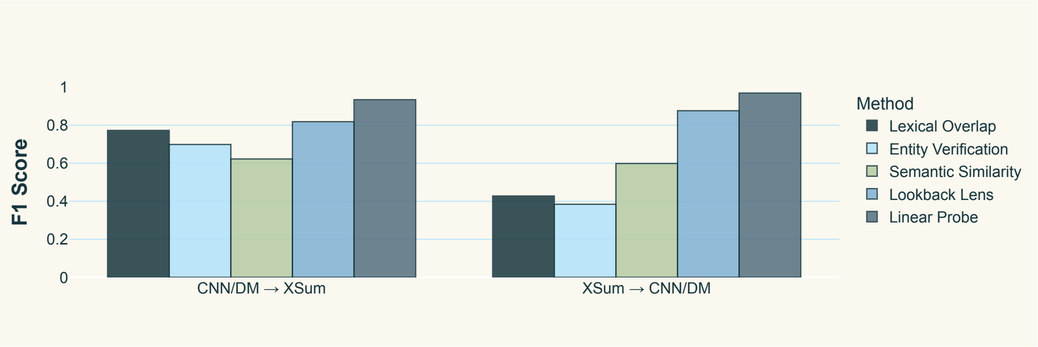

Figure 3: **Cross-domain transfer performance of hallucination detection methods.** F_1 scores for detectors trained on one news dataset (CNN/DM or XSUM) and evaluated on the other. Features were extracted from layer 20 of a Gemma-2-9B observer. The linear probe demonstrates high transferability compared to baseline methods.

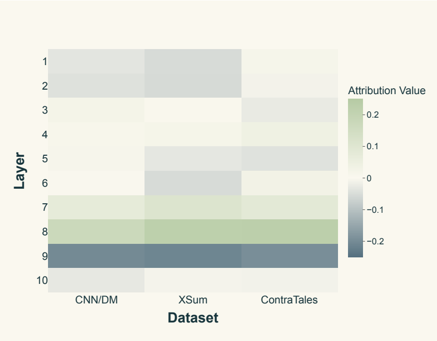

Figure 4: **Aggregated MLP layer attributions for the hallucination probe.** Mean MLP attributions ($\bar{A}^{(\ell)}_{\text{mlp}}$) per layer for a linear probe trained on layer 10 of Gemma-2-9B. Attributions are presented for evaluations on CNN/DM, XSUM, and CONTRATALES, revealing a consistent pattern across datasets in layers 7-9.

4.5 Finetuning improves internal hallucination indication

To investigate improving observer performance on a target domain without new hallucination labels, we performed unsupervised domain adaptation. Observer models (GPT-2-small, Gemma-2-2B, Gemma-2-9B) were further trained for two epochs on 1000 correct-only completions from the CONTRATALES corpus using standard SFT hyperparameters (AdamW, LR 1×10^{-5}, context length 512, batch size 8, 8xH200 GPUs, no dropout). Following this adaptation, the logistic residual-stream probe (as per §3) was retrained on the original labeled data and evaluated on an unseen test fold. As shown in Figure 6, this process improved F_1 scores for all models: by +0.10 for GPT-2-small, +0.17 for Gemma-2-2B, and +0.14 for Gemma-2-9B. For Gemma-2-9B on CONTRATALES, the F_1 score increased from 0.75 to 0.89.

5 Discussion

This work demonstrates a practical and actionable application of interpretability insights through a generator-agnostic observer paradigm: a linear probe on a transformer's residual-stream activations identifies contextual hallucinations in a single forward pass, achieving high F_1 scores. These results underscore the feasibility of leveraging internal model representations for addressing key AI challenges like hallucinations. While the F_1 scores on news benchmarks could be partially influenced by characteristics inherent in prompted synthetic hallucinations, the strong performance on CONTRATALES (a dataset designed to test unambiguous logical contradictions) mitigates this concern by showcasing the probe's ability to identify more fundamental contextual violations, arguably less susceptible to specific generation artifacts.

Our findings offer support for the linear representation hypothesis (LRH) in contextual understanding, showing how interpretability can move beyond correlation to causal intervention. The identified linear direction for hallucination's consistency across Gemma-2 model sizes (2B→9B) and transferability across news domains suggest a fundamental encoding. Crucially, its functional role is substantiated by causal interventions: a single, layer-local injection or ablation of the probe vector along this axis smoothly and monotonically modulates a generator's hallucination and repetition rates, demonstrating it as an actionable, low-dimensional, and causally effective axis for contextual-hallucination awareness. Mechanistically, gradient-times-activation attribution analyses refine this picture, pinpointing the signal for contextual inconsistency to a sparse, layer-consistent pattern of late-layer MLP activity (e.g., layers 7 and 8 positively, layer 9 negatively, for a probe on layer 10 of Gemma-2-9B), rather than diffuse attention patterns. This indicates the observer's awareness is canalised through a specific chain of late-layer feed-forward computations.

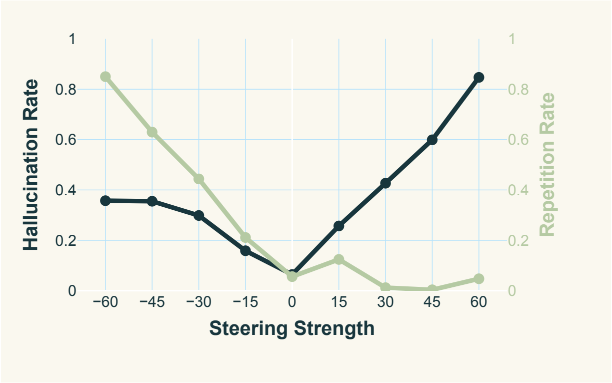

Figure 5: We use steering to generate outputs for CNNDM with the Gemma-2-2B model. We then use fuzzy string matching to determine the repetition rate, and gpt-4.1 to determine the hallucination rate.

