Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A unitary representational space for abstract and concrete concepts derived from human similarity judgments

Andrew Persichetti

andrew.persichetti@jefferson.edu

Thomas Jefferson University

Charles Zheng

National Institute of Mental Health https://orcid.org/0000-0003-3427-0845

Taylor Li

University of California - Berkeley

Stephen Gotts

National Institute of Mental Health https://orcid.org/0000-0002-7903-1832

Francisco Pereira

National Institute of Mental Health

Alex Martin

National Institute of Mental Health https://orcid.org/0000-0002-3636-2498

Article

Keywords:

Posted Date: June 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9622928/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Abstract

Human semantic memory is a vast store of knowledge about people, places, and things (i.e., concrete concepts), and about intangible concepts like norms, ideas, and relations (i.e., abstract concepts). A central goal of cognitive science is to understand how concrete and abstract concepts are represented in semantic memory. We used an approximate Bayesian method called Variational Interpretable Concept Embeddings (VICE) to embed the meaning of 751 concrete and abstract nouns in a vector space using over 1.4 million human similarity judgments. Crucially, this method is theory-agnostic, hence the dimensions in the model were not constrained to a specific number or to have predetermined meanings. The result is a 36-dimensional semantic model, in which concrete and abstract concepts are intermingled both within dimensions and across the representational space. The meanings captured by individual dimensions are interpretable and the resultant representational space is organized along two main axes of 'concreteness' and 'agency'. Furthermore, the model can incorporate new data to grow the set of words it is trained on, thus providing a powerful tool for testing theories about the mental organization of human knowledge.

INTRODUCTION

Understanding the structure of semantic memory[1] – i.e., the set of all concepts that a person knows about – requires a theory that specifies how the meanings of those concepts are represented. The most prevalent models specify a latent representational space with *n* dimensions, each corresponding to a semantic property. Each concept is assigned a coordinate value in each dimension and is represented as a vector in the multidimensional space[2–8]. Thus, to uncover the neural code that represents knowledge in the human brain, we first need to accurately identify a set of dimensions that captures the meanings of diverse concepts, from concrete entities (i.e., "dog" and "carrot") to intangible, abstract concepts (i.e., "love" and "uncertainty"). However, the possible kinds of dimensions that structure the representational space of semantic memory are vast[4]. For example, semantic dimensions might range from those that reflect low-level, sensorimotor-related properties – like shape, color, and motion – to dimensions that represent higher-level information that can be abstracted away from low-level properties and capture contextual associations between concepts[9–12]. Therefore, the central challenge for semantic theories that posit a multidimensional representation of knowledge is to identify and interpret a set of dimensions that can represent the entire range of concepts, from the most concrete to the most abstract, a difficult task given the myriad differences between these types of concepts[13–33]. Once the structure of the representational space is identified, the next step is to determine the values of the coordinates for each concept within the space.

Historically, two types of methods have been used to uncover the dimensions that define a representational space in semantic memory. The first approach asks people to make explicit judgments about the properties (also referred to as features) of concepts along dimensions that are predetermined by the experimenter (e.g., "do you associate 'tomato' with a color?")[5,18,26,34–36]. These dimensions often

reflect low-level properties but may also include any type of dimension that the researcher sees fit. The key issues with this method are that it is biased by how the researcher decides which are the relevant property dimensions to inquire about[4,37], and it is not clear how to formulate dimensions that have the potential to capture shared properties of concrete and abstract concepts. A common approach for avoiding these problems has been to instead ask people to list as many properties as they can about a set of concepts and then derive dimensions from the lists[29,38,39]. However, when asked to freely list the semantic properties of concepts, people often tend to generate only a small subset of relevant properties while omitting many important ones[14,17,19,22,38,40]. For example, a person might focus on properties that can be easily deemed to be categorically present or absent, rather than present to a degree. Furthermore, this approach works much better for concrete concepts that have perceivable properties compared to more abstract concepts that lack such properties.

The second approach does not directly rely on human judgments but instead uses large language models (LLMs) to derive a dimensional representation of words – also called a "word embedding" – based on the statistics of language use in large corpora (e.g. word co-occurrence or prediction of masked words)[41,42]. Word embeddings free the researcher from making a priori assumptions about the relevant dimensions of the semantic space. However, the result is often a high number of dimensions optimized to reflect the many meanings of the words in the training corpus, and dimensions are not intrinsically interpretable as properties. While there are some indications that directions in such a space correspond to semantic properties[8], it is not clear how far this extends. Therefore, choosing to use either a property-rating approach or a word co-occurrence model approach results in a crisis of interpretability – the first approach results in a small number of interpretable but biased dimensions, while the latter results in a much larger number of unbiased but difficult-to-interpret dimensions.

More recently, a machine learning method called Variational Interpretable Concept Embeddings (VICE)[43] has been used to simultaneously infer dimensions under constraints that encourage interpretability, and to locate concepts in that space. The method requires the collection of behavioral judgments from an odd-one-out triad task, where participants select the odd one out among three randomly selected words from the full set. Given that information, it will embed each word in the space such that geometric proximity predicts human odd-one-out decisions. Intuitively, each dimension corresponds to some (semantic) property shared by a subset of the stimuli, and which plays a role in the decisions made across many triads. In this way, VICE avoids the thorny problems faced by the methods described above. Furthermore, the triad task has been used to discover concept embeddings for a broad range of concept types (e.g., common objects, actions, foods) and stimuli formats (e.g., pictures, movies, words)[43–47]. This flexibility of VICE makes it ideal for the difficult task of identifying a set of dimensions that represents the meanings of concrete and abstract concepts in a unitary representational space.

In this study, we introduce a semantic model that embeds the meanings of 751 diverse words from the Abstract Conceptual Feature (ACF) database[34]. The model was obtained by applying VICE to 1.4 million crowdsourced human similarity judgments made during randomly chosen odd-one-out triads from the

751-word list. This data-driven approach is agnostic to the representational principles of semantic memory. For that reason, it can derive a latent representational space in which the dimensions are free to vary in several important ways: 1) the dimensions can reflect low-level properties of concepts or higher-level information; 2) the dimensions can be equally applied to concrete and abstract concepts; 3) the dimensions are equally able to capture categorical distinctions between concepts or reflect graded semantic features, and 4) the set is not constrained to a specific number of dimensions. After building the semantic model, we used an LLM (ChatGPT 5.2 Pro – OpenAI) to provide multiple descriptions and labels for each dimension. Importantly, we chose the ACF words because the database includes a 14-dimensional semantic model that was created by asking people to explicitly rate a diverse set of words that samples the full gamut of nouns, ranging from abstract to concrete concepts, along experimenter-chosen dimensions that spanned "exogenous" (i.e., sensorimotor), "endogenous" (e.g., related to social, emotional, and cognitive domains), and "magnitude" (e.g., space and time) properties. Thus, basing our model on the set of ACF words allowed us to directly compare the dimensions from our method to those from a conventional rating scheme in an effort to understand how they differ, and to evaluate the types of information captured by the dimensions in our model. Being able to compare our semantic model to a model that is representative of commonly used methods is critical since we are introducing a new approach to this important problem.

RESULTS

Choosing a set of words and building a model of semantic knowledge.

Our goal was to build an intelligible model of semantic knowledge representations that is based on a large and diverse set of words. Thus, we based our experiment on the set of 751 words from the ACF database because it has several desirable features. First, the ACF database used concreteness ratings from the Medical Research Council (MRC) Psycholinguistic Database[48] to choose a diverse set of nouns from across the concreteness continuum, thus ensuring that the word set includes middling words that are not easily identifiable as concrete or abstract (e.g., ceremony) in addition to words from the extreme ends of the continuum (Fig. 1a). The concreteness ratings also provided a quantitative method of identifying concrete and abstract words in our experiment as needed. Second, the authors of the ACF database removed derivative words (e.g., boat vs. lifeboat) and used norms from the Subtlex American Word Frequency norms[49] and the MRC familiarity norms[48] to ensure that the words were frequent and familiar. Third, as mentioned above, the ACF database includes a 14-dimensional semantic model that was created using explicit ratings. Since the models from the ACF database and our computational approach are based on the same set of words and both embed the full set of word meanings in unitary semantic spaces, we can easily relate the ACF model to our results.

