

Activation-Based Active Learning for In-Context Learning: Challenges and Insights

Yaseen M. Osman, Geoff V. Merrett, Stuart E. Middleton

University of Southampton, UK

y.m.osman@soton.ac.uk, gvm@ecs.soton.ac.uk, sem03@soton.ac.uk

Abstract

Deep active learning has previously been explored for LLM in-context sample selection, but not with methods that utilise recent advances in understanding of transformer activations. In this paper, we test the hypothesis that model activations could provide a fine-grained signal to optimise the selection of in-context examples. We present a comprehensive analysis of MLP activation-based deep active learning methods applied to in-context learning, including how different attention masking strategies impact active learning across diverse classification and generative datasets, using both Llama-3.2-3B and Qwen2.5-3B base models. However, we find a negative result: MLP and embedding layer outputs, viewed through the lenses of massive activations or the first four moments, do not correlate with example quality or task performance. Specifically, the absolute Spearman correlation coefficient is at most 0.33 for all tasks and models we tested, showing that such activation-based sampling should not be used for in-context learning. We hypothesise that this may be due to superposition, whereby models represent more features than they have dimensionality, suggesting that methods like Sparse Autoencoders (SAEs) may be a promising future direction.¹²

1 Introduction

In-context learning is a popular alternative to fine-tuning for improving the performance of Large Language Models (LLMs). It is the process of providing examples to the LLM in the prompt rather than updating the model weights (Brown et al., 2020). Recent studies have looked at the impact and sensitivity of LLMs towards the provided in-context examples (Schoch and Ji, 2025; Li and Qiu,

2023; Ye et al., 2023; Qin et al., 2024; Rubin et al., 2022; Zhang et al., 2022; Yao et al., 2024; Wu et al., 2023).

Active learning (Settles, 2009; Cohn et al., 1994; Ein-Dor et al., 2020) aims to optimise the data selection phase by selecting data that is likely to yield better model training performance. Active learning has been extensively studied for machine learning models, such as Support Vector Machines (SVMs) (Tong, 2001) and neural networks (Cohn et al., 1994). Recently, deep active learning has been used with deep learning models such as transformers (Liu et al., 2022b; Zhang and Zhang, 2020; Shelmanov et al., 2019; Ein-Dor et al., 2020; Bayer et al., 2026).

Deep active learning has been also adapted for sampling in-context learning examples for LLMs through a variety of techniques, such as using model uncertainty (Margatina et al., 2023; Wang et al., 2025), data diversity (Margatina et al., 2023; Li and Qiu, 2023; Lee et al., 2025; Malik et al., 2025), mixtures of both diversity and uncertainty (Mavromatis et al., 2023; Treerath and Pittorino, 2026), similarity between in-context examples and test samples encoded through sentence encoders (Liu et al., 2022a; Margatina et al., 2023), training a specialised retrieval-based model (Luo et al., 2023), Monte Carlo sampling (Schoch and Ji, 2025) and reinforcement learning (Zhang et al., 2022).

The hypothesis of this paper is that model activations could provide a fine-grained signal to optimise the selection of in-context examples, grounded in both the transformer layer weights, which encode LLM pre-trained knowledge, and the relevance of the in-context example to the test sample. In this paper, we focus on MLP layer activations given the crucial role of the MLP layer in knowledge memorisation (Dong et al., 2025; Bricken et al., 2023; Nanda et al., 2023). We also explore adjusting causal masking (BehnamGhader et al., 2024; Lin et al., 2025) to better capture com-

¹Our code and a link to our project page containing more results are available at <https://github.com/yaseenosman/activation-based-al>.

²An arXiv.org version of this work is available at <https://arxiv.org/pdf/2606.05134>.

plex relationships and enable greater information flow between the candidate in-context learning example and test sample.

The contribution in this paper is a comprehensive analysis of MLP activation-based active learning methods applied to in-context learning to date, covering a diverse range of datasets comprising both classification and generative tasks. We include a study examining how different LLM masking strategies impact active learning, including tailored masking methods for in-context learning sampling. Our results show that activation-based methods for active learning do not correlate directly with task performance. This is a negative result for our hypothesis. However, this study provides a contribution in terms of useful insights and challenges that we derive from our results to guide future research in this area.

