") and the unrelated prime *conclusion* ("conclusion"). Participants made a lexical decision about a prime word (e.g., deciding whether the string of letters *aspirine* forms a word), followed by a lexical decision about the target. The priming effect is the difference in z-scored RT between the related and unrelated condition for each target word. For example, French participants recognized *morphine* faster after reading *aspirine* (628 ms, on average) than after reading *conclusion* (690 ms).

To examine how tokenization affects LLM representations of word meanings, we segmented each prime and target into GPT-4o tokens using the tiktoken Python package, from OpenAI, and into Llama 3 tokens using the transformers

package, from Hugging Face (Wolf et al., 2020). We excluded English because out of 2,000 prime–target pairs, only one or two share a GPT-4o or Llama 3 token. Of 16,736 related pairs in the 17 other languages, 1,928 (11.5%) share at least one GPT-4o token, and 2,395 (14.3%) share a Llama 3 token. For example, GPT-4o and Llama 3 each segment *aspirine* and *morphine* into two tokens: *_aspir* + *ine* and *_morph* + *ine*. To validate any relationship between tokenization and orthographic priming in these languages, we identified shared tokens using two other tokenizers trained on over a hundred languages each, Multilingual BERT and Multilingual T5 (Devlin et al., 2018; Xue et al., 2020).

Distributional semantic representations. To measure the distributional semantic similarity of primes and targets, we used FastText embeddings from Grave, Bojanowski et al. (2018). (Buchanan et al., 2021, provided their own FastText embedding similarity measure, and the pattern of effects holds when instead using that measure, but there are missing values for over 1,600 items.) FastText is a neural network like word2vec, which generates embeddings based on the contexts where words occur (Bojanowski, Grave, et al., 2017). We z-scored the FastText embeddings within each language because different languages have different embedding spaces. To estimate priming effects, we subtracted the cosine similarity of unrelated prime–target pairs (e.g., .08 for *conclusion* and *morphine*) from related pairs (.54 for *aspirine* and *morphine*). These sorts of static embeddings pervade cognitive science (e.g., Günther et al., 2019), and if LLM representations are reliable in these 17 languages, they should correlate with FastText embeddings. FastText incorporates subwords into embeddings (e.g., the representation of *morphine* comprises *morph*, *orphi*, *rphin*, *phine*), and because some tokens will be FastText subwords, FastText may underestimate how much shared tokens inflate the similarity of LLM representations (i.e., it makes for a more severe test than a model such as word2vec would).

LLM representations. We measured the similarity of GPT-4o representations of prime–target pairs in two ways, using the OpenAI API. First, we conducted a **semantic relatedness task**, asking GPT-4o to rate how closely related pairs of word meanings are, on a scale of 1 to 7. To account for the probabilistic nature of LLM output, we averaged the likelihood of GPT-4o responding with the values 1 through 7 (i.e., summing over each scale value multiplied by the likelihood of that scale value; see OSF). We subtracted ratings for unrelated pairs from related pairs. Larger differences should predict larger priming effects.

Second, for a more implicit measure, we conducted an **odd-one-out task**, asking GPT-4o which of the target, related prime, and unrelated prime is not like the other two words. We extracted the log probability of GPT-4o deciding that the related prime was the odd word out. For the French *morphine* item, GPT-4o has a -11.37 log probability of deciding that *aspirine* is the odd one out, compared to a log probability approaching 0 for *conclusion*. (Recall that the log of 1 is 0.) As the similarity of the target to the related prime increases, the probability of deciding that the related prime is the odd one out should decrease. For ease of comparison with the other metrics, we took the absolute value of the log probabilities. These non-negative log probabilities should have a positive correlation with semantic priming effects.

