TrendScape 1.0: Exploring the latent spaces of language models

Anonymous ACL submission

Abstract

This study introduces `TrendScape 1.0`, a method for exploring and visualising the latent conceptual spaces within language models. The `TrendScape` method maps words to latent space, enabling the exploration of relationships among concepts through a five-step process.By visualising conceptual "paths" between works and analysing the resulting networks, `TrendScape 1.0` offers insights into the AI conceptual understanding.

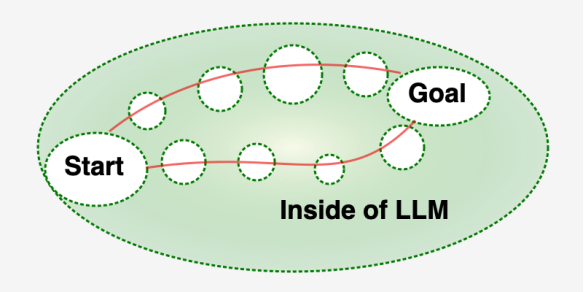


Figure 1: Conceptual diagram of `TrendScape`

1 Introduction

Research on concepts and representations has been a longstanding subject of inquiry within the field of philosophy [9]. As exemplified by Plato's "Allegory of the Cave", there exists a notion that the world perceived by humans is merely a shadow of reality. This implies that the concepts we comprehend are distinct from the actual world. In recent years, with the development of AI and deep learning models, this classical question has been revisited in novel forms. Huh et al. [5] have highlighted that the representation of concepts in AI models, particularly in language models, bears similarities to Plato's philosophy. They have termed this the "Platonic Representation Hypothesis".

In contrast, within the realm of linguistics, Saussure argued that concepts are defined through their "differences" from other concepts. Contemporary computational linguistics has built upon these ideas, advancing methods to represent words and sentences within vector spaces. For example, the skip-gram model learns distributed representations of words, enabling analogies such as King - Man + Woman = Queen.

Moreover, efforts to visualise and analyse the internal conceptual representations of language models have progressed [2]. These studies aim to elucidate the relationships among acquired concepts within models, enhancing the interpretability of AI.

This study proposes `TrendScape 1.0`, a new approach to analyse the latent conceptual space within language models through a graph-based interpretation. One advantage of a graph-based search is that the latent space of the language model is supposed to be continuous, so a simple continuous exploration ends up with morphing (e.g. GAN's glide): We expect to extract interpretable intermediate concepts by mapping the continuous latent space into a (discrete) graph structure which investigates how language models recognise and process concepts, serving as an exploration of their underlying "world models" [4].

It aligns with a broader line of research examining the structure of conceptual spaces in language models [10, 6], contributing to elucidating the mechanisms underlying AI systems' conceptual understanding and fostering the development of more refined natural language processing technologies.

2 Method

In this study, we propose a method called `TrendScape` to analyse the latent conceptual space in language models. This method maps input to the latent space, enabling the exploration and visualisation of relationships among concepts. The proposed approach consists of five primary steps. First, the input is aligned with its embedding rep-

resentation in the latent space and stored. Next, a sampling of the exploratory space is conducted, specifically extracting the vector space relationship from concept A to concept B, which is then structured into a network. The third step involves finding a conceptual path from A to B. Subsequently, a neighbourhood exploration is performed around the path, resampling the exploratory space. Finally, the relationships within the space are reconstructed and displayed.

This method employs two distinct approaches: the Word2Vec [8] version and the Sparse Autoecncoder (SAE) [12, 7] version. By combining these perspectives, it enables a more comprehensive analysis of concepts. The Word2Vec version utilizes the "word2vec-google-news-300" model's word embedding space. In this approach, "word2vec-google-news-300" serves as the model, and its linguistic space is treated as the word embedding space. Concepts for exploration are determined by topic modelling the words characterizing the text.

In contrast, the SAE version employs the Gemma SAE model, treating its intermediate layers as linguistic space. A notable feature of this method is its handling of context-laden words, managing approximately 16,000-dimensional features.

