Compositional Representations of Logic in Large Language Models

April 30, 2025

Abstract

We investigate the internal mechanisms of large language models (LLMs) in how they learn concepts formulated in logical rules (e.g. Blue or Green triangle) through in-context learning. We hypothesize that, in learning and representing the logical concepts, they represent and employ logical operators, predicates, and their compositions, akin to what the literature calls Language of Thought (LoT) for humans. The existence of latent logical structure within LLMs would have an important implication for LLM interpretability, for it suggests the existence of an interpretable hidden "mental language," and we can examine its relationship to the "surface language"—relevant to Chain of Thought (CoT) reasoning. Second, it would serve as a supporting evidence that the human mind employ a LoT and that it is learnable.

We first ask Llama 3, to verbally report the logical rule they learned during the task and see if it is a faithful representation of how they actually solved the task. Faithfulness would imply both the usage of a systematic logical representation in solving the tasks, and the reliability of verbal report for safety purposes. Second, we map out the actual causal mechanisms defining the logical rule that emerged within the LLMs during in-context learning, to show the existence of logical operators and predicates, their generalizability across rules, and their compositional structure. We use causal interventions for these investigations, specifically activation patching to distinguish causal relationships and Boundless Distributed Alignment Search to discover specific activations responsible for the logical representations.

1 Introduction

1.1 Motivation

One approach to Large Language Model (LLM) interpretability is to test its mechanisms against well-studied literature on human cognition. The Language of Thought (LoT) hypothesis is a working theory of the human mind which proposes that human thoughts are represented and processed in a mental language, similar to natural language, with a syntax and semantics that allow for complex thought and reasoning. This means that although human beings do not think in a language, there are primitives in the human mind that form a language-like systematic structure within which thoughts can occur. The logical rule learning task is among the several paradigms that have been well studied on human participants to gain insight into the logical operators and predicates that constitute the LoT. Researchers have created Bayesian inference frameworks with explicit logical primitives to solve these tasks in a human-like way, these models serving as models of human cognition ([Piantadosi(2016)]). These studies gives us a reliable and well-studied paradigm to analyze LLM's internal mechanisms. Empirical tests have found that Large Language Models (LLMs) can perform equivalently as well as both Bayesian models and human beings in the logical

rule learning tasks, and that with finetuning, they serve as a better model of how human beings solve the task than any Bayesian models ([Loo(2025)]).

This prompts the hypothesis that LLMs learned logical operators, predicates, and a syntax of composition during training, and they employ them to learn new logical concepts during the logical rule learning tasks. If true, it would first inform the LoT hypothesis for humans, both in the viability of its account and the learnability of such primitives. Second, regardless of whether human beings actually employ a Language of Thought, the existence of such systematic logical systems within these so-called black box models is useful gives us a well-theorized framework to understand how they work, and enables us to directly access to their thought, instead of relying on proxies such as Chain of Thought (CoT) reasoning. With future work, if we could map out a system of thought representation like an LoT in LLMs, we could acquire a reliable tool to better control the LLMs for safety.

In this paper, we focus on showing the existence of logical operators, predicates, and their compositional structure used to solve the task.

1.2 Organization and Overall Methods

Following these motivations, our experiments are organized in two parts. In the first, we ask LLMs to self-report the rules they learned to solve the task in addition to solving the task itself. We examine whether the self reported explanation is faithful to actually how they solved the task, meaning that there is a single underlying representation of the rule that is both verbalized and used to solve the task. Since LLMs are trained on next word predictions, there is no a priori guarantee that their self-explanations are faithful to their thought processes. Moreover, if their mechanisms were any similar to human cognition, [Reber and Lewis(1977)] suggests that they might struggle to give an explanation altogether or ad-hoc justify their behavior instead of reporting it faithfully. However, Loo et al. 2025 finds that, in logical rule learning tasks, unlike human subjects in [Piantadosi(2016)], some LLMs are able to coherently self-report the rules they learned in a more or less logical form. Then, if we also find that these self-reports are indeed faithful to its actual mental representation, we can conclude that what the LLMs learned and used to solve the task is also in a logical form. Moreover, faithfulness implies that the LLMs' self-explanation is faithfulness to its "thought process," having implications on CoT and AI safety.

The second experiment, with pending results, search for the specific logical primitives, predicates, and their compositions that represent logical rules, by manipulating the causal mechanisms that the models learned to solve the task. If we do find such a robust system, it would be an evidence of something akin to a LoT. Mapping out the causal mechanisms will also give us an insight into the exact nature of the logical representation (e.g. how XOR is represented in terms of more basic primitives) and allow us to intervene to change model behavior.

Before reporting on these experiments, we first provide results from three additional experiments we conducted to choose a model to run the main experiments. We do so because it has implications on the existence of the internal logical structures in different types of models that vary in size and training method, and on what we can postulate about LoTs in human psychology.

For the experiments we use mechanistic interpretability approaches, specifically causal intervention. Following [Meng(2022)] and [Geiger et al.(2021)Geiger, Lu, Icard, and Potts], we use activation patching and counterfactual behaviors to distinguish causal relationship between model behavior and self-explanation in the first experiment, and between different logical operators and predicates in the second experiment. For the second experiment in particular, we use Boundless Distributed Alignment Search (DAS) from [Wu(2024)] to look for specific representations within

un-aligned subspaces of the models' representation space.

1.3 Logical Rule Learning Tasks

1.3.1 The Task

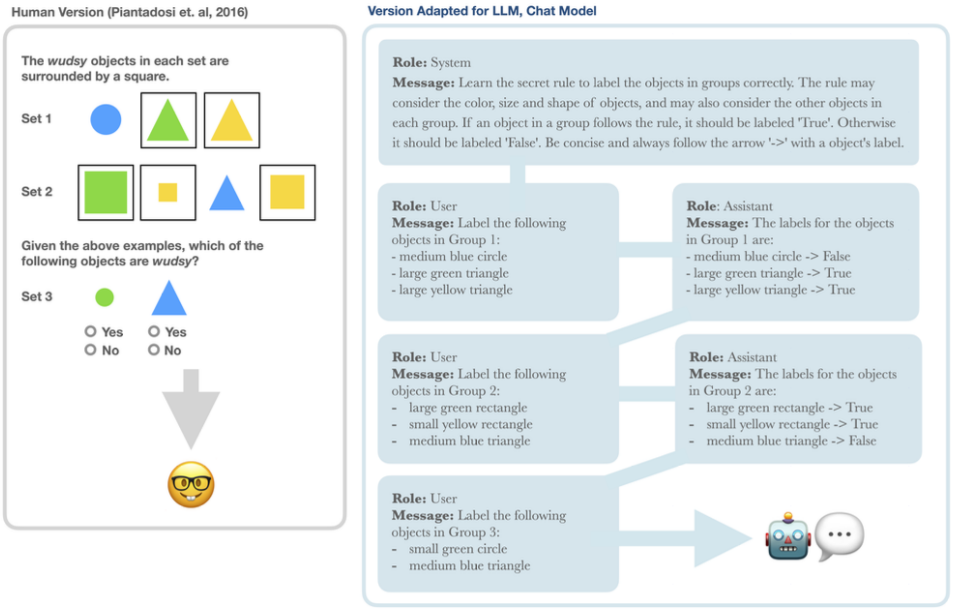


Figure 1: A logical rule learning task developed by Piantadosi et al 2016 (left) and adapted to LLMs by Loo et al. 2025 (right). Over many examples the subjects attempt to learn a concept that is formulated as a logical rule.

