Head Pursuit: Probing Attention Specialization in Multimodal Transformers

Anonymous Author(s)
Affiliation
Address
email

Abstract

Language and vision-language models have shown impressive performance across a wide range of tasks, but their internal mechanisms remain only partly understood. In this work, we study how individual attention heads in text-generative models specialize in certain semantic or visual attributes. We reinterpret the established practice of probing intermediate activations with the final decoding layer through the lens of signal processing. This lets us analyze multiple samples in a principled way and rank attention heads based on their relevance to target concepts. Our results show consistent patterns of specialization at the head level across both unimodal and multimodal transformers. Remarkably, we find that editing as few as 1% of the heads, selected using our method, can reliably impact targeted concepts in the model output.

1 Introduction

Large-scale generative models, including both language and vision-language transformers, have achieved remarkable performance on a wide spectrum of tasks, from open-ended text generation [1] to image captioning and visual question answering [2–5]. Despite these successes, the internal mechanisms by which these models organize and represent knowledge remain only partially understood. In particular, the role of individual components, such as attention heads, in mediating specific aspects of generation has been the subject of increasing interest for both interpretability and control [6, 7]. Previous studies have shown that attention heads in large language models (LLMs) often exhibit emergent roles, such as syntax tracking or copy behavior [8–10]. Interpretability tools such as the logit lens [11] and its extensions [12, 13] have provided strategies for inspecting intermediate model representations, revealing rich semantic information latent in hidden states. However, these techniques are typically applied heuristically and focus on individual examples, making it difficult to generalize findings across multiple samples or quantify the importance of specific model components in shaping the model's output.

In this work, we take a more principled approach to analyzing the specialization of attention heads in generative transformers. Specifically, we revisit a variant of Matching Pursuit (MP) [14], a classical greedy algorithm to approximate high-dimensional signals with sparse linear combinations of basis elements, and bridge it with recent interpretability techniques. By applying MP to the hidden states of text-generative models, we propose a way to identify a small set of attention heads that most strongly influence the capability of the model to generate text within a certain conceptual area (e.g., colors or numbers). Applying MP across both unimodal and multimodal models, we uncover consistent specialization patterns, with certain heads reliably governing the generation of semantically coherent token groups. We empirically validate our head selection strategy by demonstrating that targeted interventions, implemented by inverting these concept-specific heads, can selectively suppress the associated content, thereby enabling fine-grained and interpretable control over model outputs without requiring additional training.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

2 Related Work

Recent research on Transformer architectures has investigated the functional roles and specialization of attention heads. In language models, most attention heads appear redundant, with pruning studies showing that many can be removed with minimal loss in performance on NLP tasks [8, 9]. Some heads have also been linked to eliciting factual knowledge [6], promoting in-context induction [10], or suppressing lexical repetition [15].

In vision-language, similar specialization patterns have been observed in the visual encoder of CLIP-like models, by applying methods that leverage visual-textual alignment to decompose heads over sentence encodings [16, 17]. A parallel line of research adapts the mechanistic interpretability tools developed for language models to the multimodal setting. Representative works include [18] and [19], which investigate information transfer mechanisms in multimodal transformers, and [20], which extends the logit lens [11] to the analysis of visual token representations.

3 Pursuing specialized attention heads

We start our investigation by exploring whether individual attention heads of generative LLMs specialize in interpretable functions. To isolate the contribution of each head, we use a residual stream decomposition approach: following [21], we model the output written by each head into the residual stream as a matrix $\mathbf{H}_{h,l} \in \mathbb{R}^{n,d}$, where n is the number of samples in the dataset and d is the internal dimensionality of the transformer. Our aim is to identify sparse and interpretable directions for each attention head $\mathbf{H}_{h,l}$ that maximally explain its variance on a given dataset. Concretely, we seek a sparse representation of $\mathbf{H}_{h,l}$ using directions drawn from a fixed dictionary of interpretable vectors, rather than an unconstrained continuous space, to ensure that the resulting components are meaningful and grounded in known semantic structures. As a dictionary, we adopt the unembedding matrix of the language model $\mathbf{D} \in \mathbb{R}^{v,d}$, as it naturally contains directions that are aligned with semantically meaningful outputs, allowing us to ground latent structure in human-interpretable terms.