To verify the effectiveness of the proposed method, two approaches are employed. First, Word2Vec is used to construct relationships among word differences, testing whether concepts can be explored. Second, the possibility of concept exploration using tokens contextualized by sentences with SAE is examined. Through these methods, we aim to visualize and analyse conceptual representations within language models, exploring how models recognize and process concepts. Moreover, by comparing the Word2Vec and SAE approaches, we seek to unravel the mechanisms behind comprehensive conceptual understanding. These findings are expected to contribute to advances in natural language processing and improve the interpretability of AI systems.

3 Experiment

In this study, we conducted a series of experiments to validate the effectiveness of the `TrendScape` method for visualizing conceptual relationships in the latent space of language models. These experiments comprised two stages: basic validation using literary works and an extension experiment employing more advanced language models. All experiments were conducted using Google Colab A100.

3.1 Dataset

For the narrative works, We selected three representative stories from Grimm's Fairy Tales: "Hansel and Gretel", "Rapunzel", and "Snow White."

These works have different themes and are considered to offer a diverse conceptual space. The text of Grimm's Fairy Tales was downloaded from Project Gutenberg, and the HTML files were processed using BeautifulSoup to extract the text, which was then split by chapter

3.2 Experiment Setup

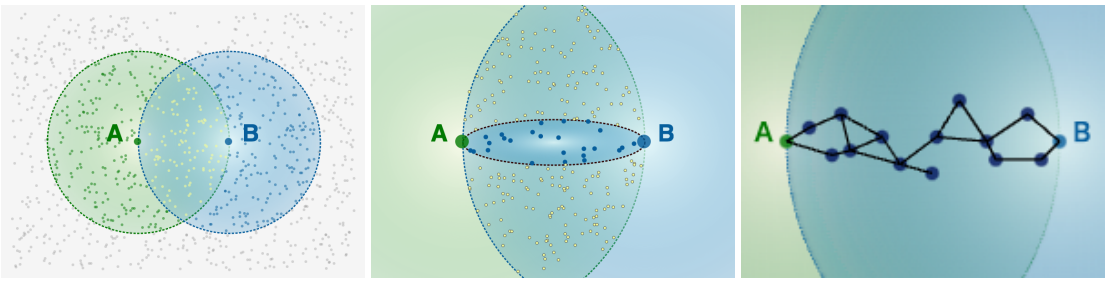
3.2.1 Topic Modelling

Topic modelling was used to model the topics within the document collection. Specifically, the LDA (Latent Dirichlet Allocation) model was applied to each work using the gensim library [3]. LDA is a probabilistic method for estimating latent topics in a document collection, and the following steps were performed in this study:

- Creation of the document-term matrix: A matrix recording the frequency of each word was generated from the preprocessed text.
- Setting the number of topics: Based on preliminary experiments, the number of topics for each work was set to 5.
- Training the LDA model: The model was optimized using Gibbs sampling.
- Visualization of topics: The resulting topics were visually represented using pyLDAvis.
- Extraction of characteristic words: The most important words in each topic were extracted to understand the characteristics of the works. The number of important words in each topic was set to 5.

3.2.2 Word Embeddings

To convert the extracted words from text analysis into a computable format, the pre-trained word embedding model word2vec-google-news-300 was used. This model was trained on the Google News dataset and represents approximately 3 million words and phrases in a 300-dimensional vector space. Using the gensim library, the word2vec-google-news-300 model was loaded, and each extracted word was converted into its corresponding 300-dimensional vector.



(a) Exploration Space Sampling (b) Exploration Space Re-Sampling (c) Concept Network Visualization

Figure 2: The TrendScape method involves several interconnected steps to analyse and visualize conceptual relationships in latent spaces. It begins by mapping and storing the input to its corresponding embedding representation in the latent space. Next, it samples the exploratory space by extracting the vector space between two concepts (A and B) and constructing a network based on the relationships in this vector representation (a). The method then finds the conceptual path from A to B within this network. To refine the results, it performs a neighbourhood exploration around the identified path, resampling the exploratory space in the process (b). Finally, the method reconstructs and displays the relationships within the explored space, providing a visual representation of the conceptual connections (c).

3.2.3 Application of the TrendScape Method

The TrendScape method, which forms the core of this study, visualises conceptual "paths" between works based on word embeddings. Its application consists of four key steps:

Construction of the Exploratory Space: To identify a set of words that conceptually bridge two works (source and target), we employed `word2vec-google-news-300`'s `closer_than` method for sampling. This method efficiently extracts words that are "close" to both the source and target. To balance computational efficiency and diversity, 500 words were randomly selected from the sampled set to form the initial exploratory space.

