Cheaper Spaces

Matthieu Moullec, Igor Douven

Minds and Machines (2025) 35:6
https://doi.org/10.1007/s11023-024-09704-x

Cheaper Spaces

Matthieu Moullec[1] · **Igor Douven**[2]

Received: 27 March 2024 / Accepted: 30 October 2024

Abstract
Similarity spaces are standardly constructed by collecting pairwise similarity judgments and subjecting those to a dimension-reduction technique such as multidimensional scaling or principal component analysis. While this approach can be effective, it has some known downsides, most notably, it tends to be costly and has limited generalizability. Recently, a number of authors have attempted to mitigate these issues through machine learning techniques. For instance, neural networks have been trained on human similarity judgments to infer the spatial representation of unseen stimuli. However, these newer methods are still costly and fail to generalize widely beyond their initial training sets. This paper proposes leveraging prebuilt semantic vector spaces as a cheap alternative to collecting similarity judgments. Our results suggest that some of those spaces can be used to approximate human similarity judgments at low cost and high speed.

Keywords Conceptual spaces · Deep learning · Multidimensional scaling · Psychological representations · Similarity judgments

1 Introduction

Over the past two decades, the conceptual spaces framework (CSF) has been gaining popularity in cognitive science and beyond (Gärdenfors, 2000, 2014; Nosofsky, 1986, 1987, 1992; Shepard, 1964, 1987). This is in large part because it offers researchers a mathematical framework for modeling concepts, concept learning, and the use of concepts in categorization and induction (see, among many other publications, Douven, 2016, 2023, 2024a, b; Douven & Gärdenfors, 2020; Douven et al., 2023; Gärdenfors, 2000; Gärdenfors & Osta-Vélez, 2023; Gärdenfors & Warglien,

✉ Matthieu Moullec
matthieu.moullec@etu.univ-paris1.fr

Igor Douven
igor.douven@univ-paris1.fr

[1] IHPST, Panthéon–Sorbonne University, Paris, France

[2] IHPST, CNRS, Panthéon–Sorbonne University, Paris, France

Published online: 14 December 2024

2012; Gärdenfors & Williams, 2001; Osta-Vélez & Gärdenfors, 2020, 2022). According to the CSF, concepts can be represented geometrically, as regions in similarity spaces, which are one- or multidimensional structures with a metric defined on them (for an overview, see Gärdenfors, 2000). The metric measures dissimilarity between items, in that the farther apart two items are in the space, the more dissimilar they are in the respect represented by the space (e.g., the more dissimilar their colors are, if the relevant space is color space; see below). The dimensions of a similarity space aim at representing measurable properties items may have, so that items can be mapped onto points in the space according to the values they assume on these properties.

The conceptual spaces approach has been applied across diverse domains of varying complexity. Examples of simple conceptual spaces are temporal space, which is represented by a singular dimension (time), and auditory space, which is characterized by the dimensions of pitch and loudness. Somewhat more complex are color spaces like CIELAB and CIELUV, which are three-dimensional, with hue, luminosity, and saturation as their dimensions. Still more complex conceptual spaces to be found in the literature are ones for actions, events, faces, tastes, scents, moral values, socio-economic status, pain, and much else (see, e.g., Bendifallah et al., 2023; Bourdieu, 1989; Castro et al., 2013; Churchland, 2012; Deauvieau et al., 2014; Douven, 2016; Gärdenfors & Warglien, 2012; Petitot, 1988; Valentine et al., 2016).

As intimated, conceptual spaces are built on top of similarity spaces. Similarity spaces can be constructed in a number of different ways. The most common approach entails gathering pairwise similarity judgments for a set of items and then using these judgments as input for a dimension-reduction technique. Multidimensional Scaling (MDS) is the most commonly used technique for this purpose (Borg & Groenen, 1999), while others such as Principal Component Analysis (PCA) and Non-Negative Matrix Factorization (NMF) are employed less frequently (Abdi & Williams, 2010; Castro et al., 2013). Other types of input data, such as confusion probabilities (indicating the likelihood of different items being mistaken for each other) and correlation coefficients (depicting the strength of correlation between items), are also occasionally used. A more recent alternative technique used for building similarity spaces is the spatial arrangement method (SpAM). This method requires participants to position items on a surface such that the distances among the items reflect the participant's similarity judgments, providing an intuitive and visual representation of perceived similarities (Goldstone, 1994).

A common procedure for transforming a similarity space into a conceptual space involves locating the prototypes of the concepts one wishes to represent within the relevant similarity space (Gärdenfors, 2000; Gärdenfors & Williams, 2001). For instance, color prototypes would be situated in CIELAB or CIELUV space. Following the identification and placement of prototypes, the mathematical technique of Voronoi tessellations (Okabe et al., 2000) is applied to segment the similarity space into distinct regions, by associating with each prototype all points in the space that are at least as close to it as they are to any of the other prototypes (Douven, 2016). Each of these regions represents a different concept (e.g., a different color concept if the underlying similarity space was CIELAB or CIELUV space).

Part of the appeal of the CSF stems from the fact that it is typically fairly straightforward to derive empirical predictions from a conceptual space (e.g., about issues like concept acquisition, or category-based induction, or graded membership, or a variety of other issues), thereby giving theories about those issues clear empirical content. That requires, of course, that the conceptual space is described in some detail. Ideally, one can load it onto one's computer and use modern software to interact with it (e.g., to measure distances in it, or to measure volumes of regions in the space). In practice, however, there is still only a limited number of conceptual spaces that are easily accessible for researchers, or even accessible at all. The root problem really concerns similarity spaces. For once we have a similarity space, its conversion into a conceptual spaces is generally rather straightforward. However, the construction of the underlying similarity space can be both very time-consuming and very expensive. In addition to this, many of the similarity spaces that *are* available are not readily generalizable to items that were not used in the process of generating the space.

Recently, a number of authors have attempted to mitigate these issues through machine learning techniques. Most notably, neural networks have been trained on human similarity judgments to infer the spatial representation of unseen stimuli (Attarian et al., 2020; Bechberger & Kühnberger, 2021; Nosofsky et al., 2017; Patel & Pavlick, 2021; Peterson et al., 2018; Sanders & Nosofsky, 2020). However, these newer methods are still costly and also do not generalize as much as would often be useful. In this paper, we look into a potentially cheap and simple alternative to collecting similarity judgments which leverages prebuilt semantic vector spaces and other tools from artificial intelligence. Our results suggest that from at least some of these spaces we can extract similarity judgments which approximate human similarity judgments to a satisfactory extent. Thereby, we can arrive at similarity spaces for some types of stimuli quickly and inexpensively.

