Learning representations of learning representations

Rita González-Márquez rita.gonzalez-marquez@uni-tuebingen.de
Hertie Institute for AI in Brain Health,
University of Tübingen, Germany

Dmitry Kobak dmitry.kobak@uni-tuebingden.de
Hertie Institute for AI in Brain Health,
University of Tübingen, Germany;
IWR, Heidelberg University, Germany

Reviewed on OpenReview: *https://openreview.net/forum?id=2OObXL3AaZ*

Abstract

The ICLR conference is unique among the top machine learning conferences in that all submitted papers are openly available. Here we present the *ICLR dataset* consisting of abstracts of all 24 thousand ICLR submissions from 2017–2024 with meta-data, decision scores, and custom keyword-based labels. We find that on this dataset, bag-of-words representation outperforms most dedicated sentence transformer models in terms of kNN classification accuracy, and the top performing language models barely outperform TF-IDF. We see this as a challenge for the NLP community. Furthermore, we use the ICLR dataset to study how the field of machine learning has changed over the last seven years, finding some improvement in gender balance. Using a 2D embedding of the abstracts' texts, we describe a shift in research topics from 2017 to 2024 and identify *hedgehogs* and *foxes* among the authors with the highest number of ICLR submissions.

1 Introduction

The International Conference on Learning Representations (ICLR) is one of the most prestigious machine learning venues: in Google Scholar Metrics it currently shares with NeurIPS the second place after CVPR. Since 2017, ICLR submissions are handled through OpenReview in a fully open way: all submitted papers are publicly visible and are eventually de-anonymized. This is not the case for most other top conferences in the field which do not make rejected papers openly visible. As the field of machine learning advances very fast, one can use ICLR submissions to study how it has changed over recent years.

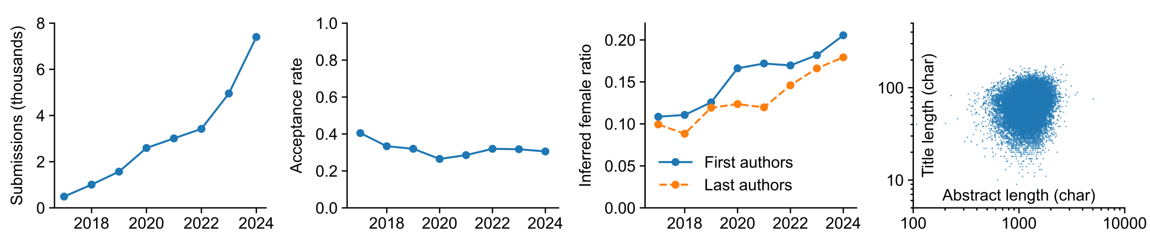

Figure 1: Summary statistics of the ICLR dataset (ICLR24v2).

Here we present the *ICLR dataset* consisting of abstracts of all ICLR submissions from 2017–2024 with meta-data and keyword-based labels (Figure 1). Our work has two goals. First, to do a metascience study of machine learning as a field, similar to how González-Márquez et al. (2024) did it for biomedicine. Second, to frame an NLP challenge: without using our labels, train a language model that would substantially surpass a naïve TF-IDF representation in terms of kNN accuracy. We found that most dedicated sentence models fared *worse* than TF-IDF, and none outperformed it by a large margin.

2 Dataset

To assemble the dataset, we queried OpenReview and downloaded titles, abstracts, author lists, keywords, reviewers' scores, and conference decisions for all 24 445 papers submitted to ICLR in 2017–2024 with intact abstracts (Figure 1). 26 submissions with placeholder abstracts (below 100 characters) were excluded. While 2024 submissions were still anonymous, we assembled the data into the ICLR24v1 dataset; after the de-anonymisation, we produced the final ICLR24v2 version. We will use the same naming convention for future updates. The data are openly available at `https://github.com/berenslab/iclr-dataset`, together with our analysis code.

