# Latent Manipulator: Steerable embedding visualization through concept-guided manipulation

Shivam Raval*
sraval@g.harvard.edu
Harvard University
Cambridge, USA

Kevin Dunnell*
dunnell@media.mit.edu
Massachusetts Institute of Technology
Cambridge, USA

Andrew Lippman
lip@media.mit.edu
Massachusetts Institute of Technology
Cambridge, USA

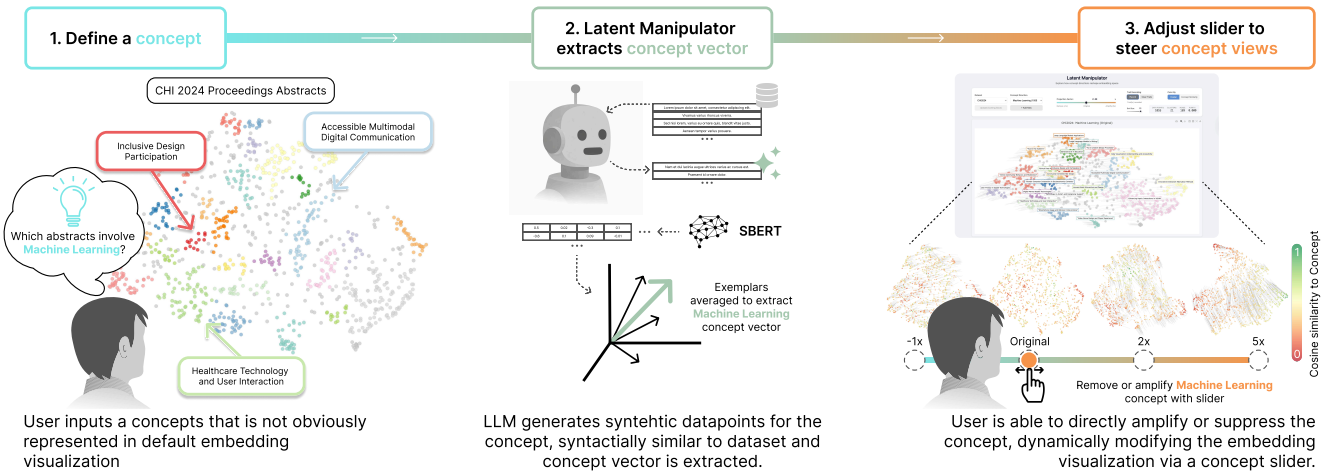


**Figure 1: Overview of concept-guided embedding manipulation. Users define a concept of interest, which Latent Manipulator converts into a semantic direction in embedding space. Interactively adjusting this direction reweights embeddings to generate alternative, concept-aligned visual organizations of the same data.**

## ABSTRACT

Embedding visualization systems reduce complex data to projection maps that support exploration and sensemaking, but only support a single static view of the data. Although changing DR methods and hyperparameters yields different projections, these views remain constrained by the concepts most salient in the upstream embeddings. We introduce concept-guided embedding manipulation, which enables users to steer visualizations by amplifying or suppressing user-defined concepts directly in embedding space. The method derives concept vectors from contrastive exemplars and reweights embeddings to generate alternative projections without retraining models. Through case studies with a collection of CHI research papers and a corpus of literary text, we show how steering can surface cross-domain methodological structure and reveal narrative patterns that are muted in default views of the datasets. Our approach helps users align embedding visualizations with their goals and exploration of multiple meaningful facets of the same dataset.

## KEYWORDS

Embedding visualizations, dimensionality reduction, steerable visualizations



## 1 INTRODUCTION

Embedding visualizations let data scientists, ML engineers, and an emerging set of consumer users "see" high-dimensional data by compressing embeddings to 2D/3D scatterplots [5, 9, 18, 32]. These maps provide a summary view and support visual exploration [6, 28]. However, current systems such as WizMap [34] and Embedding Atlas [26] typically provide a single, fixed perspective. Although DR choices can slightly vary the layout, the dominant organization is largely set by the embedding model and the concepts it encodes most strongly.