In the following sections, we explore the prospects of constructing similarity spaces by recruiting large language models (LLMs) and word embeddings. Section 2 provides theoretical background on similarity spaces and the traditional methods for constructing such spaces. Section 3 presents our methodology and starts by illustrating it using GPT-4 as the currently top LLM. In Sections 4 and 5, we look at slightly older or less powerful models which, however, have the advantage of being open source and thereby offer a more direct approach to retrieving similarity judgments.

2 Practical Limitations

A similarity space that informs us of the similarities among only a small number of items will, in general, be of merely limited theoretical interest. To make sure a similarity space has a sufficiently broad scope, it will have to be constructed on the basis of judgments concerning the similarities among a relatively large set of items. This can easily cause practical problems, however. For n items, there are $\binom{n}{2}$ pairwise similarity judgments to be made (assuming that order of appearance plays no role;

otherwise we would need twice as many). Thus, the number of similarity judgments that are required increases quadratically with n.

To make this concrete, we take as an example the first study from Douven (2016), which used as materials 49 items to arrive at a shape space, specifically a space for representing container concepts such as VASE and BOWL. While 49 is not an excessively large number of stimuli for the task at hand, had a participant had to judge each pair they would effectively have had to make $\binom{49}{2} = 1176$ pairwise similarity judgments—which is practically infeasible, given that, in all likelihood, such a task would lead to participant fatigue (and accordingly low-quality responses) and probably also to a high participant attrition rate. The problem is known in the literature, and one workaround, which was also used in Douven (2016), is to recruit a large group of participants and let each participant make a "doable" number of similarity judgments instead of having each of a small number of participants judge the entire set of items. Specifically, for the relevant study in Douven (2016), each of over 1000 participants was asked to judge the similarity of 25 pairs of items selected randomly per participant. That guaranteed that each pair of items received a fair number of similarity ratings (at minimum). For each pair of items, the responses it had received were then averaged, and these averages served as input for the MDS procedure that yielded the container space used in the further studies reported in Douven (2016). (These further studies tested hypotheses about graded membership and are unrelated to our present purposes.)

This type of workaround still comes with a downside. Specifically, given the large number of participants it requires, studies using the workaround will typically be very costly, which—especially in view of the small research budgets that are common in the humanities and also in many of the social sciences—often presents an obstacle in itself. The study from Douven (2016) that was just described was at the time it was conducted already quite expensive, but at today's rates, with Prolific—currently the main crowdsourcing platform for academic research—recommending paying participants at least £ 9 per hour, the cost would have been over £ 800 (given that participants spent on average over 5 min on the survey).

There is another way to build similarity spaces, this one quick and cheap, which uses the spatial arrangement method (SpAM). As already briefly mentioned, SpAM lets participants directly construct a similarity space by allowing them to position items on a two-dimensional surface (usually virtually on a computer screen). But SpAM has its own limitations. For one, the task is known to be cognitively demanding, meaning that the number of items that participants can be asked to locate relative to each other must be on the smaller side. For another, the task forces a participant's similarity space to be two-dimensional, even though a standard MDS procedure on the basis of pairwise similarity judgments could have shown that the participant's similarity judgments are best represented by a three- or even four-dimensional space. (For further critical discussion, see Verheyen et al. (2016); Verheyen and Storms (2021); Verheyen et al. (2022)).

Here, we want to focus on a limitation shared by the MDS method and SpAM, to wit, the problem that the spaces resulting from applications of these methods tend to generalize poorly to stimuli that are of the same type as those used to build

the similarity spaces—which were used to elicit pairwise similarity judgments in the case of the MDS method and which were to be placed relative to each other in the case of SpAM—but that were not seen by the participants. To illustrate, we use the study that Douven et al. (2023) conducted to build a mammal space. This study relied on SpAM and used as stimuli mammals from the set of mammals that had been previously used by Henley (1969). However, because—as mentioned—SpAM presents participants with a cognitively demanding task, Douven and colleagues selected twenty mammals from Henley's set instead of using the full set of thirty. That was enough for these authors' purpose, which was to use the individual spaces created by the participants to predict the degrees to which these participants were willing to accept various similarity-based inferences involving mammals taken from the set of twenty. Not only were those predictions largely successful, when the authors aggregated the individual spaces, they found that the aggregate space (reproduced in Fig. 1 here) did an even better job predicting the said degrees.

To come to the problem, we note that among the mammals in Henley's set *not* used for Douven et al.'s 2023 study are chipmunks, beavers, and raccoon. The problem immediately becomes clear when we ask where, in the mammal space shown in Fig. 1, we should place these mammals. If the dimensions of the space corresponded to measurable properties of mammals, like their average life span, or their average weight, then it would be easy to locate them in the space: look up the average life span of a raccoon, look up their average weight, and find the corresponding coordinates in the space; similarly for chipmunks, beavers, and all other mammals in Henley's set left out by Douven and colleagues. But although in the present case the dimensions are *somewhat* interpretable—the *x*-axis seems

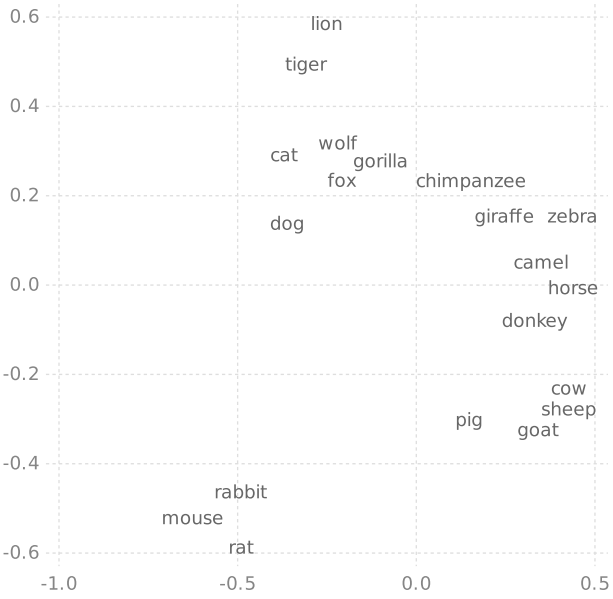

Fig. 1 Aggregate conceptual space for the Henley set, from Douven et al. (2023)

to correlate roughly with size, the *y*-axis with ferocity—it appears that they cannot be plausibly mapped 1 to 1 onto any measurable property of mammals. As a result, it is not obvious where in the space to put "chipmunk," or "beaver," or "raccoon." This problem is actually quite common: while in the ideal case the dimensions of similarity spaces correspond to measurable properties that items may have, in practice this ideal is often not met (Fig. 2).

