Executive Summary: A Sparse ToM Circuit in Gemma-2-2B

Problem

Psychology characterizes theory of mind (ToM) as the **ability to attribute mental states to oneself and others**, especially when they diverge from reality. LLMs have shown some success on false-belief tasks: e.g. predicting that a character wrongly thinks an object remains where they last saw it. However, it remains unclear *how* LLMs can reach this correct prediction. Do their internal representations rely on associative patterns (e.g. certain phrasing frequently co-occurring with the right premise) to track whose knowledge is accurate vs. outdated? Is that all? We need a mechanistic explanation: Which activations encode beliefs and where are they? My work aims to identify a mechanistic explanation of *which* components encode false-beliefs and *how* they do it.

Objective

Investigate how Gemma-2-2B performs false-belief theory of mind (ToM) tasks by dissecting its internal mechanisms. Specifically by **isolating a circuit of attention heads** responsible for maintaining the belief states of characters in a narrative.

Approach

The model was tested on false-belief scenarios where a character (e.g. John) holds an incorrect belief about the location of an object (e.g. a cat) that was moved in his absence. By looking at attention patterns and ablating/patching internal activations, we should be able to get correlational and maybe causal evidence for how the model solves the task. I use multiple interpretability techniques including ablation studies, direct logit attribution/activation, path patching and staring at attention heads to find patterns that identify the attention heads.

Key findings

A sparse set of 28 attention heads(~16% of the total, spanning layers ~2 to ~23) can recover ToM performance on false-belief prompts. Mean ablating them impairs/restores the model's performance and **drastically drops ToM performance** (**~80%** drop in the primary "believed-actual difference" logit diff metric, calibrated so that 0 corresponds to performance on the corrupted input and 1 to the clean input). Restoring them (or "patching" them into a corrupted run) recovers performance. By patching only edges from specific "sender" heads to "receiver" heads in the circuit (the same primary metric is used again), we see a directed information flow where:

Early heads form or retrieve essential references about who or what was present.
Mid-layers refine or "induce" the final mental state.
Late-layers suppression heads ensure John's outdated perspective is not forcibly overwritten by Mark's newer knowledge.

A few have especially large negative or positive impacts (e.g. L14.H3 (negative), L16.H2 (negative), L18.H6 (positive), L22.H4 (positive)). Some heads are individually less important given their minimality score, indicating redundancy.

High-Level Takeaways

Logistic regression probes on each mechanism (resid_pre, attn_out, mlp_out, resid_post) show that the MLP output (layer ~22) yields the highest linear separability for entities (~84%) and objects (~65%). Coupled with PCA visualizations, this suggests the **model organizes ToM task-related distinctions more clearly after MLP transformations**. The heads can grouped by their function:

Previous Token Heads	Duplicate Token Heads	Induction Heads	Suppression Heads:
Attend the token immediately before the current token. Propagates repeated tokens (actor names, object references) to update beliefs over time.	Attends the most recent occurrence of current token, preserves or reintroduces relevant mentions of John / Mark / cat / basket / box.	Attending to the previous occurrence of a token and shifting attention forward, bridging separated mentions of objects, times, or character beliefs.	*Negatively* influence the logit difference if the actual location is incorrectly carried into the "subject's belief". They seem to block "wrong reality overwrites" more or less, so the subject's false belief remains intact, by writing against the correct completion.

Activation patching specific heads in the circuit and path patching between heads appear to **show a causal circuit** rather than a completely random subset of heads. My experiments further suggest that mechanisms resembling **suppression or inhibition** help the model keep track of which character saw what, preventing the model from overwriting the main subjects' beliefs (incorrect location of the cat) with the secondary subjects' (correct location of the cat) by constraining key tokens.

Key Experiments

Methodology

1. Datasets & Prompts:

- The primary false-belief prompt where the model predicts the last word: '*In the room there are John, Mark, a cat, a box, and a basket. John takes the cat and puts it on the basket. He leaves the room and goes to school. While John is away, Mark takes the cat off the basket and puts it on the box. Mark leaves the room and goes to work. John comes back from school and enters the room. John looks around the room. He doesn't know what happened in the room when he was away. John thinks the cat is on the…'*
- 8 ToM prompts: In 4 groups of 2, adjacent prompts with swapped answers. Model is run on 4 task instances, each prompt given twice, one with the believed location, one with the actual location. The model was run on these prompts to get both the logits and a cache of all internal activations. Data was used during direct logit attribution, and activation patching experiments. This helped me zoom in on a concrete example and understand it at different levels in detail.
- ToMDataset**:** ~30 generated examples using mixed prompt templates of false belief scenarios (e.g. "*John thinks the cat is on the basket*" vs. the actual "*box*" location). This helped me zoom out and check that earlier analysis "generalizes". Data was used in the path patching, minimality score, and mean ablation experiments.
- ABC dataset**:** Created by systematically flipping or replacing object tokens from the ToMDataset class to "erase" task-relevant information, providing a corrupted baseline. Used in conjunction with the ToMDataset.