Yet the same embeddings can support multiple, equally valid organizations that better match different goals. Users can navigate an embedding map, but most systems offer limited ability to steer the space to surface patterns aligned with a specific objective. In open-ended settings, different users bring different lenses: for a literary corpus, one reader may want characters and places, while another may want themes or affect. If the embedding foregrounds a different axis (e.g., syntax), a single default view may fit neither intent.

To address this limitation, we introduce concept-guided embedding manipulation, which reorganizes embeddings by modulating the influence of user-defined concepts (Figure 1). A language model generates contrastive exemplars, which an embedding model encodes into a concept vector representing a semantic direction. By increasing or decreasing each item's projection along this direction,

the method produces alternative, concept-aligned views of the same data.

We instantiate this method in ***Latent Manipulator***, a prototype that integrates concept-guided manipulation into an embedding visualization workflow. A concept slider enables users to generate alternative, plausible organizations on demand. Through case studies on academic abstracts and literary text, we show how steering surfaces semantic structure that exists in the data but is not foregrounded in standard views. These examples suggest that steerable embedding visualizations can broaden what embedding maps express and support more intentional exploration.

## 2 RELATED WORK

**Dimensionality reduction and embedding visualization systems.** Dimensionality reduction techniques enable visual inspection of high-dimensional embeddings by projecting them into 2- or 3-D space. While linear methods such as PCA [24] preserve variance, nonlinear methods such as t-SNE [32], and UMAP [18] better preserve local neighborhoods, with UMAP offering improved scalability and extensions such as Parametric UMAP [29] enabling consistent projections for evolving datasets. Interactive systems such as the Embedding Projector [31] established core exploration techniques; subsequent systems like Nomic's Atlas [5], Latent Lab [6], WizMap [34], Soot [9], and Embedding Atlas [26] improve scalability for large datasets and provide additional interactive exploration capabilities. Despite their widespread use, these systems produce a single projection that reflects the dominant structure of the embedding space, treating this representation as fixed and offering limited mechanisms for users to reshape the underlying space itself [12].

**Semantic axes and interaction.** Prior work has explored user-defined semantic axes to steer projections. Systems such as InterAxis [14] and the semantic axes feature in the Embedding Projector [31] allow users to define conceptual directions, while others compare alternative embedding spaces or temporal evolution [3, 15]. However, these approaches typically reorient the visualization around a selected concept, thereby collapsing other semantic structures, limiting users' ability to balance multiple analytical priorities. In contrast, our approach treats concepts as continuously adjustable filters on the embedding space: users can amplify or attenuate specific concepts while maintaining the relative structure induced by other latent factors. This interaction model aligns with principles of direct manipulation interfaces [30], enabling users to iteratively explore parameter spaces through immediate, reversible feedback and align representations with their mental models and task-specific priorities [10].

## 3 CONCEPT-GUIDED EMBEDDING MANIPULATION

Embedding visualization systems typically encode data into high-dimensional embeddings, apply dimensionality reduction, and render the result for interactive exploration. While prior work [6, 26, 28, 34] adds rich interaction at the visualization stage, the embedding space itself is usually treated as fixed. We intervene earlier in this pipeline by letting users directly manipulate embeddings before projection.

### 3.1 Constructing concept vectors

A concept vector encodes a semantic direction in embedding space corresponding to a user-specified concept. To define a concept $c$, the user provides a short natural language description (e.g., "machine learning" or "morality"). A language model then generates two contrastive exemplar sets: a positive set $\mathcal{P}_c$ of texts that strongly express the concept, and a negative set $\mathcal{N}_c$ of texts that are devoid of it. These exemplars are encoded by the same pretrained embedding model $f$ used to embed the corpus. We then estimate the concept direction via mean-difference [36],

$$\mathbf{v}_c = \mu(\mathcal{P}_c) - \mu(\mathcal{N}_c), \tag{1}$$

where $\mu(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} f(x)$ is the mean embedding of a set. The normalized vector $\hat{\mathbf{v}}_c$ is the concept direction. Each data point's alignment with the concept is a scalar quantified by its projection coefficient $s_{i,c} = \mathbf{e}_i^\top \hat{\mathbf{v}}_c$, which is visualized as a concept similarity score in the interface. Because the exemplars are generated rather than drawn from the corpus itself, concept vectors can be defined for any concept expressible in language, without requiring manual annotation or additional concept-specific training data.

