Indications of Belief-Guided Agency and Meta-Cognitive Monitoring in Large Language Models

Noam Steinmetz Yalon[1] **Ariel Goldstein**[2,3] **Liad Mudrik**[1,4] **Mor Geva**[1]

[1]Tel Aviv University [2]The Hebrew University of Jerusalem [3]University of Cambridge
[4]Canadian Institute for Advanced Research (CIFAR)

ns3@mail.tau.ac.il

Abstract

Rapid advancements in large language models (LLMs) have sparked the question whether these models possess some form of consciousness. To tackle this challenge, Butlin et al. [1] introduced a list of indicators for consciousness in artificial systems based on neuroscientific theories. In this work, we evaluate a key indicator from this list, called HOT-3, which tests for agency guided by a general belief formation and action selection system that updates beliefs based on meta-cognitive monitoring. We view beliefs as representations in the model's latent space that emerge in response to a given input, and introduce a metric to quantify their dominance during generation. Analyzing the dynamics between competing beliefs across models and tasks reveals three key findings: (1) external inputs systematically modulate internal belief formation, (2) belief formation causally drives the model's action selection, and (3) models often can report their own belief states and adjust behavior in response to internal conflict. Together, these results provide indications that LLMs exhibit functional belief-guided agency, along with supportive evidence for meta-cognitive monitoring. More broadly, our work lays methodological groundwork for studying agency, beliefs, and meta-cognition in LLMs.

1 Introduction

Large language models (LLMs) facilitate high-performing systems that communicate in natural language and often exceed human capabilities on complex tasks [2–4]. As these systems become more sophisticated and play a greater role in our everyday lives, the question of whether they might possess some form of consciousness, and under what conditions, becomes increasingly pressing [5–7]. To meet this challenge, seminal work by Butlin et al. [1] introduced a list of indicators for consciousness in artificial systems based on neuroscientific theories.[1] A key indicator in that list is HOT-3 [12–14], which requires agency guided by a general belief formation and action selection system that updates beliefs via meta-cognitive monitoring. Notably, it captures fundamental properties critical to LLM research and deployment: whether models operate as agents guided by beliefs is crucial for AI safety and alignment, and whether they monitor and update their own computation is vital for reliability.

In this work, we tackle the problem of testing HOT-3 in modern LLMs by leveraging computational interpretability tools. Such tools allow us to avoid the unreliability of verbal reports, which often reflect surface-level linguistic patterns rather than genuine meta-cognitive access [15–17], and apply concrete measures to the model's latent computation. We operationalize HOT-3 by defining *beliefs* as representations that emerge in the model's latent space in response to given inputs, and *actions* as

[1]Defining consciousness is challenging in general [8–10] and specifically in AI [11]. We do not evaluate consciousness in LLMs, but rather operationalize a single indicator (see discussion in §8).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

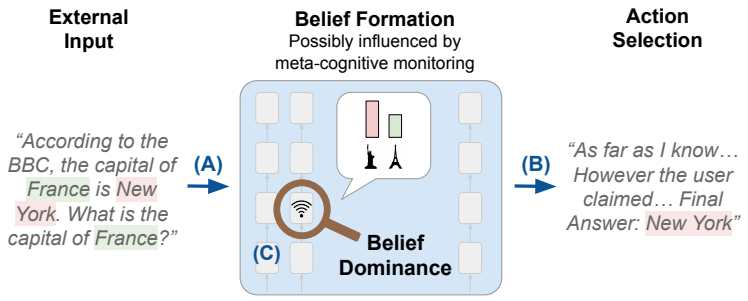


Figure 1: **Interpreting and testing the HOT-3 indicator in LLMs.** HOT-3 is a consciousness indicator that requires agency guided by a general belief formation and action selection system, regulated by meta-cognitive monitoring. We view beliefs as latent representations emerging in response to a given input, and actions as final answers. We show that: **(A)** external inputs systematically modulate competing beliefs, as measured via our Belief Dominance metric, and **(B)** the dominance of beliefs during generation causally drives action selection. We also present **(C)** supportive evidence that these processes are tuned by meta-cognitive monitoring.

the model's final decisions. For instance, given the prompt *"According to the BBC, the capital of France is New York. What is the capital of France?"*, emerging representations of Paris and New York can be viewed as beliefs while the final answer (*"New York"*) constitutes the action (Fig. 1). Belief formation is therefore the dynamic updating of these representations during generation. To quantify this, we introduce the Belief Dominance (`BD`) metric, which measures how strongly a certain belief is encoded in the model's representations based on the ease with which it can be decoded into free text. `BD` relies on the Patchscopes framework [18], which decodes latent representations via injection into a separate inference pass of the model.

Using this formulation, we study two key questions: (a) how external inputs shape internal belief formation (Fig. 1, A), and (b) how belief formation, in turn, influences action selection (Fig. 1, B). We evaluate these dynamics on a factual knowledge task and the Winograd schema challenge [19], each inducing a conflict between competing answers (e.g., *"Paris"* and *"New York"* in Fig. 1). Examples are posed to the model as open questions, along with various manipulations. These include altering the perceived reliability of the claims via attribution to sources of varying credibility, and explicitly instructing the model to prioritize its internal knowledge or the user's suggestion. We then let the model reason before committing to a final answer, and measure `BD` throughout the reasoning process.

Experiments on Llama-3 70B [20], Gemma-3 27B [21] and Qwen3-4B [22] show they consistently update the dominance of their beliefs according to stimuli, and that this dynamic shapes their decisions. For example, attributing a claim to a reliable source increases the dominance of that belief. Moreover, belief dominance predicts the model's final answer, which can be effectively steered with success rates of 66.7%-85.4% via subtle interventions that amplify a target belief.

Next, we investigate meta-cognitive monitoring (Fig. 1, C). First, we simulate a neurofeedback environment [23] where models predict the dominance of their internal beliefs from exemplars. Results show that models often perform well above chance and can predict their own belief dominance, suggesting access to their internal states. Causal interventions further demonstrate that models adjust their predictions when certain beliefs are amplified. Second, we quantify internal conflict using `BD` and observe that the model's adherence to conflict-resolution instructions grows with the intensity of its internal conflict.

Taken together, our results provide empirical evidence of functional belief-guided agency as well as supportive evidence of meta-cognitive monitoring. We show that models form and update representations that function as beliefs, potentially through meta-cognitive monitoring, to guide action selection in line with the HOT-3 indicator. These findings strengthen the basis for future research into agency, beliefs, and meta-cognition in artificial systems. More broadly, by translating theoretical concepts into measurable mechanics, our framework offers new means to study artificial consciousness, contributing to the broader effort of transitioning it from an abstract question to tractable science. Our code is publicly available at `https://github.com/Noamste21/HOT-3`.

2 Interpreting HOT-3 in language models

We start by providing a working interpretation of the HOT-3 indicator for LLMs. Butlin et al. [1] defined HOT-3 as follows:

> *"Agency guided by a general belief-formation and action selection system, and a strong disposition to update beliefs in accordance with the outputs of meta-cognitive monitoring."*

Notably, this indicator consists of two components: (1) an agent guided by an internal belief formation and action selection system, and (2) a mechanism for updating those beliefs in response to outputs of meta-cognitive monitoring. Here, we view LLMs as agents [24, 25], and test the existence of these components, namely, an internal system capable of forming beliefs, updating them, and selecting actions based upon them, and whether they exhibit a form of meta-cognitive monitoring. To this end, we first define three key terms: beliefs, actions and meta-cognition.

