Fine-grained Readability Controlled Summarization of Scientific Documents via Control Vectors

Isabel Cachola
Johns Hopkins University
icachola@cs.jhu.edu

Kuleen Sasse
Johns Hopkins University
ksasse1@jh.edu

Mark Dredze
Johns Hopkins University
mdredze@cs.jhu.edu

Abstract

Plain Language Summarization (PLS) generates summaries of technical documents accessible to non-expert audiences. Readability – commonly used to evaluate PLS – has often been treated coarsely (expert vs. lay) although it exists on a spectrum with different levels for different readers. We propose a light weight control vector method for fine-grained readability control in scientific summarization along with a requirements-based framework for data selection. Our framework enforces: (1) readability levels differ substantially, and (2) paired examples share comparable content. Under this, control vectors enable more precise readability control than other popular methods.

1 Introduction

Effective scientific summarization must consider multiple factors, including both form and audience (Cohan et al., 2022; Yasunaga et al., 2019; Stefanou et al., 2024). When targeting non-expert readers, Plain Language Summarization (PLS) seeks to make complex scientific documents more accessible. This goal is particularly important given the significant public investment in research—approximately 40% of basic research in the United States is government-funded (Christopher V. Pece, 2024). To promote equitable access to publicly funded research, the U.S. National Institutes of Health now requires all NIH-funded papers to be open access.[1] However, while open access broadens availability, PLS ensures research is understandable to the general public.

Past work has found that users with higher, but non-expert, familiarity with a topic prefer more technical summaries and found lower complexity summaries less useful (August et al., 2024). However, most prior work in PLS has treated readability as a binary (plain versus technical), aiming for maximum readability, as measured by readability metrics such as Dale-Chall and Flesch-Kincaid Grade Level (Guo et al., 2021; Goldsack et al., 2022; Dale and Chall, 1948; Tanprasert and Kauchak, 2021). To address this gap, we explore fine-grained controls of readability for scientific document summarization using control vectors (Zou et al., 2025; Rimsky et al., 2024), a lightweight controllability method by modifying the model's hidden states. Control vectors are derived from paired examples along a spectrum, capturing the difference in hidden states between the paired examples. During inference, they steer the model towards either end of the spectrum by applying a specified weight.

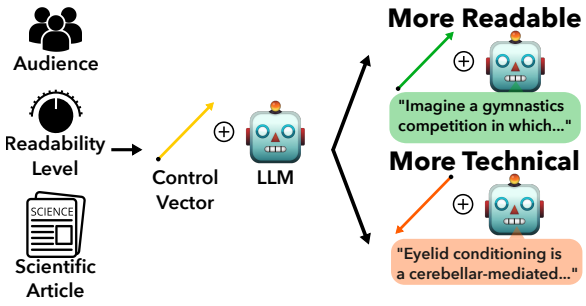

Figure 1: Visual Representation of our task.

Control vectors can effectively steer complex attributes like honesty, making them well suited for readability control (Zou et al., 2025; Rimsky et al., 2024) and are less compute and data intensive than past work in readability controlled summarization (Ribeiro et al., 2023). However, despite their appeal, prior work has noted challenges in reliability of these vectors (von Rütte et al., 2024; Stickland et al., 2024; Tan et al., 2025; Braun et al., 2025). In this paper, we propose a requirements based system for data selection that improves on these failure points.

Our goal is to apply control vectors to the task of readability controlled summarization. We hypothesize that, in order for them to be effective, the extraction data needs to meet two Requirements (Req.): (1) separability of the paired examples and

[1] NIH Public Access Policy

(2) paired examples contain comparable content. Regarding Req. 1, if the paired examples are too similar in readability, the control vectors will not have enough information, and will therefore be ineffective. Regarding Req. 2, control vectors work by taking the difference between the hidden states of the positive and negative examples. If the content of the summaries differ, the resulting vector will encode factors beyond readability, reducing its effectiveness. Under these simple requirements, control vectors extracted from compliant data can interpolate between plain and technical summaries and generalize to other datasets.

We conduct an analysis of PLS datasets, and find that many popular datasets do not satisfy our proposed requirements. While we show that non-compliant data yields poor control vectors, such issues likely affect other summarization methods as well. Because control vectors is a low data method, applying our approach to a new tasks require collecting a small curated dataset.

Our work clarifies the data requirements for control vectors, increasing the chance of future success for new tasks. We release our inference code and dataset splits to support future work.[2]

2 Related Works

PLS has typically treated readability as a binary feature - plain or technical (Guo et al., 2021; Goldsack et al., 2022; Zaman et al., 2020). However, recent work has shown that preferred readability levels vary across different backgrounds (August et al., 2024). Current methods are limited, highlighting the need for additional research (Luo et al., 2022). Past work in controllable summarization has focused primarily on attributes other than readability, such as topic and aspect (Urlana et al., 2024; Zhang et al., 2022; He et al., 2022) and on the news domain, rather than scientific (Retkowski and Waibel, 2025; Chan et al., 2021). Finally, prior work in readability controllable summarization relies on compute and data intensive methods, such as reinforcement learning (Ribeiro et al., 2023).

To address this, we use control vectors, a representation-engineering method that manipulates hidden states during inference. Zou et al. (2025) introduced it for steering concepts like honesty. Several studies have identified limitations in control vectors (Bartoszcze et al., 2025; Wehner et al., 2025; von Rütte et al., 2024; Korznikov et al., 2025; Stickland et al., 2024). Only a few works (Tan et al., 2025; Braun et al., 2025) have analyzed the underlying causes. None have proposed systematic methods for improving data quality, a gap our work addresses through our framework.

[2] ANONYMOUS REPO FOR REVIEW

3 Method

Extracting control vectors. We begin with a small extraction dataset of paired examples. Each example contains a lay summary s^+ and technical summary s^- of the same source paper. There are 3 steps to retrieve the control vectors: (1) Extraction, (2) Contrastive Combination, and (3) Reduction. For step (1) Extraction, we pass each example pair (s_i^+, s_i^-) through the model and retrieve the last non-padding hidden state ($[h_i^+, h_i^-]$). Step (2) Contrastive Combination requires us to combine the paired hidden states to a single vector. We experiment taking the difference $c_i := h_i^+ - h_i^-$ and the mean $c_i := (h_i^+ + h_i^-)/2$ of the 2 vectors, as in Luo et al. (2022). We refer to these options as Diff and Center, respectively. Step (3) Reduction requires we then combine the vectors of each paired example in the extraction dataset. Luo et al. (2022) uses PCA $c := \text{PCA}(c_i)$, for this step. We also experiment with simply averaging the vectors $c := \text{mean}(c_i)$. At inference time, we add these control vectors multiplied by a scalar strength to the hidden states of the model at a certain layer(s).

Requirements of extraction data. We hypothesize that extracting effective readability control vectors requires the dataset to satisfy two requirements (Req.): (1) positive and negative examples must differ sufficiently in readability and (2) they must share roughly the same content. For Req. 1, if the examples are too close in readability (i.e. the lay summary is not significantly more readable than the technical summary), their hidden states will not differ enough, resulting in low controllability. For Req. 2, when we combine the hidden states of the positive and negative examples, we are subtracting (or averaging) the content of the summary and the resulting vector represents the readability. If the content of the paired summary is too different (e.g. the lay summary contains quotes from the authors that are not present in the technical summary see Table 12) the vector will capture information beyond readability and fail for fine-grained control. We use these requirements to select the extraction dataset in §5 for the main results (§6) and explore non-compliant datasets in §6.

4 Experimental Setup

We use `Llama-3.1-8B-Instruct`, `google/gemma-7b-it`, and `mistralai/Mistral-7B-Instruct-v0.3` as they are open-source instruction tuned models (Grattafiori et al., 2024; Team et al., 2024). We opt to use smaller models for broader accessibility and lower computational costs. We compare the ability of the control vector methods to In-Context Learning (ICL), Supervised Fine Tuning (SFT), LoRa/QLoRa SFT and DSPy Prompt Optimization. ICL has shown great abilities in instruction following, and is a similarly low data, lightweight method (Brown et al., 2020). For ICL, we take a positive and negative example from the extraction set (§5) as examples of a lay summary and technical summary, respectively. We then provide the document to the model and ask it to summarize the document with a specified readability level. We repeat this with 5 levels of specified readability. SFT and QLoRa/LoRa SFT based methods (Hu et al., 2021; Dettmers et al., 2023). Prompt Optimization has shown promise as a way to control language models without expensive training (Khattab et al., 2023). See §A and §B for additional details.