The OpenAI API allows us to extract the likelihood that GPT-4o will produce a given token, but GPT-4o is a closed LLM, so we cannot access the embeddings that situate tokens in semantic space. We therefore used the transformers package and an open weights model, Llama 3.1 with 8 billion parameters, to measure the **similarity of contextualized embeddings**. We inserted each word into the following context (with *French* and *aspirine* as examples): *What does the French word aspirine mean?* Then, across contexts, we measured the cosine similarity of each token in the target to each token in the related prime and the unrelated prime, in each layer (e.g., *morph* and *ine* to *aspir* and *ine*, for four cosines in the related condition, in each of 32 layers). As discussed above, LLM embeddings change to reflect context, so although the initial input embeddings of the token *ine* are identical across contexts (i.e., in *morphine* and *aspirine*), the contextualized embeddings diverge. For the similarity of *ine* to persist across contexts in later layers would not be trivial; rather, it would suggest that seemingly arbitrary subword tokens (such as *ad* in *intradisciplinary*) influence how LLMs represent word meanings. We averaged the cosine similarities of all pairs of tokens for each prime–target pair in each of Llama 3's 32 layers and subtracted the mean cosine in the unrelated pair from the related pair for each item. Consistent with recent evidence that representations of word meanings emerge early in LLMs (Kaplan et al., 2024; Liu et al., 2024), we found the strongest correlation with semantic priming in the twelfth layer of Llama 3 (see OSF), so we used embeddings from that layer in our analyses.

Semantic neighbourhood density and cue-associate overlap in the word association task. In the word association task, people more often produce similar-sounding associates in response to cue words from sparse semantic neighbourhoods (i.e., words with few synonyms; Haslett & Cai, 2024; cf. Sidhu & Pexman, 2018). We therefore investigated the relationship between an English cue word's semantic neighbourhood density (from Shaoul and Westbury, 2010) and the cue word's average orthographic overlap with associates generated by humans, from the Small World of Words project (De Deyne et al., 2019), or generated by Llama 3, Claude 3, and Mistral (Anthropic et al., 2024; Jiang et al., 2023), from the LLM World of Words project (Abramski et al., 2024). We measured overlap as the length (in letters) of the longest token shared by a cue–associate pair or the length of the longest overlap in pairs that share no tokens, divided by the length of the shorter of the cue or associate (i.e., normalized for the maximum possible overlap). We then calculated the average overlap for each cue word. To avoid trivial overlaps and to emphasize phonologically motivated relationships, we investigated only monomorphemic cues and associates (e.g., *quick* but not *quickly*; see OSF for the same results when including all cues and associates).

Analysis plan

First, we measured the correlation of the FastText similarity of SPAM-L prime–target pairs with the similarity of their LLM representations. If shared tokens inflate the similarity of LLM representations, the correlations should be weaker when pairs share tokens. These embeddings are a standard, quantifiable representation of word meaning, but they provide an incomplete picture of how humans represent word meanings. For instance, embeddings overlook taxonomical relationships (e.g., Erk, 2016) and sensorimotor information (e.g., Andrews et al., 2009). So, second, we measured the correlations of LLM representations with semantic priming effects in humans. If tokens corrupt LLM representations, the correlations should be weaker when pairs share tokens.

The SPAM-L project was designed to detect priming between semantically related words, but orthographically related words prime, too. Shared word endings (e.g., *morphine* and *aspirine*) reliably facilitate recognition (e.g., Radeau et al., 1995), whereas shared onsets (e.g., *morphine* and *mobile*) have inconsistent effects, such that short overlaps prime and longer overlaps compete (e.g., Slowiaczek & Hamburger, 1992; for a summary, see Dufour, 2008). To examine whether shared tokens account for human priming effects, we regressed priming effects on orthographic overlap (i.e., the number of initial or final letters shared by prime–target pairs, in separate models), whether pairs share a final or initial token, and their interaction. To compare pairs that share tokens to pairs that share other strings, we only included pairs that overlap in at least one character. If tokens are psychologically salient, shared endings should predict larger priming effects when those endings are tokens (i.e., a main effect), and shared onsets should have a steeper negative slope when those onsets are tokens (i.e., an interaction effect).