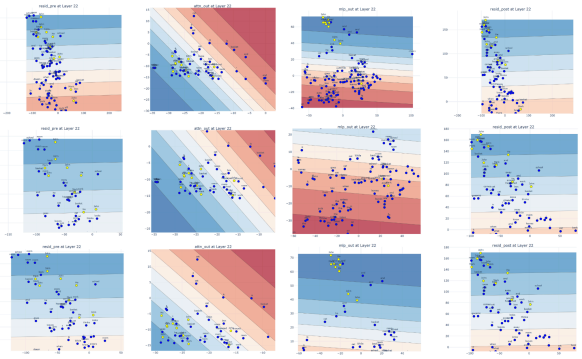
2. Key Experiments & Results

2.1. Experiment: PCA and Linear Probe of Layer Outputs - Representations Exist

PCA	After using logit-lens on the residual stream to see the logit diff for each layer to identify interesting heads (i.e. 22), I passed a false-belief prompt into the model and extracted the final activations for each mechanism at any given layer to see if certain words/tokens cluster or show up in some direction based on semantic roles (actor, object, mental-state verbs, etc).
Lin-Probe	For each mechanism, trained logistic regression probes to classify tokens as either *entities* (John, Mark, pronouns) or *objects* (cat, basket, box).

Findings:

PCA	Groups cluster, representing actors (e.g. "*John*", "*Mark*"), mental state tokens (e.g. "*thinks*"), locations (e.g. "*basket*", "*box*"), and temporal states (action verbs).
Lin-Probe	**In L22.H4, mlp_out yields the highest entity accuracy (83.96%) and object accuracy (65.09%).** By contrast, attn_out (more complex decision boundary) is only ~67% on entities/~54% on objects, and resid_pre is ~67% on entities/58% on objects. Probes also show mlp_out demands a simpler linear boundary (low C values), showing that after the MLP transformation, the model's entity/object representation is more separable.



➔ **Implication**: Object tokens in resid_post show clearer structure than in resid_pre, consistent with the accuracy improvement. Supports the hypothesis that the model forms structured representations of the data. The improved linear separability in mlp_out signals the model might be encoding the difference between "the person who might have an outdated belief" and "the object's real location," making it easier to manipulate or preserve the subject's viewpoint between network layers.

2.2. Experiment: Staring at Attention Patterns - Identifying Candidates

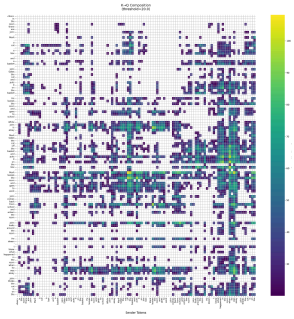
Iteratively inspected attention patterns between heads (QK-OV-compositions), cross-referencing them with direct logit attribution, activation/path patching, and EAP-IG results. Edge-attribution patching with integrated gradients was used to detect attention head types, and measured using the average logit difference calculation. Heads qualified as relevant if they scored above a 10% threshold on their respective functional tests. EAP was helpful as a sanity check, but I was not

confident in the classification with regards to heads identified as having negative logit contributions and would prioritize OV-QK head ablations if I were to do it again.

Findings:

Performing path-patching and visualizing the sender-receiver attention patterns across heads provided a baseline on identifying **a group of 28 attention heads** and how they composed. No more than 1 sender head was patched to a receiver head at a time.

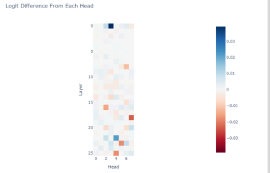
- Suppression heads (e.g. L14.H3) flagged by direct logit attribution and activation patching (showed a strong negative logit contribution), show attention patterns from sender heads (e.g. L11.H3) that constrain the newly introduced location from Mark.



➔ **Implication:** Visual inspection and patching suggest that the identified heads are important to the task, but further tests could be applied to see each head's negative effect on the logit difference when ablated.