We used the full set of 751 words as stimuli in a triad odd-one-out judgment task. During each trial of the task, a participant is shown a triad of words (e.g., {"irony", "rebel", "independence"}) and asked to pick the odd-one-out. This task requires the participant to consider the three pairs in the triad (e.g., {"irony","rebel"}, {"irony","independence"}, {"rebel","independence"}). The participant decides which pair is

the most similar, then selects the third word as the odd-one-out (Fig. 1b). Thus, the triad task requires the participant to (at least implicitly) consider which dimensions of the words' meanings determine the similarity between a pair of words in the context of a third word. In the above example trial, a person may choose "irony" as the odd-one-out, thus judging "rebel" and "independence" to be similar to one another. Across the experiment, "rebel" and "independence" will be presented together in many more triads, each time with a new third word providing a new context and a new opportunity for the pair to be deemed similar or dissimilar. For example, when "rebel" and "independence" are displayed with "freedom", a person might now choose "rebel" as the odd-one-out. In the end, we quantified the pairwise similarity of words as the probability that participants will keep them together across a range of contexts provided by the third word. Another benefit of the triad task is that a similarity judgment can be made in seconds, thus allowing us to collect a large amount of data and present a given pair of words in many contexts. A total of 14,742 participants (7,176 female) successfully completed the odd-one-out similarity experiment on Amazon Mechanical Turk. Each iteration of the experiment was a set of twenty unique triads and four catch trials. Participants could complete as many sets as they liked. After discarding data if the participant missed a catch trial or did not finish the experiment, we were left with data from 1,421,200 triads (2.02% of all possible triad combinations of the 751 words used in the experiment).

To build a model of semantic knowledge, we used the VICE method to embed the 751 words in a vector space using the large set of human similarity judgments collected in the triad task. VICE uses variational inference to generate candidate dimensions, and assign sparse, non-negative values to all the words in each dimension, with uncertainty estimates for those values. The estimates are used iteratively to automatically select the dimensions that best explain the behavioral data (Fig. 1c). VICE allows the values in dimensions to vary from absent to present in a gradual way, rather than forcing a dimension to be either absent or present (as in semantic property norms). It captures the relative importance of dimensions by not capping the range of the positive end, thus "bigger" dimensions play a larger role in the dot product calculations that predict human similarity judgments. These features give the VICE model the expressiveness to capture dimensions that can look like categories, component properties, or graded semantic properties. The result is an unbiased model of semantic knowledge that captures mental representations for a diverse set of word meanings.

VICE produces a stable and predictive model of semantic knowledge.

Since VICE solutions are sensitive to the initial starting point, we ran the VICE model 20 times with random initializations to get a sense of the stability of the dimensions. On average, there were 35.9 dimensions across the 20 models (min-max = 33–39) with 36 being the modal number of dimensions (occurring 11 times). All the prediction results reported were obtained in each of these 20 models, and we report the means and confidence intervals for each result.

First, we tested the predictive performance of each model on an independent set of human similarity judgments. The independent dataset comprised similarity judgments over 1,000 random word triads that were not included in our initial dataset. This tests whether the semantic characteristics associated with

each learned dimension are operative in novel word combinations. Given that there is no "correct choice" for each triad, there is necessarily variability in the answers that human participants can give. Thus, the best performance possible at predicting human behavior for each of the 1,000 triads would be that of a model predicting the most common, or highest probability, choice across participants. To make it possible to estimate this, the independent set contains 20 independent judgments for each of the 1,000 triads. These were obtained by asking 200 participants to make a decision on 100 of the 1,000 triads. We then calculated the unpredictability (or “noise”) in the data by measuring the variability in human judgments across the 20 samples for each of the 1,000 triads. After averaging across the triads, we determined the upper limit in fitting trial-level behavior from the data was 52.03% (± 1.07%), while the chance level would be 33.3%.We used this upper limit as a noise ceiling to compare against the performance of the twenty VICE-derived semantic models. Next, we used the dimensions from the models to predict the human similarity judgments from the independent dataset. Accuracy was remarkably consistent across the twenty semantic models and all models performed well-above the chance-level of 33.3%. On average, the models correctly predicted 41.76% (± 0.3%) of the individual trials from the independent triads (Fig. 2a). When compared to the noise ceiling, the average performance of the models was 80.3% of the best possible prediction any model could achieve for our set of 751 words. We selected the model with the highest prediction accuracy for further analyses of the semantic model derived using VICE. This example model achieved 81.5% of the best possible prediction and comprised 36 dimensions (i.e., the modal number of dimensions across the models).

In addition to using the semantic model to predict trial-level human behavior, we evaluated its ability to model behaviorally measured representational similarity across a set of words. Since our set of triads sampled a small fraction of the triads used to produce the 751x751 word similarity matrix (2.02%), these data cannot be used to reasonably assess how well the model can predict the behaviorally measured representational similarity in our full word set. Therefore, we used a previously collected dataset, in which all triads for a subset of 50 words from our full set (i.e., 19,600 triads) were sampled twice[37]. The 50x50 similarity matrices generated from each set of human judgments were highly correlated with one another (Pearson’s r = 0.91). Therefore, we generated a single 50x50 similarity matrix by combining the two sets of triad data and compared it to a similarity matrix created using the dimension values from the VICE-derived semantic model (Fig. 2b). The two matrices were strongly correlated (Pearson’s r = 0.83) with performance approaching the correlation between the two behavioral similarity matrices, thus demonstrating that the VICE-derived semantic model can accurately reproduce behaviorally measured similarity between word pairs.

Dimensions in the semantic model reflect conceptual properties that span the concreteness continuum.

Thus far, the results demonstrate the reproducibility of the semantic model and its ability to predict human judgments and achieve human-like representational similarity between word pairs. At this point, we examined whether each dimension represented a characteristic related primarily to abstract or to concrete words alone or, instead, something shared by words across the concreteness continuum. First,

we used the concreteness scores provided in the ACF database to calculate the ratio of concrete and abstract words in each dimension. We labelled a dimension as concrete or abstract if more than two-thirds of the words with weightings on that dimension were one or the other. Using this rule for identifying the concreteness of the thirty-six dimensions, we found five predominantly concrete dimensions and thirteen abstract dimensions, while the remaining eighteen dimensions contained a more even mix of concrete and abstract words **(**Fig. 3a**)**. Thus, a large portion of the dimensions in the semantic model capture information that is common across concrete and abstract words.

Next, we asked whether the dimensions were intelligible and what aspects of word meaning were captured by each. Figure 3b displays the top words from a few examples of abstract, concrete, and mixed dimensions with labels given to the dimensions by ChatGPT (see Online Supplementary Materials for the full list of labels for all dimensions and the rationale given by the AI for why the labels were chosen). As one might expect, the dimensions that were identified as concrete reflect concepts related to things in the world like “animals” and “weapons”, while abstract dimensions reflect concepts like “prosocial” and “reasoning”. Regardless of the degree of concreteness, each dimension reflects high-level conceptual information rather than low-level properties of concepts. For example, “zoo”, “frog”, and “gorilla” all have strong weights in the “animal” dimension, even though they refer to a place, an amphibian, and a mammal, respectively, and thus share few low-level properties.