Finally, to account for human behavior in word production, we examined the correlation between the semantic neighbourhood density of a cue in the word association task and the orthographic overlap between that cue word and its associates. If tokens reflect the sorts of overlaps that lead people to think of similar-sounding words, the negative correlation between semantic neighbourhood density and cue–associate overlap should be stronger when measuring overlap in tokens than in other strings of letters. However, any such correlation would be due, in part, to the fact that words in denser semantic neighbourhoods are more common, on average, and that common English words tend to be single tokens, which precludes those words from sharing tokens with their associates. This is an important way that compression algorithms align with human word processing—people decompose rare words into subword constituents more often than they decompose common words (e.g., Alegre & Gordon, 1999)—but for the purposes of this analysis, we treat word frequency as a confound. We therefore investigated only multi-token cue words, and in hierarchical regression models, we regressed the semantic neighbourhood density of a cue word on its frequency (from Brysbaert and New, 2009), its length in letters, its length in tokens, and its mean orthographic overlap with its associates, in Step 1, and then in Step 2, we added token overlap to the model. If tokens are psychologically salient, token overlap should have a significant negative relationship with semantic neighbourhood density after accounting for word frequency, length, and other orthographic overlaps, and it should significantly improve the variance in semantic neighbourhood density explained by the regression models.

LLM representations

Correlations with distributional semantics

The similarities of LLM representations of SPAM-L prime–target pairs have the expected correlations with the similarities of their distributional semantic representations: FastText similarity has a positive relationship with GPT-4o semantic relatedness rating, Llama 3 embedding similarity, and negative GPT-4o odd-one-out log probability. Crucially, shared tokens weaken these correlations (see Table 1). When pairs have dissimilar FastText representations and share no tokens (i.e., toward the left of the red lines in the top half of Figure 1), they tend to receive low relatedness ratings, have dissimilar embeddings, and have high probabilities of being the odd word out, but when those pairs share tokens (i.e., toward the left of the blue lines), they tend to receive high relatedness ratings, have similar embeddings, and are unlikely to be the odd word out.

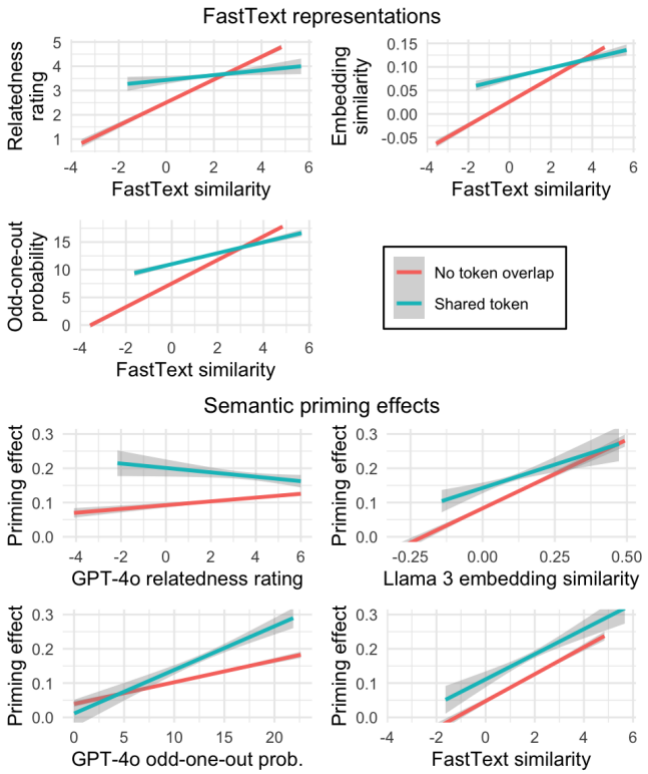

Figure 1: LLM representations against FastText representations, and semantic priming against LLM representations. For FastText (bottom right), shared tokens refer to Llama 3 tokenization. Odd-one-out probability refers to negative log probability. Shaded bands indicate 95% CIs.