We use the *logical rule learning* paradigm to test the existence and usage of LoT within LLMs because it has been used extensively with humans ([Bruner(1956)], [Feldman(2000)], [Kemp(2012)], [Piantadosi(2016)]). This provides us certainty that these tasks can give meaningful results, but also allow for comparisons with a wealth of human data including learning curves and error patterns. Logical rule learning paradigm basically involves objects that are categorized into concepts formulated in an underlying rule, and subjects are tasked to induce the underlying rule based on these categorizations.

We will outline the specific set of tasks developed by [Piantadosi(2016)] because it is the basis by which we are investigating logical primitives in LLMs. Each object is specified in terms of shape, color, and size, and each concept is formulated with first-order logic in terms of these predicates. For example, in the example in the figure above, the concept *wudsy* might mean "objects that are green or yellow," or "objects that have the same shape as the green object in the set."

For each round of the task, the subjects are given a set of objects that they need to categorize whether each object belongs to a concept (i.e. *wudsy*) or not. Then, they are given an answer to the correct categorizations, and move on to the next round to be tasked to categorize another set of objects. This procedure repeats and the subjects gain more and more data points as to what the concept might be, refining their mental hypothesis and predictions.

1.3.2 The Learning Curve

Since the subjects refine their hypothesis on what the logical rule that constitutes the concept might be over the course of the experiment, we can observe how their categorizations change over time which is a proxy of how their mental representation of the concept changes over time. This dynamic change is called the learning curve, and will be the basis with which to judge how close a model is at approximating human mental representations of logical primitives.

1.3.3 Categorization and Rule Elicitation

The main experiments in [Piantadosi(2016)] and [Loo(2025)] involved asking the subjects to categorize the objects into whether they fit the concept they have learned or not, but they were sometimes also asked to verbalize the rule that they have learned. We call the former main task "categorization" and the latter verbalization task "rule elicitation." As will be further discussed later, human beings can achieve high accuracy in categorization while struggling to verbalize the rules.

We note that both human and LLM subjects might represent logical rules to formulate the concept that they are learning, but do so subconsciously without being able to verbalize them. Moreover, they might both attempt to post-hoc justify their own behavior when asked for verbalization, resulting in a non-faithful elicitation of the rule they have learned.

2 Related Works

[Reber and Lewis(1977)]

[Reber and Lewis(1977)] is one of the earlier papers on artificial grammar learning, which ask human subjects to discriminate strings that are generated by an underlying grammar rule. Other than being one of the early rule learning task studies, their investigation showed that participants who are proficient at solving the tasks were, in surprisingly high frequency, unable to verbalize a formal rationale for their choices or gave an irrelevant one. This shows that the learned rule might be part of what people theorize as implicit knowledge, where the content of the knowledge is both highly systematic and subconscious. The implications of this is the highest for our second experiment, where we test whether LLMs can faithfully self-report the rules they have learned. If it is the case, we have a further evidence for the robustness of the internal representation of logical primitives, and also for the reliability of CoT. If it is not the case, we have further reason to investigate LoT above CoT

[Piantadosi(2016)]

[Piantadosi(2016)] showed an empirical method to test specific Language of Thought (LoT) theories. They searched for the most likely set of primitives of LoT that human subjects might have used in solving a set of large-scale concept learning tasks that they developed, described in the previous section. To do this, they developed psychologically realistic approximate Bayesian inference models with different built-in LoTs and tested which model best approximated the learning curve of human subjects throughout the trials. They showed that the inferences of human subjects were the most consistent with a rich set of Boolean operations including first-order quantification. Their results give us the specific task and benchmark with which to test the LLMs and also the hypothesis

that LLMs might employ an LoT that resemble first-order logic.

[Loo(2025)]

[Loo(2025)] tested the same logical rule learning tasks on LLMs and discovered that LLMs performed as good as Bayesian models and human subjects across tasks. Second, they discovered that, after fine-tuning, LLMs can approximate human learning curves of these logical concepts just as well as Bayesian models.

These results directly prompt the research hypothesis of our own experiments which is whether LLMs also use LoTs in solving these tasks. If true, it will add to the evidence that the human mind employ them, and also that LoTs are learnable, since, unlike Bayesian models, LLMs are trained with *carte blanche*, without specified logical primitives.

[Loo(2025)] also tested GPT4 on rule elicitation, and found that elicited rules were highly consistent with categorization. This prompted our second experiment, which is to use mechanistic interpretability methods to fine the actual causal relationship between the two.

[Tan(2023)]

[Tan(2023)] showed cases where the LLM's verbalization is faithful to the causal mechanism they actually employed in their generation processes. They also employed causal abstraction, which is used in our second experiment, but is testing elicitation faithfulness, which we examine in the second experiment. Hence our results will complement their results, and their results indicate that the elicited rules is likely to be faithful.

[Yang et al.(2025)Yang, Campbell, Huang, Wang, Cohen, and Webb]

[Yang et al.(2025)Yang, Campbell, Huang, Wang, Cohen, and Webb] attempts to resolve the long-standing debate between symbolic and neural network approaches to reasoning capabilities, suggesting that emergent reasoning in neural networks depends on the emergence of symbolic mechanisms. They used mechanistic interpretability methods to discover emergent symbolic architecture within LLMs that allow it to abstract the logical variables, induce rules, and apply the rule to a new instance.

The results give us evidence that LLMs do acquire symbolic architecture during its training process, and employ it to solve tasks, supporting out LoT hypothesis. They also indicate that we might find the induced representation of rules for Pinatadosi et al.'s rule learning tasks to be found in the middle layers of the model.

3 Model Selection: Verbalization Proficiency and Correspondence with Human Responses

For the two experiments, we need a model that is open-source to allow for mechanistic interpretability, able to solve the task in very high accuracy, and large enough to self-report the rules it learned. [Loo(2025)] tested categorization and rule elicitation on GPT4, but since it is a closed-source model, we benchmarked Llama 3 models and Gemma 2 models on the same tasks as GPT4. We also desired a model where its learning curve throughout the in-context learning process matched that of the human data, since [Piantadosi(2016)] used learning curve correspondence to find a Bayesian models

with LoTs with logical primitives that most likely matched that of humans. Hence, if a model performs well and reliably self-reports rules but has a divergent learning curve from humans, it may have learned a different set of logical primitives than that of humans.