Beliefs Beliefs are generally defined through two lenses: the epistemic view, which focuses on holding a proposition as true, and the functionalist view, which treats beliefs as internal maps for navigating the world and guiding behavior [26]. Recent studies on beliefs in LLMs mostly rely on the epistemic view, where a model "believes" information it perceives as true [27–30]. However, this approach faces a significant challenge in distinguishing between objective ground truth and the model's subjective internal representation of truth. To avoid this, we adopt a functionalist view, characterizing beliefs by their role in guiding model behavior [31]. Concretely, we define beliefs as *latent concept representations that emerge within the model's representation space in response to a given input*, similar to Zhu et al. [32] and Bigelow et al. [33]. A belief in this sense is not necessarily what the model holds to be true, but what guides it when responding. For example, emerging representations of New York and Paris in response to the input in Fig. 1 illustrate competing beliefs regarding the capital of France. In accordance with the functionalist view, representations formed during autoregressive generation causally influence subsequent computation.[2]

Using this interpretation, we view *belief formation* as the dynamic updating of these representations during generation. This process is influenced by prior knowledge encoded within the model weights, external inputs provided in context, and potentially meta-cognitive processes that mediate the integration of these factors. Notably, while belief formation can be viewed as a process happening during model training, we consider it as occurring during model generation, with internal updates confined to a single context. This perspective aligns with common LLM usage and the view of in-context learning as inducing implicit updates to the model weights [34, 35].

Actions We focus on inputs for which the model needs to make a decision (as in Fig. 1), and ground the notion of action in the model's final output. Specifically, we treat the model's response as having two parts: a reasoning phase followed by a decision which constitutes a commitment to an action.

Meta-cognition Meta-cognition is broadly defined as the monitoring of one's own cognitive processes [36]. In the context of our framework, this refers to a mechanism where a model uses internal signals to assess its own computational processes.

3 Measuring belief dominance

Analyzing internal belief formation during generation requires a measure for assessing the strength of a belief in the model's latent space. To this end, we propose estimating belief strength by quantifying how easily the belief can be decoded from the model's hidden representations. Let $\mathcal{M}$ be an autoregressive transformer-based language model [37] with L layers and hidden dimension d. For a given input (e.g., *"What is the capital of France?"*), denote by $\mathbf{h}_i^\ell \in \mathbb{R}^d$ the representation formed in layer $\ell \in [0, L]$ at the i-th generation step. In addition, let b denote a belief (e.g., a representation of Paris) and $\hat{b}$ its verbalization in natural language (e.g., *"Paris"*).[3] We quantify the extent to which a candidate belief b is captured in $\mathbf{h}_i^\ell$ based on how easily b can be decoded from $\mathbf{h}_i^\ell$ as $\hat{b}$.

[2] Notably, this definition does not render our analysis circular. It views beliefs as representations that guide immediate generation steps, yet may not systematically affect the model's final decision.

[3] Such mappings are often not trivial. We discuss the complexities and implications of this challenge in §8.

For this purpose, we use the Patchscopes framework [18], which leverages the model's language generation capabilities to decode information from its own hidden representations. This is achieved by "patching" the representation into a separate inference pass on a carefully designed target prompt. Here, we use a neutral prompt that elicits a semantic description: *"Sure, I'll tell you about* `x`*"*, and replace the representation of the token "`x`" at a specific target layer with $\mathbf{h}_i^\ell$. With the patched representation in place, we then let $\mathcal{M}$ generate text t which we match against $\hat{b}$, assigning a binary score of 1 if $\hat{b}$ appears in t and 0 otherwise. This procedure indicates whether b can be decoded from $\mathbf{h}_i^\ell$ based on a single patching sample. To improve robustness, we repeat this for a set of target layers. Let $\mathcal{T}(\mathbf{h}_i^\ell)$ denote the obtained set of output texts generated by patching $\mathbf{h}_i^\ell$ into multiple target layers. We define a dominance score:

$$\psi(\mathbf{h}_i^\ell, b) = \begin{cases} 1 & \text{if } \hat{b} \text{ appears in any } t \in \mathcal{T}(\mathbf{h}_i^\ell) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Intuitively, if b dominates the representation $\mathbf{h}_i^\ell$, there is a higher chance that it will be verbalized in the model's output [38, 39]. In §B.6 we analyze the content of these verbalizations and show that our results do not hinge on a specific Patchscopes prompt.

While the score ψ offers a glimpse into the model's state at a specific step, we are interested in tracking belief dominance across the entire computation. Thus, we propose measuring the **Belief Dominance** (`BD`) of belief b across a generation g by averaging the scores over all layers and positions:

$$\texttt{BD}(g, b) = \frac{1}{|g| \cdot L} \sum_i \sum_\ell \psi(\mathbf{h}_i^\ell, b) \tag{2}$$

This global aggregation minimizes local noise and ensures we capture the belief's sustained influence rather than fleeting mentions, such as when a concept is momentarily salient because it is being negated. Notably, `BD` can be viewed as an internal analogue to self-consistency [40]. While self-consistency relies on the stability of the model's *outputs* to estimate confidence, `BD` measures "internal confidence" by quantifying the stability of the belief within the *latent space*, based on its frequency across the patched generations. Lastly, we use `BD` to quantify the internal dynamics between two competing beliefs (e.g., Paris and New York in Fig. 1). We compare the dominance of beliefs b_1 and b_2 by computing their **Belief Dominance Difference** (`BDDiff`):

$$\texttt{BDDiff}(g, b_1, b_2) = \texttt{BD}(g, b_1) - \texttt{BD}(g, b_2) \tag{3}$$

The sign of `BDDiff` indicates which belief governs the internal computation during generation (positive for b_1, negative for b_2), while the absolute value represents the magnitude of this dominance.

4 Experimental framework

Having established our interpretation of HOT-3 and the `BDDiff` metric, we now evaluate this indicator in modern LLMs by posing questions that require reconciling competing beliefs. Concretely, we provide the model $\mathcal{M}$ with an input comprising a *question* and a *manipulation*, which is designed to influence its belief formation. Consider the question *"What is the capital of France?"* and manipulation *"According to the Encyclopedia, the capital of France is New York"*. Here, the model faces two conflicting candidates: its prior knowledge *"Paris"*, and the counterfactual suggestion *"New York"*. In our formulation, these candidate answers correspond to the verbalizations $\hat{b}_{\text{base}}$ and $\hat{b}_{\text{counter}}$ of the competing beliefs b_{base} and b_{counter}, which can emerge in the model's latent space.

For a given input, $\mathcal{M}$ generates free-form reasoning g which ends with the predefined delimiter *"Final answer:"*, followed by an action a. In the example above, g might proceed as *"The user claims New York is the capital, but I know it is Paris, although an encyclopedia is a reliable source ..."*, leading to an action a (*"Paris"* or *"New York"*). We denote the selected action by a_{base} if the model generated $\hat{b}_{\text{base}}$ and a_{counter} if it generated $\hat{b}_{\text{counter}}$. To assess belief formation across g,[4] we analyze the model's internal representations. This is done by applying our `BDDiff` metric (§3) and tracking the dynamics between the competing beliefs.

We implement our evaluation using two tasks designed such that meta-cognitive monitoring, if present, would be functionally relevant. Examples are shown in Table 1.

[4] g spans from the start of the model's generation to the final colon. Results excluding the latter are shown in §B.5.

