Mechanistic Interpretability Meets Cognitive Linguistics: Modelling Locative *Image Schemas* in the Circuit Framework

Mattia Proietti[♣♠*], **Afra Alishahi**[♠], **Grzegorz Chrupała**[♠], **Alessandro Lenci**[♣]
[♠]Cognitive Science and Artificial Intelligence Research Centre, Tilburg University,
[♣]CoLing Lab, Department of Philology, Literature and Linguistics, University of Pisa,
mattia.proietti@phd.unipi.it,
{a.alishahi, g.a.chrupala}@tilburguniversity.edu,
alessandro.lenci@unipi.it

Abstract

Large Language Models are often considered the best computational testbeds for linguistic theorisation at our disposal. However, their inner workings remain largely opaque, and the mechanisms behind their behaviour cannot always be easily connected with theoretical linguistic assumptions. Mechanistic Interpretability (MI) is surging as a specialised field to reverse engineer models' internals and shed light on the causal relationships happening under the hood. Nevertheless, MI is predominantly focused on AI-Safety problems, and the attempts to understand linguistically motivated behaviours with these tools are still limited. In this work, we investigate whether an LLM, namely LlaMA-3.2-1b, has developed specialised mechanisms governing the selection of the locative preposition in simple copular clauses. To frame the problem as a *next-token prediction* objective, we introduce the *Stranded Locative Preposition Selection* task along with a small dataset aptly curated to test it. We make use of several MI tools to scan the model's internals and relate their mechanisms to classic theory in Cognitive Linguistics, which assumes that the two basic locative prepositions *in* and *on* are the respective linguistic encoding of two different *Image Schemas*: CONTAINMENT and SURFACE.

Keywords: Mechanistic Interpreability, Semantics, Image Schemas, Prepositions, Spatial Semantics, Cognitive Linguistics

1. Introduction

Large Language Models (LLMs) have been one of the greatest revolutions of the last ten years, bearing the promise of offering the most advanced models and theories of language at our disposal (Piantadosi and Hill, 2022; Piantadosi, 2023; Millière, 2024). While this is controversial (Bender and Koller, 2020; Katzir, 2023; Pavlick, 2023; Lenci, 2023), it is undeniable that they constitute a new testbed for linguistic theorisation, allowing to computationally assess theoretical assumptions about language phenomena. However, this task is hindered by the lack of understanding about the causal mechanisms responsible for their behaviours. **Mechanistic Interpretability** (MI) (Olah et al., 2020; Geiger et al., 2025; Rai et al., 2024; Sharkey et al., 2025; Somvanshi et al., 2025) is recently emerging as a field to elucidate the inner workings of LLMs, particularly of Transformers (Vaswani et al., 2017), to uncover causal theories about their functioning. While efforts in MI are currently mainly devoted to research in AI safety (Bereska and Gavves, 2024), the tools and techniques coming from the field may be useful to lead scientific investigations on how language phenomena are encoded into LLMs representations and parameters (Mueller et al., 2024). On the other hand, linguistically informed problems may provide useful edge cases to test the soundness and utility of such tools, expanding the pool of possible tasks to test beyond the usual ones. In this work, we draw on literature in the Cognitive Linguistics tradition (Lakoff, 1987; Johnson, 1987; Tyler and Evans, 2003; Clausner and Croft, 1997) and developmental psychology (Mandler, 1994; Mandler and Pagán Cánovas, 2014), turning our attention to ***Spatial Image Schemas***, which we investigate in an LLM through the lens of MI tools. We introduce the task of *Locative Stranded Preposition Prediction* along with a small dataset of manually curated stimuli to test it. In English, a *stranded preposition* occurs when it is separated from its object, like shown in Figure 1. To correctly solve the task, a model has to continue a sentence with the correct locative preposition, relying on the information provided by previous tokens. The task

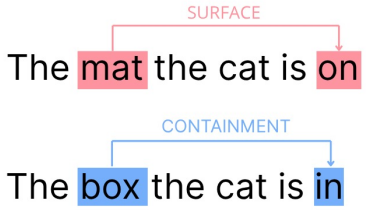

Figure 1: Stranded locative preposition and *Image Schemas* alternation.

*Work done while visiting Tilburg University, Department of Cognitive Science and Artificial Intelligence

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11320–11331
11–16 May 2026.

requires access to important semantic information related to the *Image Schema*, which abstracts from a real-world spatial configuration between two entities, a FIGURE and a LANDMARK. The former is a movable object located with respect to the latter, which is a *place* in the background (Herskovits, 2009). The nature of the LANDMARK plays an important role in the selection of the preposition, as certain LANDMARKS will be conceptualized as enabling CONTAINMENT and others SURFACE *schemas*. Consider the distinction between *mat* and *box* in the minimal pair in Figure 1: the former instantiates a SURFACE *schema*, selecting the preposition *on*, and the latter a CONTAINMENT selecting *in*, while both conveying the same general locative meaning.