Naturally, we could take recourse to a brute force solution by simply extending the set of items and rerunning the experiment, but that would be a bad idea precisely for the reason why Douven et al. (2023) chose only a subset of Henley's items. A more sensible approach, which is more in line with the proposal to be made in this paper, would be to follow Sanders and Nosofsky (2020), who trained a deep neural network on a similarity space and then used the trained net to predict where unseen items are to be placed in the space. While we believe their approach to be an important step forward, it could be argued that it still does not offer an optimal solution to the aforementioned problems. After all, in their approach, one still needs to construct a similarity space—which, as said, can be costly—and then train a neural net on the stimuli and similarity space, which is likely to add further costs.

We here would like to draw attention to the possibility of cutting down on expenses by using *pre-trained* models to obtain similarity judgments and letting them serve as input for an MDS procedure, or even to extract similarity spaces directly from those models. This could work on the assumption that, after pre-training, the models have come to already encapsulate what Sanders and Nosofsky's net was specifically trained to predict. It is worth considering this possibility, given that a wide variety of models have become available in recent years, some only as paid services, but some also for free thanks to open source contributions. In the remainder of the paper, we have a look at a number of the best-known

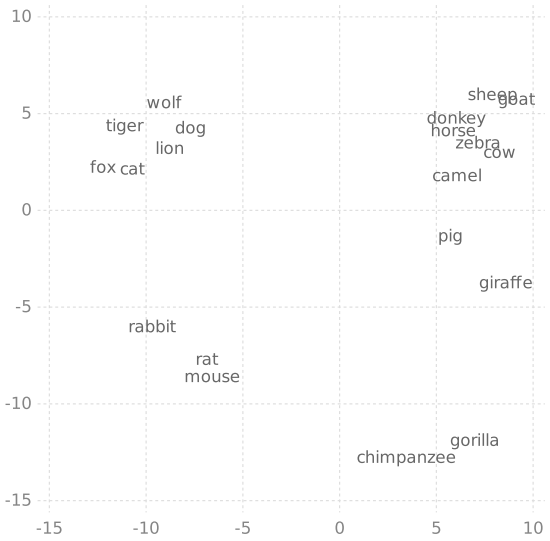

Fig. 2 Aggregate conceptual space for the Henley set based on GPT-4 similarity judgments

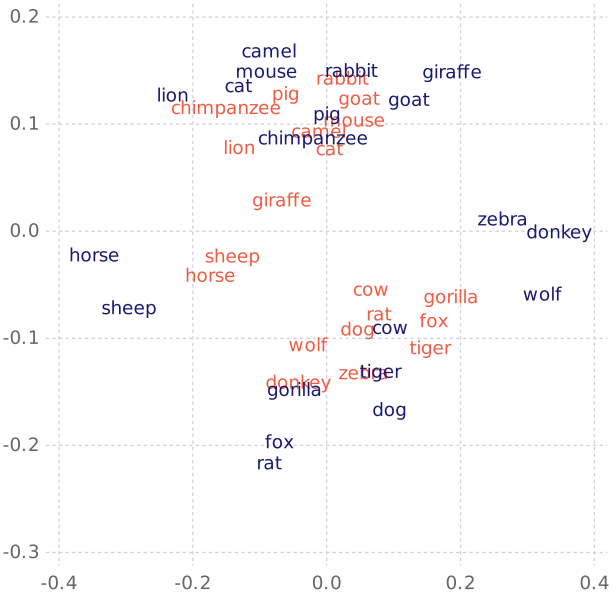

Fig. 3 Aggregate conceptual spaces for the Henley set based on human similarity judgments (in blue) and based on similarity judgments elicited from GPT-4 (in red) after Procrustes coordination

pre-trained models and will be interested in the extent to which they are able to help overcome the problem highlighted in this section.

3 Prompting GPT

The obvious model to start with is, of course, the currently widely popular, even if only commercially available, GPT-4, which is a state-of-the-art language model developed by OpenAI and based on the transformer deep learning architecture (OpenAI, 2023). Precisely because this is proprietary software, however, we are not able to inspect its layers and their activation functions, the importance of which will become relevant later on. But we can still ask it to generate similarity matrices for us—which is what we did, specifically for the set of twenty mammals used in the experiments reported in Douven et al. (2023).

Previous studies have shown encouraging results obtained from comparing GPT-4 with human judgments in inductive reasoning tasks through direct prompting (Han et al., 2024). These findings suggest a potential for similar success in tasks involving similarity judgments. We prompted GPT-4 five times, at different points in time, to create a similarity matrix for this set of mammals, explicitly asking it to use a scale from 0 to 10, with 0 indicating maximum dissimilarity and 10 indicating maximum similarity, and also to assign 10 only in the case of identity. We received each time a symmetric matrix, though the results differed slightly each time. By way

of example, Fig. 3 shows the space we obtained by conducting an MDS procedure with one of the matrices as input.[1]

The space certainly makes sense. But it will be more informative to compare it formally with the aggregate space from Douven et al. (2023) shown earlier. This might seem difficult, if only because the two spaces are on different scales. Also, while, for instance, "chimpanzee" and "gorilla" appear close together in both spaces, they appear in the bottom right corner in the GPT-4 space but in the upper right corner in the space from Douven et al. (2023). There is a way around these difficulties, however, stemming from the fact that similarity spaces are identical up to similarity transformations, meaning that they are not affected by any combination of the operations of shifting, rescaling, rotating, and mirroring, all of which preserve relative distances. This fact is exploited by a technique known as "Procrustes analysis" (Schönemann, 1966), which uses similarity transformations to align different spaces as closely as possible. The technique thus allows for the adjustment and alignment of the spaces, thereby addressing variances in orientation, position, and scale, and, as a result, facilitating a meaningful comparison.

More specifically, using the `protest` function from the `vegan` package for the statistical computing language `R`, we can compute correlations between similarity spaces after Procrustes coordination. For each of the five similarity matrices given by GPT-4, we find a very high and significant correlation between that space and the aggregate space from Douven et al. (2023); all correlations are at least close to .9, and two matrices even yielded a correlation of .92 (all ps $< .0001$). To formally check the consistency of GPT-4, we also looked at the correlations among spaces from different GPT-4 matrices, finding correlations of .95 and higher (all ps $< .0001$).

Instead of comparing the similarity spaces, we can also compare more directly the matrices on which they are based, that is, compare the similarity matrices from GPT-4 with the one we can extract from the aggregate space shown in Fig. 1. To do this, we can use the so-called Mantel test (Mantel, 1967), which is specifically meant to determine the correlation between pairs of matrices with the same dimensions. This test showed that the similarity matrices obtained from GPT-4 all correlated highly with the similarity matrix based on the aggregate space from Douven et al. (2023): correlations were all close to .85 and highly significant (all ps $< .001$), with one exception for which $r = .78$ (which is still high).

These results strongly suggest that GPT-4 may help us solve both the generalizability problem and the cost problem described earlier: instead of running another experiment with a larger set of materials (e.g., including the items from Henley's set of mammals, if we are interested in extending the mammal space from Douven et al. (2023)), or of training a neural net in the manner of Sanders and Nosofsky (2020), we can simply ask GPT-4 to create a similarity matrix for our materials. At least at today's prices, the latter method would not only be significantly faster than the former but also much less expensive.

