Language Resources and Evaluation
https://doi.org/10.1007/s10579-025-09866-9

ORIGINAL PAPER

# Multilingual prediction of semantic norms with language models: a study on English and Chinese

**Bo Peng**[1] · **Yu-yin Hsu**[1] · **Emmanuele Chersoni**[1] · **Le Qiu**[1] · **Chu-Ren Huang**[1]

Received: 5 March 2025 / Accepted: 8 July 2025


## Abstract

Lexical semantic norms characterize each lexical concept in terms of a set of semantic features for the words of a language. They provide essential resources for behavioral, computational, and neuro-cognitive studies of language and human cognition. Recent research advocate for the need for cognitively motivated feature sets, arguing that semantic representations grounded in human cognition can facilitate cross-linguistic modeling and even enable the prediction of a word's semantic features based on its translation in another language. In this study, we present a new dataset of brain-based, Binder-style semantic norms for Chinese. Using the corresponding English dataset and the representational power of multilingual language models, we conduct systematic experiments on semantic norm prediction both within and across languages. We evaluate monolingual and English-Chinese cross-lingual norm prediction using two different methods: embedding-based regression vs. prompting with large language models. Our results show that bidirectional models from the BERT family and GPT-4 achieve a good level of accuracy, with moderate-to-high correlations with human ratings. Notably, in the cross-lingual setting, the best and the worst predicted features align with the higher and lower end of levels of human agreement when comparing norms of words between translated words. Our results support a novel computational approach for supplementing and expanding cognitive semantic norms, highlighting the potential of language models to bridge cross-linguistic semantic representations.

**Keywords** Semantic norms · Large language models · Word embeddings · Psycholinguistics · Multilinguality · Cognitive modeling

Extended author information available on the last page of the article

Published online: 06 September 2025

## 1 Introduction

How can we systematically describe the meaning of words? Linguists have long sought to break down word meanings into more fundamental components, with the aim of identifying a finite set of semantic primitives that can account for the vast and diverse lexicon of natural languages. These features are often symbolic in nature (Jackendoff, 1992; Wierzbicka, 1996) and used in categorical representations, in which each feature can be either present or not (e.g. *girl* as [+HUMAN -MAN -ADULT...]). In addition to the difficulty of establishing objective criteria for choosing a universal set of primitives, another challenge arises from the inherently gradient nature of lexical meaning. Features within a concept may vary in their degrees of prototypicality for that concept (Murphy, 2004), making discrete symbolic models inadequate for fully capturing the semantics of words.

In psycholinguistics and cognitive psychology, it is common to collect *semantic feature norms* (McRae et al., 2005; Vinson & Vigliocco, 2008; Devereux et al., 2014; Buchanan et al., 2019) by asking human subjects to produce verbal descriptors for lexical concepts (e.g., HAS_FEATHERS, CAN FLY, LAYS EGGS for the concept of *bird*) and using a weight derived from the frequency of the association between a descriptors and a concept to indicate the salience of a given feature. These norms have the advantage of immediate explainability, since they directly reflect human intuitions and are highly interpretable, but they are also highly subjective, making cross-linguistic comparison difficult. Even when the same concepts are studies, semantic norms collected in English may differ substantially from those collected in other languages, e.g., German (Kremer & Baroni, 2011), Italian (Kremer & Baroni, 2011; Montefinese et al., 2013), Spanish (Vivas et al., 2017), Finnish (Kivisaari et al., 2023).

An alternative approach was introduced by Binder et al. (2016), who proposed a set of 65 cognitively-motivated semantic primitives (henceforth **Binder features**). These features are considered brain-based, because each of them is associated with a specific neural motivated processing, with the evidence drawn from the neuroscientific literature. Since these features are grounded in human cognition, they should theoretically be extended across languages: if lexical meanings emerge from human experience and from fundamental neural processes, Binder features should be capable of capturing the core semantic components across different languages[1]. However, Binder-style norms were available only for the English language, with Chinese datasets adopting the same framework now emerging (Wang et al., 2023; Qiu et al., 2023).

Semantic features have also been used in the field of computational linguistics, but their limitations in terms of coverage of the lexicon of natural languages led to the adoption of a different approach to meaning representation. **Distributional Semantic Models** (DSMs) and their successors, **word embeddings**, are data-driven representations aiming at modeling the semantic content of a word through vector derived from statistical distributions in a large corpus of texts (Turney & Pantel, 2010; Baroni et al., 2014). Traditional models assigned each word a single vector that summa-

[1] This hypothesis is further supported by the recent finding (De Varda et al., 2025) that multilingual models can be trained to predict brain activity (fMRI scans) in a set of languages and make efficient predictions in unseen languages in a zero-shot setting scenario.

rized its entire distributional history, disregarding contextual variations in meaning, i.e., the same word may assume different meanings in specific contexts (Apidianaki, 2023). However, modern language model architectures introduced a groundbreaking innovation, **contextualized word representations** (Peters et al., 2018; Devlin et al., 2019), enabled models to generate specific word embeddings for words depending on the sentence context they appear. The Transformer architecture (Vaswani et al., 2017) was widely adopted by the Natural Language Processing (NLP) community as a standard for language models, and research in multilingual training showed strong cross-lingual transfer capacities in models simultaneously pre-trained on multiple languages (Devlin et al., 2019; Karthikeyan et al., 2019; Pires et al., 2019; Conneau et al., 2020).

While word embeddings are a data-driven, automatically-induced representation and, in this sense, and their features are not the product of a subjective choice, they are not interpretable, because it is not immediately evident which shades of meaning are encoded into the vectors' dimensions. This motivated the flourishing of the literature on *probing tasks* to inspect the information contained in the models' representations indirectly, by analyzing their performance in simple classification tasks targeting some specific types of linguistic knowledge [see Linzen et al. 2016; Tenney et al. 2019; Hewitt & Liang 2019; Liu et al. 2019; Wu et al. 2020; Chersoni et al. 2021; Geiger et al. 2021; Kauf et al. 2023, *inter alia*]. Another approach to the interpretability of the embeddings consists instead in mapping their dimensions onto human-readable features, such as semantic norms (Fagarasan et al., 2015; Utsumi, 2020; Chersoni et al., 2021). In this case, the dimensions of the embedding of a given word are provided as input features for a regressor, with the goal of predicting the norm ratings for the same word that were elicited from humans in a psycholinguistic dataset.

To our knowledge, most work predicting semantic norms on the basis of word embeddings operated in monolingual settings. A key research question remains whether embeddings in one language can be used to model semantic feature norms in another language. Binder-style norms, given their claim of cognitive grounding, are excellent candidates for this type of experiment because features that are based on neural processing systems should have higher cross-lingual validity than traditional semantic primitives. Moreover, the task of predicting norms for one language by means of training on the embeddings and human ratings for another language could be highly significant for psycholinguistics, as this would pave the way to the automatic collection of norm databases in low-resource languages.

In this paper, we make the following contributions:

- **Binder-zh 2.0**: We introduce an updated version of the Chinese Binder-style norms, featuring a larger number of native speakers' ratings per word. Additionally, we provide a detailed statistical analysis of the collected data.
- **Monolingual and cross-lingual norm prediction**: We conduct a comparative evaluation of several monolingual and cross-lingual language models on the task of norm prediction. This includes monolingual prediction in English and Chinese and a zero-shot cross-lingual setting, where we train on one language and test on the other one. Our results reveal that: (a) in both languages, norms can be reliably

predicted by both word embedding projection – particularly with bidirectional models such as BERT – and prompting with Large Language Models, where GPT-4 achieving the highest accuracy in this latter case. (b) In the cross-lingual setting, correlations with human ratings are significant, though they vary widely across features; for example, visual features tend to be predicted with higher accuracy, while spatial and temporal features present greater challenging.

Codes and norm data are available at https://anonymous.4open.science/r/norms_correlation-F185.

## 2 Related work

Since neural word embeddings replaced traditional DSMs in NLP, researchers have debated about the 'black box' nature of these representations. On one hand, dense vector models like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) showed robust performance and were easier to train and integrate in downstream applications. On the other hand, their dimensions were not interpretable, because they no longer correspond to specific word co-occurrence contexts, unlike the sparse spaces of DSMs (Baroni et al., 2014).

A line of research attempted to ground vector representations in perceptual data, usually using regression algorithms or neural networks to map them onto interpretable features such as semantic (Fagarasan et al., 2015; Bulat et al., 2016; Li & Summers-Stay, 2019), modality (Chersoni et al., 2020) and concreteness norms (Thompson & Lupyan, 2018; Flor, 2024). Notably, two studies, Utsumi (2020) and Chersoni et al. (2021), specifically used the Binder norms as the feature set for the projection; the latter study was the first to employ contextualized vector representations from language models in this task (i.e., vectors extracted from ELMo and BERT), showing that they perform comparably or better than the best static embedding models.

It should be noticed that the primary goal of this research trend has always been assessing the interpretability of the embeddings, although the mapping process itself has interesting applications, as it has automatically predicted norms for words unseen in the training data as a byproduct. In light of this, when the dominant paradigm in NLP shifted towards Large Language Models (LLMs) (Brown et al., 2020; Achiam et al., 2023; Touvron et al., 2023) that could be queried via natural language instructions (the so-called *prompts*), researchers started to explore the possibility of eliciting semantic norms with the new technology.

Xu et al. (2023) used the words of the Glasgow (Scott et al., 2019) and of the Lancaster norms (Lynott et al., 2020) to prompt ChatGPT 3.5 and 4 and verify to what extent LLMs can predict the original human ratings. They found that LLM predictions are well-aligned with humans in non-sensorimotor, and abstract domains, but they struggle with sensorimotor concepts; however, GPT-4 achieved better results than the previous model in some of the vision-related areas, probably due to its additional training with images.

