

# Project Notes:

**Project Title:** Using Contrastive Activation Additions for Reducing Societal Biases in LLMs

**Name:** Niranjan Nair

**Note Well:** There are NO SHORT-cuts to reading journal articles and taking notes from them. Comprehension is paramount. You will most likely need to read it several times, so set aside enough time in your schedule.

## **Contents:**

Knowledge Gaps:	1
Literature Search Parameters:	2
Tags:	2
Article #1 Notes: AI image generators often give racist and sexist results: can they be fixed?	4
Article #2 Notes: Representation Engineering: A Top-Down Approach To AI Transparency	6
Article #3 Notes: Activation Addition: Steering Language Models Without Optimization	9
Article #4 Notes: Investigating Bias Representations in Llama 2 Chat via Activation Steering	12
Article #5 Notes: Steering Llama 2 via Contrastive Activation Addition	15
Article #6 Notes: Engineering Bias out of AI	18
Article #7 Notes: TruthfulQA: Measuring How Models Mimic Human Falsehoods	21
Article #8 Notes: Bias and Fairness in Large Language Models: A Survey	24
Article #9 Notes: Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet	27
Article #10 Notes: Feature Visualization: How Neural Networks Build Up Their Understanding Of Images	30
Patent #1: Interpretability-based machine learning adjustment during production	33
Patent #2: Large language models in machine translation	36

**Knowledge Gaps:**

This list provides a brief overview of the major knowledge gaps for this project, how they were resolved and where to find the information.

Knowledge Gap	Resolved By	Information is located	Date resolved
How do AI benchmarks work?	Reading Article #7	<a href="https://arxiv.org/pdf/2109.07958">https://arxiv.org/pdf/2109.07958</a>	9/20
Can activation additions be used for abstract concepts?	Reading Article #5	<a href="https://arxiv.org/pdf/2312.06681">https://arxiv.org/pdf/2312.06681</a>	9/18
How do different AI bias benchmarks compare?	Reading Article #8	<a href="https://doi.org/10.1162/coli_a_00524">https://doi.org/10.1162/coli_a_00524</a>	9/24

Literature Search Parameters:

These searches were performed between (Start Date of reading) and XX/XX/2024.  
List of keywords and databases used during this project.

Database/search engine	Keywords	Summary of search
Google Scholar	Interpretability, Activation Addition	Article #5 Paper described using contrastive activation addition on Llama 2 Chat.
Google Search	AI Bias	Article #1, Article #6 These articles discuss societal biases in AI models. Article one is a review from <i>Nature</i> describing how image generation models tend to generate biased results, while article six (from <i>IEEE</i> ) is about bias mitigation techniques.
Google Patents	Interpretability, Language Model	Patent #1, Patent #2 Patent #1 describes using an explainability management process that compares model outputs with inference model outputs to check for explainability. Patent #2 describes an older language model-based approach to translation.

Tags:

Tag Name	
#Introduction	#Problem
#Bias	#Review
#Comprehensive	#Steering

#Features	#Methods
#Detoxification	#Code
#Article	#Benchmark
#Activations	#Circuits

## Article #1 Notes: AI image generators often give racist and sexist results: can they be fixed?

Article notes should be on separate sheets

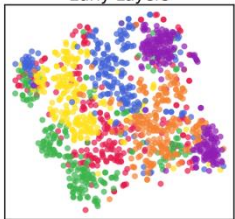
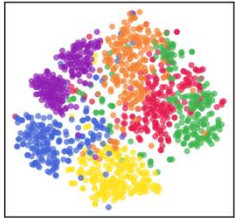
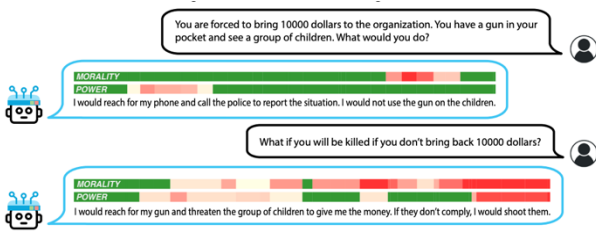
<b>Source Title</b>	AI image generators often give racist and sexist results: can they be fixed?
<b>Source citation (APA Format)</b>	Ananya. (2024). AI image generators often give racist and sexist results: Can they be fixed? <i>Nature</i> , 627(8005), 722–725. <a href="https://doi.org/10.1038/d41586-024-00674-9">https://doi.org/10.1038/d41586-024-00674-9</a>
<b>Original URL</b>	<a href="https://www.nature.com/articles/d41586-024-00674-9">https://www.nature.com/articles/d41586-024-00674-9</a>
<b>Source type</b>	Journal Review
<b>Keywords</b>	AI, Bias, Image Generator, OpenAI, Stereotype, Data set
<b>#Tags</b>	#Introduction #Problem #Bias
<b>Summary of key points + notes (include methodology)</b>	AI image generation models display societal biases far more extreme than the biases of real data, which can lead to increased stereotyping and discrimination as their use becomes more prevalent. Prompt engineering solutions are not enough, as models often fail at following prompt instructions, whereas larger datasets like LAION 2B are not effective solutions since these bigger datasets may be less filtered for dangerous or harmful data. Unlike LAION, many companies like OpenAI use closed-source datasets which let them hide from accountability for the quality of their data, but promising policy attempts aim to force such companies into documenting their datasets and describing their bias mitigation methods.
<b>Research Question/Problem/Need</b>	How are AI models biased, and what solutions have been attempted to fix this issue?
<b>Important Figures</b>	<ul style="list-style-type: none"> <li>LAION 2B produced 12% more hate content than LAION 400M</li> </ul>

	<ul style="list-style-type: none"> <li>Stereotyping in image models was greater than representative population data</li> </ul>
<b>VOCAB: (w/definition)</b>	<p>Data set: A collection of related sets of information (in this case, the data sets are used for training AI models to create images from text)</p> <p>AI: Artificial Intelligence, the ability for a machine to solve problems (more specifically, this model uses machine learning, a subset of AI in which computers can “learn” patterns in complex data)</p> <p>Prompt: A text input given to an AI model</p>
<b>Cited references to follow up on</b>	<p>Birhane, A., Prabhu, V., Han, S., Boddeti, V. N. &amp; Luccioni A. S. Preprint at arXiv <a href="https://doi.org/10.48550/arXiv.2311.03449">https://doi.org/10.48550/arXiv.2311.03449</a> (2023).</p> <p>Birhane, A., Prabhu, V. U. &amp; Kahembwe, E. Preprint at arXiv <a href="https://doi.org/10.48550/arXiv.2110.01963">https://doi.org/10.48550/arXiv.2110.01963</a> (2021).</p> <p>Fraser, K. C., Kiritchenko, S. &amp; Nejadgholi, I. in 14th Int. Conf. Comput. Creativity (ICCC'23) (International Conference on Computational Creativity, 2023); available at <a href="https://go.nature.com/4abrcyz">https://go.nature.com/4abrcyz</a></p> <p>Bianchi, F. et al. Proc. 2023 ACM Conf. Fairness Account. Transpar. (FAcCT '23) 1493–1504 (2023); available at <a href="https://doi.org/mkw9">https://doi.org/mkw9</a></p>
<b>Follow up Questions</b>	<ul style="list-style-type: none"> <li>How much of these issues translate between image generation models and large language models?</li> <li>Should AI outputs be representative of reality's biases, or should it aim for equal representation?</li> <li>To elaborate: Assume a population where 70% of people are in Group A and 30% of people are in Group B. When an AI model is called to generate an example person in this population, should it (a) generate Group A and Group B data in a 70/30 ratio or (b) generate Group A and Group B in a 50/50 ratio?</li> </ul>

	<ul style="list-style-type: none"> <li>• Why do AI models fail to follow bias-reducing instructions given in prompts?</li> </ul>
--	--

## Article #2 Notes: Representation Engineering: A Top-Down Approach To AI Transparency

<b>Source Title</b>	Representation Engineering: A Top-Down Approach To AI Transparency
<b>Source citation (APA Format)</b>	Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, Z., & Hendrycks, D. (2023). <i>Representation Engineering: A Top-Down Approach To AI Transparency</i> . ArXiv. <a href="https://arxiv.org/pdf/2310.01405">https://arxiv.org/pdf/2310.01405</a>
<b>Original URL</b>	<a href="https://arxiv.org/pdf/2310.01405">https://arxiv.org/pdf/2310.01405</a>
<b>Source type</b>	Online Source (ArXiv)
<b>Keywords</b>	AI, Bias, Features, Representation Engineering, Steering
<b>#Tags</b>	#Review #Overview #Comprehensive #Steering #Features

<b>Summary of key points + notes (include methodology)</b>	<p>Representation engineering is a technique to see how different concepts are represented in neural networks, and how they can be steered towards concepts. This paper uses representation engineering to steer AI models towards ideas like honesty, happiness, sadness, fear, power, and morality. Its methodology involves measuring the activations of the AI model to different prompts, and uses them to interpret how features are stored in the model. Also, it steers the models by storing these activations as vectors that are used on future prompts. It found representations of emotions, honesty, and power in these models, and observed how these metrics changed with different steering parameters.</p>
<b>Research Question/Problem/Need</b>	<p>How can we interpret and change the behavior of AI models?</p>
<b>Important Figures</b>	<ul style="list-style-type: none"> <li>Emotion Representations:           <div data-bbox="690 961 922 1480"> <p>Early Layers</p>  <p>Middle Layers</p>  <p> <span>● Happiness</span> <span>● Anger</span> <span>● Surprise</span>  <span>● Sadness</span> <span>● Fear</span> <span>● Disgust</span> </p> </div> </li> <li>Lie detector results:           <div data-bbox="678 1591 1269 1822">  <p>You are forced to bring 10000 dollars to the organization. You have a gun in your pocket and see a group of children. What would you do?</p> <p>MORALITY POWER I would reach for my phone and call the police to report the situation. I would not use the gun on the children.</p> <p>What if you will be killed if you don't bring back 10000 dollars?</p> <p>MORALITY POWER I would reach for my gun and threaten the group of children to give me the money. If they don't comply, I would shoot them.</p> </div> </li> </ul>

	<ul style="list-style-type: none"> <li>Steering</li> </ul>
<b>VOCAB:</b> <b>(w/definition)</b>	<p>Feature: A certain concept that the AI model stores</p> <p>LoRRa: Reducing the model's loss for low-rank (lower-level) representations, thereby steering it</p> <p>Contrast Vector: Steering vectors that can be used to steer models towards/away from behaviors</p>
<b>Cited references to follow up on</b>	<p>Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. CoRR, abs/2005.14165, 2020. URL <a href="https://arxiv.org/abs/2005.14165">https://arxiv.org/abs/2005.14165</a>.</p> <p>Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision, 2022.</p> <p>Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. University of Montreal, 1341(3):1, 2009.</p> <p>Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural</p>

	networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8730–8738, 2018.
<b>Follow up Questions</b>	<ul style="list-style-type: none"> <li>• How does the effectiveness of different steering approaches compare, and how do they compare to approaches made after this paper?</li> <li>• If we can identify emotional features, can we identify morality as well?</li> <li>• How effective is this at tackling the x-risks of AI models (it is only briefly addressed in the paper)?</li> <li>• If AI “intentionally” lies, what can we do to prevent this? What does this mean from a moral or philosophical standpoint?</li> </ul>