[1] To be more exact, we used the `MutlivariateStats.jl` package for the Julia language (Bezanson et al., 2017) to carry out classical MDS.

Alas, our "Just Ask GPT" approach quickly hit a roadblock when we asked it to generate a similarity matrix for the full set of Henley items. GPT-4 then let us know that our set of mammals is too large and that it is only able to generate similarity matrices for smaller sets. Although the inner workings of GPT-4 itself have not been described, it is known that transformer architectures depend upon an inner representation which typically scales quadratically with the number of tokens in both prompt and generation (Lin et al., 2022; Vaswani et al., 2017). Because each token attends to all other tokens, scaling the size of the desired output can easily become unmanageable. While we expect this issue to be overcome in future versions of GPT, at the moment no fix appears to be available.[2,3]

Of course, GPT-4 is not the only available commercial LLM. Another well-known LLM is Bard, Google's state-of-the-art transformer model. Rerunning the above experiment with this model gave disappointing results, however. We had to try repeatedly just to get a symmetric similarity matrix, and once we had a couple, they turned out to correlate poorly with the similarity matrix from Douven et al. (2023) (for details, see the Supplementary Materials). Other commercial LLMs, such as Cohere's Command or Llama2 70b, did not do any better. These findings motivated us to look at a simpler approach, which we describe in the following.

4 Spaces from Embeddings

At the core of the simpler approach is the idea that we can extract vector representations (or embeddings) from language models and compare these directly with human similarity judgments, instead of asking the model to generate a similarity matrix. For GPT-4 or Bard, this is unfortunately not possible: because they are not open source, there is no way to access the underlying vector representation of the input. The good news is that there are several similar models that *are* open source, although these models are known to be less powerful than the two aforementioned ones.

The landscape of open source language models is vast, presenting a broad array of architectures and functionalities. The ones that seem most relevant to our purposes are the so-called word embedding models (Almeida & Xexéo, 2019; Incitti et al., 2023; Mikolov et al., 2013). Word embedding models are trained to represent

[2] For instance, there is no way for users to change the attention mechanism used by GPT-4. But even if that were possible, it might be inadvisable, given that the currently available alternative attention mechanisms appear to yield less accurate results (Niu et al., 2021).

[3] Independent of the limitation mentioned here, there is a concern one may have about the "Just Ask GPT" approach, and about the approach proposed in this paper more generally. As a referee noted, if we had added "rose" to the set of mammals, GPT-4 might have rated the similarity between roses and each of the mammals in the set, which we could then have turned into a similarity space in which roses also would have been represented, together with the mammals, which intuitively would make little sense. Here, it is to be noted that there are general adequacy criteria for similarity spaces, one of which is the interpretability of the resulting dimensions (see, e.g., Borg & Groenen, 1999; Douven, 2021; Douven et al., 2022). It is safe to speculate that the said similarity space would not satisfy this criterion.

inputs which tend to co-occur in linguistic corpora, which as an objective already sounds somewhat similar to what similarity spaces aim to achieve.

One of the first major neural word embedding models was Word2Vec (Mikolov et al., 2013). The model was trained on the Google New Dataset, which contains about 6 billion words, and it aims to predict surrounding words for a given input word. It tries to achieve this objective by minimizing the so-called Skip–Gram loss function[4]:

$$J_{\text{Skip–Gram}} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log \Pr(w_{t+j} \mid w_t),$$

where T is the total number of words used for training, c is the context window size (i.e., the chunks of text preceding and following the center word), and $\Pr(w_{t+j} \mid w_t)$ is the probability of predicting word w_{t+j} given center word w_t.

To see whether Word2Vec might be able to help where GPT-4 had to pass, we started by obtaining the Word2Vec embeddings for each item in the materials from Douven et al. (2023), which yielded an array of twenty vectors. While it would not be wrong to think of the Word2Vec vector space as a Euclidean space, it is standard in the literature to measure distances among vectors in this space using the cosine distance (Manning, 2009; Manning & Schutze, 1999; Mikolov et al., 2013), which is defined as

$$1 - \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

for vectors $\mathbf{u}$ and $\mathbf{v}$. We followed standard practice and used this distance in our experiments.

We applied an MDS procedure to the thus obtained similarity matrix, which yielded the space shown in the left panel of Fig. 4. For a formal comparison with the similarity matrix from Douven et al. (2023), we conducted again both a Procrustes analysis and a Mantel test. While not as good as for the GPT-4 matrices, the results were still quite satisfactory, getting a correlation of .83 out of the Procrustes analysis and one of .61 from the Mantel test (both ps $< .0001$). And with Word2Vec, there is no impediment to obtaining similarities for larger sets of items. Indeed, we were able to get similarities for the full set of thirty items in the materials of Henley (1969) without any difficulty. Using the matrix of the larger set as input for an MDS procedure gave us the space shown in the right panel of Fig. 4.

These results are *especially* encouraging in view of the fact that Word2Vec dates back to 2013. It is reasonable to expect that newer models are able to give better results still. A popular later model is FastText, launched by Facebook Research

[4] The function stated here is simpler than the one proposed in Mikolov et al. (2013), but it is commonly used in many popular machine learning libraries. We also note that the Skip–Gram architecture is only one of two primary versions of Word2Vec. The other version is known as "Continuous Bag-of-Words" (CBOW). While Skip–Gram predicts context words from a target word, CBOW predicts a target word from a bag of context words. The difference between the two architectures is immaterial for our present purposes.

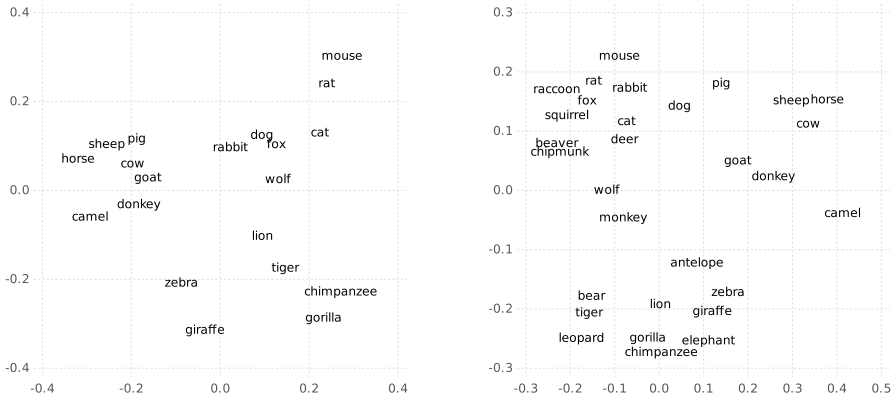

Fig. 4 Word2Vec results for the set of mammals used in Douven et al. (2023) (left) and for the full Henley set (right)