Hinging on that basic distinction, we ask the following questions: **RQ1**) *which role does the LANDMARK token play in selecting the preposition in LLMs?* and **RQ2**) *Can we localise a small subgraph in the model that is responsible for the preposition prediction behaviour?*

Our main contributions are threefold: i) we propose a framework to evaluate LLMs' linguistic competence leveraging MI techniques and pairing them with theoretical linguistics assumptions; ii) we introduce the *stranded preposition prediction* task as a new case study to test the framework, and iii) we introduce a new curated dataset to benchmark it.[1]

1.1. Case Study: *Image Schemas* and Preposition Stranding

Some theories of conceptual representations rely on the assumption that humans build concepts, and hence develop a semantic theory, by leveraging abstract mental structures sometimes defined as *Image Schemas* (Lakoff, 1987; Johnson, 1987; Mandler and Pagán Cánovas, 2014), pre-linguistic constructions which abstract frequent perceptual patterns reflecting actual real-world configurations. These conceptual structures are often introduced as the representation of the interaction of two fundamental elements, one of which is a movable entity, while the other is a fixed reference or background (Levinson, 2003). While such elements have been called differently in the literature, their nature and essential substance do not vary much across different namings and definitions. The movable entity may be called a FIGURE, while the background can be referred to as a LANDMARK[2]. Starting from the spatial configuration of these two elements as well as their dynamic interplay, meaning is mapped from a source domain, which is physical and perceptual, to a target domain, conceptual and abstract in nature. This mapping from the experiential to the conceptual realm functions as an abstraction of frequent real-world patterns perceived through the body and the senses and serves as the scaffold for meaning at different degrees of adherence to spatial reality, including conceptual metaphors. An example of a well-known *Image Schema* is the CONTAINMENT schema. This schema abstracts from our knowledge of objects that work as containers in real life and projects the containment function to other areas of semantics, enabling the generation of meanings that may be more or less metaphorical. For example, when saying something like "I have something *in my head*", a speaker would leverage the CONTAINMENT *Image Schema* through the MIND AS A CONTAINER conceptual metaphor, operating a mapping between her perceptual knowledge of containers and a conceptual abstraction that relate the head/mind to an object bearing properties like boundedness and interior/exterior opposition (Johnson, 1987). A much less metaphorical, though abstract, usage of such a schema can be found in expressions as "Bob is *in town*", where the subject is presented as enclosed, and thus *contained*, into a location, even though the "town", acting here as a LANDMARK for the FIGURE "Bob", is not a proper container in reality.

As such, *Image Schemas* would not be only abstract conceptual representations, but would also be encoded into language structure and could be partially recovered from it by looking at the formal patterns they produce in linguistic data. In several languages, including English, a prominent example of this phenomenon is the selection of prepositions. For example, locative expressions may be realised through the prepositions *in* and *on*, which are selected in relation to the *schema* the speaker intends to express (Tyler and Evans, 2003), namely: CONTAINMENT, (expressed through *in*) and SURFACE, (expressed through *on*).[3]. Both *schemas* are locative, expressing the meaning of "*something being somewhere*". Still, they differ in subtle yet important aspects related to the nature of the LANDMARK (see Figure 1). In fact, the nature of the spatial configuration is largely determined by the LANDMARK, in particular by the shape and affordances of its referent, which impose constraints on the spatial scene conceptualisation (Zwarts, 2017; Herskovits, 1985). In that respect, CONTAINMENT involves a LANDMARK that offers some form of *inclusion* to the FIGURE, while SURFACE presupposes the notions of *contiguity* and *support* between the two Thus, for example, a LANDMARK as *mat* enables the conceptualisation of a SURFACE, determining the selection of the preposition *on*. On the con-

[1] Code and data available at https://github.com/aittam9/stranded_prep.

[2] Other naming have been proposed in the literature like *trajectory* or *ground*, but the substance remains essentially the same. See for example Talmy (2000)

[3] There are some notable exceptions like in *on a bus/train*.

trary, a LANDMARK as *box*, yields an example of CONTAINMENT, requiring the preposition *in*. Therefore, while *Image Schemas* are elements of conceptual structures and are formed through experience and multimodal perception, they are also coded in language through specific markers. The hypothesis underlying this work is that the selection of the locative preposition in sufficiently unambiguous sentences largely depends on the semantics of the LANDMARK, its nature and affordances. While this is perfectly acceptable, although debatable, for humans, who may rely on their abstract conceptual structures derived from perception to build certain linguistic representations, it is less obvious if and how LLMs, relying only on textual data, can achieve similar outcomes. Additionally, it is hard to test preposition selection in autoregressive LMs, which are trained through the *next-token prediction* objective, as the LANDMARK usually comes after the preposition. To cope with that, we leverage the *preposition stranding* construction (Hoffmann, 2011), occurring in sentences where the preposition is displaced with respect to its object, and appears *after* it. These constructions are usually found in *wh*-clauses (Mussemann, 2024; Hoffmann, 2011), like *which mat was the cat on?*. However, our interest is not in investigating the phenomenon of *preposition stranding per se*, but rather to leverage it to frame the preposition selection as a *next-word prediction* task. Our primary interest concerns how the meaning of a LANDMARK interacts with the preposition selection, helping to distinguish *in* from *on*, and what components allow for that in LLMs.

2. Related Works

The Circuits Framework A basic tenet of MI is that LLMs can be seen as directed computational graphs (Ferrando et al., 2024; Geiger et al., 2021; Elhage et al., 2021) implementing their behaviours through smaller, identifiable sub-graphs called *circuits* (Olah et al., 2020; Elhage et al., 2021). Recent works (Goldowsky-Dill et al., 2023; Meng et al., 2022; Conmy et al., 2023; Syed et al., 2024; Hanna et al., 2024) have developed techniques rooted in causal interventions (Pearl, 2000; Geiger et al., 2025) to manipulate model internals and discover important components forming *circuits* that are causally relevant to models' outputs. So far, *circuits* have been found for several tasks, such as Indirect Object Identification (Wang et al., 2023), syllogistic reasoning (Kim et al., 2025), greater-than reasoning (Hanna et al., 2023), colored objects identification (Merullo et al., 2023), and other factual-recall tasks (Yu et al., 2023). Ideally, for each task a model can consistently perform, there may be a *circuit* responsible for that.