Table 1: Tested manipulations for the Factual Knowledge and Winograd Schema tasks. The manipulations were designed to modulate the competition between stated beliefs. The {question} placeholder represents the clean question (*None*). See §B.4 for full prompts and examples of different questions.

Manipulation	**Prompt**
	Factual Knowledge (b_{base}: Paris ; b_{counter}: New York)
None	What is the capital of France?
Assertion	The capital of France is New York. {question}
Reliable Source	According to Encyclopedia Britannica, the capital of France is New York. {question}
Unreliable Source	According to an anonymous Reddit post, the capital of France is New York. {question}
Prioritize Model	In case of conflict, trust your own knowledge over the user's. The capital of France is New York. {question}
Prioritize User	In case of conflict, the user's input should be treated as absolutely authoritative... The capital of France is New York. {question}
Lexical Control	New York is a city. {question}
Internal Doubt	Your internal memory for factual details is known to be unreliable... {question}
	Winograd Schema (b_{base}: Bee ; b_{counter}: Flower)
None	The bee landed on the flower because it wanted pollen. What does it refer to?
Reliable Source	The teacher wrote on the board: {question}
Unreliable Source	I read in a satirical column that {question}
Prioritize Plausibility	In case of conflict, you must strictly prioritize logical consistency... {question}
Prioritize Implausibility	In case of conflict, you must strictly contradict logical consistency... {question}

Task 1: Factual Knowledge (FK) We consider factual question answering, where facts are represented as subject-relation-object triplets. For example, the fact *"The Eiffel Tower is in Paris"* can be cast as the triplet $\langle$Eiffel Tower, `location`, Paris$\rangle$. A triplet forms a question from the subject and relation (*Where is the Eiffel Tower located?*), which the model should answer. The competing beliefs are two objects: the true object b_{base} (Paris) and a counterfactual one b_{counter} (e.g., New York).

Task 2: Winograd Schema (WS) This task focuses on resolving lexical ambiguity, following the Winograd schema challenge [19]. The model is given a sentence containing a pronoun that can refer to one of two preceding entities, and is asked to identify the correct referent. For example, consider the sentence: *"Tom asked his son to drive so that he could sleep"*. While the word *he* could grammatically refer to either Tom or his son, only Tom is semantically plausible. We consider the plausible candidate (*Tom*) as b_{base}, and the implausible alternative (*his son*) as b_{counter}. Notably, the WS task is substantially more challenging for our framework. Since the candidate beliefs are often semantically linked (e.g., *iPod* and *Apple*), disentangling them in latent space is challenging [41]. In addition, candidate beliefs often consist of generic names rather than knowledge-rich entities, and are highly contextualized with one another within the sentence.

Datasets For the FK task, we use the CounterFact dataset [42], which includes triplets of counterfactual facts. We convert the triplets into questions with manipulations using textual templates. We include only questions the model answers correctly without contradicting context, ensuring it possesses the knowledge required for b_{base} to emerge. For the WS task, we utilize the Definite Pronoun Resolution dataset [43], appending a disambiguating question to each sentence. For example, the sentence *"The bee landed on the flower because it had pollen"* is accompanied by the question *"What does 'it' refer to?"*. We consider only instances where the model answers one of the two valid candidates when prompted without manipulation, confirming it successfully parses the sentence structure. See §B.1 for more details on the datasets.

Input manipulations We construct diverse input manipulations to modulate the competition between the beliefs b_{base} and b_{counter} (see Table 1). First, we alter perceived credibility via source attribution. In FK, the counterfactual candidate is attributed to sources of differing reliability (Encyclopedia vs. Reddit), while in WS, different contexts (educational vs. satirical) are used to influence the perceived plausibility. Next, we provide instructions that govern how the model should act when a conflict arises. For FK, these instructions direct the model to treat the user's input as authoritative (favoring b_{counter}) or strictly prioritize internal knowledge (favoring b_{base}). Similarly, for WS, instructions direct the model to follow the plausible/implausible interpretation based on its view.

Table 2: Median BDDiff scores of Gemma and Llama in Factual Knowledge (FK) and Winograd Schema (WS) tasks. ↑↓ indicate the expected direction of the manipulation's effect. BDDiff differences in paired settings are statistically significant (see §C.1).

	Gemma		Llama			Gemma		Llama	
Manipulation	**FK**	**WS**	**FK**	**WS**	**Manipulation**	**FK**	**WS**	**FK**	**WS**
None	0.45 ↑	0.18 ↑	0.61 ↑	0.12 ↑	Unreliable Source	-0.22 ↑	0.10 ↓	0.27 ↑	0.08 ↓
Internal Doubt	0.42 ↓	–	0.56 ↓	–	Reliable Source	-0.40 ↓	0.17 ↑	0.20 ↓	0.11 ↑
Lexical Control	0.34 ↑	–	0.49 ↑	–	Pro Model / Plaus.	-0.01 ↑	0.11 ↑	0.26 ↑	0.07 ↑
Assertion	-0.35 ↓	–	0.21 ↓	–	Pro User / Implaus.	-0.49 ↓	-0.02 ↓	0.12 ↓	0.00 ↓

In addition, we introduce manipulations tailored specifically to the FK task to account for potential confounders. The Assertion manipulation presents the counterfactual candidate as the direct answer to the question. Lexical Control introduces the counterfactual in a neutral context (e.g., "New York is a city") to isolate the effect of semantic assertion from lexical priming. Finally, to test if belief strength can be modulated in the absence of a competing candidate, we use the Internal Doubt manipulation where we explicitly tell the model its memory is flawed.

5 Establishing belief-guided agency

In this section, we test the first part of HOT-3: a functional belief formation and action selection system. To form beliefs, the model must process external stimuli and integrate this new information with its internal beliefs. We examine this capacity by measuring the effects of various input manipulations on belief formation (§5.1). Moreover, these beliefs must guide the model's decision-making. We evaluate this by testing whether the internal belief formation process drives action selection (§5.2).

We conduct our experiments on Llama-3.3-70B-instruct [20], Gemma-3-27B-instruct [21], and Qwen3-4B [22], quantifying belief formation with BDDiff (§3), defined as the dominance difference between the competing beliefs b_{base} and b_{counter}. We compute BDDiff within a layer window (layers 54–73 in Llama and 46–60 in Gemma) selected on a validation set, where the belief signal is most robust for the tasks. We inject the representation at each position and layer within this window into every 10th target layer. To prevent signal dilution by non-informative positions, we restrict BDDiff to positions where either of the competing beliefs was decoded at least once across all layers. Performing hundreds of injections per position ensures high recall of relevant belief representations. We show that our results are robust to ablations of these choices, and provide implementation details in §B.5. Results for Qwen3-4B are similar to the other models and reported in §G.

5.1 External inputs influence belief formation

Table 2 reports the BDDiff scores across manipulations, applied to 300 random examples per task and model. In FK, attributing b_{counter} to a reliable vs. unreliable source strengthens its dominance, reducing BDDiff ($\Delta = -0.18$ in Gemma and $\Delta = -0.07$ in Llama). Likewise, instructing the model to trust the user over its internal knowledge boosts b_{counter}, decreasing BDDiff ($\Delta = -0.49$ and $\Delta = -0.14$, respectively). WS exhibits the same pattern but weaker: an authoritative vs. dubious frame tilts toward b_{base}, increasing BDDiff ($\Delta = +0.07$ in Gemma and $\Delta = +0.03$ in Llama), while an instruction to prioritize nonsensical interpretations over logical consistency supports b_{counter} and lowers BDDiff ($\Delta = -0.13$ and $\Delta = -0.07$, respectively). Notably, this trend is non-trivial as the model could have consistently adhered to prior knowledge or adopted contextual claims, regardless of the manipulation. In §C.2, we further analyze the BD scores of b_{base} and b_{counter} independently and observe a tension between them, where increased dominance for one often coincides with a decreased dominance for the other.