3.1 Model Performance on Rule Elicitation

We first replicate the second experiment of [Loo(2025)] on Llama 3.3 models and Gemma 2 models, testing how models perform on categorization and rule elicitation. For [Loo(2025)], rule elicitation was only possible with large non-open-source models such as GPT4, for verbalizing a rule in itself is a difficult task that human beings also struggle with ([Reber and Lewis(1977)]) and the task that required both verbalization and categorization was a complicated task. Since we aimed to perform mechanistic interpretability methods on verbalization and categorization, we needed to find a newer and perhaps larger open source model that would perform as well as GPT4.

3.1.1 Methods

Prompting. In each prompt, we included a system message of instructions, sets of labeled exemplars, and then the new object set that is to be labeled. The system message instructs GPT4 to infer a labeling rule from the labeled exemplars, explain the labeling rule concisely, and then provide labels for the new object set.

Evaluation Metrics. Their output was measured in three metrics. Rule Correctness refers to the proportions of elicited rules that accurately captured the correct categorizations. The GPT4 data is drawn from [Loo(2025)] and uses a slightly different metric—it measures the average proportion of correct categorizations that each rule accounted for. Categorization correctness refers to whether the label the model gave matched the ideal correct label in the dataset. Consistency refers to whether the categorization was consistent with the models' self-reported rules.

Evaluation Methods. Every model except for GPT4 was graded by ChatGPT and hence might contain error. For GPT4, its self-reported rules were converted to a Python code by GPT3.5 which was then analyzed.

3.1.2 Results

Figure 2 shows that both Llama 3.1 405b and Llama 3.3 70b outperforms GPT4, making it a good candidate for further mechanistic interpretability experiments as open-source models. We find that performance across the metrics heavily depend on model size, with smaller model performing poorly, especially considering that both of the small models performed well when asked only to categorize the objects. We suspect that this is due to some sort of a *cogntivie overload* on a small model.

3.2 Learning Curve Correspondence with Human Data

We replicate the third experiment of [Loo(2025)] on the same models, calculating the R^2 correlation between LLM and human categorizations to compare learning curves. Since [Piantadosi(2016)] used learning curves to determine the most likely logical primitives that is present in the human mind, low correlation of learning curves might mean that, even if LLMs do employ logical primitives to learn new concepts, they might not possess a similar set of primitives to humans.

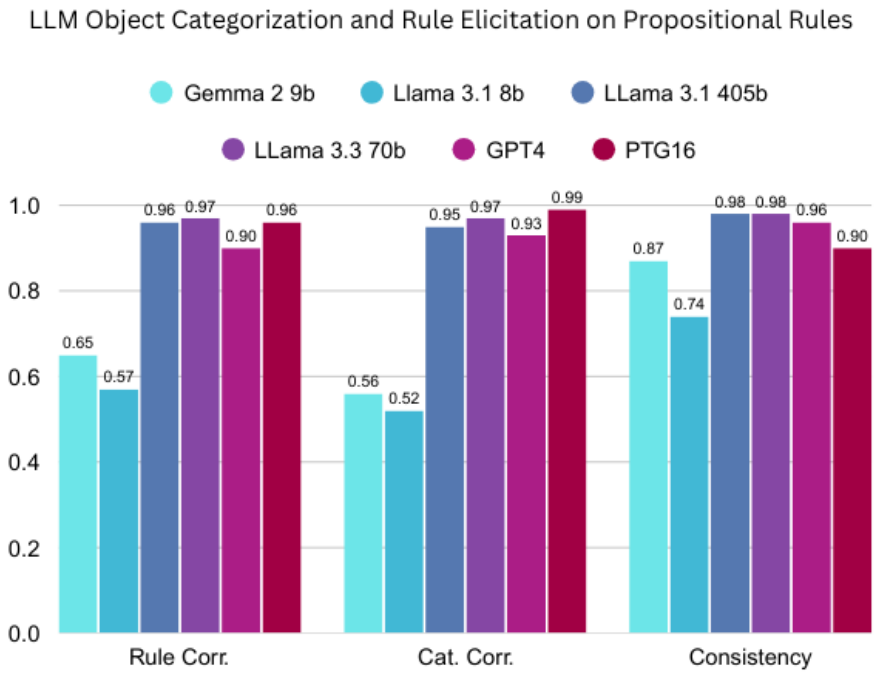


Figure 2: Instruction-tuned LLMs were evaluated on how well it categorized objects, the correctness of the self-reported rules, and the consistency between its categorization and self-reported rule.

3.2.1 Methods

Prompting. The model a system message of instructions, sets of labeled exemplars, and then the new object set that is to be labeled. The model was tasked to label the new object set without verbalizing the rule it learned.

Evaluation Metric. The model's probability of outputting 'True' as the next token was used as a proxy to the model's confidence of the answer. This was compared to the proportion of human subjects who answered 'True' which was a proxy of each individual's confidence in the answer. These numbers across all rules and objects were compared by computing the R^2 score (square of Pearson's correlation coefficient).

3.2.2 Results

Figure 3 shows that responses from any LLM that is not finetuned on human data on rule learning tasks is not highly correlated with human responses.

3.3 Comparing Predictability of Human Responses by LLMs and Bayesian Models

Since we want to see that high correlation between LLM and human answers do not only result from high task accuracy from both agents, we instead put LLMs against the best Bayesian model in how well they predict human responses. Since Bayesian models also have high task accuracy, the predictability will be a signifier of high correlation with human responses independent to accuracy.

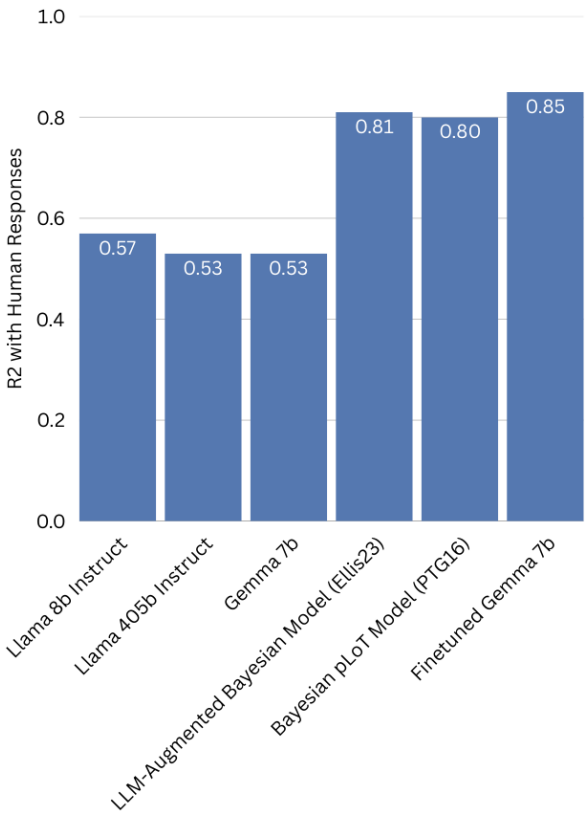


Figure 3: Enter Caption

3.3.1 Methods

Hence we performed an analysis where, given an LLM and a Bayesian model trained on human responses, we looked at occasions both where LLMs and humans agreed on five consecutive categorizations and where Bayesian models and humans agreed on five consecutive categorizations. Then, we looked at which is better at matching the human response on the sixth categorization task. LLMs that are finetuned are shown to do better than Bayesian models. We hope that instruction tuning will also bring this result, but have not tested it yet.