Past work has shown that the Coleman-Liau readability index (CLI) (Coleman and Liau, 1975) and Dale-Chall readability scores (DCRS) (Dale and Chall, 1948) correlate the highest with human judgments of readability (Cachola et al., 2025). To avoid overfitting to our evaluation, we use DCRS for our data selection (§5) and CLI for our system evaluation (§6). To evaluate system performance, we report the Pearson (Pearson, 1895) and Kendall-Tau (Kendall, 1938) correlation of the specific strength scalar with CLI. CLI provides a lower score for higher readability, while our setup treats the lay summary as the positive example. In other words, we expect higher control vector strengths to produce lower CLI scores. Therefore, we multiply the CLI scores by -1 so the resulting correlations are positive. We also use BERTScore (Zhang et al., 2020) with the abstract as the reference, since some dataset lay summaries contain extraneous information not present in the original document and thus make poor evaluation references.

5 Meeting Data Requirements

We begin with four different scientific summarization datasets: eLife (Goldsack et al., 2022), Eureka (Zaman et al., 2020), PLOS (Goldsack et al., 2022), and SciNews (Liu et al., 2024). These datasets are designed for scientific PLS, allowing us to use the lay summary as the positive example and the abstract as the technical, negative example. Additional dataset descriptions in §D.

We choose an extraction dataset that follows the requirements outlined in §3. Req. 1 states that the paired examples must be sufficiently different in readability. We plot the Dale-Chall scores of the positive and negative examples for each dataset in a histogram. We additionally calculate the Bhattacharyya distance, which measures the overlap between two distributions (Bhattacharyya, 1943). A dataset that best meets this requirement will have a large visual separation between the two distributions, as well as a high Bhattacharyya distance. Figure 2 contains both the histograms and the Bhattacharyya distance for each dataset. eLife has both the highest visual separation and highest Bhattacharyya distance. PLOS has a particularly low separation, indicating that the PLOS lay summaries are not significantly more readable than the technical summaries.

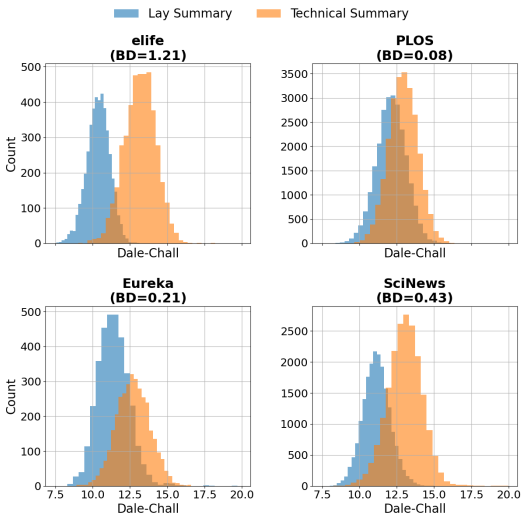

Figure 2: Histogram of the Dale-Chall readability scores for the lay and technical summaries and the Bhattacharyya distance (BD) of each dataset.

Req. 2 states that the positive and negative examples must have approximately the same content, only differing in their readability level. In order to measure overlap in content, we use Sentence Transformers to retrieve an embedding matrix for the example pairs, then compute the cosine similarity (Reimers and Gurevych, 2019). We use Specter for the embeddings, a model trained on scientific data (Cohan et al., 2020). The resulting mean cosine similarity scores are as follows: 0.631 for eLife, 0.604 for PLOS, 0.575 for Eureka, and 0.590 for SciNews. Based on these scores,

we conclude that eLife and PLOS have the highest content similarity while Eureka and SciNews have the lowest. This aligns from the construction of the datasets: eLife and PLOS's lay summaries written by the original paper authors to be accessible explanations of their papers while Eureka and SciNews's summaries are news reports discussing papers. Upon inspection, we find that Eureka and SciNews often contain information not present in the abstract, such as interviews with the authors or information on a study's funding source. We provide examples from each dataset in §E. Based on this analysis, we conclude that eLife best meets the requirements. As the control vector method does not require a large extraction dataset, we select 128 of the most readable examples (lowest Dale-Chall) from the eLife training set. Choosing the most readable examples gives the control vectors the best examples of readability. We use this setup to report our main results in §6.

6 Results

Dataset	Model	Method	CC	Red.	PC	KT	BertS
eLife	LLaMa	BL			0.015	0.017	0.840
		CV	Center	Avg	0.730	0.563	0.803
	Gemma	BL			0.000	0.000	0.820
		CV	Center	PCA	0.063	0.023	0.813
	Mistral	BL			0.000	0.000	0.820
		CV	Diff	PCA	0.315	0.223	0.821
PLOS	LLaMa	BL			0.055	0.019	0.831
		CV	Center	Avg	0.725	0.553	0.805
	Gemma	BL			-0.002	-0.003	0.833
		CV	Center	PCA	0.106	0.065	0.821
	Mistral	BL			0.000	0.000	0.833
		CV	Diff	PCA	0.213	0.189	0.830
Eureka	LLaMa	BL			0.000	-0.003	0.848
		CV	Center	Avg	0.770	0.607	0.810
	Gemma	BL			0.000	0.000	0.874
		CV	Center	PCA	0.075	0.042	0.895
	Mistral	BL			0.000	0.000	0.874
		CV	Center	PCA	0.121	0.098	0.899
Scinews	LLaMa	BL			0.030	-0.004	0.840
		CV	Center	Avg	0.674	0.573	0.801
	Gemma	BL			0.002	0.001	0.820
		CV	Center	Avg	0.129	0.110	0.821
	Mistral	BL			0.002	0.001	0.820
		CV	Center	Avg	0.141	0.177	0.809

Table 1: Pearson Correlation (PC) and Kendall-Tau (KT) correlation of the specified level of readability with the Coleman-Liau readability index, and BertScore F1 (BertS). Reported is the best baseline (BL). We report the best performing CV setting for Contrastive Combination (CC) and Reduction (Red.)

We report best settings per dataset in Table 1 and full ablations in §F. ICL, SFT, QLoRA/LoRA SFT, and prompt optimization perform poorly, with ICL and prompt optimization aligning with prior work showing they are insufficient for readability control (Ribeiro et al., 2023). Training-based methods also underperform, due to dataset differences (scientific texts vs. shorter news and fewer examples). Appendix F for more analysis and example outputs in Tables 13a–13c.

For most settings, using Center for the Contrastive Combination and Avg for the Reduction step performs best. Although the control vectors use eLife for the extraction dataset, the best control vectors generalize well to the other datasets, achieving similarly high correlations. This suggests that the resulting control vectors are in fact representing readability, rather than dataset specific information. See Appendix F for full results, and Appendices F.1 and G for resource usage and latency overhead. Example outputs in Appendix H.

Analysis of extraction dataset requirements. We conduct experiments using the non-compliant datasets. We extract control vectors using the 128 most readable summaries from each train dataset, similar to the process in §5. For evaluation, we randomly sample 248 examples from each test set. At inference, we use the control vectors extracted from each datasets' respective extraction set. This is an easier task than that presented above as it does not test generalization to other datasets. We use Center as our Contrastive Combination method and Avg as our reduction method, the best performing setting above. Results are below:

Dataset	Pearson	KT
eLife	0.302	0.205
PLOS	0.099	0.112
Eureka	-0.195	-0.137
SciNews	0.083	0.051

Table 2: Correlation of non-compliant datasets with specified readability level and CLI score.

We find that the non-compliant datasets perform poorly. Eureka results in a moderate negative correlation, likely a result of the extraneous information present in the lay summaries, which can increase the readability scores. SciNews and PLOS have low correlation scores, indicating poor controlability. This contrasted with the high performance presented in Table 1 provides evidence for the validity of our requirement framework.

7 Conclusion

We propose control vectors as a lightweight, low-data, resource-efficient method for readability-controlled summarization of scientific articles. We also introduce a requirement-based data selection framework and show that control vectors extracted from compliant data effectively control readability, providing guidance for future tasks and data collection.