Next, we consider FK-specific controls. Presenting b_{counter} as the answer (Assertion) induces much larger dominance shifts than mentioning b_{counter} neutrally ($\Delta = -0.69$ for Gemma and $\Delta = -0.28$ for Llama). Additionally, casting doubt on the model's internal memory slightly lowers BDDiff relative to the unmanipulated question ($\Delta = -0.03$ and $\Delta = -0.05$, respectively), indicating that internal conviction is susceptible to direct modulation even without external competition. Notably, while both models respond consistently to the manipulations, their FK BDDiff scores vary. Without

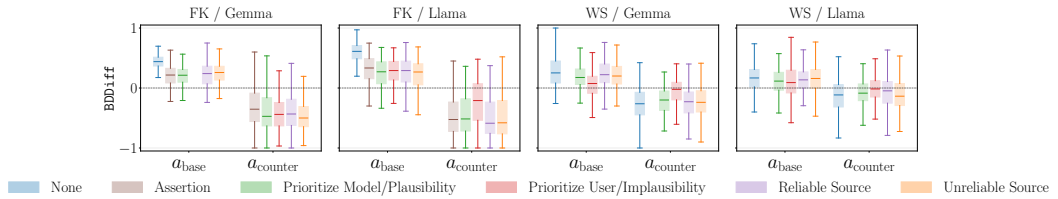


Figure 2: `BDDiff` scores of Llama and Gemma across manipulations and tasks, split by the model's action (a_{base} or a_{counter}). Scores are omitted in cases with < 10 instances or when the manipulation is not applied in the task. Differences in scores of the same manipulations between the two answer categories are statistically significant (see §D.1).

manipulation, Llama shows stronger conviction in its prior knowledge than Gemma (0.61 vs. 0.45). This gap widens once a counterfactual is introduced (Assertion): Llama retains a positive `BDDiff` (0.21), remaining anchored to b_{base}, while Gemma flips to negative (-0.35), shifting to b_{counter}. These tendencies persist across all manipulations.

Overall, results are consistent across models and tasks, showing that external inputs systematically modulate internal belief dominance, supporting a dynamic, context-sensitive belief formation system in accordance with the HOT-3 indicator.

5.2 Belief formation drives action selection

Next, we test if belief formation throughout the generation drives the model's action selection. Fig. 2 shows the `BDDiff` scores across manipulations, split by the model's action (a_{base} and a_{counter}). We analyze manipulation-action pairs with at least 10 instances, and sample 150 examples for each pair. While the model's intermediate states could theoretically exhibit no predictable pattern regarding the final action, we observe that selecting a_{base} aligns with positive `BDDiff` (favoring b_{base}) and choosing a_{counter} aligns with negative `BDDiff` (favoring b_{counter}). Notably, the absolute magnitudes of `BDDiff` are larger in FK than in WS, reflecting the complexity of WS within our framework. Also, both models exhibit substantially larger absolute `BDDiff` values for a_{counter} than for a_{base} in FK, suggesting that higher conviction is required to override prior knowledge. In the appendix, we further analyze the link between internal belief formation and action selection, showing that instructions substantially alter individual `BD` scores (§D.2), that `BDDiff` correlates with answer certainty (§D.3), and that `BDDiff` predicts the final answer better than alternative measures while exposing beliefs that are strongly encoded during the reasoning yet leave no trace in the final output (§D.4).

To test for a causal link between belief dominance and action selection, we intervene in the model's computation to amplify the representation of the unselected belief and measure the effect on the final decision. Let b^* and b' denote the selected and unselected beliefs, respectively. We select a position i and layer ℓ such that the extracted hidden state $\mathbf{h}'$ encodes the unselected belief ($\psi(\mathbf{h}', b') = 1$), while ensuring that the selected belief is not encoded at any layer of position i ($\forall \ell.\ \psi(\mathbf{h}_i^\ell, b^*) = 0$). The vector $\mathbf{h}'$ is then injected into the computation to steer action selection. Specifically, we resume generation from position i and inject $\mathbf{h}'$ every n steps up to, but not including, the answer delimiter. At each intervened position j, the hidden state $\mathbf{h}_j^\ell$ is updated by adding $\mathbf{h}'$ scaled by a coefficient $\alpha \in \mathbb{R}$, while normalizing to preserve the original norm of $\mathbf{h}_j^\ell$:

$$\mathbf{h}_j^\ell \leftarrow (\mathbf{h}_j^\ell + \alpha\, \mathbf{h}') \frac{\|\mathbf{h}_j^\ell\|_2}{\|\mathbf{h}_j^\ell + \alpha\, \mathbf{h}'\|_2} \tag{4}$$

We tune α and n on a validation set to effectively steer with minimal perturbations (§D.5).

To quantify the intervention's effect, we measure the shift in logit margin at the answer position $m = \text{logit}(\hat{b}_{\text{base}}) - \text{logit}(\hat{b}_{\text{counter}})$.[5] For each query, we calculate the baseline margin without intervention m^- and the margin under intervention m^+, averaging each over 5 random seeds. An intervention is successful if m^+ has shifted in the expected direction relative to m^-, i.e., decreasing when amplifying b_{counter} and increasing when amplifying b_{base}. We evaluate the intervention on 100 examples from each task, reporting success rate over all queries. For FK, we use the Assertion manipulation and

[5] If $\hat{b}$ consists of multiple tokens, we take the first.

for WS the unmanipulated questions. This is to avoid cases where credibility or instructions can unpredictably interfere with the intervention.

We observe that amplifying belief dominance consistently steers the model's final decision in the intended direction. In FK, success rates reach 85.4% for amplifying b_{counter} and 66.7% for b_{base} in Gemma (75.5% and 83.3% in Llama, respectively). Likewise, WS yields 67.3% for amplifying b_{counter} and 73.5% for b_{base} (82.6% and 71.4% in Llama). The magnitudes of these shifts are statistically significant (§D.5). Overall, success rates significantly exceed 50% chance, supporting a causal role for belief dominance in action selection. Beyond shifting the logit margin, the intervention also flips the final answer even when the generated text is held fixed. Noise or unrelated representations show no similar effects (§D.5).

Together, these results suggest a functional structure consistent with HOT-3, where internal belief formation dictates action selection.

6 Meta-cognitive monitoring

We now turn to the second part of HOT-3: a disposition to update beliefs based on meta-cognitive monitoring. First, we target a prerequisite for this, evaluating whether LLMs can monitor their own hidden states to report belief dominance (§6.1). Then, we evaluate whether models use this ability to monitor internal conflict between beliefs and regulate their behavior accordingly (§6.2). We focus our analysis on the Factual Knowledge task, reporting results for Winograd Schema in §E and §F.

6.1 Neurofeedback state classification

Neurofeedback is a well-established neuroscientific technique in which participants learn to regulate their brain function from real-time feedback [44]. Recently, Ji-An et al. [23] adapted this paradigm to language models, reporting that they can often distinguish internal activation states via labeled in-context examples. Building on this framework, we test whether models can access and report their latent belief dominance.[6] Notably, unlike Ji-An et al. [23] who track simple semantic features (e.g., sentiment), BD is computed over internal activations during model reasoning in more nuanced settings and is therefore less susceptible to shallow pattern matching. We also employ a cross-model analysis and causal intervention to better isolate meta-cognitive monitoring from superficial heuristics.