To get a better understanding of how each semantic dimension is related to low-level properties, we leveraged the explicit feature ratings from the ACF database. Each of the fourteen dimensions from the ACF semantic model reflects a specific property that can ostensibly be grounded in either sensorimotor, social, cognitive, emotional, or magnitude systems. To quantify how much of the information in our model is present in the ACF model, we used a ridge regression to predict each of the VICE dimensions from the fourteen dimensions in the ACF (Fig. 4a). We first considered which types of ACF properties explained variance in each VICE-derived semantic dimension (Fig. 4b). Pearson’s correlations between select semantic dimensions from our model and ACF property dimensions show an intuitive correspondence between the models. For example, the “food” dimension is most correlated with “taste/smell” from the ACF model ($r_{(749)} = 0.60$), while the “reasoning” and “literacy” dimensions are most correlated with the “thought” dimension from the ACF model ($r_{(749)} = 0.39$ and 0.32, respectively –**Figure 4c**). However, much variance in the semantic dimensions is left unexplained by the ACF model. One possibility is that the ACF model contains broadly the same information as ours but apportions it differently between dimensions. To quantify this, we first need to characterize how much of the variability in VICE dimensions is predictable (rather than due to noise). Therefore, we ran ridge regression using the dimensions from the nineteen other VICE-derived semantic models to measure stability across the VICE-derived models and to obtain a noise ceiling to contrast with the results from the ACF-property regression. The VICE-derived semantic models were highly stable, with all but five dimensions achieving an $R^2 > 0.5$ (Fig. 4b, pink line in left panel) and the average R^2 across dimensions was 0.78 (Fig. 4b, pink bar in right panel). By contrast, when using the ACF model, only one dimension achieved an $R^2 > 0.5$ (Fig. 4b, gray line in left panel) and the average R^2 across dimensions was 0.22 (Fig. 4b, gray bar in right

panel). Thus, the ACF properties leave a large portion of variance unexplained in the VICE-derived semantic dimensions. Thus, while the dimensions in our semantic model capture aspects of word meaning that correspond to relevant ACF properties, they go well beyond these component low-level properties.

Emergent category structure in the semantic embedding.

In the previous section, we characterized each dimension by considering the values of all the words in that dimension. Conversely, we can characterize each word as a set of values corresponding to its coordinates along these interpretable dimensions and use them to understand how its meaning is represented in the multidimensional space. This is illustrated for a few words as rose plots in Fig. 5, which highlight the fact that the meaning of a word is represented as a combination of a subset of dimensions. We can also visualize the overall organization of the space by reducing the 36-dimensional semantic embedding to a two-dimensional map that captures similarity between the words, using t-distributed stochastic neighborhood embedding (t-SNE) (Fig. 6). The representational map reveals clear global organizing principles that give rise to local category-like clusters. The global organization runs along two main axes of 'concreteness' and 'agency'. Along the concreteness axis, the words shift from concrete categories like people, animals, food, scenery, and artifacts, to abstract ones like emotions, measurements, ideas, and qualities. Also of note, we do not see a sharp divide between concrete and abstract words, as might be expected based on influential models of word concreteness, like the Dual Coding Theory[50]. Instead, we see a gradual shift from concrete to abstract words – i.e., a continuum of concreteness[34]. Along the agency axis, concrete words shift from foods, objects and scenery to animals and people, while the abstract words move from measurements and statements (e.g., "fact" and "definition") to emotions and ideas. In addition to these dominant axes, valence also influences the organization of the representational space. For example, weapons are separated from other artifacts and positive emotions and ideals are separable from negative emotions and transgressions.

Category-like structures emerge from the global organization of our semantic embedding (Fig. 6). These clusters range from oft-described categories like food, animals, people, and emotions to less common ones related to money, entertainment, performing arts, and geography. The integration of concrete and abstract words is clear at this local level, as well. For example, words for conspecifics, like "boy", "girl", "sister", and "father" are found adjacent to more abstract words that are related to people, like "mankind", "gender", and "youth". Another example cluster contains the words "food", "menu", "pie", and "coffee". Overall, our results demonstrate that human similarity judgments can be used to derive a unitary semantic space that integrates concrete and abstract words, and organizes meaning in a way that captures both the properties of concepts and category-like relationships between them.

DISCUSSION

Understanding how knowledge is stored in semantic memory requires a characterization of the dimensions that give meaning to concepts in a representational space. In the current report, we

introduce a 36-dimensional model of semantic memory in which the meanings of both concrete and abstract concepts are represented as sets of dimension weights. Crucially, the dimensions were not postulated a priori by us or by other researchers. Instead, the number of dimensions and the information they captured were learned from over a million human similarity judgments collected in an online odd-one-out triad task. We chose a data-driven approach because we did not want to make a priori decisions that constrained the model to a specific number of dimensions with predetermined meanings. Instead, we built a computational model of semantic knowledge by approaching the problem with four basic assumptions: 1) human knowledge comprises concepts that span a continuum of concreteness – from concrete objects and properties to abstract ideas. 2) A model of semantic memory should be a unitary multidimensional representational space that accommodates all concepts. 3) Making similarity judgments in an odd-one-out triad task requires people to consider the relevant dimensions that make two concepts similar to one another in the context of the third (odd-one-out) concept in each trial. 4) The set of dimensions that people use to assess similarity is sufficient to capture the meaning of the concept. Thus, collecting many triads using diverse concepts provides a large dataset to train and test a computational model that will naturally find the semantic dimensions that best explain human knowledge about the full set of concepts[43,45]. We chose the VICE method because it requires dimensions to be sparse (i.e., each concept will not have weightings on every dimension) and non-negative (i.e., all dimensional weights are positive, so combining dimensions will not cancel each other out) to help ensure that dimensions will be interpretable. We demonstrate that this data-driven modelling approach yields intelligible dimensions that can be used to accurately predict independent human judgments and reproduce the representational similarity of word pairs derived from human behavior. Furthermore, we show that our model organizes knowledge along two primary axes of concreteness and agency in a way that gives rise to local category structure.

In contrast to our data-driven approach, many commonly used semantic modelling approaches do rely on a priori property selection to determine which dimensions to include in the model. In these approaches researchers can choose to base the dimensions on a range of principles – for example, empirical observations of functional selectivity in the brain[3] or a specific theoretical framework[34] – and then ask people to rate individual concepts on how well they align to each dimension. While the methodology differs between these approaches and ours, we do not think that the methods or the results are adversarial to one another. Quite the contrary, we chose to use the word set from the ACF database because models with prespecified property dimensions can be help us to understand the dimensions in our semantic model (and vice versa), and the models can even potentially be combined. For example, in the current paper, we used component property dimensions from the ACF model to help us understand which types of information are being captured by the dimensions in our model. We found intuitive results, such as the “taste/smell” dimension from the ACF model explains some, but not all, of the variance in the “food” dimension from our model (Fig. 4d). This type of analysis offers insights into how the dimensions in our semantic model relate to specific low-level properties, thus providing another method of interpreting the dimensions, and the potential for refining and guiding the model moving forward.

We believe the main limitation to our model is the relatively small number of words that it was trained on. The current set of 751 English nouns is a small sample of nouns, not to mention other parts of speech. However, increasing the size of the input dataset for our semantic model can be achieved rather easily, by collecting new triad data with new sets of words that contain a reasonable number of words from the existing set that serve as a bridge between the sets. In fact, we collected the data used in this iteration of the model by "bridging" word sets (see Methods section). That is, we collected triads for the 378 most abstract words separately from the triads for the 373 most concrete words. We then collected a set of 370 words that were from the middle of the concreteness scale (i.e., half of the words were from the abstract set and the other half from the concrete set). The 370-word set worked as a bridge to combine the concrete and abstract sets of words and thus provided a larger set of human similarity judgments that could be modelled together. Thus, we, and other researchers, can add to the existing model by collecting triad data in a new set of words that includes a subset of words from the existing model training set. In this way, we can incorporate other word classes into the model, in addition to more nouns. All data are available online so researchers can download the most up-to-date triad set and add on to it (LINK). Newer iterations of the semantic model will provide opportunities to test the durability of the existing dimensions when new words are added and to understand how different relationships between words alter the representational space. As the model grows, so grows a powerful tool for understanding how human knowledge is represented in a multidimensional space.

In summary, we introduced a data-driven model of semantic memory that is based on 751 concrete and abstract nouns. Since the dimensions in the model are generated from human behavior and were not constrained to specific properties, they offer a novel embedding of meaning that has many potential applications. For example, these novel dimensions may better reflect how meaning is represented in complex brain networks compared to an approach of combining component low-level property dimensions. If so, our semantic model should make it possible to explain more variance in brain activation than competing models and/or help guide development of novel experiment designs that would make it possible to adjudicate between theoretical models. The concept embeddings in the model could also give therapists working with patients who suffered a stroke or neurodegenerative disease a tool for uncovering patterns of semantic deficits or naming errors that otherwise might not be obvious. These types of insights may allow the therapist to select tailored sets of concepts and guide the order of presentation in training paradigms. For both basic science and downstream applications, it will likely be necessary to expand the word set. Fortunately, the model can be expanded by collecting more triads with new sets of words that can be incorporated into the current dataset and modelling procedure. Thus, we can continue to build a model of semantic memory that accommodates more domains of human knowledge.

