Is a cute *puyfred* cute?
Context-dependent form-meaning systematicity in LLMs

Jaïr Waal
Independent Researcher
jairwaal@gmail.com

Giovanni Cassani
Department of Cognitive Science
and Artificial Intelligence
Tilburg University, The Netherlands
g.cassani@tilburguniversity.edu

Abstract

We investigate static and contextualized embeddings for English pseudowords across a variety of Large Language Models (LLMs), to study (i) how these models represent semantic attributes of strings they encounter for the very first time and how (ii) these representations interact with sentence context. We zoom in on a key semantic attribute, valence, which plays an important role in theories of language processing, acquisition, and evolution. Across three experiments, we show that pseudoword valence is encoded in meaningful ways both in isolation and in context, and that, in some LLMs, pseudowords affect the representation of whole sentences similarly to words. This highlights how, at least for most LLMs we surveyed, pseudowords and words are not qualitatively different constructs. Our study confirms that LLMs capture systematic mappings between form and valence, and shows how different LLMs handle the contextualisation of pseudowords differently. Our findings provide a first computational exploration of how sub-lexical distributional patterns influence the valence of novel strings in context, offering useful insights for theories on the form-meaning interface and how it affects language learning and processing.

1 Introduction

Recently, a few studies have focused on pseudowords – phonotactically legitimate strings in a language which however lack conventional meaning –, uncovering a rich web of associations between sub-lexical patterns and semantic dimensions (Westbury et al., 2017; Gatti et al., 2024b; Joosse et al., 2024; Gatti et al., 2024a; Cassani et al., 2020; Sabbatino et al., 2022). This perspective blurs the distinction between pseudowords and words (Hendrix and Sun, 2020; Gatti et al., 2022; de Varda et al., 2024; Ryskina et al., 2020): from a learning perspective, every word a speaker knows used to be a pseudoword. However, the first time speakers experienced a novel word, it likely appeared *in context* (Savic et al., 2022; Chaffin et al., 2001). Nonetheless, no study has yet considered how sentence context influences the semantic connotations of entirely novel words.

We aim to start exploring the interplay between systematic form-meaning mappings and linguistic context, relying on computational analyses of current LLMs. How is lexical meaning constructed for novel words upon first encounter? What role does the linguistic form of words play in the process, if any? How does it interact with context, which is a key source in learning novel words (Savic et al., 2022; Lazaridou et al., 2014; Chaffin et al., 2001)?

In addressing these questions, we rely on evidence that has challenged the notion of arbitrary form-meaning mappings (Hockett, 1960). Many studies have observed sound symbolic associations (Köhler, 1929; Sapir, 1929) in typologically diverse languages (Winter et al., 2022; Ćwiek et al., 2022; Blasi et al., 2016). These observations support the notion that form-meaning mappings are best characterized as at least partly systematic (Pimentel et al.; Blasi et al., 2016; Dingemanse et al., 2015; Sidhu and Pexman, 2018). The extent to which systematicity documented in isolated words and pseudowords interacts with the semantics of co-occurring words is an open question we aim to address in this work.

We present three experiments which investigate pseudowords by focusing on a specific semantic attribute, i.e., valence, which characterizes words by how positive or negative they are. Theoretically, this semantic dimension has been identified as a critical axis along which both people and language models structure semantic representations (Osgood et al., 1957; Westbury et al., 2024). Practically, large datasets of human ratings for both words and pseudowords are available for this specific semantic dimension, making it a viable starting point to investigate how semantic connotations of entirely

Findings of the Association for Computational Linguistics: ACL 2025, pages 18747–18769
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

novel words interact with sentence context. Our first experiment focuses on static representations learned by different LLMs and how well they capture human ratings of pseudoword valence. The second experiment studies how such static representations are affected by co-occurring words. Finally, the last experiment investigates how pseudowords influence sentence representations. Our study provides the following key contributions:

LLMs capture pseudowords' valence: we see a sizable correlation between human valence ratings and static embeddings produced by all LLMs for pseudowords, confirming that the tokenization in part-words allows LLMs to encode form-meaning mappings.

Systematicity interacts with context: LLMs behave differently, but two cognitively sensible patterns emerge. Some models are only influenced by co-occurring words, discounting form-related information; others are influenced by both, suggesting that upon first encounter, pseudowords might retain valence associations conveyed by the linguistic form.

Pseudowords affect sentence representations: across models, the valence of sentences is systematically influenced by pseudowords. Moreover, the influence of pseudowords is similar to that of words, strengthening the hypothesis that words and pseudowords are not qualitatively distinct.

2 Experiment 1: Static representations of novel words

We replicate the design from Gatti et al. (2024a), investigating to what extent the static representations that five LLMs learn for pseudowords encode valence in line with human intuitions.

2.1 Materials & Methods

We use the 1,500 English pseudowords from Gatti et al. (2024a) rated for valence using a best-worst-scaling paradigm, and the valence norms collected by Warriner et al. (2013) for more than 13,000 English words. Ratings for both words and pseudowords are on a continuous scale. Following Gatti et al. (2024a), we train a linear regressor to predict the valence of words, encoded in a variety of ways[1]. The trained regressor is then applied to pseudowords to get a predicted valence rating, which is correlated with the human ratings. This design ensures that valence is construed in the same way for words and pseudowords, strengthening the analysis' validity.

Following Gatti et al. (2024a), we re-implement the letter uni-gram model, yielding 26-dimensional vectors where each position stores the frequency of an English lowercase letter in a target string. Moreover, we replicated the FastText (Bojanowski et al., 2017) model, encoding each string as the 300-dimensional vector from a custom FastText model trained on the Corpus of Contemporary American English (CoCA Davies, 2010) using 2- to 5-grams. Unlike the original study, though, we use Ridge Regression to better account for the large dimensionality of input vectors (the best hyper parameter values are available in Table 2 in Appendix A).

We tested five LLMs, accessed using HuggingFace (Wolf et al., 2020), encompassing a variety of architectures. BERT-EN(glish), in its base configuration, (Devlin et al., 2019) is an encoder model which has been extensively used for a variety of tasks, including the investigation of pseudowords' semantics (de Varda et al., 2024). Multilingual BERT (M-BERT, Devlin et al., 2019) offers a multilingual variant: if systematic mappings between word form and meaning are cross-linguistic (Blasi et al., 2016), a model which learns from multiple languages is expected to better leverage such mappings. RoBERTa (Zhuang et al., 2021) is a variant of the BERT model, which has also been used to investigate pseudowords' semantics (de Varda et al., 2024). Moreover, we considered two autoregressive models, GPT-2 Medium (Radford et al., 2019) and LLaMa 3.2 (Grattafiori et al., 2024). These models, and the comparison with the previous ones, allow us to investigate how pseudowords' representations change between autoregressive and Masked Language Models (MLMs).

For each word and pseudoword, we extracted the corresponding static embedding[2], before positional

[1] All materials used in the analyses are supplied at this link: `https://osf.io/aeygc/`.

[2] For GPT2 and RoBERTa, we prepended a white space to each string when extracting static embeddings, since this affects tokenization: for example, 'wonderful' is tokenized as ['w', 'onder', 'ful'] whereas '_wonderful' (with _ indicating a white space for better clarity) is tokenized as ['Ġwonderful']; the pseudoword 'brogmub' would be tokenized as ['b', 'rog', 'm', 'ub'] while '_brogmub' ['Ġbro', 'gm', 'ub'], with differences not just in the first token, but also in the resulting tokens within the pseudoword. This is important to ensure consistency throughout our experiments, since in Experiments 2 and 3 pseudowords will be embedded in sentences and preceded by a white space.

Model	All pseudowords	Lowest 25%	Highest 25%	Challenging set
Letter unigrams	**0.411; p <.001** **[0.368, 0.452]**	0.188; p <.001 [0.089, 0.284]	0.100; p = 0.053 [-0.001, 0.199]	**0.549; p <.001** **[0.466, 0.623]**
FastText	0.287; p <.001 [0.239, 0.332]	0.168; p <.01 [0.068, 0.265]	0.138; p <.01 [0.038, 0.236]	0.299; p <.001 [0.194, 0.397]
BERT-EN	0.344; p <.001 [0.298, 0.388]	0.255; p <.001 [0.158, 0.347]	0.132; p <.05 [0.031, 0.230]	0.331; p <.001 [0.228, 0.427]
M-BERT	0.364; p <.001 [0.319, 0.407]	0.249; p <.001 [0.152, 0.342]	**0.187; p <.001** **[0.088, 0.283]**	0.381; p <.001 [0.281, 0.472]
RoBERTa	0.377; p <.001 [0.332, 0.419]	*0.270; p <.001* *[0.174, 0.361]*	0.102; p = 0.051 [0.000, 0.201]	0.365; p <.001 [0.264, 0.460]
GPT2-Medium	0.333; p <.001 [0.287, 0.377]	0.187; p <.001 [0.087, 0.283]	*0.169; p <.01* *[0.069, 0.265]*	0.303; p <.001 [0.198, 0.401]
LLaMa3.2	*0.409; p <.001* *[0.366, 0.450]*	**0.303; p <.001** **[0.208, 0.392]**	0.122; p <0.05 [0.021, 0.221]	*0.385; p <.001* *[0.285, 0.476]*

Table 1: Spearman rank correlation coefficients, between human ratings and model predictions based on static representations for various models and subsets of pseudowords. The best model for each subset of pseudowords is in bold, the second best in italic. 95% Confidence Intervals are provided in brackets. p-values are reported next to each correlation coefficient.

encoding is added. This representation is entirely context independent. When a string consists of multiple tokens - hence for all pseudowords - we averaged the embeddings of the relevant tokens.