3.3.2 Results

The figures 4 and 5 show that Gemma-7b finetuned on human responses on the logical rule learning tasks is a better predictor of human responses than the best Bayesian model presented by [Piantadosi(2016)]. The non-finetuned model and the finetuned smaller model (GPT2) are worse predictor than the Bayesian model.

3.4 General Discussion

These results show that performance on rule elicitation depend mostly on model size, and that correspondence with human learning curve is based on the model being trained on the data, whether an LLM was finetuned on human responses or a Bayesian model was explicitly trained on human

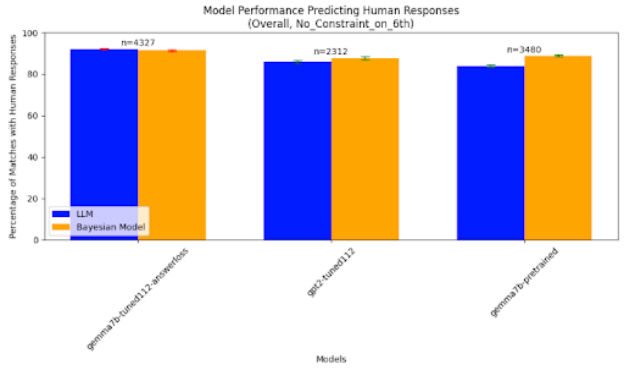


Figure 4: Comparison of PTG-16 (Bayesian model from [Piantadosi(2016)]) and GPT2, Gemma-7b, and Gemma-7b finetuned on human responses. The y-axis represents the score of how many times each model response agreed with the human response on the sixth categorization, following five consecutive agreements.

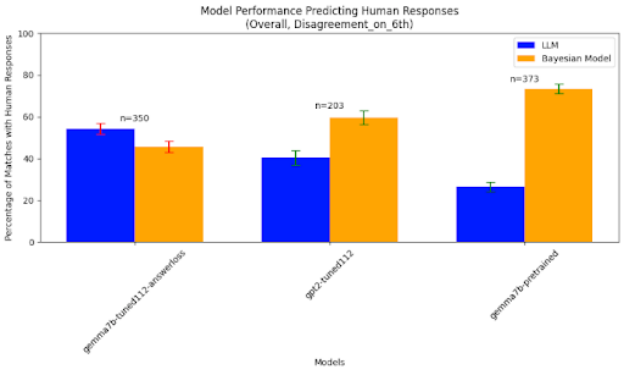


Figure 5: Same comparison but with an additional restriction that PTG-16 and each LLM had to have a disagreement on the sixth categorization, and seeing which one agreed with the human response.

data. The rule elicitation performance prompts us to perform the rest of the experiment on large Llama 3 models. We decide not to finetune it to match human data for the following reasons.

First, the relatively low correspondence of instruction-tuned LLMs and human learning curves could be due to a number of factors, including the fact that the visual prompt given to human subjects were adapted to a text format for the LLMs. Moreover, due to these translational hurdles, the higher correspondence for Bayesian models and finetuned LLMs may result simply from the fact that they had access to human data, instead of having a more similar set of logical primitives to human subjects. This would make more sense considering that it is hard to believe that finetuning on logical rule learning tasks would fundamentally change the internal representation of logical primitives in LLMs. Indeed, sharing the same logical primitives does not guarantee similar learning curves but enables it. Hence, we believe the increase in correspondence with human learning curves result from other factors rather than the logical primitives. In future studies, we could use causal intervention to see if off-the-shelf LLMs and finetuned LLMs share the same primitives.

4 First Experiment: Faithfulness of Rule Elicitation

4.1 Motivation

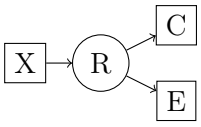
[Loo(2025)] showed that GPT4 was able to verbalize the logically structured rule that constitutes the concept it was learning, and that the categorizations it makes and the elicited rules were highly consistent. From this, they cautiously conclude that the model's stated rules and its classification behavior stem from a common representation.

However, prior work has pointed that LLMs could ad-hoc justify the behavior they have shown and that these justifications are influenced by various kinds of biases ([Turpin et al.(2023)Turpin, Michael, Perez, an This would make sense considering that instruction-tuned LLMs are trained to give answers that are helpful to users instead of being trained to solve tasks. Hence, we are left to prove the strong claim that LLM's classifications and the elicited rules stem from the same underlying representation of the logical rule that it has learned.

What the faithfulness of rule elicitation provides is an indirect insight into the nature of how LLMs represent the logical concept they learned, which, since the elicited rule is in a logical form, is likely to also be in a logical form, in terms of logical operators, predicates and their composition. Moreover, faithfulness would indicate that the internal representations of the logical rules can be used in a variety of tasks including both categorization and rule elicitations, serving as an evidence for their generalizability across tasks. Further study can investigate different tasks that involve logical concept representation and see if operators and predicates carry over.

4.2 Hypothesis Space of Causal Relationships and Experimental Design

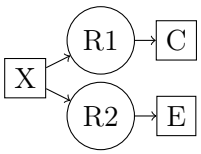
By the elicited rule being faithful to the way LLMs categorized objects, we mean that there is a single internal representation of the rule that is used to both carry out the categorizations and to self-report the rule it is using. The causal graph looks like the following:



Graph 1.

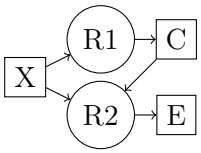
Here, 'X' stands for examples that were given before the task prompt, 'R' stands for a abstract rule representation (e.g. Blue or Rectangle), 'C' for the categorization tokens (e.g. the token 'True'), and 'E' for the elicited rule tokens (e.g. 'The object is labeled true if it is blue or a rectangle').

Out of countless possible causal graphs between these sixth variables (plus additional rule representations), many are impossible due to the ordering of the prompt tokens (e.g. 'X' cannot be influenced by other variables), the assumption that highly accurate categorization and sensible rule elicitations must be influenced by the given examples if not an abstract rule representation (i.e. 'C' and 'E' must be caused by some 'X' or 'R'). With these in mind, here are some salient examples of possible causal graphs:

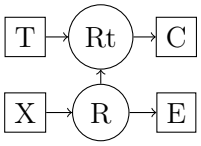


Graph 2. The above graph is where the LLM represents two (or more) rules and hence the categorization and rule elicitation are not caused by the same representation. For example, the

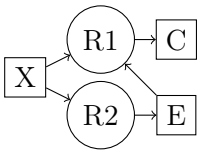
LLM could represent a rule to solve a task, and when later asked to report the rule, it can look back into the examples and come up with a new rule to share.