METHODS

Participants

A total of 15,691 participants (7672 female; mean age = 42.6, s.d.=11.4) completed the odd-one-out similarity experiment on Amazon Mechanical Turk. Each iteration of the experiment was a set of 20 unique triads and 4 catch trials. Participants could complete as many sets as they liked. We excluded data from a set in which the participant either (1) failed to supply the correct answer on a catch trial; (2) had an average trial response time of less than 500 ms; (3) or selected the same response on every trial. In the event of a rejected block, the individual was excluded from further participation in the experiment and the block was posted again. After excluding data that did not pass these quality checks, we were left with data from from 14,742 participants (7176 female) used for training and testing the computational model and a separate group of two hundred participants (97 female) that each completed 100 triads used for estimating a noise ceiling from 1,000 randomly chosen triads that were not part of the training and testing dataset. Participants were required to be high school graduates, native English speakers, and live within the United States. These participants gave informed consent under the National Institutes of Health (NIH) Institutional Review Board-approved protocol number 000589 and were compensated financially for their time.

Human similarity judgments

To collect human similarity judgments on a large set of concrete and abstract concepts, we created word triads using the full set of 751 words from the abstract conceptual feature (ACF) database. Each triad comprised 3 randomly selected words from the full word set. We randomly selected 1,421,200 triads from the total set of 70,312,375 triads (i.e., 2.02% of all possible triad combinations of 751 items). We then separated the triads into 71,060 unique sets of 20 triad trials each and posted the sets on Amazon's Mechanical Turk to obtain similarity judgements using an odd-one-out triad task. On each trial of the experiment, 3 words were presented in random order in a horizontal row under the instruction to "click on the odd-one out." The participants used a computer mouse to click on the chosen word, which was followed by a blank screen for 500 milliseconds (ms), then the next trial. Participants were given as much time as needed to make their choice. To ensure that participants were attentive during each set, we included 4 catch trials in which participants saw the words "PLUS," "MINUS," and "EQUAL" randomly ordered in a horizontal line under the instruction to "click on EQUAL." Thus, there were 24 trials in each set. Individuals were allowed to complete as many sets as they liked. Before starting the experiment, participants were presented with instructions and an example triad that included words that were not used in the experiment.

VICE MODEL

We took the resulting data set of odd-one-out judgments, consisting of 1,421,204 triads, and split it into a training set of 1,278,976 triads (which is 9,992 batches of 128 each) and a validation set of 142,228 triads. We trained VICE[43] on this data using a hyperparameter grid search consisting of the following parameters:

pi ranging from 0.1 to 0.9 in increments of 0.1 sigma_slab ranging in powers of 2 from 0.125 to 16.0, and sigma_spike ranging in powers of 2 from 0.125 to 4.0, with the constraint that sigma_spike < sigma_slab, and using 20 different random seeds for each parameter combination. For all models, we set initial dimensionality to 100, number of epochs to 2000 with no early stopping, learning rate to 0.001, batch size to 128, number of MC samples to 5, and using CPU (rather than GPU). Post-training, dimensions were pruned using threshold of 5 minimal number of items (min_k). We selected the final parameter combination (pi = 0.9, sigma_spike = 0.25, sigma_slab = 1.0) based on the average cross-entropy over 20 seeds on the validation set. For follow-up analyses, we selected the best of the 20 seeds for that parameter combination based on the best accuracy on the separate noise ceiling test set.

CHAT GPT

We used ChatGPT 5.2 Pro (OpenAI) to obtain labels that capture the meaning of each dimension in our semantic model. We gave the following prompt to ChatGPT for each dimension

"The following twenty words are the top weighted words in a semantic dimension (presented in order): <top twenty words from dim n>

Please give three high-level labels that best capture the meaning represented by the semantic dimension. Each label should be three words or less. After listing all three labels, give a short rationale for why you chose those three labels. Please answer as quickly as possible."

The average response time across the thirty-six prompts was 96.9 seconds (SD 28.2s). The labels and the rationale for selecting them given by ChatGPT can be found in the online supplemental material.

RIDGE REGRESSION

We used ridge regression on our semantic model to measure the stability of the VICE models across the 20 random seeds and to compare our semantic model with the component low-level property dimensions from the ACF model. First, we ran 19 separate ridge regressions on our semantic model, each time using the dimensions from a VICE model derived from one of the other 19 random seeds. We used the closed-form solution for the ridge regression, in which a VICE model from a random seed (X) was fit to our semantic model (*y*):

$$\mathbf{b = (XTX + II)^{-1} XTy}$$

Before running each regression, we chose the optimal value for the regularization parameter (λ) using cross-validation on a range of values from 0.1 to 40. The same method was used to fit the ACF model dimensions (X) to our semantic dimensions (*y*).

References

1. Tulving E (1972) Episodic and semantic memory. Organization of memory. Academic, Cambridge, MA, pp 381–403
2. Reilly J, Finley AM, Litovsky CP, Kenett Y (2023) N. Bigram semantic distance as an index of continuous semantic flow in natural language: Theory, tools, and applications. J Exp Psychol Gen 152:2578–2590
3. Binder JR et al (2016) Toward a brain-based componential semantic representation. Cogn Neuropsychol 33:130–174
4. Reilly J et al (2025) What we mean when we say semantic: Toward a multidisciplinary semantic glossary. Psychon Bull Rev 32:243–280
5. Crutch SJ, Troche J, Reilly J, Ridgway GR (2013) Abstract conceptual feature ratings: the role of emotion, magnitude, and other cognitive domains in the organization of abstract conceptual knowledge. Front Hum Neurosci 7
6. Barsalou LW, Dutriaux L, Scheepers C (2018) Moving beyond the distinction between concrete and abstract concepts. Philos Trans R Soc B Biol Sci 373:20170144
7. Frisby SL, Halai AD, Cox CR, Ralph L, M. A., Rogers TT (2023) Decoding semantic representations in mind and brain. Trends Cogn Sci 27:258–281
8. Grand G, Blank IA, Pereira F, Fedorenko E (2022) Semantic projection recovers rich human knowledge of multiple object features from word embeddings. Nat Hum Behav 6:975–987
9. Schwanenflugel PJ, Fabricius WV, Noyes CR, Bigler KD, Alexander JM (1994) The Organization of Mental Verbs and Folk Theories of Knowing. J Mem Lang 33:376–395
10. Barsalou LW (1999) Perceptual symbol systems. Behav Brain Sci 22:577–660
11. Mahon BZ, Caramazza A (2008) A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. J Physiol -Paris 102:59–70
12. Montefinese M (2019) Semantic representation of abstract and concrete words: a minireview of neural evidence. J Neurophysiol 121:1585–1587
13. Borghi AM (2022) Concepts for Which We Need Others More: The Case of Abstract Concepts. Curr Dir Psychol Sci 31:238–246
14. Borghi AM et al (2019) Words as social tools: Language, sociality and inner grounding in abstract concepts. Phys Life Rev 29:120–153
15. Lupyan G, Mirman D (2013) Linking language and categorization: Evidence from aphasia. Cortex 49:1187–1194
16. Gleitman LR, Cassidy K, Nappa R, Papafragou A, Trueswell JC (2005) Hard Words. Lang Learn Dev 1:23–64
17. Fini C, Era V, Da Rold F, Candidi M, Borghi AM (2021) Abstract concepts in interaction: the need of others when guessing abstract concepts smooths dyadic motor interactions. R Soc Open Sci 8:201205