We evaluate the pseudowords' representations each model produces by computing the Spearman correlation (which avoids the assumption of a linear relation between observed and predicted values) between human and model-predicted valence ratings for the entire set of pseudowords. Moreover, we zoom in on three particularly relevant subsets of pseudowords. First, we consider the 25% most negative pseudowords according to human ratings: previous studies have suggested that systematic form-meaning mappings may be particularly useful in quickly signaling negative valence as an adaptive mechanism (Adelman et al., 2018; Louwerse and Qu, 2017). Second, we zoom in on the pseudowords in the 25% most positively rated pseudowords. This probes whether model can capture differences between extremely negative and somewhat negative(positive) pseudowords beyond capturing differences between negative, neutral and positive pseudowords, offering a more fine-grained evaluation. Finally, the pseudowords in Gatti et al. (2024a) were created using Wuggy (Keuleers and Brysbaert, 2010), which creates pseudowords starting from actual words and permuting characters following phonotactic rules. This poses a problem in our study. Consider the pseudoword *toutured*, clearly derived from *tortured*. When asked to determine its valence, a participant might rate the closest match instead - a process which appears to have been at play considering that Gatti et al. (2024a) found that a model predicting a pseudoword's valence based on the valence of the word at the lowest edit distance yields a sizable correlation. It is however questionable whether the rating for *toutured* actually constitutes the rating of a pseudoword, given the presence of a such a distinct and precise nearest neighbor based on word form. To zoom in on ratings provided for less transparent pseudowords, we created a subset consisting of pseudowords with no neighbors at an edit distance 1 and at least 3 neighbors at edit distance 2 in the SUBTLEX-US dataset (74,286 words) (Brysbaert et al., 2012). This was done to minimize the chance that participants resorted to the same closely matching word when rating a pseudoword's valence. This set, which we term the *challenging set*, consisted of 309 pseudowords.

2.2 Results

In Table 1, we see that letter uni-grams outperform other approaches on the full set, with a larger margin on the *challenging set*. However, when focusing on the tails of the distribution, the neural models emerge as very competitive, even though performance declines substantially, especially for the pseudowords rated as most positive. It is hard to identify consistent trends across neural models, since all embed pseudowords in a way which is largely consistent with human ratings, with no systematic differences between MLMs and autoregressive models. LLaMa 3.2 shows strong performance across all evaluation sets, topping the chart for negatively perceived pseudowords and showing strong performance on all pseudowords and on the challenging set. In Appendix A, we provide scatter-plots of predicted versus observed values for all transformer models to better characterize the relation. Here we see that models can generally differentiate between negatively and positively perceived pseudowords, but struggle more with differentiating pseudowords with extreme valence ratings from pseudowords with moderate valence ratings in the same direction.

2.3 Discussion

In line with Gatti et al. (2024a), letter uni-grams emerged as the best featurization to capture human valence ratings produced for isolated pseudowords. When zooming in on pseudowords which do not closely resemble a specific existing word, this advantage grew. However, when we considered the tails of the distribution, focusing on the pseudowords in the top and bottom quartiles, we observed that neural models emerged as the best performing models, better capturing subtler differences among pseudowords with similar perceived valence. Our results, hence, dovetail with observations from Haslett and Cai (2024) that part-words reliably encode meaning even when they do not reflect morphemes: we further show that such part-words provide representations that reflect human intuitions on the valence of entirely novel strings.

The focus on the tails of the valence distribution also highlights that it is easier to capture negative than positive valence ratings. This aligns with multiple observations that non-arbitrary mappings encode negative emotions more strongly (Adelman et al., 2018; Louwerse and Qu, 2017), in line with an adaptive account of form-meaning systematicity. In this view, the role of non-arbitrariness is to facilitate the communication of important messages through statistically reliable correlates in word form. If we use a word whose meaning our interlocutor does not know, we can still communicate negative emotions through systematic form-meaning correspondences, and thus elicit the appropriate response.

Finally, we see an advantage for LLaMa 3.2, which is a multilingual model with a comparatively large vocabulary. This suggests that systematic mappings apply cross-linguistically (M-BERT also shows a rather strong performance, especially on the tails of the distribution). Moreover, the large vocabulary hints to the possibility that form-meaning mappings matter most in low-frequency words.

3 Experiment 2: How novel words are represented in context

After establishing that all LLMs represent isolated pseudowords reflecting human ratings on pseudowords' valence, we turn to investigate how pseudowords' representations encoded by these models are affected by co-occurring words. How do they morph, if they do at all, when context kicks in? Do they change entirely according to the other words they co-occur with or do they maintain a stable, even if somewhat blurry, semantics?

3.1 Materials & Methods

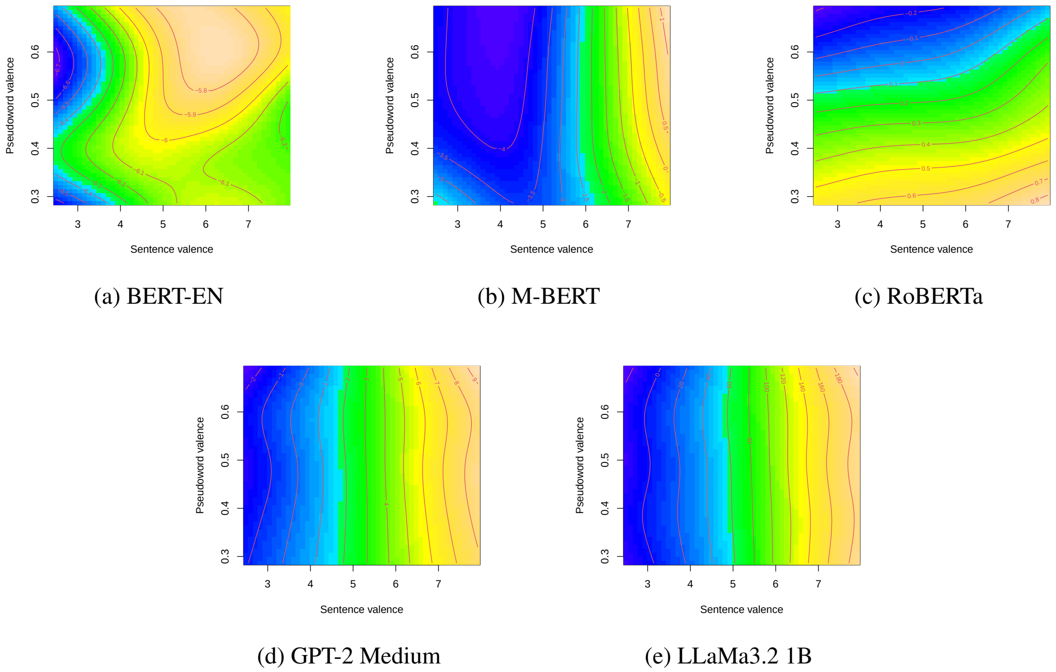
We constructed 100 sentence templates using words with valence norms from Warriner et al. (2013), in order to span the whole valence spectrum from very negative to very positive sentences. We identified sets of 4 or 5 words with similar valence and fed them to GEMMA (Team et al., 2024) with the instruction to create sentences of at most 10 words, without negation nor irony. Other content words were kept to a minimum and only used when necessary to form fluent sentences[3] Sentences were reviewed manually and modified to have an open slot at the end, to be filled by a noun.

We then randomly sampled 51[4] pseudowords from the *challenging set* to have a uniform distribution of valence ratings and avoid that results are entirely driven by the bulk of pseudowords with no clear perceived valence. We then plugged each pseudoword in each sentence template, generating a total of 5,100 sentences, and derived the fully contextualized embedding of each pseudoword in each sentence.

We feed each contextualized pseudoword embedding to the same regressor trained in Experiment 1 used to obtain the predicted valence of

[3] For example, given the words [*nauseating*, *stench*, *trashy*, *slaughterhouse*] which all share similarly negative valence ratings, we created the sentence *A nauseating stench from the trashy slaughterhouse* ***created*** *a lot of [open slot].*, where the verb *created* was needed to have a fluent sentence even though it did not belong to the original set of similarly rated words: in all these cases we used neutral words as much as possible.

[4] This is the number that yields the largest uniform distribution given the valence ratings of the pseudowords in the *challenging set*.



(a) BERT-EN

(b) M-BERT

(c) RoBERTa

(d) GPT-2 Medium

(e) LLaMa3.2 1B

Figure 1: Tensor products from GAMMs fitted to the difference between pseudoword valence predicted from static and contextualized embeddings, predicted using a Ridge Regressor trained on word valence (see Experiment 1).

pseudowords in isolation, obtaining the predicted valence for a contextualised pseudoword. The use of a same probe to make inferences on representations from different layers is consistent with studies investigating the internal representations of LLMs (Kuribayashi et al., 2025) — we further discuss this choice in Appendix B. We then subtracted an isolated pseudoword's predicted valencefrom the same pseudoword's predicted valence when plugged in each sentence, obtaining 100 differences per pseudoword. Positive differences indicate that the predicted valence was higher when a pseudoword appeared in the sentence, and thus that the sentence context made the pseudoword more positive.

This difference was entered as the dependent variable in a Generalized Additive Mixed Model (GAMM, Baayen et al., 2017), including pseudoword valence and sentence template valence (computed as the average valence rating of the words used to generate a sentence) as simple nonlinear smooths and their interaction as a partial tensor product. We also included random intercepts for sentence templates and pseudowords.

If pseudowords' representations are malleable, we should observe that the difference only depends on sentence template's valence: the more positive the sentence, the more a pseudoword will become positive. However, if the context and the pseudoword interact and thus pseudowords retain at least some of the semantics they encode in isolation when used in context, we should observe a significant effect of pseudoword valence and a significant partial tensor product. We expect a higher difference when a very negative pseudoword is plugged into a very positive sentence, reflecting the shift a pseudoword undergoes, from its original negative connotation to a more positive one (vice versa for negative differences). If the original valence of a pseudoword and that of the sentence template match, the difference should be close to 0.

3.2 Results

The analysis highlights distinct patterns across models, visible in Figure 1 which shows the tensor products[5]. Orange shades indicate higher predicted differences, blue shades indicate lower predicted differences[6], with red lines connecting predictions of the same magnitude. The plots combine the main effects and interactions in a single visualization showing the predicted values.

For M-BERT, GPT2 and LLaMa 3.2 the main

[5] All regression coefficients and plots for all smooths and tensor products are available in Appendix B, Table 3.

[6] The scale is specific to each plot, so orange can indicate negative differences and blue can indicate positive differences in different plots: what is constant is that orange denotes higher predicted differences than blue within a same plot.

effect of sentence valence and the tensor product significantly influenced the predicted difference, whereas pseudoword valence did not. These models show a clear pattern whereby the target difference is predominantly influenced by sentence valence: when a sentence features negative words, the representation of any pseudoword becomes more negative, while the opposite happens when the sentence is positive. A slight yet significant non-linear interaction with pseudoword valence in M-BERT, visible in the bottom right corner of the plot, shows that the representations of negatively perceived pseudowords require more positive templates to become more positive.