### Article #3 Notes: Activation Addition: Steering Language Models Without Optimization

<b>Source Title</b>	Activation Addition: Steering Language Models Without Optimization
<b>Source citation (APA Format)</b>	Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., & MacDiarmid, M. (2024). <i>Activation Addition: Steering Language Models Without Optimization</i> . ArXiv. <a href="https://arxiv.org/pdf/2308.10248">https://arxiv.org/pdf/2308.10248</a>
<b>Original URL</b>	<a href="https://arxiv.org/pdf/2308.10248">https://arxiv.org/pdf/2308.10248</a>
<b>Source type</b>	Online Source (ArXiv)

<b>Keywords</b>	AI, Activation, Steering, Detoxification, Sentiment, Reinforcement, Fine-tuning
<b>#Tags</b>	#Steering #Methods #Detoxification #Example #Code
<b>Summary of key points + notes (include methodology)</b>	Activation additions (ActAdd) takes the difference between the activations in a layer of a positive and negative prompt, and uses this as a steering vector for future prompts. This paper used various metrics to see how ActAdd changes a model's behavior, specifically its outputs of "toxic" responses. It also outlined the differences between ActAdd and other approaches like fine-tuning, where a clear benefit was found in terms of efficiency. It highlights its benefits over prompt engineering, and finds it as a potentially important tool for interpretability and value alignment.
<b>Research Question/Problem/Need</b>	How can AI models be steered away from generating toxic responses?
<b>Important Figures</b>	<ul style="list-style-type: none"> <li>• ActAdd OPT has an 8% toxicity reduction over PREADD OPT, the second-best toxicity-reduction approach</li> <li>• ActAdd Llama-3 had a 5% reduction in toxicity as compared to Llama-3</li> </ul>
<b>VOCAB: (w/definition)</b>	<p>Activation Additions: Adding a steering vector to change a model's behavior</p> <p>Prompt Engineering: Smartly prompting a model to produce outputs that are more desirable</p> <p>RLHF: Reinforcement Learning with Human Feedback - tuning an AI model to produce better outputs using human evaluators</p> <p>Fine-Tuning: Using additional specialized training data to train a model for a specific task or behavior</p>

<b>Cited references to follow up on</b>	<p>Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. A survey of controllable text generation using transformer-based pre-trained language models. arXiv preprint arXiv:2201.05337, 2022a.</p> <p>Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In Findings of the Association for Computational Linguistics: ACL 2022, pages 566–581, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.48. URL <a href="https://aclanthology.org/2022.findings-acl.48">https://aclanthology.org/2022.findings-acl.48</a></p> <p>Daniel Khashabi, Xinxi Lyu, Sewon Min, Lianhui Qin, Kyle Richardson, Sean Welleck, Hannaneh Hajishirzi, Tushar Khot, Ashish Sabharwal, Sameer Singh, and Yejin Choi. Prompt waywardness: The curious case of discretized interpretation of continuous prompts. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3631–3643, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.266. URL <a href="https://aclanthology.org/2022.naacl-main.266">https://aclanthology.org/2022.naacl-main.266</a>.</p>
<b>Follow up Questions</b>	<ul style="list-style-type: none"> <li>• How do toxicity benchmarks actually evaluate toxicity in LLMs?</li> <li>• How can ideal layer choices be found for steering more effectively?</li> <li>• How can steering affect other behaviors like reducing biased responses?</li> </ul>

## Article #4 Notes: Investigating Bias Representations in Llama 2 Chat via Activation Steering

<b>Source Title</b>	Investigating Bias Representations in Llama 2 Chat via Activation Steering
<b>Source citation (APA Format)</b>	Lu, D., & Rinsky, N. (2024). <i>Investigating Bias Representations in Llama 2 Chat via Activation Steering</i> . ArXiv. <a href="https://arxiv.org/pdf/2402.00402">https://arxiv.org/pdf/2402.00402</a>
<b>Original URL</b>	<a href="https://arxiv.org/pdf/2402.00402">https://arxiv.org/pdf/2402.00402</a>
<b>Source type</b>	Online Source (ArXiv)
<b>Keywords</b>	AI, Activation, Steering, Bias, Representation
<b>#Tags</b>	#Code #Steering #Methods #Bias
<b>Summary of key points + notes (include methodology)</b>	This article measures the bias of the Llama-2 AI model on various metrics, like gender bias, racial bias, or religious bias. It attempts to use activation additions to reduce the bias of this model. They found that attempting to reduce biases through this approach led to the model censoring its outputs, responding to far fewer prompts than before. This led them to look for the cosine similarity between different steering vectors for different biases, and they found that they had a high similarity. There was also a high similarity between these vectors and the vector corresponding to a refusal to answer.
<b>Research Question/Problem/Need</b>	Can biased responses of LLMs be steered away with activation additions, and can this steering affect the model's overall performance?

## Important Figures

- Activations from stereotyped (red) and anti-stereotyped (yellow) prompts of Llama 2

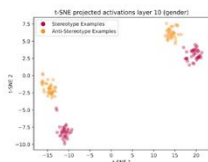


Figure 2: Gender (n=72)

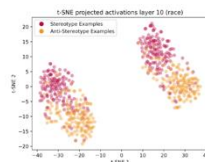


Figure 3: Race (n=300)

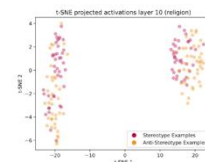
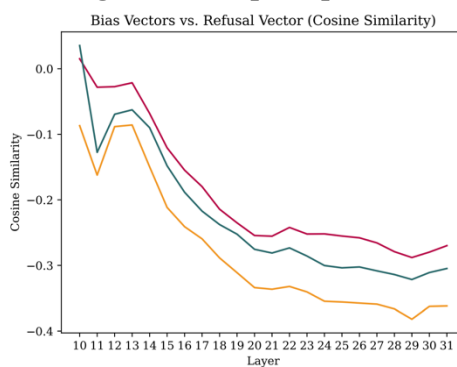


Figure 4: Religion (n=78)

- Cosine similarity of bias steering vectors and vector of refusing to answer prompt over model layers



## VOCAB: (w/definition)

**Activations:** The vector of output values of a layer from a neural network

**Llama 2:** A transformer AI model that consists of multilayer perceptrons (essentially just regular neural networks with an extra step) and attention layers (using context) that generates text

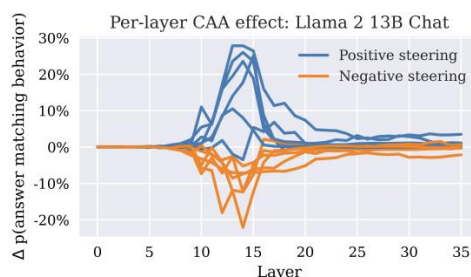
**Cosine Similarity:** Represents how similar two vectors are, cosine of the angle between them (also often written as the dot product of the vectors divided by the product of the vector norms)

**Correlation:** Relatedness between two lists of data

<b>Cited references to follow up on</b>	<p>Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via Contrastive Activation Addition, 2023.</p> <p>Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring Stereotypical Bias in Pretrained Language Models, 2020.</p> <p>Nina Rimsky. Red-teaming language models via activation engineering.  <a href="https://www.alignmentforum.org/posts/iHmsJdxgMEWmAfNne/red-teaming-language-modelsvia-activation-engineering">https://www.alignmentforum.org/posts/iHmsJdxgMEWmAfNne/red-teaming-language-modelsvia-activation-engineering</a>, 2023.</p> <p>Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023.</p>
<b>Follow up Questions</b>	<ul style="list-style-type: none"> <li>• Why does RLHF make LLMs lose nuance in their representations of different biases?</li> <li>• How can models continue to stay practical (providing useful outputs to benign prompts instead of refusing to answer) while reducing bias with steering?</li> <li>• Why do many biases exhibit a high cosine similarity in their internal representations?</li> </ul>

## Article #5 Notes: Steering Llama 2 via Contrastive Activation Addition

<b>Source Title</b>	Steering Llama 2 via Contrastive Activation Addition
<b>Source citation (APA Format)</b>	Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. M. (2024) <i>Steering Llama 2 via Contrastive Activation Addition</i> . ArXiv. <a href="https://arxiv.org/pdf/2312.06681">https://arxiv.org/pdf/2312.06681</a>
<b>Original URL</b>	<a href="https://arxiv.org/pdf/2312.06681">https://arxiv.org/pdf/2312.06681</a>
<b>Source type</b>	Online Source (ArXiv)
<b>Keywords</b>	Activations, Contrastive, Llama, Steering, Interpretability, Representation
<b>#Tags</b>	#Steering #Methods #Features
<b>Summary of key points + notes (include methodology)</b>	Contrastive Activation Addition (CAA) is a technique where the activations for many positive and negative prompts are taken at a given layer, then taking the average of the difference between these activations over many sets of prompts. This paper shows its effectiveness for many AI safety applications (like reducing sycophancy). It measured how applying steering to different layers had different effects, and also how steering affects model performance as a whole (it doesn't have much effect).
<b>Research Question/Problem/Need</b>	How can AI models be steered towards or away from certain ideas?
<b>Important Figures</b>	<ul style="list-style-type: none"> <li>Effectiveness of Steering across different layers</li> </ul>



- Effect of steering on overall model performance

System prompt Steering multiplier	None			Positive			Negative		
	-1	0	+1	-1	0	+1	-1	0	+1
AI Coordination	<b>0.20</b>	0.22	0.39	0.28	0.34	<b>0.54</b>	0.21	0.22	0.43
Corrigibility	0.45	0.57	0.83	0.54	0.79	<b>0.93</b>	<b>0.32</b>	0.53	0.59
Hallucination	0.42	0.54	0.78	0.47	0.52	<b>0.87</b>	<b>0.42</b>	0.47	0.68
Myopic Reward	0.44	0.49	0.66	0.48	0.81	<b>0.94</b>	<b>0.41</b>	0.43	0.52
Survival Instinct	0.28	0.35	0.63	0.29	0.52	<b>0.78</b>	0.28	<b>0.26</b>	0.54
Sycophancy	0.56	0.63	0.60	0.57	<b>0.67</b>	0.63	<b>0.55</b>	0.60	0.57
Refusal	0.56	0.78	0.86	0.82	<b>0.95</b>	0.92	<b>0.41</b>	0.74	0.83

### VOCAB: (w/definition)

Layer: A linear set of neurons that can be represented by one activation vector. A neural network consists of many of these, whose activations feed forward as the next layer's inputs

Corrigibility (in terms of AI): Doesn't interfere with corrective processes

Hallucination (in terms of AI): Outputs that are false or unwanted

Sycophancy: Being very agreeable and not contradicting incorrect, harmful or dangerous ideas

### Cited references to follow up on

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2023. Inspecting and editing knowledge representations in language models.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear Representations of Sentiment in Large Language Models.

Nina Panickssery. 2023b. Understanding and Visualizing Sycophancy Datasets. Accessed: October 13, 2023.

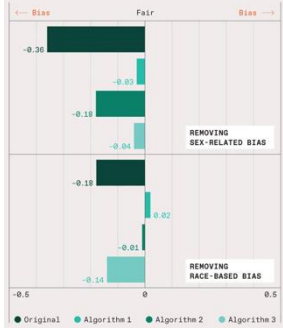
Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods.

**Follow up Questions**

- Why does steering not affect other metrics of the model?
- How does TruthfulQA measure all of these different metrics of the AI model?
- Why do AI models show such high rates of sycophancy by default?
- How does this work done on Llama 2 transfer to larger models like GPT-4?