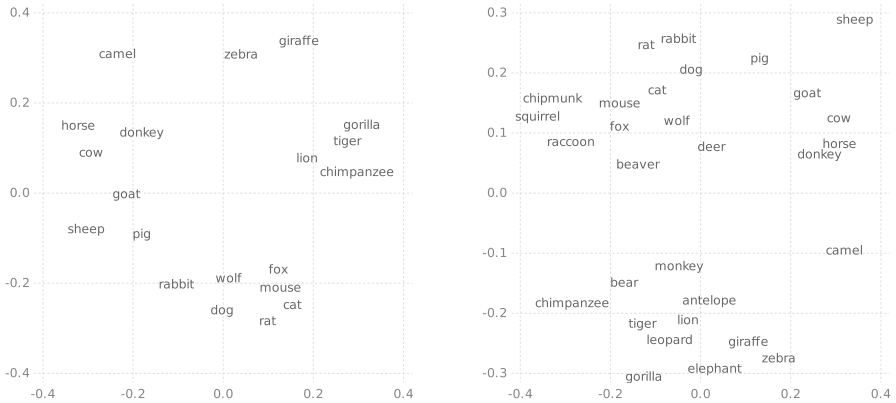

Fig. 5 FastText results for the set of mammals used in Douven et al. (2023) (left) and for the full Henley set (right)

in 2017 (Bojanowski et al., 2017), which preserves the overall approach of Word2Vec but also takes subword information into account. Rerunning the procedure just described for Word2Vec yielded results that were better indeed, with a correlation of .89 from the Procrustes analysis and one of .72 from the Mantel test (both $ps < .0001$). The similarity spaces for the limited and the full Henley set are shown in Fig. 5.

Thus, although already a bit dated, Word2Vec and FastText offer promising results. The correlations with the similarity matrix and aggregate space from Douven et al. (2023) are high enough to make us at least somewhat confident in their predictions of the locations of mammals in a more encompassing mammal space.[5]

[5] These results are also in line with those obtained for similar tasks in previous research. While testing on the UCLA Verbal Analogy Test (UCLA VAT), Snefjella et al. (2022) tasked Word2Vec to choose the

Although the correlations are lower than those achieved with GPT-4, there are several important advantages. Not only do we have the ability to inspect the model's internals, but we also obtain precise, context-independent embeddings for each word. This differs significantly from GPT-4, where the typical process of sampling the next token introduces variability into the results. Might we be able to get closer to those results by using more recent models that share GPT-4's architecture (i.e., transformer models) but are open source? We turn to these models in the next section.

5 Open Source LLMs

GPT-4 was seen to do an excellent job of predicting human similarity judgments for the mammals in the materials from Douven et al. (2023). However, currently we can only access GPT-4 via its prompt, and when we prompted it for similarity judgments for the full set of mammals from Henley (1969), it let us know that we were asking too much.

The word embedding models considered in the previous section are open source and, as a result, we have access to their layers. From these, we can readily obtain vector representations for all mammals in Henley's set, and from these representations we can then, in turn, derive similarities. It was seen that the similarities for all pairs of mammals from the set used by Douven and colleagues were, especially in the case of FastText, quite close to the human similarity judgments as documented by these authors.

Both Word2Vec and FastText count as old in the fast-moving field of AI. Since these models became available, language modeling has come to increasingly rely on the very different transformer architecture (Vaswani et al., 2017) and has progressively moved from modeling simple inputs (e.g., words as in Word2Vec) to much more complex ones like sentences and even longer text fragments. GPT-4 is a transformer model but offers, as seen, limited accessibility. However, there are open source transformer models which are not as powerful as GPT-4 but which are still much more modern than Word2Vec and FastText and which, precisely because they are open source, do allow us to access their layers directly, so that we can obtain vector representations (e.g., of all mammals in Henley's set) from them as easily as these could be obtained from Word2Vec and FastText.

It is at least a priori reasonable to expect the newer transformer models to improve on the older word embedding models. Perhaps the most innovative feature of the former type of architecture, and the feature that most clearly distinguishes it from Word2Vec and FastText, is the attention mechanism mentioned previously, which dynamically computes the importance of context words (the "surrounding" words)

Footnote 5 (continued)
correct analogy between two competing choices. The authors report that Word2Vec managed to correctly predict the right analogy with an accuracy of .69, compared to an accuracy of .84 achieved by human participants.

for a given center token (i.e., the token on which the attention is focused), where the context can have a length of hundreds, and in the latest models even thousands, of characters. This allows transformer models to pick up the significance of specific tokens, or sequences of tokens, within their context and to use this information to obtain improved embeddings. Besides architecture modification, the size, number of layers in the network, and training data of these models have also been greatly improved compared to the older models (Lin et al., 2022). In our scenario, since we are embedding individual words without any surrounding context, the full potential of the model's capabilities may not be realized. However, we can still anticipate that these models' ability to capture contextual relationships during training may result in more accurate embeddings, potentially aligning more closely with human similarity judgments.

The overarching goal of the newer models is not essentially different from that of older models, to wit, the prediction of tokens given some context. To achieve this goal, the first generation of breakthrough transformers mostly used masked word prediction, a technique that masks certain words in a sentence which the model should then try to predict from the masked words' context. Among the models that worked this way, BERT was a notable success, with widespread applications both in research and in industry (Devlin et al., 2018). Where N_{masked} is the number of tokens to predict, and x_{context} represents the surrounding tokens for a masked token x, the loss function that BERT was trained with can be specified as follows[6]:

$$\text{Loss}_{\text{MLM}} = -\frac{1}{N_{\text{masked}}} \sum_{i \in \text{masked}} \log \Pr(x_i \mid x^i_{\text{context}}).$$

While this loss function is in itself not very different from those that were used in the training of the older embedding models, a key training difference lies in the context, which, as said, can be quite large here.

As we did for the previous models, we passed the mammal names from Henley's set that were used in Douven et al. (2023) through the BERT model in order to obtain their embeddings. When we compared these embeddings with the human similarity judgments, the results proved to be quite a bit poorer than those we got from the older word embedding models. For instance, as can be seen in Fig. 6, BERT puts "horse" and "zebra" at a relatively large distance from each other, which would seem wrong: these words are semantically close in the judgment of anyone with a good command of the English language, which is also in accordance with what we got from the embedding models discussed earlier. That BERT underperforms relative to those models is confirmed by statistical tests, specifically yielding a Procrustes correlation of .43 and a Mantel correlation of .13.

Using variants of BERT that are popular in the machine learning community, notably DistilBERT and RoBERTa, yielded only marginally better results. It thus appears that the first generation of text generation models performs much worse

[6] The subscript MLM stands for "Masked Language Modeling."