We employ a classical sparse coding algorithm: Simultaneous Orthogonal Matching Pursuit (SOMP) [22]. SOMP is a multi-sample extension of Orthogonal Matching Pursuit [23], itself a refinement of the original Matching Pursuit algorithm [14]. Rather than analyzing each sample independently, SOMP jointly considers all samples in a given dataset and selects the dictionary directions that are most informative across the representation. Formally, given a head activation matrix $\mathbf{H} \in \mathbb{R}^{n,d}$ and a dictionary $\mathbf{D} \in \mathbb{R}^{v,d}$, SOMP aims to iteratively construct a column-sparse coefficient matrix $\mathbf{W}^* \in \mathbb{R}^{n,v}$ such that $\mathbf{H} \approx \mathbf{W}^*\mathbf{D}$. At each iteration, the algorithm identifies the dictionary entry most correlated with the current residuals and refits the reconstruction to minimize the difference between the original head outputs and their approximation. This process continues until a predefined sparsity level is reached.

Importantly, we note a conceptual connection between our reinterpretation of SOMP and the logit lens (LL) [11], a tool widely used in mechanistic interpretability to probe internal representations of transformer models. Similarly to the method just described, LL works by projecting a single residual stream vector onto the unembedding directions to approximate the output logits of the model at intermediate layers. This is equivalent to performing a single step of matching pursuit on an individual example. Our SOMP-based method generalizes this idea in two key ways: it operates on multiple examples simultaneously, and it selects multiple dictionary directions, each capturing distinct components of the signal. This leads to a more robust and semantically structured characterization of the attention head's functional role. In Table 1, we report some examples of specialized attention heads, obtained by applying SOMP to Mistral-7B attention heads, prompted by questions from the TriviaQA dataset [24]. Before applying SOMP, the tokens from the prompt were aggregated by averaging. As we show in the table, a direct application of LL [1] in this setting results in noisier and highly redundant explanations.

Besides returning lists of latent directions and associated natural language tokens that better characterize each head, SOMP produces a reconstruction of the head representation in the space spanned by those vectors. Building on this insight, we propose a method to automatically identify the heads most relevant for a target attribute. Given a list of words related to the chosen semantic area, one

[1] We aggregate over mutiple samples by storing the 5 tokens with highest logits for each sample, and then taking the 5 most frequent tokens overall.

Table 1: Top-5 tokens identified by SOMP and LL on selected attention heads of Mistral-7B, evaluated on TriviaQA prompts.

Method	L18.H27 (Politics)	L24.H20 (Nationality)	L25.H14 (Calendar)	L30.H28 (Digits)
SOMP	COVID; Soviet; Obama; Biden; Clinton	British; American; European; German; English	February; July; October; Christmas; April	9; 1; 3; 7; five
Logit Lens	Covid; vaccine; pandemic; COVID; Biden	American; Americans; America; American; California	Sunday; October; February; Oct; breakfast	u; 8; u; n; 9

can restrict the unembedding matrix to the rows associated to these tokens and apply SOMP on this concept-specific dictionary. Then, the fraction of head variance explained by SOMP in this setting can be considered as a measure of specialization of the head, allowing us to rank and select heads by their relevance with respect to the target concept.

4 Controlling generation through specialized heads

We now evaluate how the specialization of attention heads can be leveraged to apply domain-specific targeted interventions to model behavior, effectively validating our selection. One way to do so is to disrupt the information flow from a selected subset of heads to the residual stream during the forward pass. Concretely, we apply this intervention by inverting the sign of the head representations. The key preliminary step is to identify relevant and specialized heads. To accomplish this, we apply SOMP over a restricted unembedding dictionary, filtered to include only a set of tokens associated with the target property. Then, we rank attention heads by the proportion of their variance explained by the SOMP reconstruction and intervene on the top-k ranked heads. In all of our experiments, we include a random control condition to verify the specificity of our findings. This control involves intervening on a randomly selected set of attention heads that matches the original set in both size and layer distribution but is entirely disjoint from it.