MI and Linguistic Theory - Although the literature focusing on linguistically motivated *circuits* is still limited, previous works have leveraged MI concepts and techniques to address problems rooted in linguistic theory. Following the formal/functional competence distinction proposed by Mahowald et al. (2024), Hanna et al. (2025) ran MI analyses on several LLMs, looking for a separation between functional and formal circuits, adapting previous tasks to be representative of both kinds of competence. Hanna and Mueller (2025) have questioned if LLMs process sentences in an incremental, human-like way through an investigation of garden path constructions, showing that this may be the case up to a certain extent. Boguraev et al. (2025), have studied the LMs causal mechanisms underlying English filler-gaps constructions, which are supposed to share relevant similarities. Their findings reveal that LLMs partially share causal structures across different filler-gap constructions, in line with relevant linguistic theory (Sag, 2010), while spotting new factors that may contribute to refining the theory itself, showing how the interplay between linguistic theorising and mechanistic analysis can help formulate stronger hypotheses about human language. Ferrando and Costa-jussà (2024) applied several MI techniques to *subject-verb agreement* datasets to investigate the mechanisms responsible for solving this task crosslinguistically. Arora et al. (2024) developed CausalGym, a suite to benchmark interpretability methods on linguistic tasks, adapting the already existing SyntaxGym (Gauthier et al., 2020). The present work relates to these works, which aimed at linking mechanistic interpretability methods to linguistic theory.

3. Data and Model

Data - We manually built a set of triplets consisting of three nouns, respectively representing a FIGURE and two LANDMARKS, of which one is thought to enable a CONTAINMENT, while the other a SURFACE *schema*. All the nouns are entities appearing in usual real-life scenarios and prototypical spatial configurations like `<cat, kennel, mat>` or `<fork, drawer, table>`, and so on. Starting from such triplets, we built five different prompt templates to linguistically represent the aforementioned spatial configurations, which are locative in nature, as shown in Table 1. While the prompts may differ in the type of sentence they represent (interrogative, affirmative or negative), they are all constructed around the copula *to be* as it is the most basic way to express locatives (Herskovits, 2009; Zwarts, 2017). All five templates represent a *stranded preposition* construction, with the preposition at the end, always following the schema LANDMARK-FIGURE-preposition. We built a list of 157 triplets from which we derived 314

sentences (half expecting *in* and half expecting *on*), which amount to a total of 1,570 available input sentences across all templates. For experiments requiring minimal pairs (e.g., *Activation Patching*, see below), we use both clause types, expecting *in* and *on*, as both target and counterfactual inputs in turn. This allows us to augment the data and to consider both prepositions as the correct label alternately.

Model - We tested LlaMa-3.2-1b (Grattafiori et al., 2024) as it offers a good trade-off between size and performance. It is a basic decoder-only autoregressive Transformer (Vaswani et al., 2017) with 16 hidden layers, embedding size of 2,048, 32 attention heads and a vocabulary of 128,256 tokens.[4]

4. Experiments

4.1. Overview

We carried out a set of experiments to gain insight into the mechanisms behind the model's behaviour when performing the stranded preposition prediction task. Starting with a basic performance assessment to make sure the model is able to perform the task, we applied several techniques to understand some of the mechanisms responsible for the model's behaviour. First of all, we applied the Logit Lens (nostalgebraist, 2020) to perform **Direct Logit Attribution** and to have a sense of how high-level components of the model may directly contribute to the logits (Ferrando et al., 2024; Geva et al., 2022), focusing layer-wise on the residual stream, the attention and the MLP sub-layers. This way, we can have a sense of which high-level components in the model are important to perform the task at hand, as well as understand at which point of the architecture the model starts to perform it correctly. We performed an **Activation Patching** (ActP) experiment (Zhang and Nanda, 2023; Heimersheim and Nanda, 2024), again on the same three components, to see how the information flows across them along the model's architecture. This allows us to both confirm the findings of the previous experiment and to see how the components interact with token positions, to spot which token may be important for the predictions. After that, we performed a **Probing** (Vulić et al., 2020; Belinkov, 2022) experiment using the representation of different words to predict a binary class distinguishing CONTAINMENT and SURFACE schemas to consolidate hypotheses made with ActP, and to confirm which tokens gather the important information to operate such a distinction. We then turn to apply **Edge Attribution Patching with Integrated Gradients** (EAP-IG, (Hanna et al., 2024)) to scale over simple activation patching and try to isolate computational sub-graphs (i.e., *circuits*) able to retain a relevant portion of the entire model performance. We led cross-template comparisons and tried to derive a unified circuit from the intersection of the template-related circuits. The first three experiments will be showcased in the next Section 4.2, grouped together as part of an *Exploratory Analysis* mainly covering **RQ1**, while the latter ones will be shown in the Section related to *Circuit Discovery*, covering **RQ2**. The experiments are mostly implemented using TransformerLens[5] and EAP-IG[6] Python libraries.

4.2. Exploratory Analysis

Performance assessment - We assessed the model's ability to predict the stranded preposition with two metrics: i) *Average Logit Difference* (ALD) (Wang et al., 2023), consisting of the difference in logits between a correct and an incorrect answer averaged across inputs, ii) *accuracy*, counting how many times the model assigned higher logits to the correct answer than the incorrect one. The results shown in Table 1 for these complementary metrics suggest that the model performs the task with reasonable proficiency. In line with previous literature (Wang et al., 2023; Zhang and Nanda, 2023; Hanna et al., 2024), in the remainder of the paper, ALD will be the primary metric used to assess model behaviour from a mechanistic point of view. In our case the difference will always be computed between the logits for the correct preposition vs. the incorrect one. For instance, in the sentence *Which kennel was the cat* the metric would compute $logits(in) - logits(on)$.