2 Related Work

Attention Masking Adjustments Recent work explored removing the causal masking from pre-trained decoder-only transformer models to enhance their performance in encoder-based and embedding tasks by allowing the model to process full sequences bidirectionally (BehnamGhader et al., 2024; Lin et al., 2025). Additionally, Springer et al. (2025) and Leviathan et al. (2025) achieved a similar effect by instead duplicating the prompt twice in the input. Other work focusing on studying and adjusting attention masking includes Barbero et al. (2024, 2025); Swietojanski et al. (2023); Wu et al. (2024).

In-Context Data Sampling Adjacently, for in-context learning examples sampling, Liu et al. (2022a) and Margatina et al. (2023) have utilised another model’s embeddings to encode both examples and tests, then employed the k-nearest neighbours algorithm. Furthermore, sampling using the attention weights of the queried transformer model has been explored in the context of Retrieval-Augmented Generation (RAG) (Chen et al., 2025) and Visual Instruction Tuning (Hu et al., 2025). We consider these sampling techniques to be the most relevant.

Mechanistic Interpretability of Activations and In-Context Learning Several works have used Sparse Autoencoders (SAEs) to facilitate the interpretation of activations and mitigate difficulties from superposition (Sharkey et al., 2022; Bricken

et al., 2023; Huben et al., 2024). Superposition is when the model represents more features than its dimensionality, resulting in highly uninterpretable and entangled activation matrices (Elhage et al., 2022).

Notably, Cho and Hockenmaier (2025) used SAEs to perform downstream steering to increase the in-context learning performance of a model. On the other hand, Cho et al. (2025) utilised SAEs beyond interpretability and found that SAEs reconstruction loss correlates with prompt task performance for in-context learning settings.

Moreover, Olsson et al. (2022) attributed the emergence of the in-context learning ability to specialised attention heads called induction heads.

3 Method

In this section, we present our methodology for scoring in-context examples using active learning and how different types of masking impact sampling performance.

3.1 Masking

To allow for different information-flow variants between the example and the test sample, we explored various masks. Given a 1-shot prompt, these are:

- **Causal Mask**, which blocks the attention from future tokens (i.e., default masking).
- **No Mask**, which removes the causal masking following (Lin et al., 2025).
- **Row Test-Based**, which only masks the test *queries* (i.e., rows of test tokens in attention weights).
- **Col Test-Based**, which only masks the test *keys* (i.e., columns of test tokens in attention weights).

Figure 1 depicts these maskings.

3.2 In-Context Examples Scoring

Once the masking of a model, containing L layers, has been adjusted, given a set of candidate examples M and test samples N , for each example $m \in M$, we construct $len(N)$ 1-shot prompts. Specifically, for an $n \in N$, $p_{m,n}$ is the concatenation of example m , a delimiter " $\text{\textbackslash n\textbackslash n}$ ", n . We then input this as a prompt to the model for a single forward pass, resulting in $L + 1$ output hidden states as activation matrices (i.e., one for each MLP layer

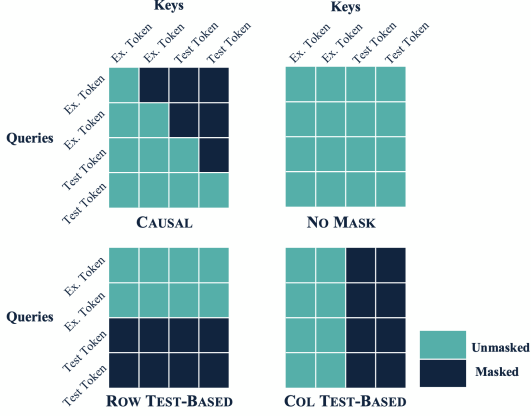


Figure 1: Shows a visualisation of the different attention maskings used in this paper using a prompt consisting of both example tokens (i.e., Ex. Token) and test tokens.

and for the embedding layer) $\{X_0, \dots, X_L\}$. The score of the prompt is given by:

$$pt_score(p_{m,n}) = \frac{1}{L+1} \sum_{l=0}^L metric(X_l), \quad (1)$$

such that $metric()$ is a metric of choice to reduce an activation matrix to a score.