RoBERTa shows the closest pattern to our initial hypothesis and is the only model where pseudoword valence significantly influences the target difference and interacts with sentence valence. The tensor product shows how the predicted target difference is positive when negatively rated pseudowords are plugged in sentences consisting of positive words, and negative in the opposite situation. Therefore, across RoBERTa's attention blocks and layers, an initially negative pseudoword grows more positive if plugged in a positive sentence (resulting in a positive target difference), and vice versa, a positive pseudoword becomes more negative in a negative sentence (resulting in a negative target difference).

BERT-EN shows no significant main effect but a significant non-linear interaction which is however hard to interpret and may depend on the specific pseudowords and sentence templates used.

3.3 Discussion

This experiment reveals a predominant pattern in how neural language models contextualise pseudowords, whereby the dominant force is the sentence context: any pseudoword, when plugged into a negative sentence, is represented more negatively, and vice versa for positive sentences. Therefore, despite the fact that all models produced static embeddings for pseudowords which correlated with human valence ratings, the semantics of pseudowords appears to be largely malleable. Non-linear interactions with pseudoword valence are present, suggesting that pseudowords are not completely inert, but visual inspection of the tensor products clearly shows how sentence valence is by far the main driver.

RoBERTa, on the contrary, deviates from this pattern: pseudowords retain most of their valence connotation, and sentence templates make them shift most when the valence encoded in the pseudoword and that coming from the context clash.

Both patterns fit with possible cognitive accounts of pseudoword processing. The former predicts that pseudowords, while encoding valence in isolation, are so weak that context fully dictates their interpretation. The latter, on the contrary, predicts that pseudowords are strong enough to retain some of their valence even when processed in a sentence. It is interesting to note that the two multilingual models, M-BERT and LLaMa 3.2, which produced very strong correlations with human ratings when considering static embeddings, rely heavily on sentence context to shape the representation of those same strings in context. Speculatively, these models might have better picked up systematic form-meaning mappings while also learning to disregard those when context is available, since the same subword tokens must encode radically different connotations depending on the co-occurring ones across languages. Further studies should consider the specific sub-lexical tokens used to encode words and pseudowords and how their representations shift in contexts changing in language and valence.

To sum up, we have seen that despite the fact that all LLMs exhibited a sizable correlation with human ratings of isolated pseudoword valence, they embed pseudowords differently in context.

4 Experiment 3: How novel words affect sentence embeddings

In this last experiment, we investigate to what extent the valence encoded in pseudowords alters the representation of whole sentences. Once again, if pseudowords retain any of the semantics they encode when considering their static embeddings, we expect that sentence valence should change when sentences include different pseudowords. We further consider what happens when plugging words with different valence ratings in the sentences: intuitively, the word *plague* should make a sentence containing positive words such as *We were all incredibly happy to celebrate the funny [open slot].* quite less positive. Will the same happen for a negatively rated pseudoword like *puyfred*?

4.1 Materials & Methods

We used the same 100 sentence templates constructed for Experiment 2. For each of the five target LLMs we retrieved the 10 likeliest words to fill the open slot that have a neutral valence in Warriner et al. (2013)'s dataset. We thus generated 1,000 sentences (10 for each template), which each model considers at least somewhat probable and where the word filling the open slot does not have a clear valence. We then derived a sentence embedding[7] for each of the 1,000 sentences. In all cases, we used the embedding at the last layer, reflecting full integration of sentential context. Thus, we obtained 1,000 sentence embeddings, each derived from a sentence template designed to have a specific valence. We then trained a Ridge Regressor to predict the average valence of each template given the sentence embedding.

In order to ensure a robust pipeline, we used two different 10-fold cross-validation procedures (the pseudo-code for both is available in Appendix C). In the first, each fold contained 10% of the sentences derived from each template (so 1 sentence from each template per fold). A Ridge Regressor was trained on sentence embeddings from all sentence templates (each appearing 9 times, each time with a different filler word) and used to predict the sentence valence given a sentence embedding derived from the same templates featuring a different filler. In the second CV pipeline, instead, each fold contained 10% of the templates, each with all the 10 sentences derived from it. This time the regressor was used to predict the average valence of entirely different sentence templates. Unsurprisingly, the latter proved more challenging, with higher Root Mean Squared Error and lower correlation between predicted and true template valence – results are detailed in the Supplementary Materials. However, since both approaches yield qualitatively similar results, we only show the former here and present results for the latter in Appendix C.

During training we thus obtained 10 different regressors, one per fold, each used to predict the valence of a specific sentence from a sentence template. We averaged these predictions to get an average predicted valence of the template. Then, we derived sentence embeddings for each of the 5,100 sentences resulting from crossing pseudowords and sentence templates. We used all 10 trained regressors to get a predicted valence for each sentence and averaged the 10 predictions. We are interested in what changes between completing templates with neutral words versus pseudowords of different valence. We thus subtracted (1) the average predicted valence of a sentence template completed with neutral words from (2) the predicted valence of a sentence template featuring a pseudoword, averaged over the predicted ratings obtained from the 10 regressors trained during cross-validation. Positive differences indicate that the predicted sentence valence is higher when the sentence template includes a pseudoword, and vice versa.

This difference was entered as the dependent variable in a GAMM, with the same structure as detailed for Experiment 2. Once again, we are interested in the pseudoword valence main effect and in partial tensor product. If pseudowords' representations are robust, we expect that a negative pseudoword should trigger a more negative difference when plugged into a positive sentence, and vice versa for positive pseudowords.

Finally, we replicated the entire pipeline using 51 words sampled to have a uniform distribution over valence ratings from Warriner et al. (2013), excluding words used to create sentences in the first place, to investigate whether words and pseudowords affect sentence templates differently. Detailed results are provided in Appendix C.

4.2 Results

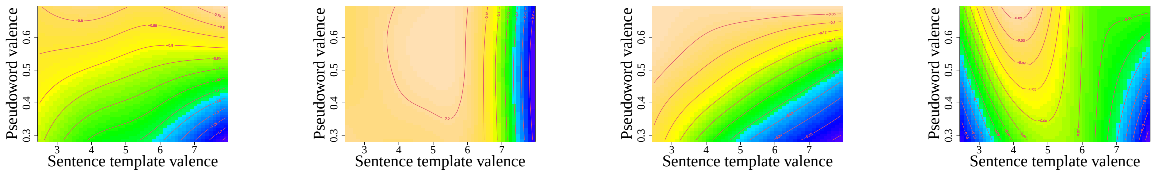
All LLMs show the same pattern, with minor differences in the exact shape and magnitude of the effects. Figure 2 displays the tensor products for all models except BERT-EN, which exhibited a weird pattern in the previous experiment[8] Figure 3 shows what happens with real words instead of pseudowords.

M-BERT and GPT-2 show very consistent patterns between words and pseudowords, with sentence embeddings becoming more negative when negative (pseudo)words are plugged into positive sentences. The effect of (pseudo)word valence is robust, with sentences becoming more positive when (pseudo)words grow more positive. Predictably, the effect of word valence is stronger than that of pseudoword valence, suggesting that these models encode words' semantics more reliably than pseudowords. This pattern dovetails with results from the previous experiment.

RoBERTa and LLaMa, on the contrary, show less consistent patterns: in both models, the most important predictor is sentence valence, although

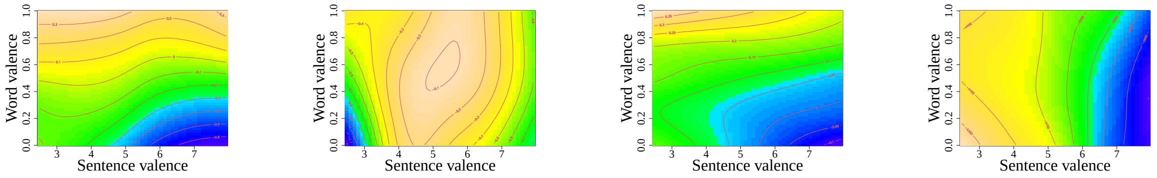
[7] With BERT-EN, M-BERT, and RoBERTa we used the embedding of the CLS token; for auto-regressive models we averaged the word embeddings of each token in a sentence.

[8] All smooths and tensor product visualizations for both CV pipelines are provided in Appendix C.



(a) M-BERT (b) ROBERTa (c) GPT-2 Medium (d) LLaMa 3.2 1B

Figure 2: Tensor products between the pseudoword valence and sentence template valence from GAMMs fitted to the difference between (a) valence predicted from sentence embeddings derived from template filled with neutral words and (b) template filled with pseudowords of different valence.



(a) M-BERT (b) ROBERTa (c) GPT-2 Medium (d) LLaMa 3.2 1B

Figure 3: Tensor products between the pseudoword valence and sentence template valence from GAMMs fitted to the difference between (a) valence predicted from sentence embeddings derived from template filled with neutral words and (b) template filled with pseudowords of different valence.

there are strong non-linear interactions with pseudoword valence that make interpretations difficult.

4.3 Discussion

In this experiment, sentence template valence plays an important role, which is expected since we analyzed sentence embeddings. Nonetheless, pseudoword valence also exerts a strong influence in most models: it is remarkable that even changing a single item in the sentence has such a systematic influence on the valence encoded by the sentence embedding. It is even more remarkable considering that (i) we are manipulating pseudowords, with no specific meaning profile the model can recognize, and (ii) a similar qualitative pattern emerges when replacing pseudowords with words.

The pattern observed for M-BERT and GPT-2 is surprising considering that in Experiment 2 the contextualized pseudoword representations they produced were almost entirely dictated by the sentence template. It is interesting then to see that the effect goes both ways: pseudoword representations change because of context (Experiment 2) but leave a trace that affects context as well (this experiment).

Even though pseudowords are entirely new, the tokens used to encode them appear to carry stable valence information that attention heads do not wash out: it is not enough to describe a pseudoword perceived to be negative as *cute* to fully convince an LLM. A positive sentence about a cute something, will be (slightly) more negative if that something is labeled as a *puyfred*, attesting to robust form-valence mappings. In general, the presence of a pseudoword with negative valence in a positive sentence slightly shifts the sentence's overall valence, demonstrating that these models pick up on form-meaning mappings.

5 General Discussion

Across three experiments, we showed that several LLMs are sensitive to non-arbitrary form-valence mappings in the input text. In Experiment 1, we assessed to what extent the static embeddings each model assigns to pseudowords capture valence in line with human ratings. In Experiment 2 we explored how such representations are contextualized, systematically varying the valence of co-occurring words. Finally, in Experiment 3, we investigated the trace that different pseudowords leave in sentence embeddings derived from sentences manipulated to consist of words with different valence. In this endeavor we leveraged gold standard human valence ratings for both words (Warriner et al., 2013) and pseudowords (Gatti et al., 2024a).