Direct Logit Attribution (DLA) - We applied a technique called Logit Lens (LL) (nostalgebraist, 2020), to decode inner representations of the model into vocabulary space by multiplying it with the unembedding layer matrix ($\mathbf{W}_U \in \mathbb{R}^{Vd}$).[7] Thus, it is possible to gain insights into the contribution of each model's component to the formation of the logits. The LL can be performed layer-wise with different levels of granularity. A Transformer layer $l \in L$ essentially consists of two components: an attention block A^l and a Feed Forward Network FFN^l linked by the *residual stream*, that is the input sequence token embedding representation $\mathbf{x} = \langle x_1, x_2, ..., x_n \rangle$ at layer l. For each l, there are two *residual stream* locations: $\mathbf{x}^l \in \mathbb{R}$ and $\mathbf{x}^{mid,l} \in \mathbb{R}$, respectively before and after A^l. The DLA of each component/representation to the output token $w \in V$ with the logit lens is computed as:

[4] https://huggingface.co/meta-llama/Llama-3.2-1B

[5] https://transformerlensorg.github.io/TransformerLens

[6] https://github.com/hannamw/EAP-IG

[7] The notation for this Section is adapted from Ferrando et al. (2024)

Template key	Template form	ALD	Accuracy
TEMPLATES0	The LANDMARK the FIGURE was	1.72	80.42%
TEMPLATES_QUESTIONS	Do you know which LANDMARK the FIGURE was	2.03	86.01%
TEMPLATES_QUESTIONS2	Which LANDMARK was the FIGURE	0.92	66.78%
TEMPLATES_RELATIVE_NEG	I don't know which LANDMARK the FIGURE was	2.29	78.32%
TEMPLATES_RELATIVES_AFF	I saw the LANDMARK which the FIGURE was	1.71	86.01%

Table 1: Prompt templates and model performance on each of them measured with Average Logit Difference (ALD) and accuracy.

$$DLA_{f_w(\mathbf{x})\leftarrow c} = f^c(\mathbf{x}_n^l)\mathbf{W}_U \quad (1)$$

where the component c is either A or FFN and $f^c(\mathbf{x}_n^l)$ is its output representation at a given layer, and n is the last token position in the sequence. We applied the LL layer-wise to the A^l and FFN^l outputs as well as to the pre and mid representations of the *residual stream* on the last token position.

Figure 2 shows the results for the *residual stream* (left) and the A and FFN outputs (right), averaged across prompt templates. Considering the *residual stream*, it seems that the model is basically unable to perform the task until l_{11}. On the right side of the plot, it can be seen that the first notable spike in the attention line, which happens to be also the greatest, is indeed at l_{11}, while the FFN (mlp_out) line peaks at the l_{12}, somehow pairing with the left plot. Taken together, these results suggest that: i) the model starts to perform the task at l_{11}, and ii) the A^{11} output, together with FFN^{12} seems to play an important role in initiating the process of bringing the model toward the right prediction. This kind of analysis does not take into account the role played by single token representations, and it is thus impossible to understand what the model is attending to or how the information is flowing and mixing across tokens. However, since attention layers are often seen as information movers (Ferrando et al., 2024; Olsson et al., 2022; Elhage et al., 2021), one possible interpretation is that the attention sub-layer at the 11th block is the one starting to move the right information from the most informative token to the last one, which is used to finalise the prediction.

Activation Patching (ActP) (Meng et al., 2022; Heimersheim and Nanda, 2024; Zhang and Nanda, 2023) - To better understand components' contribution to the logits position-wise, we ran a causal intervention experiment. ActP is a technique consisting in replacing components' outputs $f^c(\mathbf{x})$ with the value $\tilde{\mathbf{h}}$ they would have taken if a minimally different input was given to the model, which can be expressed as $f(\mathbf{x}|(\text{do}(f^c(\mathbf{x}) = \tilde{\mathbf{h}}))$. This basically means replacing (patching) the model's activations obtained on a "corrupted" input ($\mathbf{x}$) with clean ones obtained on a clean input ($\mathbf{h}$) at each token position and each layer until the model can restore its original prediction. The idea is that if the model is able to recover the clean output at a given token position and a given component activation, then those activations are causally relevant to the production of the target output. We measure the impact of the intervention using ALD as a metric:

$$ActP_{f(\mathbf{x})\leftarrow c} = ALD(f(\mathbf{x}), f(\mathbf{x}|(do(f^c(\mathbf{x}) = \tilde{\mathbf{h}})))) \quad (2)$$

This technique assumes that the input dataset is made of minimal pairs, which in our case are sentences differing only for the nouns referring to the LANDMARK and alternating between CONTAINMENT *schemas* expecting the preposition *in*, and SURFACE *schemas* expecting *on*. For example, the sentence *The kennel the cat was* expecting *in* as a possible continuation, is patched with something like *The mat the cat was* expecting *on*, and vice versa, so as to always have the alternation of two different LANDMARKS requiring different prepositions. For space reasons, in Figure 3 we report the results of the experiment obtained from a single template on the three components we already used: *residual stream*, attention and FFN outputs. Considering the *residual stream* patching (Figure 3, left plot), we see that the information is concentrated on the LANDMARK token and stays on it until the 11th layer. Also, if we look at the attention output patching (Figure 3, central plot), we see that the slot corresponding to the last token position and the 11th layer is always particularly dark, hinting at the importance of that activation to restore the output. It is possible that the attention module at l^{11} is responsible for moving the right information from the LANDMARK toward the last token, making it usable for prediction. It seems like the information needed to perform the task resides in the LANDMARK token representation from the beginning, starting from the token embedding, but it is used only later, as already suggested by the LL experiment. The LANDMARK information seems to be moved progressively forward toward the following tokens, whose representation gets enriched by the preceding context and therefore becomes more important for the prediction. At l^{11}, something happens, and the most important token becomes the last one, where information is