### 3.2 Generating base embedding visual

Input items (e.g., a collection of documents or sentences extracted from a corpus of text) are mapped to vectors using a pretrained embedding model (SentenceBERT for text in our case [25]). Dimensionality reduction methods such as UMAP [18] or t-SNE [32] then project embeddings into 2D/3D for visualization. These projections support interactive exploration, while varying DR settings typically yields small variations of the same dominant organization. We insert a concept-guided manipulation stage between embedding generation and DR so users can amplify or attenuate semantic concepts in embedding space and generate multiple alternative views from the same data.

### 3.3 Embedding manipulation method

Prior work suggests many semantic concepts correspond to approximately linear directions in embedding spaces [1, 13, 23]. Concept directions can be estimated from contrastive examples and added to or removed from embeddings without retraining [4, 7, 8, 20, 27]. Given a corpus $\mathcal{X} = \{x_i\}_{i=1}^N$ and embedding model $f : \mathcal{X} \to \mathbb{R}^d$, each item has embedding $\mathbf{e}_i = f(x_i)$. For a concept $c$, we collect contrastive exemplars: positives $\mathcal{P}_c$ and negatives $\mathcal{N}_c$, and compute a mean-difference concept direction $\mathbf{v}_c = \mu(\mathcal{P}_c) - \mu(\mathcal{N}_c)$, normalized to $\hat{\mathbf{v}}_c$. We define a user-controlled parameter $\alpha_c$ and manipulate each embedding via $\tilde{\mathbf{e}}_i(\alpha_c) = \mathbf{e}_i + \alpha_c(\mathbf{e}_i^\top \hat{\mathbf{v}}_c)\hat{\mathbf{v}}_c$. Positive $\alpha_c$ amplifies alignment with $c$, $\alpha_c \in [-1, 0)$ attenuates it, and $\alpha_c = -1$ removes the linear contribution of $c$. Unlike projection-based semantic axes that collapse the space to concept directions, our method adjusts concept strength while retaining the remaining structure. We then apply dimensionality reduction (UMAP in our system) to the manipulated embeddings to generate the updated 2D visualization as users adjust sliders.


**Figure 2: Steering on CHI 2024 proceedings abstracts along the *Machine Learning* concept direction. Columns show the same embedding projection with the concept removed, original, and amplified (2.0×, 5.0×). Top row colors points by cosine similarity to the concept (0=red, 1=green); bottom row shows discrete topic clusters. Callouts and correspondence lines track a focal paper ("ConverSense") as it moves through the space and changes cluster association while clusters evolve and consolidate with increasing concept weight.**

## 3.4 Quantifying manipulation success using cluster purity

To quantitatively assess how concept-guided manipulation reshapes semantic organization, we evaluate cluster purity of the clusters before and after applying concept-guided transformations. We use HDBSCAN [17] to cluster the 2D projections and obtain cluster labels. Let $C = \{c_1, c_2, \ldots, c_K\}$ denote the set of valid cluster labels (excluding noise), and let $\mathcal{E}_k = \{\tilde{\mathbf{e}}_i : \ell_i = c_k\}$ represent the set of embeddings assigned to cluster $c_k$, where $\ell_i$ is the cluster label for embedding $\tilde{\mathbf{e}}_i$. For each cluster $c_k$, we compute the centroid $\mu_k$ as the mean of all embeddings in that cluster. The cluster-wise purity $c_k$ is the average cosine similarity between each embedding in the cluster and the cluster centroid:

$$p_k = \frac{1}{|\mathcal{E}_k|} \sum_{\tilde{\mathbf{e}}_i \in \mathcal{E}_k} \frac{\tilde{\mathbf{e}}_i^\top \boldsymbol{\mu}_k}{\|\tilde{\mathbf{e}}_i\|_2 \|\boldsymbol{\mu}_k\|_2}. \tag{2}$$