On a similar topic, Trott (2024) investigated whether GPT-4 can predict human ratings for a large variety of psycholinguistic datasets, including sensorimotor norms for

words in context (Trott & Bergen, 2022) (previous studies rather focused on words in isolation). The answers of GPT-4 were found to be positively correlated with human judgements, and the correlation scores often exceeded their average inter-annotator agreement. However, the author also observed that GPT-4 is better at predicting the properties of abstract rather than concrete words: the correlations with ratings for sensory domains were significant, but noticeably lower. This study raised the issue of *data leakage*, that is, the risk that the LLM might have seen the gold standard scores during pretraining. As a possible solution, the author suggested to evaluate LLMs on datasets published after the cutoff date for the collection of the training data of the LLMs.

Most recently, Martínez et al. (2024a) used ChatGPT-4o to predict ratings of concreteness, valence, arousal and dominance norms for a large set of words and multiword expressions. The results showed high correlation values with humans, although no semantic norms were included in the testing materials. A later extension of this study for the same variables (Martínez et al., 2024b) reported that the results generalize to Spanish, paving the way for the LLM-based acquisition of norms in languages other than English.

Compared to previous studies, our work focused on predicting semantic norms in an English-Chinese bilingual setting[2]. Given the availability of parallel datasets in both languages, we adopted the cognitively-motivated features proposed by Binder et al. (2016). Our primary goal is to investigate whether semantic norms in one language can be used to predict norms in another language in a full zero-shot setting.

We want to stress again the fact that, if the mapping between different languages such as English and Chinese works well, the result would be very meaningful for the acquisition of norms in low-resource languages. Works on the 'curse of multilinguality' (Conneau et al., 2020; Chang et al., 2023) showed that linguistic similarity between languages in the training data of multilingual models is a primary factor affecting language modeling performance, and thus the results of mapping between more similar languages would be likely to yield even better results.

## 3 The dataset: binder-zh 2.0

The work of Binder et al. (2016) aimed to develop a feature-based representation that captures experiential aspects central to concept acquisition. They proposed a framework that organizes human experience into 65 cognitively motivated features (e.g., *Dark*, *Shape*, *Social*), grouped into 14 domains (e.g., *Vision*, *Somatic*, *Audition*). Each domain corresponds to a set of features specialized for neural processors which have been identified in the neuroscience literature.

The Binder-zh 2.0 norms dataset comprises 535 Chinese words translated from the original English words, including 242 words from the *Knowledge Representation in*

[2] There have been previous studies aiming at predicting psycholinguistic variables in a crosslingual fashion, but they were limited to simpler types of norms, with a much more limited number of target variables, e.g. concreteness (Thompson & Lupyan, 2018) and sensory norms (Chersoni et al., 2020). To our knowledge, our study is the first one focusing on semantic norms.

*Neural Systems project* (Glasgow et al., 2016) (141 nouns, 62 verbs, and 39 adjectives) and 293 additional abstract nouns. Unlike the published English Binder norms dataset,[3] which includes 65 features per word, our Binder-zh 2.0 dataset, following an earlier version of annotation released in (Binder et al., 2016) containing 68 features per word. This is a result from splitting the feature *Temperature* into *Hot* and *Cold*, *Texture* into *Smooth* and *Rough*, and *Weight* into *Light* and *Heavy*. In addition to the feature set, the original English survey queries were adopted and translated into simplified Mandarin for human annotation. This translation process has been established by Qiu et al. (Qiu et al., 2023), and we developed the current data set following the same design and further expanding the dataset to ensure each entry was rated by at least 30 native speakers. The specific adaptation details are outlined below.

### 3.1 Stimuli

The translation was conducted by two native Mandarin speakers, both Master's students in linguistics. For both features and target words, the most common and core sense of each English term was translated. While some words had multiple colloquial senses, more specific translations were chosen to match the frequency of their polysemous English counterparts. For example, *football* can be translated as either 足球 (soccer) or 美式足球 (American football). To disambiguate, 美式足球 was selected for *football*, and 足球 for *soccer*. Additionally, since the concept of "adjective" can be less clear in Chinese, an optional adjectival suffix -的 (*-de*) was added to disambiguate cases where English past participles could function as either adjectives or verbs. This is a necessary step also to ensure good prediction quality in our computational modeling experiments, as multilingual models have been shown to struggle with the representation of ambiguous words (Rivière et al., 2024).

The final survey and target words were manually reviewed by one of the authors, a native Mandarin speaker. All 535 words have their part of speech (POS) matched with their English counter parts, with corresponding survey questions addressing the 68 cognitively motivated features. One culturally specific adaptation was made for the target word *banjo*, which was replaced with the culturally relevant Chinese instrument 二胡 (*erhu*), while the remaining words were consistent with their English counterparts.

### 3.2 Data collection

Following the Binder norms, we employed a continuous rating design to measure feature relevance for each word, using the same 0-6 Likert scale as in the original study. A score of 6 indicated high relevance of a feature to a target word, while 0 indicated irrelevance. Data were collected via the crowdsourcing platform 问卷星 (*Wenjuanxing*),[4] widely used in China. To mitigate potential subjectivity arising from personal experiences and backgrounds, a larger sample size was prioritized. In total, 16,342 rating sessions were collected, with each of the 535 target words receiv-

[3] https://www.neuro.mcw.edu/index.php/resources/brain-based-semantic-representations/

[4] https://www.wjx.cn/

ing 30-35 sets of ratings per 68 features. Participants' demographics and language backgrounds were verified before they took the survey, and each participant was compensated with RMB ¥20 upon successful completion (conditional on passing the standard attention checks).

### 3.3 Inter-annotator agreement analysis

We evaluated the reliability of our collected ratings through multiple complementary measures. The attribute-based Fleiss' kappa yielded an average of 0.097, indicating slight agreement when considering each feature independently, while the word-oriented ICC showed an average of 0.211, suggesting fair consistency in absolute ratings. Profile-based agreement revealed more substantial consensus: while correlation averaged 0.192, cosine similarity provided a robust 0.548, indicating that annotators showed moderate agreement on the relative importance of features even when their absolute ratings differed.

Notably, concrete concepts with well-defined perceptual characteristics demonstrated significantly higher agreement across all metrics. For instance, "黄瓜" (cucumber) achieved a profile correlation of 0.365 and cosine similarity of 0.631, while abstract concepts like "主题" (theme) showed much lower agreement (correlation: 0.036, cosine: 0.283). This pattern supports previous psycholinguistic findings that concrete concepts elicit more consistent feature representations across individuals than abstract concepts, which typically show greater variability McRae et al. (2005); Katja Wiemer-Hastings and Xu (2005). The higher cosine similarities relative to correlations across our dataset suggest that annotators generally agree on which features are relevant, even when they differ in their judgment of absolute feature strengths.

### 3.4 Cross-lingual feature section and correlation reports

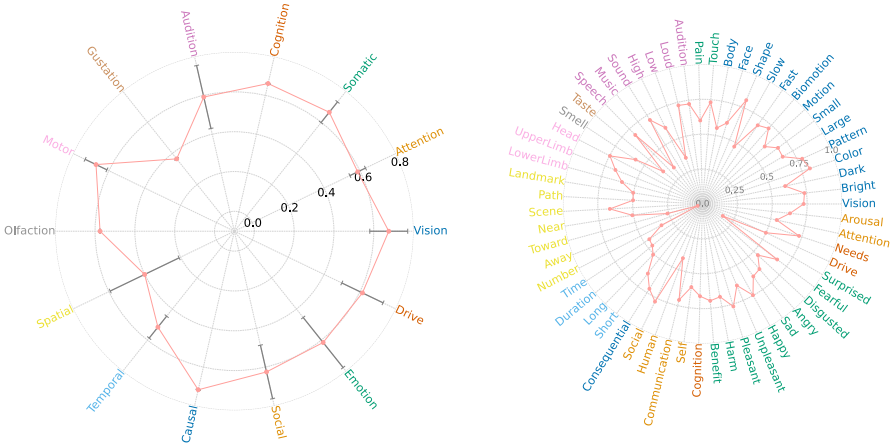
In this study, one of our objectives is to assess the predictive capabilities of language models in a cross-lingual setting. Therefore, we focus only on the 59 common features shared by both the English and Chinese datasets for our experiments, excluding those features that lack ratings for verbs and adjectives in the English Binder norms dataset; these features are *Temperature*, *Hot*, *Cold*, *Texture*, *Smooth*, *Rough*, *Weight*, *Light*, *Heavy*, *Complexity*, *Practice*, and *Caused*.

The meanings of the 59 features and their corresponding domains are listed in Table 1. Analysis of the annotated ratings in English and Chinese yield an average Spearman correlation coefficient ($\rho$) of 0.596 and a Pearson correlation coefficient ($r$) of 0.596 across all common features, indicating a moderately high level of agreement.