Correlations with human semantic priming

The similarities of LLM representations of prime–target pairs have the expected correlations with the priming effects reported in SPAM-L: Semantic priming has a positive relationship with relatedness rating, contextualized embeddings, and negative odd-one-out log probability. Shared tokens weaken correlations with semantic priming for relatedness ratings and contextualized embeddings, which again suggests that shared tokens cause LLM representations to diverge from human representations. However, the non-significant difference for odd-one-out probability trends in the opposite direction. In fact, when prime–target pairs have the same LLM representation similarity (or FastText similarity), pairs that share tokens tend to exhibit larger priming effects than pairs that do not (i.e., in the bottom half of Figure 1, the blue lines are higher than the red lines). If LLMs further inflated the similarity of pairs that share tokens (shifting those observations toward the right in the bottom panels of Figure 1), LLM representation similarity would be a better predictor of priming, not worse. The fact that people recognize targets faster after reading primes that share a token, independent of semantic similarity, demonstrates the limitations of a behavioural metric like word recognition, which collapses semantic, orthographic, and other influences. Most relevantly, words that share tokens overlap in word form, so in the next section, we consider the impact of shared tokens in terms of orthographic priming.

Table 1: Correlations of the FastText similarity of SPAM-L prime–target pairs (top) or semantic priming effects (bottom) with three measures of the similarity of LLM representations, comparing correlations when pairs share tokens versus do not. Fisher's *z* reports the difference in z-scored correlation coefficients divided by the standard error. The bottom rows report the correlation of semantic priming with FastText similarity, using GPT-4o and Llama 3 tokenization.

	Shared (*r*)	Not (*r*)	Fisher's *z*	*p*
Correlation with FastText similarity				
Relatedness	.054	.220	7.1	1e-12
Embedding	.135	.259	6.0	2e-9
Odd one out	.264	.381	5.5	4e-8
Correlation with human semantic priming				
Relatedness	-.045	.046	3.9	1e-4
Embedding	.081	.152	3.3	8e-4
Odd one out	.185	.142	-1.8	.066
FastText (GPT-4o)	.141	.154	0.6	.569
FastText (Llama 3)	.140	.151	0.5	.589

Orthographic priming

Typically, shared word endings facilitate word recognition, whereas shared onsets facilitate when they are short but compete as they increase in length (see Dufour, 2008). To investigate whether tokenization captures these characteristics of human word processing, we regressed priming effects on the length of final overlap or initial overlap (i.e., strings of characters shared by targets and related primes) and whether those pairs share their final token or initial token, respectively. First, as reported in Table 2 and illustrated in Figure 2, the main effect of shared final tokens indicates that people recognize targets faster following primes that end with the same token. This boost in priming over other overlaps in word ending is further supported by ANOVA comparing variance explained in models with versus without the shared token factor ($F > 30$, $p <$ 1e-13, for each of GPT-4o, Llama 3, mBERT, and mT5). Second, the negative interactions of shared initial token with length of initial overlap indicates that shorter shared onsets facilitate whereas longer shared onsets compete but that, crucially, this is true only of onsets that are segmented into tokens. Tokenization supplies valuable information by detecting overlapping segments that are psychologically salient: It selectively identifies subword constituents that affect word processing in humans.

Table 2: Coefficients (β) from eight linear regression models, regressing priming on length of initial (top) or final (bottom) orthographic overlap and whether pairs share initial (top) or final (bottom) tokens, according to four tokenizers.

Tokenizer	GPT-4o	Llama 3	mBERT	mT5
Initial overlap				
Intercept	.11 ***	.12 ***	.12 ***	.11 ***
Token	.04 ***	.04 ***	.02	.05 ***
Length	.00	.01	.00	-.01
Len:Token	**-.04** ***	**-.06** ***	**-.03** ***	**-.02** *
Final overlap				
Intercept	.14 ***	.14 ***	.14 ***	.15 ***
Token	**.06** ***	**.06** ***	**.07** ***	**.06** ***
Length	.02 ***	.02 ***	.02 ***	.02 ***
Len : Token	-.01	-.01 *	-.01	-.01