Network Structuring: To structure the relationships among the extracted words, we performed hierarchical clustering using the AgglomerativeClustering function from the scikit-learn library. Based on the clustering results, edges were created between words, with edge weights set as the square of the distance between words that computed as 1 - their cosine similarity. This approach resulted in a network structure where semantically similar words were strongly connected.

Pathfinding: On the constructed network, the shortest path from the source to the target was explored. While the shortest path algorithm from the networkx library served as the foundation, we implemented a custom heuristic function that prioritized semantic relevance. Specifically, a probabilistic search was conducted by considering the top-k (k=3) candidates at each step, mitigating the risk of falling into local optima.

Result Integration: To enhance stability and comprehensiveness, the process was repeated multiple times, and the results were integrated. The final network was visualized using the plotly library, allowing for intuitive understanding of the relationships between concepts.

Through these steps, we identified the sequence of words that act as conceptual bridges between the works and visualized their relationships.

3.3 Results

The results of the TrendScape visualisation of the concept space are shown in Figure 3. Only SNOWDROP with HANSEL AND GRETEL results are shown due to space constraints, for other results see Appendix A.1.

As mentioned earlier, naively transitioning across the conceptual space of an LLM could result in morphing based on superficial similarity. For example, when transitioning from "Iceland" to "Ireland," the transition may proceed through intermediate forms like "Iceland," "Ieeland," "Ipeland," or "Iqueland," which do not necessarily correspond to meaningful transitions in the latent space. In contrast, as shown in Figure 3, two semantically related concepts, such as "came" and "went," are connected by a path at a short distance on the graph. This suggests that TrendScape may provide a more appropriate visualization of the LLM's world model and how the LLM semantically interprets narratives.

Looking at this diagram, we can see that words like "winter", "snow", and "snows" form a distinct

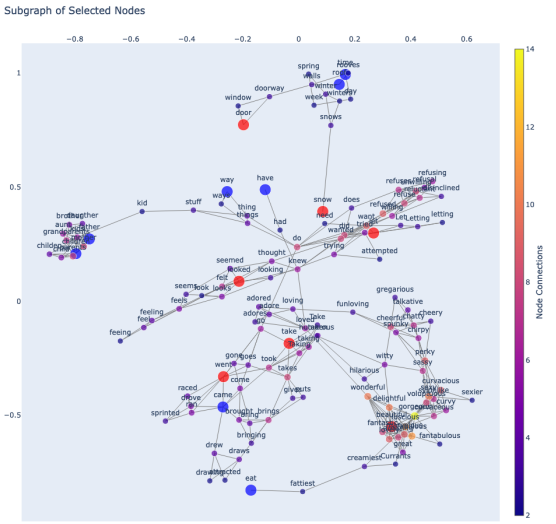


Figure 3: `TrendScape` between “SNOWDROP” and “HANSEL AND GRETEL”

cluster, which can be considered unique to SNOWDROP. On the other hand, expressions like “children” and “brother” appear to represent a cluster more characteristic of HANSEL AND GRETEL. Meanwhile, words such as “do” and “knew” are general terms likely to appear in both stories, serving as connecting points between many clusters. This result might offer insights into how we perceive the relationship between these two different stories.

3.4 Additional Experiment: Extension Using SAE

Although word2vec is a traditional approach, we also attempted to apply the latest language models to `TrendScape`. Specifically we employed a more advanced language model *Sparse Autoencoder (SAE)* [12, 7], to construct conceptual networks based on sentence-level embeddings.

The SAE in the current study is Gemma-2b based one. Using the output of layer 18 from Google’s large-scale language model Gemma-2b, each sentence was represented as a 16,384-dimensional feature vector. Each sentence was truncated to a maximum of 512 tokens before being input into the model.