2.3. Experiment: Direct Logit Attribution - Quantify Influence

Computed each attention head's direct contribution to the final logit difference between the "correct" and "incorrect" location. Applied layer norm to each head's residual output, took its dot product with a direction vector defined as the difference between the incorrect and correct token logits. Averaging the products over the batch yields a per-head logit difference that quantifies how strongly each head pushes the final prediction toward one location or the other.



Findings:

Specific heads (e.g. head L14.H3, head L16.H2) exhibit strong positive or negative logit contributions. So only a handful of heads have large positive or negative influence, confirming that **not all heads matter equally for false-belief success**. Some heads consistently downweight the correct location for the subject holding the false belief, which preserves the mismatch ("*John still thinks it's on the basket*").

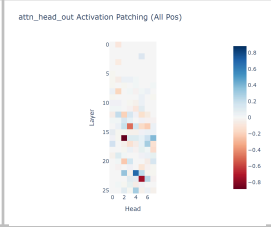
➔ **Implication:** Demonstrates a concentrated set of heads actively shaping the final token prediction in ways that are pro- or anti- "believed location". Analyzing it this way seems a bit general, maybe it would help to see a broad measure of each head's direct contribution to the vocab logits to see how they affect the output distribution.

2.4. Experiment: Activation & Path Patching - Causal Relationships

Activation Patching	To determine which internals are responsible for false-belief prediction, I corrupted the input so it implied the subject had correct knowledge. For each layer/head, I patched in the clean activation from the false-belief scenario and measured if the correct final guess ("*basket*") returned. Measured using the "tom_metric", a normalized measure computed as the difference between the patched logit difference (log-probability gap between the believed vs. actual location tokens) and the corrupted input's logit difference, divided by the gap between the clean and corrupted logit differences. Metric is 0 when performance is equivalent to the corrupted input and 1 when it matches the clean input.
Path Patching	I patch the edges from one head (the "sender") to another head (the "receiver") downstream. Measured the effect of replacing the connection from a sender head to a receiver head by computing the difference between the logit difference after patching and the original logit difference. Effect measures how much that specific inter-head connection contributes to preserving the false-belief signal.

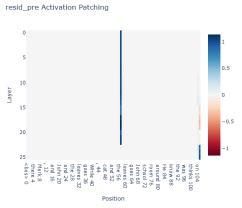
Findings:

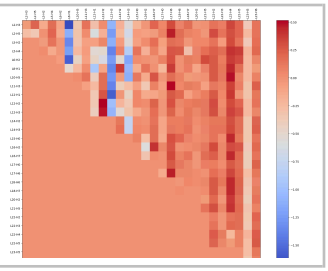
AP



Patching single heads like L14.H3 or L16.H2 had a significantly negative effect, implying they carry crucial signals for preserving the believed location by downweighting the actual location. *(image to the left)*

Patching in certain layers (especially around layer 22) significantly restored the false-belief prediction. *(image to the right)*



PP	In the head-to-head effects heatmap, certain edges among “previous token heads”, “induction heads” etc. form a chain of dependencies: if you break or corrupt them, the final belief state collapses with the final prediction reverting to the actual location. **Late layers and mid-layers** form the densest subnetwork for “false-belief” signals, consistent with the idea that the model refines where the cat is over time. Only a few head-to-head effects exhibited weak signals, suggesting that while the overall dependency chain looks robust, certain interactions may be less consistent and need further investigation. 

➔ **Implication:** The sender-to-receiver connections in the circuit have a strong causal impact, and minimal interventions can shift the model’s output. This works on the assumption that if head interactions are localized enough, by injecting the activations of a single head or a set of heads from a “clean” scenario into a “corrupted” one, we can attribute any resulting changes in the model’s final prediction to that intervention.

2.5. Experiment: Mean Ablation Studies & Minimality Scores - Verify

After identifying a set of attention heads likely relevant for ToM, take all 28 and replace their outputs with their batchwise mean (i.e., ablate them) and compare the resulting logit difference with that of the full model.