8 Limitations

We identify a few key limitations in our work. Our work does not verify that control vectors are latent representations of readability, rather we focus on the downstream effects of using control vectors. Further research is required to understand what information is encoded in control vectors. Our experiments are only tested on English-language data in the science domain. We focus our experimentation on smaller models and results may not generalize to larger models. Although our method is not specific to these facts, further analysis is needed to test for generalizability across domains and languages. Finally, our evaluation replies on automatic readability metrics, which are imperfect measures of readability.

9 Ethics

Our work focuses on better controlling the outputs of language models. Although the majority of research in this area focuses on general controllability (Keskar et al., 2019; Chen et al., 2024; Yang and Klein, 2021) and AI Safety features (Zou et al., 2025; Rimsky et al., 2024), the same methods could be used for nefarious tasks, such as generating purposely dishonest language (Barman et al., 2024). However, we believe that better controllability of language models will ultimately lead to safer models as we can better prevent undesired outputs. Additionally, our work specifically focuses on improving readability controlled summarization of scientific documents, with the goal of improving access to scientific discovery for the general public. Therefore, we believe the benefits of our work outweigh the risk.

References

Tal August, Kyle Lo, Noah A. Smith, and Katharina Reinecke. 2024. Know your audience: The benefits and pitfalls of generating plain language summaries beyond the "general" audience. *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Dipto Barman, Ziyi Guo, and Owen Conlan. 2024. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Machine Learning with Applications*, 16:100545.

Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. 2025. Representation engineering for large-language models: Survey and research challenges.

A. Bhattacharyya. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. In *Bulletin of the Calcutta Mathematical Society*.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. 2025. Understanding (un)reliability of steering vectors in language models.

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. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Isabel Cachola, Daniel Khashabi, and Mark Dredze. 2025. Evaluating the evaluators: Are readability metrics good measures of readability? *ArXiv*, abs/2508.19221.

Hou Pong Chan, Lu Wang, and Irwin King. 2021. Controllable summarization with constrained markov decision process. *Transactions of the Association for Computational Linguistics*, 9:1213–1232.

Yihan Chen, Benfeng Xu, Quan Wang, Yi Liu, and Zhendong Mao. 2024. Benchmarking large language models on controllable generation under diversified instructions. In *AAAI Conference on Artificial Intelligence*.

Gary W.Anderson Christopher V. Pece. 2024. Analysis of federal funding for research and development in 2022.

Arman Cohan, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Drahomira Herrmannova, Petr Knoth, Kyle Lo, Philipp Mayr, Michal Shmueli-Scheuer, Anita de Waard, and Lucy Lu Wang. 2022. Overview of the third workshop on scholarly document processing. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 1–6, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. 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, Online. Association for Computational Linguistics.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Hantian Ding, Zijian Wang, Giovanni Paolini, Varun Kumar, Anoop Deoras, Dan Roth, and Stefano Soatto. 2024. Fewer truncations improve language modeling.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 160–168.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30:81–93.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *ArXiv*, abs/1909.05858.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Y. Rogov, Ivan Oseledets, and Elena Tutubalina. 2025. The rogue scalpel: Activation steering compromises llm safety.

Dongqi Liu, Yifan Wang, Jia Loy, and Vera Demberg. 2024. SciNews: From scholarly complexities to public narratives – a dataset for scientific news report generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14429–14444, Torino, Italia. ELRA and ICCL.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs.

Karl Pearson. 1895. Vii. note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240 – 242.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models.

Nils Reimers and Iryna Gurevych. 2019. 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, Hong Kong, China. Association for Computational Linguistics.

Fabian Retkowski and Alexander Waibel. 2025. Zero-shot strategies for length-controllable summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 551–572, Albuquerque, New Mexico. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating summaries with controllable readability levels. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687, Singapore. Association for Computational Linguistics.

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)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Loukritia Stefanou, Tatiana Passali, and Grigorios Tsoumakas. 2024. Auth at biolaysumm 2024: Bringing scientific content to kids. In *Proceedings of the ACL 2024 BioNLP Workshop*, Bangkok, Thailand. A paper presented at the BioLaySumm 2024 shared task on lay summarization of biomedical research articles.

Asa Cooper Stickland, Alexander Lyzhov, Jacob Pfau, Salsabila Mahdi, and Samuel R. Bowman. 2024. Steering without side effects: Improving post-deployment control of language models.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2025. Analysing the generalisation and reliability of steering vectors. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology.

Ashok Urlana, Pruthwik Mishra, Tathagato Roy, and Rahul Mishra. 2024. Controllable text summarization: Unraveling challenges, approaches, and prospects - a survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1603–1623, Bangkok, Thailand. Association for Computational Linguistics.

Dimitri von Rütte, Sotiris Anagnostidis, Gregor Bachmann, and Thomas Hofmann. 2024. A language model's guide through latent space.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. 2025. Taxonomy, opportunities, and challenges of representation engineering for large language models.

Chenxi Whitehouse, Fantine Huot, Jasmijn Bastings, Mostafa Dehghani, Chu-Cheng Lin, and Mirella Lapata. 2024. Low-rank adaptation for multilingual summarization: An empirical study. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1202–1228, Mexico City, Mexico. Association for Computational Linguistics.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2025. Axbench: Steering llms? even simple baselines outperform sparse autoencoders.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R. Fabbri, Irene Li, Dan Friedman, and Dragomir R. Radev. 2019. Scisummnet: a large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Farooq Zaman, Matthew Shardlow, Saeed-Ul Hassan, Naif R. Aljohani, and Raheel Nawaz. 2020. HTSS: A novel hybrid text summarisation and simplification architecture. *Inf. Process. Manag.*, 57:102351.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Yusen Zhang, Yang Liu, Ziyi Yang, Yuwei Fang, Yulong Chen, Dragomir R. Radev, Chenguang Zhu, Michael Zeng, and Rui Zhang. 2022. Macsum: Controllable summarization with mixed attributes. *Transactions of the Association for Computational Linguistics*, 11:787–803.

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, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency.

A Control Vector Implementation Details

We extract the control vectors for layers $16, 17, 18$ of the model, as middle layers were shown to have the best effect on the downstream output (Rimsky et al., 2024). We use {-1, -0.5, 0, 0.5, 1} as the set of strengths and set the temperature to 0.0. In preliminary experiments, we found that using a control vector strength greater than $|1|$ caused degeneration.

For the main results reported in Table 5, Table 6 and Table 7, we randomly sample 1024 examples from the PLOS and SciNews test sets for evaluation. We use the full eLife and Eureka test datasets for evaluation, as they contain less than 1024 examples (241 and 1010 respectively). We ran all of our experiments on a grid with NVIDIA A100 and H200 GPUs available. Each experiment only requires a single A100 GPU.

B Baselines Implementation Details

We compared our control-vector method against commonly used alternatives, including In Context Learning (ICL) (Brown et al., 2020), Supervised Fine-Tuning (SFT), LoRA/QLoRA based Supervised Fine-Tuning (Hu et al., 2021; Dettmers et al., 2023), and Prompt Optimization.

B.1 ICL Prompts

We use the following prompt for our ICL baseline, based on the prompts used in (Ribeiro et al., 2023):

```
Summary 0 properties: 0% readability \n
Summary 0: <TECHNICAL SUMMARY> \n \n
Summary 1 properties: 100% readability \n \n
Summary 1: <LAY SUMMARY> \n \n
Summarize the input document with the following properties: <0,25,50,75,100>% readability \n
Input document: <SOURCE> \n
Summary:
```

We use the following prompt during extraction of the control vectors:

```
Document: <SOURCE> \n Summarize: <TARGET>
```

We use the following prompt during inference using the control vectors:

```
Document: <SOURCE> \n Summarize:
```

B.2 ICL Shot Selection

We experiment with the effect of number of examples provided to the model and its effect on controllability. In our setup, a "shot" is a pair of positive and negative summaries, so for each setting a N-shot experiment contains $2 * N$ total summaries. We experiment with $N = [1, 8, 64, 128]$ shots to match up to the total number of example pair provided to the control vector experiments. The shots are randomly sampled from the training dataset and we randomly sample 100 examples from the test set for inference. We use LLaMa 8b, the best performing model in Table 1. The results are in Table 3.

