Steering Prepositional Phrases in Language Models: A Case of *with*-headed Adjectival and Adverbial Complements in Gemma-2

Stefan Arnold and **Rene Gröbner**
Friedrich-Alexander-Universität Erlangen-Nürnberg
Lange Gasse 20, 90403 Nürnberg, Germany
(stefan.st.arnold, rene.edgar.gröbner)@fau.de

Abstract

Language Models, when generating prepositional phrases, must often decide for whether their complements functions as an instrumental adjunct (describing the verb adverbially) or an attributive modifier (enriching the noun adjectivally), yet the internal mechanisms that resolve this split decision remain poorly understood. In this study, we conduct a targeted investigation into Gemma-2 to uncover and control the generation of prepositional complements. We assemble a prompt suite containing *with*-headed prepositional phrases whose contexts equally accommodate either an instrumental or attributive continuation, revealing a strong preference for an instrumental reading at a ratio of 3:4. To pinpoint individual attention heads that favor instrumental over attributive complements, we project activations into the vocabulary space. By scaling the value vector of a single attention head, we can shift the distribution of functional roles of complements, attenuating instruments to 33% while elevating attributes to 36%.

1 Introduction

Transformer-based (Vaswani et al., 2017) Language Models (LMs) (Devlin et al., 2019; Brown et al., 2020; Chowdhery et al., 2023) internalize rich inventories of dependency grammar (Jawahar et al., 2019; Hu et al., 2020) and deploy this structural knowledge to generate grammatically coherent sentences. Targeted evaluations showed that these models reliably choose the correct lexeme from a set of grammatically minimally different continuations in variety of syntactic constructions, including *agreement* (Linzen et al., 2016; Goldberg, 2019; Finlayson et al., 2021), *licensing* (Wilcox et al., 2018; Warstadt et al., 2019), *binding* (Marvin and Linzen, 2018), and the structure of *arguments* (Kann et al., 2019; Conia and Navigli, 2022).

In this study, we examine modifier attachment in prepositional phrases (PP). A PP typically contains a preposition as head immediately followed by a complement, which can serve as either an adverbial modifier or adjectival modifier (Nakashole and Mitchell, 2015). Figure 1 presents an illustrative example of a PP with distinct functional roles of their complements. In the adverbial role, the PP (i.e., *saw*) attaches to the verb phrase, denoting an instrument that describes an action. In the adjectival role, the PP (i.e., *notch*) attaches to the noun phrase, providing an attribute to an object. Both functional roles can compete during generation, and in some scenarios it might be desirable that the model prioritizes an instrumental reading, while in other scenarios, it might be more desirable to emphasize an attributive reading. Yet the internal mechanisms that govern whether an autoregressive LM produces a verb-modifying or noun-modifying adjunct have not been probed.

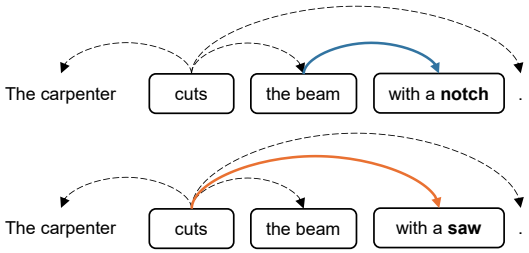

Figure 1: Example for the ambiguity of prepositional phrases: ↑ attaches the complement (*notch*) to the noun phrase (*beam*), acting as a adjectival modifier that supplies an attribute, whereas ↓ attaches the complement (*saw*) to the verb phrase (*cut*), serving as an adverbial modifier that introduces an instrument.

Recent work in the field of mechanistic interpretability (Wang et al., 2023; Geiger et al., 2025) is aimed at reverse engineering the internal processes carried out by LMs in terms of intelligible circuits. These circuits represent faithful simplifications of opaque computations, allowing us to interpret the contributions of individual model components to the final generated text. Circuit-level analyses have recently pinpointed components re-

Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 69–78

sponsible for performing induction (Olsson et al., 2022; Wang et al., 2023) and reasoning (Brinkmann et al., 2024), storing and retrieving factual knowledge (Meng et al., 2022; Geva et al., 2023; Merullo et al., 2023; Wang and Xu, 2025), and enforcing structural well-formedness (Finlayson et al., 2021).

Contribution. To dissect a circuit that characterizes the selectional preferences for preposition-led modifiers, we make three key contributions.

1. We design a controlled task by manually authoring prompts centered on *with*-introduced PP, whose context evokes both an instrument and an attribute, forcing a model to choose between these equally plausible but competing completions. We then assess this task on `Gemma-2` (Riviere et al., 2024) and note a preference toward adverbial instruments over adjectival attributes, occurring in a ratio of 3:4.

2. To localize the promotion of instrumentive and attributive continuations, we employ logit attribution (Belrose et al., 2023). By projecting internal activations into the vocabulary space, this technique allows us to make claims about the model components most responsible for favoring instruments or attributes. When applied to attention heads, we reveal a single head that consistently delivers the dominant direction towards instruments.

