Short Paper

Are formal and functional linguistic mechanisms dissociated in language models?

Michael Hanna[1*], Yonatan Belinkov[2], Sandro Pezzelle[1]

[1] Institute for Logic, Language, and Computation
University of Amsterdam
m.w.hanna@uva.nl, s.pezzelle@uva.nl

[2] Technion – Israel Institute of Technology
belinkov@technion.ac.il

Although large language models (LLMs) are increasingly capable, these capabilities are unevenly distributed: they excel at formal linguistic tasks, such as producing fluent, grammatical text, but struggle more with functional linguistic tasks like reasoning and consistent fact retrieval. Inspired by neuroscience, recent work suggests that to succeed on both formal and functional linguistic tasks, LLMs should use different mechanisms for each; such localization could either be built-in or emerge spontaneously through training. In this paper, we ask: do current models, with fast-improving functional linguistic abilities, exhibit distinct localization of formal and functional linguistic mechanisms? We answer this by finding and comparing the "circuits", or minimal computational subgraphs, responsible for various formal and functional tasks. Comparing 5 LLMs across 10 distinct tasks, we find that while there is indeed little overlap between circuits for formal and functional tasks, there is also little overlap between formal linguistic tasks, as exists in the human brain. Thus, a single formal linguistic network, unified and distinct from functional task circuits, remains elusive. However, in terms of cross-task faithfulness—the ability of one circuit to solve another's task—we observe a separation between formal and functional mechanisms, suggesting that shared mechanisms between formal tasks may exist.

1. Introduction

A wide body of research has argued that language and thought are dissociated in the human brain (Fedorenko, Ivanova, and Regev 2024). That is, such research argues that the regions of the brain that respond differentially to well-formed linguistic input are distinct from those that respond to other structured inputs, such as mathematics, music, and code (Amalric and Dehaene 2019; Chen et al. 2023; Ivanova et al. 2020). They are also distinct from regions that respond to language-adjacent capabilities such as theory of mind and reasoning (Shain et al. 2022; Monti, Parsons, and Osherson 2009).

* Corresponding author

Action editor: {action editor name}. Submission received: DD Month YYYY; revised version received: DD Month YYYY; accepted for publication: DD Month YYYY.

These regions of the brain, termed the *language network*, are thus selective for language (perhaps narrowly defined), and language alone.

In recent work, Mahowald et al. (2024) argue that these two types of stimuli, to which the language network does and does not respond, correspond to two distinct types of linguistic competence: *formal* and *functional*. Formal linguistic competence is necessary for "getting the form of language right". It involves the correct structuring of language at the sub-word, lexical, and sentence level; in other words, phonology, morphology, syntax, and lexical semantics are all domains of formal linguistic competence. Past neuroscientific work has shown all of these to be supported by the language network (Regev et al. 2024; Shain et al. 2024; Hu et al. 2022).

In contrast, functional linguistic competence involves abilities that allow speakers to achieve goals in the world, but may involve non-linguistic cognition. For example, speakers may engage in formal or pragmatic reasoning, or employ world knowledge in conversation, without such abilities being intrinsic to language. Similarly, speakers may use situation modeling skills to make sense of a narrative, or engage in theory of mind reasoning to understand their interlocutors' point of view, although language use need not entail the use of these abilities. Moreover, exercising these abilities does not engage the brain's language network.

Mahowald et al. furthermore claim that this distinction between formal and functional linguistic competence is reflected in the performance of today's large language models (LLMs). In particular, LLMs have strong formal linguistic competence as evidenced by their strong performance on syntax benchmarks and their general ability to output fluent and natural text (Hu et al. 2020; Gauthier et al. 2020; Warstadt et al. 2020). However, their functional linguistic competence is markedly worse: LLMs frequently output false reasoning, hallucinate untrue facts, and fail at complex social reasoning (Dziri et al. 2023; Xu, Jain, and Kankanhalli 2024; Strachan et al. 2024).

Much work has attempted to ameliorate these problems via retrieval-augmented generation or the use of chain-of-thought reasoning (Gao et al. 2024; Wei et al. 2022). However, Mahowald et al. offer another solution: perhaps LLMs would have stronger functional linguistic abilities if formal and functional linguistic abilities were as distinct in LLMs as they are in the human brain. Such a dissociation in LLMs could come about in two ways: either it could be explicitly built into them, or it could arise naturally via training, or the model's inductive biases. Today's LLMs have no explicit formal-functional modularity, but whether any has arisen despite that fact is an open question.

In this paper, we seek to answer that question: to what extent are formal and functional linguistic competence dissociated in the internals of today's LLMs? If the two are not dissociated, new architectures or training procedures that bias models towards a formal-functional dissociation may be necessary to achieve this. If the two are already dissociated (despite the fact that LLMs struggle more with functional linguistic competence), this might indicate that dissociation does not suffice to improve LLMs' abilities. This question of dissociation is also relevant due to the increase in work that uses LLMs to explicitly model language in the human brain, oftentimes by predicting activations within the brain using those from LLMs (Tuckute, Kanwisher, and Fedorenko 2024; Sucholutsky et al. 2024). The presence or absence of a dissociated language network in LLMs could help us judge whether such comparisons are licensed. Rather than comparing activations within models and brains, though, we propose to characterize mechanisms within models, and verify whether they are organized in the same way as the human brain.

To investigate whether formal and functional competence are distinctly localized within the LLMs, we draw on techniques from LLM interpretability, and in particular, *mechanistic interpretability*, which studies the mechanisms that underlie LLMs' behavior using low-level causal methods (Ferrando et al. 2024). Concretely, we study the formal-functional distinction using **circuits**: a circuit is a small path through the LLM that contains all of the mechanisms underlying its behavior on a task of interest (Olah et al. 2020; Wang et al. 2023). Circuits provide causal guarantees that the localization found is correct, as all parts of the model outside of the circuit can be ablated without changing model behavior; see Section 3 for more details. This targeted, causal evidence for the correctness of our localization is a notable advantage of our framework, as such evidence is hard to come by in human brains. In humans, one must rely on either natural experiments (e.g., individuals with brain damage) or coarse-grained experiments using transcranial magnetic stimulation; optogenetics allows for finer-grained causal manipulation of neurons, but only in non-human subjects.

We thus translate Mahowald et al.'s hypothesis about emergent dissociation between formal and functional linguistic abilities in LLMs into the LLM circuit analysis framework. First, we identified 5 tasks involving formal linguistic competence, and 5 tasks involving functional linguistic competence. We next selected 5 LLMs, and found the circuits responsible for their behavior in these tasks. We then measured the similarity between each pair of task circuits, focusing in particular on similarities both within and across the formal and functional task groups; see Figure 1 for an overview.

But what does it mean for the two task circuits to be similar or dissociated? One way to operationalize this is to measure the overlap between formal and functional circuits, in terms of the components (and connections between them). If formal and functional networks are dissociated, we expect that formal and functional circuits should have low overlap, while formal circuits should have high overlap with one another. Our findings suggest formal and functional language competence are not dissociated in LLMs when this is measured via circuit overlap. Circuits for formal tasks have small but non-zero overlap with circuits for functional tasks. More importantly, circuits for formal tasks do not have an especially high overlap with one another.

We also measure similarity dissociation via the ability of one circuit to perform another's task, or *cross-task faithfulness*, as past work has found measuring overlap and cross-task faithfulness to yield different results (Hanna, Pezzelle, and Belinkov 2024). In this setting, one circuit is similar to another if it can perform the other's task well. Thus, if formal and functional competence are dissociated in LLMs, we would expect formal task circuits to perform formal tasks well, but perform functional tasks poorly (and vice-versa).Under this metric, a formal-functional divide appears more plausible: formal task circuits indeed perform other formal tasks better than they perform functional ones; moreover, functional task circuits perform other functional tasks better than they perform formal ones. We conclude that although a unified formal network in terms of overlap does not exist, formal language circuits are indeed more similar to each other in terms of the tasks that they can perform.

In summary, we find evidence for emerging formal-functional dissociation in LLMs, though only in terms of cross-task faithfulness. That is, formal task circuits perform other formal tasks better than they perform functional ones; functional tasks similarly struggle to perform formal tasks. However, our overlap studies suggest against the existence of one unified region for all language tasks. In performing these analyses, we apply mechanistic interpretability techniques to a question from neuroscience in LLMs for the first time. We moreover conduct a study of circuits across LLMs, employing both the greatest number and largest size LLMs of such a study to date. We release the code for our experiments, as well as the efficient circuit-finding tools within them, at `https://github.com/hannamw/formal-functional-dissociation`.

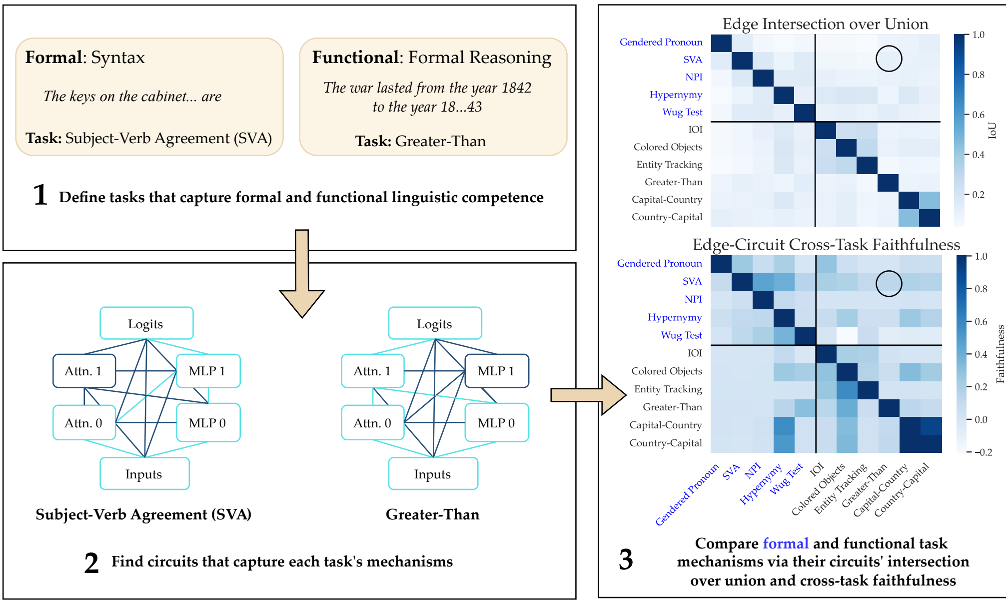

Figure 1
Our experimental pipeline. First, we define tasks that capture various aspects of formal or functional linguistic linguistic competence. Then, we find circuits for these tasks, which describe the model mechanisms responsible for them. Finally, we compare those circuits, measuring the similarity between formal and functional task mechanisms (as measured by e.g. their circuits' intersection over union or cross-task faithfulness).

2. Background

2.1 The Language Network in the Human Brain

The language network, as described by Fedorenko, Ivanova, and Regev (2024), has three main characteristics: (1) it responds in an undifferentiated fashion to various types of language use; (2) it is consistent across modalities and languages; (3) and it is robustly dissociated from both low- and high-level networks with non-language roles.