In the case of Word2Vec, word groups are created first and then embedded. On the other hand, SAE processes sentences and generates feature spaces for each token, embedding them accordingly. SAE performs context-aware word embeddings, unlike Word2Vec. Additionally, while Word2Vec represents features across all 300 dimensions and uses cosine similarity, SAE treats each feature as independent. Therefore, cosine similarity is not considered suitable for comparison in SAE, and Wasserstein distance was used to measure the distance. Furthermore, SAE only connects edges when the same features are activated between two entities, indicating that they share the same characteristics.

As shown in Fig. 7, unlike previous cases, the same word appears multiple times. This is because SAE embeddings differ from Word2Vec and similar models, as they are context-aware embeddings. Even a single word like “forest” can have various roles depending on the context, and each occurrence of “forest” is visualized accordingly. The interpretation of the result is in

4 Limitation and Potential Risk

In this paper, we conducted only a qualitative analysis on a small amount of data only once. Applying the proposed method to larger datasets, such as those used in research, and quantitatively evaluating the results would allow for a more accurate validation of its effectiveness. Furthermore, this study analyzed only two models: Word2Vec and SAE. It is also important to apply this approach to a wider range of recently proposed LLMs. Since the analysis performed with this method was limited, deeper exploration of the conceptual spaces within LLMs is required. Additionally, as we have not explicitly compared our approach with existing methods, conducting such comparisons will be necessary in future work.

Our research focuses on LLM interpretability, but as it advances, exposing the model’s inner workings may enable malicious attacks. Thus, it’s crucial to both understand these mechanisms and prevent misuse

5 Conclusion

`TrendScape 1.0` introduces a novel method for visualizing and exploring latent conceptual spaces in language models. By mapping input to latent space and employing a five-step process, our approach offers insights into AI conceptual understanding. Experiments using Word2Vec and SAE models demonstrated how `TrendScape`’s enables us to analyse the concept space of language models.

References

[1] Dana H Ballard, Geoffrey E Hinton, and Terrence J Sejnowski. 1983. Parallel visual computation. *Nature*, 306(5938):21–26.

[2] Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[4] Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

[5] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

[6] Genji Kawakita, Ariel Zeleznikow-Johnston, Naotsugu Tsuchiya, and Masafumi Oizumi. 2024. Gromov–wasserstein unsupervised alignment reveals structural correspondences between the color similarity structures of humans and large language models. *Scientific Reports*, 14(1):15917.

[7] Dmitrii Kharlapenko, neverix, Neel Nanda, and Arthur Conmy. 2024. Self-explaining sae features. Accessed: 2024-10-16.

[8] Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[9] Jun Otsuka. 2024. What machine learning tells us about the mathematical structure of concepts.

[10] Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. 2024. The geometry of categorical and hierarchical concepts in large language models. *arXiv preprint arXiv:2406.01506*.

[11] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

[12] Adly Templeton. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet.

A Appendix

A.1 Visualization for Other Data

In the main text, only the visualization results for RAPUNZEL and SNOWDROP were shown. Here, we also present the results for HANSEL AND GRETEL with RAPUNZEL (Fig. 4), as well as RAPUNZEL with SNOWDROP (Fig. 5).

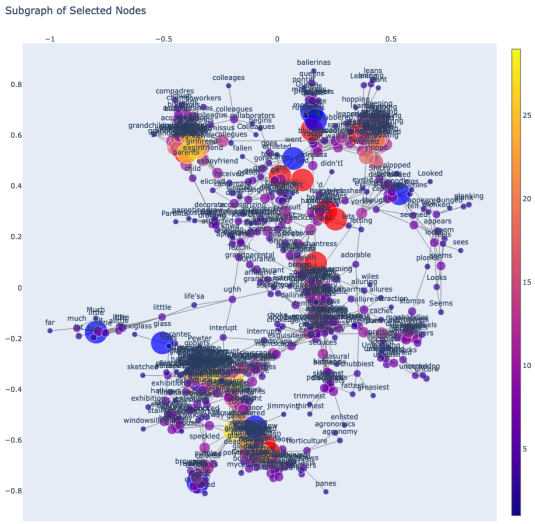


Figure 4: TrendScape between "HANSEL AND GRETEL" and "RAPUNZEL"

A.2 Visualization for SAE

The visualization of SAE is shown in .

We extracted the word "forest" from the texts of HANSEL AND GRETEL and RAPUNZEL to explore the differences in their respective concepts. While Word2Vec assigns unique features to individual words, it is important to note that these instances of "forest" carry features that reflect the context of the respective narratives.