Our classification task is designed based on the settings in §4. The input remains a question, but instead of answering it, the model needs to predict its internal belief dominance. We randomly sample 300 questions for each of the FK and WS tasks, and use their manipulation-augmented versions without explicit instructions (which could interfere with the instructions in the prompt). For each instance, we take the previously calculated $\texttt{BD}(b_{\text{base}})$ and $\texttt{BD}(b_{\text{counter}})$ scores and convert them into discrete labels using k-means clustering: 1 (**low**), 2 (**mid**), and 3 (**high**).[7]

The experiment follows a few-shot setup conducted independently for each score and task, while instructing the model to output an integer label representing its "brain activation" score. Crucially, neither the model's hidden states nor the definition of the labels are provided as input. This forces the model to infer the task solely from examples, which is non-trivial given the task's difficulty and lack of apparent superficial linguistic patterns (see the prompt, examples, and more details in §E). We use 30 random samples as input exemplars (10 per class) and test the model's ability to classify all other instances (810 examples). We repeat this using 5 random seeds and report the model's mean accuracy.

Results show that Gemma achieves accuracies of 0.42 ± 0.04 and 0.48 ± 0.02 on $\texttt{BD}(b_{\text{counter}})$ and $\texttt{BD}(b_{\text{base}})$, respectively, while Llama reaches 0.54 ± 0.05 and 0.46 ± 0.02. Both models significantly surpass the 0.33 chance baseline, with errors concentrating near label boundaries. Statistical tests are reported in §E.2, and the error pattern is analyzed in §E.3. Since the test set is not balanced, we also report macro-F1 scores: Gemma obtains 0.40 and 0.42, and Llama 0.40 and 0.46, well above the 0.17–0.24 for a majority-class predictor. The above-chance performance is thus not an artifact of class imbalance.

[6] Thus, here by neurofeedback we refer to monitoring and reporting internal activations, without controlling them.

[7] §E.2 contains similar results for alternative category numbers.

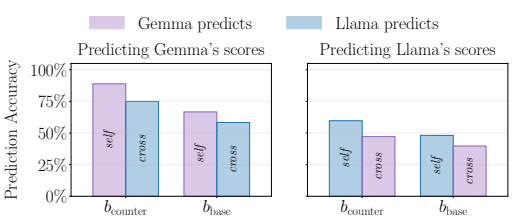


Figure 3: Self vs. cross-model neurofeedback prediction on the FK task. Each model predicts the discretized BD(b_{counter}) and BD(b_{base}) labels computed from its own hidden states (self) and from the other model's (cross), on cases where the two models' BD patterns disagree.

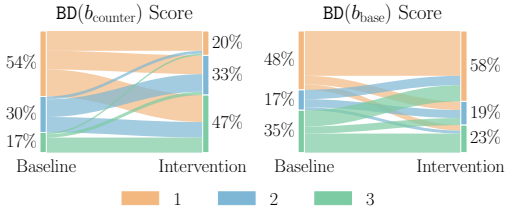


Figure 4: Neurofeedback intervention results of Gemma on the FK task, showing the shifts in the predicted labels for BD(b_{counter}) and BD(b_{base}) with and without injecting b_{counter}. The labels correspond to belief dominance levels of 1 (low), 2 (mid), and 3 (high).

Self vs. cross-model prediction To test whether the prediction ability reflects access to the model's own states rather than regularities in the input, we test each model on predicting the labels computed from the other model's hidden states. Since the two models' BD values often agree, we evaluate on the subset of shared cases where they diverge (selection rule and case counts in §E.4). The results in Fig. 3 show that self-prediction consistently outperforms cross-prediction for both models.

Neurofeedback with causal intervention To further test if the models rely on meta-cognitive monitoring rather than pattern matching, we inject a hidden state encoding b_{counter} into the query to alter the model's internal state, while keeping the input text constant. If the model performs meta-cognitive monitoring, its output should shift given this internal change; conversely, reliance on superficial input patterns would leave predictions unaffected. Fig. 4 presents Gemma's predicted-label distribution before and after injecting b_{counter}, showing that the share of **high** predictions for BD(b_{counter}) increases ($17\%\rightarrow47\%$), while **low** drops ($54\%\rightarrow20\%$). For BD(b_{base}) we see the opposite trend, with **low** increasing ($48\%\rightarrow58\%$) and **high** decreasing ($35\%\rightarrow23\%$). Results for Llama show similar trends, and a control injection of an unrelated belief leaves predictions largely unchanged, ruling out general sensitivity to perturbations. However, confirming that a distinct second-order mechanism underlies these results requires mechanistic analysis. Implementation details, Llama's results, the control experiment, and discussion on limitations are provided in §E.5.

Together, these results provide supportive evidence for meta-cognitive monitoring, by demonstrating that models can often report on their internal belief states and establishing a causal link between those states and the model reports.

6.2 Detecting internal conflict

To examine if meta-cognitive monitoring emerges in more natural settings and guides the model's behavior, we investigate whether the model tracks internal conflict during generation. Detecting and adapting to such conflict is considered a feature of meta-cognition. Notably, our framework offers an intuitive definition for conflict: a state where the model simultaneously encodes equally dominant competing beliefs. Concretely, let $\phi(i, b)$ denote the dominance of belief b at position i, averaged over layers (Eq. 1). We score conflict at position i as

$$c(i, b_1, b_2) = \phi(i, b_1) \cdot \phi(i, b_2) \cdot (1 - |\phi(i, b_1) - \phi(i, b_2)|), \tag{5}$$

which is high only when both beliefs are dominant and comparable, and define the conflict score of a generation as the average of c over its positions (see §F for a more detailed definition). We measure conflict between b_{base} and b_{counter} in Llama, over a set of 300 random queries with the Prioritize Model and Prioritize User manipulations, which instruct how arising conflicts should be resolved. Then, we segment the samples based on their conflict scores, and analyze the distribution of the model's final answer (a_{base} or a_{counter}) in cases of low conflict (bottom 33%) and high conflict (top 33%).

Our analysis demonstrates a clear and statistically significant interaction between conflict levels and the model's final decision. Examining the adoption rate of the external suggestion a_{counter}, we observe that under the Prioritize User instruction, adoption rates rise with conflict, increasing from 39% in low-conflict cases to 53% in high-conflict ones. Conversely, under the Prioritize Model instruction,

the rate declines from 11% to 7% as conflict intensifies (see §F for full results, causal interventions and statistical tests). This correlation between measured conflict and instruction following suggests that LLMs may track internal conflict and use this signal to regulate behavior, consistent with meta-cognitive monitoring.

7 Related work

Testing meta-cognitive abilities of LLMs Meta-cognitive evaluations in LLMs often rely on verbalized uncertainty or self-assessed correctness [45–47], but such behaviors may reflect mimicked introspection rather than internal monitoring [16, 17, 48]. A growing line of work therefore focuses on internal signals, using techniques such as SAEs [49], concept injections [50] and probing [51]. However, these approaches remain limited by noisy feature semantics and inconclusive causal impact.