This cluster-wise purity measures the degree to which items clustered together in the visualization share a similarity within themselves, providing a grounded proxy for semantic coherence. Higher purity values indicate that embeddings within a cluster have a low variance in embedding similarity and are more consistently and semantically clustered. The final purity score is calculated as a weighted average across all clusters, where the weights correspond to the cluster sizes. This metric provides a single scalar value in the range $[0, 1]$ that quantifies the global concept alignment of the clustering. An increase in cluster purity upon enhancing a concept suggests a better coherent semantic organization in the manipulated embeddings.

# 4 CASE STUDIES

We illustrate how *Latent Manipulator* supports task-aligned exploration by generating alternative, plausible organizations of the same dataset. Across both case studies, steering makes specific interpretive lenses more visually legible.

## 4.1 CHI Research Papers

We visualize 1,055 CHI 2024 paper abstracts embedded using SentenceBERT and projected with UMAP, as shown in Figure 2. In the baseline view, papers cluster primarily by research topics (e.g., Generative AI in Education, Digital Mental Health Technologies, Innovative Interactive Fabrication Methods, etc.), reflecting dominant topical signals in the embeddings. While useful for overview, this organization obscures cross-cutting conceptual and methodological patterns.

*4.1.1 Manipulation along Machine Learning concept direction.* To examine the methodological structure, we amplify the concept of *Machine Learning* (Figure 2). In the baseline view, ML-heavy contributions are often absorbed into domain clusters because their abstracts discuss application settings alongside methods. For instance, "ConverSense: An Automated Approach to Assess Patient-Provider Interactions using Social Signals" describes automated assessment pipelines and social-signal modeling, yet initially sits within a mental health / care-oriented neighborhood [2]. "Towards Building Condition-Based Cross-Modality Intention-Aware Human-AI Cooperation under VR Environment" similarly foregrounds modeling and intention-aware cooperation within a VR context [2]. "GustosonicSense: Towards understanding the design of playful