To get the overall score of m , we take the mean of all prompt scores that consist of m , as follows:

$$st_score(m) = \frac{1}{len(N)} \sum_{n \in N} pt_score(p_{m,n}). \quad (2)$$

3.3 Metric Choices

The choices of metrics used in Equation 1, given X_l , are the first four moments: mean, variance, skewness, and kurtosis. These moments allow the capturing of various activation patterns. For the mean, we take the absolute value of X_l before calculating its mean. This is because we empirically observe that the mean of MLP activations is close to zero, owing to the presence of many positive and negative activation values. We do not make any changes to the other three moments.

We use the largest massive activation, given its importance in model performance (Sun et al., 2024), as an additional metric by selecting the activation with the largest absolute magnitude in X_l . We refer to this as the max activation. For completeness, we also explore the max activation, but for each feature vector and take the mean across all vectors in X_l . We refer to it as per-vector max.

4 Experimental Setup

In this section, we outline our experimental setup in detail before presenting our results. We conduct all our evaluations using the LM evaluation harness framework (Gao et al., 2023).

4.1 Models

We mainly experimented with two models, which are Llama-3.2-3B (Grattafiori et al., 2024), and Qwen2.5-3B (Team, 2024; Yang et al., 2024). These provide examples for two different families that are frequently used. We also explored Llama-3.1-8B and Llama3.1-8B-Instruct.

4.2 Datasets

We use a diverse set of datasets, covering classification-based (i.e., two multiple-choice tasks and a true/false task) and generative-based tasks (i.e., a math reasoning task). These are ARC-Challenge (ARC-C) (Clark et al., 2018), OpenBookQA (OBQA) (Mihaylov et al., 2018), BoolQ (Clark et al., 2019; Wang et al., 2019), and GSM8K (Cobbe et al., 2021), with the number of test samples, $len(N)$, being 1172, 500, 3270, and 1319, respectively.

4.3 Sampling and Evaluation

For each dataset, we randomly sample a subset M of 1000 in-context example candidates from its training split. However, for ARC-C, we use the full training split of 1119 samples as it is close to 1000.

Moreover, we calculate the st_score for all examples in M . We also evaluate each of the resulting $p_{m,n}$ 1-shot prompts on the task. Specifically, we evaluate task performance using exact match with strict match filtering for GSM8K and using accuracy for the other datasets. Finally, we compute the Spearman correlation coefficient between the task performance and st_score through utilising the SciPy implementation (i.e., `scipy.stats.spearmanr`) (Virtanen et al., 2020).

5 Results

We focus solely on 1-shot in-context learning to simplify the exploration and reduce the computational cost of our experiment, thereby enabling a finer-grained analysis. However, our scoring method can be applied to few-shot settings. Lastly, we use a batch size of 1 for all our experiments on an Nvidia L4, except for GSM8K, which we run