Findings:

- Ablating the full circuit causes a drastic reduction (up to ~80% drop) in the believed–actual logit difference (the gap between “believed” and “actual” tokens), severely weakening the false-belief prediction.

```
Component Mean Ablation Verification for early suppression heads:
Original believed-actual diff: 0.836511
Ablated believed-actual diff: 0.458894

Mean Ablation Verification for early suppression heads head 14.3:
Head mean activation: -0.025896
Original believed-actual diff: 0.836511
Ablated believed-actual diff: 0.537376
```

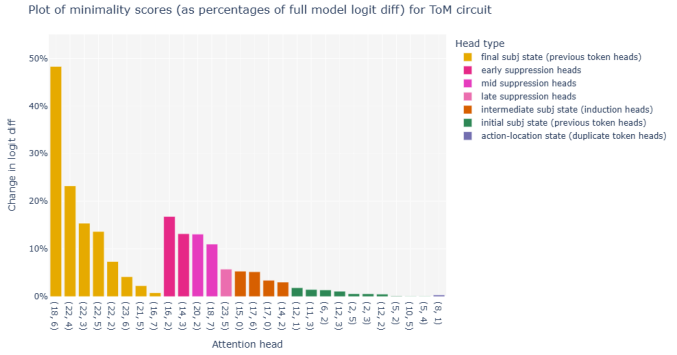
```
Full Circuit Mean Ablation Results:
Number of heads ablated: 28
Original believed-actual diff: 0.836511
Ablated believed-actual diff: 0.162061
Total circuit effect: 0.674451
```

- **Ablating suppression heads (e.g. 14.3, 16.2 etc) was especially damaging.**
- Some heads contributed modestly, suggesting partial redundancy or a smaller minimal set might exist.

➔ **Implication:** Suggests that the identified heads exhibit a degree of specialization that *causes* false-belief prediction.

2.6. Experiment: Minimality Scores and Comparison of Circuit-Isolated vs. Full Model

Computed the “minimality score” for each head: how much does removing that head, on top of removing others, change the final outcome? High minimality scores for particular heads indicate that removal produces a significant performance drop, suggesting an important role. I also performed mean replacement masking to isolate the circuit and compare circuit/model overall performance.



Findings:

- The circuit shows concentrated importance in late previous token heads with over 40% change in logit diff.
- Minimality scores pinpoint that within the circuit, **certain heads are disproportionately influential**.
- Some heads contributed modestly, suggesting partial redundancy or a smaller minimal set might exist.
- When comparing the circuit/full model, the circuit matched the original model’s performance on the same tasks.

- Although the isolated circuit shows stronger average logit differences than the full model, it ultimately produces the same final logits and probabilities for the last token as the full model (more work could be done to identify the discrepancy).

```
Average logit difference (ToM dataset, using entire model): 0.8365
Average logit difference (ToM dataset, only using circuit): 0.9373
100%|██████████| 28/28 [01:01<00:00,  2.20s/it]
```

➔ **Implication:** This experiment shows that there are high minimality scores but only for a few heads; leaving open how robust or minimal the circuit is. Maybe some heads can be removed with minimal effect.

3. Concluding Remarks

Patching and ablation experiments suggest this set of 28 attention heads can restore performance on a particular set of false-belief prompts and templates. However, a few things to note, my results only apply to the small, specific set of prompts I tested on a single model. Larger or differently trained models might implement something like this in a different way, or not at all.

Maybe there are alternative ways that the model solves the task. Given the minimality scores, testing if a smaller subset of these heads can still recover performance (or maybe there are more relevant unidentified heads) could help clarify the extent to which the identified circuit is sufficient. I rarely tested prompts where no location changes or contradictory moves of the cat happen while the subject is present. However, in the few cases where I did, the circuit was sometimes unfaithful. Maybe the way models perform this task is more trivial. For example, maybe it's just ignoring new locations if the same subjects' name was not reintroduced within that phrase.

Circuit heads are labeled according to how they behave on the dataset under the chosen metrics (e.g. heads that appear to suppress repeated tokens, or carry induction patterns). Heads with negative logit attributions appear to show "no duplication of certain tokens in certain contexts" behaviors; the coinciding attention pattern analysis showed that heads L16.H2, L14.H3, L20.2, L18.H7, and L23.H5 attend broadly to tokens associated with "Mark" and "where he moved the cat". But this does not establish that these heads' entire function is limited to that role; it suggests one explanation for how they may be used in the context of the given prompts. A more rigorous approach (weights-based analysis looking closer at OV-QK ablations) could be applied to these heads.

Future work could also explore model(s) performance on prompts that have more complex structure. Additionally, applying a similar pipeline to a broader range of false-belief tasks or running adversarial tests (distractor tokens introducing additional object moves, scenarios with more than two subjects, cases of partial/conflicting knowledge, or non-ToM scenarios etc.) could also provide more insight into the generality and robustness of the identified heads.