18. Villani C, Lugli L, Liuzza MT, Borghi AM (2019) Varieties of abstract concepts and their multiple dimensions. Lang Cogn 11:403–430
19. Della Rosa PA, Catricalà E, Vigliocco G, Cappa SF (2010) Beyond the abstract—concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words. Behav Res Methods 42:1042–1048
20. Ponari M, Norbury CF, Vigliocco G (2018) Acquisition of abstract concepts is influenced by emotional valence. Dev Sci 21:e12549
21. Reilly J, Kean J (2007) Formal Distinctiveness of High- and Low-Imageability Nouns: Analyses and Theoretical Implications. Cogn Sci 31:157–168
22. Paivio A et al (2011) Dual coding theory, word abstractness, and emotion: A critical review of Kousta. *J. Exp. Psychol. Gen.* 142, 282–287 (2013)
23. Kelly AE et al (2024) Conceptual structure of emotions. Emotion 24:1550–1561
24. Newcombe PI, Campbell C, Siakaluk PD, Pexman PM (2012) Effects of Emotional and Sensorimotor Knowledge in Semantic Processing of Concrete and Abstract Nouns. Front Hum Neurosci 6
25. Connell L, Lynott D, Banks B (2018) Interoception: the forgotten modality in perceptual grounding of abstract and concrete concepts. Philos Trans R Soc B Biol Sci 373:20170143
26. 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. Behav Res Methods 52:1271–1291
27. Villani C, Lugli L, Liuzza MT, Nicoletti R, Borghi AM (2021) Sensorimotor and interoceptive dimensions in concrete and abstract concepts. J Mem Lang 116:104173
28. Harpaintner M, Sim E-J, Trumpp NM, Ulrich M, Kiefer M (2020) The grounding of abstract concepts in the motor and visual system: An fMRI study. Cortex 124:1–22
29. Harpaintner M, Trumpp NM, Kiefer M (2018) The Semantic Content of Abstract Concepts: A Property Listing Study of 296 Abstract Words. Front Psychol 9:1748
30. Kiefer M, Pielke L, Trumpp NM (2022) Differential temporo-spatial pattern of electrical brain activity during the processing of abstract concepts related to mental states and verbal associations. NeuroImage 252:119036
31. Trumpp NM, Ulrich M, Kiefer M (2024) Experiential grounding of abstract concepts: Processing of abstract mental state concepts engages brain regions involved in mentalizing, automatic speech, and lip movements. NeuroImage 288:120539
32. Zdrazilova L, Sidhu DM, Pexman PM (2018) Communicating abstract meaning: concepts revealed in words and gestures. Philos Trans R Soc B Biol Sci 373:20170138
33. Desai RH, Reilly M, van Dam W (2018) The multifaceted abstract brain. Philos Trans R Soc B Biol Sci 373:20170122
34. Troche J, Crutch SJ, Reilly J (2017) Defining a Conceptual Topography of Word Concreteness: Clustering Properties of Emotion, Sensation, and Magnitude among 750 English Words. Front

Psychol 8
35. Devereux BJ, Tyler LK, Geertzen J, Randall B (2014) The Centre for Speech, Language and the Brain (CSLB) concept property norms. Behav Res Methods 46:1119–1127
36. Banks B, Connell L (2023) Multi-dimensional sensorimotor grounding of concrete and abstract categories. Philos Trans R Soc B Biol Sci 378:20210366
37. Persichetti AS, Shao J, Denning JM, Gotts SJ, Martin A (2024) Taxonomic structure in a set of abstract concepts. Front Psychol 14:1278744
38. McRae K, Cree GS, Seidenberg MS, Mcnorgan C (2005) Semantic feature production norms for a large set of living and nonliving things. Behav Res Methods 37:547–559
39. Buchanan EM, Valentine KD, Maxwell NP (2019) English semantic feature production norms: An extended database of 4436 concepts. Behav Res Methods 51:1849–1863
40. Plaut DC, Shallice T (1993) Deep dyslexia: A case study of connectionist neuropsychology. Cogn Neuropsychol 10:377–500
41. Kumar P (2024) Large language models (LLMs): survey, technical frameworks, and future challenges. Artif Intell Rev 57:260
42. Miller GA (1995) WordNet: a lexical database for English. Commun ACM 38:39–41
43. Muttenthaler F, Zheng PL, McClure CY, Vandermeulen P (2022) R. A. ,. Hebart, M. N. VICE: Variational Interpretable Concept Embeddings. Adv Neural Inf Process Syst
44. Hebart MN et al (2019) A database of 1,854 object concepts and more than 26,000 naturalistic object images. PLOS ONE 14 THINGS:e0223792
45. Hebart MN, Zheng CY, Pereira F, Baker C (2020) I. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. Nat Hum Behav 4:1173–1185
46. Crutch SJ, Connell S, Warrington EK (2009) The different representational frameworks underpinning abstract and concrete knowledge: Evidence from odd-one-out judgements. Q J Exp Psychol 62:1377–1390
47. Bockes A, Hebart MN, Lingnau A (2025) Revealing Key Dimensions Underlying the Recognition of Dynamic Human Actions. Commun Psychol 3:149
48. Coltheart M, The MRC, Psycholinguistic Database (1981) Q J Exp Psychol Sect A 33:497–505
49. Brysbaert M, New B (2009) Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behav Res Methods 41:977–990
50. Paivio A (1991) Dual coding theory: Retrospect and current status. Can J Psychol Rev Can Psychol 45:255–287