Experiment 1 showed that all models develop representations that capture human ratings across a number of evaluation sets which targeted pseudowords with extreme ratings (either very positive or very negative) as well as pseudowords with no clearly similar words that might have functioned as attractors and influenced the rating. While letter unigrams top the chart when considering all pseudowords, they fail to capture differences in perceived valence across positive and negative pseudowords. On the contrary, neural models fare better, particularly LLaMa 3.2 1B. Our evaluation thus complements and further qualifies findings from Gatti et al. (2024a), highlights the importance of using appropriate and informative evaluation sets, and speaks to the tension between single letters and broader systematic patterns in language when considering the semantic connotations elicited by entirely novel strings (Sabbatino et al., 2022; Gatti et al., 2022; Joosse et al., 2024; Sidhu and Pexman, 2018; Westbury et al., 2017). The slight yet consistent advantage of multilingual models further suggests that systematic form-meaning mappings may be partly cross-linguistic (Blasi et al., 2016; Ćwiek et al., 2022), benefiting a model which is not limited to learn from English form-meaning patterns. At the same time, Experiment 2 shows that M-BERT and LLaMa discard pseudowords' valence when these appear in context, possibly indicating that a model trained on multiple languages learns to downplay information coming from the make-up of a new word to prioritize contextual information, highlighting a tension between systematicity and arbitrariness across languages.

Our results further highlight the role of sub-word tokenization in LLMs (Kudo et al., 2024). FastText (Bojanowski et al., 2017) has been extensively used to gauge the likely meaning of novel strings thanks to its decomposition of every string into its constituent, overlapping n-grams (Gatti et al., 2022; Hendrix and Sun, 2020; Joosse et al., 2024; Sabbatino et al., 2022). Sub-word tokenization offers an alternative approach, which splits only *less frequent* words into *non-overlapping* segments: this study confirms that part-words do encode systematic form-meaning mappings (de Varda et al., 2024; Cai et al., 2024). Haslett (2025) has explored the role of part-words produced by sub-lexical tokenization in shaping representations considering semantic radicals in Chinese: a model's vocabulary, and hence the degree to which it segments word forms into part-words, has direct influences on the representations the model produces, underscoring how form-meaning mappings play an important role in LLMs. Future work should further investigate differences between these two strategies (overlapping n-grams from any string *versus* non-overlapping tokens derived only from those strings that the model does not store as whole words) in capturing non-arbitrary form-meaning relations. The performance of LLaMa — a multilingual model, with a large vocabulary — hints to the possibility that learning form-meaning mappings from low-frequency words only might provide more signal.

Finally, our evidence points to small yet systematic effects of both the perceived valence of a pseudoword and the valence of the words it co-occurs with, with two profiles emerging. On the one hand, M-BERT, GPT-2 and LLaMa 3.2 almost exclusively rely on sentence context to represent pseudowords in context (Savic et al., 2022). On the other hand, RoBERTa is predominantly sensitive to the valence of the pseudoword itself but is also influenced by the valence of co-occurring words. At the same time, in some models, sentences tend to absorb the valence of pseudowords, which emerge as having a sufficiently strong semantics to survive the contextualisation process. It is in this sense telling that the effect found when using words of varying valence instead of pseudowords is similar, showing that some LLMs consider novel strings consisting of part-words akin to words. This fits with behavioral evidence showing semantic effects for pseudowords in (primed) lexical decision tasks (Hendrix and Sun, 2020; Gatti et al., 2022), suggesting the existence of a shared cognitive encoding of words and pseudowords.

Future work should focus on collecting behavioral data from speakers of typologically distinct languages to understand how humans process pseudowords in context. Our results provide principled predictions about how novel words might be represented in context, to further study how our cognitive system deals with this very common situation. Especially during acquisition, most words are akin to pseudowords for learners (Cassani et al., 2020; Cassani and Limacher, 2021). Understanding how humans as well as cutting edge NLP models handle the ever present situation of quickly having to encode an entirely novel string in context and derive flexible and generalizable representations is an important step in the process of modeling human and artificial language learning (Weissweiler et al., 2023), which will further aid in developing theories of how linguistic form interfaces with meaning to shape this process (Sidhu and Pexman, 2018; Monaghan et al., 2014).

6 Limitations

This study has a few notable limitations which narrow the generalisability of the findings. First, the study only considers English pseudowords plugged in English sentences. This choice is primarily dictated by which resources are available, with large databases of valence ratings for both words and pseudowords being currently only available for English. This evidently limits the scope of our findings and leaves several questions open. Closely related, valence is certainly an important dimension of meaning, but does not encompass meaning: future studies should investigate whether the patterns we observed also apply to other salient semantic connotations.

Moreover, the lack of human behavioral data about processing of pseudowords in context limits the scope of our findings. At this stage, we can only describe what different LLMs do, but not yet adjudicate which model behavior more closely reflects human behavior. Such data would be crucial to develop cognitive theories of how linguistic form and sentence context affect the first stages in the process by which lexical meaning develops and becomes entrenched in semantic memory. At the same time, a comparison with human behavior would allow to understand which learning mechanisms deployed in language models are responsible for the encoding of pseudowords' semantic associations and their update in context. Crucially, we do not assume that a model whose behavior fits that of humans is necessarily human-like. We see and study these models as powerful information processing systems that can illuminate the presence and possible influence of statistical patterns in the input. Observing similar patterns of behavior in models and humans can then point to interesting directions to better chart how both behave and uncover fundamental similarities and differences in learning mechanisms and how they leverage input information.

Another limitation has to do with the fact that we probed text-based models and written pseudowords. Exploration of systematicity in spoken and signed pseudowords would be necessary to develop precise theories of how meaning is constructed in context for novel, word-like signs.

Tied to the issue of the role of form-meaning mappings, our study assumes a full-fledged vocabulary: the question of whether developmentally plausible vocabularies in terms of size and content also support stable and useful systematic form-meaning mappings remain open and cannot be answered based on our simulations.

Related to data availability, we relied on word valence ratings collected approximately 15 years ago. We believe it is unlikely that the valence of a substantial part of the words in this dataset shifted in this time period, but more recent ratings would improve the validity of the estimates and the modeling.

7 Ethical statements

This work uncovers possible biases in LLMs representations that bear relevance in the following areas. First, because of sub-lexical semantic associations, these models may encode certain names differently than other, with cascading effects. Second, for the same reason, this study shows how LLMs could be leveraged to craft product and brand names with specific semantic associations that could influence people's behavior in subtle ways.

The use of automatically generated sentences in Experiments 2 and 3 could pose ethical concerns, which we tried to mitigate by manually reviewing all sentences before using them.

Finally, the pseudoword and word valence ratings were made freely accessible by the authors: no license was stated for either dataset. No dataset used in this study contains unique identifiers that could disclose personal identities.

8 Acknowledgments

We thank Afra Alishahi for their help in conceptualising the study, Federico Bianchi for their critical feedback on a draft version of this paper, and Peter Hendrix for useful discussions throughout. We used ChatGPT for writing the pseudocode, which was then manually reviewed and edited.

References

James S. Adelman, Zachary Estes, and Martina Cossu. 2018. Emotional sound symbolism: Languages rapidly signal valence via phonemes. *Cognition*, 175:122–130.

R. Harald Baayen, Shravan Vasishth, Reinhold Kliegl, and Douglas Bates. 2017. The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, 94:206–234.

Damián E. Blasi, Søren Wichmann, Harald Hammarstrom, Peter F. Stadler, and Morten H. Christiansen. 2016. Sound-meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39):10818–23.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Marc Brysbaert, Boris New, and Emmanuel Keuleers. 2012. Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behav. Res. Methods*, 44(4):991–997.

Zhenguang Cai, Xufeng Duan, David Haslett, Shuqi Wang, and Martin Pickering. 2024. Do large language models resemble humans in language use? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 37–56.

Giovanni Cassani, Yu Ying Chuang, and R. Harald Baayen. 2020. On the semantics of nonwords and their lexical category. *Journal of Experimental Psychology-Learning Memory and Cognition*, 46(4):621–637.

Giovanni Cassani and Niklas Limacher. 2021. Not just form, not just meaning: Words with consistent form-meaning mappings are learned earlier. *Quarterly Journal of Experimental Psychology*, page 17470218211053472.

Roger Chaffin, Robin K Morris, and Rachel E Seely. 2001. Learning new word meanings from context: a study of eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1):225.

Aleksandra Ćwiek, Susanne Fuchs, Christoph Draxler, Eva Liina Asu, Dan Dediu, Katri Hiovain, Shigeto Kawahara, Sofia Koutalidis, Manfred Krifka, Pärtel Lippus, et al. 2022. The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B*, 377(1841):20200390.

Mark Davies. 2010. The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and linguistic computing*, 25(4):447–464.

Andrea Gregor de Varda, Daniele Gatti, Marco Marelli, and Fritz Günther. 2024. Meaning beyond lexicality: Capturing pseudoword definitions with language models. *Computational Linguistics*, 50(4):1313–1343.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mark Dingemanse, Damian Blasi, Gary Lupyan, Morten H. Christiansen, and Padraic Monaghan. 2015. Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, 19(10):603–615.

Daniele Gatti, Marco Marelli, and Luca Rinaldi. 2022. Out-of-vocabulary but not meaningless: Evidence for semantic-priming effects in pseudoword processing. *PsyArXiV*.

Daniele Gatti, Laura Raveling, Aliona Petrenco, and Fritz Günther. 2024a. Valence without meaning: investigating form and semantic components in pseudowords valence. *Psychonomic Bulletin & Review*, pages 1–13.

Daniele Gatti, Francesca Rodio, Luca Rinaldi, and Marco Marelli. 2024b. On humans'(explicit) intuitions about the meaning of novel words. *Cognition*, 251:105882.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, and Tobias Speckbacher. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, arXiv:2407.21783.

David A Haslett. 2025. Tokenization changes meaning in large language models: Evidence from chinese. *Computational Linguistics*, pages 1–30.

David A Haslett and Zhenguang G Cai. 2024. How much semantic information is available in large language model tokens?

Peter Hendrix and Ching Chu Sun. 2020. A word or two about nonwords: Frequency, semantic neighborhood density, and orthography-to-semantics consistency effects for nonwords in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Charles F. Hockett. 1960. The origin of speech. *Scientific American*, 203(3):88–97.