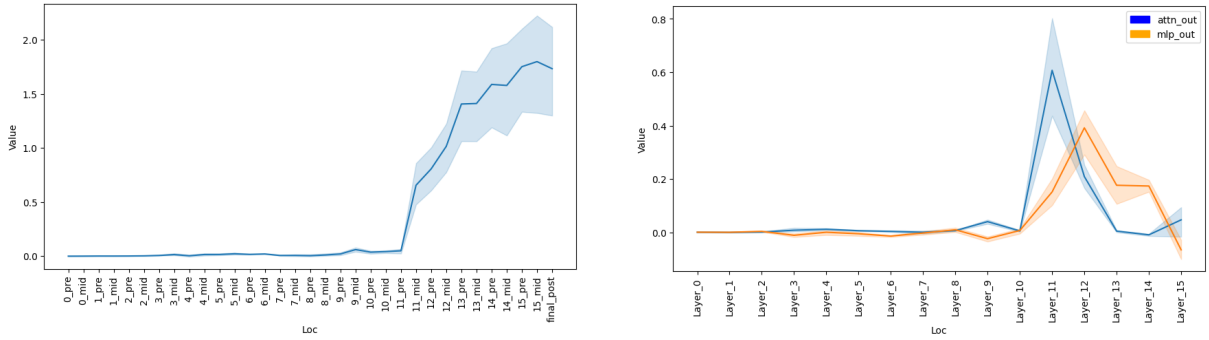

Figure 2: Logit lens applied to the residual stream (left) and to attention and MLP outputs (right) at each layer

concentrated for the prediction. Other components' activations seem to have some relevance, which may suggest that the relevant information is somehow spread across the sentence and the network, and its flow may not be neatly identifiable. However, the majority of the components activation that show high scores in this patching experiment are relative to the LANDMARK token or the last token, which is compatible with the hypothesis already pinpointed in the previous experiment: *the information necessary to predict the correct preposition resides in the LANDMARK representation since the beginning of the model architecture and it is moved to the last token (maybe by A^{11}) to make use of it.*

Probing (Köhn, 2015; Gupta et al., 2015; Belinkov, 2022) - We trained a simple binary classifier on the representations of different tokens in our sentences to predict the binary distinction between CONTAINMENT and SURFACE *schemas*. This helped us decode the semantic contents of the LANDMARK tokens and compare those representations with those of other surrounding words in the sentence. In particular, we focus on three target tokens: i) the LANDMARK token, ii) the last token and iii) a random token between the previous two. Formally, the probe maps the output of a model's layer $f^l(\mathbf{x})$ to binary labels as follows $p : f^l(\mathbf{x}) \rightarrow z$, where the two classes are defined as $\{f^l(\mathbf{x}) : \mathbf{x} \in Containment\}$ and $\{f^l(\mathbf{x}) : \mathbf{x} \in Surface\}$. To alleviate the data scarcity problem, we trained a linear classifier, namely a Logistic Regressor in the *sci-kit-learn* implementation (Pedregosa et al., 2011), with a k-fold cross-validation regime setting $k = 5$. Cross-validation has been done for each prompt template layer-wise for all three target words. We averaged the accuracy scores across folds and templates, as shown in Figure 4. It can be seen that the classifier can distinguish between the two classes since the very beginning of the model architecture (i.e., from the embedding layer) when LANDMARK tokens are used, while it is unable to make the distinction for both the random and the last token at the beginning. However, both random and last token representations seem to approach the LANDMARK one pretty quickly in the early and middle layers of the model, and the probe performances tend to converge toward the latest layers for all the analysed tokens. This may be caused again by the fact that, initially, the needed information is stored in the LANDMARK token, and it is then spread across the following tokens, as they become progressively more contextualised. This experiment shows how the two classes of LANDMARK can be separated based on the semantic content of their embeddings, along a clear distinction between CONTAINMENT and SURFACE *schemas*.

To sum up, in this Section, we have seen that the model is able to perform the task starting from the 11th layer, that two components seem particularly involved in allowing that, namely attention output at 11th layer and Feed Forward Network output at the 12th layer and that the LANDMARK token seems to contain the necessary information to operate the *in/on* distinction since the initial layers of the architecture. In the next Section, we'll turn our attention to automatically discover minimal sub-graphs that may be responsible for performing the task.

4.3. Circuit Discovery

Edge Attribution Patching with Integrated Gradients (EAP-IG, Hanna et al. (2024)) - EAP-IG is an automatic circuit discovery technique which scales over activation patching and improves over simple attribution patching (Syed et al., 2024), incorporating gradient integration (Sundararajan et al., 2017), already used to improve gradient methods for input features attribution (Shrikumar et al., 2017; Wang et al., 2024). This allows for automatically scoring the importance of edges linking components in the net, evaluating their contribution to the logits.

Edge activations are scored by cumulating the gradients on a straight line path going from a baseline input representation z' to z, which measures