The domain-level and feature-level Spearman correlation scores are shown in Fig. 1. Notably, high cross-lingual agreement ($\rho$ >0.75) was observed for visual related features such as *Shape*, *Color*, and *Pattern*, and features like *Social* and *Pleasant*. However, lower agreement was found for features, such as *Away* and *Surprised*, where $\rho$ values were close to 0. Overall, visual features exhibited relatively high cross-linguistic correlations, as did features in the domains of *Causal*, *Social*, *Cogni-*

**Table 1** List of domains and meaning components (features) in Binder et al. (2016)

| Domain | Meaning components (features) |
|---|---|
| Vision | Vision, Bright, Dark, Colour, Pattern, Large, Small, Motion, Biomotion, Fast, Slow, Shape, Complexity, Face, Body |
| Somatic | Touch, Hot, Cold, Smooth, Rough, Light, Heavy, Pain |
| Audition | Audition, Loud, Low, High, Sound, Music, Speech |
| Gustation | Taste |
| Olfaction | Smell |
| Motor | Head,Upper Limb, Lower Limb, Practice |
| Spatial | Landmark, Path, Scene, Near, Toward, Away, Number |
| Temporal | Time, Duration, Long, Short |
| Causal | Caused, Consequential |
| Social | Social, Human, Communication, Self |
| Cognition | Cognition |
| Emotion | Benefit, Harm, Pleasant, Unpleasant,Happy, Sad, Angry, Disgusted, Fearful, Surprised |
| Drive | Drive, Needs |
| Attention | Attention, Arousal |




(a) Domain level (b) Feature level

**Fig. 1** Spearman correlation coefficient ($\rho$) between English and Chinese human-annotated feature ratings

*tion*, and *Emotion*. In contrast, the lowest correlation values were observed for spatio-temporal features.

## 4 Methodology

In the following, we present two experiments, the first uses embedding-based regression, and the second uses prompting with large language models.

Our first approach to cross-lingual norm prediction consists of using word embeddings for both Chinese and English words from shared pretrained language models (PLMs). To account for contextual variability, we apply position-wise mean pooling to a set of contextual representations for each target word, generating its embedding. Formally, for a target word $w_i$, the word embedding $e\left(w_i\right)$ is calculated as:

$$e\left(w_i\right) = \frac{1}{N} \sum_{j=1}^{N} e_j\left(w_i\right) \tag{1}$$

where $e_j\left(w_i\right)$ denotes the contextual representation of the word $w_i$ in sentence $S_j \in \{S_1, S_2, ..., S_N\}$. These contextual embeddings are extracted from the last hidden states of the PLM, indexed by the word's position. If a target word is tokenized into multiple subword tokens, its embedding is derived by averaging the hidden states of those tokens.

The generated word embeddings serve as input features for predicting norms ratings. We first train a regressor on either English or Chinese data, using the word embeddings as input variables and the human-annotated ratings as target responses. To evaluate cross-lingual performance in a zero-shot setting, we then use the embeddings from the other language (Chinese or English) as input to the regressor trained on the first language.

Prior research has shown that pretraining multilingual language models enhances performance across a variety of cross-lingual transfer tasks Conneau and Lample (2019); Conneau et al. (2020). This experiment aims to determine whether such models also exhibit strong cross-lingual validity for predicting cognitive neural features.

In our second experiment, we explore prompting LLMs to annotate ratings for target words. We conduct prompting experiments using three LLMs: GPT-4o (Achiam et al., 2023), Llama3.1-8B-Instruct Grattafiori et al. (2024), and Qwen2.5-7B-InstructHui et al. (2024). Since the original Llama3.1-8B model does not support Chinese, we use Llama3.1-8B-Chinese-Chat,[5] a version fine-tuned specifically for Chinese. Both Llama3.1-8B and Qwen2.5-7B are loaded in brain floating point (bfloat16) format and run locally, while GPT-4o is accessed via its online interface on August 6th, 2024. Further details on the prompt design are provided in Section 5.3.

[5] https://huggingface.co/shenzhi-wang/Llama3.1-8B-Chinese-Chat

## 5 Implementation details

### 5.1 Word embedding acquisition

Word embeddings are obtained by applying average pooling to a set of last hidden state tensors encoded by pretrained language models (PLMs) using various input sentences containing the target word. Concretely, we retrieve up to 1 million sentences for each Chinese and English word from the May 2024 dumps of Wikipedia in the respective language and then randomly select up to 500 sentences containing the target word. During processing, we observe that sometimes the number of sentences for a target word is lower than 500: most English words can be paired with 500 sentences, with only a few exceptions (e.g. the word "Joviality" has the lowest sentence frequency of 10); on the other hand, the frequency distribution for Chinese words is more variable, with more than half of the words occurring in a number of sentences between 250 to 500, and the remainder occurring in less than 250 sentences. The minimal sentence frequency is 2, observed for the words “一角钱-dime", “布满灰尘的-dusty", “发刷-hairbrush", “午后场-matinee", and “尖锐声-screech." However, it is worth noting that previous research has shown that out-of-context contextualized embeddings can be effectively constructed using a relatively small number of token embeddings (ranging from 10 to 100), as performance improvements from sampling a higher number of contexts are negligible Vulić et al. (2020); Lenci et al. (2022). The distribution of the number of sentences is shown in Fig. 2.

For a comprehensive assessment, we conducted experiments on contextual word embeddings extracted from both masked language models (MLMs) and causal language models (CLMs). We selected embedding models from the BERT Devlin et al. (2019) and GPT-2 Radford et al. (2019) series, as these are foundational and representative of their respective model architectures. To ensure comparability, we used models with similar backbones and parameter sizes whenever feasible. Table 2 lists model names, references and HuggingFace IDs[6] for the models used in our experiments. Multilingual models were employed to extract word embeddings for both Chinese and English, while monolingual models were used to obtain embeddings

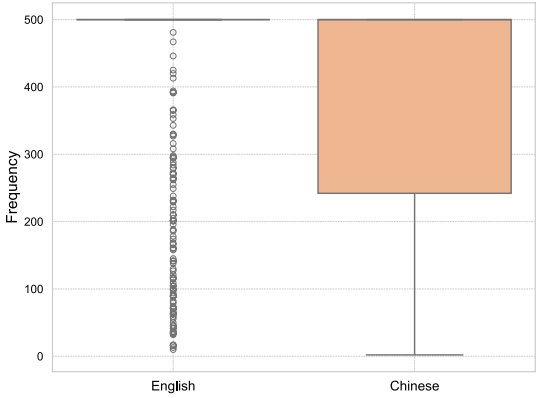



**Fig. 2** The frequency distribution of selected English and Chinese sentences

[6] https://huggingface.co/models

**Table 2** PLMs used in the experiments

| Type | Name | ID |
|---|---|---|
| CLM | mGPT(Tan et al., 2022) | THUMT/mGPT |
| | $GPT2_{EN}$ (Radford et al., 2019) | openai-community/gpt2 |
| | $GPT2_{ZH}$ (Zhao et al., 2019) | uer/gpt2-chinese-cluecorpussmall |
| MLM | mBERT(Devlin et al., 2019) | google-bert/bert-base-multilingual-cased |
| | $BERT_{EN}$ (Devlin et al., 2019) | google-bert/bert-base-cased |
| | $BERT_{ZH}$(Devlin et al., 2019) | google-bert/bert-base-chinese |

exclusively for either Chinese or English. Additionally, we included randomly initialized word embeddings as reference baselines, with their dimension set to 768 (same as BERT).

Additionally, we train fastText Bojanowski et al. (2017) embeddings using the Skip-gram architecture on the same Wikipedia corpora. For Chinese text, we preprocess the raw sentences by first segmenting words using Jieba.[7] For English text, we use standard tokenization (splitting on whitespace and punctuation). The Skip-gram model is trained with a dimensionality of 300, a context window size of 5, and 5 negative sampling iterations. For Chinese, subword units are extracted using character n-grams (range: 3–6) to handle out-of-vocabulary terms and morphological variations, while English embeddings rely on word-level n-grams.

To enable cross-lingual analysis, we align the monolingual Chinese and English fastText embeddings into a shared vector space using the MUSE Lample et al. (2017) framework. We use the publicly available Chinese-English dictionary from MUSE for this process, ensuring compatibility with standard cross-lingual benchmarks. The resulting model henceforth will be referred to as xfastText (cross-lingual fastText).

### 5.2 Regression setting

The extracted embeddings served as independent variables of a Ridge regression algorithm, which was chosen for its capability of handling high collinearity between predictors (we expect this to be the case with word embedding dimensions) Hoerl and Kennard (2000). Using the embeddings as our set of independent variables, we predict the 59 Binder features as dependent variables in a multivariate regression setting.

We evaluated the multilingual embedding models in a cross-lingual scenario, where the regressor was trained solely on English or Chinese data and then tested on the other language. Concurrently, both multilingual and monolingual embedding models were assessed in a monolingual setting, where the regressor was trained and tested within the same language using 5-fold cross-validation.

We recorded the average values for mean squared error (MSE) and Spearman's correlation coefficient ($\rho$) respectively across all the features and across all the words, and we presented the results in Table 4.

[7] https://github.com/fxsjy/jieba

### 5.3 Prompt design

Figure 3 illustrates the templates for English and Chinese prompts. Drawing on the survey design from the original study (Binder et al., 2016), the prompts are structured to assess the relevance of target words to conceptual features, as evaluated by LLMs. The verbal relation connects the target word to the item-specific query. Table 3 provides examples of these queries and their Chinese translations for different grammatical classes and features. Additionally, each prompt includes a pair of high-value and medium-value examples, along with their corresponding explanations. A high-value example represents a concept that would receive a high rating, while a medium one corresponds to a concept with a moderate rating. The English examples are available in the research materials[8] provided by Binder et al. (2016). The Chinese examples and explanations were translated from the English versions.

## 6 Results and discussion

### 6.1 Norm prediction with embedding models

**Overall performance**. As shown in Table 4, the correlation scores for random baseline embeddings indicate that even randomly initialized word embeddings can exhibit weak correlations with Binder feature vectors at the word level. However, these correlations are much lower and typically not significant when analyzed at the feature level.