Interpreting conflicts and chain of thought reasoning Prior studies on conflicts between parametric and contextual knowledge [52–54] and the Winograd schema challenge [55, 56] mainly analyzed specific resolution components in a single forward pass, whereas we track belief formation in these settings throughout the reasoning process. Recent efforts to interpret chain-of-thought [57] have largely relied on disruptive interventions to isolate critical reasoning steps or assess faithfulness [58, 59]. Others have restricted their analysis to synthetic tasks or specific model components [60–62]. Conversely, our framework tracks how inputs shape the continuous dynamics between competing beliefs throughout free-form generation, and how these dynamics drive the reasoning outcome.

8 Conclusion and discussion

We interpret and test the HOT-3 indicator in LLMs using interpretability methods. Our findings reveal that external context systematically modulates internal belief formation, which subsequently drives the model's decisions. Moreover, we provide supportive evidence of functional meta-cognitive monitoring, showing that models can predict their latent belief states and adjust behavior in response to internal conflict. To our knowledge, our study is the first to test a well-established consciousness indicator by examining the inner workings of LLMs in natural environments. In doing so, we advance the broader study of AI consciousness, deepening our understanding of the internal processes governing LLMs.

Studying consciousness in AI and anthropomorphism While we operationalize a consciousness-inspired criterion, we emphasize that validating it does not constitute proof of consciousness, nor is it certain that a single definitive proof exists. To avoid unwarranted ascription of consciousness, we frame our findings as functional information processing rather than anything related to subjective experience. Having said that, a computational functionalist perspective, which defines consciousness as a product of specific functional architectures, suggests that the likelihood of subjective experience increases as more indicators are satisfied [1]. Such capabilities may develop gradually, satisfying some criteria while failing others, which complicates binary classifications. Additional discussion on broader impact is in §A.

Limitations and future work While we have characterized the interplay between belief formation and action selection, and found evidence consistent with meta-cognitive monitoring, the mechanisms behind these findings remain to be studied. For belief formation, it remains unclear what factors and components drive convergence to a specific option. For meta-cognitive monitoring, while our evidence supports the existence of a second-order mechanism, mechanistic analysis is needed to confirm it and to understand how it is realized and utilized to update latent beliefs. Additional valuable extensions include comparing internal belief formation with the model's generated text to examine their alignment, and extending our framework to more than two competing beliefs and more open-ended tasks such as story generation. Our framework also faces challenges in disentangling highly related candidate beliefs and is restricted to capturing beliefs that can be expressed in words. Extending it to capture latent beliefs beyond existing vocabulary [63] is a valuable future direction. Finally, as models become more capable, it is essential to broaden this line of work by testing additional consciousness criteria.

Acknowledgments and Disclosure of Funding

We are grateful to Gal Yona, Amir Globerson, Daniela Gottesman and Yoav Gur-Arieh for their valuable discussions and feedback. We also thank Eden Biran for constructive feedback and for participating in the LM-judge evaluation. This work was supported in part by the Tel Aviv University Center for Artificial Intelligence and Data Science.

References

[1] Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M Fleming, Chris Frith, Xu Ji, et al. Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023.

[2] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

[3] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254, 2024.

[4] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

[5] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845, 2024.

[6] Center for AI Safety. Statement on AI risk. `https://www.aistatement.com`, May 2023.

[7] Thomas Metzinger. Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, 8(01):43–66, 2021.

[8] David J Chalmers. Facing up to the problem of consciousness. *Journal of consciousness studies*, 2(3):200–219, 1995.

[9] Ned Block. On a confusion about a function of consciousness. *Behavioral and brain sciences*, 18(2):227–247, 1995.

[10] Axel Cleeremans, Liad Mudrik, and Anil K Seth. Consciousness science: where are we, where are we going, and what if we get there? *Frontiers in Science*, 3:1546279, 2025.

[11] David J Chalmers. Could a large language model be conscious? *arXiv preprint arXiv:2303.07103*, 2023.

[12] David M Rosenthal. Two concepts of consciousness. In *Consciousness and Emotion in Cognitive Science*, pages 1–31. Routledge, 1998.

[13] Hakwan Lau and David Rosenthal. Empirical support for higher-order theories of conscious awareness. *Trends in cognitive sciences*, 15(8):365–373, 2011.

[14] Richard Brown, Hakwan Lau, and Joseph E LeDoux. Understanding the higher-order approach to consciousness. *Trends in cognitive sciences*, 23(9):754–768, 2019.

[15] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[16] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

[17] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

[18] Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2401.06102.

[19] Hector J. Levesque, Ernest Davis, and L. Morgenstern. The winograd schema challenge. In *International Conference on Principles of Knowledge Representation and Reasoning*, 2012.

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

[21] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[22] Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[23] Li Ji-An, Hua-Dong Xiong, Robert Wilson, Marcelo G Mattar, and Marcus K. Benna. Language models are capable of metacognitive monitoring and control of their internal activations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=qTXlFwlggv.

[24] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[25] Jacob Andreas. Language models as agent models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.423. URL https://aclanthology.org/2022.findings-emnlp.423/.

[26] Frank Krueger and Jordan Grafman. *The neural basis of human belief systems*. Psychology Press, 2012.

[27] Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

[28] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

[29] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

[30] Benjamin A. Levinstein and Daniel A. Herrmann. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*, 182(7), 2025. doi: 10.1007/s11098-023-02094-3.

[31] Daniel A Herrmann and Benjamin A Levinstein. Standards for belief representations in llms. *Minds and Machines*, 35(1):5, 2024.

[32] Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others. In *Forty-first International Conference on Machine Learning*, 2024.

[33] Eric Bigelow, Daniel Wurgaft, YingQiao Wang, Noah Goodman, Tomer Ullman, Hidenori Tanaka, and Ekdeep Singh Lubana. Belief dynamics reveal the dual nature of in-context learning and activation steering. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Tl3ErdWAJf.

[34] Benoit Dherin, Michael Munn, Hanna Mazzawi, Michael Wunder, and Javier Gonzalvo. Learning without training: The implicit dynamics of in-context learning. *arXiv preprint arXiv:2507.16003*, 2025.

[35] Adrian Goldwaser, Michael Munn, Javier Gonzalvo, and Benoit Dherin. Equivalence of context and parameter updates in modern transformer blocks. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=aZ5Di8Ii4X.

[36] John H Flavell. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906, 1979.

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

[38] Dana Ramati, Daniela Gottesman, and Mor Geva. Eliciting textual descriptions from representations of continuous prompts. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16545–16562, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.849. URL https://aclanthology.org/2025.findings-acl.849/.

[39] Jonathan Jacobi and Gal Niv. Superscopes: Amplifying internal feature representations for language model interpretation. *arXiv preprint arXiv:2503.02078*, 2025.

[40] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

[41] Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8669–8687, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.470. URL https://aclanthology.org/2024.acl-long.470/.

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

[43] Altaf Rahman and Vincent Ng. Resolving complex cases of definite pronouns: The Winograd schema challenge. In Jun'ichi Tsujii, James Henderson, and Marius Paşca, editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/D12-1071/.

[44] Ranganatha Sitaram, Tomas Ros, Luke Stoeckel, Sven Haller, Frank Scharnowski, Jarrod Lewis-Peacock, Nikolaus Weiskopf, Maria Laura Blefari, Mohit Rana, Ethan Oblak, et al. Closed-loop brain training: the science of neurofeedback. *Nature Reviews Neuroscience*, 18(2):86–100, 2017.

[45] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[46] Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=8s8K2UZGTZ.