That is, the language network responds differentially to language as opposed to non-language stimuli, but it responds equally to syntax as it does to (lexical) semantics; the two do not activate distinct regions of the brain (Fedorenko et al. 2020; Shain et al. 2024). Moreover, the language network activates on both linguistic input and output (Menenti et al. 2011). It activates whether the language is heard or read, and is similar across languages (Regev et al. 2013; Malik-Moraleda et al. 2022). However, it does not overlap with perceptual or motor regions, and also excludes regions for higher-level non-linguistic competence, such as cognitive control and theory of mind. (Li, Hiersche, and Saygin 2024; Pritchett et al. 2018; Quillen, Yen, and Wilson 2021).

The bulk of these findings come from experiments using a relatively simple setup used to localize brain regions responsible for a given competence. In such studies, brain activations (typically measured using functional magnetic resonance imaging; (fMRI) are measured in response to two contrasting stimuli (or while performing two contrasting tasks), where only one is relevant to the phenomenon of interest. Areas that differentially respond to the phenomenon of interest are inferred to be involved

Competence	Subdomain	Description	Example
Formal Linguistic Competence — Needed to get the form of language right	**Phonology**	Constraints on wordforms	*The new species was called a* ***drick*** / ***dmick****.*
	Morphology	(Novel) morpheme ordering constraints	***TikTok-ifi-cation*** / ***TikTok-ation-ify***
	Syntax	Agreement, word order, etc.	*The keys on the cabinet* ***are*** / ***is*** *mine.*
	Lexical Semantics	Parts of speech, word meanings, etc.	*Would you like soup or* ***salad*** / ***soft****?*
Functional Linguistic Competence — Needed to achieve goals in the world	**Formal Reasoning**	Mathematics, logic, planning, etc.	*Two birds joined the three on the branch. Now there are* ***five*** / ***four*** *birds on the branch.*
	Social Reasoning	Pragmatics, theory of mind, etc.	*Bob took a bite and said, "I've had better." Bob thought the food was* ***bad*** / ***good****.*
	Situation Modeling	Discourse coherence, narrative structure, etc.	*Ann doesn't own a dog. She loves her* ***cat*** / ***dog****.*
	World Knowledge	Facts, concepts, common sense, etc.	*The capital of France is* ***Paris*** / ***Rome****.*

Table 1
Subdomains of formal and functional linguistic competence. Each subdomain includes a sentence that is **correct** or **incorrect** with respect to it, depending on the word chosen. Figure adapted from Mahowald et al. (2024)

in its processing. For example, Fedorenko et al. (2010) localize the language network by finding regions that activate on well-formed sentences as opposed to lists of non-words.

We note that the aforementioned views on the language network are somewhat controversial. For example, Murphy and Woolnough (2024) criticize this body of work for its (over-)reliance on fMRI selectivity evidence, while others question the idea of separating formal and functional language use at all (Forkel and Hagoort 2024). In this paper, we make no claims about the formal language network in the brain. Rather, we seek to translate the Mahowald et al.'s claim into LLMs, and test whether it holds.

2.2 Formal and Functional Linguistic Competence in Language Models

Building on the literature regarding the language network in the brain, Mahowald et al. (2024) propose a related distinction between *formal* and *functional* linguistic competence. They use *formal linguistic competence* to refer to linguistic abilities necessary to get the *form* of language right; *formal linguistic competence* refers to language abilities that we use to achieve goals or otherwise *function* with language. Notably, while formal linguistic competence cleanly maps onto the language network, functional linguistic competence includes only some non-language-network abilities; others, like music, are excluded.

Mahowald et al. divide formal and functional linguistic competence into subdomains (Table 1). Formal language competence includes subdomains like **phonology**, **morphology**, **syntax**, and **lexical semantics**. For Mahowald et al., phonology corresponds to the rules governing valid wordforms (i.e., phonotactics), while morphology involves the correct ordering of morphemes. Syntax involves not only correct word order, but also higher-level abilities like agreement (e.g. between subjects and verbs).

Lexical semantics entails using words correctly according to their part of speech, lexical category, or meaning. Mahowald et al. distinguish this category from semantics more broadly: general conceptual knowledge belongs to functional language competence.

In contrast, functional language competence consists of **formal reasoning**, **world knowledge**, **situation modeling**, and **social reasoning**. Formal reasoning includes math and logical abilities, while world knowledge includes facts and commonsense knowledge. Situation modeling entails the ability to track the state of a discourse, and the structure of narratives. Finally, social reasoning covers pragmatics and theory of mind.

Mahowald et al. note that LLMs have strong formal linguistic competence, succeeding on tests of syntactic ability and lexical semantics (Chang and Bergen 2024). It is more challenging to measure English LLMs' abilities in phonology, given that LLMs seldom produce novel phonemes, and morphology, given the relative simplicity of English morphology, but LLMs do seem to generate valid novel morphemes (McCoy et al. 2023). In contrast, models often struggle with functional tasks (Dziri et al. 2023; Strachan et al. 2024), though recent LLMs have markedly improved reasoning abilities due to intensive post-training. Still, there remains a clear gap between the relative ease of learning formal linguistic competence, and the ongoing challenge of functional linguistic competence.

The solution, according to Mahowald et al., is to induce modularity in LLMs, just as it exists in the human brain. Such modularity could take two forms: *architectural modularity*, which is explicitly built into a model's architecture, and *emergent modularity*, which occurs naturally due to e.g. the model's inductive bias and training process. They note, as do we, that transformers are well-suited for this sort of emergent modularity: past work has found LLM attention heads (imperfectly) dedicated to certain syntactic relations (Vig and Belinkov 2019; Clark et al. 2019), as well as heads that add as induction, succession, and copy suppression modules (Olsson et al. 2022; Gould et al. 2024; McDougall et al. 2024). How, though, can we localize such modules if they exist?

2.3 Causal Localization in LLMs

Localizing the regions of a model that perform a given task or ability is a key question in the interpretability of NLP models. Modern interpretability work often uses causal interventions (Pearl 2009) to do so. The core idea behind these is that a unit—e.g., a parameter in a weight matrix, or neuron in a model activation—is important if perturbing (or *intervening on*) it causes the relevant model behavior to change. For example, if setting a neuron's activation to zero in a model causes the model to be unable to recall a country's capital, we conclude that the neuron played a role in the model's capital-recall ability. Causal interventions thus allow us to infer the function of units within a model.

A wide body of work performs localization in the parameter space of models. As zeroing-out parameters one by one is prohibitively expensive, it is common to learn a binary mask over parameters, indicating which parameters are important (Han et al. 2015; Frankle and Carbin 2019; Prasanna, Rogers, and Rumshisky 2020). Unimportant parameters are set to zero; such masks are learned to maximize both sparsity and model performance under this regime. These techniques were initially developed to increase model efficiency, but have since been used to locate modules within models (Csordás, van Steenkiste, and Schmidhuber 2021), find language-specific and knowledge-critical subnetworks (Lin et al. 2021; Choenni, Garrette, and Shutova 2023; Bayazit et al. 2024).

Other work instead localizes mechanisms in activation space, looking for neurons, components (such as transformer models' attention heads or multi-layer perceptrons), or entire layers that are important to task abilities. Such work has measured whole layers' importance by perturbing them with Gaussian noise (Meng et al. 2022), zeroed out entire attention head activations (Voita et al. 2019; Olsson et al. 2022), or computed linear approximations of ablation effects (Nanda 2023).

However, one must be cautious when performing causal interventions. For one, performing the right intervention is essential. While zero ablations are common and intuitive, they are harmful because model activations (and parameters) are seldom zero; zeroing them out may bring the activations out of distribution, causing harm (Hase, Xie, and Bansal 2021; Chan et al. 2022). Thus, if we zero ablate a given unit and observe a drop in model performance, we cannot determine if this stems from the unit's importance, or the out-of-distribution issue (Li and Janson 2024). Mean ablations, which replace activations with their mean across a dataset, are less harmful, but can still fall out-of-distribution. Activation patching, which intervenes on a model by replacing a component's activation on one example, with an activation on another example, avoids this issue (Vig et al. 2020; Geiger et al. 2021); to our knowledge, no similar technique has been developed for parameter localization.

Moreover, causal interventions of this sort tell us only which units are *necessary*; they do not prove that the localized units are *sufficient* to perform the task of interest, or that we have captured all relevant units. This issue and the issue with zero ablations have cast some doubt on older causal localization studies. Fortunately, recent work has developed a framework for the causal localization of mechanisms in transformer models that avoids many of these problems: circuits.

3. Circuits

We characterize the mechanisms behind LLMs' formal and functional linguistic competence using circuits (Olah et al. 2020; Elhage et al. 2021). Circuits are small subgraphs of a model—generally no more than 5% thereof—that capture how it performs a given task. Crucially, circuits are both necessary and sufficient for models to perform tasks; they destroy task performance when ablated, and suffice to perform the task when everything outside them is ablated. That is, circuits aim to capture entire task mechanisms. We propose to compare the similarity of task mechanisms by comparing their circuits.

3.1 Definitions

A model's **circuit** for a given task is the minimal computational subgraph of the model that is faithful to its behavior on the task (Wang et al. 2023; Hanna, Liu, and Variengien 2023). That is, even if all parts of the model outside of the circuit are ablated (or instead corrupted), model behavior will not change. See Hanna, Pezzelle, and Belinkov (2024) or Miller, Chughtai, and Saunders (2024) for reviews of circuits literature.

Computational Graph. In this study, we focus on LLMs using the transformer architecture (Vaswani et al. 2017). Such a model's computational graph describes the computations it performs, and can be viewed as a directed graph that begins at the model's inputs, flows through its components, i.e. attention heads and multi-layer perceptrons (MLPs), and ends at its logits (Conmy et al. 2023; Hanna, Liu, and Variengien 2023). In most autoregressive transformers, each component's input is directly added to its output via a residual connection; as a result, each component's input is the sum of the outputs of previous components. Thus, in our computational graph, every component has an edge to components that come later in the model, and the input to a given node v is the sum of the output of all nodes u with an edge to v. A circuit should identify those edges that

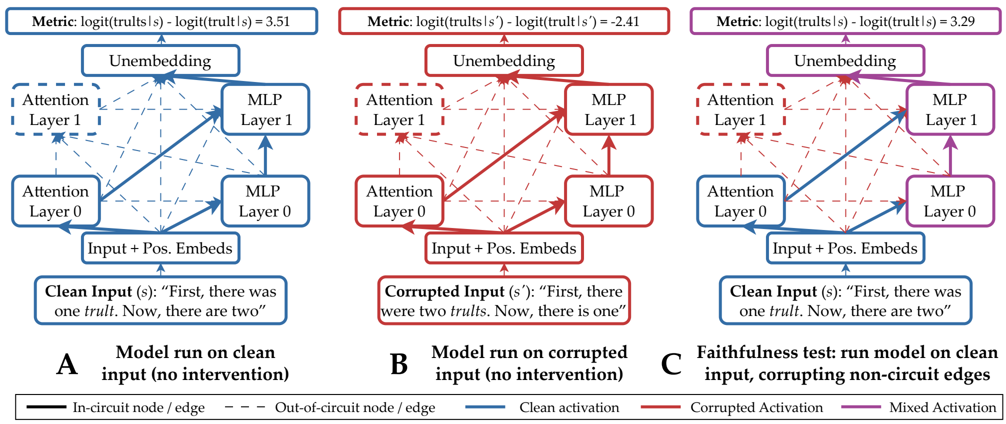

Figure 2
A toy circuit in a 2-layer transformer, where nodes are attention and MLP layers. **A**: We first run our whole model on the clean inputs to establish baseline behavior. **B**: We then run the model on a corrupted input that elicits very different behavior, and save the corrupted activations. **C**: To test a circuit's faithfulness, we run the model on clean inputs, replacing non-circuit node / edge activations with corrupted activations; behavior should stay the same as in (A).

are important for the model's ability to perform the task at hand. See Figure 2 for a toy example of a circuit in a 2-layer transformer model's computational graph.