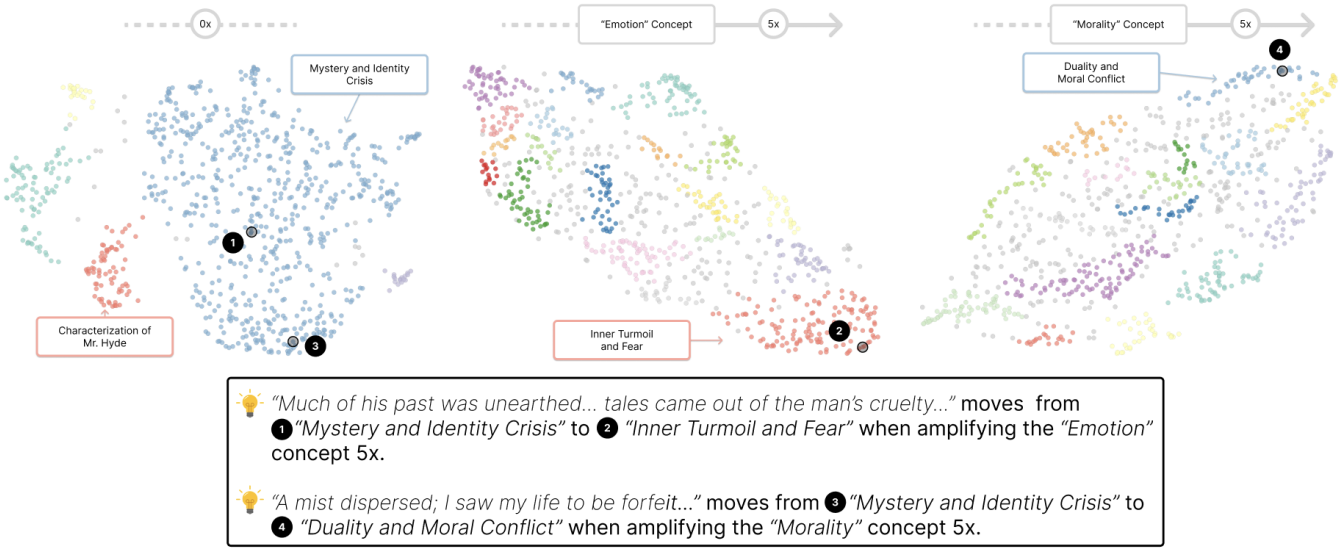


**Figure 3: Concept steering on sentence-level embeddings from *The Strange Case of Dr. Jekyll and Mr. Hyde*. The original embeddings (left) are compared with views produced by amplifying the Emotion (middle) and Morality (right) concepts (5.0×).**

gustosonic eating experiences" begins as fabrication/experience-oriented, but its ML-relevant language becomes more consequential under this lens [33]. "Time2Stop: Adaptive and Explainable Human-AI Loop for Smartphone Overuse Intervention" likewise shifts toward this methodological region when ML is emphasized [22]. Across examples, steering surfaces a methodological layer that is present but not foregrounded in the default, topic-first organization.

*4.1.2 Manipulation along Agency & Control concept direction.* Agency-related language is diffuse in the baseline view: papers that foreground user choice, autonomy, or self-regulation appear scattered among domain clusters. For example, "Choosing What You Want Versus Getting What You Want: An Experiment with Choice in Video Ad Placement" explicitly centers choice and control in algorithmic ad delivery, yet appears near domain-specific clusters in the default layout [11]. Similarly, ""I finally felt I had the tools to control these urges": Empowering Students to Achieve Their Device Use Goals With the Reduce Digital Distraction Workshop" emphasizes empowerment and control over device use goals, but initially reads as a wellbeing intervention within a topical neighborhood [16]. After amplifying *Agency & Control*, these papers relocate toward a more coherent region with other work discussing user-directed outcomes and constraints in interactive systems. Notably, this lens can also increase proximity to papers framed around agentic interaction paradigms, e.g., "Apple's Knowledge Navigator: Why Doesn't that Conversational Agent Exist Yet?" [21], "Join Me Here if You Will: Investigating Embodiment and Politeness Behaviors When Joining Small Groups of Humans, Robots, and Virtual Characters" [35], and "Empowering Calibrated (Dis-)Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style" [19], highlighting that "agency" may operationalize through multiple adjacent framings (choice/control, self-regulation, or agent interaction).

## 4.2 Literature

We embed 849 sentence-level excerpts from *The Strange Case of Dr. Jekyll and Mr. Hyde*. In the baseline visualization, clusters align with a small set of plot- and character-centric groupings (e.g., identity crisis, Hyde characterization, dual identity, experimentation), reflecting dominant narrative signals.

*4.2.1 Manipulation along emotion direction.* Several emotionally intense passages initially appear within broader narrative clusters, most commonly *Mystery and Identity Crisis*, despite containing strong affective language. Examples include "...that insurgent horror was knit to him closer than a wife, closer than an eye; ... lay caged in his flesh...", "...something abnormal and misbegotten... seizing, surprising and revolting...", and "Much of his past was unearthed... tales came out of the man's cruelty...". Under 5× amplification of the *Emotion* concept, these excerpts relocate into a more coherent region labeled *Inner Turmoil and Fear* (middle of Figure 3). This reorganization foregrounds emotional intensity as an organizing principle, rather than narrative function or character role.