Aäron Yme Joosse, Gökçe Kuscu, and Giovanni Cassani. 2024. You sound like an evil young man: A distributional semantic analysis of systematic form-meaning associations for polarity, gender, and age in fictional characters' names. *Journal of Experimental Psychology: Learning, Memory, & Cognition*.

Emmanuel Keuleers and Marc Brysbaert. 2010. Wuggy: A multilingual pseudoword generator. *Behavior research methods*, 42:627–633.

Keito Kudo, Hiroyuki Deguchi, Makoto Morishita, Ryo Fujii, Takumi Ito, Shintaro Ozaki, Koki Natsumi, Kai Sato, Kazuki Yano, Ryosuke Takahashi, Subaru Kimura, Tomomasa Hara, Yusuke Sakai, and Jun Suzuki. 2024. Document-level translation with LLM reranking: Team-J at WMT 2024 general translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 210–226, Miami, Florida, USA. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large language models are human-like internally. *arXiv preprint arXiv:2502.01615*.

Wolfgang Köhler. 1929. *Gestalt Psychology*. Liveright, New York, NY.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414, Baltimore, Maryland. Association for Computational Linguistics.

Max Louwerse and Zhan Qu. 2017. Estimating valence from the sound of a word: Computational, experimental, and cross-linguistic evidence. *Psychon Bull Rev*, 24(3):849–855.

Padraic Monaghan, Richard C. Shillcock, Morten H. Christiansen, and Simon Kirby. 2014. How arbitrary is language? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 369(1651):20130299.

Charles E Osgood, George J Suci, and Percy H Tannenbaum. 1957. The measurement of meaning.

Tiago Pimentel, Arya D. McCarthy, Damian Blasi, Brian Roark, and Ryan Cotterell. Meaning to form: Measuring systematicity as information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1751–1764. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Maria Ryskina, Ella Rabinovich, Taylor Berg-Kirkpatrick, David Mortensen, and Yulia Tsvetkov. 2020. Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 367–376, New York, New York. Association for Computational Linguistics.

Valentino Sabbatino, Enrica Troiano, Antje Schweitzer, and Roman Klinger. 2022. "splink" is happy and "phrouth" is scary: Emotion intensity analysis for nonsense words. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 37–50, Dublin, Ireland. Association for Computational Linguistics.

Edward Sapir. 1929. A study in phonetic symbolism. *Journal of experimental psychology*, 12(3):225 – 239.

Olivera Savic, Layla Unger, and Vladimir M Sloutsky. 2022. Exposure to co-occurrence regularities in language drives semantic integration of new words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 48(7):1064.

David M Sidhu and Penny M. Pexman. 2018. Five mechanisms of sound symbolic association. *Psychonomic Bulletin and Review*, 25(5):1619–1643.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45:1191–1207.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.

Chris Westbury, Geoff Hollis, David M. Sidhu, and Penny M. Pexman. 2017. Weighing up the evidence for sound symbolism: Distributional properties predict cue strength. *Journal of Memory and Language*, 99:122–150.

Chris Westbury, Michelle Yang, and Kris Anderson. 2024. The principal components of meaning, revisited. *Psychonomic Bulletin & Review*, pages 1–23.

Bodo Winter, Márton Sóskuthy, Marcus Perlman, and Mark Dingemanse. 2022. Trilled/r/is associated with roughness, linking sound and touch across spoken languages. *Scientific Reports*, 12(1):1035.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Details over Experiment 1

Table 2 shows the best α for the Ridge Regressors trained on each feature representation in Experiment 1. For all models, we explored the following values: [0.01, 0.05, 0.1, 0.5, 1, 2, 5, 10, 50, 100].

Model	Best Alpha
Uni-grams	100
FastText	50
LLama 3.2	10
GPT-2 Medium	10
BERT-EN	10
M-BERT	10
RoBERTa	10

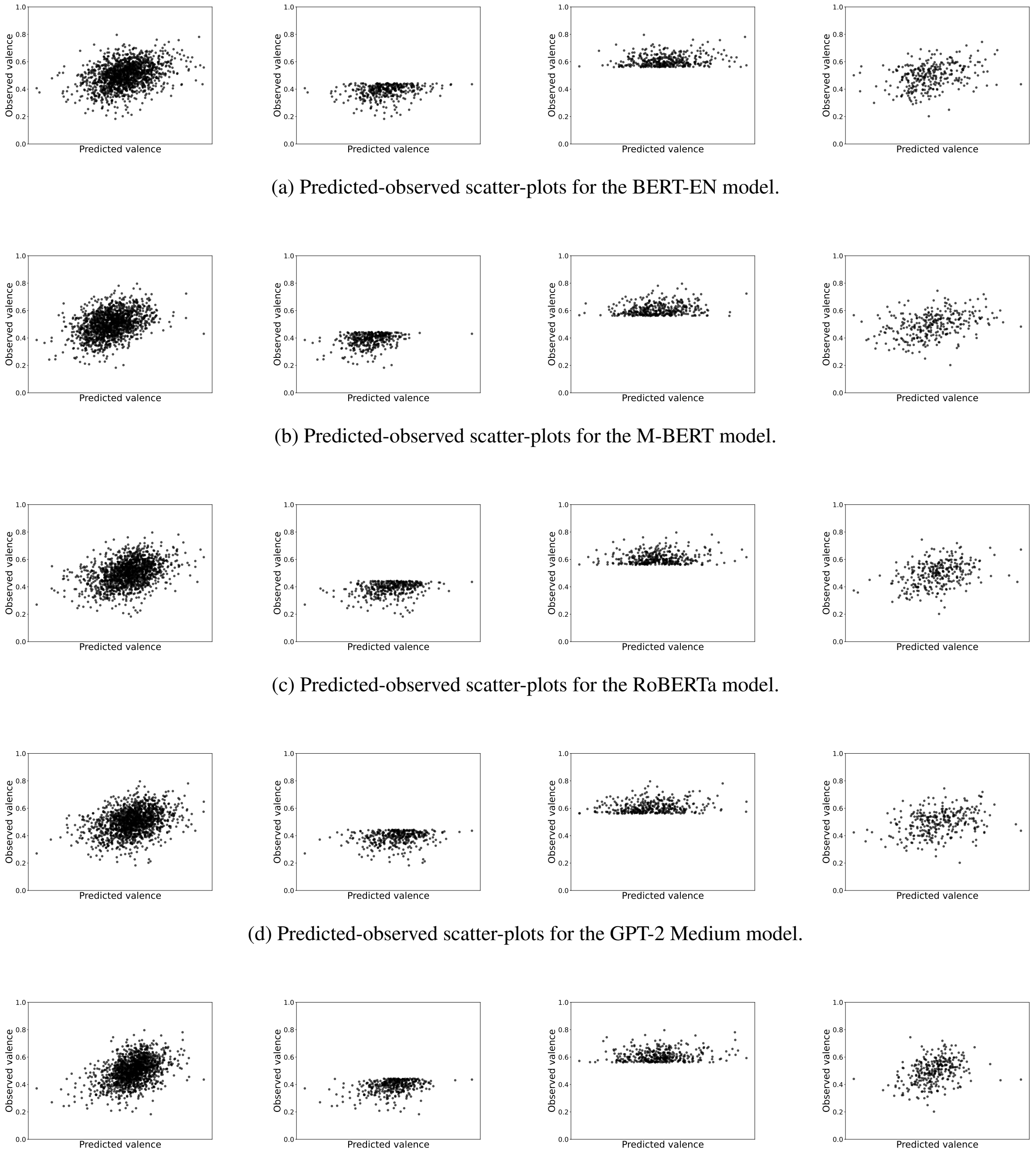
Table 2: Best alpha values and explored hyperparameter ranges for Ridge Regressors across different models.

The LLMs we used have the following number of parameters. BERT-EN: 110M; M-BERT: 110M; RoBERTa: 125M; GPT-2: 345M; LLaMa 3.2: 405B. All experiments were run using Google COLAB, with runtime set to NVIDIA A100 GPU, and a virtual machine equipped with NVIDIA RTX A4500 GPU. The amount of computing hours used in all phases of this project totals around 250 hours. Running the computational simulations used to produce the results reported takes less than 1 hour for all three experiments on the NVIDIA RTX A4500, provided that one has embedded all words, pseudowords, and sentences.

B Details over Experiment 2

We argue an important design choice in our experiment consists in training a single regressor on static embeddings and use it to make inferences for both static and contextualised embeddings, as this allows to model the direct influence that sentence context exerted on the representations. This approach is attested in the literature, with approaches like *logitlens*. Evidently, though, for this approach to be tenable, the two embedding spaces need to be directly comparable and not a rotation of each other. For models trained with weight-tying, this is built-in. For models trained with independent embedding and un-embedding matrices, this is not a guarantee. To ensure that BERT models satisfy the requirement of having comparable embedding spaces at different layers despite the many linear and non-linear transformations, we conducted the following experiment.

First, we harvested $\sim$25k words from different psycholinguistic resources to collate a representative sample, and for each we derived the static embedding, $\mathbf{w_{static}}$, using BERT-base. Then, we fetched 50 sentences from the Corpus of Contemporary American English (CoCA) for each word, and derived the contextualised word embedding (CWE) of the target word in each sentence, $\mathbf{w_{12}}$,



(a) Predicted-observed scatter-plots for the BERT-EN model.

(b) Predicted-observed scatter-plots for the M-BERT model.

(c) Predicted-observed scatter-plots for the RoBERTa model.

(d) Predicted-observed scatter-plots for the GPT-2 Medium model.

(e) Predicted-observed scatter-plots for the LLaMa3.2 model.

Figure 4: Scatterplots showing the relation between observed and predicted valence ratings in different models over different evaluation sets (from left to right: all pseudowords, the 25% pseudowords with lowest observed ratings, the 25% pseudowords with the highest observed ratings, the pseudowords in the challenging set.

obtaining 50 CWEs per word. We then averaged the 50 $\mathbf{w_{12}}$ vectors for a same word to obtain a prototype vector for each word, $\mathbf{w_{proto}}$. For each $\mathbf{w_{static}}$, we retrieved its 10 nearest neighbors among the 25k prototype embeddings, embedded at layer 12, $10nn_{static_{t}o_{p}roto}$. Then we retrieved the 10 nearest neighbors of each prototype embedding among the 25k available prototype embeddings, $10nn_{proto-to-proto}$. Given the embeddings and the nearest neighbors, we then computed (i) how often the top nearest neighbor of $\mathbf{w_{static}}$ among $10nn_{static-to-proto}$ is the word itself and (ii) the Jaccard coefficient between $100nn_{static-to-proto}$ and $10nn_{proto-to-proto}$, indicating how many of the nearest neighbors are shared between $\mathbf{w_{static}}$ and $\mathbf{w_{proto}}$. If the embedding spaces are not aligned, the prototype embedding of a word should *not* be the nearest neighbor of its static representation and the 10 nearest neighbors of the static and prototype representations among other prototype representations should not be comparable.