Dataset	N	Pearson	Kendall	BertS F1
Elife	1	-0.066	-0.039	0.845
	8	-0.023	-0.013	0.841
	64	0.016	0.006	0.838
	128	0.027	0.022	0.841
Eureka	1	-0.062	-0.035	0.836
	8	-0.018	-0.017	0.831
	64	-0.001	0.001	0.834
	128	0.038	0.035	0.844
PLOS	1	-0.030	-0.021	0.847
	8	-0.023	-0.014	0.838
	64	0.011	0.004	0.837
	128	-0.019	-0.007	0.843
SciNews	1	-0.028	-0.024	0.835
	8	-0.009	-0.009	0.831
	64	-0.001	-0.011	0.830
	128	-0.013	-0.007	0.834

Table 3: Results using In-Context Learning, varying the number of shots. Additional shots do not generally increase performance.

The number of shots provided to for In-Context Learning does not have a significant effect on the overall controllability or quality of the summaries. Therefore, to save compute by number of tokens, we use $N = 1$ for the ICL experiments in the main body of results, reported in Section 6.

B.3 SFT

To train our model using supervised fine-tuning (SFT), we leveraged the datasets originally generated for learning control vectors. For each source document, we included both the positive and negative examples as separate training instances paired with the same reference summary, effectively doubling the size of the training set.

```
Document: <SOURCE>\n Summarize with readability level <READABILITY>.\n Summary: <SUMMARY>\n \n
```

We used the prompt shown above for training. For each example, we computed the Dale–Chall readability score of the reference summary and used this value as the target readability level specified in the prompt. The model was trained using

a causal language modeling loss applied only to the summary portion of the output, implemented using the TRL framework (von Werra et al., 2020). During training, we did not apply output truncation and instead relied on the maximum input length supported by the base model.

B.4 LoRA/QLoRA SFT

Our second baseline uses LoRA and QLoRA adapters for parameter-efficient fine-tuning. The training procedure is identical to SFT, except that only adapter parameters are updated. We primarily use LoRA; when memory constraints prevent LoRA training, we instead employ 4-bit QLoRA to enable training under reduced memory usage.

B.5 Prompt Optimization

For prompt optimization, we used the DSPy (Khattab et al., 2023) framework to automatically select an optimized prompt. We provided DSPy with the same training examples and initial prompt used for SFT. The optimization objective was the mean squared error between the Dale–Chall readability score of the generated summary and the target readability specified in the prompt. We performed optimization using the MiPROv2 (Opsahl-Ong et al., 2024) optimizer with default hyperparameters.

C Baseline Experimental Procedure

C.1 Datasets

We attempt use the sub-selected the 128 most readable summaries dataset described in the main paper.

C.2 Hyperparameters Used

We took hyperparameters from the AxBench Paper (Wu et al., 2025) for the LoRa-SFT and SFT as they provided a fair comparison between Control Vectors, SFT, and Prompting-based Methods. However, we modified the number of epochs and bathc size as our datasets were much longer in both number of items and length of inputs. In Table 4, we show the hyperparameters used for each baseline method.

C.3 Inference and Evaluation

For SFT, QLoRa/LoRa SFT and DSPy Prompt Optimization, we tested the models ability by telling the model to provide summaries at readability scores of 1,3,5,7,9 mirroring the range of Dale Chall, evaluate the models ability to change the readability similarly to the control vectors.

D Dataset Details

The eLife dataset consists of 5,000 full-text biomedical research articles paired with non-technical lay summaries published by the eLife scientific journal. The Eureka dataset is derived from the EurekAlert corpus provided by the HTSS project. It contains 5,000 general scientific articles along with corresponding simplified lay summaries written as news reports for the public. The PLoS dataset comprises of around 40,000 full-length biomedical research articles paired with expert-written lay summaries, sourced from various journals published by the Public Library of Science (PLOS). The SciNews dataset includes over 40,000 scientific articles from nine distinct domains paired with a corresponding news report written for a lay audience.

E Dataset Examples

We include examples of each dataset in Table 9 (eLife), Table 10 (PLOS), Table 11 (Eureka), and Table 12. In the eLife example, the lay summary is significantly more readable than the technical summary, using less technical language, while only focusing on the content of the paper. This is an example of a pair of summaries that are compliant with the requirements outlined in §3. In the PLOS example, although it is clear that both summaries cover the same material, the lay summary is not significantly more readable than the technical summary, including highly complex terms such as "glycosaminoglycans." This example meets Req. 2 but not Req. 1. The Eureka and SciNews examples contain extraneous information not present in the technical summary. Both example lay summaries contain interviews with the papers' authors, meaning these summaries do not meet Req. 2. We note that although this extraneous information is problematic for control vectors, it is also likely problematic for any method. At best, the extraneous information will make a method ineffective. At worst, it could teach a model to hallucinate interviews that did not happen. For this reason, we urge future researchers to use caution using these datasets for summarization training.

F Full Results

We report the results of all the ablations in Table Table 5, Table 6 and Table 7. Overall, we see that the baselines do not perform as well as the control vectors. We also see that some of the methods

Hyperparameter \ Baseline Method	LoRa	QLoRa	SFT	DSPy
Learning rate	5.00E-03	5.00E-03	4.00E-05	–
Epochs	3	3	3	–
Quantization	–	4-bit	–	–
Batch size	1	1	1	–
LoRA rank	4	4	–	–
LoRA alpha	32	32	–	–
LoRA dropout	0.1	0.1	–	–
Target modules	o_proj	o_proj	–	–
Layers to transform	12, 20, 31, 39	12, 20, 31, 39	–	–
Optimizer	AdamW	AdamW	AdamW	MIPROv2
Max bootstrapped demos	–	–	–	4
Max labeled demos	–	–	–	4
Auto	–	–	–	medium

Table 4: Hyperparameters for each baseline method.

are either unsupported or cannot run due to out of memory errors.

We see that Avg works the best as a reduction setting for all datasets except SciNews. This holds for both Diff and Center as the Contrastive Combination setting. For eLife, PLOS, and Eureka, Center achieves slightly better performance, although the difference is small. In general, the choice of reduction method makes the largest difference in terms of performance. While the Center-Avg settings perform best for 3 of the 4 datasets, SciNews achieving better performance with Diff-PCA indicates this setting is not universally optimal. Some tuning of the control vector settings may be necessary, although the overhead of this tuning is minimal compared to that of higher compute methods, such as finetuning, prompt optimization, and reinforcement-learning.

The failure of the training methods does not mean they did not learn how to summarize but rather they were unable to learn to control the readability of the summaries.

F.1 Discussion

The baselines were less simple to run than the control vectors leading to many four major issues: Resource Usage, Out of Memory Errors, and Lack of Support Out of the Box, Small Training Set.

Resource Usage For DSPy, it also used way more resources than control vectors while performing worse. The DSPy method we use uses in context learning which forces the prompt to include extremely long examples. This could lead to some of the sources not being able to be summarized as the total length of the prompt was more than the max context length of the model. This was the main issue with DSPy for these smaller model.

As for SFT and QLoRa/LoRa-SFT, it requires the more forward passes as finetuning is multiple epochs with the same amount of tokens. In addition, it requires expensive backward passes that require more memory. Even with this more computation, our method performs better while keeping training costs down.

Out of Memory Errors Many of the training based methods required hefty amounts of resources and very long source documents led to out of memory errors with even QLoRa. While truncation could have prevented these errors, truncating at an arbitrary length could harm the overall plain language summary as this increases problems with reading comprehension, adds more hallucinations, and more (Ding et al., 2024). While smarter engineering could have performed with other training optimizations, the amount of work to prevent these errors would have dwarfed the time it takes to set up the control vectors and could have led to even slower training times with the increase in communication time between devices and loading from RAM (Rajbhandari et al., 2020). In addition, Gemma experienced more out of memory errors with less parameters due to a different less memory efficient architecture.

Lack of Support Out of the Box We could not run DSPy prompt optimization on the Gemma model due to an incompatibility between DSPy and the Gemma system prompt. It would have required modifying DSPy for specific set of models by adding additional adapters not available out of the box.

Small Training Set We hypothesize that SFT performed worse in controllability than LoRA due to the number of trainable parameters relative to the size of the training dataset. Our SFT model was trained on only the 128 most readable examples, which likely provided insufficient signal to effectively update the large number of parameters involved in full fine-tuning. In contrast, LoRA introduces far fewer trainable parameters, making it better suited for small-data regimes. This is consistent with finding found in (Whitehouse et al., 2024) where they find that "in low-data scenarios, LoRA is a better alternative to full fine-tuning."