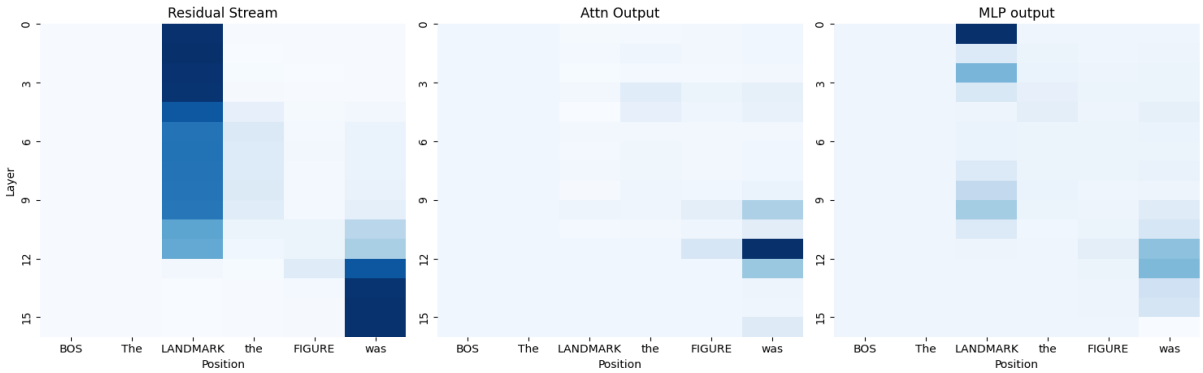

Figure 3: Activation patching results example on a single template

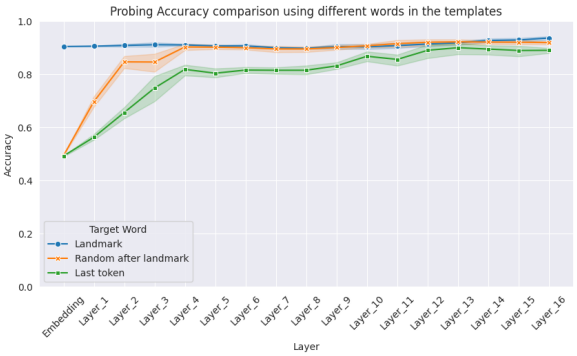

Figure 4: Accuracy of the probe averaged across folds (k=5) and prompts

the degree of change in the contribution to the logits caused by the change in the input. Gradients integration is usually replaced with a sum of m steps for computational reasons (Sundararajan et al., 2017; Hanna et al., 2024) and the full scoring equation is reported in equation 3.[8]

$$(z'_u - z_u) = \frac{1}{m}\sum_{k=1}^{m}\frac{\partial L(z' + \frac{k}{m}(z' - z))}{\partial z_v} \quad (3)$$

After edges are scored according to equation 3, we can retain the top-k ones and isolate a circuit. We applied EAP-IG to our task, running the model over all five templates independently, to isolate, for each of them, a circuit retaining 85% of the full model performance. We firstly scored the model's edges with EAP-IG, then we started from a threshold of 50k (15% of model size), best performing edges and pruned them until we reached a circuit retaining around 80-85% of the model performance. The performance of the circuits we obtained is shown in Table 2 along with the retained value with respect to the baseline (full-model) and their size, both in terms of the number of edges

[8]For complete details about this implementation we refer the reader to Hanna et al. (2024)

	Performance		Edges included	
Template	ALD	Retained	#	%
TEMPLATES0	1.47	85%	25,909	6.70%
TEMPLATES_QUESTIONS	1.73	85%	48,965	12.66%
TEMPLATES_QUESTIONS2	0.73	80%	48,972	12.66%
TEMPLATES_RELATIVE_NEG	1.96	85%	32,952	8.52%
TEMPLATES_RELATIVE_AFF	1.44	84%	48,952	12.66%

Table 2: Average Logit Difference (ALD) obtained with the identified circuits and percentage of retained performance with respect to the baseline (full-model). Each circuit has been derived from a specific template.

included and their proportion with respect to the whole model size. We then perform several cross-template comparisons, measuring the structural similarity among circuits derived by different templates and the transferability of one circuit's performance to templates different from the one used to derive it in the first place. Lastly, we tried to derive a single *core circuit* from the intersection of the single template-related circuits to have a minimal common subgraph to be tested across all templates.

4.3.1. Cross-template comparisons and "core circuit"

Following Hanna et al. (2024), we computed both the structural similarity between pairs of circuits, in terms of edges overlapping through Intersection Over Union (IoU) and Edge Recall (ER) as shown in equation 4 and Cross-Template Faithfulness (CT-Faith) in terms of the performance of a circuit derived from a certain template on a different template:

$$ER(C_1, C_2) = \frac{C_1 \cap C_2}{C_2}$$
$$IoU(C_1, C_2) = \frac{C_1 \cap C_2}{C_1 \cup C_2} \quad (4)$$

The CT-Faith metric shows how much better or worse a circuit derived from a template X performs

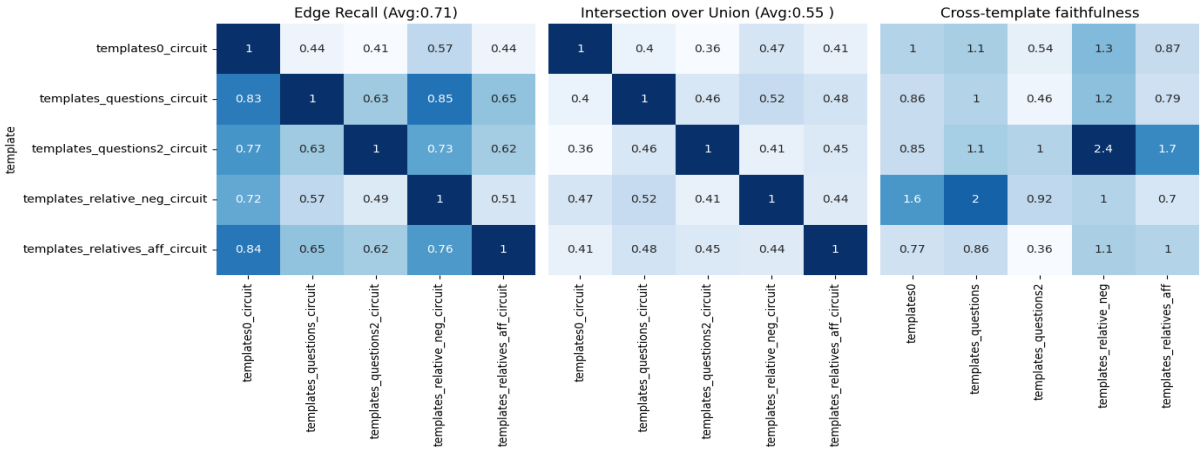

Figure 5: Comparisons among circuits with EDGE RECALL (left), INTERSECTION OVER UNION (center) and cross-template faithfulness (right)

