MLVis 2026: Machine Learning Methods in Visualisation for Big Data (2026)
D. Archambault, I. Nabney and J. Peltonen (Editors)

Visual Analysis of Semantic Paraphrase Embebdding Projection Stability

M. Schmidt[1], D. A. Keim[1], and F. L. Dennig [1]

[1]University of Konstanz

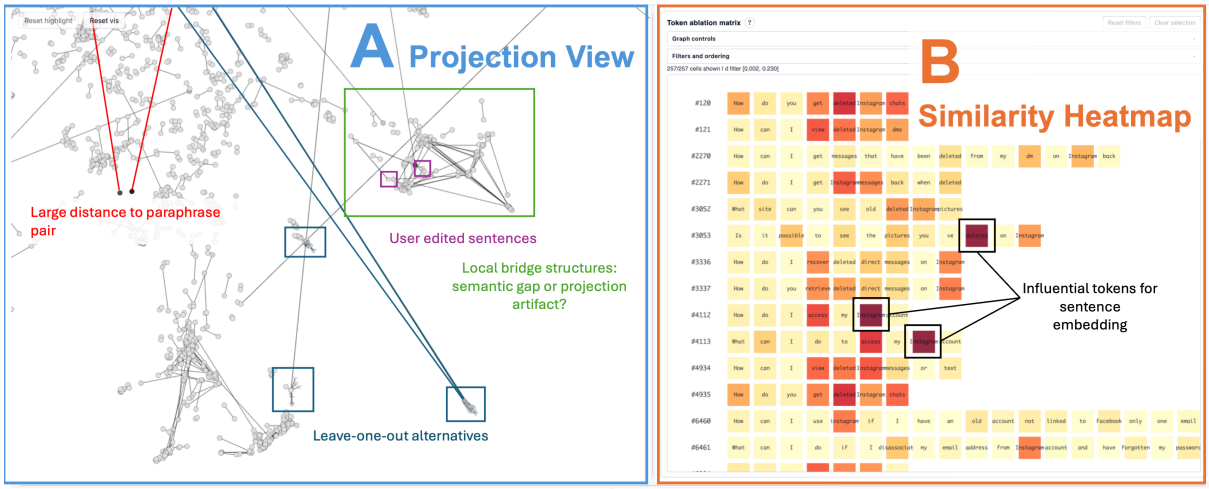

Figure 1: *(A) UMAP projection of sentence embeddings for paraphrase pairs and user-edited variants. Most paraphrase pairs remain locally coherent, but some exhibit large separations. Leave-one-out alternatives and small bridge-like local structures highlight cases where apparent neighborhoods may reflect either genuine semantic transitions or projection artifacts. (B) Token-level similarity heatmap for a selected set of sentences. The matrix exposes tokens with disproportionate influence on the sentence embedding, explaing local clustering structure in the projection.*

Abstract

Sentence embeddings are widely used as a proxy for semantic similarity in applications such as passage retrieval for question answering. A common approach to inspect such embeddings is to project them into a lower dimension. However, neither small embedding distance nor proximity in a low-dimensional projection guarantees semantic equivalence. We present an interactive analysis tool for exploring this mismatch on paraphrase pairs from the Quora Question Pairs dataset. The system combines an interactive UMAP projection, editable sentence variants, and a heatmap for token sensitivity analysis. This enables users to inspect whether local neighborhoods in the projection reflect relationships in the original embedding space and to identify tokens influencing the sentence representation. In a qualitative case study, we show failure modes, including paraphrase pairs with unexpectedly large embedding distances and non-equivalent questions in which local neighborhoods are not defined by semantic similarity. Our results show that visual proximity should be treated as an exploratory cue rather than a semantic equivalence.