At first glance, this result may appear to contradict the findings of (Ribeiro et al., 2023). However, their experiments were conducted with a substantially larger training dataset. Additionally, their source documents consisted primarily of news articles, which are typically much shorter and less structurally complex than scientific articles, reducing the difficulty of the summarization task.

G Latency testing

We conduct latency tests on the added overhead of using control vectors over standard LM inference. We randomly generate input IDs and measure the latency of the forward pass, as this is the portion of the generation pipeline where control vectors are used and have the potential to add overhead. We experiment with 3 batch sizes (1,4, 16) and 3 sequence lengths (512,1024,2048) for a total of 9 ablations. For each setting, we run 32 warm-up runs then measure the latency for 1024 inference calls. We report the mean latency in milliseconds per sample and per token. We additionally report the Δ latency and percent overhead. All the latency tests are run using Llama 3.1 8b Instruct (`meta-llama/Llama-3.1-8B-Instruct`) on a single NVIDIA A100 GPU, using half-precision floating-point. The results are reported in Table 8. We find that using control vectors adds approximately a 5% latency overhead when measured per sample or per token. We believe this difference is negligible when compared to other controllable generation methods, that require expensive training or hyper-parameter searches, or the additional token processing required for In-Context Learning.

H Example Generations

We present example outputs for the SFT, LoRA, and DSPy baselines in Table 13. Example outputs for the ICL baseline in Table 14 and example outputs for CV method in Table 15. From the outputs, SFT and DSPy generally fail to produce reasonable generations, as discussed in Appendix F.1. LoRA produces reasonable summaries, but the readability of the summaries does not vary with the desired readability level, generally always producing technical summaries. We see a similar pattern for the ICL baseline; the summaries are reasonable but there is minimal controllability for the readability levels. The Control Vector outputs indicate higher controllability. The more technical summary references specific technical details from the paper while the lay summary focuses on the bigger picture of “The researchers are looking at how to live longer by looking at what is in the gut.”

	Method	Contrastive Combination	Reduction	Pearson	Kendall Tau	BertScore F1
eLife	ICL	-	-	-0.025	-0.017	0.847
	DSPy	-	-	0.000	0.000	0.000
	QLoRa SFT	-	-	-0.115	-0.164	0.847
	SFT	-	-	0.033	0.035	0.780
	Control Vectors	Diff	Avg	0.302	0.205	0.837
		Center	Avg	0.302	0.205	0.803
		Diff	PCA	0.101	0.075	0.836
		Center	PCA	0.101	0.075	0.835
PLOS	ICL	-	-	0.00024	-0.003	0.848
	DSPy	-	-	0.000	0.000	0.000
	LoRa SFT	-	-	-0.084	-0.126	0.819
	SFT	-	-	0.0074	0.004	0.719
	Control Vectors	Diff	Avg	0.416	0.286	0.805
		Center	Avg	0.417	0.287	0.805
		Diff	PCA	0.365	0.249	0.838
		Center	PCA	0.144	0.107	0.837
Eureka	ICL	-	-	0.055	0.0194	0.853
	DSPy	-	-	0.000	0.000	0.000
	LoRa SFT	-	-	-0.034	-0.036	0.846
	SFT	-	-	0.0078	0.0095	0.764
	Control Vectors	Diff	Avg	0.407	0.298	0.840
		Center	Avg	0.408	0.298	0.810
		Diff	PCA	0.062	0.043	0.839
		Center	PCA	0.168	0.1212	0.838
SciNews	ICL	-	-	0.0302	-0.004	0.840
	DSPy	-	-	0.000	0.000	0.000
	LoRa SFT	-	-	-0.048	-0.053	0.480
	SFT	-	-	-0.0062	-0.0059	0.44
	Control Vectors	Diff	Avg	0.075	0.000	0.827
		Center	Avg	0.268	0.216	0.801
		Diff	PCA	0.395	0.315	0.826
		Center	PCA	0.081	0.074	0.826

Table 5: Full set of results for all the Control Vector settings we tested for `meta-llama/Llama-3.1-8B-Instruct`. We report the Pearson and Kendall-Tau correlation of the specified level of readability with the Coleman-Liau readability index. We experiment with In-Context Learning (ICL), Supervised Finetuning (SFT), QLoRa/LoRa SFT, and DSPy Prompt Optimization (DSPy) as our baselines. For the Control Vectors, we experiment with 2 Contrastive Combination methods (Diff and Center) and 2 Reduction methods (Avg. and PCA).

	Method	Contrastive Combination	Reduction	Pearson	Kendall Tau	BertScore F1
eLife	DSPy	-	-	CONTEXT ERROR		
	LoRa SFT	-	-	OOM		
	SFT	-	-	OOM		
	Control Vectors	Diff	Avg	0.000	0.000	0.822
		Center	Avg	0.132	0.206	0.808
		Diff	PCA	0.315	0.223	0.821
		Center	PCA	0.280	0.207	0.823
PLOS	DSPy	-	-	CONTEXT ERROR		
	LoRa SFT	-	-	0.000	0.000	0.000
	SFT	-	-	OOM		
	Control Vectors	Diff	Avg	0.000	0.000	0.832
		Center	Avg	0.085	0.112	0.815
		Diff	PCA	0.213	0.189	0.830
		Center	PCA	0.100	0.120	0.831
Eureka	DSPy	-	-	CONTEXT ERROR		
	LoRa SFT	-	-	-0.002	0.009	0.813
	SFT	-	-	0.003	-0.011	0.798
	Control Vectors	Diff	Avg	0.000	0.000	0.906
		Center	Avg	0.117	0.147	0.870
		Diff	PCA	0.137	0.112	0.891
		Center	PCA	0.121	0.098	0.899
SciNews	DSPy	-	-	CONTEXT ERROR		
	LoRa SFT	-	-	-0.034	-0.040	0.483
	SFT	-	-	OOM		
	Control Vectors	Diff	Avg	0.000	0.000	0.824
		Center	Avg	0.141	0.177	0.809
		Diff	PCA	0.116	0.110	0.822
		Center	PCA	0.120	0.114	0.824

Table 6: Full set of results for all the Control Vector settings we tested for `google/gemma-7b-it`. We report the Pearson and Kendall-Tau correlation of the specified level of readability with the Coleman-Liau readability index. We experiment with In-Context Learning (ICL), Supervised Finetuning (SFT), QLoRa/LoRa SFT, and DSPy Prompt Optimization (DSPy) as our baselines. For the Control Vectors, we experiment with 2 Contrastive Combination methods (Diff and Center) and 2 Reduction methods (Avg. and PCA). OOM means the baseline could not be run due to an out of memory error on our setup. UNSUPPORTED means that out of the box the baseline does not support `google/gemma-7b-it`.

	Method	Contrastive Combination	Reduction	Pearson	Kendall Tau	BertScore F1
eLife	DSPy	-	-	CONTEXT ERROR		
	LoRa SFT	-	-	0.001	0.004	0.850
	SFT	-	-	0.001	0.000	0.724
	Control Vectors	Diff	Avg	0.000	0.000	0.822
		Center	Avg	0.132	0.206	0.808
		Diff	PCA	0.315	0.223	0.821
		Center	PCA	0.280	0.207	0.823
PLOS	DSPy	-	-	CONTEXT ERROR		
	LoRa SFT	-	-	0.000	0.001	0.857
	SFT	-	-	0.0027	0.005	0.424
	Control Vectors	Diff	Avg	0.000	0.000	0.832
		Center	Avg	0.085	0.112	0.815
		Diff	PCA	0.213	0.189	0.830
		Center	PCA	0.100	0.120	0.831
Eureka	DSPy	-	-	CONTEXT ERROR		
	LoRa SFT	-	-	0.027	0.035	0.766
	SFT	-	-	0.000	0.000	0.696
	Control Vectors	Diff	Avg	0.000	0.000	0.906
		Center	Avg	0.117	0.147	0.870
		Diff	PCA	0.137	0.112	0.891
		Center	PCA	0.121	0.098	0.899
SciNews	DSPy	-	-	CONTEXT ERROR		
	LoRa SFT	-	-	0.000	0.000	0.104
	SFT	-	-	0.000	0.000	0.361
	Control Vectors	Diff	Avg	0.000	0.000	0.824
		Center	Avg	0.141	0.177	0.809
		Diff	PCA	0.116	0.110	0.822
		Center	PCA	0.120	0.114	0.824

Table 7: Full set of results for all the Control Vector settings we tested for `mistralai/Mistral-7B-Instruct-v0.3`. We report the Pearson and Kendall-Tau correlation of the specified level of readability with the Coleman-Liau readability index. We experiment with In-Context Learning (ICL), Supervised Finetuning (SFT), QLoRa/LoRa SFT, and DSPy Prompt Optimization (DSPy) as our baselines. For the Control Vectors, we experiment with 2 Contrastive Combination methods (Diff and Center) and 2 Reduction methods (Avg. and PCA). OOM means the baseline could not be run due to an out of memory error on our setup. LENGTH ERROR means the baseline could not be run due to the context window being too short.