3. Through activation scaling (Merullo et al., 2023), we demonstrate a targeted intervention that reliably steers prepositional complements. By applying a scalar to the value of the attention head, we can shift the biased preference into a near-balanced distribution of instrumentive and attributive modifiers. We are able to raise the rate of attributes to 36% while reducing the amount of instruments to 33% by downweighting a single attention head, adjusting only a minimal number of parameters.

2 Related Work

2.1 Model Introspection

The interpretability community has devised a diverse set of techniques for understanding internal representations. These techniques span a spectrum of evidence: from *behavioral evaluations* that infer internals from output observations, through *structural correlations* that relate internals to formal linguistics, to *causal interventions* that manipulate internals to edit model behavior.

Behavioral evaluations study model outputs under carefully designed stimuli to reveal syntactic rules or semantic abilities. Linzen et al. (2016) assembled texts containing curated stimuli for which perplexity is evaluated as evidence of the presence or absence of linguistic knowledge.

Structural interpretations aim to identify linguistic properties captured in hidden states through auxiliary models applied at sentence-level (Adi et al., 2016; Conneau et al., 2018) and word-level (Tenney et al., 2019), while Hewitt and Manning (2019) particularly designed a structural probe to recover parse trees from hidden states.

Our work is built upon probing via vocabulary projections (Ghandeharioun et al., 2024) in which hidden representations are inspected in the vocabulary space by mapping them directly through the unembedding matrix rather than an auxiliary model. By applying the identity function, we can map representations from any layer to the final layer, interpreting every hidden state into a distribution over the vocabulary. Extensions apply linear (Pal et al., 2023) or affine (Belrose et al., 2023) mapping to improve the interpretability in the vocabulary space.

Causal interventions seek to locate the mechanisms that mediate behavior. By contrasting clean and corrupted inputs and patching intermediate activations from the clean run into the corrupted run, *activation patching* (Vig et al., 2020; Meng et al., 2022) puts forth an intervention for isolating functional circuits. The granularity of localization spans individual neurons (Finlayson et al., 2021), attention heads (Merullo et al., 2023; Wang et al., 2023), feedforward sublayers (Meng et al., 2022), and the residual stream (Belrose et al., 2023).

Apart from localization, interventions can also be used for model steering. Meng et al. (2023) modify weight matrices of feedforward layers to edit factual associations, whereas Rimsky et al. (2024) add a learned direction to the attention heads.

Recent interpretability efforts have derived functional mechanism responsible for performing tasks described in-context (Olsson et al., 2022; Hou et al., 2023; Brinkmann et al., 2024; Singh et al., 2024), storage and recall of factual knowledge and other forms of memories (Geva et al., 2023; Merullo et al., 2023; Wang and Xu, 2025), and concrete mechanism tailored to linguistic inflection (Finlayson et al., 2021), arithmetic calculation (Stolfo et al., 2023), and numerical comparison (Hanna et al., 2023). These findings reinforce that trans-

former internals can be decomposed into reusable motifs: attention heads copy and move information (Wang et al., 2023), feedforward layers serve as key–value memories (Geva et al., 2021), and the residual stream linearly accumulates the contributions from all model components.

2.2 Phrase Attachment

PP is typically framed from a parsing perspective, concerned with deciding whether a PP modifies the verb phrase or the noun phrase. Ambiguity arises because both attachments are often syntactically valid when semantic priors that license any reading are absent (Karamolegkou et al., 2025).

To disambiguate PP attachment, approaches rely on lexicalization (Ratnaparkhi et al., 1994; Pantel and Lin, 2000), contextualization (Ratnaparkhi and Kumar, 2021), and the integration of world knowledge (Nakashole and Mitchell, 2015). Although PP attachment has been studied extensively in parsing, the mechanisms by which LMs express selectional preferences over PP complements have not undergone a targeted analysis. We recast **PP attachment** to **PP completion** where the model must choose between two unambiguous adjuncts, both of which are made plausible by the preceding context.

3 Task Formulation

PP attachment is a classic problem in syntactic parsing, concerned with deciding whether a PP modifies the verb phrase (high attachment) or the noun phrase (low attachment). The ambiguity arises because both options are often syntactically licit in the absence of licensing context (Karamolegkou et al., 2025). Given a PP occurring within a sentence where multiple attachment sites are possible, the goal is to select the most plausible site.

We formalize PP attachment through an ordered quadruple (V, N, P, C), where V is the main verb, N is the head noun of its direct object, P is the preposition, and C is the candidate complement of the PP. The task is to decide whether the PP (P, C) attaches *adverbially* to V (instrumental adjunct) or *attributively* to N (nominal modifier). These roles yield distinct structures and meanings: an *instrument* specifies how the action is performed, whereas an *attribute* describes a property of the noun. Prior work typically treats the PP as a unit to be attached over a closed set of candidate parses.