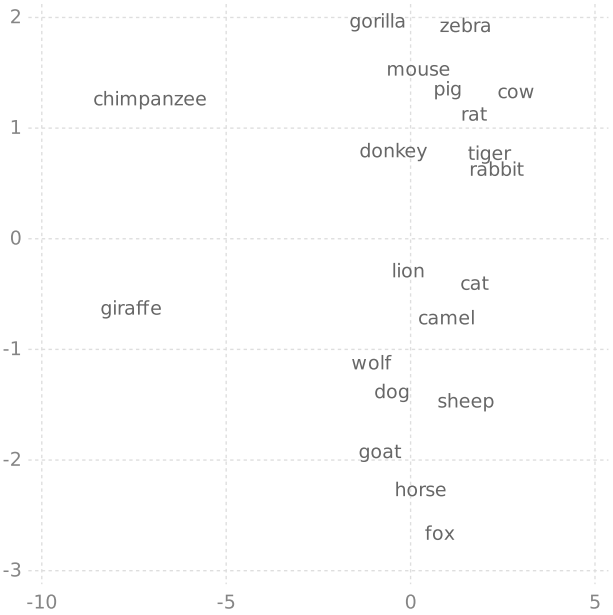

Fig. 6 BERT results for the set of mammals used in Douven et al. (2023)

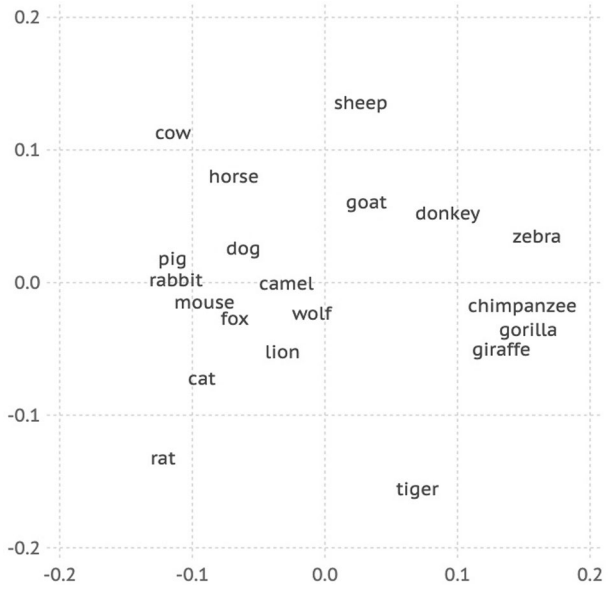

Fig. 7 Llama 3.2 1B Instruct results for the set of mammals used in Douven et al. (2023)

than older text-embedding models when it comes to predicting human similarity judgments.[7]

Given the rapid advancements in open source LLMs, we extended our analysis to more recent models. Specifically, we tested Meta's latest release, Llama−3.2 1B Instruct. While the results (see Fig. 7) demonstrated a notable improvement over older models, yielding a Procrustes correlation of .74 and a Mantel correlation of .3, they are still much worse than those from the older word embeddings.

These results are not just disappointing but also surprising: why would the newer models perform worse than Word2Vec and FastText on what appears to be a rather standard natural language processing (NLP) task? One possibility is that the larger context windows used by the newer models, as well as the attention mechanism, made these models attend to features that, while perhaps relevant to a host of NLP tasks, do not map well onto human similarity judgments.[8,9] This problem has come to be known as the "representation degradation problem" (Gao et al., 2019).

If the above speculation is along the right lines, then a solution to the suboptimal performance of BERT and its ilk might be to further fine-tune the models on sentences that are explicitly deemed semantically similar by humans. This approach *has* actually been taken by the NLP community, culminating in a training method known as "contrastive learning" (Qiu et al., 2022). Contrastive learning aims at minimizing the distance between similar tokens (words or sentences), much in the way in which Word2Vec architectures do this, where the similarity of the tokens is similarity as judged by human observers (i.e., in contrastive learning, the dataset on which a model is trained includes human similarity judgments; see Jaiswal et al., 2020).[10]

The combination of transformer models and contrastive learning appears quite promising and two widely used models resulting from this combination are all-MPNet-base-v2 and text-embedding-ADA-002 (Neelakantan et al., 2022; Reimers & Gurevych, 2019). The latter is one of OpenAI's latest embedding model; the former builds on the MPNet model from Microsoft, which follows an architecture quite similar to BERT, but having been fine-tuned with a set of sentence similarity judgments, the model achieved state of the art in sentence similarity upon its release.

[7] These results are consistent with the results reported in Snefjella et al. (2022), which concerned a different task, to wit, that of judging relational similarity. In that task, too, major LLMs failed to outperform Word2Vec; see also Ushio et al. (2021).

[8] A more technical (though still speculative) explanation relates the problem to the occasional overspecificity of the attention mechanism of transformer models, which, as argued in Demeter et al. (2020) and Ushio et al. (2021), can occur due to a limited number of hidden units exhibiting large activations, a phenomenon which often results in suboptimal performance in tasks that require calculating distances between network states (Sajjad et al., 2021). See on this also Timkey and van Schijndel (2021), whose authors refer to the phenomenon as the occurrence of "rogue dimensions," which they identify as key factors in the distortion or "obscuring" of representations in transformer models.

[9] While recognizing the strengths of LLMs, Lappin (2023) argues that these models still fall short in areas like natural language inference, analogical reasoning, and understanding figurative language. These limitations could extend to the challenge of accurately predicting human similarity judgments, which often involve nuanced understanding and contextual interpretation.

[10] For more on this, see Reimers and Gurevych (2019), who trained a variety of transformer models with contrastive learning, reaching state of the art results on the Hugging Face Model Evaluation Toolkit leaderboard.