Among the embedding models, MLMs consistently outperform CLMs in norm prediction tasks. BERT achieves the highest correlation scores and lowest MSE for English norm prediction, while mBERT performs best for Chinese norms. Conversely, GPT-based models show weaker performance for both languages, likely due to differences in attention mechanisms. Specifically, MLMs, with their bidirectional

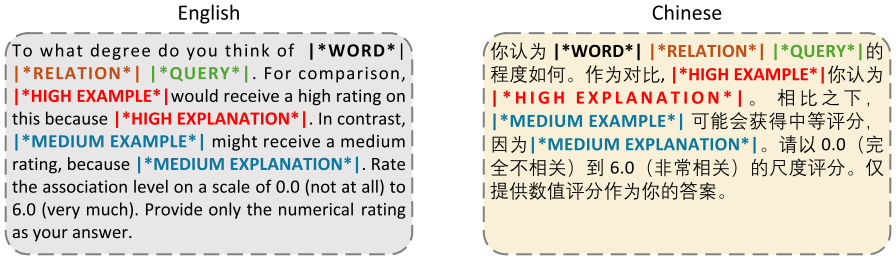



**Fig. 3** English and Chinese prompt templates. The placeholders |*WORD*|, |*RELATION*|, and |*QUERY*|were replaced with the target word, verbal relation, and general query stem, respectively. *Note*: The relation and query are selected based on the POS of target word. The placeholders |*HIGH EXAMPLE*|, |*MEDIUM EXAMPLE*|, |*HIGH EXPLANATION*|, and |*MEDIUM EXPLANATION*|correspond to the examples of high or medium relevance of features and their associated explanations

[8] https://www.neuro.mcw.edu/index.php/resources/brain-based-semantic-representations/

**Table 3** Examples of queries and their corresponding Chinese translations

| Feature | POS | \|*RELATION*\| | \|*QUERY*\| | Chinese Translation |
|---|---|---|---|---|
| Color | noun | Having | A characteristic or defining color? | 具有某种特征或标志性颜色？ |
| | verb | Being Associated with | Color or change in color? | 与颜色或色变化相关？ |
| | adj | Describing | A quality or type of color? | 描述颜色的量或类型？ |
| Lower Limb | noun | Being Associated with | Actions using the leg or foot? | 与使用腿或脚的动作相关？ |
| | verb | Being | An action/activity using the leg or foot? | 用腿或脚进行的动作或活动？ |
| | adj | Being Related to | Actions of the leg or foot? | 与腿或脚的动作相关？ |
| Fearful | noun | Being | Someone/something causing fear? | 是让你感到害怕的人或事？ |
| | verb | Being Associated with | Feeling afraid? | 与感到害怕相关？ |
| | adj | Being Related to | Feeling afraid? | 与感到害怕相关？ |

attention, can capture both prefix and suffix contexts, producing richer contextual embeddings. In contrast, CLMs, which rely on left-to-right attention, are limited to prefix information, resulting in less comprehensive semantic representations. This discrepancy aligns with prior findings that contextual embeddings from CLMs tend to underperform in semantic tasks (Springer et al., 2024; BehnamGhader et al., 2024). The static embedding models put together a solid performance: they are generally worse than the best contextualized models (BERT or mBERT), but they are much better than the baseline and on par or better than the other models; it is also interesting to notice that the cross-lingual mapping from Chinese to English even works slightly better with the static vectors.

In supervised monolingual settings, monolingual models (BERT, GPT, and fastText) outperform their multilingual (mBERT and mGPT) and cross-lingual aligned (xfastText) counterparts for English norm prediction. However, the trend reverses for Chinese norms, where multilingual models (mBERT and mGPT) and cross-lingual aligned embedding achieve slightly higher correlation scores and lower MSE than their monolingual counterparts. This result suggests that multilingual pretraining (as in mBERT), multilingual fine-tuning (as in mGPT), and cross-lingual alignment (as in xfastText) may compromise performance on English while offering marginal benefits for Chinese.

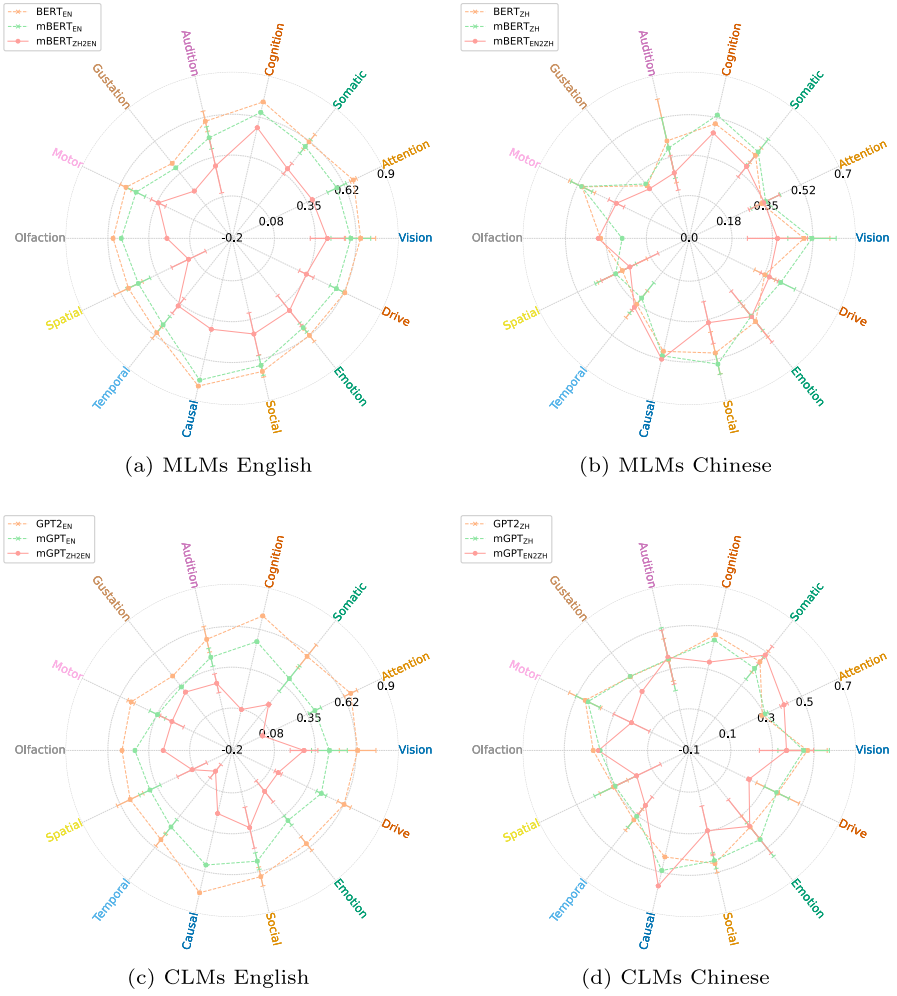
**Table 4** Overall results for the regression task. For each language and metric, the best average score has been highlighted in **bold**. The scores for monolingual prediction of English norms are shown in the upper section, the ones for monolingual prediction of Chinese norms are in the middle section, while the cross-lingual prediction scores are in the lower section

| | Feature | | Word | |
|---|---|---|---|---|
| | MSE | $\rho$ | MSE | $\rho$ |
| $\text{Random}_{EN}$ | 2.168 ± 0.923 | − 0.006 ± 0.042 | 2.168 ± 0.776 | 0.322 ± 0.212 |
| $\text{fastText}_{EN}$ | 1.223 ± 0.482 | 0.591 ± 0.094 | 1.223 ± 0.611 | 0.662 ± 0.160 |
| $\text{xfastText}_{EN}$ | 1.132 ± 0.532 | 0.573 ± 0.108 | 1.123 ± 0.587 | 0.646 ± 0.127 |
| $\text{BERT}_{EN}$ | **1.048 ± 0.358** | **0.622 ± 0.090** | **1.047 ± 0.496** | **0.691 ± 0.148** |
| $\text{mBERT}_{EN}$ | 1.367 ± 0.489 | 0.551 ± 0.105 | 1.367 ± 0.655 | 0.627 ± 0.169 |
| $\text{GPT2}_{EN}$ | 1.164 ± 0.400 | 0.594 ± 0.100 | 1.164 ± 0.641 | 0.672 ± 0.156 |
| $\text{mGPT2}_{EN}$ | 2.975 ± 1.140 | 0.431 ± 0.092 | 2.975 ± 2.209 | 0.532 ± 0.224 |
| $\text{Random}_{ZH}$ | 0.733 ± 0.257 | −0.018 ± 0.055 | 0.733 ± 0.428 | 0.417 ± 0.207 |
| $\text{fastText}_{ZH}$ | 0.713 ± 0.191 | 0.406 ± 0.104 | 0.713 ± 0.331 | 0.591 ± 0.186 |
| $\text{xfastText}_{ZH}$ | 0.697 ± 0.179 | 0.410 ± 0.114 | 0.697 ± 0.309 | 0.600 ± 0.176 |
| $\text{BERT}_{ZH}$ | 0.745 ± 0.167 | 0.426 ± 0.123 | 0.745 ± 0.354 | 0.608 ± 0.173 |
| $\text{mBERT}_{ZH}$ | **0.653 ± 0.155** | **0.434 ± 0.118** | **0.653 ± 0.308** | **0.628 ± 0.180** |
| $\text{GPT2}_{ZH}$ | 0.856 ± 0.195 | 0.392 ± 0.102 | 0.856 ± 0.486 | 0.571 ± 0.199 |
| $\text{mGPT2}_{ZH}$ | 0.792 ± 0.190 | 0.396 ± 0.122 | 0.792 ± 0.414 | 0.591 ± 0.171 |
| $\text{fastText}_{ZH2EN}$ | 3.773 ± 1.971 | 0.369 ± 0.109 | 3.773 ± 0.970 | 0.530 ± 0.291 |
| $\text{mBERT}_{ZH2EN}$ | 3.773 ± 1.546 | 0.354 ± 0.158 | 3.773 ± 1.098 | 0.399 ± 0.204 |
| $\text{mGPT}_{ZH2EN}$ | 6.145 ± 4.660 | 0.190 ± 0.131 | 6.145 ± 2.111 | 0.335 ± 0.176 |
| $\text{fastText}_{EN2ZH}$ | 4.931 ± 1.937 | 0.340 ± 0.154 | 4.931 ± 2.419 | 0.359 ± 0.165 |
| $\text{mBERT}_{EN2ZH}$ | 3.642 ± 2.060 | 0.358 ± 0.118 | 3.642 ± 1.063 | 0.471 ± 0.171 |
| $\text{mGPT}_{EN2ZH}$ | 18.880 ± 16.236 | 0.323 ± 0.148 | 18.880 ± 4.837 | 0.244 ± 0.189 |

Additionally, multilingual models demonstrate some cross-lingual generalization. Both mBERT and mGPT achieve higher correlations than random embeddings in zero-shot cross-lingual settings; mBERT shows greater consistency in capturing language-independent conceptual features. However, mGPT struggles with cross-lingual generalization, showing lower correlations and higher MSE likely reflecting its English-centric pretraining and subsequent adaptation to other languages Tan et al. (2022).