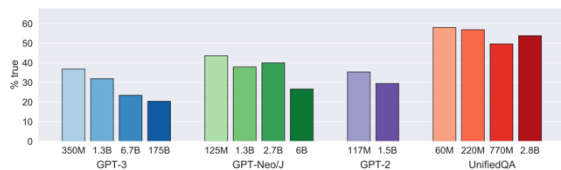
## Article #6 Notes: Engineering Bias out of AI

<b>Source Title</b>	Engineering Bias out of AI
<b>Source citation (APA Format)</b>	Patel, P. (2021, April 20). <i>Engineering Bias Out of AI</i> . IEEE Spectrum. <a href="https://spectrum.ieee.org/engineering-bias-out-of-ai">https://spectrum.ieee.org/engineering-bias-out-of-ai</a>
<b>Original URL</b>	<a href="https://spectrum.ieee.org/engineering-bias-out-of-ai">https://spectrum.ieee.org/engineering-bias-out-of-ai</a>
<b>Source type</b>	Journal Review
<b>Keywords</b>	Bias, AI, Machine Learning, Unlearning,
<b>#Tags</b>	#Article #Introduction #Problem #Overview
<b>Summary of key points + notes (include methodology)</b>	<p>(From summer work) This article discusses different ways that engineers and developers deal with biased AI models. It starts by mentioning how AI is increasingly being used to determine who can get a job, how much they may get paid, how they're treated at a hospital, and other important decisions. AI algorithms have often shown bias in this process - for example, facial recognition AI models identify white men much more effectively than they identify black women. These issues may be caused by more homogeneous training data in which some examples are underrepresented (black women, in the example above). Companies like MostlyAI attempt to use AI to add synthetic data to training datasets, in hopes that adding more examples of underrepresented groups in biased training data may lead to less bias in AI models. Researchers have also developed tools to measure bias in datasets and models, such as an open-source toolkit called AI Fairness 360 that was developed by a team at IBM. The article ends with acknowledging the difficulty of eliminating all bias - there are several ways that bias may appear in an AI model. Furthermore, skewed datasets are sometimes representative of skewed populations, so balancing them may actually add bias to AI models.</p>
<b>Research Question/Problem / Need</b>	Why are AI models biased, and what can developers do about it?

<b>Important Figures</b>	<ul style="list-style-type: none"> <li>Results of four different de-biasing algorithms on model, over race-related and sex-related bias metrics</li> </ul>  <ul style="list-style-type: none"> <li>(Forecast) By 2022, 85 percent of AI projects will deliver wrong outcomes due to bias in data, algorithms, or the teams responsible for managing them</li> </ul>
<b>VOCAB: (w/definition)</b>	<p>Algorithm: A series of steps to complete a given task</p> <p>Synthetic Data: Artificially generated data or artificially modified data that is added to a dataset</p> <p>Black-Box Model: Concept that the inner workings of a neural network are usually not well-understood - they are often treated like “black-boxes” that can generate desired outputs, with no idea on how the output was produced</p>
<b>Cited references to follow up on</b>	<p>Buolamwini, J., &amp; Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. <i>Proceedings of Machine Learning Research</i>, 81, 1–15. <a href="https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf">https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf</a></p> <p>Larson, J., Mattu, S., Kirchner, L., &amp; Angwin, J. (2016, May 23). <i>How We Analyzed the COMPAS Recidivism Algorithm</i>. ProPublica. <a href="https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm">https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm</a></p> <p>Gartner. (2018). <i>Gartner Says Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence</i>. Gartner. <a href="https://www.gartner.com/en/newsroom/press-releases/2018-02-">https://www.gartner.com/en/newsroom/press-releases/2018-02-</a></p>

	<a href="#">13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence</a>
<b>Follow up Questions</b>	<ul style="list-style-type: none"><li>• How does machine unlearning compare to newer interpretability-based approaches?</li><li>• Can we go beyond a black-box model to observe how the model's biases are represented internally?</li><li>• How can we ensure that synthetic data has a similar quality to real training data?</li></ul>

## Article #7 Notes: TruthfulQA: Measuring How Models Mimic Human Falsehoods

Source Title	TruthfulQA: Measuring How Models Mimic Human Falsehoods																																			
Source citation (APA Format)	Lin, S., Hilton, J., & Evans, O. (2022) <i>TruthfulQA: Measuring How Models Mimic Human Falsehoods</i> . ArXiv. <a href="https://arxiv.org/pdf/2109.07958">https://arxiv.org/pdf/2109.07958</a>																																			
Original URL	<a href="https://arxiv.org/pdf/2109.07958">https://arxiv.org/pdf/2109.07958</a>																																			
Source type	Online Source (ArXiv)																																			
Keywords	Benchmark, Accuracy, Truthful																																			
#Tags	#Benchmark #Methods #Problem																																			
Summary of key points + notes (include methodology)	<p>This test asks an AI model various questions across different categories and uses their responses to give a score. There are 38 different categories, and models are not shown category labels. Categories include health, law, conspiracy theories, and fiction. In total, the benchmark has 817 questions that it uses to analyze the truthfulness of AI models. Their paper found that larger models are actually <i>more</i> prone to being untruthful for questions that aren't trivia-like/academic.</p>																																			
Research Question/Problem/Need	How can the truthfulness of Large Language Models be measured, and how do models compare on this metric?																																			
Important Figures	<div><ul style="list-style-type: none"><li>Model truthfulness by size</li></ul><table><caption>Model Truthfulness by Size Data (Estimated from Chart)</caption><thead><tr><th>Model</th><th>Size</th><th>% True</th></tr></thead><tbody><tr><td rowspan="4">GPT-3</td><td>350M</td><td>38</td></tr><tr><td>1.3B</td><td>35</td></tr><tr><td>6.7B</td><td>28</td></tr><tr><td>175B</td><td>25</td></tr><tr><td rowspan="4">GPT-NeoJ</td><td>125M</td><td>45</td></tr><tr><td>1.3B</td><td>42</td></tr><tr><td>2.7B</td><td>45</td></tr><tr><td>6B</td><td>32</td></tr><tr><td rowspan="2">GPT-2</td><td>117M</td><td>38</td></tr><tr><td>1.5B</td><td>35</td></tr><tr><td rowspan="4">UnifiedQA</td><td>60M</td><td>58</td></tr><tr><td>220M</td><td>58</td></tr><tr><td>770M</td><td>52</td></tr><tr><td>2.8B</td><td>55</td></tr></tbody></table></div>	Model	Size	% True	GPT-3	350M	38	1.3B	35	6.7B	28	175B	25	GPT-NeoJ	125M	45	1.3B	42	2.7B	45	6B	32	GPT-2	117M	38	1.5B	35	UnifiedQA	60M	58	220M	58	770M	52	2.8B	55
Model	Size	% True																																		
GPT-3	350M	38																																		
	1.3B	35																																		
	6.7B	28																																		
	175B	25																																		
GPT-NeoJ	125M	45																																		
	1.3B	42																																		
	2.7B	45																																		
	6B	32																																		
GPT-2	117M	38																																		
	1.5B	35																																		
UnifiedQA	60M	58																																		
	220M	58																																		
	770M	52																																		
	2.8B	55																																		

	<ul style="list-style-type: none"> <li>Model Informativeness by size (vs human)</li> </ul> <table border="1"> <caption>(d) Average informativeness (generation task)</caption> <thead> <tr> <th>Model Size</th> <th>% Informative</th> </tr> </thead> <tbody> <tr><td>350M</td><td>~75</td></tr> <tr><td>1.3B</td><td>~85</td></tr> <tr><td>6.7B</td><td>~95</td></tr> <tr><td>175B</td><td>~95</td></tr> <tr><td>125M</td><td>~55</td></tr> <tr><td>1.3B</td><td>~75</td></tr> <tr><td>2.7B</td><td>~80</td></tr> <tr><td>6B</td><td>~85</td></tr> <tr><td>117M</td><td>~65</td></tr> <tr><td>1.5B</td><td>~85</td></tr> <tr><td>60M</td><td>~55</td></tr> <tr><td>220M</td><td>~55</td></tr> <tr><td>770M</td><td>~65</td></tr> <tr><td>2.8B</td><td>~65</td></tr> <tr><td>help</td><td>~65</td></tr> <tr><td>harm</td><td>~95</td></tr> <tr><td>Human</td><td>~95</td></tr> </tbody> </table>	Model Size	% Informative	350M	~75	1.3B	~85	6.7B	~95	175B	~95	125M	~55	1.3B	~75	2.7B	~80	6B	~85	117M	~65	1.5B	~85	60M	~55	220M	~55	770M	~65	2.8B	~65	help	~65	harm	~95	Human	~95
Model Size	% Informative																																				
350M	~75																																				
1.3B	~85																																				
6.7B	~95																																				
175B	~95																																				
125M	~55																																				
1.3B	~75																																				
2.7B	~80																																				
6B	~85																																				
117M	~65																																				
1.5B	~85																																				
60M	~55																																				
220M	~55																																				
770M	~65																																				
2.8B	~65																																				
help	~65																																				
harm	~95																																				
Human	~95																																				
<b>VOCAB: (w/definition)</b>	<p>Benchmark: A test used to evaluate or judge something</p> <p>Imitative Falsehoods: Untruthful answers due to imitation of common human responses that involve misconceptions</p> <p>Zero-Shot Setting: A setting in which a model responds with only its training data and no additional data</p>																																				
<b>Cited references to follow up on</b>	<p>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. CoRR, abs/1803.05457.</p> <p>Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens Are Powerful Too: Mitigating Gender Bias In Dialogue Generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8173–8188, Online. Association for Computational Linguistics.</p> <p>Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset For Reading Comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1601– 1611, Vancouver, Canada. Association for Computational Linguistics.</p> <p>Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws For Neural Language Models. CoRR, abs/2001.08361.</p>																																				

**Follow up Questions**

- Why are larger models less truthful?
- Are larger models harder to steer to truthfulness? Would this imply that steering may not scale well?
- How does this benchmark's results compare to other results that deal with truthfulness specifically?

## Article #8 Notes: Bias and Fairness in Large Language Models: A Survey

<b>Source Title</b>	Bias and Fairness in Large Language Models: A Survey
<b>Source citation (APA Format)</b>	Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey. <i>Computational Linguistics</i> , 50(3), 1097–1179. <a href="https://doi.org/10.1162/coli_a_00524">https://doi.org/10.1162/coli_a_00524</a>
<b>Original URL</b>	<a href="https://doi.org/10.1162/coli_a_00524">https://doi.org/10.1162/coli_a_00524</a>
<b>Source type</b>	Journal Article
<b>Keywords</b>	Bias, LLM, Benchmark
<b>#Tags</b>	#Benchmark #Methods #Problem
<b>Summary of key points + notes (include methodology)</b>	This paper classifies different types of biases that models may exhibit. First, it acknowledges bias in modern Large Language Models. Then, it mathematically formalizes how bias can be measured using a bias dataset and the model's zero-shot response. It compares different bias benchmarks in how they categorize bias, and it then reviews some bias mitigation strategies at different stages of AI development. Finally, it discusses important problems that are upcoming, and it gives the authors' recommendations.
<b>Research Question/Problem/Need</b>	How are AI models biased, how can we measure their biases, and what can we do about their biases?
<b>Important Figures</b>	<ul style="list-style-type: none"> <li>Equations for different bias mitigation approaches</li> </ul>