Table 1 Summary of models, ranked on the basis of their performance (in terms of Mantel's *r*)

Model	Architecture	Training data	Dimensions	*r* (Mantel)	Cost	Generalizability
BERT	Transformer	BooksCorpus + Wikipedia	768	.14	Free	✓
RoBERTa	Transformer	WebText	768	.19*	Free	✓
DistilBERT	Transformer	BooksCorpus + Wikipedia	768	.26*	Free	✓
Llama3.2 1B Instruct	Transformer	Unknown	4096	.30**	Free	✓
text-embedding-ada-002	Transformer	Unknown	1536	.52	Low	✓
all-MPNet-base-v2	Transformer	Multi-source	768	.55**	Free	✓
Word2Vec	Skip–Gram	Google News	300	.61**	Free	✓
FastText	Skip–Gram	Common Crawl	300	.72**	Free	✓
GPT-4	Transformer	Unknown	Unknown	.92**	Low	×

Note: $*p < .05$, $**p < .001$

We subjected both models to the same test as the models discussed previously, meaning that we obtained the embeddings of the items that served as the materials for Douven et al. (2023) and turned these into a similarity matrix, again using the cosine distance. Here it appeared that the improved training method had indeed paid off, for in a comparison with the similarities from Douven et al. (2023) we found Procrustes correlations of .82 for both models and a Mantel score of.54 for all-MPNet-base-v2 and of.52 for text-embedding-ADA-002. But while this indicates a marked improvement over the earlier transformer models, the results still disappoint, not only absolutely speaking but also when compared especially with those we got from FastText.[11]

So again, why do the newer models so much worse in our tests than the older word embeddings like Word2Vec and FastText? It could be that, while BERT and related models are designed to capture deep contextual relationships within text, their full potential is unlocked only with *task-specific* fine-tuning. *With* such additional training, these newer models' representations might get better aligned with the similarity judgments we seek to model. Only further experimentation can tell whether this speculation is in the right direction. Note, however, that even if it is, such further training of BERT and related models could become costly and so whichever spaces we might be able to get from these models would no longer be free.

To conclude this section, Table 1 gives an overview of the results, ranking the various models on the basis of their performance. It is clear that GPT-4, currently the top language model, outperforms the other models by a lot, which could well

[11] We do not show the MDS models for the responses we got from all-MPNet-base-v2 and text-embedding-ADA-002. Interested readers are referred to the Supplementary Materials.

be due to the enormous resources which GPT-4 uses, as well as the latest research being directly applied. But while it is still low-cost (not entirely free), it was seen not to fully address the generalizability problem. It seems safe to speculate, however, that newer versions of this model that we can expect to see in the near future will not be limited in this way. Nevertheless, researchers working on conceptual spaces who need a similarity space right *now* may want to give FastText a try, provided an approximation of human similarity judgments is good enough for their purposes, and provided also that their items are text-based.

6 Conclusion

We have critically discussed the main methods for constructing similarity spaces as well as the prospects of using tools from artificial intelligence, like large language models and similar models, as an alternative method. Multidimensional scaling is still the golden standard for creating similarity spaces, but while it is a well validated approach, it is not without drawbacks, most notably, the high costs associated with extensive data collection and limited generalizability when applied to new sets of stimuli. These constraints hinder the method's usefulness, especially in disciplines operating under stringent budgetary limitations. The spatial arrangement method (SpAM) offers a less expensive alternative, allowing for the direct construction of similarity spaces by participants. However, this method also suffers from the generalizability problem. Moreover, SpAM is cognitively demanding, limiting the number of items that can be effectively processed, and it enforces a two-dimensional representation which may not always align with the dimensionality of the given similarity judgments.

Our research findings suggest that transformer models (such as GPT-4) and word embeddings may offer an alternative to generating similarity spaces that is *not* beset by the above challenges. Our experiments with these models have produced some similarity spaces that not only approximate one based on human judgments, but have also done so with remarkable speed (no need to get approval from an ethics committee, no tedious programming of a survey, no waiting time until enough participants have been recruited) and little (e.g., in the case of GPT-4) or no (e.g., in the case of FastText) costs. In a comparison with similarity data from Douven et al. (2023), GPT-4 clearly stood out. At the same time, we encountered limitations in the size of the item sets GPT-4 could process. The probably most practical recommendation coming from our research was that, for now, researchers who are looking for a way to create similarity spaces at no cost and with a potential to use a large set of items may be best off using FastText, at least if a moderately high correlation with human similarity judgments is enough for their purposes.

One limitation of our study is that we tested the models using only a single dataset, the Henley mammal set. One would hope that the computational methods discussed in this paper apply broadly, including to non-verbal domains and to concepts with affective dimensions (see, e.g., Stolier et al., 2018, 2020), but there is reason

to be cautious in this regard (see De Deyne et al., 2020).[12] Nevertheless, with the increasing capabilities of multimodal models, such as GPT-4, which already processes some images (albeit with limitations), it may soon be possible to assess the limitations of our approach (if any) empirically. Meanwhile, our findings provide an initial indication that large language models (LLMs) and word embeddings hold significant promise for addressing the dual challenges of generalizability and cost in the construction of similarity spaces-an issue that motivated our research.[13]

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics, 2*, 433–459.

Almeida, F., & Xexéo, G. (2019). Word embeddings: A survey. Preprint retrieved from http://arxiv.org/abs/1901.09069

Attarian, M., Roads, B. D., & Mozer, M. C. (2020). Transforming neural network visual representations to predict human judgments of similarity. Preprint retrieved from http://arxiv.org/abs/2010.06512

Bechberger, L., & Kühnberger, K. U. (2021). Grounding psychological shape space in convolutional neural networks. *International Conference on Software Engineering and Formal Methods, 2021*, 86–106.

Bendifallah, L., Abbou, J., Douven, I., & Burnett, H. (2023). Conceptual spaces for conceptual engineering? Feminism as a case study. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-023-00708-7

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review, 59*, 65–98.

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.

Borg, I., & Groenen, P. J. F. (1999). Modern multidimensional scaling: Theory and applications. *Journal of Educational Measurement, 40*, 277–280.

Bourdieu, P. (1989). Social space and symbolic power. *Sociological Theory, 7*(1), 14–25.

Castro, J. B., Ramanathan, A., & Chennubhotla, C. S. (2013). Categorical dimensions of human odor descriptor space revealed by non-negative matrix factorization. *PloS ONE, 8*(9), e73289.

Churchland, P. M. (2012). *Plato's camera*. MIT Press.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The small world of words english word association norms for over 12,000 cue words. *Behavior Research Methods, 51*, 987–1006.

[12] See in this connection also De Deyne et al. (2019), who derive similarities from word associations. As these authors argue, this approach may capture aspects of meaning grounded in perception and emotion better than text-based models do.

[13] We are greatly indebted to two anonymous referees for valuable comments on a previous version of this paper.

De Deyne, S., Cabana, A., Li, B., Cai, Q., & McKague, M. (2020). A cross-linguistic study into the contribution of affective connotation in the lexico-semantic representation of concrete and abstract concepts. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd annual conference of the cognitive science society* (pp. 2776–2782). Cognitive Science Society.

Deauvieau, J., Penissat, É., Brousse, C., & Jayet, C. (2014). Les catégorisations ordinaires de l'espace social français. *Revue Fran ç aise de Sociologie, 55*, 411–457.

Demeter, D., Kimmel, G., & Downey, D. (2020). Stolen probability: A structural weakness of neural language models. Preprint retrieved from https://arxiv.org/abs/2005.02433

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. Preprint retrieved from http://arxiv.org/abs/1810.04805.

Douven, I. (2016). Vagueness, graded membership, and conceptual spaces. *Cognition, 151*, 80–95.