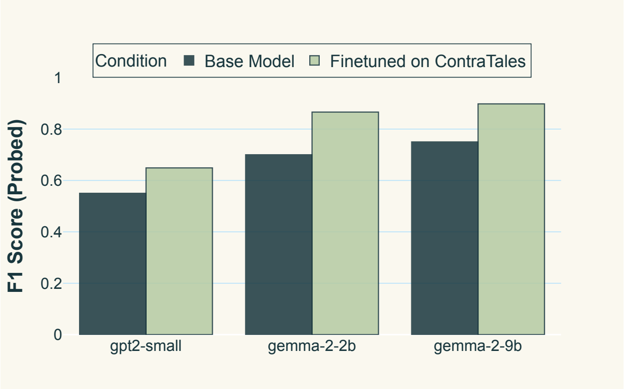

Figure 6: Unsupervised domain adaptation boosts probe accuracy. Bars report the F_1 of the residual-stream logistic probe before (base) and after (FT) two-epoch SFT on 1000 *correct* CONTRATALES continuations.

Furthermore, the practical utility of this internal representation is practically enhanced by unsupervised domain adaptation. Finetuning the observer model on in-domain, correct-only text from CONTRATALES (without new hallucination labels) significantly improved the logistic probe's F_1 score (e.g., for Gemma-2-9B, from 0.75 to 0.89). This implies that additional in-domain language modelling sharpens the internal distinction between logically supported and unsupported statements, a distinction the linear probe can then more effectively exploit, offering an inexpensive, label-free path to enhanced single-pass detection capabilities: important for real-world deployment.

Limitations A primary limitation of this work is the reliance on synthetically generated hallucinations for training and evaluating the majority of our detectors, particularly on the news and medical datasets. While necessary for creating labeled data at scale, continuations prompted from large language models may exhibit patterns or artifacts predictable to an observer model trained on similar data, which might not be representative of naturally occurring "in-the-wild" hallucinations generated by various models under different conditions. This could potentially lead to an overestimation of the detector's performance and generalisability beyond the specific generation methods used here. Although the CONTRATALES dataset offers a valuable benchmark for pure logical contradictions, it also represents a specific type of structured inconsistency. Further validation on datasets containing a broader spectrum of organically generated, human-verified hallucinations would strengthen claims regarding real-world applicability.

Beyond the nature of the training data, the evaluation of the steering experiments relies on a `gpt-4.1` judge to determine hallucination rates. While this is a common method, LLM-based evaluations can be subject to noise, bias, and variability, which may affect the precision of the quantitative steering results. Additionally, while the linear probe detector itself is lightweight and efficient, the observer paradigm necessitates deploying and running a sufficiently capable base transformer model, which still carries significant computational costs compared to purely surface-level detection heuristics. Finally, our study focuses specifically on **intrinsic** hallucinations that contradict or are unsupported by the provided source context; the applicability of the discovered hallucination direction and detection method to **extrinsic** hallucinations, which introduce novel but unverifiable information from outside sources, remains untested.

This research provides compelling evidence for a single, transferable, and causally effective linear direction within transformer activations that corresponds to contextual hallucination. This direction is primarily processed by a sparse and consistent MLP sub-circuit and can be leveraged for both lightweight detection and controlled generation, advancing interpretability by providing concrete methods for building more reliable AI systems. To facilitate further research, we release the CONTRATALES benchmark.

Impact Statement

This paper advances mechanistic interpretability by improving AI reliability and safety through evidence of generator-agnostic methods for detecting and controlling contextual hallucinations via mechanistic insights and causal steering. This offers potential for positive societal impact, enabling the deployment of more trustworthy AI in domains where accuracy is essential. However, this work also carries potential risks; the steering technique, while intended for mitigation, could be misused for generating misinformation, and the detector's performance may be influenced by biases present in its training data.

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Azaria, A. and Mitchell, T. The Internal State of an LLM Knows When It's Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, Dec 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering Latent Knowledge in Language Models Without Supervision. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., Schubert, L., Voss, C., Egan, B., and Lim, S. K. Thread: Circuits. *Distill*, 2020. doi: 10.23915/distill.00024. URL https://distill.pub/2020/circuits.

Chuang, Y.-S., Qiu, L., Hsieh, C.-Y., Krishna, R., Kim, Y., and Glass, J. R. Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps, 2024. URL https://arxiv.org/abs/2407.07071.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Dodds, Z. H., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022. URL https://arxiv.org/abs/2209.10652.

Engels, J., Liao, I., Michaud, E. J., Gurnee, W., and Tegmark, M. Not all language model features are linear. *arXiv e-prints*, pp. arXiv–2405, 2024.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Ferrando, J., Obeso, O., Rajamanoharan, S., and Nanda, N. Do I Know This Entity? Knowledge Awareness and Hallucinations in Language Models. *arXiv preprint arXiv:2411.14257*, 2024.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A., et al. spacy: Industrial-strength natural language processing in python. 2020.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 1(1):1–58, 2024. doi: 10.1145/3703155.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., and Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730.

Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., and Gal, Y. Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs, 2024. URL https://arxiv.org/abs/2406.15927.

Makhzani, A. and Frey, B. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Manakul, P., Liusie, A., and Gales, M. J. F. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, Dec 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

Minder, J., Du, K., Stoehr, N., Monea, G., Wendler, C., West, R., and Cotterell, R. Controllable Context Sensitivity and the Knob Behind It. *arXiv preprint arXiv:2411.07404*, 2024.