4.1 Mitigation of toxic content

Experimental setting As a first experiment, we focus on toxicity mitigation: specifically, reducing the occurrence of offensive words in text generated by Mistral-7B [25]. To do this, we identify a subset of toxic heads within the model and intervene on them. We consider two datasets, RealToxicityPrompts (RTP) [26], which contains naturally occurring Web prompts, and Thoroughly Engineered Toxicity (TET) [27], a benchmark with carefully constructed test cases, both of which are designed to elicit harmful responses from LLMs. For both datasets, we extract and label toxic words from Mistral's responses using Llama3.3 [28]. From the list, we select the 100 most frequent toxic words, and randomly choose 70% of these to identify the toxic heads. To evaluate effectiveness, we measure how often the remaining 30% of toxic words appear in Mistral's outputs after inversion, relative to their frequency before the intervention.

Table 2: Normalized frequency of held-out toxic words after intervention. Lower values indicate better mitigation. Targeted heads reduce toxicity, while random heads often increase it.

Dataset	Top 16 heads	Top 32 heads	Random 16 heads	Random 32 heads
RTP	0.77	0.72	1.26 ± 0.48	1.19 ± 0.31
TET	0.66	0.48	1.09 ± 0.17	1.41 ± 0.49

Result analysis The results we obtain by inverting the sign of toxic head activations are displayed in Table 2, for 16 and 32 heads. In both RTP and TET, intervening on such heads noticeably reduces the frequency of toxic words, even if they were *not used* for the head selection. Moreover, intervening on randomly chosen control heads tends to increase the frequency of toxic words. This trend is expected, as we are randomly picking and disrupting heads that are deemed non-toxic (we impose that they are distinct from the targeted ones), thus implicitly reinforcing toxic behavior.

4.2 Targeted control of visual attributes

We now move to evaluating the extent and implications of head specialization in the LLM backbones within generative Vision-Language models (VLMs). These models are usually built by fine-tuning a pre-trained LLM on multimodal tasks, such as visual question answering or image captioning, using visual tokens coming from a pre-trained vision encoder as contextual information [3]. In line with recent works [20] that have successfully applied the logit lens to visual tokens of LLaVA, a prominent example of VLM, we investigate head specialization by applying our MP-based analysis on the head representations of image patches, averaged over tokens.

Experimental setting For this experiment we benchmark LLaVA-NeXT-7B [4] (from now on just LLaVA for short) on a range of image classification datasets, including: MNIST [29], SVHN [30], GTSRB [31], Eurosat [32], RESISC45 [33] and DTD [34]. For each dataset, we begin by selecting the set of k most relevant heads. Heads are selected by applying SOMP to the unembedding matrix restricted to tokens corresponding to class names, and sorting heads by the fraction of variance explained by the SOMP reconstruction. In this experiment, we prompt the model to classify the image, and evaluate the generated output in terms of exact match with the ground truth class label.