CCS Concepts

• ***Human-centered computing*** → ***Visualization;*** • ***Computing methodologies*** → *Machine learning;*

1. Introduction

Large language models (LLMs), such as GPT-5 [Ope26] and related transformer-based architectures, have become the dominant paradigm for building modern AI systems. A central mechanism underlying these models is the use of dense vector representations—embeddings—that encode linguistic inputs into high-dimensional semantic spaces. Within these spaces, similarity between texts is operationalized through geometric proximity, enabling a wide range of downstream applications, including semantic search, clustering, and retrieval-augmented generation (RAG) [LPP*20]. In particular, RAG

DOI: 10.2312/kjtgq,0.04/..2

systems critically depend on the assumption that semantically equivalent or related inputs are mapped to nearby regions in embedding space, such that relevant information can be reliably retrieved.

Paraphrases are sentences that differ syntactically while preserving semantic content and therefore a natural test case for this assumption. Prior work has demonstrated that incorporating paraphrastic variation can improve robustness in tasks such as question answering and summarization [DMRL17]. However, empirical evidence also suggests that modern models remain sensitive to surface-level perturbations, leading to variability in embedding representations and, consequently, in downstream behavior [MLY*20]. This raises two fundamental questions: (1) to what extent does proximity in high-dimensional embedding space faithfully reflect semantic equivalence, and (2) to what extent are these high-dimensional relationships preserved in loq-dimensional projections such as UMAP?

A common approach to interrogating embedding spaces is dimensionality reduction, with techniques such as UMAP [MHSG18] providing visually interpretable projections of high-dimensional structures. These projections are widely used to support qualitative analysis, debugging, and model comparison. However, UMAP is a non-linear method that prioritizes local structure preservation while distorting global distances. As a result, two distinct mismatches may arise: low-dimensional projection proximity may fail to preserve true high-dimensional embedding relationships, and high-dimensional embedding proximity itself may fail to correspond to semantic equivalence. Thus, visual closeness in UMAP does not guarantee semantic equivalence.

In this work, we separately investigate (1) the relationship between embedding distance and semantic equivalence, and (2) the relationship between high-dimensional embedding neighborhoods and their preservation in low-dimensional UMAP projections, in the context of paraphrase detection. Using the Quora Question Pairs (QQP) dataset [SGNE19] as a case study, we develop an interactive visualization system that enables detailed exploration of UMAP projections of sentence embeddings. Our approach allows users to inspect clusters of paraphrastic and non-paraphrastic pairs, analyze local neighborhood stability under projection, and identify instances where geometric proximity fails to align with semantic similarity. Rather than proposing a new embedding model, dimensionality-reduction method, or quantitative benchmark, our contribution is analytical and diagnostic: in contrast to general-purpose embedding visualization systems, we focus on annotated paraphrase pairs and combine high-dimensional neighborhood analysis, UMAP-based inspection, and token-removal probes to identify whether observed failures arise from the embedding space, the projection, or local textual factors. We contribute:

(1) An examination of the consistency between embedding-space distances and UMAP-projected neighborhoods.
(2) An exploration tool for failure modes where paraphrases are not reliably grouped, or non-paraphrases grouped.
(3) Qualitative insights into the limitations of relying on low-dimensional visualizations for assessing semantic structure.

These failure modes extend beyond paraphrase datasets to embedding-based RAG, semantic search, clustering, and duplicate detection. Our findings have direct implications for the design and evaluation of such systems, particularly in high-stakes applications where robustness to linguistic variation is essential. A publicly accessible version of our application is available under `embedding-projections.schmidt.dbvis.de`.

2. Related Work

2.1. Sentence Embedding Visualizations

Sentence embeddings are a standard representation for semantic similarity, retrieval, and paraphrase analysis. Early approaches learned global sentence representations directly [LFdS*17], while BERT established contextualized language representations as the dominant paradigm [DCLT19]. Sentence-BERT made transformer-based sentence embeddings practical for similarity search and clustering [RG19], and embedding-based retrieval is now a core component of systems such as retrieval-augmented generation [LPP*20].