We used the `gender` package (Blevins and Mullen, 2015) to infer genders of the first and the last author based on their first names. We could infer genders for 41.8% of the first authors and 49.9% of the last authors; note that the inference model fails at inferring gender for many non-Western names. We observed a steady increase in the inferred female ratio (based only on papers with inferred genders) that almost doubled since 2017: from 11% in 2017 to 21% in 2024 for the first authors, and from 10% to 18% for the last authors.

To label the dataset, we relied on the author-provided keywords and used them to assign papers to 45 non-overlapping classes (Table S1). We took the 200 most frequent keywords, combined some of them together into one class (e.g. *attention* and *transformer*), disregarded very broad keywords (e.g. *deep learning*), and assigned papers to rarer keywords first. Using this procedure, we ended up labeling 53.4% of all papers.

Reviewed papers had on average 3.7 reviews, with 93% having either 3 or 4 reviews. Across all 244 226 possible pairs of reviews of the same paper, the correlation coefficient between scores was 0.40. This was substantially higher than what had been reported for computational neuroscience conferences — 0.16 for CCN (Goodman, 2023) and 0.25 for Cosyne (Ostojic, 2020), — but note that the ICLR scores are not entirely independent as the reviewers are allowed to update them after discussion.

3 Embedding challenge

To obtain an embedding of each abstract, we used classic bag-of-words representations (Schmidt, 2018) as well as modern sentence transformers, pre-trained on large amounts of text data. We evaluated all of them using kNN classification accuracy ($k = 10$ and 10-fold cross-validation). As our main application is 2D visualisation (Section 4) which is based on the kNN graph, we consider kNN accuracy one of the most relevant metrics quantifying representation quality.

Table 1: kNN accuracies in high-dimensional spaces, and in the 2D space after t-SNE. All values should be interpreted with an uncertainty of ∼1% (see text).

Model	**High-dim.**	**2D**
TF-IDF	59.2%	52.0%
SVD	58.9%	55.9%
SVD, L^2 norm.	**60.7%**	**56.7%**
SimCSE	45.1%	36.3%
DeCLUTR-sci	52.7%	47.1%
SciNCL	58.8%	54.9%
SPECTER2	58.8%	54.1%
ST5	57.0%	52.6%
SBERT	**61.6%**	**56.8%**
Cohere v3	61.1%	56.4%
OpenAI v3	**62.3%**	**57.1%**

We used TF-IDF representation with log-scaling as implemented in scikit-learn (Pedregosa et al., 2011), which showed the best results in our prior benchmark (González-Márquez et al., 2022). Its kNN accuracy was 59.2% (Table 1). It decreased to 58.9% after SVD to 100 dimensions, but increased to 60.7% after L^2 normalisation in the SVD space (or, equivalently, using the cosine metric for kNN search).

As sentence transformers, we used three models which were specifically trained to produce representations of scientific abstracts: DeCLUTR-sci (Giorgi et al., 2021), SciNCL (Ostendorff et al., 2022), and SPECTER2 (Cohan et al., 2020). We also used SimCSE (Gao et al., 2021), ST5 (Ni et al., 2022), and the latest version of SBERT (Reimers and Gurevych, 2019). The SBERT model (`all-mpnet-base-v2`) was trained on over one billion documents from different domains and holds state-of-the-art results in recent benchmarks among all models of its size (Muennighoff et al., 2023). These six models all have `bert-base` architecture with 110 M parameters and 768-dimensional embeddings. To get the representation of each abstract, we used the representation that each model had been fine-tuned for: either average pooling over all tokens (SBERT, DeCLUTER-sci, ST5) or the classification token `[CLS]` (SciNCL, SPECTER2, SimCSE). All models were downloaded from Hugging Face. We also benchmarked two commercial models: one by Cohere (`embed-english-v3.0` in `clustering` mode; 1024-dimensional embeddings), and one by OpenAI (`text-embedding-3-large`; 3072-dimensional embeddings). For all models we report kNN accuracy using the Euclidean metric; the cosine metric gave very similar results.