Douven, I. (2021). Fuzzy concept combination: An empirical study. *Fuzzy Sets and Systems, 407*, 27–49.

Douven, I. (2023). The role of naturalness in concept learning: A computational study. *Minds & Machines, 33*, 695–714.

Douven, I. (2024a). The learnability of natural concepts. *Mind & Language*. https://doi.org/10.1111/mila.12523

Douven, I. (2024b). Social learning in neural agent-based models. *Philosophy of Science*. https://doi.org/10.1017/psa.2024.33

Douven, I., & Gärdenfors, P. (2020). What are natural concepts? A design perspective. *Mind & Language, 35*, 313–334.

Douven, I., Elqayam, S., Gärdenfors, P., & Mirabile, P. (2022). Conceptual spaces and the strength of similarity-based arguments. *Cognition, 218*, 104951.

Douven, I., Verheyen, S., Elqayam, S., Gärdenfors, P., & Osta-Vélez, M. (2023). Similarity-based reasoning in conceptual spaces. *Frontiers in Psychology, 14*, 1.

Gao, J., He, D., Tan, X., Qin, T., Wang, L., & Liu, T.-Y. (2019). Representation degeneration problem in training natural language generation models. Preprint retrieved from http://arxiv.org/abs/1907.12009.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. MIT press.

Gärdenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. MIT Press.

Gärdenfors, P., & Osta-Vélez, M. (2023). Reasoning with concepts: A unifying framework. *Minds & Machines, 33*, 451–485. https://doi.org/10.1007/s11023-023-09640-2

Gärdenfors, P., & Warglien, M. (2012). Using concept spaces to model actions and events. *Journal of Semantics, 29*, 487–519.

Gärdenfors, P., & Williams, M.-A. (2001). Reasoning about categories in conceptual spaces. *IJCAI, 2001*, 385–392.

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers, 26*, 381–386.

Han, S. J., Ransom, K. J., Perfors, A., & Kemp, C. (2024). Inductive reasoning in humans and large language models. *Cognitive Systems Research, 83*, 101155.

Henley, N. M. (1969). A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior, 8*(2), 176–184.

Incitti, F., Urli, F., & Snidaro, L. (2023). Beyond word embeddings: A survey. *Information Fusion, 89*, 418–436.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., & Makedon, F. (2020). A survey on contrastive self-supervised learning. *Technologies, 9*(1), 2.

Lappin, S. (2023). Assessing the strengths and weaknesses of large language models. *Journal of Logic, Language and Information.*[SPACE]https://doi.org/10.1007/s10849-023-09409-x

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open, 3*, 111–132.

Manning, C. D. (2009). *An introduction to information retrieval*. Cambridge University Press.

Manning, C. D., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Mantel, N. A. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research, 27*, 209–220.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26*, 1.

Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., Yuan, Q., Tezak, N., Kim, J. W., Hallacy, C., et al. (2022). Text and code embeddings by contrastive pre-training. Preprint retrieved from http://arxiv.org/abs/2201.10005.

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing, 452*, 48–62.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 87–108.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology, 43*, 25–53.

Nosofsky, R. M., Sanders, C. A., Gerdom, A., Douglas, B., & McDaniel, M. A. (2017). On learning natural-science categories that violate the family-resemblance principle. *Psychology Science*. https://doi.org/10.1177/0956797616675636

Okabe, A., Boots, B., Sugihara, K., & Chiu, S. N. (2000). *Concepts and applications of voronoi diagrams*. Wiley.

OpenAI. (2023). GPT-4 technical report. *ArXiv*, *abs/2303.08774*. https://api.semanticscholar.org/CorpusID:257532815

Osta-Vélez, M., & Gärdenfors, P. (2020). Category-based induction in conceptual spaces. *Journal of Mathematical Psychology*. https://doi.org/10.1016/j.jmp.2020.102357

Osta-Vélez, M., & Gärdenfors, P. (2022). Analogy as a search procedure: A dimensional view. *Journal of Experimental and Theoretical Artificial Intelligence*. https://doi.org/10.1080/0952813X.2022.2125081

Patel, R., & Pavlick, E. (2021). Mapping language models to grounded conceptual spaces. *International Conference on Learning Representations*.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science, 42*(8), 2648–2669.

Petitot, J. (1988). Morphodynamics and the categorical perception of phonological units. *Theoretical Linguistics, 15*, 25–72.

Qiu, R., Huang, Z., Yin, H., & Wang, Z. (2022). Contrastive learning for representation degeneration problem in sequential recommendation. *Proceedings of the fifteenth ACM international conference on web search and data mining*, 813–823.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Preprint retrieved from http://arxiv.org/abs/1908.10084.

Sajjad, H., Alam, F., Dalvi, F., & Durrani, N. (2021). Effect of post-processing on contextualized word representations. Preprint retrieved from http://arxiv.org/abs/2104.07456.

Sanders, C. A., & Nosofsky, R. M. (2020). Training deep networks to construct a psychological feature space for a natural-object category domain. *Computational Brain & Behavior*. https://doi.org/10.1007/S42113-020-00073-Z

Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika, 31*, 1–10.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1*, 54–87.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317–1323.

Snefjella, B., Ichien, N., Holyoak, K., & Lu, H. (2022). Predicting human judgments of relational similarity: A comparison of computational models based on vector representations of meaning. *Proceedings of the Annual Meeting of the Cognitive Science Society, 44*, 44.

Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences, 115*(37), 9210–9215.

Stolier, R. M., Hehman, E., & Freeman, J. B. (2020). Trait knowledge forms a common structure across social cognition. *Nature Human Behaviour, 4*, 361–371. https://doi.org/10.1038/s41562-019-0800-6

Timkey, W., & van Schijndel, M. (2021). All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. Preprint retrieved from http://arxiv.org/abs/2109.04404.

Ushio, A., Espinosa-Anke, L., Schockaert, S., & Camacho-Collados, J. (2021). BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? Preprint retrieved from http://arxiv.org/abs/2105.04949

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *Quarterly Journal of Experimental Psychology, 69*, 1996–2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł, & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 10.

Verheyen, S., & Storms, G. (2021). Whether the pairwise rating method and the spatial arrangement method yield comparable dimensionalities depends on the dimensionality choice procedure. *Methods in Psychology, 5*, 100060.

Verheyen, S., Voorspoels, W., Vanpaemel, W., & Storms, G. (2016). Caveats for the spatial arrangement method: Comment on Hout, Goldinger, and Ferguson (2013). *Journal of Experimental Psychology: General, 145*, 376–382.

Verheyen, S., White, A., & Storms, G. (2022). A comparison of the spatial arrangement method and the total-set pairwise rating method for obtaining similarity data in the conceptual domain. *Multivariate Behavioral Research, 57*, 356–384.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH ("Springer Nature").

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users ("Users"), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use ("Terms"). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com