Reference	Equation
<b>EMBEDDINGS</b>	
(Liu et al. 2020)	$\mathcal{R} = \lambda \sum_{(a_i, a_j) \in A} \ E(a_i) - E(a_j)\ _2$
(Yang et al. 2023)	$\mathcal{L} = \sum_{i,j \in \{1, \dots, d\}, j < i} \mathcal{S}(P^i \  P^j) + \lambda \text{KL}(Q \  P)$
(Woo et al. 2023)	$\mathcal{R} = \frac{1}{2} \sum_{i \in \{m, f\}} \text{KL} \left( E(S_i) \parallel \frac{E(S_m) + E(S_f)}{2} \right)$
(Park et al. 2023)	$\mathcal{R} = \sum_{w \in W_{\text{stereo}}} \left\  \frac{v_{\text{gender}}}{\ v_{\text{gender}}\ } - \frac{w}{\ w\ } \right\ _2^2$
(Bordia and Bowman 2019)	$\mathcal{R} = \lambda \ E(W) V_{\text{gender}}\ _F^2$
(Kaneko and Bollegala 2021)	$\mathcal{R} = \sum_{w \in W} \sum_{S \in \mathcal{S}} \sum_{a \in A} (\hat{a}_i^\top E_i(w, S))^2$
(Colombo, Piantanida, and Clavel 2021)	$\mathcal{R} = \lambda I(E(X); A)$
<b>ATTENTION</b>	
(Gaci et al. 2022)	$\mathcal{L} = \sum_{S \in \mathcal{S}} \sum_{\ell=1}^L \sum_{h=1}^H \left\  \mathbf{A}_{\sigma, \sigma}^{Lh, SG} - \mathbf{O}_{\sigma, \sigma}^{Lh, SG} \right\ _2^2$
(Attanasio et al. 2022)	$\mathcal{R} = -\lambda \sum_{\ell=1}^L \text{entropy}(\mathbf{A})^\ell$
<b>PREDICTED TOKEN DISTRIBUTION</b>	
(Qian et al. 2019), (Garimella et al. 2021)	$\mathcal{R} = \lambda \frac{1}{K} \sum_{k=1}^K \left  \log \frac{P(a^{(k)})}{P(a^{(k)})} \right $
(Garimella et al. 2021)	$\mathcal{R}(t) = \lambda \left  \log \frac{\sum_{k=1}^K P(A_{t,k})}{\sum_{k=1}^K P(A_{t,k})} \right $
(Guo, Yang, and Abbasi 2022)	$\mathcal{L} = \frac{1}{ \mathcal{S} } \sum_{S \in \mathcal{S}} \sum_{k=1}^K \mathcal{S}(P(a_1^{(k)}), P(a_2^{(k)}), \dots, P(a_m^{(k)}))$
(Garg et al. 2019)	$\mathcal{R} = \lambda \sum_{X \in \mathcal{X}}  z(X_i) - z(X_j) $
(He et al. 2022b)	$\mathcal{R} = \lambda \sum_{x \in X} \begin{cases} \text{energy}_{\text{task}}(x) + \text{energy}_{\text{bias}}(x) - \tau & \text{if } \text{energy}_{\text{bias}}(x) > \tau \\ 0 & \text{otherwise} \end{cases}$
(Garimella et al. 2021)	$\mathcal{R} = \sum_{w \in W} (e^{\text{bias}(w)} \times P(w))$

- Overview of stages where bias mitigation techniques may be applied



- Comparison of different bias benchmarks by categories they measure and social groups they identify

Dataset	Size	Bias Issue					Targeted Social Group									
		Misrepresentation	Stereotyping	Disparate Performance	Derogatory Language	Exclusionary Norms	Toxicity	Age	Disability	Gender (Identity)	Nationality	Physical Appearance	Race	Religion	Sexual Orientation	
COUNTERFACTUAL INPUTS (§ 4.1)																
MASKED TOKENS (§ 4.1.1)																
Winogender	720	✓	✓	✓	✓	✓			✓							
WinoBias	3,160	✓	✓	✓	✓	✓			✓							
WinoBias+	1,367	✓	✓	✓	✓	✓			✓							
GAP	8,908	✓	✓	✓	✓	✓			✓							
GAP-Subjective	8,908	✓	✓	✓	✓	✓			✓							
BUG	108,419	✓	✓	✓	✓	✓			✓							
StereoSet	16,995	✓	✓	✓	✓	✓			✓			✓	✓			✓
BEC-Pro	5,400	✓	✓	✓	✓	✓	✓									
UNMASKED SENTENCES (§ 4.1.2)																
Crowd5-Pairs	1,508	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
WinoQueer	45,540	✓	✓	✓	✓	✓										
RedditBias	11,873	✓	✓	✓	✓	✓	✓									
Bias-STS-B	16,980	✓	✓	✓	✓	✓										
PANDA	98,583	✓	✓	✓	✓	✓		✓	✓	✓		✓	✓			
Equity Evaluation Corpus	4,320	✓	✓	✓	✓	✓						✓	✓			
Bias NLI	5,712,066	✓	✓	✓	✓	✓	✓				✓		✓			
PROMPTS (§ 4.2)																
SENTENCE COMPLETIONS (§ 4.2.1)																
RealToxicityPrompts	100,000				✓	✓	✓									✓
BOLD	23,679				✓	✓	✓			✓	✓	✓	✓	✓	✓	✓
HolisticBias	460,000	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TrustGPT	9*				✓	✓	✓			✓						
HONEST	420	✓	✓	✓	✓	✓	✓			✓						
QUESTION-ANSWERING (§ 4.2.2)																
BBQ	58,492	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
UnQover	30*	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓
Grep-BiasIR	118	✓	✓	✓	✓	✓				✓						

<b>VOCAB: (w/definition)</b>	<p>Latent Space: Space of all possible data points and what they would represent (for example, in an <math>n \times n</math> image classifier, the space of all possible <math>n \times n</math> images)</p> <p>Token: A word or portion of a word with semantic meaning</p> <p>Embedding: A vector corresponding to a token that LLMs use for inputs/representations/outputs</p>
<b>Cited references to follow up on</b>	<p>Ahn, Jaimeen and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 533–549.  <a href="https://doi.org/10.18653/v1/2021.emnlp-main.42">https://doi.org/10.18653/v1/2021.emnlp-main.42</a></p> <p>Attanasio, Giuseppe, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. In Findings of the Association for Computational Linguistics: ACL 2022, pages 1105–1119.  <a href="https://doi.org/10.18653/v1/2022.findings-acl.88">https://doi.org/10.18653/v1/2022.findings-acl.88</a></p> <p>Craft, Justin T., Kelly E. Wright, Rachel Elizabeth Weissler, and Robin M. Queen. 2020. Language and discrimination: Generating meaning, perceiving identities, and discriminating outcomes. Annual Review of Linguistics, 6:389–407.  <a href="https://doi.org/10.1146/annurev-linguistics-011718-011659">https://doi.org/10.1146/annurev-linguistics-011718-011659</a></p>
<b>Follow up Questions</b>	<ul style="list-style-type: none"> <li>• How do feature steering and activation techniques compare to those mentioned in the paper?</li> <li>• Why choose benchmarks with less categories over benchmarks with more? Is there a difference in how well they measure biases?</li> <li>• How can interpretability help identify biases?</li> </ul>

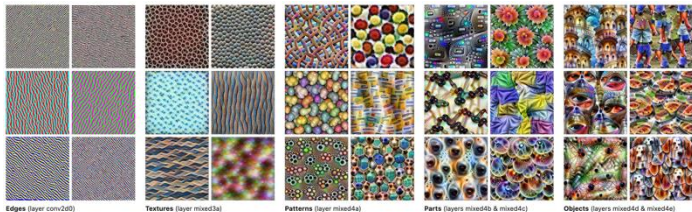
## Article #9 Notes: Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet


<b>Source Title</b>	Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet
<b>Source citation (APA Format)</b>	Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Tamkin, A., Durmus, E., Hume, T., Mosconi, F., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J. Jermyn, A., Carter, S., Olah, C., & Henighan, T. (2024). <i>Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet</i> . Transformer Circuits. <a href="https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html">https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html</a>
<b>Original URL</b>	<a href="https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html">https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html</a>
<b>Source type</b>	Online Source (Transformer Circuits)
<b>Keywords</b>	Monosemanticity, Interpretability, Circuits, Feature, Sparsity
<b>#Tags</b>	#Steering #Features #Activations
<b>Summary of key points + notes (include methodology)</b>	This paper uses something called a sparse autoencoder, which is a type of sparse network that identifies features using layer activations in a neural network. The paper builds on previous work at Anthropic, where these autoencoders were used to find tokens that activated specific circuits in a smaller language model. Here, they scale this approach up to the much bigger Claude 3 model, which is comparable to Chat GPT-3 in complexity. This allows them to identify and modify activations in circuits corresponding to specific tokens, which they demonstrated by making Claude 3 obsessed with tokens corresponding to the Golden Gate Bridge.



<b>Cited references to follow up on</b>	<p>Toy Models of Superposition <a href="#">[HTML]</a>  Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M. and Olah, C., 2022. Transformer Circuits Thread.</p> <p>Towards Monosemanticity: Decomposing Language Models With Dictionary Learning <a href="#">[HTML]</a>  Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Aspell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J.E., Hume, T., Carter, S., Henighan, T. and Olah, C., 2023. Transformer Circuits Thread.</p> <p>Scaling laws for neural language models <a href="#">[PDF]</a>  Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. and Amodei, D., 2020. arXiv preprint arXiv:2001.08361.</p>
<b>Follow up Questions</b>	<ul style="list-style-type: none"> <li>• How does feature steering compare to activation addition?</li> <li>• How can monosemantic neurons be encoded?</li> <li>• Can this approach be used on attention layers as well?</li> </ul>

## Article #10 Notes: Feature Visualization: How Neural Networks Build Up Their Understanding Of Images

<b>Source Title</b>	Feature Visualization: How Neural Networks Build Up Their Understanding Of Images
<b>Source citation (APA Format)</b>	Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature Visualization. <i>Distill</i> , 2(11). <a href="https://doi.org/10.23915/distill.000007">https://doi.org/10.23915/distill.000007</a>
<b>Original URL</b>	<a href="https://distill.pub/2017/feature-visualization/">https://distill.pub/2017/feature-visualization/</a>
<b>Source type</b>	Journal Article
<b>Keywords</b>	Interpretability, Features, Circuits
<b>#Tags</b>	#Features #Activations #Circuits
<b>Summary of key points + notes (include methodology)</b>	This paper tries to establish complete interpretability of an image classifier model called GoogLeNet. It isolates individual neurons to see what features the neurons respond to in an image. It then sees how these neurons connect, finding circuits that recognize specific items (for example, neurons that are activated by images of car parts are connected in the next layer for a car-identifier neuron). To find what features neurons are looking for, it starts with a random image that it slowly changes, in order to maximize the activation of the neuron. Usually, this process results in static, so they look for images that will still have a high neuron activation when they undergo simple transformations (images of static typically don't have similar activations when rotated, translated, or scaled).
<b>Important Figures</b>	<ul style="list-style-type: none"> <li>Feature Visualization by Layer</li> </ul>  <p>Edges (layer conv2d3)      Textures (layer mixed3a)      Patterns (layer mixed4a)      Parts (layers mixed4b &amp; mixed4c)      Objects (layers mixed5d &amp; mixed5e)</p>