Visualization has long been used to inspect such spaces [NDKS22]. The Embedding Projector popularized interactive exploration of high-dimensional embeddings through projections such as PCA and t-SNE [STN*16]. For Natural Language Processing (NLP), Liu et al. studied visual exploration of semantic structure in word embeddings [LBT*18], and Berger extended this direction to contextualized embeddings [Ber20]. More recent systems analyze sentence embeddings and transformer internals more directly, including USEVis [JTH*21], LMFingerprints [SKB*22], and LayerFlow [SGSEA25]. Across this literature, a recurring issue is that dimensionality reduction can distort neighborhood structure and thus mislead interpretation [JLK*25, HWKK23].

Our work focuses on this problem for paraphrase analysis. While prior work has shown that paraphrastic variation can improve question answering [DMRL17], models remain sensitive to surface-level perturbations [JL17, RSG18, RWGS20]. We therefore study whether local proximity in a UMAP projection reflects high-dimensional embedding distance, and whether either of them is a reliable indicator of semantic equivalence.

2.2. Attributions in Embedding

A complementary line of work investigates which input tokens most influence a model representation or prediction. Attribution methods are commonly divided into gradient-based and perturbation-based approaches [LACB24]. Gradient-based methods such as Integrated Gradients are widely used when model internals are accessible [STY17]. In contrast, perturbation-based methods are model-agnostic and often easier to interpret.

We omit tokens from each sentence, recompute the sentence embedding, and measure the resulting change. This type of erasure analysis has been used in NLP interpretability to estimate the contribution of words or internal components [LMJ16], and recent work has further formalized the properties of removal-based attribution methods [LCL23]. We use this strategy not as a full attribution framework, but as a lightweight probe to explain local changes in embedding geometry and projected neighborhoods.

3. Interactive Projection Visualization

To analyze semantic differences between paraphrases, we develop a web-based application for interactive visualization of question pairs from subsets of the QQP dataset that are labeled as paraphrases. We compute sentence embeddings using *qwen3-embedding:8b* [ZLL*25], served through Ollama with the default `Q4_K_M` quantization. These embeddings form the basis for all projection, similarity, and attribution views in the system.

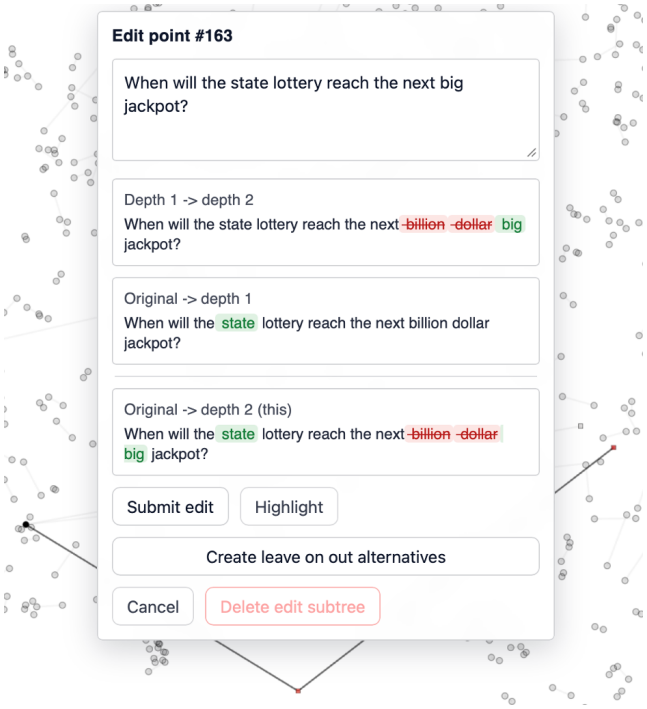

Figure 2: *Context menu for an individual point in the projection view. It supports editing the underlying sentence, highlighting the point and its derived edits, generating leave-one-out variants, removing edits, and viewing text diffs.*