We found that the three models specifically trained to represent scientific abstracts all had lower kNN accuracy than TF-IDF. Only SBERT and both commercial models could outperform TF-IDF, and only marginally, by less than 2 percentage points (Table 1). SBERT (61.6%) was only surpassed by the OpenAI embedding model (62.3%) with the performance gap below 1 percentage point. Note that all reported values should be interpreted with an

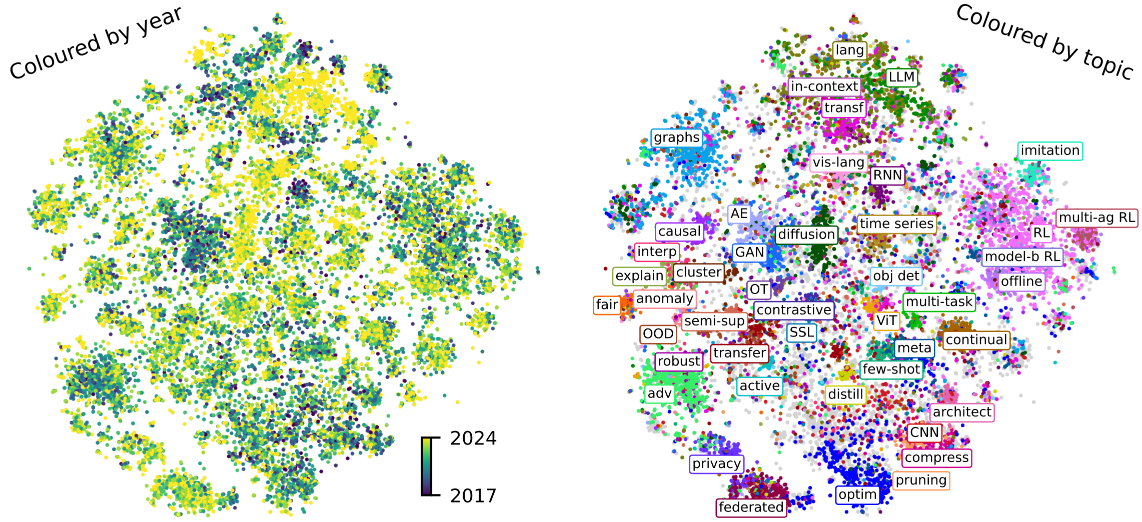

Figure 2: t-SNE embedding of the SBERT representation of ICLR abstracts (2017–2024). Left: coloured by year; right: coloured by topic.

error of around $\pm 1\%$, corresponding to the binomial standard deviation $100\sqrt{p(1-p)/n}$ for test set size $n \approx 2500$ and accuracy $p \approx 0.6$.

These results were surprising for us, because sentence transformers are complex models pre-trained with masked language modeling (Devlin et al., 2019) and fine-tuned with contrastive loss functions on large corpora. Yet their representations were not (much) better than bag-of-words representations that capture nothing beyond word counts. Modern benchmarks evaluate embedding models using various metrics and do find that sentence transformers outperform bag-of-words models (Muennighoff et al., 2023). However, the kNN graph quality is the only metric relevant for our application (see below), and here the modern models were not much better than TF-IDF, at least on our dataset.

We hope that this well-defined and practically relevant benchmark will act as a challenge for the NLP community.

4 Trends in machine learning

For data exploration, we used the SBERT representation and applied t-SNE (van der Maaten and Hinton, 2008) to embed the 768-dimensional representation in 2D. We chose t-SNE rather than UMAP (McInnes et al., 2018) because t-SNE performs the best in terms of kNN classification and kNN recall (González-Márquez et al., 2022, 2024). We used openTSNE (Poličar et al., 2019) with default parameters. In 2D, kNN classification was 56.8% (Table 1): very close to what we got using TF-IDF (56.7%) and using the OpenAI model (57.1%).