Figures

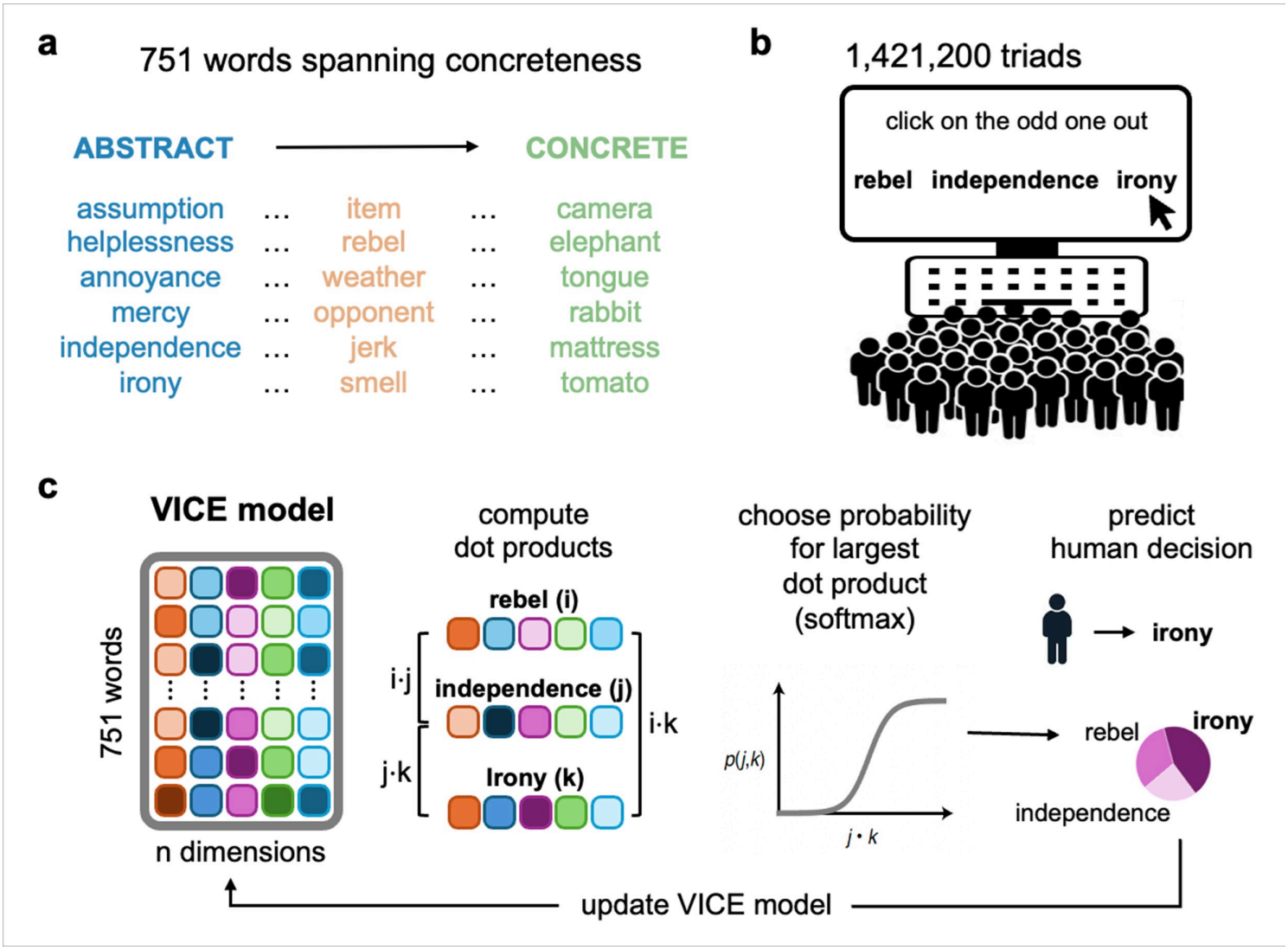


Figure 1

Online task and VICE modelling procedure. **a**, We used the complete word set from the ACF database to span the gamut of concreteness. **b**, Nearly 15,000 online participants completed a combined total of 1,421,200 triad odd-one-out similarity judgments on 751 words (i.e., 2.02% of all possible triads). Choosing the odd one-out of three words is an implicit judgment that the other two words are similar in the context of that triad. **c**, We used the VICE method to embed the word meanings in a latent representational space based on the relatively small subset of triads sampled in our experiment. The embedding was initialized with random weights and used the dot product between the pairs in each triad to predict which word pair was most similar as a way to then choose the odd-one-out. The model predictions were compared to human judgments and the model was updated accordingly via error backpropagation to the weights.

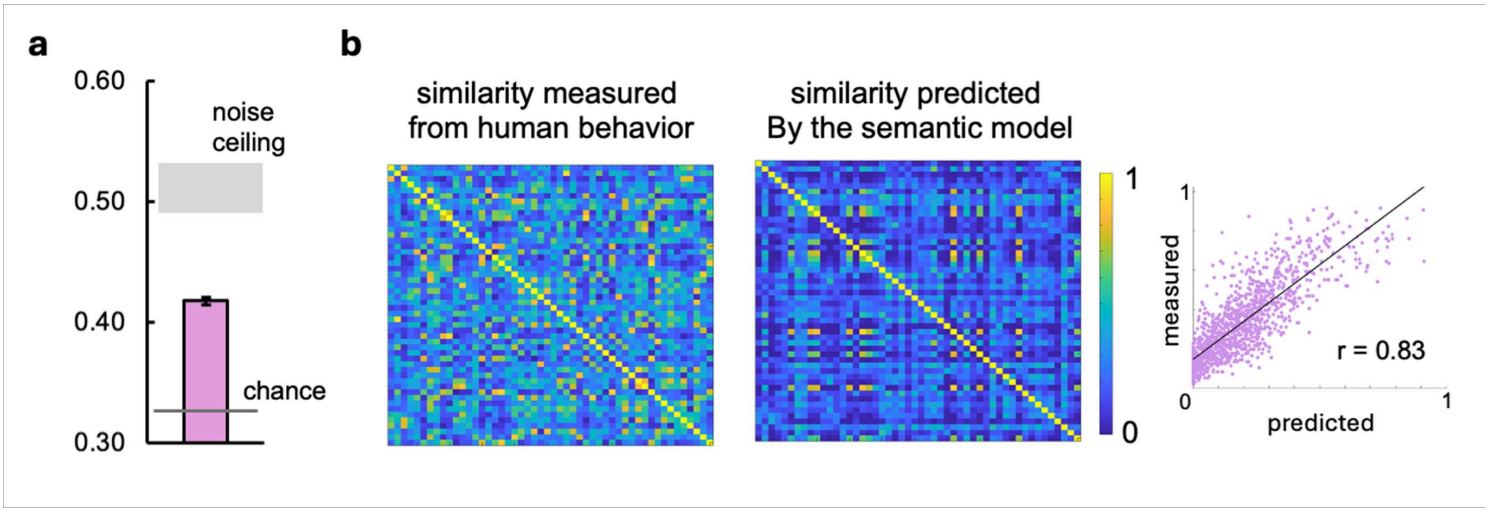


Figure 2

The semantic model predicts single-trial human judgements and representational similarity. a, Predictive performance of the semantic model was evaluated by its ability to predict individual trials in a separate set of human similarity judgments. The noise ceiling was calculated as the consistency of participants' responses to an independent set of 20,000 triads and indicates the best performance a model could achieve given the variability in the data, due to noise or (almost) evenly distributed choice probabilities. The ability of the VICE-derived semantic model to predict responses in the independent data reached 81.5% of the noise ceiling, thus demonstrating excellent predictive performance. The error bars and shaded areas denote +/- 1 standard deviation from the mean. **b,**To test the ability of the semantic model to estimate behavioral similarity, a similarity matrix created using the dimensions from the model was compared with a fully sampled behavioral similarity matrix for a subset of fifty words. The model-generated similarity matrix explained most of the variance in the behavioral data (Pearson's r = 0.83).