on a template Y with respect to the circuit originally derived from template Y, and its values have to be read as percentages. Therefore, 1 stands for exact matching of the performance between the two circuits, while lower- and greater-than-one values are to be considered respectively as worsening or increasing in performance. Ideally, if circuits capture the task nature with a high degree of generalisation, different prompt template formalisations of the same underlying task should yield equal or highly similar and similarly performing circuits. Figure 5 shows the results for all three used metrics. It can be seen that, on average, IoU is about 0.55 while ER gets higher scores, with an average of 0.77. Although both these metrics are problematic when measuring overlap among circuits of different sizes, 3 out of 5 of the compared circuits have the same size (12.66%) as shown in Table 2, which should mitigate the problem to some extent. Taken together, these three metrics suggest that the similarity among circuits derived from the different templates is not striking and that, while some circuits tested on templates different from the one used to derive them actually improved the performance over the original circuit, most of them underperform.

Lastly, we derived a circuit from the intersection of all the circuits obtained from the different templates. We call it the *core circuit* as it represents the *core* edges common across all template-derived circuits. We then tested the performance of this circuit on each of the different templates to compare the performances of the full model, the template-derived circuits and the core circuit. The size of this derived circuit is only 3.6% of the entire model but it is able to retain substantial portions of the entire model performance, ranging between 59% and 70% depending on the template, as shown in Table 3.

Template	ALD	Retained
TEMPLATES0	1.2	70%
TEMPLATES_QUESTIONS	1.4	69%
TEMPLATES_QUESTIONS2	0.65	70%
TEMPLATES_RELATIVE_NEG	1.6	70%
TEMPLATES_RELATIVE_AFF	1	59%
Core-Circuit size	3.6%	

Table 3: Average Logit Difference (ALD) obtained by the core circuit on each template and percentage of retained performance with respect to baseline. The circuit is derived from the intersection of all template-related circuits.

5. Discussion

We have investigated whether LlaMA-3.2-1b encodes important information about the distinction between CONTAINMENT and SURFACE *Image Schemas* governing the selection of locative prepositions. To do so, we have looked at models' representations and components' behaviour through a series of experiments. With **Direct Logit Attribution**, we saw that the model starts to perform the task around layer 11 and that the output of two components seems to be particularly important: attention and Feed Forward Network outputs at layer 11 and 12, respectively. At the same time, **Probing** showed that the LANDMARK token seems to contain enough information to distinguish the two *schemas* early on in the architecture, but this information is somehow taken from that token and gathered into the last one for prediction starting from layer 11th, as shown by **Activation Patching**. Through **EAP-IG**, we have been able to isolate circuits related to single templates as well as a minimal core circuit derived from the intersection of the previous one. We saw that the overlap, both in terms of edges and performance, across the different circuits is partial and performance transferability variable. However, a circuit of only 3.6% the size

of the full model can retain a substantial portion of the baseline performance (59-70%), and may be considered the backbone mechanism for the task at the end. These findings suggest that, while it is possible to discover small circuits responsible for the behaviour under analysis, formal changes in the prompt can make this unstable. An important implication of that is that a certain degree of variability in the template may be taken into account when performing circuit discovery, as a task may be expressed through different linguistic formulas, which may lead to variable results. This may also suggest that the task is not encoded at a higher abstract semantic level transcending its different prompt realisations, but it is rather influenced by the actual linguistic form of the templates. All in all, it seems that the model distinguishes between CONTAINMENT and SURFACE *schemas* and selects the locative preposition accordingly, and that some specific components are responsible for that, consistently across different prompt templates, while there is variability concerning the complete circuits employed to solve the task when presented in different forms. However, precisely interpreting the role of the components and the circuit is not easy and may be only speculative at this stage, requiring further work and analysis. In our case, applying MI techniques to a problem devised on the basis of theoretical assumptions rooted in Cognitive Linguistics has proven useful to formulate hypotheses regarding the mechanisms responsible for performing the given task and thus better understanding how specific linguistic knowledge may be encoded inside the model's internals. Specifically, these techniques offer a powerful toolbox to investigate linguistically relevant problems in LLMs, providing a newer way to evaluate the linguistic competence these models encode. Therefore, generalising this approach to other phenomena may broaden the evaluation strategies at our disposal and help gain deeper insights about the relationship between models' mechanisms and theoretical linguistics predictions.

6. Conclusions

In this work, we introduced a linguistic task, namely the *Stranded Locative Preposition Prediction*, to frame the locative preposition selection as a *next-token prediction* and investigate it with MI techniques. We have shown that the model can solve the task, and we located some of the high-level components that may be responsible for the behaviour through a series of experiments. We have seen that both template-related circuits and a minimal *core* circuit can be derived, and their performance assessed against the baseline obtained by the full model. We plan future improvements of the task by expanding the set of prepositions and constructions to test, including, for example, PATH *schemas* like SOURCE and GOAL, while leading more granular analysis on the located model's components.

7. Limitations

This work has a number of important limitations. We only experimented with a narrow phenomenon and, in the future, more *image schemas* and preposition alternations may be included. Several of the tested techniques, such as Activation Patching and EAP-IG, assume important constraints on the input data as building a dataset of carefully aligned minimal pairs. This makes it difficult to manually craft datasets to test, and therefore, we could experiment with only a relatively small dataset. Other techniques, like *value zeroing* (Mohebbi et al., 2023) for context mixing or *information flow route* (Ferrando and Voita, 2024) for circuit discovery, exist that may be complementary to the ones we used and may be useful to develop an even more complete picture of the mechanisms behind the task at hand.

Lastly, we only tested one LLM, which limits the generalization of the findings across models, as it is not granted that models use akin mechanisms to encode the same information and perform the task.