**Domain-level Performance**. The radar plots in Fig. 4 illustrate the average Spearman's correlation coefficients between predicted and human-annotated values for MLMs (subplots 4a and 4b) and CLMs (subplots 4c and 4d) across 14 feature domains in both supervised and zero-shot unsupervised cross-lingual prediction settings.

In supervised settings, monolingual models generally outperform their multilingual counterparts in terms of consistency and higher correlation scores for English norm predictions. For example, BERT and GPT2 perform better than mBERT and mGPT, reflecting the advantage of language-specific pre-training. However, when predicting Chinese norms, multilingual models show competitive or even superior performance in certain domains. For instance, mBERT outperforms BERT in vision-




(a) MLMs English (b) MLMs Chinese

(c) CLMs English (d) CLMs Chinese

**Fig. 4** The average Spearman correlation coefficient ($\rho$) and standard deviation for each attribute predicted by MLMs and CLMs in both mono- and cross-lingual norm prediction. The solid lines represent values for cross-lingual prediction, while the dashed lines with light colors represent values for monolingual prediction

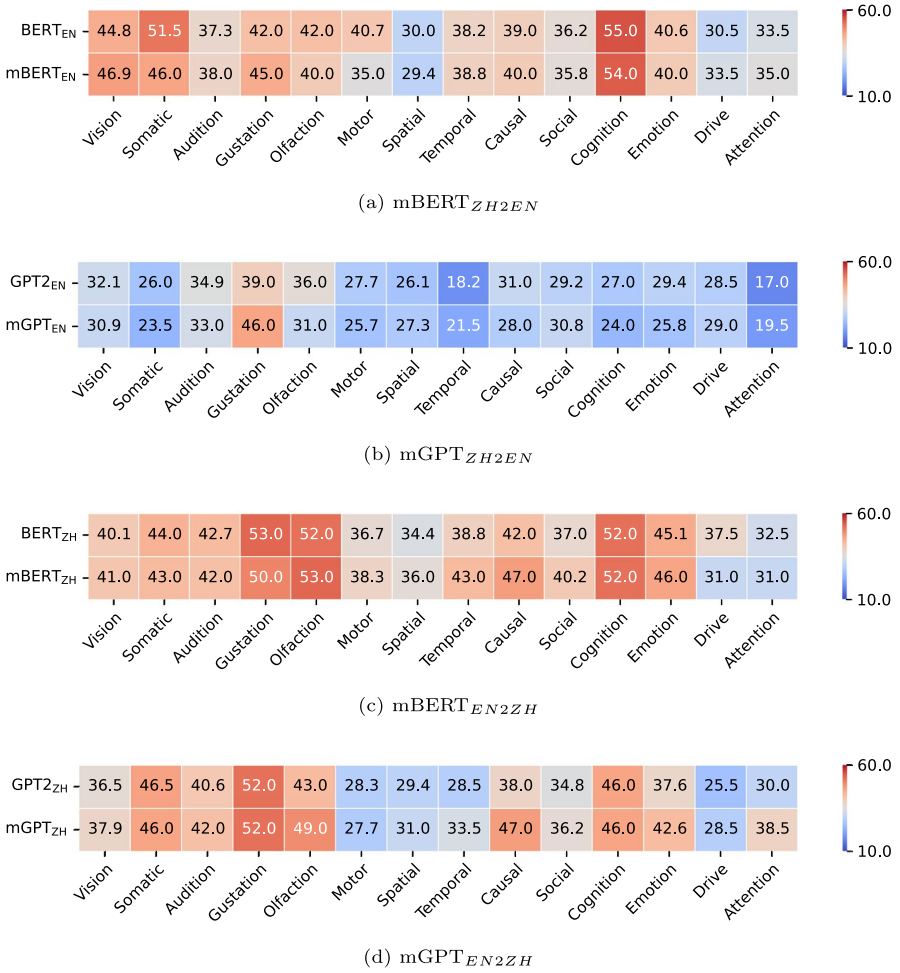
related features, while mGPT surpasses GPT2 in emotion-related features. This suggests that multilingual training provides additional benefits in capturing brain-based semantic features for Chinese. Both MLMs and CLMs show relatively consistent trends for features such as “Vision" and “Cognition", but performance varies for a lot across other features (e.g., the Causal ones).

In the unsupervised cross-lingual setting, performance fluctuates significantly across domains. For instance, when using Chinese embeddings to predict English norms, correlation scores exhibit considerable fluctuations in the “Spatial" and “Social" domains for both mBERT and mGPT, with mBERT’s standard deviation reaching zero in the "Spatial" domain. Similarly, when using English embeddings to predict Chinese norms, fluctuations appear in the “Spatial" domain for both models. Emotion-related domains also exhibit large standard deviations, indicating unstable performance on certain emotion features. Furthermore, “Olfaction" and “Motor" domains consistently show lower correlation scores, suggesting that these features are heavily influenced by language-specific contexts. In contrast, “Vision", “Cognition", and “Attention" domains yield higher and more consistent scores, reflecting their more universal conceptual nature. These findings highlight the challenges of cross-lingual alignment for complex or culturally specific features and underscore the trade-off between multilingual generalization and language-specific optimization.

Figure 5 displays the mean number of common top-100 words with the highest $\rho$ scores between zero-shot cross-lingual predictions and supervised monolingual predictions for each domain, aggregated across the features related to that domain. A higher number indicates greater agreement, with many common words among the top ones for a given feature group.

When using Chinese embeddings to predict English norms (subplots 5a and 5b), mBERT shows a greater number of common words across most domains compared to mGPT, which aligns with the domain-level correlation scores in subplots 4a and 4c. Specifically, $\mathrm{mBERT}_{ZH2EN}$ achieves a high number of common words in domains such as “Vision," “Somatic," “Gustation," and “Cognition" with both $\mathrm{BERT}_{EN}$ and $\mathrm{mBERT}_{EN}$, indicating strong alignment between monolingual supervised predictions and cross-lingual zero-shot predictions in these domains. In contrast, $\mathrm{mGPT}_{ZH2EN}$ and $\mathrm{mGPT}_{EN}$ show the highest number of common words in the "Gustation" domain, while the number of common words remains below 40 in most other domains.

When using English embeddings to predict Chinese norms (subplots 5c and 5d), the number of common words between $\mathrm{mBERT}_{EN2ZH}$ and $\mathrm{BERT}_{ZH}$, as well as $\mathrm{mBERT}_{EN2ZH}$ and $\mathrm{mBERT}_{ZH}$, is generally similar, with $\mathrm{mBERT}_{EN2ZH}$ slightly outperforming $\mathrm{BERT}_{ZH}$ in domains like "Emotion" (47.0 vs. 42.0). Similarly, in subplot 5d ($\mathrm{mGPT}_{EN2ZH}$), $\mathrm{GPT2}_{ZH}$ (monolingual supervised) and $\mathrm{mGPT}_{ZH}$ (multilingual supervised) exhibit comparable performance, with $\mathrm{mGPT}_{ZH}$ achieving higher numbers in domains like “Emotion" (47.0 vs. 38.0). The highest numbers are observed in “Vision" (53.0 for $\mathrm{BERT}_{ZH}$ and 52.0 for $\mathrm{GPT2}_{ZH}$) and “Audition" (52.0 for both), indicating strong alignment between cross-lingual zero-shot predictions and supervised predictions in these domains. This suggests that multilingual supervised predictions align well with cross-lingual zero-shot predictions for Chinese, particularly in domains like “Vision" and “Audition."