*Graph 3.*The above graph represents post-hoc elicitation, where the LLM comes up with the elicited rule by looking at both the examples and its own categorization, justifying its answer.



Graph 4. 'T' here represents the task object (e.g. Small Blue Triangle). The above graph represents the case where the task object influences the formation of the categorization representation.



Graph 5. The above graph represents the opposite case where the elicited rule influences the categorization. This was possible in the experiment in Loo et al. 2025 that tested the consistency between elicited rules and categorizations, which we also performed on new models in section 3.

If the elicited rule was faithful, this was true and we were to intervene on 'R', we would find that every time the 'C', the 'E' also changes. The converse is not true, for two different rules might lead to the same categorization of a given object. In experiment 3.1, we look for this pattern. Now, this behavior could also be seen in *Graph 3*, for every time the categorization flips, it might affect the elicited rule representation enough to alter it. One way to tell apart Graph 1 and Graph 3 is to directly intervene on 'C' and seeing if 'E' changes. This will be done in experiment 3.3.

4.3 Experiment 1.1: Co-Flipping of Categorization and Elicited Rule with Intervened Rule

4.4 Experimental Design and Methodology

We do the following to ensure intervention on rule representations, not other variables that we are considering, or some other representations that might randomly perturb the output. First, we generate a "correct" run in which the model labels the test object accurately and provides a plausible rule. Next, we create a "flipped" run by systematically perturbing label-token embeddings to reverse the categorization outcome. This way, the differences in activations across token positions and layers for these two runs will likely be only related to the rule representations.

Second, we perform intervention on token positions after the examples before the categorization. This way, it is likely that the rule representations exists at this point, and the categorization representation or categorization token has not appeared yet. We also ask LLMs to produce categorization before rule elicitations, to eliminate cases represented by *Graph 5*, which is harder to account for in a separate experiments than *Graph 4*.

The overall study design is very similar to that of ROME. We use the Pyvene library to perform activation patching: insert internal activations from the flipped run into the correct run, layer by layer and token by token (using sliding windows). After each intervention, we measure changes in:

- The label assigned by the model,
- The elicited rule, and
- The logical consistency between the new rule and label.

We record the proportion of interventions that flip the label, how often the rule flips as well (co-flip rate), and how many co-flips yield logically consistent outcomes. These metrics are visualized through heatmaps spanning different layers, token positions, and activation types (residual streams, MLP outputs, attention heads).

4.5 Preliminary Results

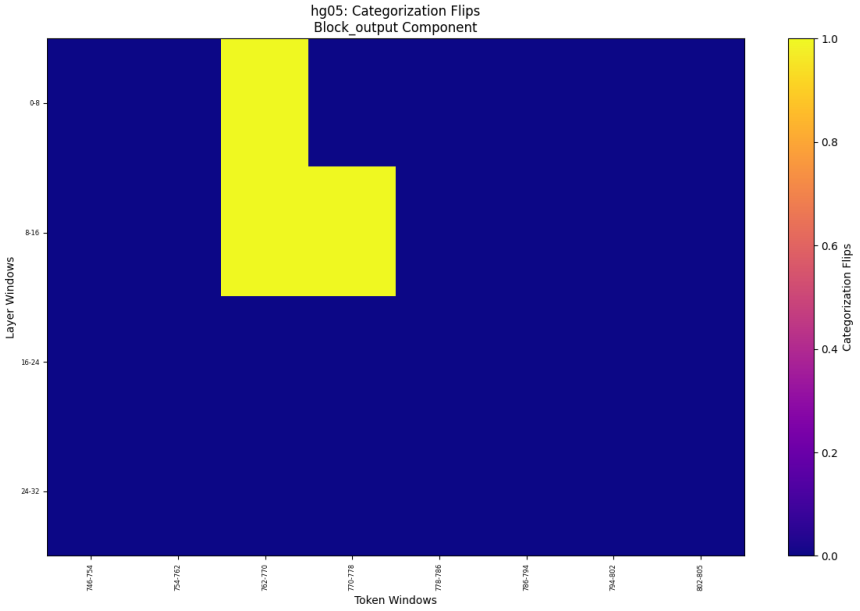


Figure 6: The rule the LLM was tasked to learn was "not a Circle". Yellow regions represent the token positions and layers in which intervening the block output from the clean run into the noised run restored the original object categorization.

For the rule "not a Circle" (Figure 6), We find that representations that causally determine the categorization lies in token positions during which the test object is introduced, and in early layers. We also find that rule never changed when categorization flipped, giving us the co-flip rate of 0. The non-existence of regions that give co-flips heavily suggest that there is no underlying representation that causes both the categorization and the elicited rule.

For the rule "Blue or a Circle" (Figure 7), we also find that there are no co-flips. The regions that cause categorization flips are more spread out, and this could be due to a number of factors: we intervened on smaller token positions and perhaps the representation is more spread out, and since the rule is more complex, perturbations on non-essential representations may have had more downstream effect on the ultimate categorization.

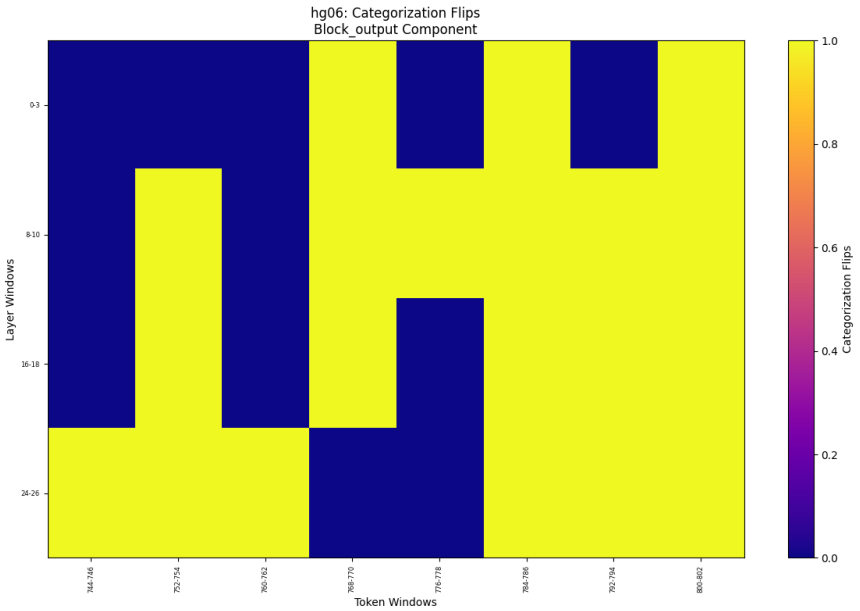


Figure 7: The rule the LLM was tasked to learn was "Blue or a Circle." Yellow regions represent the token positions and layers in which intervening from the clean run into the noised run restored the original object categorization.