To study the mechanisms LMs employ when context licenses contrastive PP continuations, we recast PP attachment as a generative zero-shot decision. Rather than choosing among candidate parses, the model must *continue* a *with*-headed PP with either an instrument or an attribute. This framing serves as a representative probe of selectional preferences. Our experimental design relies on stimuli in which role attribution of the PP adjunct is unequivocal. Because PP corpora feature equivocal constructions, we needed to construct a manually curated set of prompts in which each continuation is unambiguous in its syntactic role, yet the surrounding context is crafted to render both options equally coherent and contextually plausible.

We reformulate PP attachment for continuation so that the model is tasked with producing the PP complement. To make both readings viable while keeping the options functionally distinct, we manually curate pairs of complements that are *each* unambiguous in role identity but jointly create a context in which either completion is plausible. We constrain prompts to a structured subject-verb-object frame followed by a *with*-headed PP:

(1) The *subject verb* the *object* with a ...

This placement ensures that both instrumental and attributive continuations are plausible and the design naturally suits zero-shot prompting in which the model generates the PP complement directly.

Because PP complement generation depends on world knowledge, we augment the prompts with a minimal licensing context that associates the subject with a plausible instrument and the object with a plausible attribute. Each context introduces two entities, where the subject-associated noun acts as plausible *instrument* for the action and the object-associated noun as plausible *attribute* of the object.

(2) A *subject* has a *subject-associated noun*.

(3) A *object* has a *object-associated noun*.

For our experiments, we select `Gemma-2`, an autoregressive model with two billion parameters. `Gemma-2` is an enticing candidate for probing mechanisms of selectional preference due to its large vocabulary size with numerous reserved words, and studying it also contributes to the growing body of interpretation for the `Gemma` family (Lieberum et al., 2024).

Table 1 presents selected examples of our manually licensed prompts. Appendix A provides the complete set of prompts, comprising a total of 100. We observe that the `Gemma-2` model manifests instrumental adjuncts rather than attributive modi-

Prompts with licensing contexts			Instrument	Attribute
A carpenter has a *saw*.	A beam has a *notch*.	The carpenter cuts the beam with a	**saw**	notch
A chef has a *syringe*.	A cake has a *frosting*.	The chef decorated the cake with a	**syringe**	frosting
A florist has a *shear*.	A bouquet has a *rose*.	The florist trims the bouquet with a	**shear**	rose
A pilot has a *joystick*.	A plane has a *failure*.	The pilot lands the plane with a	**joystick**	failure
A welder has a *torch*.	A joint has a *crack*.	The welder seals the joint with a	**torch**	crack

Table 1: Excerpt from our prompt suite. *italic* indicates the candidate instrument and attribute; underline represents the phrase triplet containing a verb, noun, and preposition for which the model must decide whether it attaches the instrument or attribute; **bold** highlights the preferred prepositional complement.

fier. This observation is consistent with psycholinguistic studies reporting that humans tend to favor high attachment to verbs over low attachment to nouns (Spivey-Knowlton and Sedivy, 1995), and with recent findings indicating that language models display a bias toward instrumental rather than attributive readings (Zhou et al., 2024).

4 Logit Attribution

We aim to isolate a faithful mechanism within our language model that governs the selectional preferences for the formation of PP.

To pinpoint model components driving this selectional preference for an adverbial or adjectival reading, we turn to *logit attribution* (Belrose et al., 2023). The idea behind logit attribution is to interpret the role of a particular component in a language model for a given task in terms of the vocabulary space. This is built on the premise that the residual stream can be decomposed into the sum of contributions from every model component. Recall that each model component in the transformer adds its output onto the residual stream, and the residual stream state gets projected onto the unembedding matrix, producing the logits distribution. Due to the linearity of the residual stream, every layer of computation can be traced back as the direct effect of each sublayer to the logits up to that point.

Because PP continuation in our task boils down to copying one of two lexical adjuncts from a context that licenses an instrument and attribute equally plausibly, we apply logit attribution to attention heads. This choice is motivated by the key role of attention heads in performing copying operations (Wang et al., 2023; Merullo et al., 2023).

We obtain the direct effects of attention heads favoring instrumental or attributive readings following Merullo et al. (2023). We start by extracting the corresponding vectors in the unembedding matrix for the target adjuncts. The additive update made by the attention layer is composed of the concatenated updates of each attention head after it is passed through the output weight matrix within the attention layer. We therefore divide the weight matrix of the attention output into one component for each attention head and project the head activations into the space of the residual stream by multiplying them with the corresponding slice of weight matrix. We then dot product the projected activation of the attention head with the weight vectors of the attribute and instrument from the unembedding matrix, giving us a scalar value representing the logit for each of those continuations represented by the head. We then compute the dot product between the projected activation of a given head and the unembedding vector for each target word, yielding a scalar logit for each adjunct. By subtracting these two logit values, we get the direct attribution to the logit difference between attribute and instrument. This logit difference captures the effect the head has in promoting one word (relative to another) to be the continuation: a positive value indicates that the head writes in the direction of the attribute, promoting an adjectival reading, whereas a negative value indicates that the head writes in the direction of the instrument, promoting an adverbial reading.