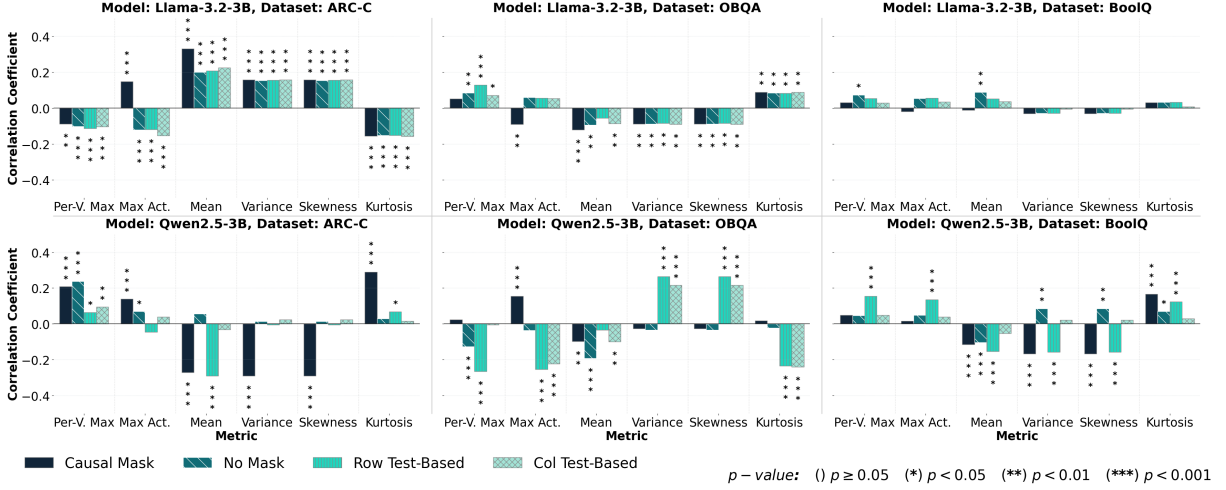


Figure 2: Shows the Spearman correlation and the associated p -values between task performance and the different activation-based metrics discussed. Results are for both Llama-3.2-3B and Qwen2.5-3B across three different classification-based datasets.

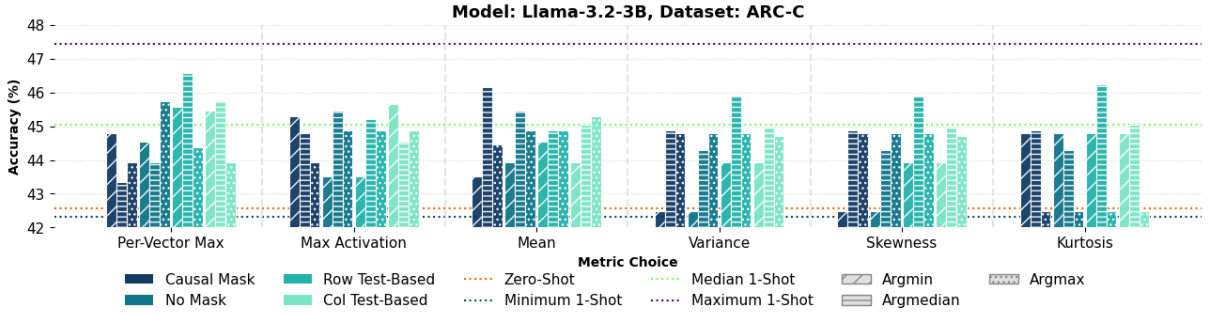


Figure 3: Shows the accuracy of Llama-3.2-3B on ARC-C when sampling examples with argmin, argmedian, and argmax metric values (column pattern) using the different masks (column colour). The accuracy does not align with argmin, argmedian, and argmax. The performances of the examples of the maximum, median and minimum accuracies are also included.

on an Nvidia L40 with a batch size of 64 due to the computational time of the generative task.

In Figure 2, we present the correlation coefficient across three classification-based datasets. For both Qwen2.5-3B and Llama-3.2-3B, a consistently low correlation coefficient is observed. Additionally, low p -values (i.e., $p < 0.001$) are observed across several experiments, reflecting a lack of correlation with high statistical significance. However, some results, especially when using Qwen2.5-3B or both models on BoolQ, have relatively higher p -values (i.e., $p \geq 0.05$), indicating lower statistical significance. We argue that the lack of correlation is persistent due to the very low correlation values in these results, which approach zero (i.e., < 0.1). Furthermore, the highest overall correlation value achieved is only around 0.33.

Figure 3 includes the performance of examples that achieve the maximum, median, and minimum

scores under each metric and mask combination on ARC-C using Llama-3.2-B, referred to as argmax, argmedian, and argmin, respectively. The true best example is never selected. Moreover, across all metrics and masks, performance and score magnitude do not align. Additionally, argmax and argmin values indicate high variability in 1-shot performance, and some examples drop performance even below the zero-shot accuracy, aligning with previous work, such as Schoch and Ji (2025).