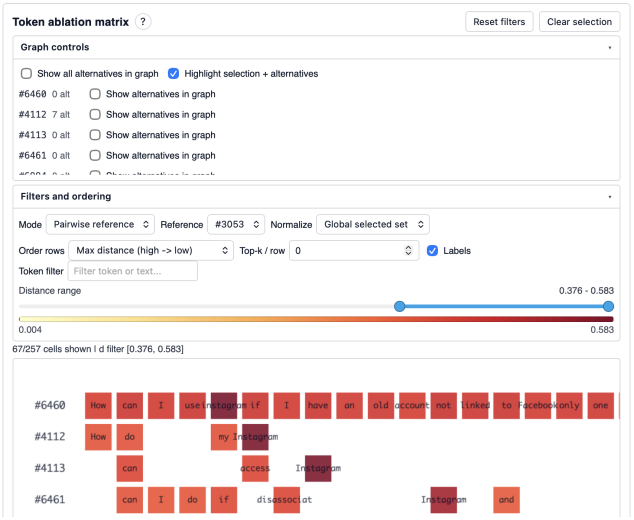

Figure 3: *The similarity heatmap showing the cosine similarity in the original high-dimensional embeddings space. It contains a control panel to enable highlighting in the projection view, filter entries based on token occurrence, pick the distance mode or reorder entries based on attributes such as the mean distance.*

3.1. Projection View

The central component of the application is the projection view (see Fig. 1 A). To visualize the high-dimensional sentence embeddings, we reduce them to two dimensions using UMAP. We choose UMAP because it is computationally efficient and typically preserves local neighborhood structure well [MHSG18]. Since the preservation of global structure is sensitive to hyperparameters and initialization [KL21], the interface provides controls for adjusting these settings interactively. The resulting projection is shown as a scatterplot in which each point represents a sentence embedding. Sentences belonging to the same paraphrase pair are linked by a line segment. The visualization is implemented with a WebGL-based rendering layer to support interactive exploration of large datasets. Points can be selected either individually by clicking or collectively through lasso selection. The system also supports interactive editing of sentences. After selecting a point, the user can modify the corresponding sentence and submit the edited version for re-embedding. The updated embedding is projected into the existing UMAP space without refitting the projection, allowing the new point to be compared directly to the original distribution. Fig. 2 shows the dialog presented to the user after clicking a point. Edited samples are visually distinguished from original dataset points by a different shape and color. A websocket connection provides live status updates during this process. When hovering over a point, the system highlights the point together with its related predecessors and successors, including edited variants.

To support distance-based analysis, point color encodes the the cosine distance $d_{\cos}(s_1, s_2)$ between paired sentences, normalized to $[0,1]$ by the range of pairwise distances in the dataset. In addition, selecting a sentence allows the user to generate and project its *leave-one-out* variants $\{s_{\setminus t} \mid t \in s\}$, each obtained by removing a single token t from the original sentence s, which are visualized with a triangle shape and blue color.

3.2. Attribution Heatmap

To analyze token-level effects on sentence representations, we provide an attribution heatmap that visualizes similarities between original sentences and their *leave-one-out* variants [LMJ16] (see Fig. 1 B). Fig. 3 displays our heatmap component with its control panel. The heatmap supports three comparison modes.

Self-Distance Mode: Each ablated sentence embedding is compared to the embedding of its own original sentence, i.e., $\mathbf{r} = e(s_i)$, where $e(s)$ denotes the sentence embedding. This mode measures the sensitivity of a sentence embedding to individual token removal.

Pairwise Reference Mode: Token ablations of one sentence are compared to the embedding of a selected reference sentence, i.e., $\mathbf{r} = e(s_r)$. This allows direct semantic comparison between two paraphrase candidates.

Corpus Aggregate Mode: Ablated embeddings are compared against the mean embedding of a selected set of sentences, i.e., $\mathbf{r} = \text{mean}_j e(s_j)$. This mode supports analysis relative to a broader local or corpus-level semantic context.

The heatmap uses global min-max normalization over the current selection. It provides tooltips on hover, top-k filtering, ordering by mean, minimum, or maximum distance, and ordering by sentence length. Additional controls allow token-based filtering and restriction of the displayed similarity range. Selections in the heatmap can be linked to the projection view, where corresponding points are highlighted and non-selected points are dimmed to reduce visual clutter.