Figure 2 visualizes the logit difference calculation for each head in every layer. Since `Gemma-2` has 26 layers and 8 heads for each layer, this totals 208 heads to test. Despite some variation in the roles of every head throughout our prompt suite, we identify a series of heads that push the model towards attributive modifiers or instrumental adjuncts. However, the heads that consistently affect PP completion are clustered in early layers, and these heads uniformly drive the model toward instrumental adjuncts. `L0H2` emerges as the principal driver of *with*-headed phrase completions, rendering it an ideal target for steering interventions.

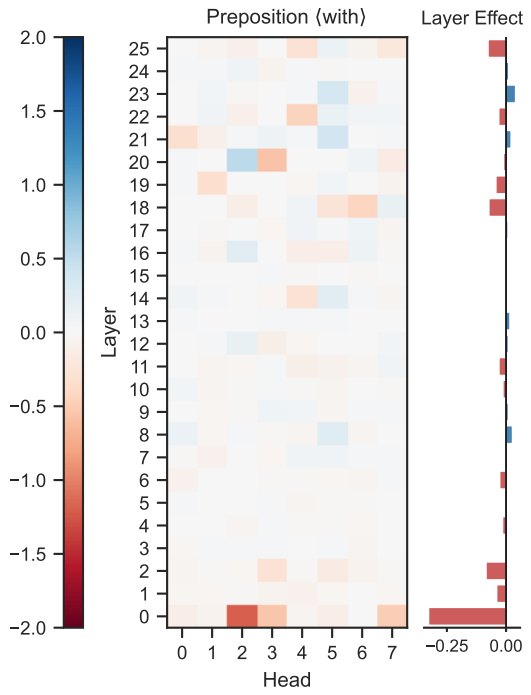

Figure 2: Attribution map of direct effects for *with*-headed PP constructions. Only a single attention head shows a strong selectional preference.

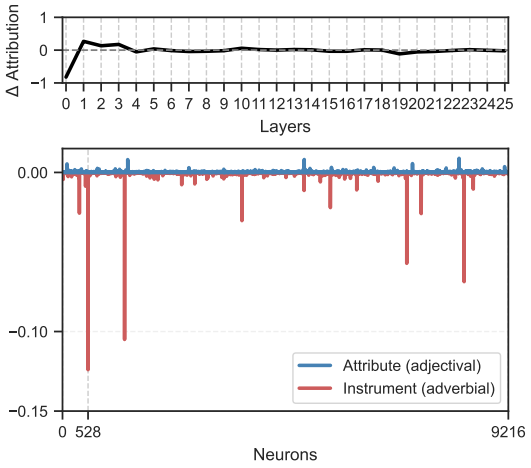

Figure 3: Attribution across feedforward layers. Only a sparse subset of neurons contribute to the choice between attributive and instrumental readings.

To validate that attention heads provide the primary signals for PP continuation, we extended our attribution analysis to feedforward layers. Figure 3 plots the averaged preference for all 26 layers along with its scores for the 9216 neurons (only for the initial layer). We observe that most feedforward layers show no consistent directional bias, contributing similarly to both readings. Any non-zero attributions are almost exclusively confined to

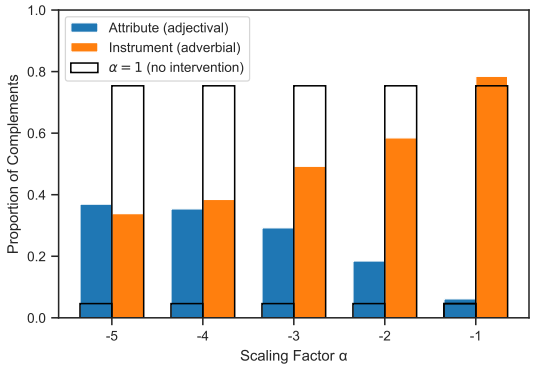

Figure 4: Proportion of shifts in selectional preferences by the multiplicative value under activation scaling.

initial feedforward layer, where only a handful of neurons register meaningful effects on the instrumental versus attributive difference in logits, and even these contributions are an order of magnitude smaller than those we observed for attention heads. This lack of directional specificity in neuron attributions allows us to exclude feedforward sublayers as primary drivers of PP completion decisions.

5 Activation Steering

Several techniques have been proposed to steer the generative process of LMs. Merullo et al. (2023) scale the activation of an attention head by a scalar, whereas Rimsky et al. (2024) add a learned direction to the activation of an attention head. Since a single attention head heavily contributes to the direction in logit, we adopt the scaling intervention.