*4.2.2 Manipulation along morality direction.* Amplifying the *Morality* concept produces a complementary reorganization. Passages that articulate vice, guilt, or ethical conflict, such as "At that time my virtue slumbered; my evil..." and "A mist dispersed; I saw my life to be forfeit...", initially appear in clusters including *Characterization of Mr. Hyde* and *Mystery and Identity Crisis*. With morality amplified, these passages migrate toward regions labeled *Inner Turmoil and Guilt* and *Duality and Moral Conflict* (right of Figure 3. This view surfaces an ethical structure that is present across the text but not primary in the default, plot-driven organization.

These qualitative observations are complemented by Figure 4, which shows that increasing concept amplification is associated with higher cluster purity along the targeted analytical dimension

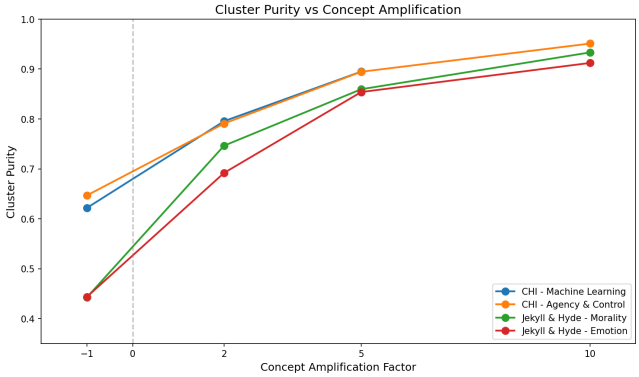


**Figure 4: Cluster purity versus concept amplification for CHI 2024 abstracts and *The Strange Case of Dr. Jekyll and Mr. Hyde*. Amplification is associated with increased purity along the targeted concept dimension, with concept-dependent variation.**

across both datasets. This complements our qualitative observations, reflecting a systematic reweighting of semantic structure rather than isolated visual effects.

## 5 CONCLUSIONS

We introduce concept-guided embedding manipulation, enabling users to interactively explore multiple meaningful organizations of the same dataset by amplifying or attenuating user-defined concepts. Through case studies on CHI abstracts and literary text, we show that concept-specific views can surface structure that is muted in default projections, such as methodological groupings across application domains and ethical or affective organization in narrative text.

Several limitations and extensions remain. Manipulation depends on the embedding model's representational capacity; concepts that are weakly encoded may not yield coherent reorganization, and even well-represented concepts may surface through a broader latent framing (e.g., in our CHI data, *Machine Learning*–aligned abstracts concentrate within a larger "Human-AI Interaction and Design" region), which can differ from a user's intended interpretation. Multi-concept steering can also be cognitively challenging, and concept cosine similarity views help, but do not fully resolve this trade-off. Finally, manipulation plus dimensionality reduction can be expensive on large datasets; in practice, precomputing common concepts enables responsive interaction.

Several future directions seem promising. A user study comparing how users interact with and explore data using static versus steerable visualizations can surface insights that guide future iterations of *Latent manipulator*. Concept-based visual re-organization is a rich avenue for further exploration, as datasets may contain concepts with binary, continuous, or hierarchical structure. Exposing these structures at multiple levels of abstraction could further engage users and reveal patterns that would remain inaccessible in a single embedding view.

## REFERENCES


[1] Steven Abreu, Sina Richter, and Abel Guimarães. 2024. Steering Large Language Models using Conceptors: Improving Addition-Based Activation Engineering. arXiv:2410.16314 [cs.CL] https://arxiv.org/abs/2410.16314

[2] Manas Satish Bedmutha, Anuujin Tsedenbal, Kelly Tobar, Sarah Borsotto, Kimberly R Sladek, Deepansha Singh, Reggie Casanova-Perez, Emily Bascom, Brian Wood, Janice Sabin, Wanda Pratt, Andrea Hartzler, and Nadir Weibel. 2024. ConverSense: An Automated Approach to Assess Patient-Provider Interactions using Social Signals. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 448, 22 pages. https://doi.org/10.1145/3613904.3641998