Task. A *task* in circuits analysis consists of clean and corrupted inputs, expected outputs, and a metric to measure task performance. For example, a subject-verb agreement task (like that of Linzen, Dupoux, and Goldberg 2016; Newman et al. 2021) might consist of clean inputs like $s =$ *The keys on the cabinet*; the expected output would be a plural-conjugated verb, like *are*. The corrupted outputs are drawn from the same task distribution, but elicit very different behavior; here, the corrupted input might be $s' =$*The key on the cabinet*, which elicits a singular-conjugated verb like *is*.

The metric could then be the difference in probabilities assigned to plural and singular verbs, given s; i.e., $m = p(\text{plural-verb}|s) - p(\text{singular-verb}|s)$. Note that the metric should be continuous, both for use with circuit-finding methods (see Section 3.2) and to allow for observation of incremental changes in model behavior as changes are made to the circuit. Tasks should be solvable by the model in which the circuit is to be found; if the model cannot solve the task, there may not be any task behavior to localize.

Faithfulness. A circuit is *faithful* if the model's task behavior stays the same, even when all nodes and edges outside the circuit are ablated. At a high level, this entails running our model on clean inputs s, but replacing the activations of the nodes and edges outside the circuit, with activations taken from corrupted inputs s'. As the vast majority of the model is outside the circuit, the model should behave as if it were being run on s'—unless we have correctly localized all task-relevant nodes and edges.

More formally, and following Hanna, Pezzelle, and Belinkov (2024), we test model faithfulness by running the model on s and performing the following causal intervention. Let v be a node in the whole model's computational graph $G = (V, E)$. Let z_u denote the activation of a given node u during the current forward pass, and z'_u denote its activation on corrupted inputs s'. Let $C = (V_C, V_E)$ denote our circuit. For each noninput node v, we set its input to

$$v_{in} = \sum_{(u,v)\in E_C} z_u + \sum_{(u,v)\in E\setminus E_C} z'_u. \tag{1}$$

If all edges into v are in the circuit, its input is $\sum_{(u,v)\in E} z_u$, the same as its input without interventions; if none are, its input is the same as it is when the model is run on s'.

Having done this, we measure m_C, the circuit's performance. Denote by $m_\emptyset$ the model's performance when entirely corrupted (i.e., when it is run on the corrupted input). We can then compute a measure of normalized faithfulness (Marks et al. 2024):

$$F = \frac{m_C - m_\emptyset}{m - m_\emptyset}. \tag{2}$$

We aim to attain a faithfulness of 1: $F < 1$ implies that we have missed task-relevant edges, while $F > 1$ suggests we have missed edges that work *against* the model's task abilities (but may nonetheless be task-relevant). In practice, circuit faithfulness trades off with size; it starts at 0, and quickly grows towards 1 as the circuit's size grows, but only reaches 1 after many less-important edges are added. As a result, it is common to study circuits with faithfulness near but below 1, which omit less important components.

3.2 Finding Circuits

A naive approach to finding circuits is to compute how much the model's performance decreases when each edge is corrupted; this is the edge's indirect effect (IE; Pearl 2001). To find a circuit of size n, one could simply take the n edges with the highest |IE|. However, as testing an edge requires one forward pass, this approach takes $O(|E|)$ forward passes, and both $|E|$ and the cost of a forward pass increase with model size.

We thus opt to linearly approximate each edge's IE via edge attribution patching (EAP; Syed, Rager, and Conmy 2024). Given an edge (u, v), EAP estimates its IE as:

$$\hat{\text{IE}}_{u,v} = (z'_u - z_u)^\top \nabla_{v_{in}} m(s), \tag{3}$$

where $(z'_u - z_u)$ indicates the change in u's activation upon corruption, and $\nabla_{v_{in}} m(s)$ is the change in metric when v's input changes. Computing the activations z'_u and z_u of all nodes u requires two forward passes, while computing $\nabla_{v_{in}} m(s)$ for all nodes v requires one backward pass. We can thus compute $\hat{\text{IE}}$ for all edges in constant time.

As EAP is often inaccurate in practice, we instead use Hanna, Pezzelle, and Belinkov's (2024) EAP with integrated gradients (EAP-IG), which improves upon EAP by computing $\nabla_{v_{in}} m$ at intermediate points between s and s'. Having thus computed each edge's IE, we find the circuit by taking the top-n edges by $|\hat{\text{IE}}|$. We take edges by the absolute value of $\hat{\text{IE}}$ to find a circuit that is a *complete* explanation of the model's task behavior, containing all relevant nodes and edges, even if they harm performance.

Recent work in circuit-finding instead learns a mask indicating which model nodes and edges are in the circuit; the mask is trained to optimize both sparsity and model performance when the out-of-circuit units are corrupted (Chintam et al. 2023; Bhaskar et al. 2024; Li and Janson 2024). Though we prefer techniques that provide IE estimates for each edge, mask-based approaches could be an interesting avenue for future work.

3.3 Why use circuits to localize formal and functional linguistic competence?

We argue for the use of circuits to localize formal and functional linguistic competence because they capture mechanisms that are both necessary *and sufficient* for models to perform formal and functional linguistic tasks. Many techniques can identify parameters or neurons that harm model performance when ablated; however, such approaches may miss units that are causally relevant to the model's behavior on formal and functional tasks, as one can harm performance without ablating all relevant units.

These issues have played an important role in prior work attempting to localize a language network in LLMs. Zhang et al. (2024), for example, localize a "core linguistic region" consisting of model parameters that harm model abilities cross-lingually when set to zero. However, such zero ablations can cause harm unrelated to the importance of the units ablated (see Section 2.3); moreover, they do not check that this zero-ablation harms language selectively. Thus, we cannot be sure whether this network is crucial for LLMs' language abilities, or their abilities in general.

AlKhamissi et al. (2024) engage in a more neuroscientifically grounded study, identifying "language network neurons" using a localizer task: AlKhamissi et al. select those neurons whose activations on non-word lists and on sentences differ most significantly. They find that these neurons' activations better predict brain data than random neurons do; moreover, zero-ablating these neurons harms model performance. Based on this, the authors argue that these neurons constitute a language network.

We argue that this inference, too, is flawed. As before, zero-ablation is a destructive technique, prone to harming model performance by throwing its activations out of distribution. However, this issue is exacerbated by the activation-difference localization method. It is known that certain outlier neurons have magnitudes up to 20x larger than others in the same layer (Timkey and van Schijndel 2021; Dettmers et al. 2022; Ahmadian et al. 2023); such neurons are prone to be found by activation difference methods due to their high magnitudes in general. Moreover, even just quantizing these neurons (i.e. imprecisely recording their values, not zeroing them) is disastrous for model performance (Lin et al. 2024). These results can thus be explained by the detection and ablation of outlier neurons, rather than a language network.

In contrast to the methods of these past studies, and others that do not engage in causal analysis at all (Kisako, Kuribayashi, and Sasano 2025), circuits provide causal guarantees about the correctness of the desired localization. They not only use more principled ablation methods, but also aim to be necessary, sufficient, and minimal. This allows us to compare the localizations found via circuits without worrying that these include unnecessary components, or miss necessary ones.

4. Tasks and Data

In our main experiments, we consider 10 tasks that give us broad coverage over most of the subdomains of formal and functional linguistic competence described by Mahowald et al.. We exclude some such categories (like phonology) as they are impossible to test in text-based LMs. For others (social reasoning), LM abilities are generally poor, while circuits are best found for tasks models perform well; LM are quite competent on all tasks we study here (see App. A for details). See Table 2 for an overview of tasks. While we introduce some tasks that are new to circuit analysis, most have previously been studied in the circuits literature. We note, however, that because prior work studied these tasks in smaller or older models, it is difficult to make direct comparisons between the circuits we find, and the circuits prior work found.

	Task	Category	Input and [Expected Output]
Formal	Subject-Verb Agreement (SVA)	Syntax	The keys on the cabinet [are]
	Gendered Pronoun Agreement	Syntax	Maria said that [she]
	Negative Polarity Items (NPIs)	Syntax	The customer that the managers liked has [never]
	Hypernymy	Lexical Semantics	Roses are a type of [flower]
	Wug Test	Morphology	First, there was one *trult*. Now there are two [*trults*]
Functional	Indirect Object Identification (IOI)	Situation Modeling	Alice and Bob went to the store. Then Alice gave a bottle of water to [Bob]
	Entity Tracking	Situation Modeling	The apple is in Box F, the computer is in Box Q,. . . , Box F contains the [apple]
	Colored Objects	Situation Modeling	On the table, I see an orange textbook, a red puzzle, and a purple cup. What color is the textbook?
	Greater-Than	Formal Reasoning	The war lasted from the year 1842 to the year 18[43, 44, . . . , 99]
	Country-Capital	World Knowledge	France, whose capital, [Paris]

Table 2
Tasks under study. The top five tasks are formal, while the bottom five are functional.

4.1 Formal Tasks

Subject-Verb Agreement (SVA) is a *Syntactic* task that gives models inputs like $s =$ "The keys to the cabinet", and expects verbs that agree with the subject *keys*. Its corrupted variant inverts the plurality of the subject; here *keys* would change to *key*. We measure task performance as the probability assigned to verbs that agree with the subject, minus that assigned to verbs that do not. For this task, we adapted Newman et al.'s (2021) SVA data.

Gendered-Pronoun Agreement is a *Syntactic* task that gives models inputs involving explicitly gendered entities, like $s =$ "The heroine went home because". The expected output is the corresponding gendered pronoun, *she*. Its corrupted variant replaces the subject with the corresponding opposite-gender noun (here, *hero*); task performance is measured as $logit(\text{she}) - logit(\text{he})$. Vig et al. (2020) identified neurons causally responsible for models' ability to perform this task in a gender-biased scenario; we adapt their data, adding explicitly gendered entities.

Negative Polarity Item (NPI) usage is a *Syntactic* task that gives models inputs that do or do not license NPIs, like *ever* or *any*. For example, the input $s =$ "The customer that the managers liked has" could be continued by "never", but not by the NPI "ever". In each corrupted variant, we add an NPI-licensing word to s, or remove it if it already exists; in our example, s' could be "*No* customer that the managers liked has", which licenses the use of the NPI *ever*. Task performance is measured as $logit(\text{never}) - logit(\text{ever})$. We adapted this task's data from the corresponding SyntaxGym task (Gauthier et al. 2020).

Hypernymy is a *Lexical Semantic* task that gives models inputs like $s =$ "Roses are a type of", and expects outputs like *flower*. Its corrupted variant replaces the hyponym (*roses*) with that of another type (e.g. *diamonds*); the resulting metric is $p(\text{flower}|s) - p(\text{gem}|s)$. We use the task data from Hanna, Pezzelle, and Belinkov (2024).

The **Wug Test** is a *Morphological/Syntactic* task that tests models' abilities to generate singular and plural forms of nonce words (Berko 1958). It gives models inputs containing nonce words like $s =$ "First, there was one *trult*. Now there are two", and expects outputs like *trults*. Its corrupted variant reverses the number of entities present, as in $s' =$ "First, there were two *trults*. Now there is one". Task performance is measured as $p(\text{trults}) - p(\text{trult})$. We generate new nonce words using Wuggy (Keuleers and Brysbaert 2010) to avoid LMs having previous exposure. The Wug test requires morphological abilities (namely, to generate the plural or singular of a nonce word) but also requires syntactic abilities (to understand that the nonce word and preceding numeral must agree in number).