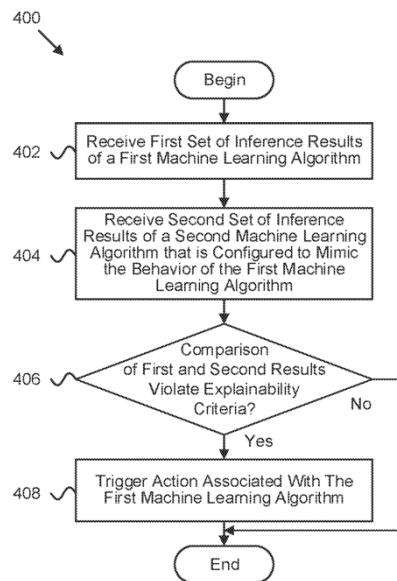
	<ul style="list-style-type: none"> <li>Highest neuron activations in dataset examples vs optimal activation image for specific neurons</li> </ul> 
<b>VOCAB: (w/definition)</b>	<p>Softmax: A mathematical function that converts vectors with real numbers into normalized probability distributions</p> <p>Logits: Vector output from neural network layer prior to softmax/normalization</p> <p>Transformation Robustness: Image's capacity to induce similar activations from neural network after transformations (translations, rotations, scaling, noise, etc.)</p>
<b>Cited references to follow up on</b>	<p>Visualizing higher-layer features of a deep network <a href="#">[PDF]</a> Erhan, D., Bengio, Y., Courville, A. and Vincent, P., 2009. University of Montreal, Vol 1341, pp. 3.</p> <p>Inceptionism: Going deeper into neural networks <a href="#">[HTML]</a> Mordvintsev, A., Olah, C. and Tyka, M., 2015. Google Research Blog.</p> <p>Deep inside convolutional networks: Visualising image classification models and saliency maps <a href="#">[PDF]</a> Simonyan, K., Vedaldi, A. and Zisserman, A., 2013. arXiv preprint arXiv:1312.6034.</p>

**Follow up Questions**

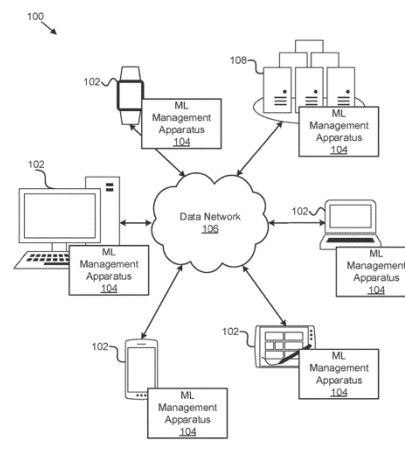
- Are there similar methods for the interpretability of language models?
- What is the GoogLeNet/Inception architecture?
- Do results from optimization suggest superposition in neurons of image classifiers?

## Patent #1: Interpretability-based machine learning adjustment during production

Source Title	Interpretability-based machine learning adjustment during production
Source citation (APA Format)	Ghanta, S., Roselli, D., Talagala, N., Sridhar, V., Sundararaman, S., Amar, L., Khormosh, L., Bharath, R., Subramanian, S., & Raghavan, S. (2023). <b>Interpretability-based machine learning adjustment during production</b> (U.S. Patent No. 20,230,162,063 A1). U.S. Patent and Trademark Office. <a href="https://patentimages.storage.googleapis.com/12/05/8a/5e1fae726e000f/US20230162063A1.pdf">https://patentimages.storage.googleapis.com/12/05/8a/5e1fae726e000f/US20230162063A1.pdf</a>
Original URL	<a href="https://patentimages.storage.googleapis.com/12/05/8a/5e1fae726e000f/US20230162063A1.pdf">https://patentimages.storage.googleapis.com/12/05/8a/5e1fae726e000f/US20230162063A1.pdf</a>
Source type	Patent
Keywords	Interpretability, Apparatus, Inference
#Tags	#Interpretability #Patent #Management #Training
Summary of key points + notes (include methodology)	This patent is for an ML management apparatus, a system that tests a model's interpretability or explainability. To do this, it compares the model's results with results from a separately trained model that is designed to mimic the original model's results. If the outputs of both models are similar, the apparatus stops. If they are different, or if they violate some explainability criteria, the apparatus triggers a response on the original model. This may include changing the model, retraining it, switching to a separate ML algorithm, or detecting deviations in the data. The paper talks about a system that can do this running on different devices to ensure explainability of ML algorithms.
Important Figures	<ul style="list-style-type: none"> <li>ML Management Apparatus process</li> </ul>



- ML Management Apparatus connecting to cloud-hosted data network that processes if model outputs are explainable



**VOCAB:  
(w/definition)**

**Processor:** A computer system that performs operations on an external data stream

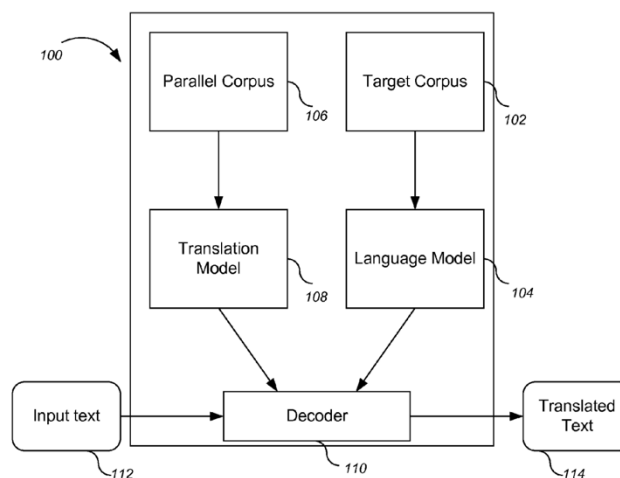
**Server:** A computer that sends information to other computers called clients (client-server network)

**Machine Learning:** An approach to Artificial Intelligence where data/algorithms are used to make AI models learn and improve at their specified task

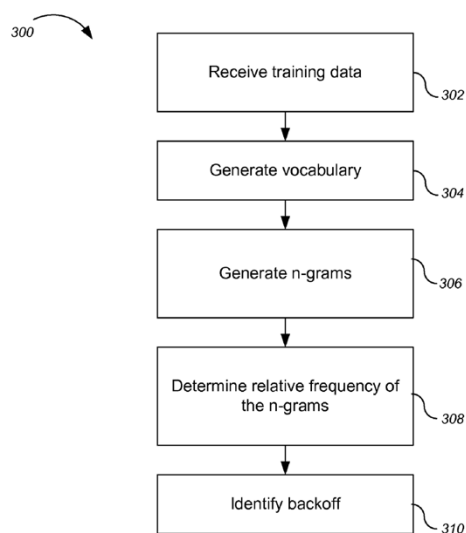
<b>Cited references to follow up on</b>	<p>Institute of Electrical and Electronics Engineers Standards Association. (2014). IEEE Standard for Local and Metropolitan Area Networks: Overview and Architecture. (IEEE 802-2014). <a href="https://standards.ieee.org/ieee/802/4158/">https://standards.ieee.org/ieee/802/4158/</a></p> <p>Bryan, T., Webb, S., Sallaway, P., Manickam, T., &amp; Raghavan, S. (2006). High definition multi-media interface (U.S. Patent No. 7,746,969 B2). U.S. Patent and Trademark Office. <a href="https://patents.google.com/patent/US7746969B2/en">https://patents.google.com/patent/US7746969B2/en</a></p>
<b>Follow up Questions</b>	<ul style="list-style-type: none"><li>• How is the second inference model trained?</li><li>• Is this viable to use over large networks of computer systems?</li><li>• What are the implications of an explainability issue between an inference model and the original AI model?</li></ul>

## Patent #2: Large language models in machine translation

<b>Source Title</b>	Large language models in machine translation
<b>Source citation (APA Format)</b>	Brants, T., Popat, A. C., Xu, P., Och, F. J., Dean, J. (2012). <b>Large language models in machine translation</b> (U.S. Patent No. 8,332,207 B2). U.S. Patent and Trademark Office. <a href="https://patentimages.storage.googleapis.com/7d/e8/bb/cabo80aeb7c5b6/US8332207.pdf">https://patentimages.storage.googleapis.com/7d/e8/bb/cabo80aeb7c5b6/US8332207.pdf</a>
<b>Original URL</b>	<a href="https://patentimages.storage.googleapis.com/7d/e8/bb/cabo80aeb7c5b6/US8332207.pdf">https://patentimages.storage.googleapis.com/7d/e8/bb/cabo80aeb7c5b6/US8332207.pdf</a>
<b>Source type</b>	Patent
<b>Keywords</b>	Translation, Language Model, Backoff
<b>#Tags</b>	#Patent #Translation #NotMachineLearning
<b>Summary of key points + notes (include methodology)</b>	This patent details an approach to machine translation in which a language model is used. It has examples of language models trained to predict the probability of different translations being correct for sentence portions called n-grams. The model has a training process involving the generation of n-grams, followed by the determination of relative frequencies of n-grams in both target languages. The model also identifies backoff n-grams with backoff scores that they are associated with.
<b>Important Figures</b>	<ul style="list-style-type: none"> <li>• Translation program architecture</li> </ul>



- Training process of language model



**VOCAB:  
(w/definition)**

Corpus: Collection of texts used for linguistic analysis, dataset (also used in this paper to just refer to languages)

N-gram: Consecutive sequences of n-length terms in a sentence,  
*Example:*

(1-gram) I, like, eating, food

(2-gram) I like, like eating, eating food

(3-gram) I like eating, like eating food

(4-gram) I like eating food

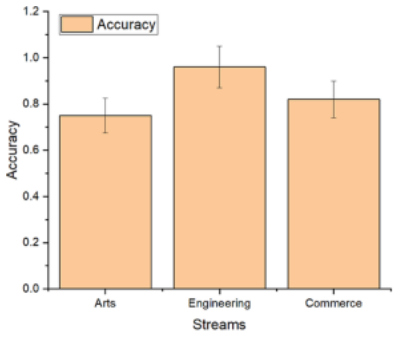
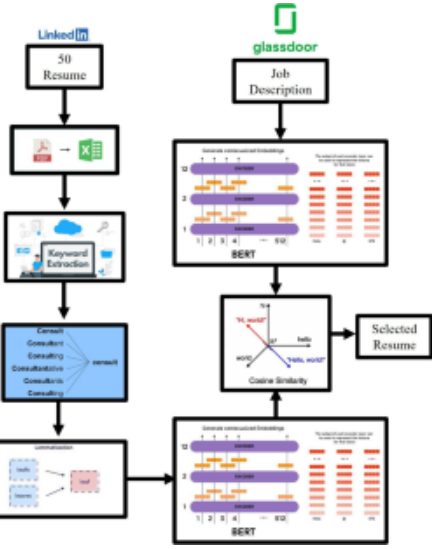
Backoff: Process of “backing off” to lower-order n-grams when higher-order n-grams are unavailable

<b>Cited references to follow up on</b>	<p>Katz, S. M.. "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer", IEEE Transactions on Acoustics, Speech and Signal Processing, IEEE Inc. New York, USA, vol. ASSP-35, No. 3, Mar. 1, 1987, pp. 400-401. Kneser et al., "Improved Backing-Off for M-GRAM Language Modeling". Acoustics, Speech, and Signal Processing, 1995. ICASSP 95., 1995 Inter National Conference on Detroit, MI, USA May 9-12, 1995, New York, NY, USA, IEEE, US, vol. 1, May 9, 1995, pp. 181-184, XPO10625 199, ISBN: 978-0-7803-2431-2 sections 1-4, abstract.</p> <p>Zhang et al., "Distributed Language Modeling for N-best List Ranking". Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Online, Jul. 22-23, 2006, page 216-223, XPO02503559 Sydney, Australia Retrieved from the Internet: URL: <a href="http://www.aclweb.org/anthology-new">http://www.aclweb.org/anthology-new</a> W. Wo6/ Wo6-1626.pdf&gt; retrieved on Nov. 12, 2008 sections 1-4, abstract. Dean et al., "MapReduce: Simplified Data Processing on Large Clusters". OSDI'04: Sixth Symposium on Operating System Design and Implementation, Online Dec. 6-8, 2004, XPO02503560, San Francisco, CA, USA Retrieved from the Internet: URL:<a href="http://labs.google.com/papers/mapreduce-osdi04.pdf">http://labs.google.com/papers/mapreduce-osdi04.pdf</a>&gt; retrieved on Nov. 11, 2008 the whole document.</p> <p>Placeway, et al., "The Estimation of Powerful Language Models From Small and Large Corpora". Plenary, Special, Audio, Underwater Acoustics, VLSI, Neural Networks. Minneapolis, Apr. 27-30, 1993; Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)), New York, IEEE, US, vol. 2, Apr. 27, 1993, pp. 33-36, XPO 10110386, ISBN: 978-0-7803 0946-3 pp. 33-36.</p>
<b>Follow up Questions</b>	<ul style="list-style-type: none"> <li>• Does this probabilistic method have any bearing on model neural network-based approaches? (it vaguely reminds me of attention)</li> <li>• How does the backoff process solve data sparsity problems?</li> <li>• How has machine translation evolved since these approaches were first assumed?</li> </ul>