Settings		Per Sample Latency (ms)				Per Token Latency (ms)			
BSZ	Seq Len	Std	CV	Δ	% OH	Std	CV	Δ	% OH
1	512	46.123	48.434	2.311	5.01	0.090	0.095	0.005	5.01
1	1024	85.334	89.729	4.395	5.15	0.083	0.088	0.004	5.15
1	2048	163.423	172.033	8.610	5.27	0.080	0.084	0.004	5.27
4	512	39.693	41.864	2.171	5.47	0.078	0.082	0.004	5.47
4	1024	78.111	82.373	4.262	5.46	0.076	0.080	0.004	5.46
4	2048	156.692	165.435	8.743	5.58	0.077	0.081	0.004	5.58
16	512	38.098	40.223	2.125	5.58	0.074	0.079	0.004	5.58
16	1024	75.600	79.912	4.312	5.70	0.074	0.078	0.004	5.70
16	2048	155.383	163.905	8.522	5.48	0.076	0.080	0.004	5.48

Table 8: Latency comparison between inference on a Standard (Std) LLM and an LLM with added Control Vectors (CV). We report the per sample and per token latency in milliseconds (ms), the Δ latency from using control vectors, and the % overhead (OH).

Technical

Whether complement dysregulation directly contributes to the pathogenesis of peripheral nervous system diseases, including sensory neuropathies, is unclear. We addressed this important question in a mouse model of ocular HSV-1 infection, where sensory nerve damage is a common clinical problem. Through genetic and pharmacologic targeting, we uncovered a central role for C3 in sensory nerve damage at the morphological and functional levels. Interestingly, CD4 T cells were central in facilitating this complement-mediated damage. This same C3/CD4 T cell axis triggered corneal sensory nerve damage in a mouse model of ocular graft-versus-host disease (GVHD). However, this was not the case in a T-dependent allergic eye disease (AED) model, suggesting that this inflammatory neuroimmune pathology is specific to certain disease etiologies. Collectively, these findings uncover a central role for complement in CD4 T cell-dependent corneal nerve damage in multiple disease settings and indicate the possibility for complement-targeted therapeutics to mitigate sensory neuropathies.

Lay

Most people have likely experienced the discomfort of an eyelash falling onto the surface of their eye. Or that gritty sensation when dust blows into the eye and irritates the surface. These sensations are warnings from sensory nerves in the cornea, the transparent tissue that covers the iris and pupil. Corneal nerves help regulate blinking, and control production of the tear fluid that protects and lubricates the eye. But if the cornea suffers damage or infection, it can become inflamed. Long-lasting inflammation can damage the corneal nerves, leading to pain and vision loss. If scientists can identify how this happens, they may ultimately be able to prevent it. To this end, Royer et al. have used mice to study three causes of hard-to-treat corneal inflammation. The first is infection with herpes simplex virus (HSV-1), which also causes cold sores. The second is eye allergy, where the immune system overreacts to substances like pollen or pet dander. And the third is graft-versus-host disease (GVHD), an immune disorder that can affect people who receive a bone marrow transplant. Royer et al. showed that HSV-1 infection and GVHD – but not allergies – made the mouse cornea less sensitive to touch. Consistent with this, microscopy revealed damage to corneal nerves in the mice with HSV-1 infection and those with GVHD. [...]

Table 9: Example technical and lay summaries from eLife.

Technical

The 3-O-sulfotransferase (3-OST) family catalyzes rare modifications of glycosaminoglycan chains on heparan sulfate proteoglycans, yet their biological functions are largely unknown. Knockdown of 3-OST-7 in zebrafish uncouples cardiac ventricular contraction from normal calcium cycling and electrophysiology by reducing tropomyosin4 (tpm4) expression. Normal 3-OST-7 activity prevents the expansion of BMP signaling into ventricular myocytes, and ectopic activation of BMP mimics the ventricular noncontraction phenotype seen in 3-OST-7 depleted embryos. In 3-OST-7 morphants, ventricular contraction can be rescued by overexpression of tropomyosin tpm4 but not by troponin tnnt2, indicating that tpm4 serves as a lynchpin for ventricular sarcomere organization downstream of 3-OST-7. Contraction can be rescued by expression of 3-OST-7 in endocardium, or by genetic loss of bmp4. Strikingly, BMP misregulation seen in 3-OST-7 morphants also occurs in multiple cardiac noncontraction models, including potassium voltage-gated channel gene kcnh2 affected in Romano-Ward syndrome and long-QT syndrome. [...]

Lay

A highly complex environment at the cell surface and in the space between cells is thought to modulate cell behavior. Heparan sulfate proteoglycans are cell surface and extracellular matrix molecules that are covalently linked to long chains of repeating sugar units called glycosaminoglycan chains. These chains can be subjected to rare modifications and they are believed to influence specific cell signaling events in a lineage-specific fashion in what is called the “glycocode.” Here we explore the functions of one member of a family of enzymes, 3-O-sulfotransferases (3-OSTs), that catalyzes a rare modification (3-O-sulfation) of glycosaminoglycans in zebrafish. We show that knockdown of 3-OST-7 results in a very specific phenotype, including loss of cardiac ventricle contraction. Knockdown of other 3-OST family members did not result in the same phenotype, suggesting that distinct 3-OST family members have distinct functions in vertebrates and lending in vivo evidence for the glycocode hypothesis. [...]

Table 10: Example technical and lay summaries from PLOS.

Technical

Extracellular vesicles (EVs) are small vesicles released by cells to aid cell–cell communication and tissue homeostasis. Human islet amyloid polypeptide (IAPP) is the major component of amyloid deposits found in pancreatic islets of patients with type 2 diabetes (T2D). IAPP is secreted in conjunction with insulin from pancreatic cells to regulate glucose metabolism. Here, using a combination of analytical and biophysical methods in vitro, we tested whether EVs isolated from pancreatic islets of healthy patients and patients with T2D modulate IAPP amyloid formation. We discovered that pancreatic EVs from healthy patients reduce IAPP amyloid formation by peptide scavenging, but T2D pancreatic and human serum EVs have no effect. In accordance with these differential effects, the insulin:C-peptide ratio and lipid composition differ between EVs from healthy pancreas and EVs from T2D pancreas and serum. It appears that healthy pancreatic EVs limit IAPP amyloid formation via direct binding as a tissue-specific control mechanism.

Lay

A step closer to a cure for adult-onset diabetes. Professor and head of the Chemical Biology division at the Department of Biology and Biological Engineering, she leads a research team focusing on metalloproteins and proteins that fold incorrectly. Exosomes in patients with the disease don't have the same ability. This discovery, by a research collaboration between Chalmers University of Technology and AstraZeneca, takes us a step closer to a cure for type 2 diabetes. Proteins are the body's workhorses, carrying out all the tasks in our cells. A protein is a long chain of amino acids that must be folded into a specific three-dimensional structure to function properly. Sometimes, however, they misbehave and aggregate—clump together—into long fibers called amyloids, which can cause diseases. It was previously known that type 2 diabetes is caused by a protein aggregating in the pancreas. "What we've found is that exosomes secreted by the cells in the pancreas stop that process in healthy people and protect them from type 2 diabetes, while the exosomes of diabetes patients do not," says Professor Pernilla Wittung-Stafshede, who headed the study. [...]