4.2 Functional Tasks

Indirect Object Identification (IOI) is a *Situation Modeling / World Knowledge* task providing inputs like "When Mary and John went to the store, John gave a bottle of milk to" and expecting the output "Mary". Its corrupted version replaces the second instance of *John* with an unrelated name like *Bob*. Task performance is measured via the difference in the logit assigned to the correct vs. incorrect entity, i.e., $logit(\text{Mary}) - logit(\text{John})$. This task, common throughout the circuits literature, requires models to recognize that, if an individual has an object, they cannot give it to themselves; this can be solved via situation modeling or potentially world knowledge. This task's data is adapted from Wang et al. (2023), who introduced it and studied its circuit in GPT-2 small (Radford et al. 2019).

Entity Tracking is a *Situation Modeling* task that gives models inputs like $s =$ "The apple is in Box F, the computer is in…the document is in Box Q. Box F contains the" and expects outputs like "apple" (Kim and Schuster 2023). In the corrupted version, the queried object (e.g. *computer in Box Q*) is different. Task performance is measured via $logit(\text{apple}|s) - logit(\text{computer}|s)$. In this task, LLMs must track entities and their state over the length of a discourse. Prakash et al. (2024) studied this task in Llama-7B (Touvron et al. 2023).

Colored Objects is a *Situation Modeling* task that gives models inputs like $s =$ "On the table, I see an orange textbook, a red puzzle, and a purple cup. What color is the textbook?"; the expected output is "orange". In the corrupted version, object colors and the queried object (e.g. "blue mug") are different. Task performance is measured via $logit(\text{orange}|s) - logit(\text{blue}|s)$. Merullo, Eickhoff, and Pavlick (2024) first studied this task in GPT-2 medium (Radford et al. 2019), showing that its circuit uses many of the same mechanisms as IOI does; however we adapt this task's data from the original

BigBench task (Srivastava et al. 2023).

Greater-Than is a *Formal Reasoning* task with inputs like $s =$"The war lasted from the year 1842 to the year 18"; we expect outputs greater than 42. Performance is measured via $\sum_{y>\texttt{YY}} p(y|s) - \sum_{y\leq\texttt{YY}} p(y|s)$. The corrupted input for this task replaces the start year `YY` (in s, `YY`="42") with "01", shifting the model's output distribution towards years < `YY`. This task's data is adapted from Hanna, Liu, and Variengien (2023), who introduced it, and studied its circuit in GPT-2 small (Radford et al. 2019).

Country-Capital is a *World Knowledge* task that gives models inputs like $s =$ "France, whose capital, " and expects outputs like "Paris". Its corrupted variant replaces *France* with another country like *Italy*; task performance is then measured via $logit(\text{Paris}|s) - logit(\text{Rome}|s)$. We also include the reverse task (**Capital-Country**), which queries the country of which a city is the capital; past work has shown that this has a similar circuit to the Country-Capital task (Hanna, Pezzelle, and Belinkov 2024).

5. Experimental Pipeline

Our experimental pipeline works as follows. For a given model, we find the circuit for each task in Section 4, using 500 examples per task. Then, for each pair of task circuits, we measure their similarity; if a formal-functional dissociation is to exist, formal circuits should be dissimilar from functional ones but similar to one another. Here, we discuss our choice of models, circuit-finding methods, and metrics to measure circuit similarity.

Models. We study state-of-the-art models from five families: Llama-3 8B (Llama Team 2024), Gemma-2 2B (Gemma Team 2024), Qwen-2 7B (Yang et al. 2024), Mistral-v0.3 7B (Jiang et al. 2023), and OLMo 7B (Groeneveld et al. 2024). We choose these models, lying in the 2-8 billion parameter range, because they are the largest models for which current circuit-finding techniques can function, due to both memory and compute constraints. Moreover, these models are capable enough to perform some situation modeling and world knowledge tasks, on which smaller models generally fail. Note that these are base models that underwent no instruction tuning, as this type of model was analyzed by Mahowald et al. and is commonly studied in the circuits literature. We speculate that results for instruction-tuned models would be similar, as recent work has shown that learned features are similar across base and instruction-tuned models (Kissane et al. 2024), and that fine-tuning mostly enhances existing circuits (Prakash et al. 2024).

Circuit Finding. For each task, we estimate the IE of each edge in the model's computational graph using EAP-IG, as described in Section 3.2. Each node in the computational graph represents an attention head, an MLP, and the model's inputs or logits, and edges indicate causal links between nodes. Given these IEs, we can then construct a circuit by taking the top-n edges with the highest absolute IE. For each task, we search for the minimum n such that the top-n circuit has a faithfulness of at least 85%, i.e., we find the minimum n such that the circuit recovers at least 85% of the full model's task performance. In this way, we find a circuit that explains almost all of the model's performance on the task, without including a long tail of low-IE nodes and edges.

Because faithfulness trades off with size, and we find circuits of a fixed faithfulness, each task's circuit may be of a different size. That is, some tasks may only rely on small circuits, while others may require the inclusion of much larger mechanisms in order to achieve 85% faithfulness. This can make ascertaining the similarity of two circuits rather challenging, as not all similarity metrics treat different-sized circuits the same.

Metrics. Given two tasks T_1, T_2 with circuits (subgraphs) C_1, C_2, how can we compute their similarity? Past work (Csordás, van Steenkiste, and Schmidhuber 2021) has used metrics such as intersection over union (**IoU**) and **recall** (with respect to C_1):

$$\text{IoU}(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}, \qquad \text{recall}(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_2|}. \tag{4}$$

Note that IoU and recall behave differently when circuits are of different sizes. If $|C_1| >> |C_2|$, IoU penalizes this heavily, as it is capped at $|C_2|/|C_1|$. While recall(C_2, C_1) is also capped at $|C_2|/|C_1|$, recall(C_1, C_2) can be as high as 1; indeed, large circuits may naturally recall more edges from smaller ones.

In contrast to these, Hanna, Pezzelle, and Belinkov (2024) suggest measuring **cross-task faithfulness**, by running the circuit C_1 on task T_2, and vice versa. A circuit that captures many of the mechanisms required for another task should have high performance on it. In the following experiments, we use all three of these metrics; however, the appropriate metric in a given scenario depends on the hypothesis being tested.

We do note that many other potential graph-based similarity metrics exist. For example, given two task circuits, and corresponding IEs for each edge, we can compute their weighted graph edit distance between the two by summing the difference in scores assigned to each edge by each task; note that as IEs have different ranges per task, these must be normalized first. Similarly, we can concatenate the edge scores of each task into a vector, and apply metrics such as cosine similarity to these. These metrics could be useful for future work; however, pilot experiments using them yielded very similar scores for all circuits, likely due to the poor IE estimates provided by EAP and EAP-IG, which are known to capture the ordering of edges better than they capture the edges' actual IEs (Syed, Rager, and Conmy 2024; Hanna, Pezzelle, and Belinkov 2024).

6. Do formal and functional networks overlap in LLMs?

In this section, we investigate the similarity of circuits for formal and functional linguistic tasks based on the degree to which the circuits' edges overlap. In Section 6.1, we conduct an analysis of circuit overlap based on intersection over union. Then, we investigate whether functional task circuits may contain formal task circuits (Section 6.2). Finally, in Section 6.3 we characterize the edges that are shared between circuits.

6.1 Main experiment

From a circuit overlap perspective, if formal linguistic competence forms a consistent network in LLMs, distinct from that of functional linguistic abilities, there should be no overlap between formal and functional task circuits, and high overlap between pairs of formal task circuits. We measure overlap using IoU, as it is a symmetric (non-directional) metric; here, we care about whether there is any overlap, not the direction in which the overlap occurs.

In Figure 3, we report the IoU between each pair of task circuits, averaged across models. We also report the average percentage of the model's edges included in each circuit, as the circuits vary in size: the SVA circuit contains only 0.71% of model edges, on average, while the Colored Objects task contains over 4x more, at 3.00%.

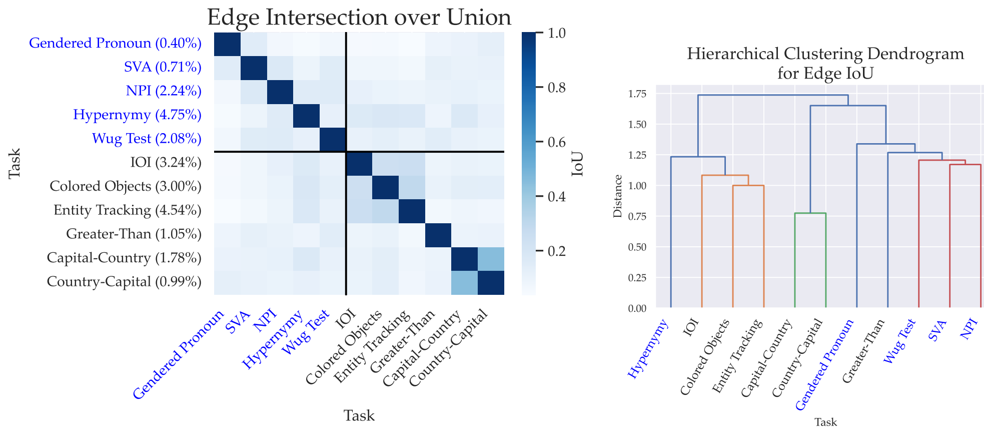

Figure 3
Left: Edge intersection over union (IoU) between tasks, averaged across models; as IoU is symmetric, this heatmap is symmetric as well. The average size of each task's circuit (as a percent of the entire model) is given in parentheses after each task's name. The IoU between most task pairs is low but non-zero. Moreover, formal tasks do not exhibit a higher level of overlap with one another than they do with functional tasks. Lines divide formal and functional tasks. **Right**: Dendrogram obtained via hierarchical (agglomerative) clustering of task IoU vectors using Euclidean distance. Tasks whose IoU vectors are more similar to one another are linked at lower levels of the dendrogram. Overall, clusters do not reflect the formal-functional divide.

The results indicate that in general, the IoU between any two pairs of circuits is low, but non-zero; the median IoU between tasks is 0.11. That is, while there are no striking examples of any formal and functional tasks having a high IoU, they are also not completely disjoint. Furthermore, the overlap between formal circuits is not especially high: while the median IoU between two distinct formal circuits (0.15) is higher than the median IoU between two distinct functional circuits (0.11) and between two formal and functional circuits (0.11), this difference is small.

Whether these IoUs are (statistically) significant depends on how we model a random, baseline circuit. Past work has considered random circuits with n edges constructed by selecting them uniformly at random (Hanna, Pezzelle, and Belinkov 2024; Shi et al. 2024). In this case, we can model the probability of two such circuits having an overlap of a given size using a hypergeometric distribution (App. B), and near all of the overlaps in Figure 3 are significantly higher than chance.

Few tasks have even moderate overlap. Capital-Country and Country-Capital have high IoU (0.43), likely because they are nearly the same task; they share the same structure (fact-retrieval) and domain (geography). Similarly, IOI, Colored Objects, and Entity Tracking all have moderately high overlap (an average IoU of 0.27), perhaps because they all involve situation modeling; Merullo, Eickhoff, and Pavlick (2024) also found that the IOI and Colored Objects circuits overlap.