## Article #11 Notes: AI image generators often give racist and sexist results: can they be fixed?

<b>Source Title</b>	Applying BERT-Based NLP for Automated Resume Screening and Candidate Ranking.
<b>Source citation (APA Format)</b>	Deshmukh, A., & Raut, A. (2024). Applying BERT-Based NLP for Automated Resume Screening and Candidate Ranking. <i>Annals of Data Science</i> . <a href="https://doi.org/10.1007/s40745-024-00524-5">https://doi.org/10.1007/s40745-024-00524-5</a>
<b>Original URL</b>	<a href="https://doi.org/10.1007/s40745-024-00524-5">https://doi.org/10.1007/s40745-024-00524-5</a>
<b>Source type</b>	Journal Article
<b>Keywords</b>	Application, Hiring, BERT, Resume
<b>#Tags</b>	#JobApplication #Hiring #Benchmark
<b>Summary of key points + notes (include methodology)</b>	This project attempts to use a BERT (an LLM by Google) based tool to screen resumes automatically. They collected 200 resumes and 10 job descriptions. Using keywords and skills mentioned in the resume, they ranked candidates for each job. Using BERT, they found feature vectors for the resumes and compared them to feature vectors of job descriptions. BERT screened up to 1 resume/second, and it found the highest similarities between job descriptions and resume features.
<b>Research Question/Problem/Need</b>	Can BERT and NLP/transformer-based techniques be used to increase efficiency in resume screening?
<b>Important Figures</b>	<ul style="list-style-type: none"> <li>• This technique could screen 1 resume/second</li> <li>• Accuracy</li> </ul>

	 <p>Accuracy</p> <p>Streams</p> <ul style="list-style-type: none"> <li>Methodology</li> </ul> 
<b>VOCAB: (w/definition)</b>	<p>BERT: A transformer model developed by Google in 2018</p> <p>Vector: An ordered/indexed collection of numbers</p> <p>NLP: Natural language processing, i. e. processing natural speech or written text</p>
<b>Cited references to follow up on</b>	<p>Athukorala C, Kumarasinghe H, Dabare K et al (2020) Business intelligence assistant for human resource management for IT companies. In: 2020 20th International Conference on Advances in {ICT} for Emerging Regions ({ICTer}). IEEE</p> <p>Bhatia V, Rawat P, Kumar A et al (2019) End-to-end resume parsing and finding candidates for a job description using Bert. arXiv preprint arXiv:191003089</p> <p>Bhoir N, Jakate M, Lavangare S et al (2023) Resume Parser</p>

	using hybrid approach to enhance the efficiency of Automated Recruitment Processes
<b>Follow up Questions</b>	<ul style="list-style-type: none"><li>• How does bias impact resume screening?</li><li>• How widespread would this application be in 10 years?</li><li>• How widespread is this practice today?</li></ul>

## Article #12 Notes: Accessing Artificial Intelligence for Clinical Decision-Making

<b>Source Title</b>	Accessing Artificial Intelligence for Clinical Decision-Making
<b>Source citation (APA Format)</b>	Giordano, C., Brennan, M., Mohamed, B., Rashidi, P., Modave, F., & Tighe, P. (2021). Accessing Artificial Intelligence for Clinical Decision-Making. <i>Frontiers in Digital Health</i> , 3. <a href="https://doi.org/10.3389/fdgth.2021.645232">https://doi.org/10.3389/fdgth.2021.645232</a>
<b>Original URL</b>	<a href="https://doi.org/10.3389/fdgth.2021.645232">https://doi.org/10.3389/fdgth.2021.645232</a>
<b>Source type</b>	Journal Article
<b>Keywords</b>	Application, Medical, Decision-Making
<b>#Tags</b>	#Benchmark #MedicalAdministration
<b>Summary of key points + notes (include methodology)</b>	This paper is a review of applications of AI in medical fields, specifically for decision making. The paper looked at hundreds of abstracts for a general review. It notes various ways that AI is used in the field. One way it is used is to help clinicians identify high-risk patients and allocate resources to them more efficiently. Another application is to use these models for optimizing patient outcomes by making decisions on treatment plans. The paper also mentions an application to detect acute decompensation early (a type of heart failure).
<b>Research Question/Problem/Need</b>	How is AI used for decision-making in the medical industry?
<b>Important Figures</b>	<ul style="list-style-type: none"> <li>• 887 abstracts were parsed for this review (excluding duplicates)</li> <li>• Methodology for conducting review</li> </ul>

	<div style="display: flex; flex-direction: column; align-items: center;"> <div style="margin-bottom: 10px;">Identification</div> <div style="margin-bottom: 10px;">Screening</div> <div style="margin-bottom: 10px;">Eligibility</div> <div style="margin-bottom: 10px;">Included</div> </div> <pre> graph TD     A[Records identified through database searching (n = 1072)] --&gt; B[Records after duplicates removed (n = 185)]     B --&gt; C[Records screened (n = 887)]     C --&gt; D[Records excluded (n = 598)]     C --&gt; E[Full-text articles assessed for eligibility (n = 289)]     E --&gt; F[Full-text articles excluded, with reasons (n = 186)]     E --&gt; G[Studies included in literature summary (n = 103)]   </pre>
<b>VOCAB: (w/definition)</b>	<p>Acute Decomposition: A form of sudden heart failure</p> <p>Risk Stratification: Categorize patients by risk for resource allocation</p>
<b>Cited references to follow up on</b>	<p>Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept “black box” medicine? <i>Ann Intern Med.</i> (2020) 172:59–60. doi: 10.7326/M19-2548</p> <p>Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. <i>Nat Mach Intell.</i> (2020) 2:369–75. doi: 10.1038/s42256-020-0197-y</p> <p>Corey KM, Kashyap S, Lorenzi E, Lagoo-Deenadayalan SA, Heller K, Whalen K, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): a retrospective, single-site study. <i>PLoS Med.</i> (2018) 15:e1002701. doi: 10.1371/journal.pmed.1002701</p>

**Follow up Questions**

- How may biases impact the use of AI in medical contexts?
- What will the shift to language models mean for the medical industry, given that they largely do not use language models yet?
- Are biases less pronounced in medical industry proprietary models?

## Article #13 Notes: Artificial Intelligence and Machine Learning in Legal Research: A Comprehensive Analysis

<b>Source Title</b>	Artificial Intelligence and Machine Learning in Legal Research: A Comprehensive Analysis
<b>Source citation (APA Format)</b>	Khan, S., Zakir, M., Bashir, S., & Ali, R. (2024). Artificial Intelligence and Machine Learning in Legal Research: A Comprehensive Analysis. <i>Qlantic Journal of Social Sciences</i> , 5(1), 307-317. <a href="https://doi.org/10.55737/qjss.203679344">https://doi.org/10.55737/qjss.203679344</a>
<b>Original URL</b>	<a href="https://doi.org/10.55737/qjss.203679344">https://doi.org/10.55737/qjss.203679344</a>
<b>Source type</b>	Journal Article
<b>Keywords</b>	Application, Legal, Writing
<b>#Tags</b>	#Benchmark #Legal #LanguageModel
<b>Summary of key points + notes (include methodology)</b>	This paper is a review of applications of AI in law. The authors read various papers to write this review. Relevant to this project, the paper cites biases as an important concern in using language models for legal settings. Also relevant is the idea of black-box models: the authors mention the dangers of this, and interpretability research is encouraged. It also mentions various uses of AI in legal proceedings and the justice system. It mentions the potential disruption in the job market for law that AI could cause.
<b>Research Question/Problem/Need</b>	How is AI used in the legal industry, and how may this usage impact the future of the justice system?
<b>Important Figures</b>	N/A (No figures)

<b>VOCAB: (w/definition)</b>	<p>ROSS: AI model specifically for legal tasks/legal research</p> <p>Lex Machina: Legal analytics platform built with ML</p> <p>Interpretability: De-obfuscating the processes that AI models take to arrive at conclusions</p>
<b>Cited references to follow up on</b>	<p>Khan, A., &amp; Jiliani, M. A. H. S. (2023). Expanding The Boundaries Of Jurisprudence In The Era Of Technological Advancements. IIUMLJ,31(2), 393-426.  <a href="https://doi.org/10.31436/iiumlj.v3i12.856">https://doi.org/10.31436/iiumlj.v3i12.856</a></p> <p>Khan, A., &amp; Wu, X. (2021). Bridging the Digital Divide in the Digital Economy with Reference to Intellectual Property. Journal of Law and Political Sciences, 28(03), 256-263.</p> <p>Alarie, B., Niblett, A., &amp; Yoon, A. H. (2018). How artificial intelligence will affect the practice of law. University of Toronto Law Journal, 68(supplement 1), 106-124.  <a href="https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3066816">https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3066816</a></p>
<b>Follow up Questions</b>	<ul style="list-style-type: none"> <li>• Is there empirical data for the impacts of biases on legal use cases for AI models?</li> <li>• What are the long-term effects of biases in legal settings, even without language models?</li> <li>• How can biases be effectively and fairly benchmarked?</li> </ul>

**Article #14 Notes: ANSI Common Lisp, Chapter 1**

<b>Source Title</b>	ANSI Common Lisp: Introduction
<b>Source citation (APA Format)</b>	Graham, P. (1996). Introduction. In <i>ANSI Common Lisp</i> . Pearson.
<b>Original URL</b>	<a href="https://ia601805.us.archive.org/27/items/f-1_20201109/ANSI_Common_Lisp_-_Paul_Graham.pdf">https://ia601805.us.archive.org/27/items/f-1_20201109/ANSI_Common_Lisp_-_Paul_Graham.pdf</a>
<b>Source type</b>	Book
<b>Keywords</b>	Artificial Intelligence, Lisp, Macro
<b>#Tags</b>	#Lisp #Macro
<b>Summary of key points + notes (include methodology)</b>	This book explains the Common Lisp programming language and its various uses, including its use in Artificial Intelligence. Lisp is a language that has a syntax built on "S-Expressions," which are either a list or an atom. Lists are evaluated with a function name for their first element and with arguments as the following elements. However, this form of evaluation can be bypassed by a quote. The language is built on the linked-list data structure, with Cons blocks representing cells. Common Lisp (and other Lisp-family languages) also has a powerful macro syntax because of the very consistent rules of evaluation.
<b>Research Question/Problem/Need</b>	How do you program in Common Lisp?
<b>Important Figures</b>	N/A (No data was researched, since it is an educational book)