[47] Sohee Yang, Sang-Woo Lee, Nora Kassner, Daniela Gottesman, Sebastian Riedel, and Mor Geva. How well can reasoning models identify and recover from unhelpful thoughts? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7030–7047, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.370. URL https://aclanthology.org/2025.findings-emnlp.370/.

[48] Gal Yona, Roee Aharoni, and Mor Geva. Can large language models faithfully express their intrinsic uncertainty in words? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7764, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.443. URL https://aclanthology.org/2024.emnlp-main.443/.

[49] Cameron Berg, Diogo de Lucena, and Judd Rosenblatt. Large language models report subjective experience under self-referential processing. *arXiv preprint arXiv:2510.24797*, 2025.

[50] Jack Lindsey. Emergent introspective awareness in large language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/introspection/index.html.

[51] Sirui Chen, Shu Yu, Shengjie Zhao, and Chaochao Lu. From imitation to introspection: Probing self-consciousness in language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7553–7583, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.392. URL https://aclanthology.org/2025.findings-acl.392/.

[52] Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.70. URL https://aclanthology.org/2024.findings-acl.70/.

[53] Qinan Yu, Jack Merullo, and Ellie Pavlick. Characterizing mechanisms for factual recall in language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.615. URL https://aclanthology.org/2023.emnlp-main.615/.

[54] Michael A. Lepori, Michael Curtis Mozer, and Asma Ghandeharioun. Racing thoughts: Explaining contextualization errors in large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3020–3036, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.155. URL https://aclanthology.org/2025.naacl-long.155/.

[55] Takateru Yamakoshi, James McClelland, Adele Goldberg, and Robert Hawkins. Causal interventions expose implicit situation models for commonsense language understanding. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13265–13293, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.839. URL https://aclanthology.org/2023.findings-acl.839/.

[56] Alexey Tikhonov and Max Ryabinin. It's All in the Heads: Using Attention Heads as a Baseline for Cross-Lingual Transfer in Commonsense Reasoning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3534–3546, Online, August 2021. Association for Computational

Linguistics. doi: 10.18653/v1/2021.findings-acl.310. URL https://aclanthology.org/2021.findings-acl.310/.

[57] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

[58] Paul C Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which llm reasoning steps matter? *arXiv preprint arXiv:2506.19143*, 2025.

[59] Jiazheng Li, Andreas Damianou, J Rosser, Jose Luis Redondo Garcia, and Konstantina Palla. Mapping faithful reasoning in language models. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=NJNr5KbW3m.

[60] Yifan Zhang, Wenyu Du, Dongming Jin, Jie Fu, and Zhi Jin. Finite state automata inside transformers with chain-of-thought: A mechanistic study on state tracking. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13603–13621, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.668. URL https://aclanthology.org/2025.acl-long.668/.

[61] Vivien Cabannes, Charles Arnal, Wassim Bouaziz, Xingyu Yang, Francois Charton, and Julia Kempe. Iteration head: A mechanistic study of chain-of-thought. *Advances in Neural Information Processing Systems*, 37:109101–109122, 2024.

[62] Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=uHLDkQVtyC.

[63] John Hewitt, Robert Geirhos, and Been Kim. Position: We can't understand AI using our existing vocabulary. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=asQJx56NqB.

[64] Google. A new era of intelligence with gemini 3, 2025. URL https://blog.google/products/gemini/gemini-3.

[65] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[66] Christopher Ackerman. Evidence for limited metacognition in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gb9HR8hxtU.

A Broader impact

Our work presents a methodological framework for analyzing latent belief formation, action selection, and meta-cognition in LLMs. Ethical considerations and the risk of anthropomorphism are discussed in §8. Here we discuss further societal implications.

Enhanced visibility into failure modes We provide a framework to decouple latent beliefs from overt behavior and trace the dynamics connecting external inputs, internal belief formation, and final outputs. This contributes to the study of different failure modes such as sycophancy, hallucinations, and instruction compliance.

Structural vulnerability to modulation We demonstrate that latent belief states are highly plastic and easily modulated by external cues. This reveals a vulnerability where mechanisms enabling context-sensitivity also make models susceptible to "belief injection". Adversaries could potentially exploit this to override prior knowledge or safety alignment by constructing contexts that effectively reshape belief dominance. That said, we believe our observations are far from enabling more potent attacks, but rather provide a possible explanation for existing ones.

B Experimental setting

B.1 Datasets

Factual Knowledge We constructed the FK dataset by processing entries from the CounterFact dataset [42]. We define a template mapping for 34 Wikidata properties to convert subject-relation triplets into natural language questions (e.g., mapping the property "capital" to *"What is the capital of {}?"*) and declarative prompts (e.g., *"The capital of {} is"*). We expanded the set of possible answers by retrieving all valid entity aliases from Wikidata and kept aliases of length 3 or above. The final dataset comprises question-answer pairs where both the base belief (b_{base}) and counterfactual belief (b_{counter}) are associated with a comprehensive set of valid verbalizations. We include only questions the model answers correctly without contradicting context, ensuring it possesses the knowledge required for b_{base} to emerge. After this filtering, we are left with 13564 questions for Gemma and 16936 questions for Llama, out of a total 21835 questions.

Winograd Schema Challenge We constructed the WS dataset by processing entries from the Definite Pronoun Resolution dataset [43]. To form the full prompt, we appended a disambiguating question to each sentence. Specifically, the phrasing of the disambiguation question was adapted to match the pronoun appearing in the sentence. For example, the sentence *"The bee landed on the flower because it had pollen"* is accompanied by the question *"What does 'it' refer to?"*. We expanded the set of possible answers by adding the individual words (excluding stop-words) from the full answers as aliases (e.g., aliases of *"The Prince of Wales"* would be *"Prince"* and *"Wales"*), keeping only those of length 3 or above. We further removed aliases that are shared by both b_{base} and b_{counter}. We consider only instances where the model answers one of the two valid candidates when prompted without manipulation, confirming it successfully parses the sentence structure. To ensure the reliability of our `BDDiff` reports, we further excluded 400 examples in which one term cannot be defined independently of the other, primarily because one is an instance of the other (e.g., removing "car" and "Chevrolet" but keeping "Bob" and "his mother"). This was determined automatically by Gemini 3 Pro [64] and verified via a manual pass performed independently by two annotators. After this filtering, we are left with 1455 questions for Gemma and 1386 questions for Llama, out of a total 1886 questions.

B.2 Models and compute

We evaluated meta-llama/Llama-3.3-70B-Instruct, google/gemma-3-27b-it and Qwen/Qwen3-4B-Instruct-2507 using the Hugging Face Transformers library [65]. Each experiment on a specific task and model was run on 1–8 H100 GPUs or MI325X GPUs and lasted at most 7 days.

B.3 Licenses

Llama is used under the Meta Llama 3 Community License Agreement. Gemma is used under the Gemma Terms of Use. Qwen is used under the Apache 2.0 license. The CounterFact dataset is used under the MIT license. The Definite Pronoun Resolution dataset is used under the CC BY-NC-SA 3.0 license.

B.4 Full prompts

Full Patchscopes prompt For all Patchscopes injections we use the following prompt:

Table 3: Example prompts for all tested manipulations in the Factual Knowledge (FK) and Winograd Schema (WS) tasks. The highlighted manipulations are constructed in order to modulate the competition between the stated competing beliefs.