Note: The token factor is sum-coded and mean-centred, such that positive main effects indicate higher intercepts when pairs share tokens, and positive interactions indicate more positive slopes. Length of final / initial orthographic overlap is log-transformed and z-scored. Only pairs that overlap in at least one character are included. To conserve space, only coefficients and significance are reported (indicates* p < *.05, *** indicates* p < *.001).*

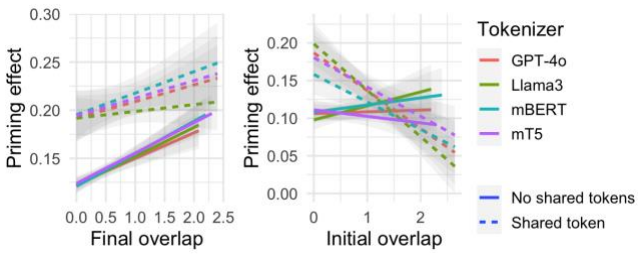

Figure 2: Semantic priming against initial orthographic overlap and final orthographic overlap. Length of overlap is log-transformed. Only pairs that overlap in at least one character are included. Shaded bands indicate 95% CIs.

Word associations

Consistent with tokens being psychologically salient, the semantic neighbourhood density of cue words has a negative correlation with cue–associate token overlap, whether associates are supplied by humans or LLMs. In short, cues with few synonyms often share tokens with their associates. To rule out the influence of word frequency (which correlates with both semantic neighbourhood density and the number of tokens per word) and to compare tokens to other strings, we regressed the semantic neighbourhood density of multi-token cues on cue frequency, cue length (in letters and tokens), and the mean length of overlapping strings in associates that share no tokens (in letters, normalized). Then, we added the mean length of overlapping tokens (in letters, normalized) to the models. In all cases, token overlap is a significant negative predictor, and it significantly improves model fit (Table 3, Figure 3). When swapping the roles of token overlap and other overlaps (i.e., including token overlap in Step 1 then adding other overlaps in Step 2), the effects are far from significant, implying that tokenization selectively identifies psychologically salient subword constituents and so accounts for human behaviour in word association task. That is, when people think of similar-sounding associates, they tend to think of words that share tokens, not other strings of letters.

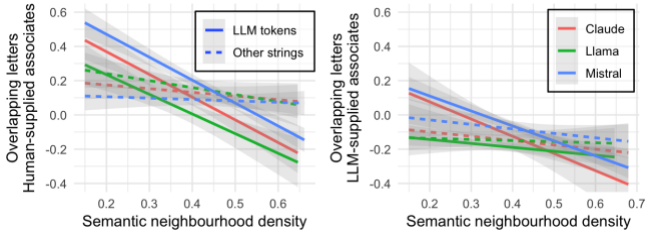

Figure 3: Cue–associate overlap in tokens versus other strings against semantic density. For clarity (i.e., to illustrate the slopes rather than the main effects), tokens and other strings are z-scored separately. Shaded bands indicate 95% CIs.

Table 3: Coefficients from 14 hierarchical regression models, regressing the semantic density of multi-token cue words on cue–associate overlap in tokens (left) or in other strings (right), after regressing semantic density on cue frequency, cue length, and overlap either in other strings (left) or tokens (right). *F* scores compare the variance explained by models with versus without overlap in tokens (left) or other strings (right) as predictors. Predictors are z-scored. For comparison, cue frequency has a coefficient of around .05 in these models.