Our results show that in 66% of cases the top nearest neighbor (1NN) of $\mathbf{w_{static}}$ in $nn_{static-to-proto}$ is indeed the word itself. In other cases, it is often an(other) inflected form of the same lemma (e.g., $\mathbf{includes_{proto}}$ as the 1NN of $\mathbf{included_{static}}$). In many other cases, it is a semantically related word type (e.g., $\mathbf{sobbing_{proto}}$ as the 1NN of $\mathbf{crying_{static}}$). For other word types, the 1NN shared one or more tokens with the target word (e.g., $\mathbf{sojourn_{proto}}$ as the 1NN of $\mathbf{adjourn_{static}}$). Finally, there are a minority of cases where the 1NN is entirely unrelated (e.g., $\mathbf{filched_{proto}}$ as the 1NN of $\mathbf{indigestion_{static}}$). As to the Jaccard coefficient, its average is 0.2, indicating that 3 to 4 nearest neighbors from $10nn_{proto-to-proto}$ also feature among the $10nn_{static-to-proto}$, confirming that the static embedding entertains consistent geometrical relations to the corresponding prototype embedding.

Crucially, when randomly permuting the dimensions—which amounts to rotating the embedding space without changing the similarities within the space—the 1NN of $\mathbf{w_{static}}$ in $10nn_{static-to-proto}$ is never the word itself and the Jaccard coefficient is consistently 0. This confirms that, while different layers use a different portion of the embedding space, in line with the increasing anisotropism across BERT's layers, the coordinates of the underlying latent space are consistent across layers, justifying our approach.

Table 3 summarizes all coefficients for the GAMMs fitted to model the difference between pseudoword valence predicted from static, a-contextual embeddings and valence predicted from fully contextualized pseudoword embeddings, when each pseudoword appeared in each of 100 sentences generated to have a specific valence. These coefficients refer to Experiment 2. Figure 5 visualizes the effects for better interpretability. The tensor products for all models but LLaMa also appear in the main text and are reported here for completeness.

C Details over Experiment 3

Algorithm 1 and Algorithm 2 detail the two cross-validation strategies used to train Ridge Regressors to estimate the average valence of a sentence template filled with neutral words given the sentence embedding and to be then used to predict the valence of a sentence template filled with pseudowords.

When using Algorithm 1, the dataset containing sentence templates with neutral words, $\mathcal{D}_{neu}$, is partitioned in *k* folds such that each fold contains *k*% of the sentences from each template. Each of the 10 sentences generated from a same template was thus assigned to a different fold. First, we derived a sentence embedding for each sentence in $\mathcal{D}_{neu}$. Then, a different Ridge Regressor is trained on the sentence embeddings in all folds but one and used to predict the valence of the sentences in the held-out fold. This process yields 10 different predicted valences for each template, which are averaged to obtain the predicted valence of a sentence template, $\hat{v}_{templ}$. First, we ran a grid search over different values for the hyper-parameter α and selected the best one by looking at how well $\hat{v}_{templ}$ approximated the template average valence considering Pearson's r. This CV pipeline, however, also yields 10 different regressors, which we used to predict the valence of sentences obtained by plugging each pseudoword in each template. We passed each sentence containing a pseudoword through all the 10 regressors trained during the CV pipeline, and averaged their predictions to obtain the predicted valence for a sentence, $\hat{v}_{templ+pw}$, so that the predicted valence did not depend on a specific subset of the training data. Finally, we derived the target difference $d_{(templ,pw)} = \hat{v}_{templ+pw} - \hat{v}_{templ}$, using the average valence of the appropriate template.

With Algorithm 2, we followed a similar approach but trained models on a different, and harder, generalization. Whereas in Algorithm 1, each regressor has seen all templates during training and is asked to predict the valence of new sentences featuring a combination of known template and a new neutral word that never occurred in training, here we ask regressors to generalize to entirely new templates. First, we randomly assigned templates to fold: since templates cover the entire valence spectrum, it is important that each fold contains templates randomly sampled from the entire valence spectrum, rather than grouping negative templates together, and separate from positive ones. Again,

BERT-EN	**Parametric Coefficients**	**Estimate**	**Std. Error**	**t**	**p**
	Intercept	-5.285	0.322	-16.4	<.001
	Smooth Terms	**edf**	**Ref. df**	**F**	**p**
	s(Template_avg_valence)	1.866	1.876	1.154	0.230
	s(Valence_filler)	1.000	1.000	0.079	0.779
	ti(Template_avg_valence, Valence_filler)	7.357	9.737	3.985	<.001
M-BERT	**Parametric Coefficients**	**Estimate**	**Std. Error**	**t**	**p**
	Intercept	1.531	0.397	3.855	<.001
	Smooth Terms	**edf**	**Ref. df**	**F**	**p**
	s(Template_avg_valence)	2.929	2.941	26.72	<.001
	s(Valence_filler)	1.000	1.000	0.00	0.984
	ti(Template_avg_valence, Valence_filler)	6.677	9.020	17.02	<.001
RoBERTa	**Parametric Coefficients**	**Estimate**	**Std. Error**	**t**	**p**
	Intercept	0.584	0.073	8.006	<.001
	Smooth Terms	**edf**	**Ref. df**	**F**	**p**
	s(Template_avg_valence)	1.859	1.867	2.877	0.096
	s(Valence_filler)	1.000	1.000	13.940	<.001
	ti(Template_avg_valence, Valence_filler)	5.810	7.780	4.386	<.001
GPT-2 Medium	**Parametric Coefficients**	**Estimate**	**Std. Error**	**t**	**p**
	Intercept	0.9774	0.3059	3.195	<.01
	Smooth Terms	**edf**	**Ref. df**	**F**	**p**
	s(Template_avg_valence)	1.993	1.995	75.330	<.001
	s(Valence_filler)	1.000	1.000	0.038	0.846
	ti(Template_avg_valence, Valence_filler)	6.195	7.735	17.377	<.001
LLaMa 3.2	**Parametric Coefficients**	**Estimate**	**Std. Error**	**t**	**p**
	Intercept	18.112	4.948	3.661	<.001
	Smooth Terms	**edf**	**Ref. df**	**F**	**p**
	s(Template_avg_valence)	2.234	2.236	99.160	<.001
	s(Valence_filler)	1.000	1.000	0.269	0.604
	ti(Template_avg_valence, Valence_filler)	7.895	9.872	8.906	<.001

Table 3: GAMM coefficients for parametric terms and smooth terms for all LLMs in Experiment 2, predicting the difference in predicted valence between the static layer and the fully contextualized layer, for pseudowords plugged into 100 sentences whose valence was manipulated to span from very positive to very negative.

CV approach	**Model**	**MSE**	**RMSE**	**MAE**	**Pearson r**
k% sentences per template per fold	BERT-EN	0.0457	0.2137	0.1563	0.9891
	M-BERT	0.1223	0.3497	0.2516	0.9708
	GPT-2 Medium	0.0055	0.0743	0.0531	0.9987
	LLama 3.2	2.1175	1.4552	1.2232	0.0421
	RoBERTa	5.9180	2.4327	1.9637	0.0354
k% templates per fold	BERT-EN	0.6106	0.7814	0.6342	0.8429
	M-BERT	0.8500	0.9219	0.7353	0.7716
	GPT-2 Medium	0.4453	0.6673	0.5424	0.8901
	LLama 3.2	2.1590	1.4694	1.2355	-0.0292
	RoBERTa	5.9667	2.4427	1.9632	0.0054

Table 4: Performance metrics for Ridge Regression models evaluated on how well they can predict the average valence of sentence templates in Experiment 3, provided separately for the two Cross Validation strategies detailed in the paper and in the pseudocode above. All models were trained using Ridge Regression with $\alpha \in \{0.0001, 0.001, 0.01, 0.1, 1, 10\}$, and the best performing α was consistently 10.

we derived sentence embeddings from each sentence. Once templates were assigned to folds, we trained k regressors, which we used to obtain a predicted valence for all sentences from unseen templates. Once again, we derived $\hat{v}_{templ}$ by averaging the valence predicted for all sentences generated from a same template. Unlike before, each regressor has seen all instances from some templates, and cannot be used to predict the valence of a sentence featuring that same template in combination with a pseudoword. Therefore, we do not need to average predictions from different regressors and simply derived $\hat{v}_{templ+pw}$ by applying the appropriate trained regressor on each sentence resulting from the combination of templates not seen in the trainig of that regressor and pseudowords. Finally, we again computed $d_{(templ,pw)} = \hat{v}_{templ+pw} - \hat{v}_{templ}$.

Table 4 summarizes the performance of the two CV pipelines in inferring the valence of a template. In general, we see that both manage to predict the average valence of templates, with sizable correlations between predicted and observed valence. As it could be easily predicted, Algorithm 1 yields a nearly perfect correlation since the generalization required of the CV pipeline is not too hard: every template appears in training after all. However, we see that the procedure works also when using Algorithm 2. LLMs differ as well, with LLaMa affording the most accurate predictions, and M-BERT the least accurate.

Importantly, however, Figure 6 and Figure 7 highlight that GAMMs fitted to model the target difference capture very similar patterns regardless of which cross-validation procedure is used to train Ridge Regressors to predict valence from sentence embeddings. This confirms that whatever pattern is observed does not depend on how regressors are trained and increases the robustness of the results.