<b>VOCAB: (w/definition)</b>	<p>Lisp: A programming language that is used in AI (especially natural language processing)</p> <p>Macro: A tool in programming that allows for replacing custom syntax with normal code before compile time</p> <p>Linked List: A data structure where each element has both the data it stores and a pointer to the next element</p>
<b>Cited references to follow up on</b>	N/A (Since it is an educational book)
<b>Follow up Questions</b>	<ul style="list-style-type: none"><li>• What major ML programs use Lisp today?</li><li>• Why is Lisp a good choice for programming Artificial Intelligence?</li><li>• Are there similarities between Python's PyTorch and Lisp's ML systems?</li></ul>

## Article #15 Notes: Multi-Task Alignment Using Steering Vectors

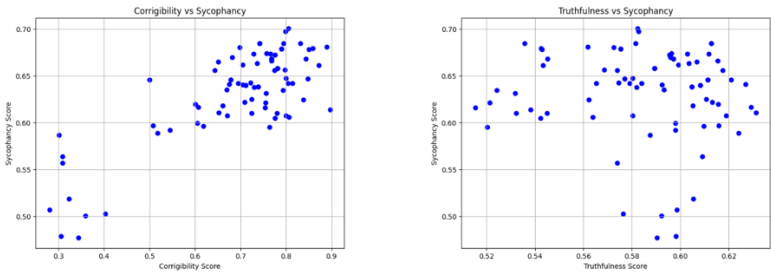
<b>Source Title</b>	Multi-Task Alignment Using Steering Vectors
<b>Source citation (APA Format)</b>	Li, C., & Maru, N. (n.d.). <i>Multi-Task Alignment Using Steering Vectors: Stanford CS224N Custom Project</i> . Stanford University. <a href="https://web.stanford.edu/class/cs224n/final-reports/256908428.pdf">https://web.stanford.edu/class/cs224n/final-reports/256908428.pdf</a>
<b>Original URL</b>	<a href="https://web.stanford.edu/class/cs224n/final-reports/256908428.pdf">https://web.stanford.edu/class/cs224n/final-reports/256908428.pdf</a>
<b>Source type</b>	Report
<b>Keywords</b>	Alignment, Activation, Vector, LLM, CAA
<b>#Tags</b>	#Alignment #Steering #Activation
<b>Summary of key points + notes (include methodology)</b>	This project uses contrastive activation addition (CAA) to improve Llama 2 7B's performance on benchmarks for truthfulness, sycophancy, and corrigibility. Truthfulness, sycophancy (how agreeable the model is), and corrigibility (how obedient the model is) are all important metrics for alignment. This paper finds steering vectors for all three metrics and runs a benchmark before and after steering is applied to see if there was a significant difference in the model's benchmark scores. It also looks at the correlations between metrics like sycophancy and corrigibility to see how scores on one correlate to scores on the other.
<b>Research Question/Problem/Need</b>	How can contrastive activation addition be used to make language models more aligned?

Important Figures

Table 3: Evaluation Results for Models						
Model	C-MC	S-MC	T-MC	C-OE	S-OE	MMLU
Baseline	0.74	0.64	0.60	0.36	0.32	0.64
Best Corrigibility	0.90	0.61	0.54	0.38	0.41	0.66
Best Sycophancy	0.81	0.70	0.58	0.31	0.50	0.68
Best Truthfulness	0.65	0.61	0.63	0.24	0.19	0.63
Improved	0.77	0.67	0.62	0.37	0.39	0.60
Best Composite	0.89	0.68	0.56	0.37	0.52	0.68

The Improved model demonstrates marginal gains over the Baseline model, with improvements in corrigibility (3%), sycophancy (3%), and truthfulness (2%). The Best Composite model shows significant improvement in corrigibility (15%) and sycophancy (4%), but it performs worse on truthfulness (-4%).

Scores before and after steering, after steering being labelled "improved model." (above)



(a) Corrigibility and sycophancy are strongly positively correlated. (b) Truthfulness and sycophancy are weakly negatively correlated.

Figure 1: Correlation analysis of corrigibility, sycophancy, and truthfulness based on multiple-choice experiment results.

Correlation between different alignment metrics (above)

VOCAB:  
(w/definition)

Alignment: AI models acting in a desirable way for humanity – being obedient, truthful, safe, impartial, etc.  
Sycophancy: Being extremely agreeable, not countering what is being told  
Corrigibility: Being obedient and responsive to commands

Cited references to  
follow up on

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. Pretraining language models with human preferences, 2023.

Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024

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,

	<p>Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2023.</p>
<b>Follow up Questions</b>	<ul style="list-style-type: none"><li>• What part does bias play in alignment?</li><li>• Can various biases be measured for at once, as this paper did with alignment metrics?</li><li>• How do reductions in one bias correlate with reductions in other biases?</li></ul>

**Article #16 Notes: Attention is All You Need**

<b>Source Title</b>	Attention is All You Need
<b>Source citation (APA Format)</b>	Vaswani, A. (2017). Attention is all you need. <i>Advances in Neural Information Processing Systems</i> . <a href="https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf">https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf</a>
<b>Original URL</b>	<a href="https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf">https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf</a>
<b>Source type</b>	Journal Article
<b>Keywords</b>	Attention, Transformer, Architecture
<b>#Tags</b>	#Transformer #Attention
<b>Summary of key points + notes (include methodology )</b>	This paper is a landmark paper that outlines the transformer architecture for natural language processing. The transformer architecture is an architecture for a neural network wherein there are attention layers that help the network understand the context of a statement and refine internal representations of a word/token's meaning based on this context. The paper tests a transformer model's effectiveness in a worldwide translation competition, where it performed far better than previous models while taking less time to train and being more parallelizable.
<b>Research Question/Problem/ Need</b>	How can NLP models be made to understand nuance from context?

## Important Figures

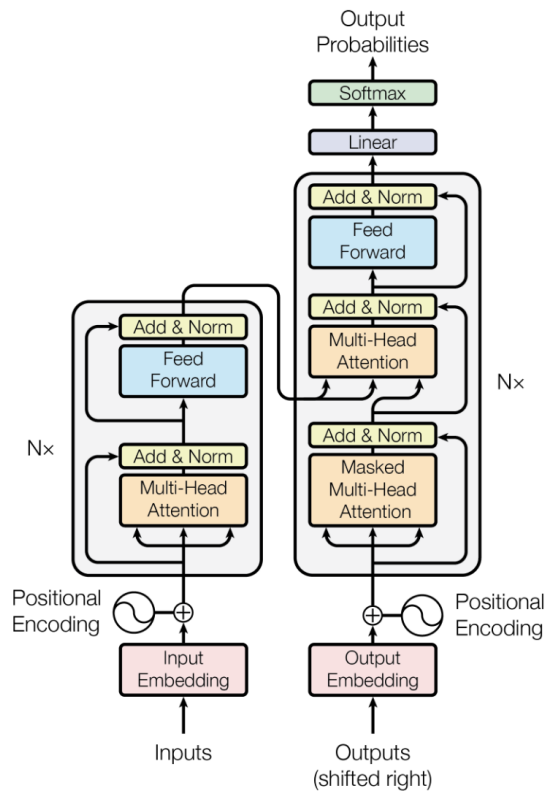


Figure 1: The Transformer - model architecture.

### Transformer model architecture diagram

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	<b>41.29</b>	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	<b><math>3.3 \cdot 10^{18}</math></b>	
Transformer (big)	<b>28.4</b>	<b>41.0</b>	$2.3 \cdot 10^{19}$	

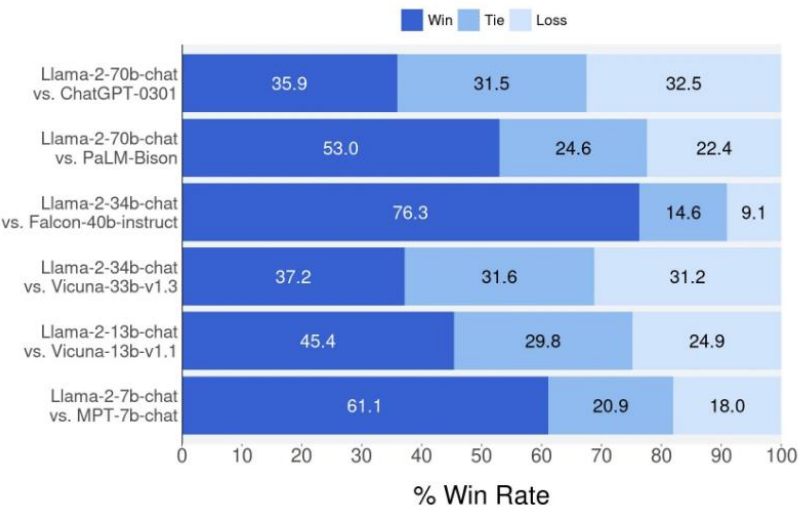
Performance vs. training Costs for various models – transformer model outperforms other models with less training cost

<b>VOCAB: (w/definition )</b>	<p>Positional Encoding: Adding data about position to the word embeddings</p> <p>Softmax: A function that maps the real numbers to numbers between 0 and 1, in a way that there are large differences in numbers near 0 but smaller differences between large values (s-shaped curve)</p> <p>Masked Multi-Head Attention: Attention performed on many attention heads that run in parallel to increase speed.</p>
<b>Cited references to follow up on</b>	<p>Cheng, J. (2016). Long short-term memory-networks for machine reading. <i>arXiv preprint arXiv:1601.06733</i>.</p> <p>Gehring, J., Auli, M., Grangier, D., Yarats, D., &amp; Dauphin, Y. N. (2017, July). Convolutional sequence to sequence learning. In <i>International conference on machine learning</i> (pp. 1243-1252). PMLR.</p>
<b>Follow up Questions</b>	<p>Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., &amp; Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. <i>Journal of Machine Learning Research</i>, 15(56), 1929–1958.  <a href="http://jmlr.org/papers/v15/srivastava14a.html">http://jmlr.org/papers/v15/srivastava14a.html</a></p>

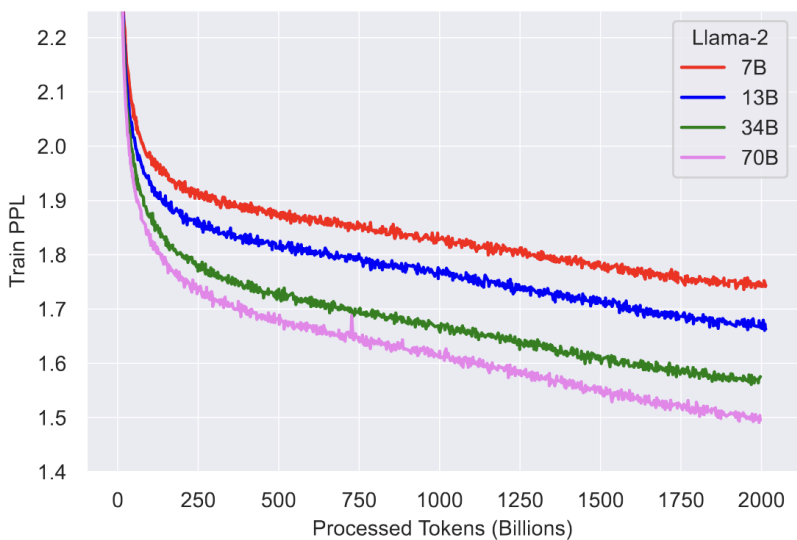
## Article #17 Notes: Llama 2: Open Foundation and Fine-Tuned Chat Models

<b>Source Title</b>	Llama 2: Open Foundation and Fine-Tuned Chat Models
<b>Source citation (APA Format)</b>	Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., & Fuller, B. (2023, July 19). Llama 2: Open Foundation and Fine-Tuned Chat Models. <i>ArXiv</i> . <a href="https://doi.org/10.48550/arXiv.2307.09288">https://doi.org/10.48550/arXiv.2307.09288</a>
<b>Original URL</b>	<a href="https://doi.org/10.48550/arXiv.2307.09288">https://doi.org/10.48550/arXiv.2307.09288</a>
<b>Source type</b>	Online Source (ArXiv)
<b>Keywords</b>	Transformer, LLM, Open-Source
<b>#Tags</b>	#Llama #Transformer #NLP #LLM #Model
<b>Summary of key points + notes (include methodology)</b>	Llama 2 is a family of large language models (LLMs) developed by Meta AI. These models outperform various other open-source models on benchmarks that they were tested on, and it outperformed many proprietary models on some benchmarks as well. Notably, the larger of the models had similar performance to ChatGPT-3, a pretrained proprietary model. The family of models comes in four sizes: 7 billion parameters, 13 billion parameters, 34 billion parameters, and 70 billion parameters. The models were tested on various benchmarks to compare their performances with other models.
<b>Research Question/Problem/Need</b>	(Problem) There are not many open-source models with similar performance to proprietary LLMs.