As a result, we identified words like "beautiful," "dangerous," and "water" that directly describe the concept of a forest. Interestingly, words related to people, such as "princess," "I," and "me," also appeared quite frequently. This may suggest that the exploration of concepts within the context of these stories brings forth these person-related terms.

A.3 Application to Research Paper Data

In an exploratory effort to test the versatility of our method, we applied TrendScape to research paper abstracts. While the primary focus of our study was on literary works, we hypothesized that the method might offer insights into the conceptual relationships within academic literature as well. This extension, although preliminary, might suggest avenues for future research in scientometric analysis and interdisciplinary knowledge discovery. As the paper data, we used the abstracts of two papers: "Learning representations by back-propagating errors" [11] and "Parallel visual computation" [1]. Fig. 7 presents the visualization results of the re-

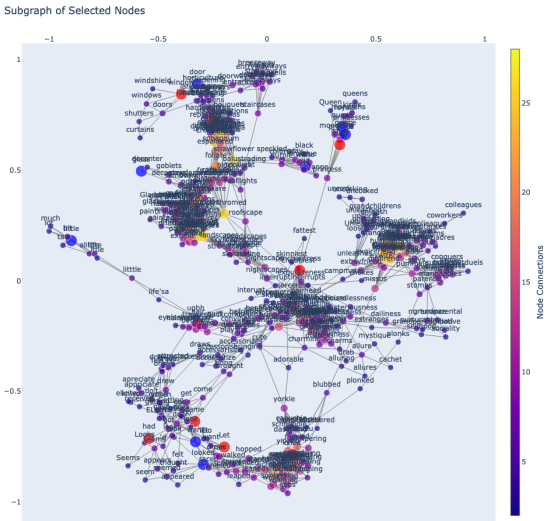


Figure 5: TrendScape between “RAPUNZEL” and “SNOWDROP”

lationships between words within the paper. The analysis of these results will be left for future work.

A.4 Checklists

Grimm’s Fairy Tales are out of copyright and in the public domain. All data used for this study is in English. We used ChatGPT and Claude for translation, proofreading, and coding assistant.

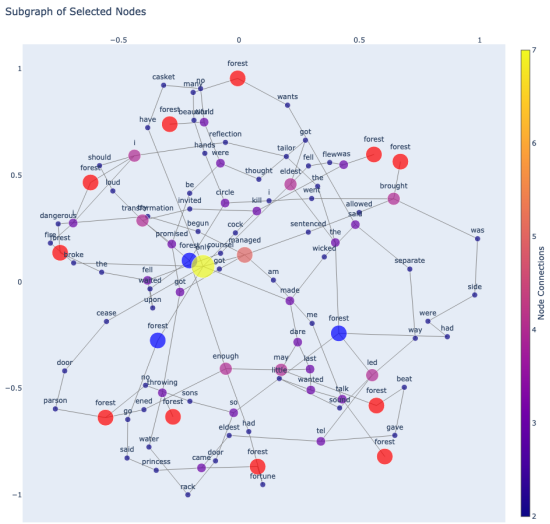


Figure 6: Trendscape using SAE latent space

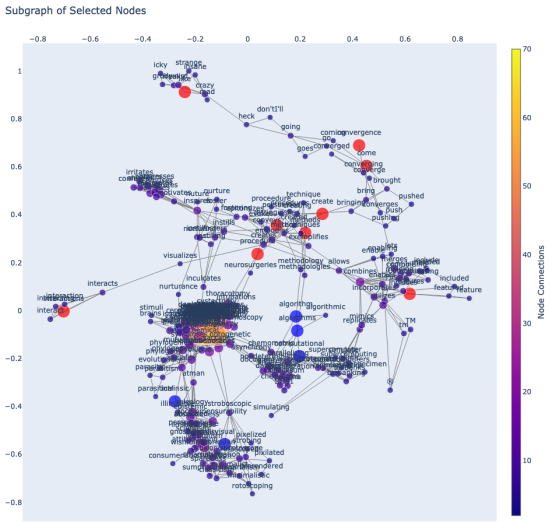


Figure 7: TrendScape between the articles