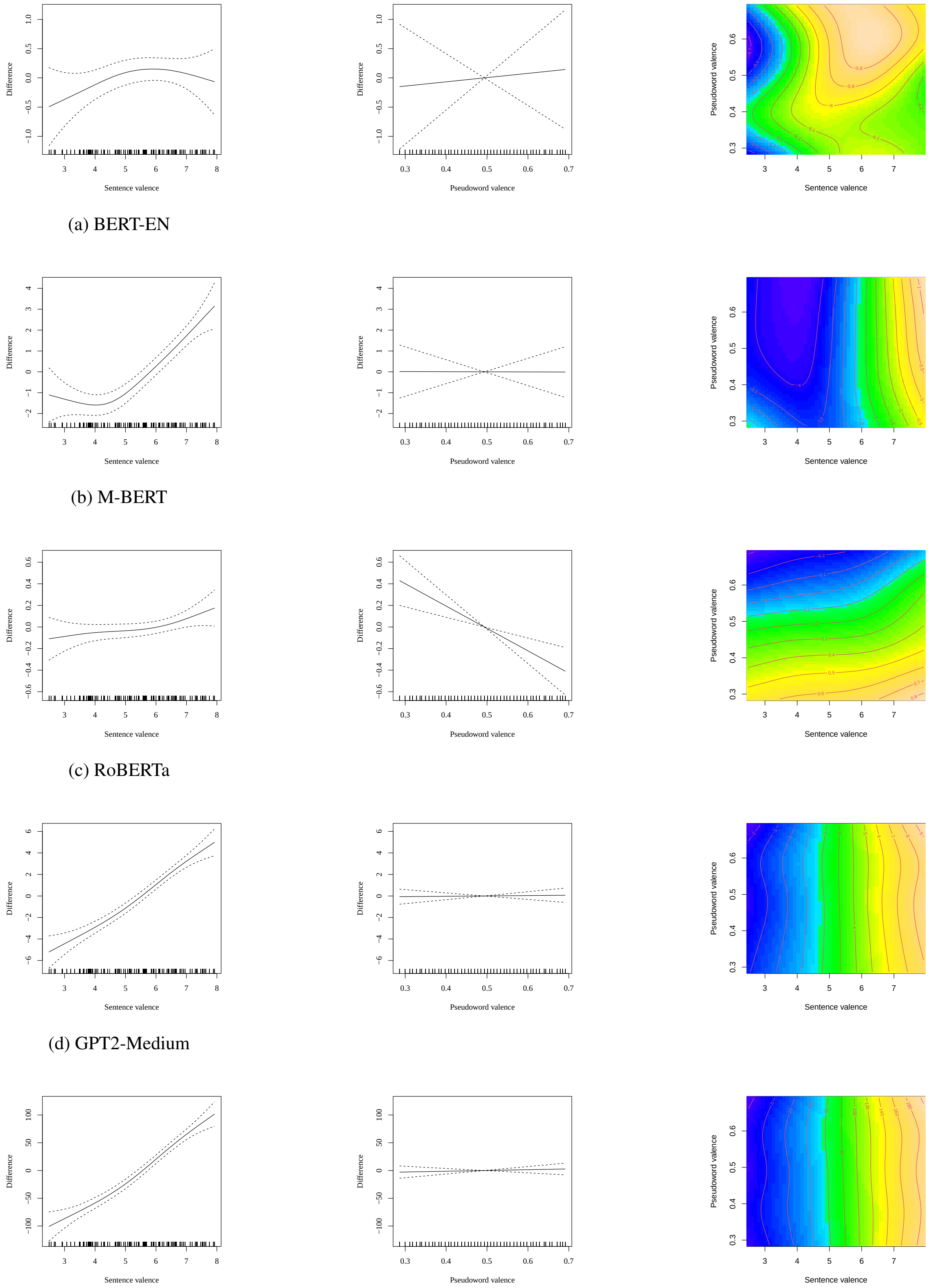
Finally, we provide the plots for smooths and tensor products when replicating pipeline 1 (with 10% of sentences from each template allocated to each fold) with 51 words rather than pseudowords, sampled to span the whole valence spectrum and not overlap with words used in sentence templates. As mentioned in the main text, the pattern emerging from the GAMMs is remarkably similar to that reported for pseudowords, confirming that for all LLMs we surveyed, pseudowords and words largely influence sentence embeddings similarly.

Algorithm 1 Cross-Validation: k% Sentences per Template per Fold

```
Input: Dataset $\mathcal{D}_{neu}$ with a sentence embedding generated from each sentence template filled with at least k neutral words, Dataset $\mathcal{D}_{pw}$ with sentence embeddings generated from the same sentence templates filled with pseudowords, Number of folds $k$.
Output: difference in predicted valence between sentence templates and sentence templates featuring each of the target pseudowords.
Step 1: Assign sentence embeddings to folds
function ASSIGN_FOLDS_KSENTENCES_PERTEMPLATE_PERFOLD($\mathcal{D}_{neu}, k$)
    for Each template $T$ in $\mathcal{D}_{neu}$ do
        Shuffle sentence embeddings of $T$
        for Each fold $f \in \{1, \ldots, k\}$ do
            Assign exactly k% sentence embeddings from $T$ to fold $f$
        end for
    end for
end function
Apply assign_folds_ksentences_perTemplate_perFold($\mathcal{D}_{neu}$, k)
Step 2: Hyperparameter tuning
for Each hyperparameter $\alpha$ in search space do
    for Each fold $f \in \{1, \ldots, k\}$ do
        Train Ridge regression model using all folds except $f$
        Predict the valence of sentences in fold $f$
        Compute Pearson correlation coefficient $r_f$
    end for
    Compute average Pearson $\bar{r}$ over all folds
    if $\bar{r}$ > best found correlation then
        Update best $\alpha$
        Save Ridge Regressors
    end if
end for
compute $\hat{v}_{templ}$ by averaging the predicted valence of all sentences from a same template
Step 3: Predict on sentences with pseudowords
function PREDICT_PSEUDOWORDS_KSENTENCES_PERTEMPLATE_PERFOLD($\mathcal{D}_{pw}, RidgeRegressors$)
    for Each fold $f \in \{1, \ldots, k\}$ do
        Use regressor from fold $f$ to predict the valence of sentence embeddings derived from templates filled with pseudowords
        Store predictions
    end for
    Compute $\hat{v}_{templ+pw}$ by averaging the valence predicted by all regressors for a sentence-pseudoword combination
end function
Use models with best $\alpha$ and predict using predict_pseudowords_ksentences_perTemplate_perFold()
Compute overall performance metrics
Compute $d = \hat{v}_{templ+pw} - \hat{v}_{templ}$ for each sentence-pseudoword combination
```

Algorithm 2 Cross-Validation: k% Templates per Fold

```
Input: Dataset $\mathcal{D}_{neu}$ with a sentence embedding generated from each sentence template filled with at least k neutral words, Dataset $\mathcal{D}_{pw}$ with sentence embeddings generated from the same sentence templates filled with pseudowords, Number of folds $k$.
Output: difference in predicted valence between sentence templates and sentence templates featuring each of the target pseudowords.
Step 1: Assign sentence embeddings to folds
function ASSIGN_FOLDS_KTEMPLATES_PERFOLD($\mathcal{D}_{neu}, k$)
    for Each fold $f \in \{1, \ldots, k\}$ do
        Assign the sentence embeddings from k% of templates to fold $f$
    end for
end function
Apply assign_folds_ktemplates_perFold($\mathcal{D}_{neu}$, k)
Step 2: Hyperparameter tuning
for Each hyperparameter $\alpha$ in search space do
    for Each fold $f \in \{1, \ldots, k\}$ do
        Train Ridge regression model using all folds except $f$
        Predict the valence of sentences in fold $f$
        Compute Pearson correlation coefficient $r_f$
    end for
    Compute average Pearson $\bar{r}$ over all folds
    if $\bar{r} >$ best found correlation then
        Update best $\alpha$
        Save Ridge Regressors
    end if
end for
Compute $\hat{v}_{templ}$ by averaging the predicted valence given each sentence embedding from a same template in $\mathcal{D}_{neu}$
Step 3: Predict on pseudowords
function PREDICT_PSEUDOWORDS_KTEMPLATES_PERFOLD($\mathcal{D}_{pw}, RidgeRegressors$)
    for Each fold $f \in \{1, \ldots, k\}$ do
        Assign each pseudoword to the fold of its corresponding template
        Use model from fold $f$ to predict on pseudowords assigned to that fold
        Store $\hat{v}_{templ+pw}$, the valence predicted for a sentence-pseudoword combination
    end for
end function
Use models with best $\alpha$ and predict using predict_pseudowords_ktemplates_perFold()
Compute overall performance metrics
Compute $d = v_{templ+pw} - \hat{v}_{templ}$ for each sentence-pseudoword combination
```