We can verify these groupings found by visual inspection using clustering. We use agglomerative clustering to find which groups of tasks are closest to one another, as measured by the Euclidean distance of their IoU vectors (i.e. the rows or columns of Figure 3, left). We then create a dendrogram using the results of this clustering. The resulting dendrogram (Figure 3, right) shows that the Capital-Country / Country-Capital grouping and the IOI / Colored Objects / Entity Tracking groupings are also found by the clustering algorithm. It also finds a purely formal cluster that is less visible on the heatmap: Wug Test / SVA / NPI. However, the broader clusters do not reflect a formal-functional split. The formal Wug Test / SVA / NPI cluster next merges with Greater-Than (functional), and the functional IOI / Colored Object / Entity-Tracking cluster merges with Hypernymy (formal).

The results of this experiment suggest that **there is low but non-zero overlap between circuits for formal and functional linguistic abilities**. At the same time, the **formal tasks do not have especially high overlap with one another**. Taken together, these two facts weigh against a potential formal-functional distinction. However, there remain open questions: could the overlap between formal and function task circuits stem from the fact that both categories of tasks are expressed in language? Moreover, what is the nature of the overlap between the formal and functional tasks? If that low overlap corresponds to a task-agnostic shared processing mechanism, akin to low-level visual systems in the brain, the formal-functional hypothesis would be more plausible.

6.2 Do functional task circuits contain formal task circuits?

Tasks involving functional linguistic competence can sometimes be presented to humans and localized in the brain without the use of language; Ivanova et al. (2021), for example, use an image-based localizer to find brain regions responsible for event semantics. However, this is less true for LLMs: all of our functional tasks are presented via linguistic input.[1] This could cause the functional mechanisms we localize to also include formal regions, leading to a misleading appearance of overlap.

Recall Analysis. One way to demonstrate that functional tasks do not contain formal tasks is to perform the same analysis as above but measure recall, a directional measure of how much a given circuit contains another one. The results of this analysis (Figure 4, left) show that functional tasks indeed do not subsume formal ones. While some formal tasks (Gendered Pronoun and SVA) with small circuits are well captured by most other task circuits, other, formal tasks like Hypernymy and NPI, are not so well captured. Moreover, Hypernymy has high coverage of many functional tasks, while the reverse is not true. In general, trends in recall seem dominated by the size of the circuits. Circuits that are large (like Hypernymy, IOI, Colored Objects, and Entity Tracking) have high recall with respect to other circuits, while being poorly covered by other circuits. Formal task circuits are not always smaller than functional task circuits: there are large formal task circuits (Hypernymy) and small functional task circuits (Country-Capital).

We also perform a clustering analysis on the reference recall vectors, using the same settings as in Section 6.1; that is, the clustering should find groups of tasks that are well recalled by similar tasks. This analysis (Figure 4, right) yields similarly negative results. Formal tasks cluster together, but imperfectly: the formal cluster includes Greater-Than, as well as Capital-Country and Country-Capital, more loosely.

Non-Language Mediated Tasks. We can also investigate whether the small formal-functional overlap we observe is due to our posing functional tasks in language, by analyzing two tasks that are purely functional, without any language involved. The first is Nikankin et al.'s (2024) basic **Math** task, which consists of simple arithmetic problems

1 Though note that this is consistent with Mahowald et al. (2024), who discuss many functional linguistic tasks presented via language to LLMs.

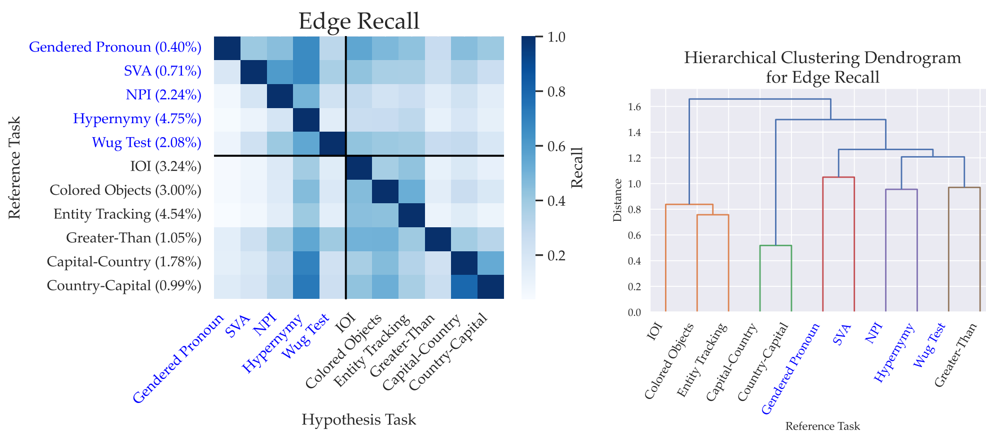

Figure 4
Left: Edge recall between tasks, averaged across models. Each square indicates how well the task circuit on the bottom (the hypothesis) captures the task circuit on the left (the reference). The average size of each task's circuit (as a percent of the entire model) is given in parentheses after each task's name. While functional task circuits may have higher recall of formal task circuits than formal task circuits do of functional ones, functional task circuits do not contain formal task circuits as a whole. Lines divide formal and functional tasks. **Right**: Hierarchical clustering dendrogram for recall vectors. Formal and functional tasks cluster somewhat separately from one another, but Greater-Than is mis-clustered.

involving addition, subtraction, and multiplication, such as *25 + 3 =*. Corrupted examples share the same operator, but different operands. We measure performance via the difference in logit assigned to the clean and corrupted equations' answers. Presented in purely symbolic form, this task should engage purely functional (formal reasoning) areas of the model, if such exist. This task is highly sensitive to tokenization,[2] and cannot easily be ported to new models, so we study it in Llama-3 (8B) alone.

The second is a string manipulation task called **Echo**, inspired by a task of the same name from Hupkes et al. (2021). In this task, models see 4 distinct tokens randomly sampled from the model's tokenizer, and must repeat the last token; a successfully solved example looks like "t_1 t_2 t_3 t_4 : t_4". We give the model two solved examples (2-shot), and then have it finish an incomplete example. In corrupted examples, the final token (t_4) is replaced with another, distinct token. We measure performance via the logit assigned to the clean and corrupted examples' final token. Since the tokens are randomly sampled, there should be no trace of linguistic structure; this task is most akin to a pattern-recognition formal reasoning task.

We find the circuits for these two tasks, and compare them to our existing task circuits using IoU, as done previously. Our results on these two tasks (Figure 5) are similar to those of other functional tasks. Neither Math nor Echo is particularly similar to any other task. Notably, although one might hypothesize that they are supported by the same mechanisms, Math and Greater-Than do not have especially high overlap. Echo also overlaps slightly more with Colored Objects, IOI, Entity Tracking, and Hypernymy

2 LLMs differ widely in how they tokenize numbers; while Llama-3 tokenizes numbers up to 3 digits in length as one token, other models (e.g. Gemma-2) tokenize them digit-by-digit, and others tokenize them in groups of 2. Single-token prediction tasks are most compatible with circuit analysis, so we use Llama-3.

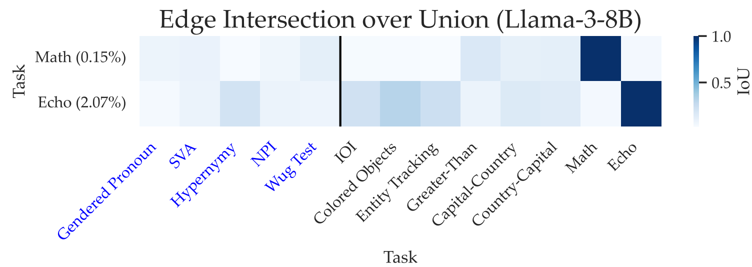

Figure 5
IoU heatmap for the Math and Echo tasks. Neither task has especially high overlap with any other task, although Echo has somewhat higher IoU with Colored Objects and Hypernymy.

than with other tasks. In general, IoU between Math / Echo and formal tasks is low but non-zero (0.09 on average). We take this as evidence that the overlap between formal and functional linguistic tasks is not due to the latter containing the former.

6.3 What components overlap between formal and functional tasks?

Our previous experiments suggest that there is formal-functional overlap, even when considering functional tasks containing no linguistic structure. But what could this overlap consist of? In order to capture entire model mechanisms for a given task, circuits must be whole paths from models' inputs to logits (Section 3). This means that, if there exist low-level task-agnostic mechanisms shared between formal and functional circuits, these might constitute a formal-functional overlap. For example, prior work suggests that transformer LMs' early layers help detokenize and contextualize words (Ghandeharioun et al. 2024). These could act as a shared low-level input processing stream, after which distinct formal or functional modules take charge.

Such low-level cross-task mechanisms also exist in the brain: brain regions responsible for vision, for example, must be active whether reading sentences or non-words. However, the localizer approach used to find the language network (Section 2.1) naturally avoids this issue by looking at the *difference* in activations in brain regions in each condition. Reading-relevant visual brain regions should activate equally in the sentence and non-word conditions, so they are not captured by the language localizer.

We check for the existence of such a task-agnostic region by looking at the intersection of all circuits, i.e. those components and edges that are relevant across every single task that we study. We then record the types of edges in this intersection circuit—what sorts of components (the inputs, attention heads, MLPs, or the logits) do they connect, and which layers are connected? We report an average across all models.

We find that the components and edges that are relevant across circuits are dominated by MLPs: all MLPs are included in every circuit. Our edge-level analysis (Figure 6, left) shows that the vast majority of edges shared across circuits involve MLPs. This is despite the fact that attention heads are far more numerous within all models. However, Figure 6 (right) shows that the components and edges shared between circuits, including the MLPs involved, span all layers. So, insofar as a shared low-level processing area should be located at early layers of the model only (or indeed layer-wise localized at all), this hypothesis is false. Rather, this shared network seems to be composed of an "MLP backbone", running up and down the model, which is essential for its functioning.

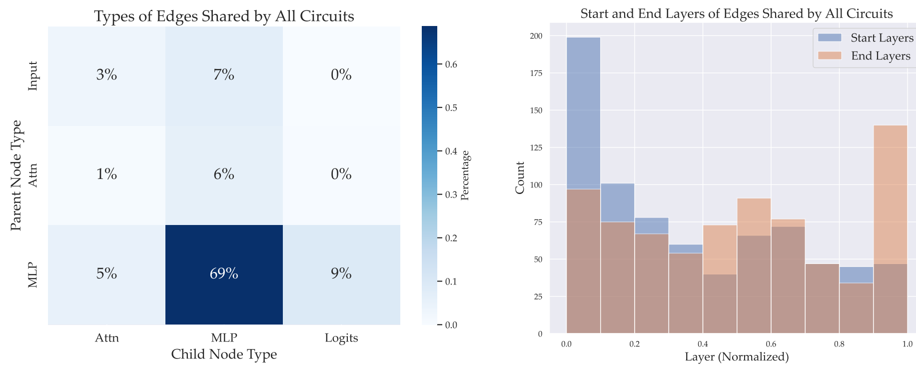

Figure 6
Left: Heatmap displaying the type of components (input, attention head, MLP, or logits node) connected by edges in the intersection circuit, averaged across models. Almost all edges connect one MLP to another. **Right**: Histogram of the start and end layer of edges in the intersection circuit, averaged across models. Because models have different numbers of layers, the reported layer is normalized (i.e. the actual layer of an edge's parent or child is divided by the model's total number of layers). Both start and end layers span the whole depth of the model, indicating that shared nodes and edges are not restricted to low-level (early-layer) processing mechanisms.