	β (token)	*F*	β (other)	*F*
Human-supplied associates, segmented by three tokenizers				
Llama 3	-.018 ***	32.1 ***	-.003	0.7
Claude 3	-.018 ***	48.2 ***	.000	0.0
Mistral	-.018 ***	72.3 ***	.000	0.0
LLM associates, segmented by the corresponding tokenizer				
Llama 3	-.007 *	4.4 *	.000	0.0
Claude 3	-.011 ***	16.5 ***	.001	0.1
Mistral	-.016 ***	53.5 ***	-.001	0.2

*Note: *** indicates* p < *.0001, * indicates* p < *.05.*

Discussion and conclusion

Across 17 languages, we found that shared tokens inflate the similarity of word meanings in GPT-4o and Llama 3: Prime–target pairs which have dissimilar FastText representations but which share tokens received high semantic relatedness ratings, had similar contextualized embeddings, and were unlikely to be the odd word out. The results were mixed when investigating the impact of shared tokens on priming effects in lexical decisions (i.e., shared tokens weakened correlations for only two of our three metrics of LLM representation similarity), but lexical decisions are sensitive to orthographic priming, which prevents us from isolating semantic priming effects. When target words share tokens with semantically *dissimilar* prime words, people recognize the targets quickly, so further inflating the similarity of pairs that share tokens would improve LLM representations as predictors of priming overall, but at the cost of conceptual alignment.

Orthographic priming confounds semantic priming, but the fact that tokens explain orthographic priming effects far better than other strings do provides strong evidence that tokens correspond to psychologically real orthographic constituents (Beinborn & Pinter, 2023). Such constituents matter because although we do not want LLMs to clumsily conflate form with meaning (e.g., *dog* is unrelated to *dig*), we do want human-aligned LLMs to infer meaning from form judiciously, when the need arises. Non-arbitrary relationships between form and meaning influence word processing and the evolution of the lexicon across many languages (e.g., Blasi et al., 2016; Ćwiek et al., 2022), and LLMs recognize sound symbolic cues (Cai et al., 2023; Loakman et al., 2024 Marklová et al., 2025; Trott, 2024), but the trick is knowing when to use those sorts of cues. For instance, people compensate for a dearth of semantically related words by thinking of similar-sounding words (e.g., Kumar et al., 2022), and we found that LLMs exhibit this human-like behaviour in the word association task. More generally, people decompose rare words to morphemes (e.g., Alegre & Gordon, 1999), and byte-pair encoding was adopted by computational linguistics to accommodate rare words (Sennrich et al., 2016), the upshot being that both LLMs and humans process familiar words holistically while segmenting unfamiliar words into chunks. Rare words make up most of any language (e.g., Baayen, 2001), and thanks to the evolved utility of word form (e.g., Dautriche et al., 2017), compression algorithms tend to segment rare words into similar chunks as people do. These psychologically salient tokens are well positioned to guide LLMs towards human-like interpretations.

To sum up, we found that LLMs overly rely on tokens to represent meanings, inflating the similarity of representations of semantically unrelated words. Clearly, there is room to improve conceptual alignment. However, we also found that tokens predict orthographic priming effects and that LLMs, like people, selectively associate words that share tokens. Tokens correspond to an intermediate level of functional subword units, which leads us to the cautiously optimistic conclusion that current tokenization methods lay the groundwork for human-like language networks.

Acknowledgements

Many thanks to the Research Grant Council of Hong Kong for supporting this work (AoE/E-601/24-N).

References

Abramski, K., Improta, R., Rossetti, G., & Stella, M. (2024). The LLM World of Words: English free association norms generated by large language models. *arXiv preprint arXiv:2412.01330*.

Alegre, M., & Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of memory and language*, *40*(1), 41-61.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological review*, *116*(3), 463.

Anthropic (2024). The Claude 3 Model Family: Opus, Sonnet, Haiku. *www.anthropic.com/news/claude-3-family*

Arnett, C., Rivière, P. D., Chang, T. A., & Trott, S. (2024). Different Tokenization Schemes Lead to Comparable Performance in Spanish Number Agreement. *arXiv preprint arXiv:2403.13754*.

Baayen, H. (2001). *Word frequency distributions*. Kluwer.

Batsuren, K., Vylomova, E., Dankers, V., Delgerbaatar, T., Uzan, O., Pinter, Y., & Bella, G. (2024). Evaluating Subword Tokenization: Alien Subword Composition and OOV Generalization Challenge. *arXiv preprint arXiv:2404.13292*.