The resulting embedding showed rich structure with many visible clusters roughly corresponding to our classes (Figure 2). Related classes were located close in the embedding, showing meaningful global organisation.

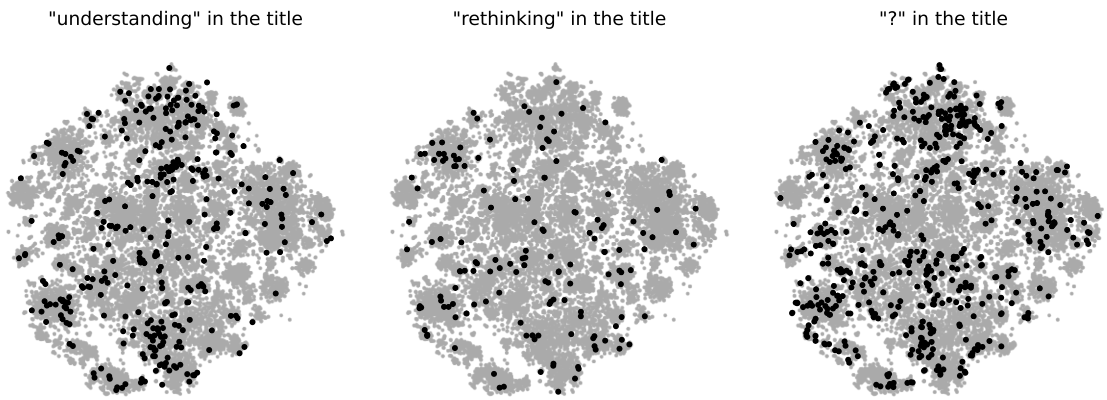

Figure 3: ICLR papers containing the words *understanding* (366), *rethinking* (155), and a question mark (550) in the title.

Overlaying the conference year and the class labels over the embedding (Figure 2) highlighted many trends in 2017–2024 machine learning. We saw that generative adversarial networks (GAN) and variational autoencoders (AE) got out of fashion while diffusion models became fashionable. Natural language processing research got dominated by large language models (LLM). Within reinforcement learning (RL), offline RL seemed to be the most recently fashionable topic. Recurrent neural networks (RNN) and adversarial examples are another two topics that lost their popularity.

We also used the 2D embedding to explore the distribution of acceptance decisions and average scores across machine learning subfields, but found no systematic differences between them (Figure S1). This suggests that ICLR's decisions were not biased towards certain topics. Similarly, we did not see any systematic differences in gender ratio between machine learning subfields (Figure S2), in stark contrast with biomedical research (González-Márquez et al., 2024) and academia as a whole (Larivière et al., 2013; Shen et al., 2018; Bendels et al., 2018).

Which subfields of machine learning are the most controversial? We investigated this question by looking at the distribution of papers containing the words *understanding*, *rethinking*, or the question mark in their titles (Figure 3). These distributions were not uniform and had local modes around language models, vision-language models, adversarial examples, and also around optimisation/distillation.

Finally, we looked at the authors with the highest number of ICLR submissions (Figure 4) and saw clear distinction between focused researchers working mostly on one topic and broad researchers working in many machine learning fields: *hedgehogs* and *foxes*, according to the famous classification by Isaiah Berlin (1953). Among the top three most prolific authors, Sergey Levine (170 submissions) and Pieter Abbeel (109) were 'hedgehogs' working mostly on reinforcement learning, while Yoshua Bengio (146) was a 'fox'. The acceptance rate among the most prolific authors in both categories was often higher than the average acceptance rate (31%).