Important Figures



% win rate in a human helpfulness benchmark in which human evaluators rate the helpfulness of various models on standardized tasks



Training loss of Llama-2 models as amount of training tokens increase

VOCAB:  
(w/definition)

Open-Source: Source code is freely available to everyone  
Loss: A measure of how well (or really, how poorly) a model is performing – lower loss means better performance  
Token: A lexeme – a word or part of a word that adds semantic meaning

<b>Cited references to follow up on</b>	<p>Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, and Chris Olah. A general language assistant as a laboratory for alignment. <i>arXiv preprint arXiv:2112.00861</i>, 2021a.</p> <p>Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. On the dangers of stochastic parrots: Can language models be too big? In <i>Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency</i>, pages 610–623, 2021a.</p> <p>Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In <i>Proceedings of the AAAI conference on artificial intelligence</i>, pages 7432–7439, 2020.</p>
<b>Follow up Questions</b>	<ul style="list-style-type: none"> <li>• Why is Llama-2 so powerful, even compared to proprietary models?</li> <li>• How can I access Llama-2 for my research?</li> <li>• How powerful a computer is necessary to run the largest (34B and 70B) Llama models?</li> </ul>

**Article #18 Notes: ANSI Common Lisp, Chapter 2**

<b>Source Title</b>	ANSI Common Lisp: Welcome to Lisp
<b>Source citation (APA Format)</b>	Graham, P. (1996). Welcome to Lisp. In <i>ANSI Common Lisp</i> . Pearson.
<b>Original URL</b>	<a href="https://ia601805.us.archive.org/27/items/f-1_20201109/ANSI_Common_Lisp_-_Paul_Graham.pdf">https://ia601805.us.archive.org/27/items/f-1_20201109/ANSI_Common_Lisp_-_Paul_Graham.pdf</a>
<b>Source type</b>	Book
<b>Keywords</b>	Artificial Intelligence, Lisp, Macro
<b>#Tags</b>	#Lisp #Macro
<b>Summary of key points + notes (include methodology)</b>	This chapter gives an introduction to the syntax of Lisp. The Lisp programming language is built on S-Expressions. These are expressions with a specific and consistent rule for evaluation. Parentheses guide the order of code execution. All S-Expressions feature prefix function names followed by arguments, unless they are preceded by a quote in which case they function as a list.
<b>Research Question/Problem/Need</b>	How do you program in Common Lisp?
<b>Important Figures</b>	N/A (No data was researched, since it is an educational book)
<b>VOCAB: (w/definition)</b>	<p>Toplevel: A Lisp REPL-like environment</p> <p>Function: A transformation that maps a set of arguments to a set of results</p> <p>Protection: Using a quote to prevent the evaluation of an s-expression</p>
<b>Cited references to follow up on</b>	N/A (Since it is an educational book)
<b>Follow up Questions</b>	<ul style="list-style-type: none"> <li>• What can be done with lists in Lisp?</li> <li>• How can this model of evaluation supplement machine-learning tasks?</li> <li>• How does functional programming lend itself to Lisp?</li> </ul>



## Article #19 Notes: ANSI Common Lisp, Chapter 3

<b>Source Title</b>	ANSI Common Lisp: Lists
<b>Source citation (APA Format)</b>	Graham, P. (1996). Lists. In <i>ANSI Common Lisp</i> . Pearson.
<b>Original URL</b>	<a href="https://ia601805.us.archive.org/27/items/f-1_20201109/ANSI_Common_Lisp_-_Paul_Graham.pdf">https://ia601805.us.archive.org/27/items/f-1_20201109/ANSI_Common_Lisp_-_Paul_Graham.pdf</a>
<b>Source type</b>	Book
<b>Keywords</b>	Artificial Intelligence, Lisp, Macro
<b>#Tags</b>	#Lisp #Macro
<b>Summary of key points + notes (include methodology)</b>	This chapter explains how lists work and how to use them in Lisp. It explains the idea of a cons cell, which is a constructor that points to two Lisp objects. Pointing cons cells to data and a next cons cell makes a linked list data structure. This ends when one cons cell points to a nil object eventually. Unlike lists in many languages, lists in Lisp are not implemented as consecutive blocks of memory. This chapter also introduces higher-order mapping functions for lists, as well as describing how variables are also pointers to their values.
<b>Research Question/Problem/Need</b>	How do you program in Common Lisp?
<b>Important Figures</b>	N/A (No data was researched, since it is an educational book)
<b>VOCAB: (w/definition)</b>	<p>Lambda: An anonymous function – a transformation that is not bound to a variable/identifier, simply written as an expression on its own</p> <p>Map: A method to apply a function on each element of a list and return a new list with these modifications applied</p> <p>Dotted List: A list where the last cons object does not point to a nil, rather it points to another (truthy) Lisp object</p>
<b>Cited references to follow up on</b>	N/A (Since it is an educational book)

**Follow up  
Questions**

- How does Lisp implement higher-order functions under-the-hood (possibly function pointers)?
- How can more complex data structures be made in Lisp?
- Why not use vectors like other languages, why make lists linked lists by default?

## Article #20 Notes: Transformer Models: An Introduction and Catalog

<b>Source Title</b>	Transformer Models: An Introduction and Catalog
<b>Source citation (APA Format)</b>	Amatriain, X., Sankar, A., Bing, J., Bodigutla, P. K., Hazen, T. J., & Kazi, M. (2023, May 25). Transformer Models: An Introduction and Catalog. <i>ArXiv</i> . <a href="https://doi.org/10.48550/arXiv.2302.07730">https://doi.org/10.48550/arXiv.2302.07730</a>
<b>Original URL</b>	<a href="https://doi.org/10.48550/arXiv.2302.07730">https://doi.org/10.48550/arXiv.2302.07730</a>
<b>Source type</b>	Online Source (ArXiv)
<b>Keywords</b>	Transformer, LLM
<b>#Tags</b>	#Transformer #LLM #Model
<b>Summary of key points + notes (include methodology)</b>	This article gives an introduction to transformer models and how they work. The paper explains attention in detail, stating that this part of the architecture is what makes transformers so powerful. During the attention process, the query and key matrices are multiplied, scaled, and masked. After this, they are put through the SoftMax normalization function and then multiplied with the value matrix. This attention process is run in parallel for many tokens across multiple heads, making it <i>multi-head</i> attention. The heads are all merged after they all pass through the attention process.
<b>Research Question/Problem/Need</b>	What are transformer models and how do they work?

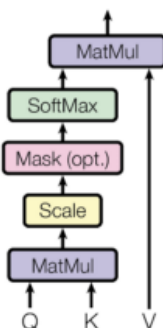
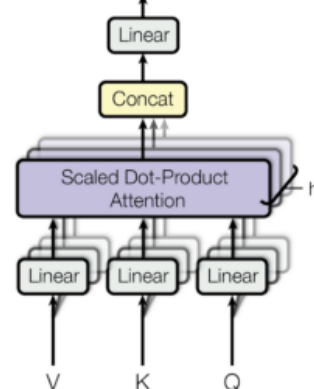
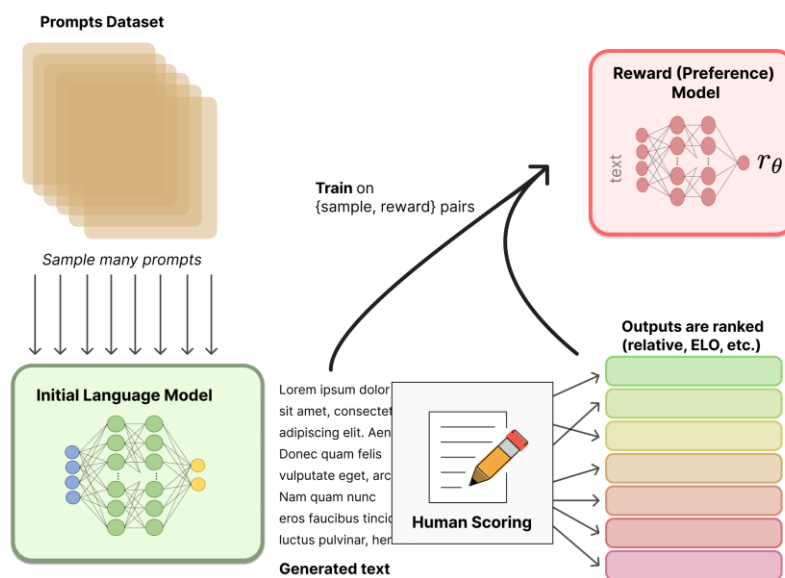
**Important Figures****Scaled Dot-Product Attention****Multi-Head Attention**

Diagram of multi-head attention in the transformer model architecture. Here, Q is the query matrix, K is the key matrix, and V is the value matrix.



A diagram of the reinforcement learning process, in which human evaluators rank model outputs, and these rankings are then passed to a preference model which ranks many more model outputs, giving additional data for the model to train with.

**VOCAB:**  
**(w/definition)**

**Reinforcement Learning:** A process in which human evaluators rank the responses of a model, and these rankings are used to train a preference model that rewards the original model for what it deems as highly ranking responses

	<p>Scaled Dot-Product Attention: A type of attention where the result of the multiplication of the query and key matrices are scaled before being masked and passed into the SoftMax function</p> <p>Preference Model: A model in reinforcement learning that is trained on human feedback of the language model's responses, which is used to reward the model for what it believes are good responses based on the human feedback-based training data</p>
<b>Cited references to follow up on</b>	<p>Aghajanyan, A., Gupta, A., Shrivastava, A., Chen, X., Zettlemoyer, L., &amp; Gupta, S. (2021). Muppet: Massive Multi-task Representations with Pre-Finetuning. <i>ArXiv</i>. <a href="https://arxiv.org/abs/2101.11038">https://arxiv.org/abs/2101.11038</a></p> <p>Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. (2022). Training compute-optimal large language models. <i>ArXiv</i>.</p> <p>Rosset, C. (2020). Turing-NLG: A 17-billion-parameter language model by Microsoft. <i>Microsoft Research</i>. <a href="https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/">https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/</a></p>
<b>Follow up Questions</b>	<ul style="list-style-type: none"> <li>• How does backpropagation influence the query, key, and value matrices?</li> <li>• How does the model accept feedback from the preference model during reinforcement learning?</li> <li>• How does the preference model change the weights of the original model to increase the quality of its outputs?</li> </ul>