4.6 Experiment 1.2: Influence of Categorization Representation on Elicited Rule Representation

4.6.1 Methodology

We prompt Llama 3.1 7b twice on the same examples and objects, with two different objects that should, based on the underlying rule, give the same label. If the models output the same rule, it means that the elicited rule representation is not affected by the test object.

We perform this on five rules and ten example-task set each. The rules tested are the following:

- *True* if BLU or REC
- *True* if not CIR
- *True* if RED and not SQR
- *True* if BLUE and REC
- *True* if RED or not SQR

For each of the ten runs of each rule, we randomly generate 15 examples given a categorization according to the underlying rule. Then, a new object is given and the LLM is asked to categorize it and report the rule.

4.6.2 Results

Figure 8 demonstrates the heavy influence of the test object, indicating a causal influence from 'T' to 'R1' that ultimately influences 'C'.

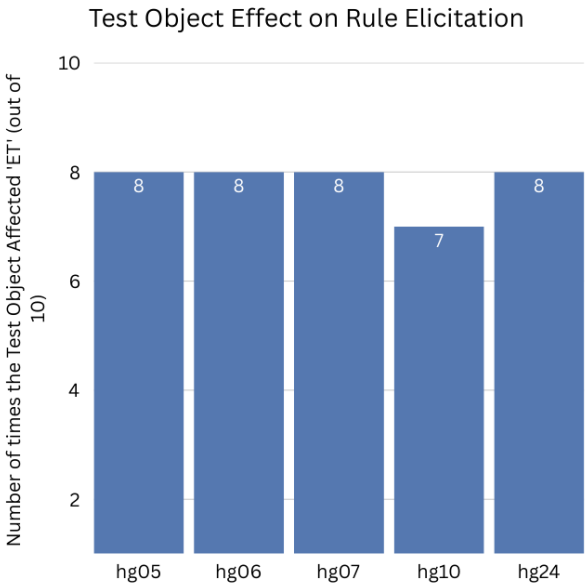


Figure 8: For each rule the rule changed meaningfully around 80% of the time when a different test object was given, even though we did not provide a label for the test object.

4.7 Experiment 1.3: Influence of the Categorization Token on Elicited Rule Representation

4.7.1 Methodology

We ask the LLM to categorize a given object and then elicit the rule, but before the rule is elicited, we swap the categorization token it produced on the test object (e.g. 'False' to 'True'). We measure how many times the elicited rule also flips.

4.7.2 Results

Figure 9 demonstrates the heavy influence of the categorization token on the elicited rule shows that there is a strong causation from 'C' to 'E'.

4.8 General Discussion

According to data not shown in this paper, the preliminary results from experiment 3.1 generalize across intervening on the block output, attention output, and MLP output, and for slightly earlier token positions. The non-existence of co-flips heavily implies that the causal relationships between the given examples, categorization, and there are two rule representations.

Based on the results, especially for the rule "not Circle", and also results from the experiment 3.2, we also conclude that there is a strong causation from 'T' to the rule that causes 'C'. From the experiment 3.3, we also find a strong causation from 'C' to the representation that causes 'E'. The best hypothesis of what the LLM is doing seems to be similar to *Graph 4*, refined below:

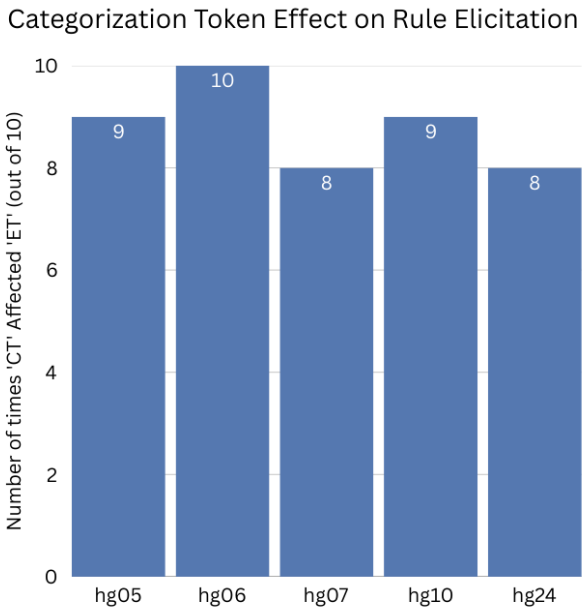
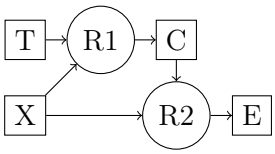


Figure 9: With the average of 90% of the time, the elicited rule changed when we forcibly changed its own categorization of the test object.



This is far from a "faithful" representation, but there are reasons why it makes sense. First, the model we tested was Llama 3.1 7b, which our model selection experiments showed was not capable of simultaneously categorizing objects and reporting rules in a reliable and consistent manner. It is likely that the model is too small for such a complicated task, and cannot maintain a long-term representation, instead relying on local-scale cross-checks that explain the causal graph above. Second, we only ran this once per rule, meaning that the result plots may be hard to interpret since some of the yellow regions may be due to inessential representations perturbing the output enough to change the categorization. This may explain the spread-out nature in Figure 7.

5 Second experiment: Compositionality and Domain-Generality of the Causal Structures of Logical Rules

The second experiment has pending results, but we included here to show the overall experiment design, of how the two experiments would complement each other.

5.1 Motivation

Even if LLMs produce faithful rules formulated with logical operators and predicates, the formulation is still in natural-language token and hence could still be a fuzzy representation of the rule that is not cleanly divided into discrete logical operators, predicates, and their composition. Conversely, even if LLMs do not produce faithful rule elicitations, they might still employ a systematic logical