Unfortunately, the role of MLPs in general is rather contested; while they have been implicated in detokenization, they have also been conceived of as key-value memories for fact storage (Geva et al. 2021), and implicated in various task-specific roles (Hanna, Liu, and Variengien 2023; Lieberum et al. 2023; Nikankin et al. 2024).

Finally, when we exclude this shared network from our overlap analysis, the size of each circuit is sharply reduced: circuits' sizes fall to half of their previous size, or less. However, some overlap remains: the median IoU shrinks, but only from 0.11 to 0.10. Moreover, as we have excluded the same shared network from every circuit, the trends in IoU and recall, regarding which circuits overlap the most, remain the same. Overall, we can say that this shared MLP backbone is what is most drives overlap, though it does not explain all of it.

7. Are Formal and Functional Task Circuits Cross-Task Faithful?

Our prior experiments provide weak evidence for a formal-functional distinction in terms of overlap. Circuits for formal and functional linguistic competence seldom overlap in today's LLMs, and what little overlap there is, is dominated by an MLP backbone. However, there does not appear to be an undifferentiated area responsible for all aspects of formal linguistic competence. In this section, we study whether these partially negative results also hold when measuring cross-task faithfulness, i.e. how well one task's circuit suffices to perform another task. If the formal-functional distinction holds, formal tasks should be able to perform other formal tasks well; moreover, they should not be able to perform functional tasks well, or vice-versa.

Examining the pair-wise cross-task faithfulness of formal and functional tasks (Figure 7, left), different trends emerge compared to earlier. While Capital-Country and Country-Capital are still similar, the IOI / Colored Objects / Entity Tracking grouping is much less clear. Moreover, while cross-task faithfulness is somewhat influenced by circuit size—the tasks whose circuits best perform other tasks (i.e., Hypernymy, Col-

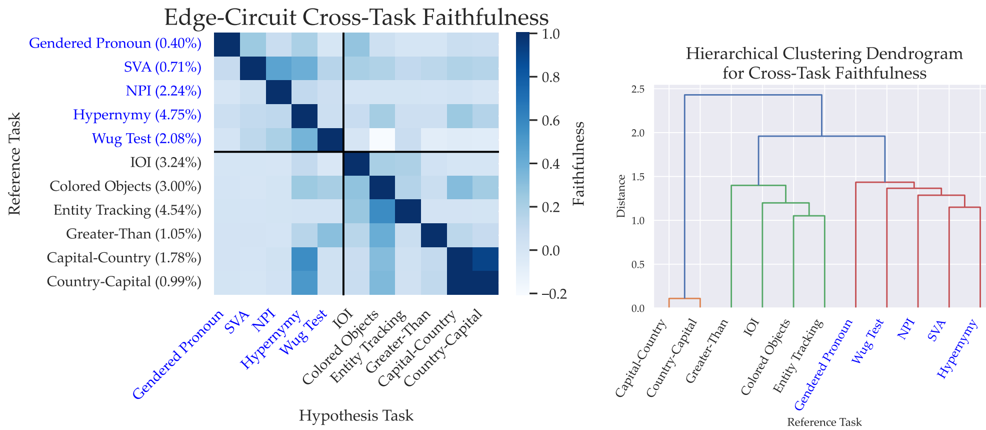

Figure 7
Left: Cross-task faithfulness between task pairs, averaged across models. Lines divide formal and functional tasks. The average size of each task's circuit (as a percent of the entire model) is given in parentheses after each task's name. **Right**: Hierarchical clustering dendrogram for cross-task faithfulness vectors. Formal and functional tasks cluster separately from one another.

ored Objects, whose columns are dark) are large—other tasks with large circuits, like Entity Tracking do not perform other tasks as well In general, the median cross-task faithfulness between two distinct formal tasks (0.11), or two distinct functional tasks (0.14) is higher than the median in formal-functional or functional-formal conditions (0.01 and 0.05), but this difference is still not great.

However, the results of our clustering analysis (Figure 7, right) provide more positive evidence. In this analysis, we cluster the cross-task similarity reference vectors, i.e. the vectors that show for a given task, which task's circuits solve it best. Surprisingly, all formal tasks form one cluster that is separate all functional tasks: Gendered Pronoun, Wug Test, NPI, SVA, and Hypernymy all cluster separately from Greater-Than, IOI, Colored Objects, and Entity Tracking. We note that Capital-Country and Country-Capital form a two-task cluster that is separate from both formal and functional tasks, but this is compatible with formal-functional dissociation and formal-formal consistency. Functional tasks need not cluster together; they need only be separate from formal tasks. Overall, these results suggest that formal tasks are more alike one another with respect to which circuits solve them. So, while there is no consistent formal region at the IoU overlap level, formal task circuits are indeed more similar to each other than functional task circuits are to them at the cross-task faithfulness level. These are promising results, but to ensure they are robust, we repeat our circuit analyses at different levels of granularity in the following section.

8. Node and Neuron-Level Circuits

While our analysis centers on circuits in a computational graph composed of nodes (attention heads and MLPs) and edges, other granularities of analysis are possible. Past work has performed causal analyses of models that focus on nodes (ignoring the edges between them) or even individual neurons (Vig et al. 2020; Finlayson et al. 2021). Recent work has studied *sparse feature circuits* composed of features from sparse autoencoders,

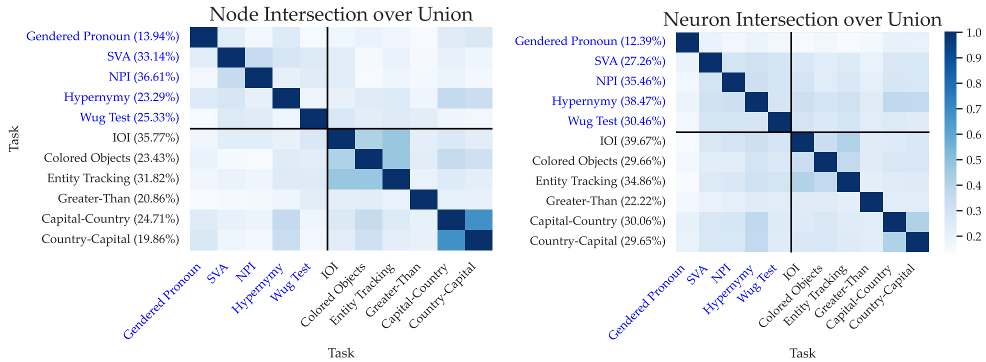

Figure 8
Left: Node-level IoU between task pairs, averaged across models. The average size of each task's circuit (as a percent of the entire model) is given in parentheses after each task's name. As with our edge-level experiments, most task pairs have low but non-zero IoU, and formal tasks exhibit no higher level of overlap with one another than they do with functional tasks. Lines divide formal and functional tasks. **Right**: Neuron-level IoU between task pairs, averaged across models. IoUs are generally quite low between all tasks.

yet more fine-grained than neurons (Marks et al. 2024). This raises a question: might we obtain different results with a computational graph of a different granularity?

To answer this, we perform our analyses again, but at the node and neuron level. We adapt EAP-IG, which produces estimates of each edge's indirect effect, to produce estimates of each node or neuron's indirect effect. We do so by replacing $\nabla_{v_{in}}$ in Equation (3) with $\nabla_{u_{out}}$, yielding (scalar) node IE estimates:

$$\hat{\mathrm{IE}}_u = (z'_u - z_u)^\top \nabla_{u_{out}} m(s). \tag{5}$$

If we replace the dot product above with element-wise multiplication, we obtain a vector $\hat{\mathbf{IE}}_u \in \mathbb{R}^{d_{model}}$ of neuron IE estimates, where each entry in the vector gives the corresponding neuron's estimated IE:

$$\hat{\mathbf{IE}}_u = (z'_u - z_u) \odot \nabla_{u_{out}} m(s) \tag{6}$$

We then compute our metrics (IoU, recall, and cross-task faithfulness) with respect to nodes or neurons, instead of edges. We omit circuits with edges between neurons due to their computational infeasibility, and sparse feature circuits because they would require sparse autoencoders for each model we study.

Our node-level IoU results (Figure 8, left) are rather similar to our edge-level results. The same circuits overlap with one another: Capital-Country and Country-Capital have highly similar circuits, while IOI, Colored Objects, and Entity Tracking are all similar to one another. All other circuits have low IoU. Overall, the circuits are much larger than in the edge scenario, including around 25% of the model on average. This is unsurprising, as nodes are much larger units than edges; including one node is essentially equivalent to including all of its outgoing edges. While we could previously only include one edge out of a node, if that edge was important, we now must include the entire node.

Neuron circuits (Figure 8, right) contrast more with edge circuits. The increased granularity of the computational graph in the neuron case does not enable greater

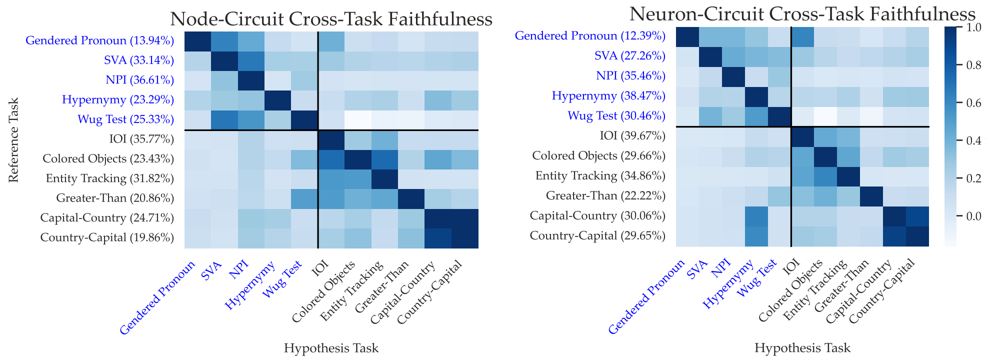

Figure 9
Cross-task faithfulness for node (**left**) and neuron (**right**) circuits. The average size of each task's circuit (as a percent of the entire model) is given in parentheses after each task's name. In both cases, formal tasks clearly have higher cross-task faithfulness with one another than with functional tasks, and vice-versa. Lines divide formal and functional tasks.