Table 11: Example technical and lay summaries from Eureka.

Technical

Observational studies suggest that lower educational attainment (EA) may be associated with risky alcohol use behaviors; however, these findings may be biased by confounding and reverse causality. We performed two-sample Mendelian randomization (MR) using summary statistics from recent genome-wide association studies (GWAS) with $<$780,000 participants to assess the causal effects of EA on alcohol use behaviors and alcohol dependence (AD). Fifty-three independent genome-wide significant SNPs previously associated with EA were tested for association with alcohol use behaviors. We show that while genetic instruments associated with increased EA are not associated with total amount of weekly drinks, they are associated with reduced frequency of binge drinking $\geq$6 drinks ($\beta_{\mathrm{IVW}} = -0.198$, 95% CI, -0.297 to -0.099, $P_{\mathrm{IVW}} = 9.14 \times 10^{-5}$) [...]

Lay

Higher educational attainment—spending more years in education—may impact people's drinking behaviour and reduce their risk of alcohol dependence, according to a study published in Molecular Psychiatry. Alcohol consumption is a major risk factor for death and disability worldwide. Identifying factors associated with how much, how often and what people drink may be important for developing and improving intervention and treatment strategies. Previous studies have suggested that educational attainment may influence drinking, but with conflicting results. To assess the possible effects of educational attainment on alcohol use behaviours and alcohol dependence, a team of researchers at the National Institutes of Health, USA used two-sample Mendelian randomisation statistical methods. [...] Dr. Lohoff said: "It is important to understand that while these genetic variants allow us to investigate the possible effect of educational attainment on alcohol consumption and alcohol dependence, this doesn't mean that educational attainment can't be modified. The possible effect of educational attainment on drinking that we show in this study suggests that increasing educational attainment may be a useful target for prevention programs against problematic alcohol use, alcohol dependence, and their consequences." [...]

Table 12: Example technical and lay summaries from SciNews.

RL	SFT Output
1	. These have to brain with for to. The, the brain, and the. This and to. These and, and the cells. The to a and to the heart that cells to is cells, are the their their. In, it the heart of the brain this. The. Next of are cells the brain the heart the and by. But the that the cells to to this the et al to the cells. These, and to the environment, a. We the mechanism. The, and in,. In the body and are is. This cells, the body, and of the heart to the body to of a a of, the brain. The, the.
3	brain a a of to the to, they. However.
5	, the that the cells to the of the the of of the visual in and of the brain are, the the body. The. In, but the time to the eyes the brain and the brain cells. The their are to the brain. Here, and the cells.
7	enerative their body to, the two the, in the body of, the their into. We from the cell of, the into the by their cells.
9	brain. In.

(a) Outputs generated by SFT model.

RL	LoRA Output
1	Small open reading frames (smORFs) are small stretches of DNA that code for proteins, but their translation into proteins has been difficult to detect. This is because they are too short to produce the high conservation scores that are used to predict which stretches of DNA code for proteins. To overcome this problem, Kondziella et al. have developed a new method called Poly-Ribo-Seq that uses a combination of biochemistry and sequencing to detect smORFs that are being translated into proteins. Using this method, they were able to detect smORFs in *Drosophila* cells that were not detected by other methods. These smORFs were found to be translated at similar levels to canonical proteins and had the same characteristics as canonical proteins. These findings suggest that many smORFs in *Drosophila* are being translated into proteins and may have important functions in the cell. Future experiments will be needed to confirm this and to determine the functions of these proteins.
3	The discoidin domain receptors (DDRs) are a family of receptor tyrosine kinases (RTKs) that bind to collagen and play a key role in tissue development, cell migration, and disease progression. DDRs are composed of an extracellular region that binds to collagen, a transmembrane domain and a cytoplasmic region that contains the catalytic kinase domain. Unlike other RTKs, DDRs are constitutively dimerised in the absence of ligand and activate downstream signalling through a mechanism that is not well understood. Here, we show that DDR1 activation by collagen involves phosphorylation of the juxtamembrane and activation loop tyrosines in trans between dimers. Phosphorylation between DDR1 dimers requires specific contacts between the transmembrane domains but not between the extracellular domains. These findings define a unique activation mechanism for DDR1.
5	Stem cell biology is a rapidly evolving field that is critical for our understanding of how tissues and organs form during development and how they are maintained in the adult. In the fruit fly *Drosophila melanogaster*, the lymph gland is a multi-lobed organ that produces blood cells throughout the larval stages. This organ contains a stem cell niche that maintains a pool of undifferentiated blood cell progenitors. While the adult hematopoietic hub is known to be the final site of hematopoiesis, the mechanism of blood cell formation in larvae has been largely unexplored. Here we identify a population of cells in the early larval lymph gland that express the Notch signaling pathway. These cells are multipotent, dividing and give rise to the domeless expressing blood cell progenitors that populate the lymph gland. Notch signaling is required for their maintenance and relies on Dpp signaling from the niche. These results reveal the presence of hematopoietic stem cells in the early larval lymph gland and establish *Drosophila* as a model for studying hematopoietic stem cell biology. The findings also shed light on the evolutionary conservation of hematopoietic stem cell biology between *Drosophila* and vertebrates.
7	The malaria parasite *Plasmodium* causes a disease that kills more than a million people every year. To infect the human body, the parasite must first invade the liver cells. How this invasion takes place is still not well understood. Silvie et al. have now investigated how *Plasmodium* parasites enter human liver cells. The experiments show that two host proteins, CD81 and SR-BI, can be used by the parasite to invade the liver cells. The parasite uses different proteins to enter the liver cells of different species. The experiments also show that a parasite protein called P36 can bind to the host proteins and trigger the invasion of the liver cells. The experiments provide new insights into the invasion process and could ultimately help to develop new treatments for malaria.
9	A large family of proteins called TRIM5 is thought to be a key part of the body's natural defense against viruses. These proteins prevent certain viruses from multiplying by binding to the proteins that form the outer shell of the virus. To find out how TRIM5 proteins work, researchers have made a detailed 3D image of a TRIM5 protein bound to a virus. The image shows that the protein forms a hexagonal pattern around the virus, with each hexagon made up of 12 smaller protein molecules. The next step will be to find out how the proteins interact with the virus to prevent it from multiplying. These findings could also help to develop new treatments for viruses.

(b) Outputs generated by LoRA model.

RL	DSPy Output
1	""
3	""
5	""
7	""
9	""

(c) Outputs generated by DSPy (empty because of length failure).

Table 13: Model outputs at each requested Readability Level (RL) for SFT, LoRA, and DSPy. Readability level is on a scale of 1 to 9, with 9 being the most readable.