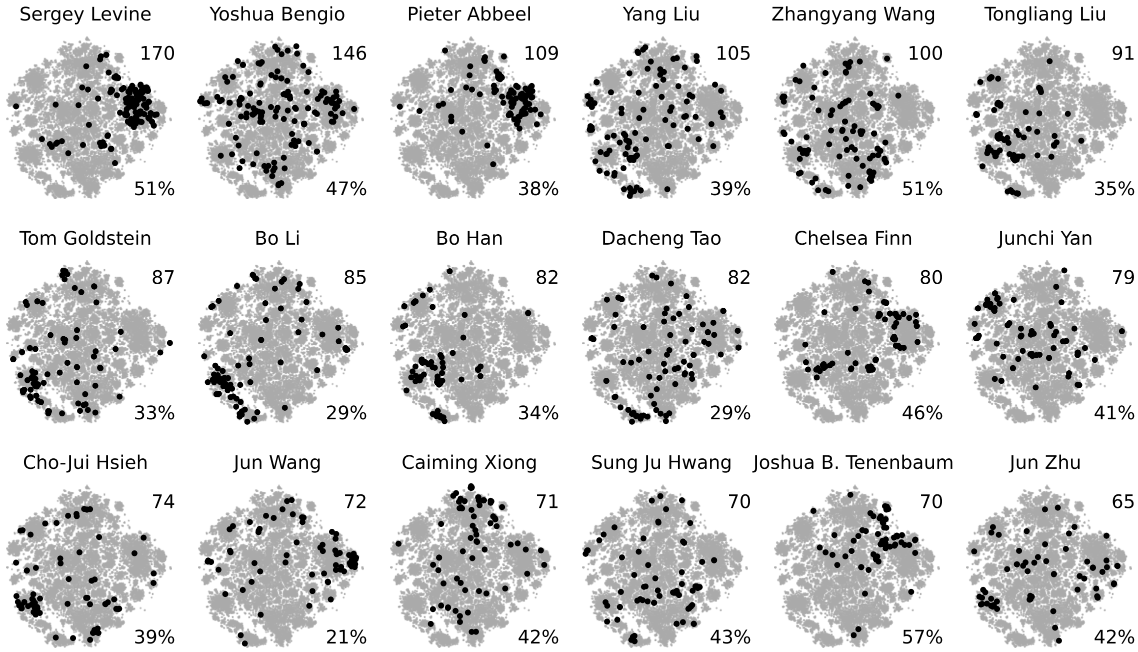

Figure 4: Top 18 authors by the total number of ICLR submissions over 2017–2024. Each panel shows the total number of submissions and the acceptance rate.

5 Conclusion

Many text datasets are available for training and benchmarking language models. The benefits of the *ICLR dataset* suggested here are (i) its compact size; (ii) it not being part of the training set of existing sentence transformer models; (iii) it covering topics very familiar to machine learning researchers, allowing qualitative assessment of embedding quality.

We demonstrated that the ICLR dataset can be used to study metascientific questions and to draw conclusions about the state of machine learning field as a whole. We also argue that substantially outperforming TF-IDF representation remains an open NLP challenge (Figures S3, S4).

Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (KO6282/2-1 and Excellence clusters 2064 "Machine Learning: New Perspectives for Science", EXC 390727645, and 2181 "STRUCTURES", EXC 390900948), by the German Ministry of Education and Research (Tübingen AI Center), and by the Gemeinnützige Hertie-Stiftung. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Rita González Márquez.

References

Michael HK Bendels, Ruth Müller, Doerthe Brueggmann, and David A Groneberg. Gender disparities in high-quality research revealed by nature index journals. *PLOS One*, 13(1): e0189136, 2018.

Isaiah Berlin. *The hedgehog and the fox: An essay on Tolstoy's view of history.* Weidenfeld & Nicolson, 1953.

Cameron Blevins and Lincoln Mullen. Jane, John... Leslie? A historical method for algorithmic gender prediction. *DHQ: Digital Humanities Quarterly*, 9(3), 2015.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, 2021.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, 2021.

Rita González-Márquez, Philipp Berens, and Dmitry Kobak. Two-dimensional visualization of large document libraries using t-SNE. In *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022.