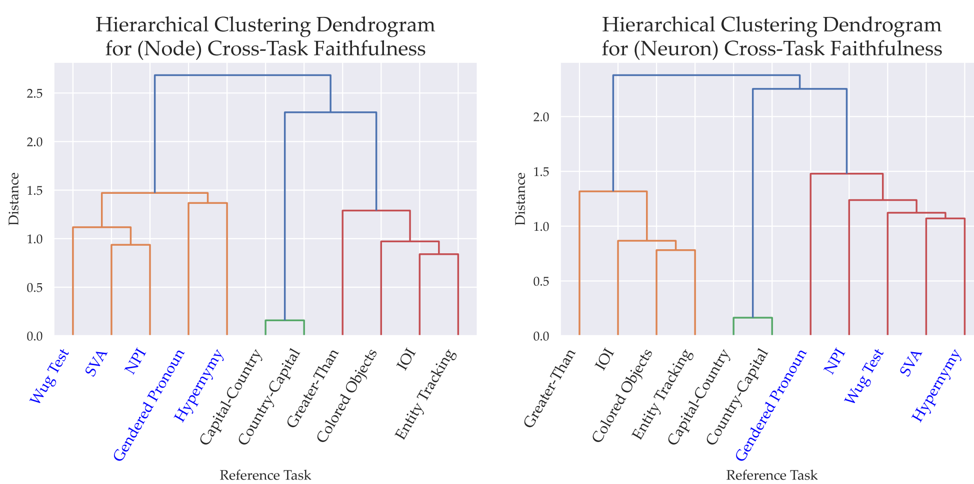

Figure 10
Cross-task faithfulness dendrograms for node (**left**) and neuron (**right**) circuits. In both cases, formal tasks clearly cluster together. Formal tasks are labeled in blue, functional tasks in black.

sparsity; neuron circuits are more similar in size (in terms of percentage of the whole model) to node than edge circuits. While some of the same trends in task overlap are visible as in past IoU analyses, neuron circuits have low IoU in general. Thus, neither IoU analysis provides further evidence for the existence of a formal linguistic region.

However, the cross-task faithfulness of node and neuron circuits (Figure 9) are more suggestive of a formal-functional distinction. In particular, formal tasks have a higher cross-task faithfulness on other formal tasks than on functional tasks, while functional tasks capture one another much better than they capture formal tasks. On the node level, the NPI and SVA circuits capture not only each other, but also the Wug Test and Gendered Pronoun tasks relatively well; on the neuron level, formal-formal similarity is more generalized, but still visible.

A clustering analysis at either level (Figure 10) yields the same conclusion. At both the node and neuron level, there is a clustering of the formal tasks as separate from the functional ones. The consistent emergence of a formal grouping as measured by cross-task faithfulness provides stronger evidence for a division between formal and functional task mechanisms in LLMs. However, the formal grouping that we uncover is defined not by the precise edges, nodes, or neurons that fall within it, but by the general ability of formal tasks to perform other formal tasks.

9. Discussion

In this paper, we have translated Mahowald et al.'s (2024) hypothesis about emergent formal-functional dissociation in LLMs into the circuit analysis framework, and tested the hypothesis. We have tested this in two ways, first measuring if formal and functional circuits contain overlapping units, and then measuring the cross-task faithfulness between the two. Our results indicate that formal and functional circuits have low but non-zero overlap. Moreover, formal circuits also have low overlap in general with one another, suggesting that there is no undifferentiated formal language region in LLMs as in the human brain. Formal circuits are also not a subset of functional ones in general, even when these functional tasks are posed using language. These results are stable across different circuit granularities. However, when we measure cross-task faithfulness, a different pattern emerges. At the edge level, we see some clustering of formal tasks as separate from functional tasks. Then, at the node and neuron level, a stronger clustering of such tasks emerges, with formal tasks clustering entirely separately from functional ones.

Which of these, overlap or cross-task faithfulness, should we trust? Overlap most closely parallels the methods used in the fMRI studies supporting the idea of the language network in the brain: first, one localizes the regions of the brain with a given function and then checks if they are the same. However, we argue that the appropriate measure for mechanistic similarity in LLMs may differ somewhat from this.

The ideal metric for mechanistic similarity should measure not only *if the same units are involved in a given pair of tasks*, but also *if they are involved to the same extent*. This could be achieved by using, e.g. a weighted recall metric, where one task's recall of another's edges is weighted by the IE that the latter assigns to each edge. Even better metrics could compute quantities such as the graph edit distance between the entire IE-scored computational graphs for two different tasks; this sort of metric would remove the need for choosing a circuit of a fixed size, and instead directly compare the causal relevance of every unit between the two tasks. However, as discussed in Section 5, such metrics are hampered by the fact that EAP and EAP-IG estimate IEs relatively poorly in absolute terms (Syed, Rager, and Conmy 2024).

How, then, can we take into account the causal importance of each unit in our model when measuring mechanistic similarity? Cross-task faithfulness measures precisely this: the more a given circuit captures causally important units for a given task, the better that circuit will perform on the task. Moreover, capturing units with a higher IE should naturally yield a larger increase in cross-task faithfulness than capturing units that are causally relevant, but have a low IE. So, we view cross-task faithfulness as a valid and important way to measure mechanistic similarity as well. Indeed, cross-task faithfulness has parallels in humans. *Dual-task interference* studies ask subjects to perform two tasks at the same time; if their performance suffers, one infers that the two tasks share the same neural resources (Watanabe and Funahashi 2014; Leone et al. 2017).

In this paper, we have also conducted a circuit study at a larger scale than most circuit analyses. In particular, many recent circuit papers study one or two tasks in GPT-2 small (Li and Janson 2024) or single models with larger parameter counts (Prakash et al. 2024; Bhaskar et al. 2024). In contrast, we study 5 different models of 2-8 billion parameters, across 10 tasks. This allows us to both build on existing results—we find that the IOI and Colored Objects tasks are not only similar to each other, as found by Merullo, Eickhoff, and Pavlick (2024), but also similar to Entity Tracking. Other results, such as the existence of an MLP backbone across circuits, and the fact that circuits have low overlap with one another in general, were possible only because of the larger models and more numerous tasks that we studied.

In conclusion, we have investigated the question of formal-functional dissociation in LLMs via a wide-ranging circuit similarity study across five formal and five functional tasks, and three granularities. In doing so we provide the first application of circuits to questions from neuroscience and discover new phenomena, such as an MLP backbone running down all circuits. However, as in prior work, we find major differences in similarity depending on whether we measure overlap and cross-task faithfulness (Hanna, Pezzelle, and Belinkov 2024). Measuring overlap shows low formal-functional similarity, but also low formal-formal similarity; in contrast, cross-task faithfulness suggests clear formal-functional dissimilarity and formal-formal similarity. Future work investigating more tasks could establish the range of tasks for which this holds, and further the use of mechanistic interpretability to gain a causally-informed low-level understanding of LLMs and their relationship with the brain.

10. Limitations

In this study, we analyzed five formal and five functional tasks, limited by the difficulty of crafting tasks that fit into the circuits framework; studying more tasks would help confirm that formal and functional tasks do indeed have distinct patterns in cross-task faithfulness. Moreover, we have studied circuits for these tasks found at a fixed faithfulness threshold, 85%; varying this threshold could cause our results to change.

We note that cross-task faithfulness has another limitation: while we generally take high cross-task faithfulness to entail high similarity between two mechanisms, cross-task faithfulness rewards circuits that find components / edges that positively contribute to target task ability. However, insofar as we want to find *complete* circuits (i.e., circuits that contain all causally relevant model units, even negatively acting ones), faithfulness is not always informative; high faithfulness may result from the exclusion of negatively acting units. In the circuit-finding scenario, where we try to find a 85% faithful circuit, we attempt to avoid this by including units based on their absolute IE; in the cross-task faithfulness scenario, however, we cannot be as sure that we have found all important components. We argue that for our purposes, this is acceptable; even if formal task circuits capture primarily positive-contribution components and edges of other formal tasks, this constitutes a formal network.

Task ∣ Model	Llama 3	Qwen 2.5	OLMo	Mistral-v0.3	Gemma 2
Gendered Pronoun	0.99	0.99	0.99	1.00	0.98
SVA	0.92	0.98	0.95	0.96	0.96
NPI	1.00	0.98	0.98	0.98	0.99
Hypernymy	0.98	0.97	0.97	0.99	0.98
Wug Test	0.89	0.87	0.86	0.93	0.90
IOI	0.55	0.72	0.69	0.65	0.79
Colored Objects	0.99	1.00	0.55	0.93	0.88
Entity Tracking	0.80	0.99	0.42	0.77	0.91
Greater-Than	1.00	1.00	0.87	1.00	1.00
Capital-Country	0.87	0.84	0.84	0.86	0.85
Country-Capital	0.97	0.95	0.95	0.96	0.93

Table A.1
Model accuracies on the tasks for which we find circuits. Though some tasks are more difficult than others—IOI and Entity tracking are particularly tough—accuracies generally surpass 90%.

Appendix A: Base Model Performance on Tasks

In this paper, we find circuits for models' task abilities. However, we must verify that models can perform these tasks; otherwise, there may be no circuit to be found. Here, we measure baseline model performance. We measure top-1 accuracy for tasks where there is one clear correct answer: Hypernymy, Wug Test, IOI, Colored Objects, Entity Tracking, Capital-Country, Country-Capital, and all other variants studied.

However, for some tasks (Gendered Pronoun, SVA, NPI, and Greater-Than), there is no one correct answer, and that answer need not always be the top output. Consider the case of SVA with the input *The keys on the cabinet...*; while *are* is clearly more correct than *is*, or any other singular-conjugated verbs, much of the model's probability may be distributed to words like *that*. Such words are neither right nor wrong, so we should not penalize models for outputting them. As such we measure accuracy by scoring each example correct if models assign more probability to the right token (or set thereof) than to the wrong token (or set thereof); alternatively, we could use accuracy with respect to the first token that is either correct or incorrect, rather than neither.

Results for the tasks shared across all models are in Table A.1. Accuracies tend high, except for IOI and Entity Tracking.

Appendix B: Random-Chance Circuit Overlap

Modeling Overlap with a Hypergeometric Distribution. One way to model the overlap that might happen between two circuits simply by chance is to consider a circuit whose edges are selected uniformly at random. Imagine two circuits $C_1 = (V_1, E_1)$ and $C_2 = (V_2, E_2)$, constructed in such a fashion; denote by $G = (V, E)$ the whole model's computational graph. We can model the probability that they overlap to a given degree using a hypergeometric distribution. If the intersection of the two circuits has size $k = |E_1 \cap E_2|$, the probability of an overlap of exactly that size is

$$p(k) = \frac{\binom{|E_2|}{k}\binom{|E|-|E_2|}{|E_1|-k}}{\binom{|E|}{|E_1|}}. \tag{B.1}$$

In words, imagine that we draw $|E_1|$ edges without replacement from our total population of $|E|$ edges; $|E_2|$ edges in this population belong to C_2. This formula computes the probability that we draw precisely $k = |E_1 \cap E_2|$ edges belonging to C_2. This same formula can be applied to node overlap, replacing E, E_1, E_2 with V, V_1, V_2. To compute the probability of an overlap of at least size k, we use the cumulative density function of the hypergeometric distribution, computed with SciPy (Virtanen et al. 2020).

References

Ahmadian, Arash, Saurabh Dash, Hongyu Chen, Bharat Venkitesh, Zhen Stephen Gou, Phil Blunsom, Ahmet Üstün, and Sara Hooker. 2023. Intriguing properties of quantization at scale. In *Advances in Neural Information Processing Systems*, volume 36, pages 34278–34294, Curran Associates, Inc.

AlKhamissi, Badr, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. 2024. The llm language network: A neuroscientific approach for identifying causally task-relevant units.

Amalric, Marie and Stanislas Dehaene. 2019. A distinct cortical network for mathematical knowledge in the human brain. *NeuroImage*, 189:19–31.

Bayazit, Deniz, Negar Foroutan, Zeming Chen, Gail Weiss, and Antoine Bosselut. 2024. Discovering knowledge-critical subnetworks in pretrained language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6549–6583, Association for Computational Linguistics, Miami, Florida, USA.

Berko, Jean. 1958. The child's learning of english morphology. *Word*, 14(2-3):150–177.