(a) mBERT$_{ZH2EN}$

(b) mGPT$_{ZH2EN}$

(c) mBERT$_{EN2ZH}$

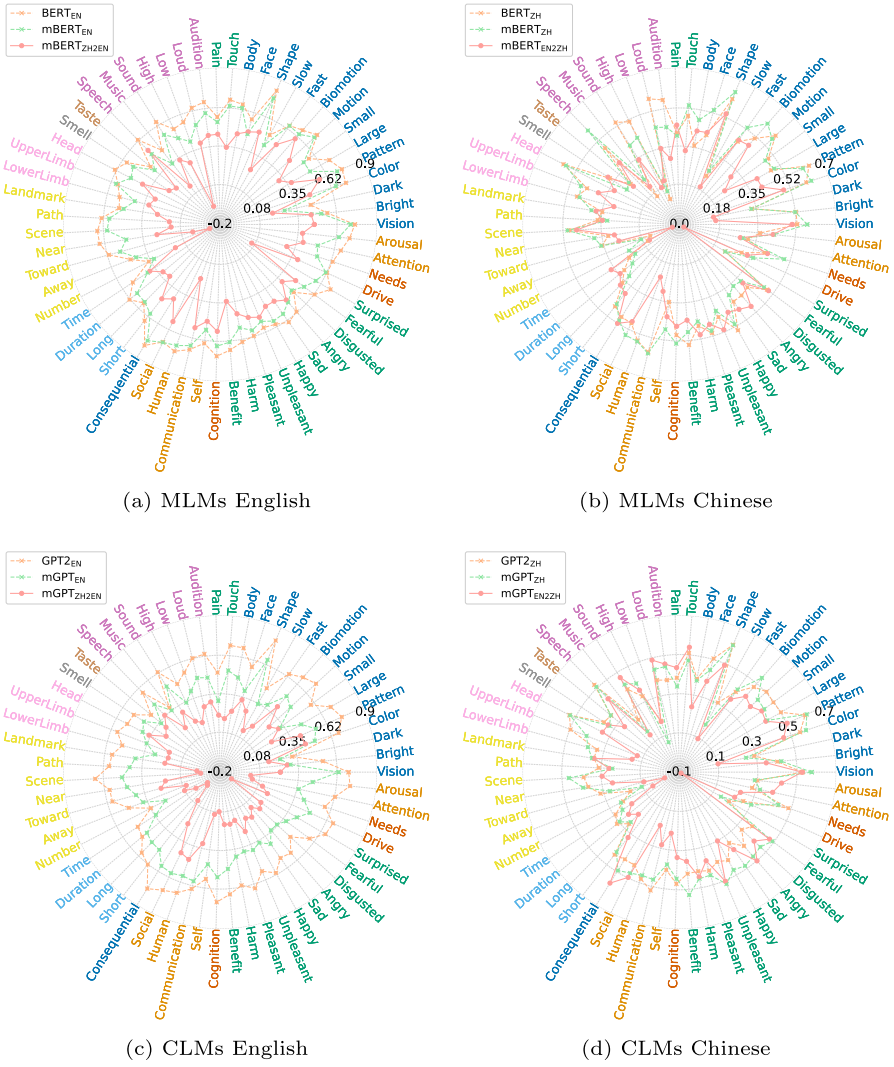
(d) mGPT$_{EN2ZH}$

**Fig. 5** Mean number of top-100 common words for each domain, averaged across semantic features within the domain, comparing multilingual models' zero-shot cross-lingual predictions with mono-/multilingual models' fully supervised predictions. Words are ranked by their Spearman correlation scores between model predictions and human ratings

Overall, the results indicate that sensory domains like "Vision" and "Audition" exhibit strong cross-lingual alignment; other domains (e.g., "Emotion") showing variability seem to be more language-specific.

**Feature-level performance**. Subplots 6a and 6b present radar plots of the average $\rho$ scores between predicted and human-annotated values for MLMs across individual features in English and Chinese, respectively. Similarly, Subplots 6c and 6d display the corresponding correlation scores for CLMs.

In supervised monolingual settings, correlation scores for MLMs and CLMs exhibit less fluctuation when predicting English norms compared to Chinese norms. For English norms, mBERT and mGPT consistently achieve lower correlations than their monolingual counterparts (BERT and GPT), albeit following similar trends. For




(a) MLMs English (b) MLMs Chinese

(c) CLMs English (d) CLMs Chinese

**Fig. 6** The Spearman correlation coefficient ($\rho$) for each feature predicted by mBERT and mGPT in both mono- and cross-lingual norm prediction. The solid lines represent values for cross-lingual prediction, while the dashed lines with light colors represent values for monolingual prediction

Chinese norms, however, the advantages of monolingual models are inconsistent. For instance, mBERT surpasses BERT in vision-related features (e.g., “Body", “Face", and “Shape"), while mGPT achieves higher scores than GPT in emotion-related features (e.g., “Fearful", “Sad", and “Happy"). These results suggest that multilingual training may enhance the ability to model semantic features for Chinese.

In the zero-shot cross-lingual prediction of norms using multilingual models, the correlation scores vary significantly across features. When using Chinese embeddings to predict English norms, mBERT’s correlation scores range from −0.121 for the feature “Away" to 0.598 for “Biomotion", while mGPT’s scores range from −

0.108 for “Supervised" to 0.450 for “Human". Similarly, when using English embeddings to predict Chinese norms, mBERT’s correlation scores range from 0.023 for the feature “Surprised" to 0.536 for “Shape", while mGPT’s scores range from −0.090 for “Supervised" to 0.567 for “Consequential".

Notably, for the “Away" and “Surprised" features, both mBERT and mGPT achieve low correlation scores regardless of whether English embeddings are used to predict Chinese norms or vice versa. This finding aligns with the correlation coefficients between English and Chinese human-annotated ratings shown in Fig. 1.

Figure 7 illustrates the performance of cross-lingual norm prediction under a zero-shot setting, focusing on the “Away" and “Surprised" features across different parts of speech (nouns, verbs, and adjectives). Subplots 7a –7d present Spearman’s $\rho$ scores, while subplots 7e –7h show mean squared error (MSE).

Both mBERT and mGPT demonstrate poor predictive performance on the “Away" and “Surprised" features, regardless of the transfer direction (predicting Chinese norms using English embeddings or vice versa). The correlation scores for these features are particularly low, with noticeable inconsistencies across parts of speech. For instance, $mBERT_{ZH2EN}$ achieves slightly higher correlations for nouns in the “Away" feature compared to verbs and adjectives, yet the overall scores remain weak. Similarly, $mGPT_{EN2ZH}$ struggles across all parts of speech, reflecting significant challenges in transferring these specific features across languages.

The MSE results in subplots 7e –7h further emphasize the poor predictive performance on these features. Both models display elevated MSE values across all parts of speech, with particularly high values for mGPT when predicting Chinese norms using English embeddings. This indicates substantial difficulty in modeling these features effectively in a cross-lingual context.

Notably, these two features exhibited the lowest correlation scores when comparing English and Chinese human ratings (see Fig. 1). The low agreement in human

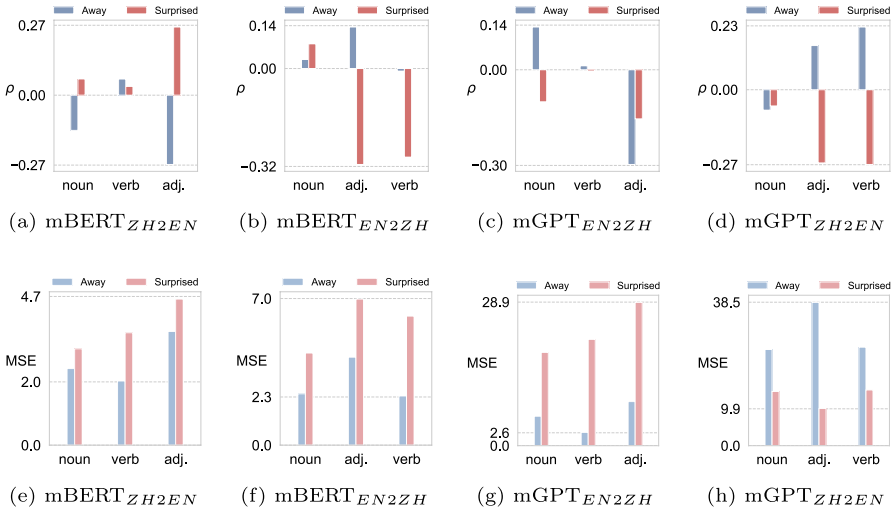



**Fig. 7** Spearman correlation scores (subplots (**a**–**d**)) and MSE (subplots (**e**–**h**)) for the "Away" and "Surprised" features across parts of speech in cross-lingual norm prediction under a zero-shot setting

data for “Away" and “Surprised" suggests inherent cross-linguistic differences, likely influenced by cultural divergence. This divergence is reflected in the contextual representations learned by the language models, further complicating cross-lingual norm prediction for these dimensions.

More broadly, the highest and lowest correlations achieved by multilingual models resemble those observed in human ratings across languages. Visual features such as Color, Shape and Pattern tend to have moderate-to-high scores, while spatial and temporal features often have coefficients close to zero. The fact that spatio-temporal semantic features may differ a lot between languages is frequently discussed in studies on linguistic relativity (Boroditsky, 2001; Casasanto & Boroditsky, 2008; Filipović & Jaszczolt, 2012), which stress the different types of conceptualization of time and space existing in different cultures. For example, a study by Boroditsky (2011) reported differences between English and Mandarin speakers in the prevailing directionality of spatiotemporal metaphors (i.e., Mandarin is mainly using “time-moving" metaphors, whereas English is mainly using “ego-moving" ones), and this is coherent with our results, in which Away and Toward are among the features with the lowest correlations across languages.