Rita González-Márquez, Luca Schmidt, Benjamin M Schmidt, Philipp Berens, and Dmitry Kobak. The landscape of biomedical research. *Patterns*, page 100968, 2024.

Dan Goodman, 2023. URL `https://twitter.com/neuralreckoning/status/1659185421568335874`.

Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R Sugimoto. Bibliometrics: Global gender disparities in science. *Nature*, 504(7479):211–213, 2013.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2006–2029, 2023.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, 2022.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688, 2022.

Srdjan Ostojic, 2020. URL `https://twitter.com/ostojic_srdjan/status/1215675748444528640`.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Pavlin G Poličar, Martin Stražar, and Blaž Zupan. openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. *BioRxiv*, page 731877, 2019.

Nils Reimers and Iryna Gurevych. 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)*, pages 3982–3992, 2019.

Benjamin Schmidt. Stable random projection: Lightweight, general-purpose dimensionality reduction for digitized libraries. *Journal of Cultural Analytics*, 2018.

Yiqin Alicia Shen, Jason M Webster, Yuichi Shoda, and Ione Fine. Persistent underrepresentation of women's science in high profile journals. *BioRxiv*, page 275362, 2018.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008.

Appendix A. Supplementary Figures

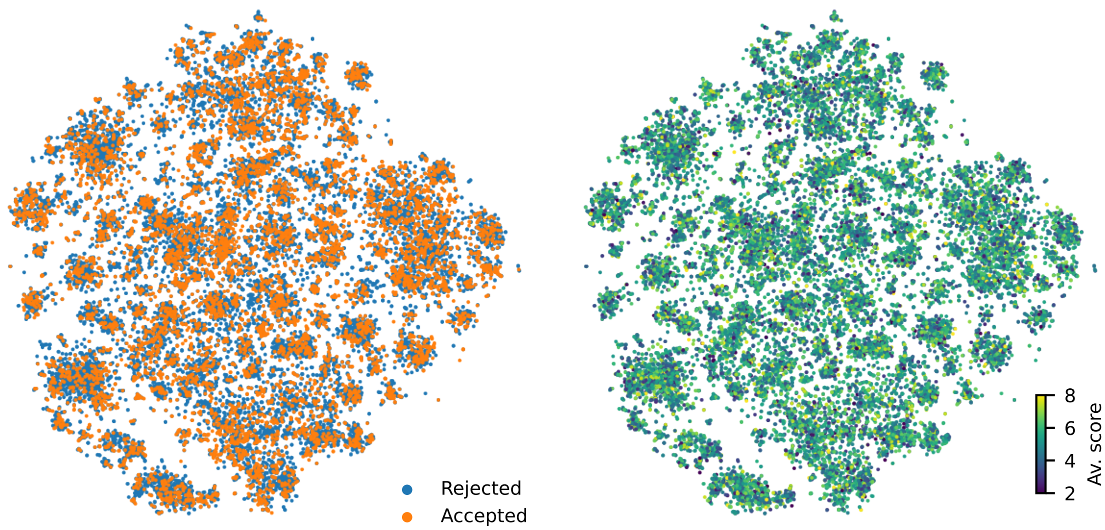

Figure S1: Acceptance decisions and average scores. Left: accepted papers are shown on top. Right: papers are shown in randomized order.

Figure S2: Inferred genders of the first and the last authors. Papers are plotted in the following order: unknown gender, male, female. Female markers are larger.