Bhaskar, Adithya, Alexander Wettig, Dan Friedman, and Danqi Chen. 2024. Finding transformer circuits with edge pruning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chan, Lawrence, Adrià Garriga-Alonso, Nicholas Goldwosky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. 2022. Causal scrubbing, a method for rigorously testing interpretability hypotheses. *AI Alignment Forum*.

Chang, Tyler A. and Benjamin K. Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350.

Chen, Xuanyi, Josef Affourtit, Rachel Ryskin, Tamar I Regev, Samuel Norman-Haignere, Olessia Jouravlev, Saima Malik-Moraleda, Hope Kean, Rosemary Varley, and Evelina Fedorenko. 2023. The human language system, including its inferior frontal component in "broca's area," does not support music perception. *Cerebral Cortex*, 33(12):7904–7929.

Chintam, Abhijith, Rahel Beloch, Willem Zuidema, Michael Hanna, and Oskar van der Wal. 2023. Identifying and adapting transformer-components responsible for gender bias in an English language model. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 379–394, Association for Computational Linguistics, Singapore.

Choenni, Rochelle, Dan Garrette, and Ekaterina Shutova. 2023. Cross-lingual transfer with language-specific subnetworks for low-resource dependency parsing. *Computational Linguistics*, pages 613–641.

Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Association for Computational Linguistics, Florence, Italy.

Conmy, Arthur, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Csordás, Róbert, Sjoerd van Steenkiste, and Jürgen Schmidhuber. 2021. Are neural nets modular? inspecting functional modularity through differentiable weight masks. In *International Conference on Learning Representations*.

Dettmers, Tim, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332, Curran Associates, Inc.

Dziri, Nouha, Ximing Lu, Melanie Sclar, Xiang (Lorraine) Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems*, volume 36, pages 70293–70332, Curran Associates, Inc.

Elhage, Nelson, 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*.

Fedorenko, Evelina, Idan Asher Blank, Matthew Siegelman, and Zachary Mineroff. 2020. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, 203:104348.

Fedorenko, Evelina, Po-Jang Hsieh, Alfonso Nieto-Castanon, Susan L. Whitfield-Gabrieli, and Nancy G. Kanwisher. 2010. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104 2:1177–94.

Fedorenko, Evelina, Anna A. Ivanova, and Tamar I. Regev. 2024. The language network as a natural kind within the broader landscape of the human brain. *Nature reviews. Neuroscience*.

Ferrando, Javier, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A primer on the inner workings of transformer-based language models.

Finlayson, Matthew, 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, Association for Computational Linguistics, Online.

Forkel, Stephanie J and Peter Hagoort. 2024. Redefining language networks: connectivity beyond localised regions. *Brain Structure and Function*, pages 1–6.

Frankle, Jonathan and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*.

Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

Gauthier, Jon, 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, Association for Computational Linguistics, Online.

Geiger, Atticus, Hanson Lu, Thomas F Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size.

Geva, Mor, 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, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic.

Ghandeharioun, Asma, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Forty-first International Conference on Machine Learning*.

Gould, Rhys, Euan Ong, George Ogden, and Arthur Conmy. 2024. Successor heads: Recurring, interpretable attention heads in the wild. In *The Twelfth International Conference on Learning Representations*.

Groeneveld, Dirk, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Association for Computational Linguistics, Bangkok, Thailand.

Han, Song, Jeff Pool, John Tran, and William J. Dally. 2015. Learning both weights and connections for efficient neural networks. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1135–1143, MIT Press, Cambridge, MA, USA.

Hanna, Michael, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Hanna, Michael, Sandro Pezzelle, and Yonatan Belinkov. 2024. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *First Conference on Language Modeling*.

Hase, Peter, Harry Xie, and Mohit Bansal. 2021. The out-of-distribution problem in explainability and search methods for feature importance explanations. In *Advances in Neural Information Processing Systems*, volume 34, pages 3650–3666, Curran Associates, Inc.

Hu, Jennifer, 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, Association for Computational Linguistics, Online.

Hu, Jennifer, Hannah Small, Hope Kean, Atsushi Takahashi, Leo Zekelman, Daniel Kleinman, Elizabeth Ryan, Alfonso Nieto-Castañón, Victor Ferreira, and Evelina Fedorenko. 2022. Precision fmri reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *Cerebral Cortex*, 33(8):4384–4404.

Hupkes, Dieuwke, Verna Dankers, Mathijs Mul, and Elia Bruni. 2021. Compositionality decomposed: how do neural networks generalise? (extended abstract). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Ivanova, Anna A., Zachary Mineroff, Vitor Zimmerer, Nancy Kanwisher, Rosemary Varley, and Evelina Fedorenko. 2021. The language network is recruited but not required for nonverbal event semantics. *Neurobiology of Language*, 2(2):176–201.

Ivanova, Anna A, Shashank Srikant, Yotaro Sueoka, Hope H Kean, Riva Dhamala, Una-May O'Reilly, Marina U Bers, and Evelina Fedorenko. 2020. Comprehension of computer code relies primarily on domain-general executive brain regions. *eLife*, 9:e58906.

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

Keuleers, Emmanuel and Marc Brysbaert. 2010. Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42:627–633.

Kim, Najoung and Sebastian Schuster. 2023. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855, Association for Computational Linguistics, Toronto, Canada.

Kisako, Riku, Tatsuki Kuribayashi, and Ryohei Sasano. 2025. On representational dissociation of language and arithmetic in large language models.

Kissane, Connor, Robert Krzyzanowski, Arthur Conmy, and Neel Nanda. 2024. Saes (usually) transfer between base and chat models. Alignment Forum.

Leone, Carmela, Peter Feys, Lousin Moumdjian, Emanuele D'Amico, Mario Zappia, and Francesco Patti. 2017. Cognitive-motor dual-task interference: A systematic review of neural correlates. *Neuroscience & Biobehavioral Reviews*, 75:348–360.

Li, Jin, Kelly J. Hiersche, and Zeynep M. Saygin. 2024. Demystifying visual word form area visual and nonvisual response properties with precision fmri. *iScience*, 27(12):111481.

Li, Maximilian and Lucas Janson. 2024. Optimal ablation for interpretability. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Lieberum, Tom, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. 2023. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla.

Lin, Ji, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, pages 87–100.

Lin, Zehui, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Learning language specific sub-network for multilingual machine translation. 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 293–305, Association for Computational Linguistics, Online.

Linzen, Tal, 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.

Llama Team. 2024. The Llama 3 herd of models.

Mahowald, Kyle, 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.

Malik-Moraleda, Saima, Dima Ayyash, Jeanne Gallée, Josef Affourtit, Malte Hoffmann, Zachary Mineroff, Olessia Jouravlev, and Evelina Fedorenko. 2022. An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, 25(8):1014–1019.

Marks, Samuel, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2024. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models.

McCoy, R. Thomas, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN. *Transactions of the Association for Computational Linguistics*, 11:652–670.

McDougall, Callum Stuart, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. 2024. Copy suppression: Comprehensively understanding a motif in language model attention heads. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 337–363, Association for Computational Linguistics, Miami, Florida, US.

Menenti, Laura, Sarah M. E. Gierhan, Katrien Segaert, and Peter Hagoort. 2011. Shared language: Overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional mri. *Psychological Science*, 22(9):1173–1182. PMID: 21841148.

Meng, Kevin, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Merullo, Jack, Carsten Eickhoff, and Ellie Pavlick. 2024. Circuit component reuse across tasks in transformer language models. In *The Twelfth International Conference on Learning Representations*.

Miller, Joseph, Bilal Chughtai, and William Saunders. 2024. Transformer circuit faithfulness metrics are not robust.

Monti, Martin M, Lawrence M Parsons, and Daniel N Osherson. 2009. The boundaries of language and thought in deductive inference. *Proceedings of the National Academy of Sciences*, 106(30):12554–12559.

Murphy, Elliot and Oscar Woolnough. 2024. The language network is topographically diverse and driven by rapid syntactic inferences. *Nature reviews. Neuroscience*.

Nanda, Neel. 2023. Attribution Patching: Activation Patching At Industrial Scale.

Newman, Benjamin, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021. Refining targeted syntactic evaluation of language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3710–3723, Association for Computational Linguistics, Online.

Nikankin, Yaniv, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. 2024. Arithmetic without algorithms: Language models solve math with a bag of heuristics.

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

Olsson, Catherine, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac

Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Pearl, Judea. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 411–420, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press, USA.

Prakash, Nikhil, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. 2024. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *The Twelfth International Conference on Learning Representations*.

Prasanna, Sai, Anna Rogers, and Anna Rumshisky. 2020. When BERT Plays the Lottery, All Tickets Are Winning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3208–3229, Association for Computational Linguistics, Online.

Pritchett, Brianna L., Caitlyn Hoeflin, Kami Koldewyn, Eyal Dechter, and Evelina Fedorenko. 2018. High-level language processing regions are not engaged in action observation or imitation. *Journal of Neurophysiology*, 120(5):2555–2570. PMID: 30156457.

Quillen, Ian A., Melodie Yen, and Stephen M. Wilson. 2021. Distinct neural correlates of linguistic and non-linguistic demand. *Neurobiology of Language*, 2(2):202–225.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Regev, Mor, Christopher J. Honey, Erez Simony, and Uri Hasson. 2013. Selective and invariant neural responses to spoken and written narratives. *Journal of Neuroscience*, 33(40):15978–15988.

Regev, Tamar I, Hee So Kim, Xuanyi Chen, Josef Affourtit, Abigail E Schipper, Leon Bergen, Kyle Mahowald, and Evelina Fedorenko. 2024. High-level language brain regions process sublexical regularities. *Cerebral Cortex*, 34(3):bhae077.

Shain, Cory, Hope Kean, Colton Casto, Benjamin Lipkin, Josef Affourtit, Matthew Siegelman, Francis Mollica, and Evelina Fedorenko. 2024. Distributed sensitivity to syntax and semantics throughout the language network. *Journal of Cognitive Neuroscience*, 36(7):1427–1471.

Shain, Cory, Alexander Paunov, Xuanyi Chen, Benjamin Lipkin, and Evelina Fedorenko. 2022. No evidence of theory of mind reasoning in the human language network. *Cerebral Cortex*, 33(10):6299–6319.

Shi, Claudia, Nicolas Beltran-Velez, Achille Nazaret, Carolina Zheng, Adrià Garriga-Alonso, Andrew Jesson, Maggie Makar, and David Blei. 2024. Hypothesis testing the circuit hypothesis in LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth,

C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germàn Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Sophie Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Strachan, James WA, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pages 1–11.

Sucholutsky, Ilia, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J. Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. 2024. Getting aligned on representational alignment.

Syed, Aaquib, 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, Association for Computational Linguistics, Miami, Florida, US.

Timkey, William and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Tuckute, Greta, Nancy Kanwisher, and Evelina Fedorenko. 2024. Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47(Volume 47, 2024):277–301.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, Curran Associates, Inc.

Vig, Jesse and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Association for Computational Linguistics, Florence, Italy.

Vig, Jesse, 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, Curran Associates, Inc.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

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

Wang, Kevin Ro, 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 *International Conference on Learning Representations*.

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Watanabe, Kei and Shintaro Funahashi. 2014. Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. *Nature neuroscience*, 17(4):601–611.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Curran Associates Inc., Red Hook, NY, USA.

Xu, Ziwei, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models.

Yang, An, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report.

Zhang, Zhihao, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Unveiling linguistic regions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6228–6247, Association for Computational Linguistics, Bangkok, Thailand.