Beinborn, L., & Pinter, Y. (2023, December). Analyzing Cognitive Plausibility of Subword Tokenization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 4478-4486).

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, *113*(39), 10818-10823.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, *5*, 135-146.

Bostrom, K., & Durrett, G. (2020, November). Byte Pair Encoding is Suboptimal for Language Model Pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4617-4624).

Bowey, J. A. (1990). Orthographic onsets and rimes as functional units of reading. *Memory & Cognition*, *18*(4), 419-427.

Brysbaert, M., & 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.

Cai, Z. G., Haslett, D. A., Duan, X., Wang, S., & Pickering, M. J. (2023). Does ChatGPT resemble humans in language use? *osf.io/preprints/psyarxiv/s49qv*.

Cassani, G., Günther, F., Attanasio, G., Bianchi, F., & Marelli, M. (2023). Meaning Modulations and Stability in Large Language Models: An Analysis of BERT Embeddings for Psycholinguistic Research.

Church, K. W. (2020). Emerging trends: Subwords, seriously? *Natural Language Engineering*, *26*(3), 375-382.

Claxton, G. L. (1974). Initial consonant groups function as units in word production. *Language and Speech*, *17*(3), 271-277.

Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., ... & Winter, B. (2022). The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B*, *377*(1841), 20200390.

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. T. (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition*, *163*, 128-145.

De Boer, B. (2000). Self-organization in vowel systems. *Journal of phonetics*, *28*(4), 441-465.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805.*

Dufour, S. (2008). Phonological priming in auditory word recognition: When both controlled and automatic processes are responsible for the effects. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *62*(1), 33.

Edman, L., Schmid, H., & Fraser, A. (2024, November). CUTE: Measuring LLMs' Understanding of Their Tokens. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 3017-3026).

Erk, K. (2016). What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics*, *9*, 17-1.

Ethayarajh, K. (2019, November). How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 55-65).

Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences*, *103*(32), 12203-12208.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In *Studies in Linguistic Analysis*, 1–32. Oxford: Blackwell.

Gage, P. (1994). A new algorithm for data compression. *The C Users Journal*, *12*(2), 23-38.

Gatti, D., Marelli, M., & Rinaldi, L. (2023). Out-of-vocabulary but not meaningless: Evidence for semantic-

priming effects in pseudoword processing. *Journal of Experimental Psychology: General*, *152*(3), 851.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in cognitive sciences*, *23*(5), 389-407.

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *Quarterly Journal of Experimental Psychology*, *69*(4), 626-653.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006-1033.

Gutiérrez, B. J., Sun, H., & Su, Y. (2023, July). Biomedical Language Models are Robust to Sub-optimal Tokenization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks* (pp. 350-362).

Gutierrez-Vasques, X., Bentz, C., & Samardžić, T. (2023). Languages through the looking glass of BPE compression. *Computational Linguistics, 49*(4), 943-1001.

Harris, Z. (1954). Distributional structure. *Word, 10*(2/3), 146–162.

Haslett, D. A., & Cai, Z. G. (2024). Wayward associations: When and why people think of similar-sounding words. *Journal of Memory and Language, 138*, 104537.

Haslett, D. A. & Cai, Z. G. (2025). How much semantic information is available in large language model tokens? *Transactions of the Association of Computational Linguistics, 13*, 408-423.

Haslett, D. A. (2025). Tokenization changes meaning in large language models: Evidence from Chinese. *Computational Linguistics*, 1-30.

Hofmann, V., Pierrehumbert, J., & Schütze, H. (2021, August). Superbizarre Is Not Superb: Derivational Morphology Improves BERT's Interpretation of Complex Words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 3594-3608).

Hofmann, V., Schütze, H., & Pierrehumbert, J. (2022, May). An Embarrassingly Simple Method to Mitigate Undesirable Properties of Pretrained Language Model Tokenizers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 385-393).

Itzhak, I., & Levy, O. (2021). Models in a spelling bee: Language models implicitly learn the character composition of tokens. *arXiv preprint arXiv:2108.11193*.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jones, M. N., Kintsch, W., & Mewhort, D. J. (2006). High-dimensional semantic space accounts of priming. *Journal of memory and language*, *55*(4), 534-552.