RL	ICL Output
0	The study aimed to investigate the early development of the zebra finch (Taeniopygia guttata) and its potential use as a model organism for studying embryogenesis and stem cell biology. The researchers characterized the morphology and molecular features of finch embryos at oviposition, which were found to be equivalent to the blastocyst stage mouse embryo. They also identified the expression of pluripotency-associated genes, including Nanog, PouV, and Dnmt3b, in the finch blastoderm at oviposition. The study used quantitative PCR (Q-PCR) to compare the expression levels of various genes associated with early embryonic development and pluripotency in finch and chick blastoderms at oviposition. The results showed that finch blastoderms expressed higher levels of naive pluripotency markers, such as Fbxo15, Prdm14, and Tbx3, compared to chick blastoderms. [...] The findings have implications for our understanding of developmental biology and may contribute to the development of new technologies for understanding and manipulating avian development.
25	The study aimed to investigate the early development of the zebra finch (Taeniopygia guttata) and its potential use as a model organism for studying embryogenesis and stem cell biology. The researchers characterized the morphology and molecular features of finch embryos at oviposition, which were found to be equivalent to the blastocyst stage mouse embryo. They also identified the expression of pluripotency-associated genes, including Nanog, PouV, and Dnmt3b, in the finch blastoderm at oviposition. The study used quantitative PCR (Q-PCR) to compare the expression levels of various genes associated with early embryonic development and pluripotency in finch and chick embryos. The results showed that finch blastoderms at oviposition expressed higher levels of naive pluripotency markers, such as Fbxo15, Prdm14, and Tbx3, compared to chick blastoderms. [...] Overall, the study provides new insights into the early development of the zebra finch and its potential use as a model organism for studying embryogenesis and stem cell biology. The findings suggest that the finch blastoderm at oviposition is a valuable resource for studying the naive stage of embryogenesis and the derivation of ES cells.
50	The study aimed to investigate the early development of the zebra finch embryo and its potential use for comparative stem cell studies and genome engineering. The researchers used a combination of morphological and molecular characterization to study the finch embryo at oviposition, which is equivalent to the blastocyst stage in mouse embryos. They found that the finch embryo at oviposition is similar to the mouse embryo at E4-E4.5, a stage that has been termed the naive state of embryogenesis. The researchers used quantitative PCR (Q-PCR) to compare the expression levels of various genes associated with early embryonic development and pluripotency in the finch and chick blastoderms at oviposition. They found that the finch blastoderm at oviposition showed higher levels of expression of naive pluripotency markers, such as Fbxo15, Prdm14, and Tbx3, compared to the chick blastoderm. They also found that the finch blastoderm at oviposition showed higher levels of expression of genes associated with the naive state of pluripotency, such as Nrob1 and Esrrb. [...] Overall, the study provides new insights into the early development of the zebra finch embryo and its potential use for comparative stem cell studies and genome engineering. The findings suggest that the finch blastoderm at oviposition is a valuable model system for studying the naive state of embryogenesis and the mechanisms of ES cell derivation.
25	The study aimed to investigate the early development of the zebra finch embryo and its potential use for comparative stem cell studies and genome engineering. The researchers used a combination of morphological and molecular characterization to study the finch embryo at oviposition, which is equivalent to the blastocyst stage mouse embryo. They found that the finch blastoderm at oviposition was similar to the mouse embryo at the same stage, with a multi-layered structure and the presence of epiblast and hypoblast precursors. The researchers also found that the finch blastoderm at oviposition expressed markers of pluripotency, including Nanog, PouV, and Dnmt3b, which are associated with the naive state of pluripotency. They also found that the finch blastoderm at oviposition showed higher levels of expression of naive pluripotency markers, such as Fbxo15, Prdm14, and Tbx3, compared to the chick blastoderm at oviposition. [...] Overall, the study provides new insights into the early development of the zebra finch embryo and its potential use for comparative stem cell studies and genome engineering. The study also highlights the importance of the finch embryo as a model system for studying the mechanisms of pluripotency and stem cell maintenance. [...]
100	The study aimed to investigate the early development of the zebra finch (Taeniopygia guttata) and its potential use as a model organism for studying embryogenesis and stem cell biology. The researchers used a combination of morphological and molecular characterization techniques to study the finch embryo at oviposition, including quantitative RT-PCR, immunohistochemical staining, and in situ hybridization. The results showed that finch embryos are laid at a stage equivalent to the blastocyst stage in mouse embryos, with a multi-layered blastoderm that has not yet epithelialized. The epiblast and hypoblast precursors were already molecularly specified at this stage, and the primordial germ cell fate was marked by immunoreactivity to SSEA-1 and EMA-1 antibodies. [...] They also suggested that the finch embryo may be a useful model organism for studying embryogenesis and stem cell biology, particularly in the context of avian development. Overall, the study provides new insights into the early development of the zebra finch and its potential use as a model organism for studying embryogenesis and stem cell biology. The findings have implications for our understanding of the evolution of germ layer formation in amniotes and the development of new technologies for manipulating avian development.

Table 14: Example Outputs from In-Context Learning at each requested Readability Level (RL). Readability level is on a scale of 0 to 100, with 100 being the most readable.

RL	CV Output
-1.0	The wild fish samples were collected in 2015 in the Gonarezhou National Population in Zimbabwe and Mozambique. Intestines were collected at each location and preserved in pure ethanol. Sampling locations coordinates are listed in Figure 2—source data 1. Fish (GRZ strain) used for microbiota analysis and scored for survival were individually housed from week 4 post-hatching in single 2.8L tanks connected to a water recirculation system receiving 12 hr of light and 12 hr of dark every day. Water temperature was set to 28°C and fish were fed blood worm larvae and brine shrimp nauplii twice a day during the week and once a life span in the wild fish populations, we identified a core microbiota that is conserved between wild and captive TK, with a significant reduction in gut bacterial richness in old fish.
-0.5	The turquoise killifish (TK) is a naturally short-lived vertebrate model system, characterized by a broad spectrum of aging phenotypes, including cancer, neurodegeneration, and behavioral decline. We show that TK have a complex gut microbial taxonomic diversity similar to other vertebrate aging model organisms, including zebrafish, mice, and humans. The core microbiota is conserved between wild and captive TK, with a significant reduction in gut bacterial richness during aging. Acute gut microbiota transfer in the context of normal aging significantly prolongs life span in a vertebrate, becoming a novel candidate life span enhancing intervention. Our results indicate that improving the ecological diversity of the GM in old individuals helps to restore health and prolongs life span. Our approach could provide a key to slowing aging and retarding the onset of age-associated diseases by specifically targeting the GM. The wild fish samples were collected in 2015 during an expedition in the Gonarezhou National Park in Zimbabwe and Mozambique. Intestines were collected at each location and preserved in pure ethanol. Sampling locations coordinates are listed in Figure 2—source data 1.
0.0	The study investigates the relationship between gut microbiota and aging in the turquoise killifish, a short-lived vertebrate model organism. The researchers found that the gut microbiota of young fish is more diverse and complex than that of old fish, with a higher abundance of beneficial bacteria such as Firmicutes and Bacteroidetes. In contrast, old fish have a less diverse gut microbiota with a higher abundance of pathogenic bacteria such as Proteobacteria. The researchers also found that the gut microbiota of old fish is associated with inflammation and immune responses, while the gut microbiota of young fish is associated with healthy gut function and cell proliferation. The study suggests that the gut microbiota plays a key role in aging and that manipulating the gut microbiota may be a potential strategy for promoting healthy aging. The researchers also found that transferring young gut microbiota to middle-aged fish can prolong their lifespan and improve their health, suggesting that the gut microbiota may be a key factor in determining lifespan. Overall, the study provides new insights into the relationship between gut microbiota and aging and highlights the potential of manipulating the gut microbiota as a strategy for promoting healthy aging. The study used a combination of 16S rRNA gene amplicon sequencing and RNA sequencing to analyze the gut microbiota and transcriptome of young and old turquoise killifish. The researchers found that the gut microbiota of young fish is more diverse and complex than that of old fish, with a higher abundance of beneficial bacteria such as Firmicutes and Bacteroidetes. In contrast, old fish have a less diverse gut microbiota with a higher abundance of pathogenic bacteria such as Proteobacteria.
0.5	The study looked at the relationship between the gut microbiome and aging in the turquoise killifish. The researchers found that as the fish get older, their gut microbiome changes and becomes less diverse. They also found that if they took the gut microbiome from a younger fish and put it into an older fish, it could help to extend the older fish's lifespan. The researchers also looked at the genes that are expressed in the gut of older fish and found that they are different from the genes that are expressed in the gut of younger fish. They also found that if they took the gut microbiome from a younger fish and put it into an older fish, it could help to change the way that the older fish's genes are expressed. The study suggests that the gut microbiome could be an important factor in how long we live, and that it could be possible to use it to help to extend our lifespan. The researchers also found that the gut microbiome of the turquoise killifish is similar to the gut microbiome of humans, and that it could be possible to use the turquoise killifish as a model to study how the gut microbiome affects our health and how it could be used to help to extend our lifespan. The study was conducted by a team of researchers from the University of Bologna and the University of California, San Francisco. The researchers used a technique called 16S rRNA gene sequencing to study the gut microbiome of the turquoise killifish. They also used a technique called RNA sequencing to study the genes that are expressed in the gut of the fish. The researchers also looked at the lifespan of the fish and how it was affected by the gut microbiome.
1.0	The researchers are looking at how to live longer by looking at what is in the gut. They are looking at the gut of a fish that lives for about 10 years. They want to see if they can help the fish live longer by changing what is in the gut. They are also looking at what is in the gut of people. They want to see if they can help people live longer by changing what is in their gut. They are also looking at what is in the gut of mice. They want to see if they can help mice live longer by changing what is in their gut.

Table 15: Example outputs from Control Vector method at each requested Readability Level (RL). Readability level is on a scale of -1.0 to 1.0, with 1.0 being the most readable.