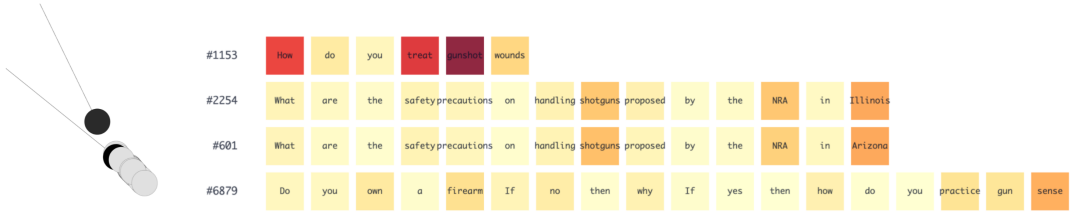

Figure 4: *Cluster in the projected QQP dataset regarding firearm safety but also accidents.*

4. Case Study

We use the system to qualitatively analyze QQP paraphrase pairs in embedding space and their UMAP projection.

A first pattern is paraphrase pairs that are semantically close yet have relatively large embedding distance. Examples include *"Is the politics of Quora inclined towards the left wing?"* and *"Is Quora an Island for left wing politics on the internet?"*, as well as *"Who do you think is going to replace Rosberg at Mercedes?"* and *"Who will replace Nico Rosberg for F1 2017 season?"* Here, similar intent is offset by differences in framing, wording, or specificity, leading to a high embedding distance.

A second pattern is paraphrase pairs with asymmetric information content. For example, *Do you own a firearm? If no then why? If yes then how do you practice 'gun sense'?"* and *Do you own a firearm? If so, why?"* are labeled as paraphrases, although the first adds safety-related content and is therefore placed closer to questions about precautions. Fig. 4 illustrates how similar topics merge into one cluster and how individual tokens can pull a question closer to a topical neighborhood than to its labeled counterpart. More generally, the projection reveals local structures such as chains, bridges, and dense neighborhoods. The attribution heatmap helps explain them by showing which omitted tokens most affect the embedding. A further pattern is paraphrase pairs that differ mainly in scope or specificity. For example, *What are the expected best upcoming Hollywood movies in 2017?"* and *What are the most anticipated movies of 2017?"* are closely related, but the first is restricted to Hollywood productions whereas the second is broader. Although close in the projection, the pair still differs semantically. Examples like these where single tokens have a big influence on the representation can be achieved by ordering and filtering the heatmap component as demonstrated in Fig. 3.

Overall, the projection view and attribution heatmap reveal several failure modes: paraphrases with high embedding distance, non-equivalent questions that remain geometrically close, and neighborhoods that reflect topic rather than meaning.

5. Discussion

This work is limited by its qualitative focus and by the use of a single dataset and embedding model; the findings should therefore not be overgeneralised. Still, the recurring patterns suggest that the observed mismatches are not isolated cases.

First, QQP is a convenient basis for analysis, but its binary labels do not capture omitted content, added constraints, or shifts in specificity. Some pairs labeled as paraphrases are therefore better understood as partially overlapping or topically related rather than strictly equivalent. The visualization does not resolve this ambiguity, but makes such cases easier to detect, which is relevant when using human-labeled datasets to evaluate models in sensitive applications.

Second, UMAP projections must be treated as exploratory views rather than direct evidence of semantic structure. Projected neighborhoods may preserve meaningful local relations, but bridge structures or dense clusters can also result from projection artifacts or shared lexical anchors. The attribution heatmap partly mitigates this by exposing token-level sensitivity through leave-one-out analysis, although token omission remains only a lightweight probe and does not fully explain semantic contribution.

These observations extend to retrieval and semantic search, where nearest neighbors are often treated as semantically substitutable. Our examples show that this assumption is too strong: close items may share topic or vocabulary without preserving meaning, while paraphrases may be separated by framing or scope differences. Geometric closeness should therefore be treated as a heuristic, not a semantic guarantee.