We expanded our Llama-3.2-3B to GSM8K example reasoning-based generative tasks in Figure 4. However, the lack of correlation persists across all masks. The NaN value resulting from using causal masking with the max activation metric is due to the largest massive activation value not changing across all prompts. This results in zero variance, which in turn results in an undefined Spearman correlation.

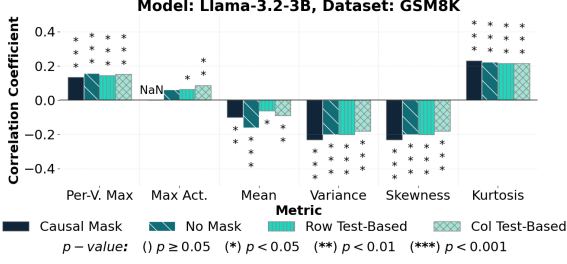


Figure 4: Shows the Spearman correlation and p -values between task performance and different activation-based metrics for GSM8K using Llama-3.2-3B.

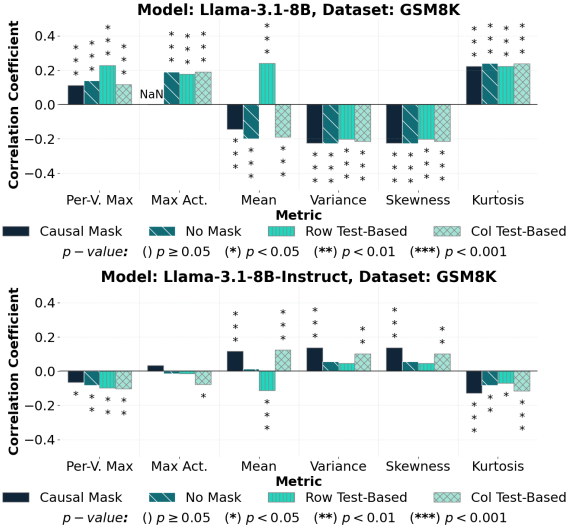


Figure 5: Shows the Spearman correlation and p -values between task performance and different activation-based metrics for GSM8K using Llama-3.1-8B and Llama-3.1-8B-Instruct.

Additionally, we run our GSM8K experiments on both Llama-3.1-8B and Llama-3.1-8B-Instruct to cover a wider variety of model sizes and variants. The observation of no correlation is still maintained, as shown in Figure 5. The absence of the NaN value in the case of the instruct variant (i.e., Llama-3.1-8B-Instruct) when computing the max activation metric for the causal mask is due to extremely low variance (i.e., around $7e-9$) in the example scores. This is in contrast to the zero variance in the cases of Llama-3.1-8B and Llama-3.2-3B.

Based on the above, we conclude that there is no reliable correlation between activation patterns and in-context example performance across the metrics we have explored. We hypothesise that this is due to the superposition phenomenon, resulting in each feature dimension representing complex, interconnected features between dimensions, requiring a non-linear filtering approach (Elhage et al.,

2022). Hence, we highlight the use of Sparse Autoencoders (SAEs) (Huben et al., 2024; Cho et al., 2025) being an alternative to the metrics used in this paper, as a promising future direction to advance further.

6 Conclusion

In this work, we showed the significantly low correlation between task performance and in-context example scores using MLP and embedding layer activations (i.e., per-layer output hidden states) across a diverse set of tasks. Thus, we argue that MLP activations should not be used to sample in-context examples without further disentangling the MLP, and that practitioners should use other sampling methods that yield more reliable scores (e.g., model uncertainty or attention weights). We hope these findings contribute to both LLMs’ explainability and active learning efforts.

Limitations

Due to computational costs and to provide finer-grained analysis, we examine only 1-shot prompts in this paper. We expect results to hold for higher shot counts, but leave it for future work to confirm this.