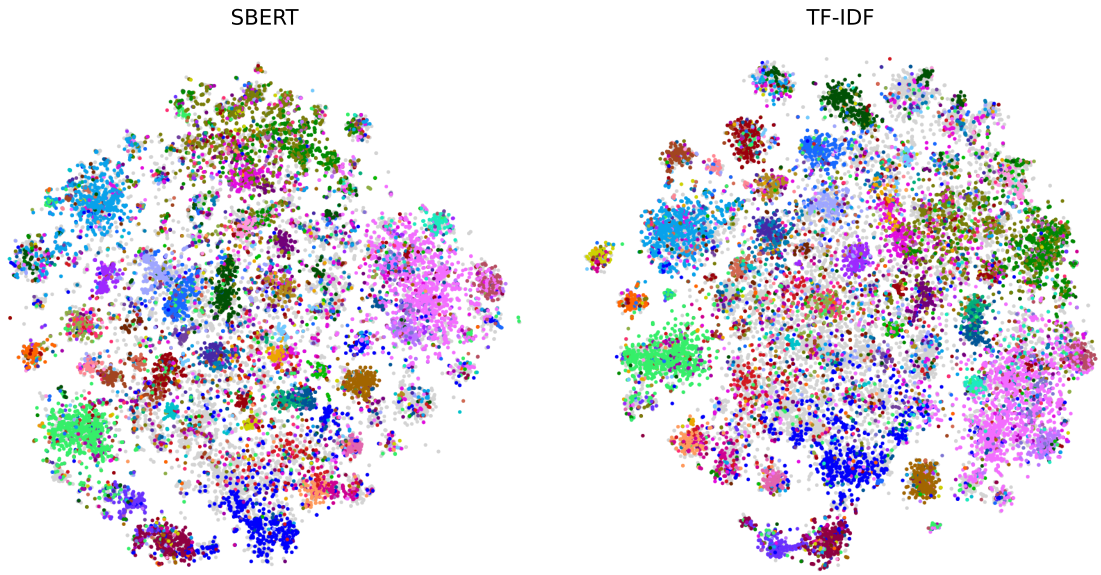

Figure S3: t-SNE embeddings of the SBERT representation (left) and of the row-normalized SVD (100 components) of the TF-IDF representation (right). The embedding on the right was rotated by 90° and flipped to align it to the SBERT embedding. Colours as in Figure 2. Unlabeled papers are shown in gray in the background.

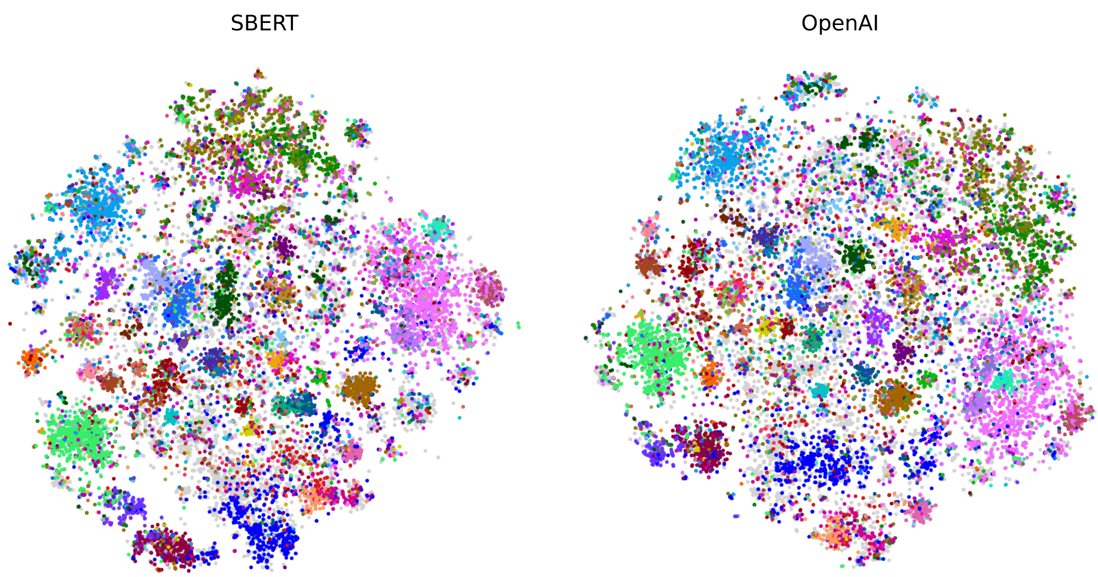

Figure S4: t-SNE embedding of the SBERT representation (left) and of the OpenAI's model representation (right). The embedding on the right was rotated by 90° and flipped. Colours as in Figures 2 and S3.