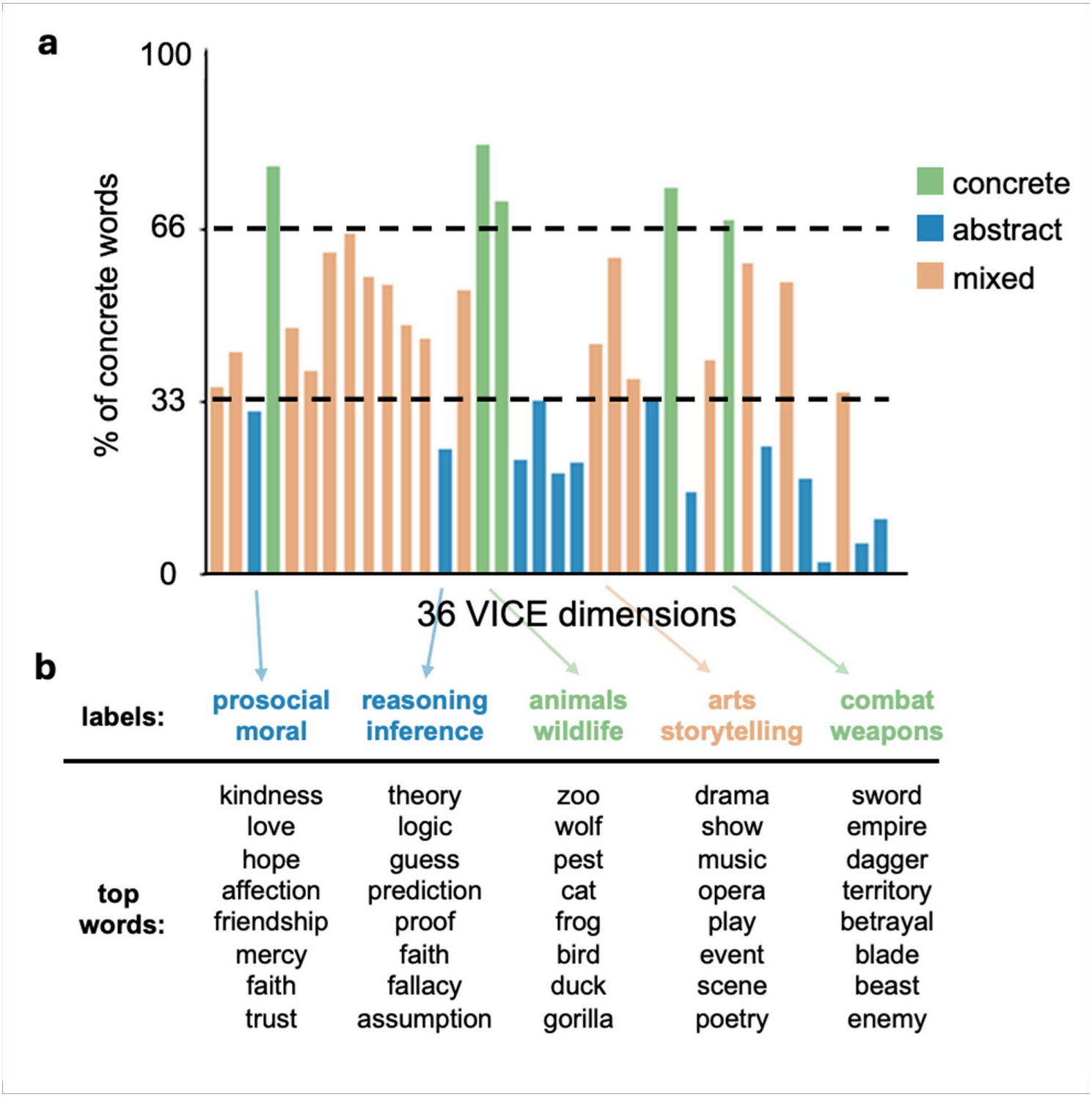


Figure 3

Concreteness of the semantic dimensions and descriptive labels. a, Of the thirty-six dimensions in the semantic model, five were predominantly concrete dimensions and thirteen were abstract dimensions, while the remaining eighteen dimensions contained a more even mix of abstract and concrete words. Note that the percentage of concrete words in each dimension is highly correlated with the mean concreteness value of each dimension (Pearson's r = 0.99). **b,** Five example dimensions with the top eight words from each dimension and labels given to the dimensions by Chat-GPT. The full list of labels

for all thirty-six dimensions and the rationale given by the AI for why the labels were chosen can be found in the Online Supplementary Materials.

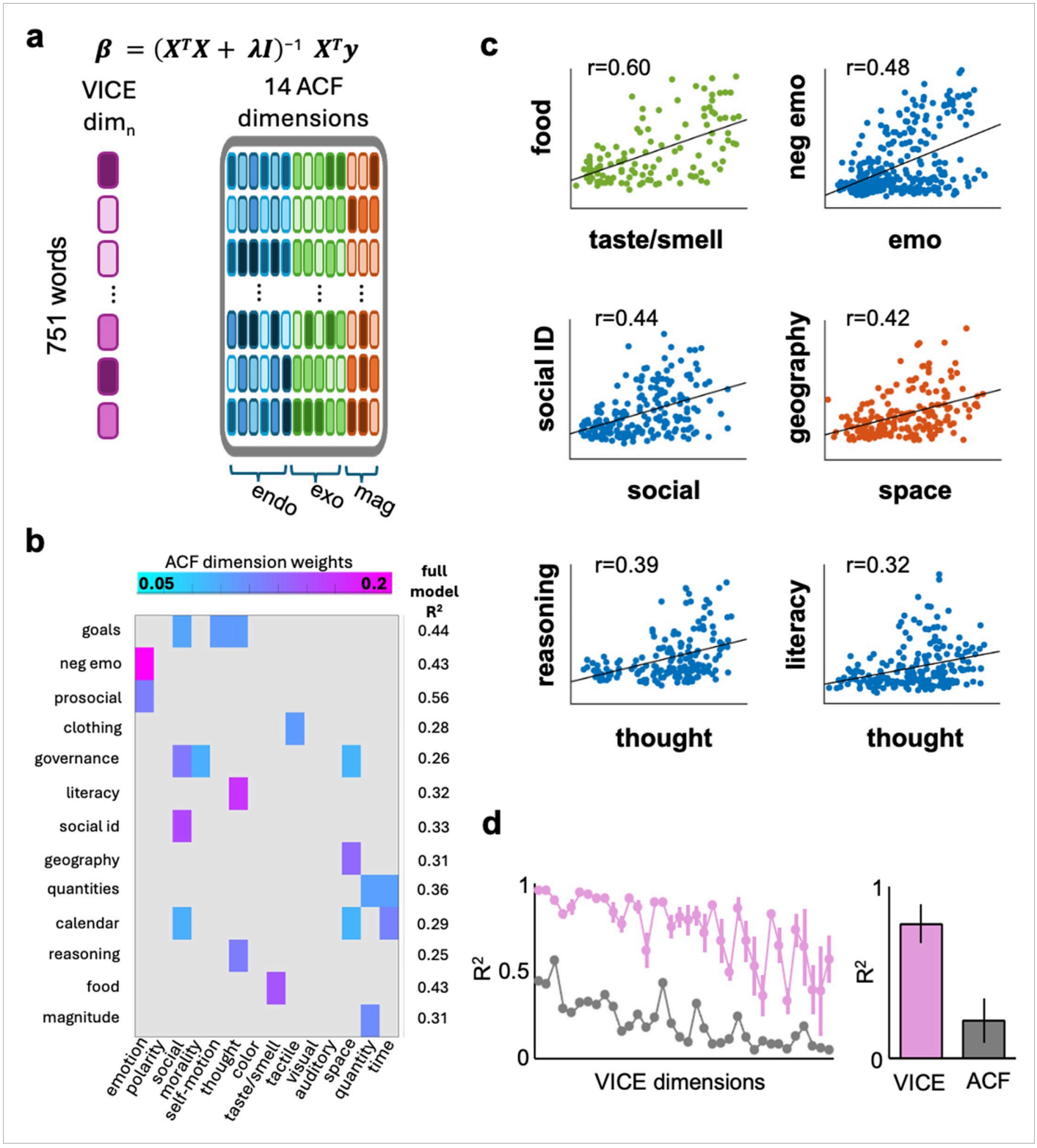


Figure 4

VICE semantic dimensions contain information beyond that in ACF properties. a, Ridge regression was used to fit the fourteen component property dimensions from the ACF model to each dimension in our

computational semantic model. The ACF-property dimensions can be divided into 6 “endogenous” (i.e., abstract), 5 “exogenous” (i.e., concrete), and 3 “magnitude” dimensions. We used ridge regression on each of the 36 dimensions in turn to ask how much variance is explained in each dimension by explicitly defined ACF properties (e.g., color, shape, size, emotion). **b,** Next, we investigated which types of properties explained variance in each dimension of our semantic model. The rows in Figure C[SG1] are the thirteen dimensions from our semantic model in which the R^2 from the ACF-property regression was greater than 0.25 with the labels provided by the generative LLM. The columns are the fourteen ACF properties. The cells in the matrix show where the beta coefficient for a property was greater than 0.05. **c,** Pearson’s correlations between VICE-derived (y-axis) and ACF-property (x-axis) dimensions highlight the correspondence between dimensions from our semantic model and related properties from the ACF model. **d,** VICE dimensions are stable across random seeds**.** Before fitting the ACF-property model to each VICE dimension (gray line and bars), we ran ridge regression using the dimensions from the nineteen other VICE models derived from random seeds (pink line and bar) to ask about the stability of the VICE models and to obtain a noise ceiling to contrast with the results from the feature regression. The VICE models were highly stable, with all but five dimensions achieving an $R^2 > 0.5$ (line graph, left panel) and the average R^2 across dimensions was 0.78 (bar graph, right panel). By contrast, when using the ACF-property model, only one dimension achieved an $R^2 > 0.5$ (line graph, left panel) and the average R^2 across dimensions was 0.22 (bar graph, right panel). Error bars = +/- 1 standard deviation from the mean.