[3] Angie Boggust, Brandon Carter, and Arvind Satyanarayan. 2022. Embedding Comparator: Visualizing Differences in Global Structure and Local Neighborhoods via Small Multiples. arXiv:1912.04853 [cs.HC] https://arxiv.org/abs/1912.04853

[4] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. arXiv:1607.06520 [cs.CL] https://arxiv.org/abs/1607.06520

[5] Brandon Duderstadt, Vincent Giardina, Andriy Mulyar, Ben Schmidt, Yuvanesh Anand, Richard Guo, Gegi Janiashvili, Lakshay Kansal, Paige Lee, Robert Lesser, Wilson Marcilio Jr, Aaron Miller, and Adam Treat. 2024. *Atlas: Scalable Information Cartography*. Technical Report. Nomic AI. https://static.nomic.ai/atlas_tech_report.pdf Accessed: 2025-04-18.

[6] Kevin Dunnell, Trudy Painter, Andrew Stoddard, and Andy Lippman. 2023. Latent lab: Large language models for knowledge exploration. *arXiv preprint arXiv:2311.13051* (2023).

[7] Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding Undesirable Word Embedding Associations. arXiv:1908.06361 [cs.CL] https://arxiv.org/abs/1908.06361

[8] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. arXiv:1903.03862 [cs.CL] https://arxiv.org/abs/1903.03862

[9] Jake Harper and Amol Kapoor. 2024. SOOT: A Visual-First Spatial File System. https://www.soot.com/. Accessed: 2025-04-18.

[10] Jeffrey Heer and Ben Shneiderman. 2012. Interactive dynamics for visual analysis. *Commun. ACM* 55, 4 (2012), 45–54.

[11] Silas Hsu and Karrie Karahalios. 2024. Choosing What You Want Versus Getting What You Want: An Experiment with Choice in Video Ad Placement. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 737, 9 pages. https://doi.org/10.1145/3613904.3642869

[12] Zeyang Huang, Daniel Witschard, Kostiantyn Kucher, and Andreas Kerren. 2023. VA + Embeddings STAR: A State-of-the-Art Report on the Use of Embeddings in Visual Analytics. *Computer Graphics Forum* (2023). https://doi.org/10.1111/cgf.14859

[13] Yibo Jiang, Bryon Aragam, and Victor Veitch. 2024. On the Origins of Linear Representations in Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, Vol. 235. PMLR. https://proceedings.mlr.press/v235/jiang24q.html

[14] Hannah Kim, Jaegul Choo, Haesun Park, and Alex Endert. 2016. InterAxis: Steering Scatterplot Axes via Observation-Level Interaction. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 131–140. https://doi.org/10.1109/TVCG.2015.2467615

[15] Yang Liu, Eunice Jun, Qisheng Li, and Jeffrey Heer. 2019. Latent Space Cartography: Visual Analysis of Vector Space Embeddings. *Computer Graphics Forum (Proc. EuroVis)* (2019). https://doi.org/10.1111/cgf.13672

[16] Ulrik Lyngs, Kai Lukoff, Petr Slovak, Michael Inzlicht, Maureen Freed, Hannah Andrews, Claudine Tinsman, Laura Csuka, Lize Alberts, Victoria Oldemburgo De Mello, Guido Makransky, Kasper Hornbæk, Max Van Kleek, and Nigel Shadbolt. 2024. "I finally felt I had the tools to control these urges": Empowering Students to Achieve Their Device Use Goals With the Reduce Digital Distraction Workshop. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 251, 23 pages. https://doi.org/10.1145/3613904.3642946

[17] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software* 2, 11 (2017), 205. https://doi.org/10.21105/joss.00205

[18] Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [stat.ML] https://arxiv.org/abs/1802.03426

[19] Luise Metzger, Linda Miller, Martin Baumann, and Johannes Kraus. 2024. Empowering Calibrated (Dis-)Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*

(Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 481, 19 pages. https://doi.org/10.1145/3613904.3642122