(a) BERT-EN

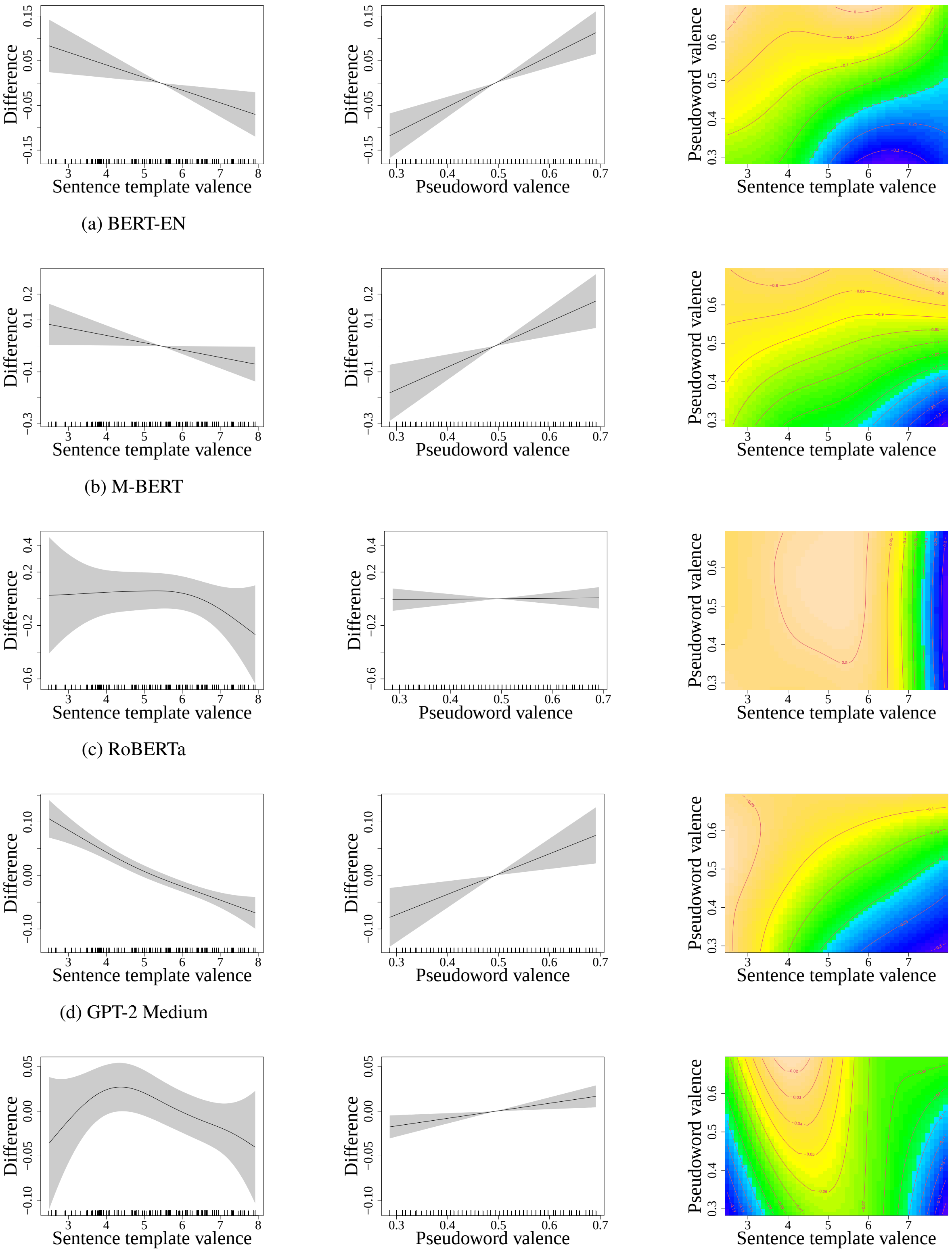
(b) M-BERT

(c) RoBERTa

(d) GPT2-Medium

(e) LLaMa 3.2 1B

Figure 5: Each row showcases the smooths (left: valence of the sentence template; center: valence of the pseudoword) and tensor product for a different LLM, as emerging from a GAMM fitted to model the difference in predicted valence between the static layer and the fully contextualized layer, for pseudowords plugged into 100 sentences whose valence was manipulated to span from very positive to very negative.



(a) BERT-EN

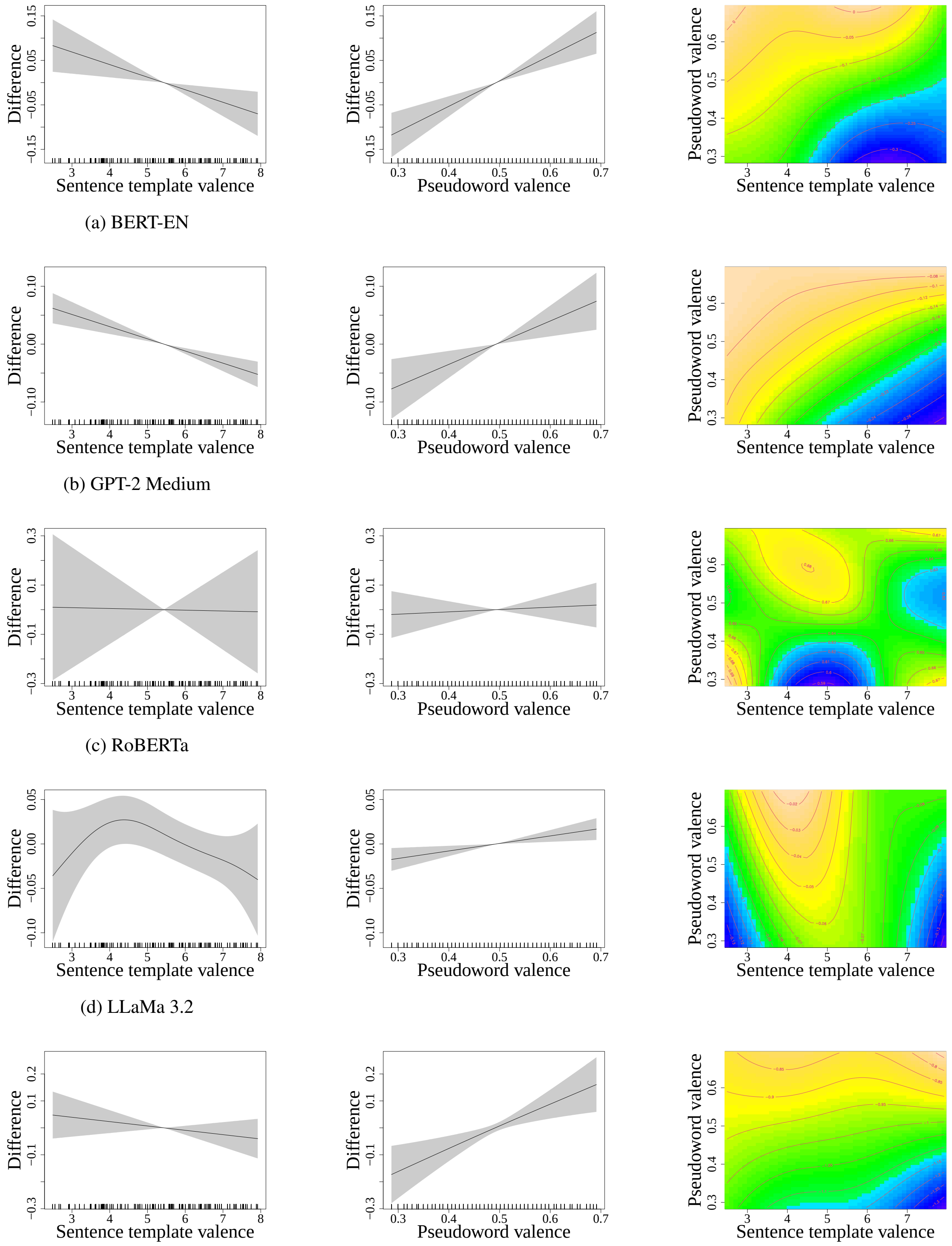
(b) M-BERT

(c) RoBERTa

(d) GPT-2 Medium

(e) LLaMa 3.2

Figure 6: Visualizations of the effects of template valence, **pseudoword** valence, and their partial tensor product as estimated by GAMMs for five different LLMs in Experiment 3. The target difference is computed using the *k% sentences per template per fold* pipeline (Algorithm 1).



(a) BERT-EN

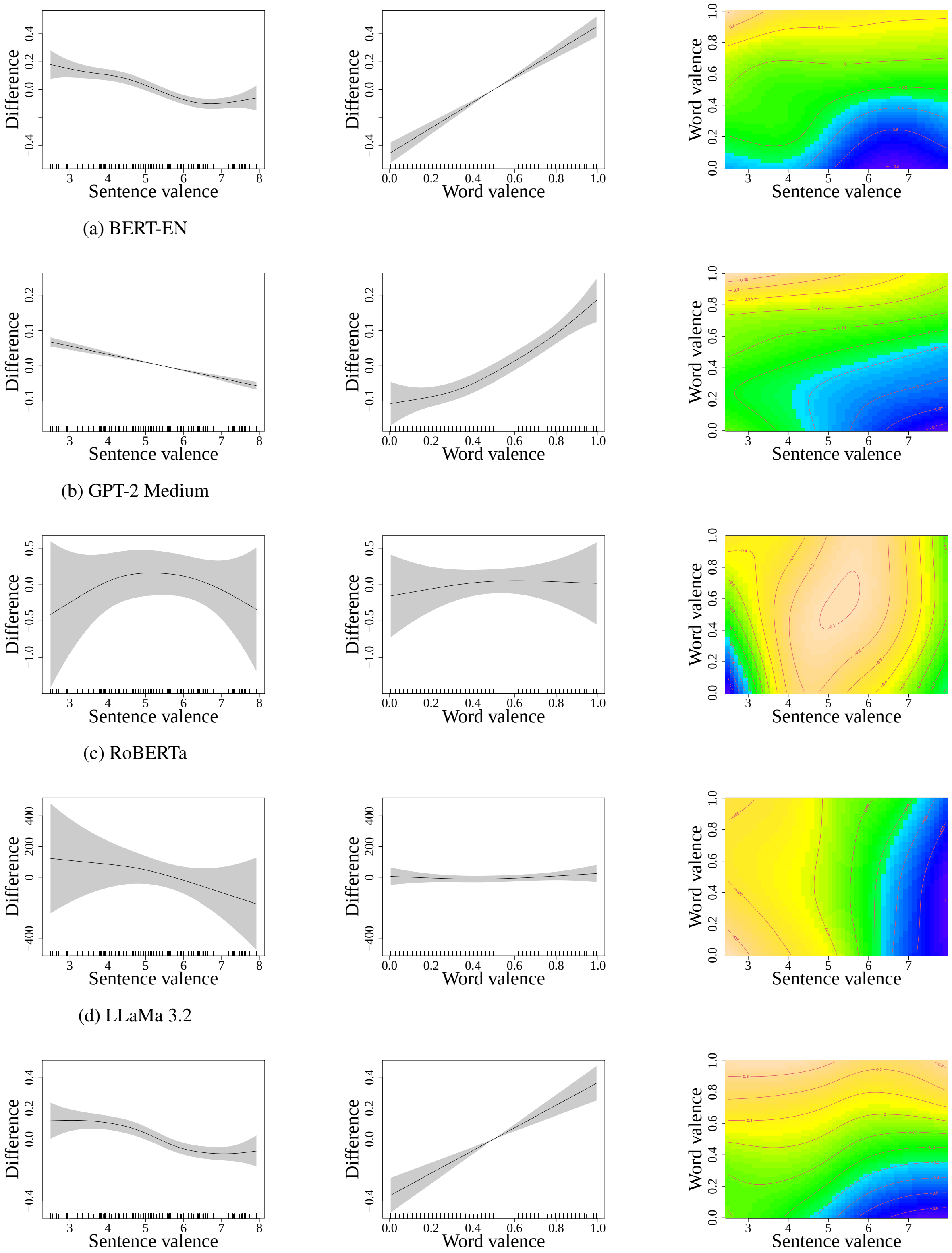
(b) GPT-2 Medium

(c) RoBERTa

(d) LLaMa 3.2

(e) M-BERT

Figure 7: Visualizations of the effects of template valence, **pseudoword** valence, and their partial tensor product as estimated by GAMMs for five different LLMs in Experiment 3. The target difference is computed using the *k% templates per fold* pipeline (Algorithm 2).



(a) BERT-EN

(b) GPT-2 Medium

(c) RoBERTa

(d) LLaMa 3.2

(e) M-BERT

Figure 8: Visualizations of the effects of template valence, **word** valence, and their partial tensor product as estimated by GAMMs for five different LLMs in Experiment 3. The target difference is computed using the *k% sentences per templates per fold* pipeline (Algorithm 1).