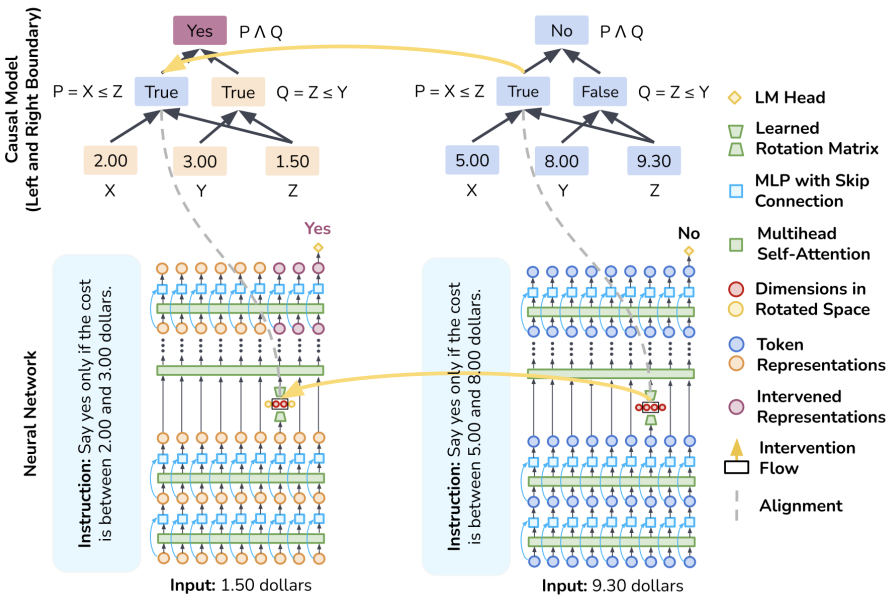


Figure 10: This is a plot from [Wu(2024)] that shows causal abstraction, the technique of learning the correspondence between an abstract causal graph and the neuronal subspaces of LLMs.

representation but only for the categorization. This is perhaps what humans also do.

5.2 General Experimental Design and Methodology

Our second experiment is inspired by [Wu(2024)], which uses Boundless Distributed Alignment Search (DAS) to automatically score how well a given causal abstraction represent the internal mechanism and find exactly which neurons each variables in the abstract causal graph correspond to. Boundless DAS achieves this by going through each token position and layer, learning the neuron subspaces and their linear combinations that correspond to a given abstract variable. The automatic learning is done with gradient descent on a dataset that define the abstract variable.

5.3 Experiment 2.1: Structure of XOR

5.3.1 Methodology

In this experiment, we test whether the model's internal reasoning for an XOR task corresponds to a recognizable logical abstraction.

We define a task in which an object must be labeled *True* if it is "yellow XOR a circle," using descriptions such as "yel cir." We compare two candidate causal models:

- **Model A (Unitary XOR):** The model implements XOR as a single, direct operation.
- **Model B (Compositional Logic):** The model composes simpler logic operations (e.g., OR, AND, NOT) to realize XOR.

Using Boundless DAS, we train a rotation to map from internal representations to each of the candidate causal models, using IIA to assess which yields a better match.

5.3.2 Preliminary Results

[To be completed]

6 General Discussion on Future Results

Overall, the pending results of these experiments could strengthen the case that LLMs can exhibit interpretable and faithful rule-based reasoning. In Experiment 1, we expect large models to show that interventions leading to label changes typically produce corresponding rule changes, indicating that explanations are causally tied to the categorization process. Experiment 2 further intends to directly demonstrate the existence of logical operators, predicates, and their compositional structure, as confirmed by strong Interchange Intervention Accuracy (IIA) scores. We also hope to explore the extent to which these primitives and the compositional structure generalize across rules and tasks.

7 Conclusion

Results are pending, but we designed the study to investigate the existence of logical operators, predicates, and their compositions in learning and representing concepts formulated in logical rules. With the first experiment, we intend to verify whether rule elicitations are causally linked to categorization decisions and whether model computations can be well-described by interpretable logical structures. If the verification fails, we still obtain a support for growing literature on Chain of Thought unfaithfulness, and will inform the interpretation of human behavior in rule learning tasks. It is likely that the result is more nuanced—changing based on model size, task complexity, and prompt ordering—based on our model performance benchmarks and previous literature such as [Geiger et al.(2021)Geiger, Lu, Icard, and Potts]. Our model selection experiment will help such analysis.

With the second experiment, we hope to find the actual representation, and the compositional structure. We believe that we will find some mechanism because of the near-perfect model performance and previous literature that showed symbolic mechanism in LLMs such as [Yang et al.(2025)Yang, Campbell, The nuanced result would be how generalizable these operators and predicates are, which is crucial to determining rather these systems count as a Language of Thought or not.

In short, we believe that this paper will have an impact on our current understanding of LLM interpretability and the systematic mechanisms that emerge within them, faithfulness of Chain of Thought reasoning, and the Language of Thought hypothesis.

References

[Bruner(1956)] Goodnow J. J. Austin G. A. Bruner, J. S. *A study of thinking*. Routledge, 1956.

[Feldman(2000)] J. Feldman. Minimization of boolean complexity in human concept learning. *Nature*, 407:630–633, 2000.

[Geiger et al.(2021)Geiger, Lu, Icard, and Potts] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.02997.

[Kemp(2012)] C. Kemp. Exploring the conceptual universe. *Psychological Review*, 119(4):685–722, 2012.

[Loo(2025)] Ellie Pavlick Roman Feiman Loo, Alyssa. Using llms to investigate the learnability of logical concepts. 2025.

[Meng(2022)] Kevin et al. Meng. Locating and editing factual associations in gpt. In *Neural Information Processing Systems*, 2022.

[Piantadosi(2016)] Joshua B. Tenenbaum Noah D. Goodman Piantadosi, Steven T. The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review 2016*, 2016.

[Reber and Lewis(1977)] Arthur S. Reber and Selma Lewis. Implicit learning: An analysis of the form and structure of a body of tacit knowledge. *Cognition*, 5(4):333–361, 1977. ISSN 0010-0277. doi: https://doi.org/10.1016/0010-0277(77)90020-8. URL https://www.sciencedirect.com/science/article/pii/0010027777900208.

[Tan(2023)] Juanhe (TJ) Tan. Causal abstraction for chain-of-thought reasoning in arithmetic word problems. In Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi, editors, *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 155–168, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.blackboxnlp-1.12. URL https://aclanthology.org/2023.blackboxnlp-1.12/.

[Turpin et al.(2023)Turpin, Michael, Perez, and Bowman] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. 2023. URL https://arxiv.org/abs/2305.04388.

[Wu(2024)] Zhengxuan et al. Wu. Interpretability at scale: Identifying causal mechanisms in alpaca. In *Neural Information Processing Systems*, 2024.

[Yang et al.(2025)Yang, Campbell, Huang, Wang, Cohen, and Webb] Yukang Yang, Declan Campbell, Kaixuan Huang, Mengdi Wang, Jonathan Cohen, and Taylor Webb. Emergent symbolic mechanisms support abstract reasoning in large language models, 2025. URL https://arxiv.org/abs/2502.20332.

8 Appendix A: Other Experimental Results

8.1 Preliminary mechanistic interpretability of rule learning

8.1.1 Replication of representation noising with rudimentary rule learning example

We replicate the methodology of the "space needle" causal tracing example using Pyvene to examine how outputs depend on representations at different layers and token positions.

GPT-2 Large is fed the following prompt: "respond to yellow with ????. blue: !!!, yelow: ". We measure the output token probabilities and, as expecred, find the "????" token highest at 0.53 and the "!!!" token second at 0.19.