Acknowledgments

This work was supported by the UK Research and Innovation Centre for Doctoral Training in Machine Intelligence for Nano-electronic Devices and Systems [EP/S024298/1]. The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

References

- Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Petar Veličković, Razvan Pascanu, and Michael M. Bronstein. 2025. [Why do LLMs attend to the first token?](#) In *Second Conference on Language Modeling*.
- Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, João G.M. Araújo, Alex Vitvitskyi, Razvan Pascanu, and Petar Veličković. 2024. [Transformers need glasses! information oversquashing in language tasks](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 98111–98142. Curran Associates, Inc.

- Markus Bayer, Justin Lutz, and Christian Reuter. 2026. [ActiveLLM: Large language model-based active learning for textual few-shot scenarios](#). *Transactions of the Association for Computational Linguistics*, 14:1–22.
- Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. [LLM2vec: Large language models are secretly powerful text encoders](#). In *First Conference on Language Modeling*.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. [Towards monosemanticity: Decomposing language models with dictionary learning](#). *Transformer Circuits Thread*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Shijie Chen, Bernal Jimenez Gutierrez, and Yu Su. 2025. [Attention in large language models yields efficient zero-shot re-rankers](#). In *International Conference on Learning Representations*, volume 2025, pages 73047–73069.
- Ikhyun Cho and Julia Hockenmaier. 2025. [Toward efficient sparse autoencoder-guided steering for improved in-context learning in large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28961–28973, Suzhou, China. Association for Computational Linguistics.
- Ikhyun Cho, Gaeul Kwon, and Julia Hockenmaier. 2025. [On the versatility of sparse autoencoders for in-context learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19531–19538, Suzhou, China. Association for Computational Linguistics.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. [BoolQ: Exploring the surprising difficulty of natural yes/no questions](#). 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)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- David Cohn, Zoubin Ghahramani, and Michael Jordan. 1994. [Active learning with statistical models](#). In *Advances in Neural Information Processing Systems*, volume 7. MIT Press.
- Yihe Dong, Lorenzo Noci, Mikhail Khodak, and Mufan Li. 2025. Is random attention sufficient for sequence modeling? disentangling trainable components in the transformer. *arXiv preprint arXiv:2506.01115*.
- Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. [Active Learning for BERT: An Empirical Study](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.
- Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. [Toy models of superposition](#). *Transformer Circuits Thread*.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2023. [A framework for few-shot language model evaluation](#).
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Jucheng Hu, Suorong Yang, and Dongzhan Zhou. 2025. δ -attnmask: Attention-guided masked hidden states for efficient data selection and augmentation. *arXiv preprint arXiv:2508.09199*.
- Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. [Sparse autoencoders find highly interpretable features in language models](#). In *The Twelfth International Conference on Learning Representations*.
- Jiho Lee, Kyungjae Lee, and Eunwoo Kim. 2025. [Contrastive in-context learning with active memory for](#)