Future work should compare additional embedding models, projection algorithms, and parameter settings, evaluate the interface in a user study, and consider datasets with more fine-grained semantic labels. More complete assessment of semantic equivalence may require decomposing sentences into atomic semantic units rather than relying on binary paraphrase labels.

6. Conclusion

We presented an interactive visual analysis system for inspecting sentence embeddings of paraphrase pairs through linked projection and attribution views. Using the QQP dataset as a case study, we showed that neither low embedding distance nor close UMAP proximity is a reliable indicator of semantic equivalence. The combined visual and interaction-based analysis reveals mismatches between dataset labels, embedding geometry, and projected neighborhoods, and helps explain them through token-level sensitivity patterns. These results underline the need to validate embedding behavior beyond geometric distance alone. More broadly, our work demonstrates that interactive visualization can serve as a useful diagnostic tool for assessing the stability and semantic reliability of embedding-based representations.

Acknowledgements

This work was partially funded by the German Research Foundation (DFG), Project-ID 551281144 – SFB 1760 (Project C05) and the Federal Ministry of Research, Technology and Space of Germany (BMFTR) in the project *REGAiT* (FKZ: 13N17424 to 13N174247).

References

[Ber20] BERGER M.: Visually analyzing contextualized embeddings. *arXiv preprint arXiv:2009.02554* (2020). doi:10.48550/arXiv.2009.02554. 2

[DCLT19] DEVLIN J., CHANG M.-W., LEE K., TOUTANOVA K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (2019), Association for Computational Linguistics, pp. 4171–4186. doi:10.18653/v1/N19-1423. 2

[DMRL17] DONG L., MALLINSON J., REDDY S., LAPATA M.: Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (2017), Association for Computational Linguistics, pp. 875–886. doi:10.18653/v1/D17-1091. 2

[HWKK23] HUANG Z., WITSCHARD D., KUCHER K., KERREN A.: VA + embeddings STAR: A state-of-the-art report on the use of embeddings in visual analytics. *Computer Graphics Forum 42*, 3 (2023), 539–571. doi:10.1111/cgf.14859. 2

[JL17] JIA R., LIANG P.: Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (2017), Association for Computational Linguistics, pp. 2021–2031. doi:10.18653/v1/D17-1215. 2

[JLK*25] JEON H., LEE H., KUO Y.-H., YANG T., ARCHAMBAULT D., KO S., FUJIWARA T., MA K.-L., SEO J.: Unveiling high-dimensional backstage: A survey for reliable visual analytics with dimensionality reduction. *arXiv preprint arXiv:2501.10168* (2025). doi:10.48550/arXiv.2501.10168. 2

[JTH*21] JI X., TU Y., HE W., WANG J., SHEN H.-W., YEN P.-Y.: USEVis: Visual analytics of attention-based neural embedding in information retrieval. *Visual Informatics 5*, 2 (2021), 1–12. doi:10.1016/j.visinf.2021.03.003. 2

[KL21] KOBAK D., LINDERMAN G. C.: Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology 39*, 2 (2021), 156–157. doi:10.1038/s41587-020-00809-z. 3

[LACB24] LYU Q., APIDIANAKI M., CALLISON-BURCH C.: Towards faithful model explanation in NLP: A survey. *Computational Linguistics 50*, 2 (2024), 657–723. doi:10.1162/coli_a_00511. 2

[LBT*18] LIU S., BREMER P.-T., THIAGARAJAN J. J., SRIKUMAR V., WANG B., LIVNAT Y., PASCUCCI V.: Visual exploration of semantic relationships in neural word embeddings. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 553–562. doi:10.1109/TVCG.2017.2745141. 2

[LCL23] LIN C., COVERT I., LEE S.-I.: On the robustness of removal-based feature attributions. In *Advances in Neural Information Processing Systems* (2023), vol. 36, Curran Associates, Inc., pp. 79613–79666. URL: https://papers.nips.cc/paper_files/paper/2023/hash/fbbda4e85a6641bf425be3a6cfd84d20-Abstract-Conference.html. 2