We then noise the "yellow" token so that the input is now effectively "respond to [noise] with ????. blue: !!!, yellow: ". We successfully equalise the output token probabilities such that "!!!" is 0.35 and "????" is 0.32.

We then patch in representations from the original run into the noised run at different layer and token positions, finding that representations from key layer and token positions restore the output of the "????" token.

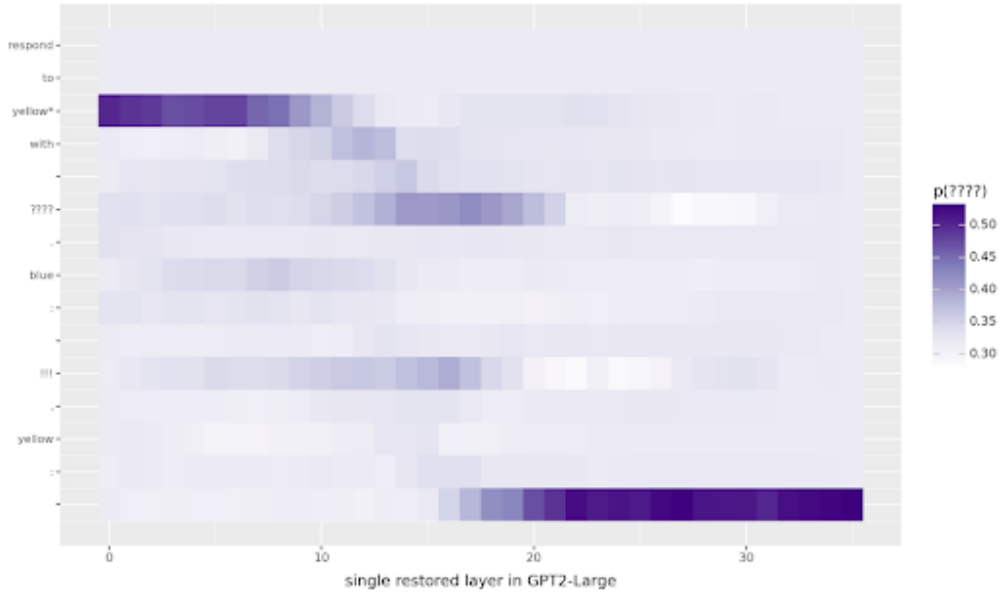


Figure 11: Probability of the original "????" output when representations are restored from the clean into the corrupted run, combining attention and MLP.

As expected, MLPs show a more local pattern, while attention shows longer range dependency. Effects are highest on the key token where the word to match is introduced as "yellow". Additionally, late token representations in late layers (especially when focussing on attention) are critical as processing completes in late layers and late tokens attend back over the sentence.

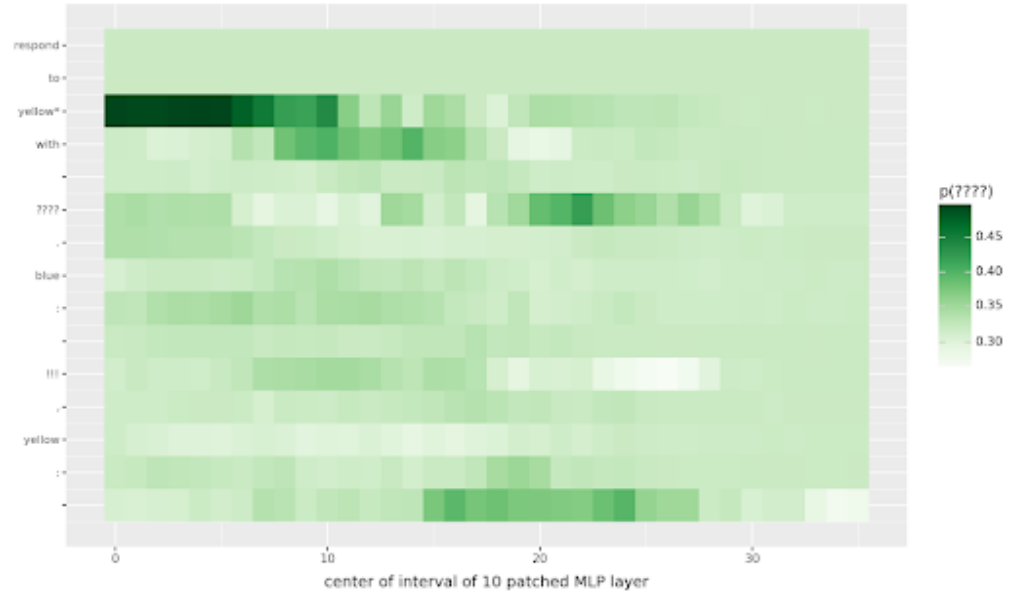


Figure 12: Probability of the original "????" output when representations are restored from the clean into the corrupted run, only transferring MLP.

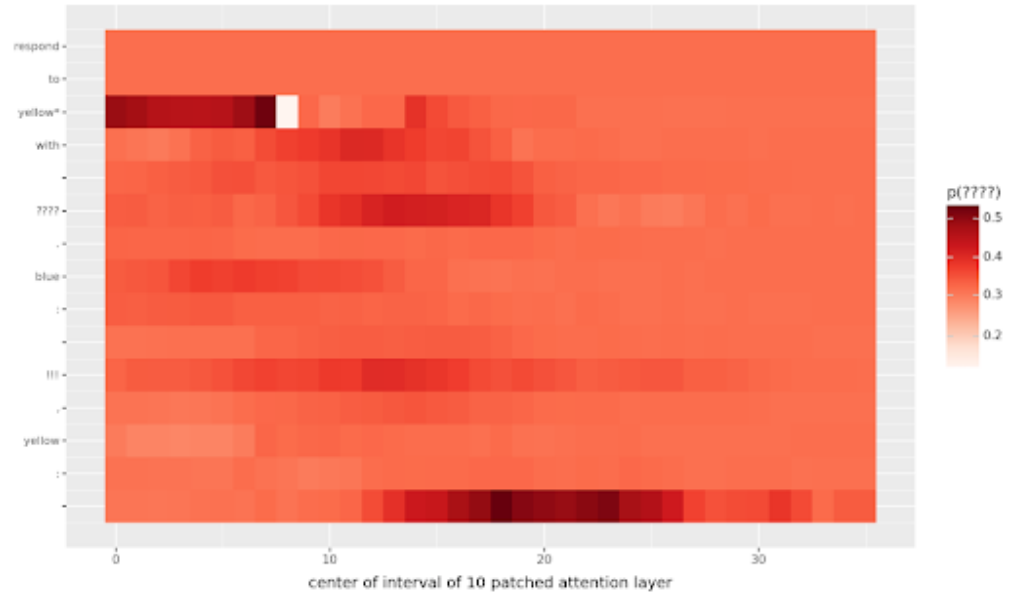


Figure 13: Probability of the original "????" output when representations are restored from the clean into the corrupted run, only transferring attention.