We hypothesize that downweighting `L0H2` will enable us to suppress instrumental readings and boost attributive readings. To test our hypothesis, we apply a multiplicative factor $\alpha < 0$ to the value vector of `L0H2`. The effect of this intervention is measured by the proportion of times the model flips the functional role of the PP complement from an instrumental modifier to an attributive modifier.

Figure 4 presents the proportions of instruments and attributes as a function of the scaling factor $\alpha \in [-5, -4, -3, -2, -1]$. Without any intervention, the model selects the instrument in 75%, an attribute in 4%, and deploys unlicensed words as PP complements in 21%, suggesting a marked bias in the functional roles of PP complements. We find that scaling down attention head `L0H2` has a strong effect on flipping prepositional complements [1]. We

[1] We note that scaling other attention heads (not depicted in Figure 4) produces markedly weaker shifts, underscoring the unique role of `L0H2` on selectional preferences.

can attenuate instruments to 33% while elevating attributes to 36%, as we decrease the multiplicative value of α. However, even extreme downweighting does not eliminate instrumental readings entirely, and a substantial portion of completions fall into alternative adjuncts outside the two intended options, indicating that aggressive steering can divert the model toward unlicensed complements.

6 Conclusion

Through a controlled prompt design, logit attribution, and activation scaling, we isolated and manipulated a single attention head in `Gemma-2` that exerts a dominant influence on the balance between adverbial and adjectival continuations following from prepositional phrases. Our findings provide a principled proof of concept for steering model output in contexts where multiple continuations are syntactically and semantically plausible, bridging interpretable and controllable generation. This study advances our understanding of how autoregressive models internally resolve functional role preferences and demonstrates that such steering can be achieved with minimal parameter interventions.

Limitations. We acknowledge two main limitations. First, we conduct all experiments on a single model. This choice is motivated by its large vocabulary size, which facilitates controlled and targeted mechanistic analysis. However, it does not guarantee that our identified mechanism generalizes to other model architectures. Second, we restrict our investigation to PPs as a representative case of selectional preference under ambiguity. Although this scope offers a controlled and well-documented testing ground, it represents only one narrow subset of the model's broader selectional preferences.

We plan to extend our mechanistic understanding of selectional preferences of ambiguous continuations to garden-path effects (Amouyal et al., 2025) and control over the assignment of predicate-argument structure like *purpose*, *location*, or *time*, going beyond adverbial and adjectival modifiers.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Samuel Joseph Amouyal, Aya Meltzer-Asscher, and Jonathan Berant. 2025. When the LM misunderstood the human chuckled: Analyzing garden path effects in humans and language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8235–8253, Vienna, Austria. Association for Computational Linguistics.

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

Jannik Brinkmann, Abhay Sheshadri, Victor Levoso, Paul Swoboda, and Christian Bartelt. 2024. A mechanistic analysis of a transformer trained on a symbolic multi-step reasoning task. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4082–4102, Bangkok, Thailand. Association for Computational Linguistics.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Simone Conia and Roberto Navigli. 2022. Probing for predicate argument structures in pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4622–4632, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on*

Natural Language Processing (Volume 1: Long Papers), pages 1828–1843, Online. Association for Computational Linguistics.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. 2025. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36:76033–76060.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. 2023. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4902–4919, Singapore. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 287–297.

Antonia Karamolegkou, Oliver Eberle, Phillip Rust, Carina Kauf, and Anders Søgaard. 2025. Trick or neat: Adversarial ambiguity and language model evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18542–18561, Vienna, Austria. Association for Computational Linguistics.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Jack Merullo, Qinan Jack Yu, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959.

Ndapandula Nakashole and Tom M. Mitchell. 2015. A knowledge-intensive model for prepositional phrase attachment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 365–375, Beijing, China. Association for Computational Linguistics.

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

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron C Wallace, and David Bau. 2023. Future lens: Anticipating subsequent tokens from a single hidden state. *arXiv preprint arXiv:2311.04897*.

Patrick Pantel and Dekang Lin. 2000. An unsupervised approach to prepositional phrase attachment using contextually similar words. In *Proceedings of the 38th annual meeting of the Association for Computational Linguistics*, pages 101–108.

Adwait Ratnaparkhi and Atul Kumar. 2021. Resolving prepositional phrase attachment ambiguities with contextualized word embeddings. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 335–340.

Adwait Ratnaparkhi, Jeff Reynar, and Salim Roukos. 1994. A maximum entropy model for prepositional phrase attachment. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 15504–15522.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Aaditya K Singh, Ted Moskovitz, Felix Hill, Stephanie CY Chan, and Andrew M Saxe. 2024. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 45637–45662.

Michael Spivey-Knowlton and Julie C Sedivy. 1995. Resolving attachment ambiguities with multiple constraints. *Cognition*, 55(3):227–267.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052, Singapore. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Zijian Wang and Chang Xu. 2025. Functional abstraction of knowledge recall in large language models. *arXiv preprint arXiv:2504.14496*.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Lingling Zhou, Suzan Verberne, and Gijs Wijnholds. 2024. Tree transformer's disambiguation ability of prepositional phrase attachment and garden path effects. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12291–12301, Bangkok, Thailand. Association for Computational Linguistics.