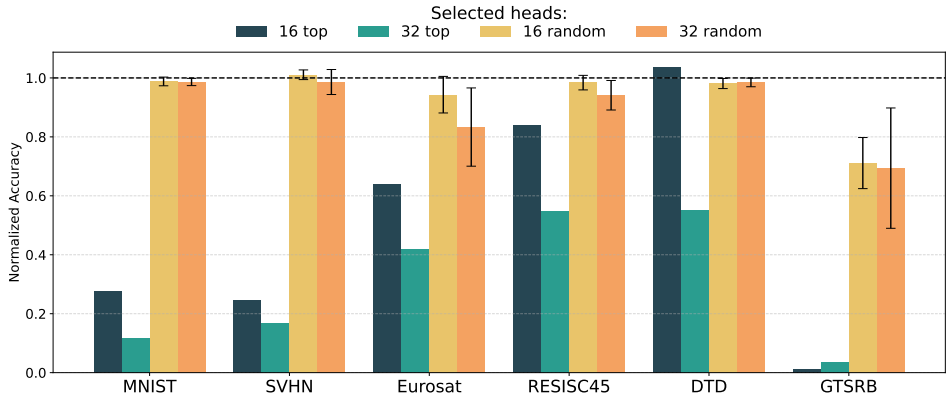

Figure 1: Classification results under different head selection strategies: (blue) 16 heads with highest variance ratio explained by SOMP; (green) 32 heads with highest explained variance ratio; (yellow) 16 random heads, with the same layer-wise counts of top 16; (orange) 32 random heads, with the same layer-wise counts of top 32.

Result analysis We report the classification results in Figure 1, normalized for each dataset with respect to the accuracy obtained by LLaVA when no intervention is applied to its forward pass. For all datasets, inverting the the top 32 heads identified by our method is sufficient to significantly disrupt the classification performance, while inverting 32 random heads at equivalent layers has substantially lower to no impact on performance. At $k = 16$ the picture is similar with the exception of DTD, whose performance is unaffected, hinting at higher head redundancy on this task. Summing up, intervening on a small set of attention heads selected via SOMP significantly disrupts classification performance across diverse datasets, confirming head-level specialization in LLaVA.

5 Discussion

In this work, we investigated the specialization of attention heads in large generative models through a sparse, interpretable decomposition of their outputs. Using Simultaneous Orthogonal Matching Pursuit (SOMP) over the model's unembedding space, we identified directions aligned with semantically meaningful attributes and used them to recover sets of specialized heads across different tasks and modalities. Our approach offered a multi-sample generalization of the logit lens, allowing us to move beyond single-token analysis toward more stable, dataset-level structure. We showed that the selected heads could be ranked by their explained variance and that intervening on a small number of them produced targeted changes in generation. These findings held across text and vision-language settings, supporting the utility of head-level analysis and intervention for model understanding and control.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[4] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[5] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[6] Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8420–8436, 2024.

[7] Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Talking heads: Understanding inter-layer communication in transformer language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[8] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, 2019.

[9] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in Neural Information Processing Systems*, 32, 2019.

[10] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[11] nostalgebraist. interpreting gpt: the logit lens. *LessWrong*, 2020.

[12] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

[13] Mansi Sakarvadia, Arham Khan, Aswathy Ajith, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard, and Ian Foster. Attention lens: A tool for mechanistically interpreting the attention head information retrieval mechanism. *arXiv preprint arXiv:2310.16270*, 2023.

[14] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

[15] Callum Stuart McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding a motif in language model attention heads. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen, editors, *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 337–363, Miami, Florida, US, November 2024. Association for Computational Linguistics.

[16] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting CLIP's image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*, 2024.

[17] Lorenzo Basile, Valentino Maiorca, Luca Bortolussi, Emanuele Rodolà, and Francesco Locatello. Residual transformer alignment with spectral decomposition. *Transactions on Machine Learning Research*, 2025.

[18] Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[19] Alessandro Serra, Francesco Ortu, Emanuele Panizon, Lucrezia Valeriani, Lorenzo Basile, Alessio Ansuini, Diego Doimo, and Alberto Cazzaniga. The narrow gate: Localized image-text communication in vision-language models. *arXiv preprint arXiv:2412.06646*, 2024.

[20] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[21] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

[22] Joel A Tropp, Anna C Gilbert, and Martin J Strauss. Algorithms for simultaneous sparse approximation. part i: Greedy pursuit. *Signal processing*, 86(3):572–588, 2006.

[23] Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, pages 40–44. IEEE, 1993.

[24] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 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)*. Association for Computational Linguistics, 2017.

[25] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[26] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, 2020.

[27] Tinh Luong, Thanh-Thien Le, Linh Ngo, and Thien Nguyen. Realistic evaluation of toxicity in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1038–1047, 2024.

[28] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[29] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[30] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada, 2011.

[31] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011.

[32] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[33] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[34] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.