[20] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff (Eds.). Association for Computational Linguistics, Atlanta, Georgia, 746–751. https://aclanthology.org/N13-1090/

[21] Amanda K. Newendorp, Mohammadamin Sanaei, Arthur J Perron, Hila Sabouni, Nikoo Javadpour, Maddie Sells, Katherine Nelson, Michael Dorneich, and Stephen B. Gilbert. 2024. Apple's Knowledge Navigator: Why Doesn't that Conversational Agent Exist Yet?. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 153, 14 pages. https://doi.org/10.1145/3613904.3642739

[22] Adiba Orzikulova, Han Xiao, Zhipeng Li, Yukang Yan, Yuntao Wang, Yuanchun Shi, Marzyeh Ghassemi, Sung-Ju Lee, Anind K Dey, and Xuhai Xu. 2024. Time2Stop: Adaptive and Explainable Human-AI Loop for Smartphone Overuse Intervention. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 1–20. https://doi.org/10.1145/3613904.3642747

[23] Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The Linear Representation Hypothesis and the Geometry of Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, Vol. 235. PMLR, 39566–39599. https://proceedings.mlr.press/v235/park24c.html

[24] Karl Pearson. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, 11 (1901), 559–572. https://doi.org/10.1080/14786440109462720 arXiv:https://doi.org/10.1080/14786440109462720

[25] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084 [cs.CL] https://arxiv.org/abs/1908.10084

[26] Donghao Ren, Fred Hohman, Halden Lin, and Dominik Moritz. 2025. Embedding Atlas: Low-Friction, Interactive Embedding Visualization. arXiv:2505.06386 [cs.HC] https://arxiv.org/abs/2505.06386

[27] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 15504–15522. https://doi.org/10.18653/v1/2024.acl-long.828

[28] Dominik Sacha, Leishi Zhang, Michael Sedlmair, John Aldo Lee, Jaakko Peltonen, Daniel Weiskopf, Stephen C. North, and Daniel A. Keim. 2017. Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis. *IEEE Transactions on Visualization and Computer Graphics* 23 (2017), 241–250. https://api.semanticscholar.org/CorpusID:7798441

[29] Tim Sainburg, Leland McInnes, and Timothy Q Gentner. 2021. Parametric UMAP Embeddings for Representation and Semisupervised Learning. *Neural Computation* 33, 11 (2021), 2881–2907.

[30] Ben Shneiderman. 1983. Direct manipulation: A step beyond programming languages. *Computer* 16, 8 (1983), 57–69.

[31] Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B. Viégas, and Martin Wattenberg. 2016. Embedding Projector: Interactive Visualization and Interpretation of Embeddings. arXiv:1611.05469 [stat.ML] https://arxiv.org/abs/1611.05469

[32] Laurens van der Maaten and Geoffrey Hinton. 2008. Viualizing data using t-SNE. *Journal of Machine Learning Research* 9 (11 2008), 2579–2605.

[33] Yan Wang, Humphrey O Obie, Zhuying Li, Flora D. Salim, John Grundy, and Florian 'Floyd' Mueller. 2024. GustosonicSense: Towards understanding the design of playful gustosonic eating experiences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 361, 12 pages. https://doi.org/10.1145/3613904.3642182

[34] Zijie J. Wang, Fred Hohman, and Duen Horng Chau. 2023. WizMap: Scalable Interactive Visualization for Exploring Large Machine Learning Embeddings. arXiv:2306.09328 [cs.LG] https://arxiv.org/abs/2306.09328

[35] Sahba Zojaji, Andrii Matviienko, Iolanda Leite, and Christopher Peters. 2024. Join Me Here if You Will: Investigating Embodiment and Politeness Behaviors When Joining Small Groups of Humans, Robots, and Virtual Characters. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 595, 16 pages. https://doi.org/10.1145/3613904.3642905

[36] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2023. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405 [cs.LG] https://arxiv.org/abs/2310.01405

Received 20 January 2025