[LFdS*17] LIN Z., FENG M., DOS SANTOS C. N., YU M., XIANG B., ZHOU B., BENGIO Y.: A structured self-attentive sentence embedding. In *International Conference on Learning Representations (ICLR)* (2017). URL: https://openreview.net/forum?id=BJC_jUqxe. 2

[LMJ16] LI J., MONROE W., JURAFSKY D.: Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220* (2016). doi:10.48550/arXiv.1612.08220. 2, 3

[LPP*20] LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W.-T., ROCKTÄSCHEL T., RIEDEL S., KIELA D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* (2020), vol. 33, Curran Associates, Inc., pp. 9459–9474. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html. 1, 2

[MHSG18] MCINNES L., HEALY J., SAUL N., GROSSBERGER L.: UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software 3*, 29 (2018), 861. doi:10.21105/joss.00861. 2, 3

[MLY*20] MORRIS J., LIFLAND E., YOO J. Y., GRIGSBY J., JIN D., QI Y.: TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (2020), Association for Computational Linguistics, pp. 119–126. doi:10.18653/v1/2020.emnlp-demos.16. 2

[NDKS22] NGO Q. Q., DENNIG F. L., KEIM D. A., SEDLMAIR M.: Machine learning meets visualization - Experiences and lessons learned. *it - Information Technology 64*, 4–5 (2022), 169–180. URL: https://www.degruyterbrill.com/document/doi/10.1515/itit-2022-0034/html, doi:10.1515/ITIT-2022-0034. 2

[Ope26] OPENAI: OpenAI GPT-5 System Card. *arXiv preprint arXiv:2601.03267* (2026). doi:10.48550/arXiv.2601.03267. 1

[RG19] REIMERS N., GUREVYCH I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019), Association for Computational Linguistics, pp. 3982–3992. doi:10.18653/v1/D19-1410. 2

[RSG18] RIBEIRO M. T., SINGH S., GUESTRIN C.: Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2018), Association for Computational Linguistics, pp. 856–865. doi:10.18653/v1/P18-1079. 2

[RWGS20] RIBEIRO M. T., WU T., GUESTRIN C., SINGH S.: Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), Association for Computational Linguistics, pp. 4902–4912. doi:10.18653/v1/2020.acl-main.442. 2

[SGNE19] SHARMA L., GRAESSER L., NANGIA N., EVCI U.: Natural language understanding with the quora question pairs dataset. *arXiv preprint arXiv:1907.01041* (2019). doi:10.48550/arXiv.1907.01041. 2

[SGSEA25] SEVASTJANOVA R., GERLING R., SPINNER T., EL-ASSADY M.: Layerflow: Layer-wise exploration of LLM embeddings using uncertainty-aware interlinked projections. *Computer Graphics Forum 44*, 3 (2025). doi:10.1111/cgf.70123. 2

[SKB*22] SEVASTJANOVA R., KALOULI A.-L., BECK C., HAUPTMANN H., EL-ASSADY M.: LMFingerprints: Visual explanations of language model embedding spaces through layerwise contextualization scores. *Computer Graphics Forum 41*, 3 (2022), 295–307. doi:10.1111/cgf.14541. 2

[STN*16] SMILKOV D., THORAT N., NICHOLSON C., REIF E., VIÉGAS F. B., WATTENBERG M.: Embedding projector: Interactive visualization and interpretation of embeddings. *arXiv preprint arXiv:1611.05469* (2016). doi:10.48550/arXiv.1611.05469. 2

[STY17] SUNDARARAJAN M., TALY A., YAN Q.: Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning* (2017), vol. 70 of *Proceedings of Machine Learning Research*, PMLR, pp. 3319–3328. URL: https://proceedings.mlr.press/v70/sundararajan17a.html. 2

[ZLL*25] ZHANG Y., LI M., LONG D., ZHANG X., LIN H., YANG B., XIE P., YANG A., LIU D., LIN J., HUANG F., ZHOU J.: Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176* (2025). doi:10.48550/ARXIV.2506.05176. 2