8. Bibliographical References

Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14638–14663, Bangkok, Thailand. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*.

Sasha Boguraev, Christopher Potts, and Kyle Mahowald. 2025. Causal interventions reveal

shared structure across english filler-gap constructions. *arXiv preprint arXiv:2505.16002*.

Timothy C Clausner and William Croft. 1997. Productivity and schematicity in metaphors. *Cognitive science*, 21(3):247–282.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352.

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. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Javier Ferrando and Marta R. Costa-jussà. 2024. On the similarity of circuits across languages: a case study on the subject-verb agreement task. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10115–10125, Miami, Florida, USA. Association for Computational Linguistics.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R Costa-Jussà. 2024. A primer on the inner workings of transformer-based language models. *arXiv preprint arXiv:2405.00208*.

Javier Ferrando and Elena Voita. 2024. Information flow routes: Automatically interpreting language models at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17432–17445, Miami, Florida, USA. Association for Computational Linguistics.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, 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.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*.

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

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. 2015. Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 12–21, Lisbon, Portugal. Association for Computational Linguistics.

Michael Hanna, Yonatan Belinkov, and Sandro Pezzelle. 2025. Are formal and functional linguistic mechanisms dissociated in language models? *Computational Linguistics*, pages 1–40.

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.

Michael Hanna and Aaron Mueller. 2025. Incremental sentence processing mechanisms in autoregressive transformer language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3181–3203.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. 2024. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

Annette Herskovits. 1985. Semantics and pragmatics of locative expressions. *Cognitive Science*, 9(3):341–378.

Annette Herskovits. 2009. *Language and spatial cognition: an interdisciplinary study of the prepositions in English*, paperback re-issue edition. Studies in natural language processing. Cambridge Univ. Press, Cambridge.

Thomas Hoffmann. 2011. *Preposition Placement in English: A Usage-based Approach*, 1 edition. Cambridge University Press.

Mark Johnson. 1987. The body in the mind: The bodily basis of meaning, imagination, and reason. *The Personalist Forum*, 5(1):58–60.

Roni Katzir. 2023. Why large language models are poor theories of human linguistic cognition: A reply to piantadosi. *Biolinguistics*, 17:1–12.

Geonhee Kim, Marco Valentino, and Andre Freitas. 2025. Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10074–10095, Vienna, Austria. Association for Computational Linguistics.

Arne Köhn. 2015. What's in an embedding? analyzing word embeddings through multilingual evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Lisbon, Portugal. Association for Computational Linguistics.

George Lakoff. 1987. *Women, fire, and dangerous things: what categories reveal about the mind*, paperback ed., [nachdr.] edition. The Univ. of Chicago Press, Chicago.

Alessandro Lenci. 2023. Understanding natural language understanding systems. *Sistemi intelligenti*, 35(2):277–302.

Stephen C. Levinson. 2003. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Language Culture and Cognition. Cambridge University Press.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in cognitive sciences*, 28(6):517–540.

Jean M. Mandler. 1994. Precursors of linguistic knowledge. *Philosophical Transactions: Biological Sciences*, 346(1315):63–69.

Jean M. Mandler and Cristóbal Pagán Cánovas. 2014. On defining image schemas. *Language and Cognition*, 6(4):510–532.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2023. Circuit component reuse across tasks in transformer language models. In *The Twelfth International Conference on Learning Representations*.

Raphaël Millière. 2024. Language models as models of language. *arXiv preprint arXiv:2408.07144*.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400, Dubrovnik, Croatia. Association for Computational Linguistics.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, et al. 2024. The quest for the right mediator: A history, survey, and theoretical grounding of causal interpretability. *arXiv preprint arXiv:2408.01416*.

Victoria Mussemann. 2024. A topic which i want to know more about – preposition placement in finite wh-relative clauses in world englishes. *English Language and Linguistics*, 28(2):341–370.

nostalgebraist. 2020. interpreting gpt: the logit lens. *LessWrong*.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*. Https://distill.pub/2020/circuits/zoom-in.

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*.

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, 381(2251):20220041.

Judea Pearl. 2000. *Models, reasoning and inference*. Cambridge University Press, Cambridge.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Steven Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*.

Steven T Piantadosi. 2023. Modern language models refute chomsky's approach to language. *From fieldwork to linguistic theory: A tribute to Dan Everett*, 15:353–414.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*.

Ivan A. Sag. 2010. English filler-gap constructions. *Language*, 86(3):486–545.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, et al. 2025. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR.

Shriyank Somvanshi, Md Monzurul Islam, Amir Rafe, Anannya Ghosh Tusti, Arka Chakraborty, Anika Baitullah, Tausif Islam Chowdhury, Nawaf Alnawmasi, Anandi Dutta, and Subasish Das. 2025. Bridging the black box: A survey on mechanistic interpretability in ai. *Available at SSRN 5345552*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Aaquib Syed, Can Rager, and Arthur Conmy. 2024. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416, Miami, Florida, US. Association for Computational Linguistics.

Leonard Talmy. 2000. *Toward a cognitive semantics: Concept structuring systems*, volume 1. MIT press.

Andrea Tyler and Vyvyan Evans. 2003. *The Semantics of English Prepositions: Spatial Scenes, Embodied Meaning, and Cognition*. Cambridge University Press.

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.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

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*.

Yongjie Wang, Tong Zhang, Xu Guo, and Zhiqi Shen. 2024. Gradient based feature attribution in explainable ai: A technical review. *arXiv preprint arXiv:2403.10415*.

Qinan Yu, Jack Merullo, 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.

Fred Zhang and Neel Nanda. 2023. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*.

Joost Zwarts. 2017. Spatial semantics: Modeling the meaning of prepositions. *Language and Linguistics Compass*, 11(5):e12241.