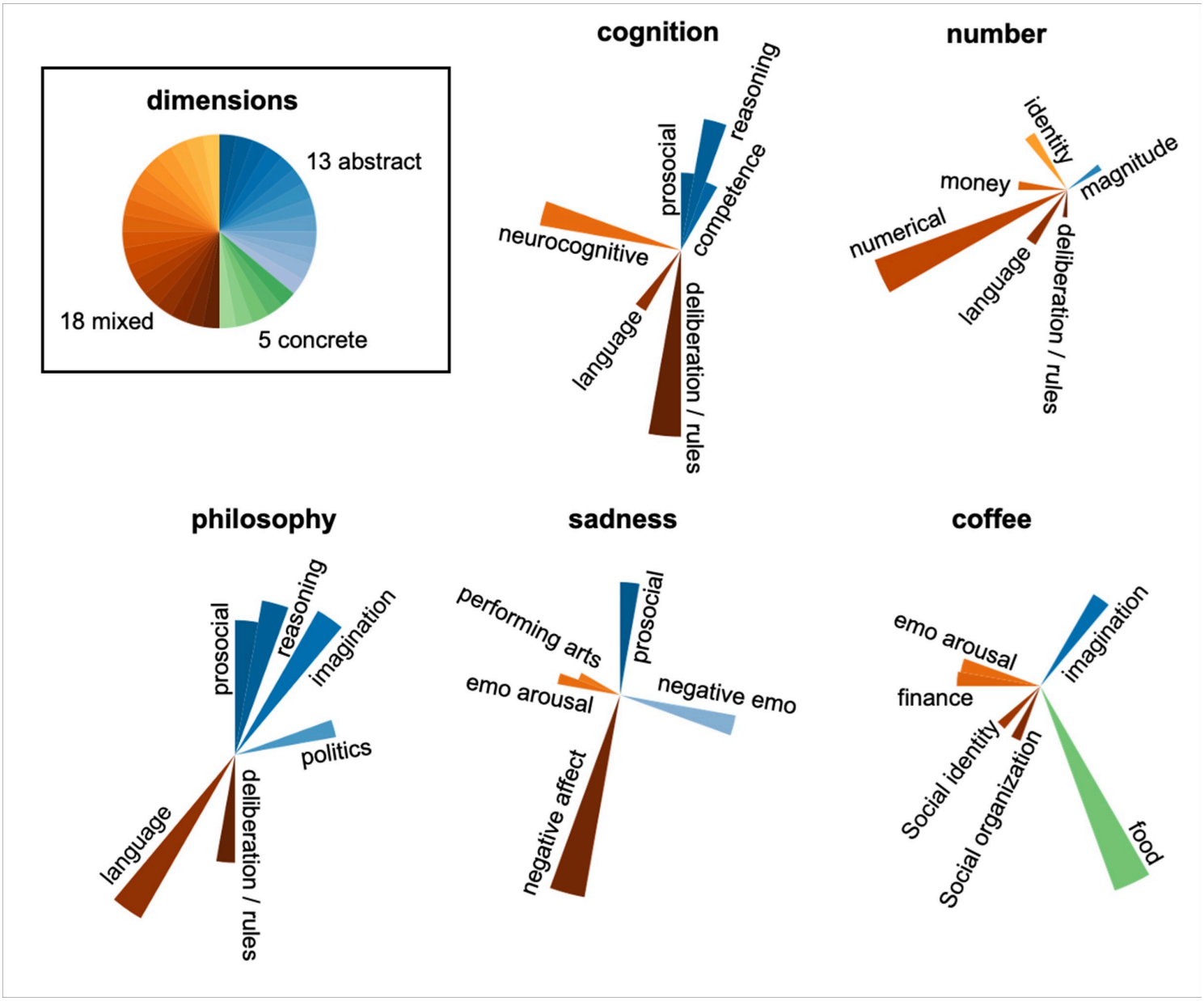


Figure 5

Rose plots for five example words show the dimensional weights that express word meaning in the semantic model. The petals on each rose plot are organized by the concreteness of the dimensions (see the legend in the upper left panel). The length of each petal signifies the strength of the weight along that dimension and the label given to each petal is from ChatGPT. The combination of these weights allows the meaning of a concept to be influenced by both dominant (big petals) and subtle (small petals) dimensions. For example, the word “number” gets its meaning mostly from the high weighting on the “numerical” dimension, but weaker weightings on dimensions like “identity” and deliberation/rules” also influence the meaning of the word and its relation to other words in the common semantic space. The full set of rose plots and dimension labels are provided in the online supplemental material.

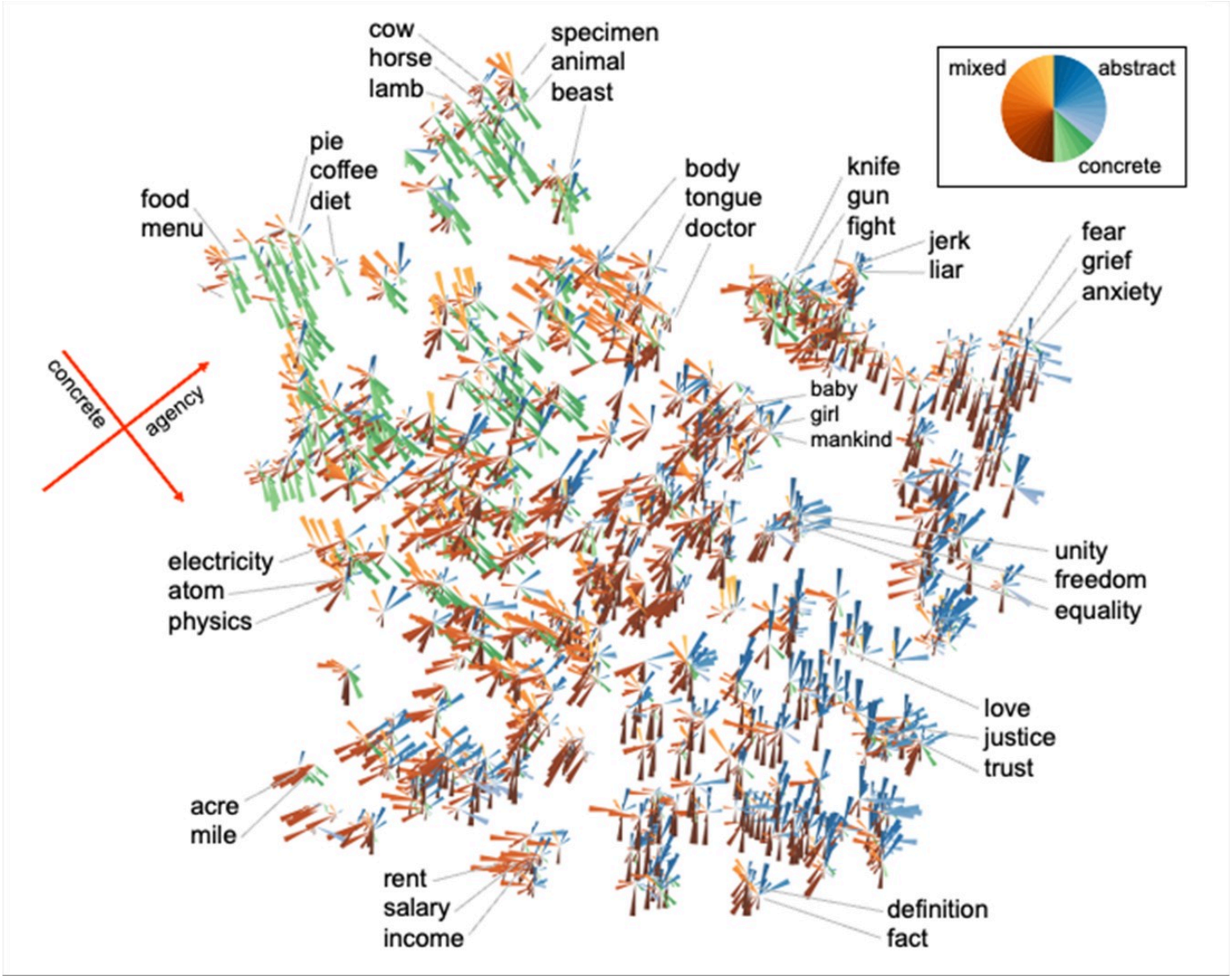


Figure 6

The 751 rose plots situated in a two-dimensional rendering of the word embeddings from the semantic model. Dimensionality of the semantic space was reduced from 36 to 2 dimensions using t-distributed stochastic neighborhood embedding. The rose plots represent the dimensional weights that give each word meaning. A subset of rose plots are labelled to illustrate the organizing principles that structure the representational map. The full set of rose plots and the code used to generate the t-SNE plot are provided in the online supplemental materials.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- viceWordssuppMatfinal.xlsx