A Prompt Suite

Subject-Instrument	Object-Attribute	Subject-Verb-Object	Adjuncts
⟨baker, whisk⟩	⟨bowl, lump⟩	⟨baker, stirs, bowl⟩	⟨whisk, lump⟩
⟨banker, spreadsheet⟩	⟨portfolio, stock⟩	⟨banker, edits, portfolio⟩	⟨spreadsheet, stock⟩
⟨barber, scissor⟩	⟨beard, fringe⟩	⟨barber, trims, beard⟩	⟨scissor, fringe⟩
⟨barista, portafilter⟩	⟨cappuccino, foam⟩	⟨barista, prepares, cappuccino⟩	⟨portafilter, foam⟩
⟨bartender, shaker⟩	⟨cocktail, garnish⟩	⟨bartender, prepares, cocktail⟩	⟨shaker, garnish⟩
⟨biologist, pipette⟩	⟨tube, liquid⟩	⟨biologist, tranfers, tube⟩	⟨pipette, liquid⟩
⟨bonesetter, splint⟩	⟨patient, fracture⟩	⟨bonesetter, stabilizes, patient⟩	⟨splint, fracture⟩
⟨brewer, keg⟩	⟨beer, trademark⟩	⟨brewer, dispenses, beer⟩	⟨keg, trademark⟩
⟨builder, spatula⟩	⟨wall, crack⟩	⟨builder, repairs, wall⟩	⟨spatula, crack⟩
⟨butcher, cleaver⟩	⟨steak, marbling⟩	⟨butcher, cuts, steak⟩	⟨cleaver, marbling⟩
⟨carpenter, saw⟩	⟨beam, notch⟩	⟨carpenter, cuts, beam⟩	⟨saw, notch⟩
⟨carpenter, chisel⟩	⟨plank, groove⟩	⟨carpenter, deepens, plank⟩	⟨chisel, groove⟩
⟨cartographer, compass⟩	⟨map, legend⟩	⟨cartographer, aligns, map⟩	⟨compass, legend⟩
⟨chef, ladle⟩	⟨pot, broth⟩	⟨chef, serves, pot⟩	⟨ladle, broth⟩
⟨chef, spoon⟩	⟨egg, shell⟩	⟨chef, cracks, egg⟩	⟨spoon, shell⟩
⟨chef, syringe⟩	⟨cake, frosting⟩	⟨chef, decorates, cake⟩	⟨syringe, frosting⟩
⟨chef, spatula⟩	⟨meal, marinade⟩	⟨chef, flips, meal⟩	⟨spatula, marinade⟩
⟨chef, spice⟩	⟨soup, flavor⟩	⟨chef, seasons, soup⟩	⟨spice, flavor⟩
⟨chemist, pipette⟩	⟨reaction, precipitate⟩	⟨chemist, measures, reaction⟩	⟨pipette, precipitate⟩
⟨chemist, flask⟩	⟨reaction, catalyst⟩	⟨chemist, conducts, reaction⟩	⟨flask, catalyst⟩
⟨cleaner, vacuum⟩	⟨carpet, crumb⟩	⟨cleaner, cleans, carpet⟩	⟨vacuum, crumb⟩
⟨coach, whistle⟩	⟨team, streak⟩	⟨coach, signals, team⟩	⟨whistle, streak⟩
⟨conductor, baton⟩	⟨orchestra, listener⟩	⟨conductor, directs, orchestra⟩	⟨baton, listener⟩
⟨cosmologist, telescope⟩	⟨planet, moon⟩	⟨cosmologist, observes, planet⟩	⟨telescope, moon⟩
⟨cosmonaut, spacesuit⟩	⟨capsule, porthole⟩	⟨cosmonaut, abandones, capsule⟩	⟨spacesuit, porthole⟩