Appendix B. Supplementary Tables

Table S1: The list of all 45 classes. From the list of all existing keywords, we selected the 200 most frequent ones and manually grouped some of them together into classes. We left out too general keywords (e.g. *deep learning*) as well as all resulting classes with fewer than 50 papers. Table continues on the next page.

Class	Samples	Keyword	Frequency
RL	1266	reinforcement learning	1608
		deep reinforcement learning	298
Adversarial	870	adversarial training	217
		adversarial attacks	106
		adversarial defense	50
		adversarial examples	196
		adversarial learning	93
		adversarial machine learning	54
		adversarial robustness	241
		adversarial	60
		adversarial attack	121
Language models	802	question answering	59
		reasoning	85
		language modeling	85
		machine translation	91
		language models	151
		nlp	166
		natural language processing	433
		language model	105
Optimization	790	optimization	410
		gradient descent	86
		combinatorial optimization	69
		bayesian optimization	64
		stochastic gradient descent	77
		stochastic optimization	56
		convex optimization	57
		sgd	86
		non-convex optimization	66
Graphs	730	gnn	64
		graph	48
		graph representation learning	85
		graph neural networks	563
		graph neural network	230
Transformers	557	transformer	340
		self-attention	73
		attention	183
		attention mechanism	53
		transformers	261
LLMs	538	llm	80
		prompting	48
		large language model	210
		large language models	447
Diffusion models	443	diffusion models	280
		diffusion model	167
		diffusion	69
Transfer learning	419	transfer learning	388
		domain generalization	124
		domain adaptation	176
GANs	380	generative adversarial networks	190
		gan	168
		generative adversarial network	70
		gans	91
Autoencoders	330	variational autoencoders	83
		autoencoders	52
		autoencoder	63
		variational autoencoder	93
		vae	71
Continual learning	313	lifelong learning	82
		continual learning	339

Class	**Samples**	**Keyword**	**Frequency**
Federated learning	298	federated learning	485
Out-of-distribution	275	out-of-distribution generalization distribution shift out-of-distribution detection out-of-distribution	59 96 92 53
Meta learning	275	meta learning meta-learning	121 301
Self-supervised learning	259	self-supervised learning	473
RNNs	250	lstm recurrent neural networks recurrent neural network rnn	66 114 48 65
CNNs	247	convolutional neural network convolutional neural networks cnn	76 130 88
Contrastive learning	244	contrastive learning	344
Privacy	215	differential privacy privacy	154 99
Compression	214	model compression compression	135 121
Causality	202	causal inference causality causal discovery	104 80 53
Explainability	194	explainable ai explainability	92 131
Offline RL	184	offline rl offline reinforcement learning	55 150
Interpretability	177	interpretability	356
Semi-supervised learning	176	semi-supervised learning	253
Robustness	175	robustness	411
Few-shot learning	157	few-shot learning	218
Multi-agent RL	151	multi-agent reinforcement learning	162
Knowledge distillation	150	knowledge distillation	211
Imitation learning	144	imitation learning	171
Time series	140	time series time series forecasting	129 54
Neural architecture search	138	neural architecture search	180
Pruning	133	network pruning pruning	48 140
Fairness	133	fairness	182
Optimal transport	132	optimal transport	165
ViTs	130	vision transformers vision transformer	51 98
Multi-task learning	121	multi-task learning	141
Active learning	111	active learning	131
Vision-language models	108	vision-language models clip	48 70
Object detection	106	object detection	125
Model-based RL	105	model-based reinforcement learning	111
Clustering	97	clustering	116
Anomaly detection	87	anomaly detection	109
In-context learning	87	in-context learning	105