- [task planning](#). In *Workshop on Making Sense of Data in Robotics: Composition, Curation, and Interpretability at Scale at CoRL 2025*.
- Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2025. Prompt repetition improves non-reasoning llms. *arXiv preprint arXiv:2512.14982*.
- Xiaonan Li and Xipeng Qiu. 2023. [Finding support examples for in-context learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6219–6235, Singapore. Association for Computational Linguistics.
- Ziyong Lin, Haoyi Wu, Shu Wang, Kewei Tu, Zilong Zheng, and Zixia Jia. 2025. [Look both ways and no sink: Converting LLMs into text encoders without training](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22839–22853, Vienna, Austria. Association for Computational Linguistics.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022a. [What makes good in-context examples for GPT-3?](#) In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.
- Mingyi Liu, Zhiying Tu, Tong Zhang, Tonghua Su, Xiaofei Xu, and Zhongjie Wang. 2022b. [Ltp: A new active learning strategy for crf-based named entity recognition](#). *Neural Processing Letters*, 54:2433–2454.
- Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y Zhao. 2023. [Dr.ICL: Demonstration-retrieved in-context learning](#). In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.
- Vijit Malik, Atul Pande, and Anirban Majumder. 2025. [Iclforge: Enhancing in-context learning with evolutionary algorithms under budgeted annotation](#). In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM '25*, page 5931–5938, New York, NY, USA. Association for Computing Machinery.
- Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. [Active learning principles for in-context learning with large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5011–5034, Singapore. Association for Computational Linguistics.
- Costas Mavromatis, Balasubramaniam Srinivasan, Zhengyuan Shen, Jiani Zhang, Huzefa Rangwala, Christos Faloutsos, and George Karypis. 2023. Which examples to annotate for in-context learning? towards effective and efficient selection. *arXiv preprint arXiv:2310.20046*.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. [Can a suit of armor conduct electricity? a new dataset for open book question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.
- Neel Nanda, Senthooan Rajamanoharan, János Kramár, and Rohin Shah. 2023. Fact finding: Attempting to reverse-engineer factual recall on the neuron level. In *Alignment Forum*, volume 6.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>.
- Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. 2024. [In-context learning with iterative demonstration selection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7441–7455, Miami, Florida, USA. Association for Computational Linguistics.
- Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. [Learning to retrieve prompts for in-context learning](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.
- Stephanie Schoch and Yangfeng Ji. 2025. [Monte Carlo sampling for analyzing in-context examples](#). In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 63–78, Albuquerque, New Mexico. Association for Computational Linguistics.
- Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.
- Lee Sharkey, Dan Braun, and Beren Millidge. 2022. [\[interim research report\] taking features out of superposition with sparse autoencoders](#).
- Artem Shelmanov, Vadim Liventsev, Danil Kireev, Nikita Khromov, Alexander Panchenko, Irina Fedulova, and Dmitry V. Dylov. 2019. [Active learning with deep pre-trained models for sequence tagging of clinical and biomedical texts](#). In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 482–489.
- Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2025. [Repetition improves language model embeddings](#). In *The Thirteenth International Conference on Learning Representations*.

- Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. 2024. [Massive activations in large language models](#). In *First Conference on Language Modeling*.
- Pawel Swietojanski, Stefan Braun, Dogan Can, Thiago Fraga Da Silva, Arnab Ghoshal, Takaaki Hori, Roger Hsiao, Henry Mason, Erik McDermott, Honza Silovsky, Ruchir Travadi, and Xiaodan Zhuang. 2023. [Variable attention masking for configurable transformer transducer speech recognition](#). In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Simon Tong. 2001. [Active learning: Theory and applications](#). Ph.D. thesis, Stanford University.
- Wilailuck Treerath and Fabrizio Pittorino. 2026. Active in-context learning for tabular foundation models. *arXiv preprint arXiv:2603.27385*.
- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and 16 others. 2020. [SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python](#). *Nature Methods*, 17:261–272.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. [Superglue: A stickier benchmark for general-purpose language understanding systems](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Xiaorong Wang, Ting Yang, Zhu Zhang, Shuo Wang, Zihan Zhou, Liner Yang, Zhiyuan Liu, and Maosong Sun. 2025. Monocle: Hybrid local-global in-context evaluation for long-text generation with uncertainty-based active learning. *arXiv preprint arXiv:2505.20195*.
- Xinyi Wu, Amir Ajorlou, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. 2024. [On the role of attention masks and layernorm in transformers](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 14774–14809. Curran Associates, Inc.
- Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. [Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1423–1436, Toronto, Canada. Association for Computational Linguistics.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Bingsheng Yao, Guiming Chen, Ruishi Zou, Yuxuan Lu, Jiachen Li, Shao Zhang, Yisi Sang, Sijia Liu, James Hendler, and Dakuo Wang. 2024. [More samples or more prompts? exploring effective few-shot in-context learning for LLMs with in-context sampling](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1772–1790, Mexico City, Mexico. Association for Computational Linguistics.
- Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. [Compositional exemplars for in-context learning](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 39818–39833. PMLR.
- Leihan Zhang and Le Zhang. 2020. [An ensemble deep active learning method for intent classification](#). In *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence, CSAI '19*, page 107–111, New York, NY, USA. Association for Computing Machinery.
- Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. [Active example selection for in-context learning](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.