⟨dentist, mirror⟩	⟨tooth, cavity⟩	⟨dentist, examines, tooth⟩	⟨mirror, cavity⟩
⟨designer, tablet⟩	⟨product, stamp⟩	⟨designer, creates, product⟩	⟨tablet, stamp⟩
⟨detective, lens⟩	⟨scene, clue⟩	⟨detective, inspects, scene⟩	⟨lens, clue⟩
⟨diver, camera⟩	⟨reef, fish⟩	⟨diver, captures, reef⟩	⟨camera, fish⟩
⟨doctor, thermometer⟩	⟨child, disease⟩	⟨doctor, checks, child⟩	⟨thermometer, disease⟩
⟨draughtsman, ruler⟩	⟨blueprints, balcony⟩	⟨draughtsman, edits, blueprints⟩	⟨ruler, balcony⟩
⟨driver, wheel⟩	⟨road, curve⟩	⟨driver, navigates, road⟩	⟨wheel, curve⟩
⟨driver, wrench⟩	⟨car, tire⟩	⟨driver, repairs, car⟩	⟨wrench, tire⟩
⟨farmer, plow⟩	⟨field, furrow⟩	⟨farmer, cuts, field⟩	⟨plow, furrow⟩
⟨firefighter, ladder⟩	⟨cat, collar⟩	⟨firefighter, saves, cat⟩	⟨ladder, collar⟩
⟨fisherman, net⟩	⟨crab, shell⟩	⟨fisherman, captures, crab⟩	⟨net, shell⟩
⟨florist, shear⟩	⟨bouquet, rose⟩	⟨florist, trims, bouquet⟩	⟨shear, rose⟩
⟨gardener, rake⟩	⟨garden, tree⟩	⟨gardener, grooms, garden⟩	⟨rake, tree⟩
⟨gardener, can⟩	⟨plant, stem⟩	⟨gardener, waters, plant⟩	⟨can, stem⟩
⟨gardener, shovel⟩	⟨soil, worms⟩	⟨gardener, digs, soil⟩	⟨shovel, worms⟩
⟨gardener, shears⟩	⟨hedge, nest⟩	⟨gardener, prunes, hedge⟩	⟨shears, nest⟩
⟨gardener, spade⟩	⟨garden, border⟩	⟨gardener, outlines, garden⟩	⟨spade, border⟩
⟨geologist, scale⟩	⟨rock, fissure⟩	⟨geologist, measures, rock⟩	⟨scale, fissure⟩
⟨guard, weapon⟩	⟨property, fence⟩	⟨guard, protects, property⟩	⟨weapon, fence⟩
⟨hunter, rifle⟩	⟨forest, deer⟩	⟨hunter, targets, forest⟩	⟨rifle, deer⟩
⟨janitor, mop⟩	⟨floor, scuffing⟩	⟨janitor, cleans, floor⟩	⟨mop, scuffing⟩
⟨jeweler, cloth⟩	⟨ring, diamond⟩	⟨jeweler, examines, ring⟩	⟨cloth, diamond⟩
⟨journalist, recorder⟩	⟨politician, controversy⟩	⟨journalist, interviews, politician⟩	⟨recorder, controversy⟩
⟨judge, hammer⟩	⟨trial, verdict⟩	⟨judge, concludes, trial⟩	⟨hammer, verdict⟩
⟨laboratorian, centrifuge⟩	⟨sample, contamination⟩	⟨laboratorian, separates, sample⟩	⟨centrifuge, contamination⟩
⟨lawyer, highlighter⟩	⟨contract, clause⟩	⟨lawyer, reviews, contract⟩	⟨highlighter, clause⟩
⟨librarian, scanner⟩	⟨book, cover⟩	⟨librarian, catalogs, book⟩	⟨scanner, cover⟩