Figure 8 is showing the Spearman correlation scores by part of speech for the MLMs (subplots 8a-8b) and CLMs (subplots 8c-8d) language models. In English, we observe consistent results with previous studies (Chersoni et al., 2021): nouns

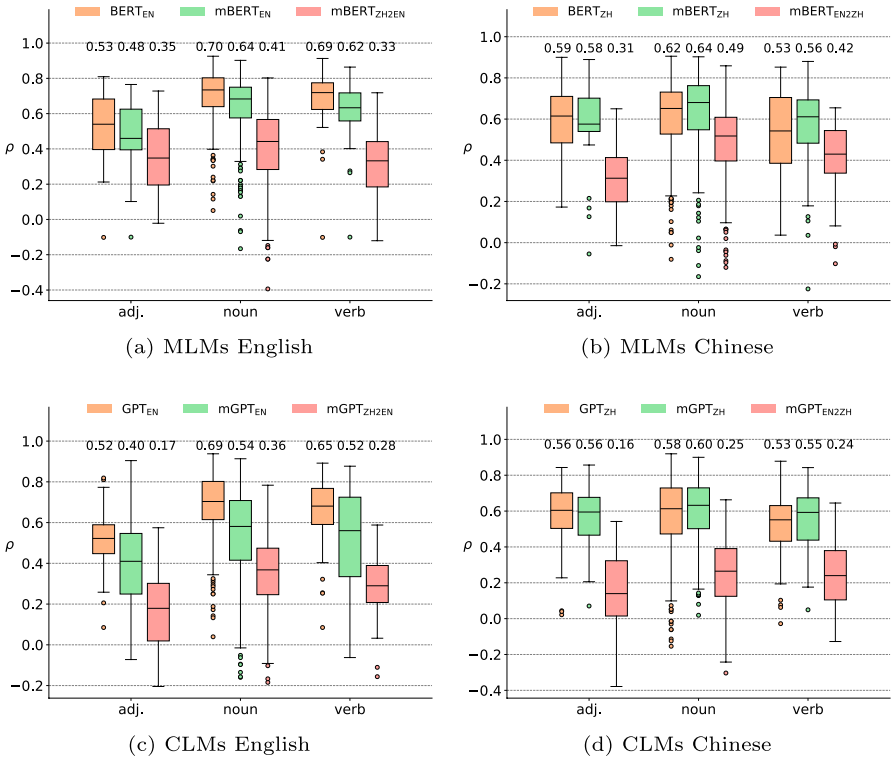



**Fig. 8** Average Spearman correlation scores split by model/training setting and by parts of speech

**Table 5** The average MSE and Spearman's $\rho$ scores across features and words for LLMs prompting

| | Feature Correlation | | Word Correlation | |
|---|---|---|---|---|
| | MSE | $\rho$ | MSE | $\rho$ |
| Llama3.1$_{EN}$ | 3.927 ± 1.522 | 0.376 ± 0.143 | 3.927 ± 1.367 | 0.413 ± 0.151 |
| Qwen2.5$_{EN}$ | 2.376 ± 0.916 | 0.500 ± 0.147 | 2.376 ± 0.889 | 0.553 ± 0.150 |
| GPT4*EN* | **1.060 ± 0.538** | **0.692 ± 0.093** | **1.060 ± 0.528** | **0.752 ± 0.099** |
| Llama3.1$_{ZH}$ | 3.683 ± 2.988 | 0.244 ± 0.135 | 3.684 ± 7.855 | 0.339 ± 0.206 |
| Qwen2.5$_{ZH}$ | **1.449 ± 0.564** | 0.406 ± 0.145 | **1.449 ± 0.406** | 0.448 ± 0.155 |
| GPT4$_{ZH}$ | 2.135 ± 0.667 | **0.473 ± 0.137** | 2.135 ± 0.577 | **0.588 ± 0.163** |

The top and bottom sections present the results of norm prediction for LLMs using English and Chinese prompts

are generally better predicted by both types of models, closely followed by verbs, whereas adjective norms are the most challenging ones to predict; in Chinese, nouns are still the easiest part of speech but adjectives come as a close second, being slightly better predicted than verbs (with the exception of the worse performing mGPT in the cross-lingual setting). It should be noticed that the edge of the nouns above the other parts-of-speech would be larger, if it was not for a relatively large number of outliers in the left tail of the distribution.

### 6.2 LLMs prompting

Prompting LLMs to predict the ratings of Binder features also showed some interesting results. First of all, it can be noticed that in English GPT-4 achieves similar results to a fine-tuned BERT Base model (see Table 4 and 5), with slightly worse MSE scores but higher correlation coefficients. The two open models are outperformed by the bidirectional BERT models, but are competitive or better than the GPT models. In Chinese, we can observe that Qwen2.5 is the best model in terms of MSE, while the correlations of GPT-4 are better both at the feature and at the word level. Compared to embedding-based models, both trail behind a Multilingual BERT fine-tuned on Chinese data for the MSE metrics, but they are more competitive with the correlation ones, especially GPT-4, which actually achieves a better correlation than mBERT at the feature level.

The solid performance of GPT-4o on Chinese is interesting, and to our knowledge this is the first time that this model is tested on a norm prediction task for a language other than English. Notice also that the task involves a much higher number of features/dimensions than in Xu et al. (2023), where the model was tested on English sensorimotor norms (18 semantic dimensions in total, against 59 features in our study) (Glasgow et al., 2016; Lynott et al., 2020). We want to stress that the first version of the Chinese dataset by Qiu and colleagues (Qiu et al., 2023) was published only in late summer 2023, after the cutoff training point for GPT-4o, so it is very unlikely that the results for GPT-4o have been affected by any sort of data contamination. Given the recent results obtained by Martínez et al. (2024b) on Spanish, it is possible that by further refining the prompts and by including a few more examples with GPT-4o

could lead to even higher performances. Looking at the open models, it is also noticeable that, without any supervised training and with just a pair of examples, a smaller architecture such as Qwen2.5-7B surpasses mGPT across all metrics for English and achieves a higher average $\rho$ score for Chinese.

In our view, it is important that research on the actual capabilities of LLMs systematically evaluates both state-of-the-art closed models and open models. Considering the lack of transparency regarding the architecture behind GPT-4o, it would be overly optimistic to generalize findings from specific closed-source models to LLMs as a whole. There is a crucial distinction between "what GPT-4 can do" and "what LLMs can do", and the remarkable performance gap between GPT-4o and the other LLMs at this stage should work as a strong reminder of this point. Moreover, there is no guarantee that closed models will remain static – they may be retrained in the future, with drastic performance changes (Rogers, 2023). Finally, in the specific case of Chinese, some closed models might not be officially available.

We believe that these considerations should encourage NLP researchers to work to enhance performance of open LLMs, which could be achieved through proper fine-tuning and more carefully designed instructions. In this sense, the recent development and release of GPT-4-level open models such as DeepSeek-V3 (Liu et al., 2024) is an extremely significant development. If prompting LLMs to predict norms can outperform the extraction of contextual representations, the collection of norming data for less resourced languages could be made much easier and cheaper for everyone. Therefore, we think research on this topic should focus more on improving open source LLMs and less on closed ones, due to reproducibility and lack of transparency concerns.

## 7 Conclusions

The analysis of lexical meanings in different languages requires a set of shared features with cross-linguistic validity that can be used to describe word semantics. In our work, the Binder features (Binder et al., 2016) served this role, given their cognitive motivation, therefore we have collected a new dataset of human ratings for Chinese words based on such features, and we make it available for future research. The dataset was collected by closely following the methodology of the original Binder collection, in order to have a comparable resource. A correlation analysis revealed that agreement between speakers for the same word in the two languages varies across features: visual, causal, social, cognition and emotion features tend to have a relatively high level of agreement, while spatial and temporal features have the lowest one.

After completing the norm collection, we conducted experiments to automatically predict norms using two methodologies: (1) projection of contextualized embeddings in both monolingual and cross-lingual settings, and (2) prompting LLMs. Each approach has its strengths and limitations. Embeddings projection generally provides higher prediction accuracy and is, at the moment, particularly useful for low-resource languages for which the most recent LLMs may not be available. On the other hand,

LLM prompting simplifies the data collection process by eliminating the need for model training, making it a promising tool for rapid norm acquisition.

For embeddings projection, we observed that monolingual models, such as BERT, generally outperform multilingual models in supervised settings, especially for English. However, multilingual models demonstrate competitive performance in zero-shot cross-lingual settings, particularly for a few specific features (e.g. “Vision", “Audition"). This suggests that multilingual models offer valuable generalization capabilities across languages, providing support to the hypothesis that the same concepts in different languages share some similarities in their representation (De Varda et al., 2025). We also found interesting that the lowest correlations are observed for spatial and temporal features, on which also between human speakers of different languages the agreement was low. Consistently with this finding, the linguistic conceptualization of space and time in different culture is often a point of discussion in studies of linguistic relativity (Boroditsky, 2011).

In the case of LLM prompting, we found that GPT-4 achieves performance comparable to a fine-tuned BERT model for English and demonstrates strong performance on Chinese, even though it had not been previously tested on a norm prediction task for a non-English language. Open models, such as Qwen2.5-7B, also show promising results, surpassing some of the supervised models in both English and Chinese, and performing better than GPT-4 itself for the error-based metrics in Chinese.

In conclusion, our work demonstrates the feasibility of using both embeddings projection and LLM prompting for semantic norm prediction, each with its own strengths and limitations. The availability of our new Chinese dataset, along with the insights gained from our experiments, provides a valuable resource for future research in cross-linguistic semantic analysis. Future work could focus on refining these methodologies, particularly in improving the cross-lingual alignment of embeddings and enhancing the domain-specific performance of LLMs.

**Acknowledgements** This work has been supported by the Start-up Fund for new recruits of the Hong Kong Polytechnic University (1-BE8G) and the RGC Direct Allocation Grant (A-PB1C). We would also like to thank the two reviewers for their constructive feedback.

**Author contributions** B. P. is the main authors of the code used to run the experiments and wrote the part about the experimental procedure. Y.Y.H. and L.Q. took care of statistical analysis, and Y.Y.H. also wrote the section about data collection. E.C. conceptualized the study and wrote the remaining sections of the manuscript. C.H. worked on the final manuscript revision.

**Funding** Open access funding provided by The Hong Kong Polytechnic University

**Data availability** Data and code to reproduce the study are provided at the following link: https://anonymous.4open.science/r/norms_correlation-F185

## Declarations

**Conflict of interest** The authors declare no Conflict of interest.

## References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., & Avila, R. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774

Apidianaki, M. (2023). From word types to tokens and back: A survey of approaches to word meaning representation and interpretation. *Computational Linguistics, 49*(2), 465–523.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors. In: Proceedings of ACL.

BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., & Reddy, S. (2024). LLM2Vec: Large language models are secretly powerful text encoders. In: Proceedings of COLM.

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology, 33*(3–4), 130–174.

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.

Boroditsky, L. (2011). How languages construct time. In S. Dehaene & E. Brannon (Eds.), *Space, time and number in the brain* (pp. 333–341). Elsevier.

Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive Psychology, 43*(1), 1–22.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & Agarwal, S. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Buchanan, E. M., Valentine, K. D., & Maxwell, N. P. (2019). English semantic feature production norms: an extended database of 4436 concepts. *Behavior Research Methods, 51*, 1849–1863.

Bulat, L., Kiela, D., & Clark, S.C. (2016). Vision and Feature Norms: Improving Automatic Feature Norm Learning Through Cross-Modal Maps. In: Proceedings of NAACL-HLT.

Casasanto, D., & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition, 106*(2), 579–593.

Chang, T.A., Arnett, C., Tu, Z., & Bergen, B.K. (2023). When is multilinguality a curse? Language modeling for 250 high-and low-resource languages. arXiv preprint arXiv:2311.09205

Chersoni, E., Xiang, R., Lu, Q., & Huang, C.-R. (2020). Automatic learning of modality exclusivity norms with crosslingual word embeddings. In: Proceedings of *SEM.

Chersoni, E., Santus, E., Huang, C.-R., & Lenci, A. (2021). Decoding word embeddings with brain-based semantic features. *Computational Linguistics, 47*(3), 663–698.

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In: Proceedings of the international conference on neural information processing systems. Curran Associates Inc., Red Hook, NY, USA.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In: Proceedings of ACL.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In: Proceedings of ACL.

De Varda, A.G., Malik-Moraleda, S., Tuckute, G., & Fedorenko, E. (2025). Multilingual computational models reveal shared brain responses to 21 languages.

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The centre for speech, language and the brain (CSLB) concept property norms. *Behavior Research Methods, 46*(4), 1119–1127.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL.

Fagarasan, L., Vecchi, E.M., & Clark, S. (2015). From distributional semantics to feature norms: Grounding semantic models in human perceptual data. In: Proceedings of IWCS.

Filipović, L., & Jaszczolt, K. M. (2012). *Space and time in languages and cultures: Language, culture, and cognition*. De Gruyter.

Flor, M.M. (2024). Three studies on predicting word concreteness with embedding vectors. In: Proceedings of the LREC-COLING workshop on cognitive aspects of the Lexicon.

Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. In: Proceedings. of NeurIPS.

Glasgow, K., Roos, M., Haufler, A., Chevillet, M., & Wolmetz, M. (2016). Evaluating semantic models with word-sentence relatedness. arXiv preprint arXiv:1603.07253

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., & Ma, Z. (2024). The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783

Hewitt, J., & Liang, P. (2019). Designing and interpreting probes with control tasks. arXiv preprint arXiv:1909.03368

Hoerl, A. E., & Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 42*(1), 80–86.

Hui, B., Yang, J., Cui, Z., Yang, J., Liu, D., Zhang, L., Liu, T., Zhang, J., Yu, B., Dang, K., Yang, A., Men, R., Huang, F., Ren, X., Ren, X., Zhou, J., & Lin, J. (2024). Qwen2.5-Coder Technical Report, arXiv:2407.21783 [cs]

Jackendoff, R. (1992). *Semantic structures* (Vol. 18). MIT Press.

Karthikeyan, K., Wang, Z., Mayhew, S., & Roth, D. (2019). Cross-lingual ability of multilingual BERT: An empirical study. arXiv preprint arXiv:1912.07840

Katja Wiemer-Hastings, K., & Xu, X. (2005). Content differences for abstract and concrete concepts. *Cognitive Science, 29*(5), 719–736.

Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., Fedorenko, E., & Lenci, A. (2023). Event knowledge in large language models: The gap between the impossible and the unlikely. *Cognitive Science, 47*(11), 13386.

Kivisaari, S.L., Hultén, A., Vliet, M., Lindh-Knuutila, T., & Salmelin, R. (2023). Semantic Feature Norms: A Cross-method and Cross-language Comparison. Behavior Research Methods, 1–10.

Kremer, G., & Baroni, M. (2011). A set of semantic norms for German and Italian. *Behavior Research Methods, 43*, 97–109.

Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. arXiv preprint arXiv:1711.00043

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.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics, 4*, 521–535.

Li, D., & Summers-Stay, D. (2019). Mapping distributional semantics to property morms with deep neural networks. *Big Data and Cognitive Computing, 3*(2), 30.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., & Dai, D. (2024). Deepseek-V3 Technical Report. arXiv preprint arXiv:2412.19437

Liu, N.F., Gardner, M., Belinkov, Y., Peters, M.E., & Smith, N.A. (2019). Linguistic Knowledge and Transferability of Contextual Representations. In: Proceedings of NAACL.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster sensorimotor norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods, 52*, 1271–1291.

Martínez, G., Conde, J., Reviriego, P., & Brysbaert, M. (2024). AI-generated Estimates of Familiarity, Concreteness, Valence, and Arousal for over 100,000 Spanish Words. Quarterly Journal of Experimental Psychology.

Martínez, G., Molero, J.D., González, S., Conde, J., Brysbaert, M., & Reviriego, P. (2024). Using large language models to estimate features of multi-word expressions: Concreteness, valence, arousal. arXiv preprint arXiv:2408.16012

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, 37*(4), 547–559.
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781
Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2013). Semantic memory: A feature-based analysis and new norms for Italian. *Behavior Research Methods, 45*, 440–461.
Murphy, G. (2004). *The big book of concepts*. MIT Press.
Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In: Proceedings of EMNLP.
Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In: Proceedings of NAACL.
Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? In: Proceedings of ACL.
Qiu, L., Hsu, Y.-Y., & Chersoni, E. (2023). Collecting and predicting neurocognitive norms for Mandarin Chinese. In: Proceedings of IWCS.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog.
Rivière, P.D., Beatty-Martínez, A.L., & Trott, S. (2024). Evaluating contextualized representations of (Spanish) ambiguous words: A new lexical resource and empirical analysis. arXiv preprint arXiv:2406.14678
Rogers, A. (2023). Closed AI Models Make Bad Baselines. Towards Data Science.
Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The glasgow norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods, 51*, 1258–1270.
Springer, J.M., Kotha, S., Fried, D., Neubig, G., & Raghunathan, A. (2024). Repetition Improves Language Model Embeddings. arXiv preprint arXiv:2402.15449
Tan, Z., Zhang, X., Wang, S., & Liu, Y. (2022). MSP: Multi-stage prompting for making pre-trained language models better translators. In: Proceedings of ACL.
Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R.T., Kim, N., Durme, B.V., Bowman, S., Das, D., & Pavlick, E. (2019). What do you learn from context? Probing for sentence structure in contextualized word representations. In: Proceedings of ICLR.
Thompson, B., & Lupyan, G. (2018). Automatic estimation of lexical concreteness in 77 languages. In: Proceedings of the annual meeting of the cognitive science society, vol. 40.
Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., & Rodriguez, A. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971
Trott, S. (2024). Can large language models help augment english psycholinguistic datasets? Behavior research methods, 1–19.
Trott, S., & Bergen, B. (2022). Contextualized Sensorimotor Norms: Multi-Dimensional Measures of Sensorimotor Strength for Ambiguous English Words in Context. arXiv preprint arXiv:2203.05648
Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*, 141–188.
Utsumi, A. (2020). Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science, 44*(6), 12844.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. In: Advances in Neural Information Processing Systems, pp. 5998–6008.
Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods, 40*(1), 183–190.
Vivas, J., Vivas, L., Comesaña, A., Coni, A. G., & Vorano, A. (2017). Spanish Semantic Feature Production Norms for 400 Concrete Concepts. *Behavior Research Methods, 49*, 1095–1106.
Vulić, I., Ponti, E.M., Litschko, R., Glavaš, G., & Korhonen, A. (2020). Probing pretrained language models for lexical semantics. In: Proceedings of EMNLP.
Wang, S., Zhang, Y., Shi, W., Zhang, G., Zhang, J., Lin, N., & Zong, C. (2023). A large dataset of semantic ratings and its computational extension. *Scientific Data, 10*(1), 106.
Wierzbicka, A. (1996). *Semantics: Primes and universals*. Oxford University Press.
Wu, Z., Chen, Y., Kao, B., & Liu, Q. (2020). Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In: Proceedings of ACL.

Xu, Q., Peng, Y., Nastase, S.A., Chodorow, M., Wu, M., & Li, P. (2023). Does conceptual representation require embodiment? Insights from large language models. arXiv preprint arXiv:2305.19103

Zhao, Z., Chen, H., Zhang, J., Zhao, X., Liu, T., Lu, W., Chen, X., Deng, H., Ju, Q., & Du, X. (2019). UER: an open-source toolkit for pre-training models. In: Proceedings of EMNLP-IJCNLP: System demonstrations.



## Authors and Affiliations

**Bo Peng[1] · Yu-yin Hsu[1] · Emmanuele Chersoni[1] · Le Qiu[1] · Chu-Ren Huang[1]**

✉ Emmanuele Chersoni
emmanuele.chersoni@polyu.edu.hk

Bo Peng
peng-bo.peng@polyu.edu.hk

Yu-yin Hsu
yu-yin.hsu@polyu.edu.hk

Le Qiu
lani.qiu@connect.polyu.hk

Chu-Ren Huang
churen.huang@polyu.edu.hk

[1] Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, 11 Yuk Choi Road, Hong Kong, China