Kaplan, G., Oren, M., Reif, Y., & Schwartz, R. (2024). From Tokens to Words: On the Inner Lexicon of LLMs. *arXiv preprint arXiv:2410.05864*.

Kaushal, A., & Mahowald, K. (2022, July). What do tokens know about their characters and how do they know it? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2487-2507).

Kumar, A. A., Lundin, N. B., & Jones, M. N. (2022). Mouse-mole-vole: The inconspicuous benefit of phonology during retrieval from semantic memory. In *Proceedings of the annual meeting of the Cognitive Science Society*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Lenci, A., Sahlgren, M., Jeuniaux, P., Cuba Gyllensten, A., & Miliani, M. (2022). A comparative evaluation and analysis of three generations of Distributional Semantic Models. *Language resources and evaluation*, *56*(4), 1269-1313.

Liu, Z., Kong, C., Liu, Y., & Sun, M. (2024). Fantastic Semantics and Where to Find Them: Investigating Which Layers of Generative LLMs Reflect Lexical Semantics. *arXiv preprint arXiv:2403.01509*.

Loakman, T., Li, Y., & Lin, C. (2024, November). With Ears to See and Eyes to Hear: Sound Symbolism Experiments with Multimodal Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 2849-2867).

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, *28*(2), 203-208.

MacKay, D. G. (1972). The structure of words and syllables: Evidence from errors in speech. *Cognitive Psychology*, *3*(2), 210-227.

Mandera, P., Keuleers, E., & 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.

Marklová, A., Milička, J., Ryvkin, L., Bennet, Ľ. L., & Kormaníková, L. (2025). Iconicity in Large Language Models. *arXiv preprint arXiv:2501.05643*.

Marslen-Wilson, W., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological review*, *101*(1), 3.

Meta (2024). Introducing Meta Llama 3: The most capable openly available LLM to date. *ai.meta.com/blog/meta-llama-3.*

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of experimental psychology*, *90*(2), 227.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Niswander, E., Pollatsek, A., & Rayner, K. (2000). The processing of derived and inflected suffixed words during reading. *Language and Cognitive Processes*, *15*(4-5), 389-420.

OpenAI (2024). GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Peters, M., Neumann, M., Iyyer, M., Gardner, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237.

Radeau, M., Morais, J., & Segui, J. (1995). Phonological priming between monosyllabic spoken words. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1297.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic bulletin & review*, *11*, 1090-1098.

Sennrich, R., Haddow, B., & Birch, A. (2016, August). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1715-1725).

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, *42*(2), 393-413.

Sidhu, D. M., & Pexman, P. M. (2018). Lonely sensational icons: semantic neighbourhood density, sensory experience and iconicity. *Language, Cognition and Neuroscience*, *33*(1), 25-31.

Slowiaczek, L. M., & Hamburger, M. (1992). Prelexical facilitation and lexical interference in auditory word recognition. *Journal of experimental psychology: Learning, memory, and cognition*, *18*(6), 1239.

Stolz, W. S., & Tiffany, J. (1972). The production of "child-like" word associations by adults to unfamiliar adjectives. *Journal of Verbal Learning and Verbal Behavior*, *11*(1), 38-46.

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of verbal learning and verbal behavior*, *14*(6), 638-647.

Treiman, R. (1989). The internal structure of the syllable. In *Linguistic structure in language processing* (pp. 27-52). Dordrecht: Springer Netherlands.

Trieman, R., & Chafetz, J. (1987). Are there onset-and rime-like units in written words. *Attention and performance XII: The psychology of reading*, 281-298.

Trott, S. (2024). Can large language models help augment English psycholinguistic datasets? *Behavior Research Methods*, 1-19.

Wolf, T. (2020). Transformers: State-of-the-Art Natural Language Processing. *arXiv preprint arXiv:1910.03771*.

Xue, L. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.