Subject-Instrument	Object-Attribute	Subject-Verb-Object	Adjuncts
⟨lifeguard, whistle⟩	⟨swimmer, monofin⟩	⟨lifeguard, signals, swimmer⟩	⟨whistle, monofin⟩
⟨lineman, multimeter⟩	⟨circuit, voltage⟩	⟨lineman, tests, circuit⟩	⟨multimeter, voltage⟩
⟨locksmith, dietrich⟩	⟨keyway, vulnerability⟩	⟨locksmith, enters, keyway⟩	⟨dietrich, vulnerability⟩
⟨magician, wand⟩	⟨mirage, misdirection⟩	⟨magician, conjures, mirage⟩	⟨wand, misdirection⟩
⟨mason, level⟩	⟨wall, cladding⟩	⟨mason, trues, wall⟩	⟨level, cladding⟩
⟨mathematician, chalkboard⟩	⟨formula, mistake⟩	⟨mathematician, derives, formula⟩	⟨chalkboard, mistake⟩
⟨mechanic, wrench⟩	⟨engine, leak⟩	⟨mechanic, fixes, engine⟩	⟨wrench, leak⟩
⟨midwife, doppler⟩	⟨woman, complication⟩	⟨midwife, monitors, woman⟩	⟨doppler, complication⟩
⟨miner, axe⟩	⟨rock, gem⟩	⟨miner, breaks, rock⟩	⟨axe, gem⟩
⟨musician, tuner⟩	⟨piece, pitch⟩	⟨musician, tunes, piece⟩	⟨tuner, pitch⟩
⟨musician, turntable⟩	⟨song, rhythm⟩	⟨musician, streches, song⟩	⟨turntable, rhythm⟩
⟨neurologist, penlight⟩	⟨pupil, dilation⟩	⟨neurologist, assesses, pupil⟩	⟨penlight, dilation⟩
⟨nurse, syringe⟩	⟨arm, vaccine⟩	⟨nurse, treats, arm⟩	⟨syringe, vaccine⟩
⟨painter, brush⟩	⟨wall, patch⟩	⟨painter, overpaints, wall⟩	⟨brush, patch⟩
⟨painter, roller⟩	⟨canvas, sketch⟩	⟨painter, brushes, canvas⟩	⟨roller, sketch⟩
⟨performer, script⟩	⟨scene, prop⟩	⟨performer, enters, scene⟩	⟨script, prop⟩
⟨pharmacist, mortar⟩	⟨leaf, stem⟩	⟨pharmacist, grinds, leaf⟩	⟨mortar, stem⟩
⟨photographer, flash⟩	⟨portrait, shadow⟩	⟨photographer, illuminates, portrait⟩	⟨flash, shadow⟩
⟨photographer, camera⟩	⟨scene, horizon⟩	⟨photographer, captures, scene⟩	⟨camera, horizon⟩
⟨physiotherapist, spirometer⟩	⟨patient, symptom⟩	⟨physiotherapist, screens, patient⟩	⟨spirometer, symptom⟩
⟨pilot, joystick⟩	⟨plane, failure⟩	⟨pilot, controls, plane⟩	⟨joystick, failure⟩
⟨plumber, wrench⟩	⟨pipe, leak⟩	⟨plumber, seals, pipe⟩	⟨wrench, leak⟩
⟨policeman, handcuff⟩	⟨criminal, scar⟩	⟨policeman, arrests, criminal⟩	⟨handcuff, scar⟩
⟨prehistorian, shovel⟩	⟨fossil, patina⟩	⟨prehistorian, excavates, fossil⟩	⟨shovel, patina⟩
⟨programmer, keyboard⟩	⟨database, password⟩	⟨programmer, accesses, database⟩	⟨keyboard, password⟩
⟨programmer, debugger⟩	⟨codebase, bug⟩	⟨programmer, debugs, codebase⟩	⟨debugger, bug⟩
⟨ranger, tranquilizer⟩	⟨tiger, wound⟩	⟨ranger, paralyzes, tiger⟩	⟨tranquilizer, wound⟩
⟨receptionist, telephone⟩	⟨visitor, question⟩	⟨receptionist, calls, visitor⟩	⟨telephone, question⟩
⟨roofer, harness⟩	⟨rope, knot⟩	⟨roofer, fastens, rope⟩	⟨harness, knot⟩
⟨scientist, microscope⟩	⟨slide, specimen⟩	⟨scientist, examines, slide⟩	⟨microscope, specimen⟩
⟨sculptor, chisel⟩	⟨block, grain⟩	⟨sculptor, carves, block⟩	⟨chisel, grain⟩
⟨singer, microphone⟩	⟨stage, spotlight⟩	⟨singer, performs, stage⟩	⟨microphone, spotlight⟩
⟨sniper, scope⟩	⟨hideout, threat⟩	⟨sniper, targets, hideout⟩	⟨scope, threat⟩
⟨statistician, notebook⟩	⟨datasets, schema⟩	⟨statistician, analyzes, datasets⟩	⟨notebook, schema⟩
⟨stenographer, headset⟩	⟨speech, message⟩	⟨stenographer, transcribes, speech⟩	⟨headset, message⟩
⟨student, pen⟩	⟨textbook, diagram⟩	⟨student, marks, textbook⟩	⟨pen, diagram⟩
⟨surgeon, knife⟩	⟨tumor, mass⟩	⟨surgeon, removes, tumor⟩	⟨knife, mass⟩
⟨tailor, needle⟩	⟨suit, tear⟩	⟨tailor, mends, suit⟩	⟨needle, tear⟩
⟨tailor, thread⟩	⟨fabric, seam⟩	⟨tailor, stitches, fabric⟩	⟨thread, seam⟩
⟨tailor, tape⟩	⟨dress, pattern⟩	⟨tailor, measures, dress⟩	⟨tape, pattern⟩
⟨teacher, pointer⟩	⟨presentation, figure⟩	⟨teacher, points, presentation⟩	⟨pointer, figure⟩
⟨teacher, chalk⟩	⟨board, equation⟩	⟨teacher, writes, board⟩	⟨chalk, equation⟩
⟨topographer, theodolite⟩	⟨bridge, camber⟩	⟨topographer, surveys, bridge⟩	⟨theodolite, camber⟩
⟨translator, dictionary⟩	⟨text, term⟩	⟨translator, translates, text⟩	⟨dictionary, term⟩
⟨vet, stethoscope⟩	⟨pet, stroke⟩	⟨vet, monitors, pet⟩	⟨stethoscope, stroke⟩
⟨waiter, cloth⟩	⟨table, decoration⟩	⟨waiter, cleans, table⟩	⟨cloth, decoration⟩
⟨welder, torch⟩	⟨joint, crack⟩	⟨welder, seals, joint⟩	⟨torch, crack⟩
⟨writer, pen⟩	⟨manuscript, flaw⟩	⟨writer, corrects, manuscript⟩	⟨pen, flaw⟩