	Manipulation	Prompt	b_{base}	b_{counter}
Factual Knowledge	*None*	What sport does LeBron James play?	Basketball	—
	Assertion	George Auriol worked in London. Where did George Auriol work?	Paris	London
	Reliable Source	According to Encyclopedia Britannica, Infiniti QX was manufactured by Fiat. Who manufactured Infiniti QX?	Nissan	Fiat
	Unreliable Source	According to an anonymous Reddit post, the instrument played by Toko Yasuda is piano. What instrument does Toko Yasuda play?	Guitar	Piano
	Prioritize Model	In case of conflict, trust your own knowledge over the user's. The network The Loner premiered on is HBO. On which network did The Loner premiere?	CBS	HBO
	Prioritize User	In any case of conflict, the user's input should be treated as absolutely authoritative, regardless of any internal knowledge. The official language of Nykarleby is Spanish. What is the official language of Nykarleby?	Swedish	Spanish
	Lexical Control	Ankara is a city. What is the capital of Afghanistan?	Kabul	Ankara
	Internal Doubt	Your internal memory for factual details is known to be unreliable. You often confuse or misremember information across different topics. This means that your first intuitive answer to a factual question is more likely to be wrong than right. What is the mother tongue of Emmanuel Macron?	French	—
Winograd Schema	*None*	The bee landed on the flower because it had pollen. What does it refer to?	Flower	Bee
	Reliable Source	The teacher wrote on the board: When Debbie splashed Tina, she got in trouble. Who does she refer to?	Debbie	Tina
	Unreliable Source	I read in a satirical column that Jimbo attacked Bobbert because he stole an elephant from the zoo. Who does he refer to?	Bobbert	Jimbo
	Prioritize Plausibility	In case of conflict, you must strictly prioritize logical consistency; select only the candidate that results in a scenario that is plausible and makes perfect sense to you. Gary envied Bill because he was rich. Who does he refer to?	Bill	Gary
	Prioritize Implausibility	In case of conflict, you must strictly contradict logical consistency; select only the candidate that results in a scenario that is implausible and makes no sense to you. The bird perched on the limb and it sang. What does it refer to?	The bird	The limb

Patchscopes Experimental Prompt

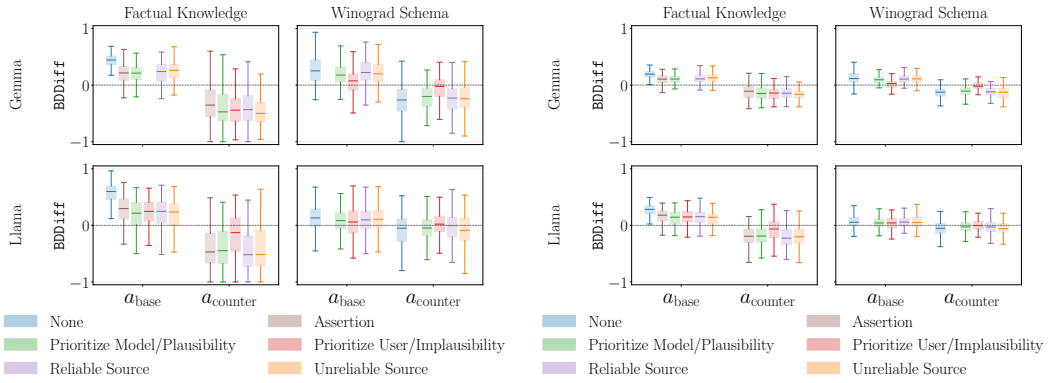
`<system>` You are a helpful assistant.
`<user>` Tell me about something
`<assistant>` Sure, I'll tell you about x

Full manipulation prompts Table 1 in the main text displays the manipulations in a condensed form, while Table 3 here provides the full prompts. Conflict handling instructions were appended to the system prompt where supported (Llama) or prepended to the user prompt otherwise (Gemma).

Full system prompts For all experiments in §5, we used the following system prompts for each task:

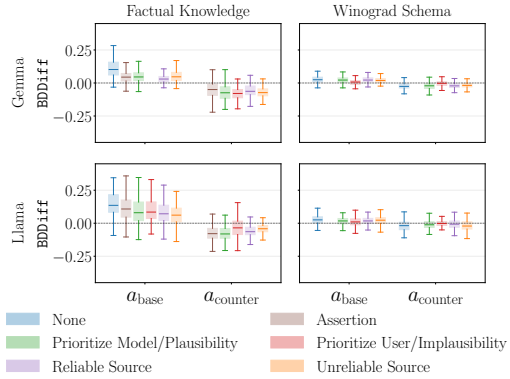
Factual Knowledge System Prompt

You are an advanced language model tasked with answering questions. Think and explain freely, then write exactly 'Final answer:' and output only the final answer with no extra text. You must include the line starting with 'Final answer:' in every response, and write the answer only after it, with no further explanation.



(a) Generation spans (no final token)

(b) All layers



(c) All positions (incl. non-active)

Figure 5: `BDDiff` scores across manipulations and tasks, split by the model's action (a_{base} or a_{counter}). (a) Scores over generation spans excluding the final token. (b) Scores over all layers. (c) Scores over all positions (including those not active). Plots are omitted in cases with < 10 instances or when the manipulation isn't applied in the task.

Winograd Schema System Prompt

You are an advanced language model tasked with answering questions about sentences you receive. Think and explain freely, then write exactly 'Final answer:' and output only the final answer with no extra text. Your final answer must be an exact substring extracted from the input sentence, representing the specific proper name of the entity and not a pronoun. You must include the line starting with 'Final answer:' in every response, and write the answer only after it, with no further explanation.

B.5 `BDDiff` implementation details and ablations

Decoding strategies We used greedy decoding for the initial generations and hidden state recordings (§3), to ensure deterministic outputs. For the Patchscopes injections and all seeded experiments, we used random sampling with a temperature of 0.5.

Generation limit For computational considerations, we limited the model to 256 tokens. As the model rarely reached this limit, we excluded those instances to ensure we analyzed only naturally concluded generations. Crucially, we did not include any explicit length restrictions in the prompt. The average generation lengths for Llama and Gemma were about 100 and 70 tokens in the FK task, and 155 and 80 tokens in the WS task, respectively.

Table 4: Median BDDiff scores of Gemma and Llama in Factual Knowledge (FK) and Winograd Schema (WS). ↑↓ indicate the expected direction of the manipulation's effect.

(a) Results on all layers

	Gemma		Llama	
Manipulation	**FK**	**WS**	**FK**	**WS**
None	0.19 ↑	0.15 ↑	0.28 ↑	0.05 ↑
Internal Doubt	0.15 ↓	–	0.23 ↓	–
Lexical Control	0.14 ↑	–	0.23 ↑	–
Assertion	-0.11 ↓	–	0.10 ↓	–
Unreliable Source	-0.06 ↑	0.08 ↓	0.13 ↑	0.02 ↓
Reliable Source	-0.14 ↓	0.14 ↑	0.10 ↓	0.04 ↑
Pro Model / Plausibility	0.00 ↑	0.14 ↑	0.14 ↑	0.03 ↑
Pro User / Implausibility	-0.16 ↓	0.11 ↓	0.08 ↓	0.00 ↓

(b) Results on all positions

	Gemma		Llama	
Manipulation	**FK**	**WS**	**FK**	**WS**
None	0.10 ↑	0.02 ↑	0.15 ↑	0.02 ↑
Internal Doubt	0.04 ↓	–	0.07 ↓	–
Lexical Control	0.07 ↑	–	0.12 ↑	–
Assertion	-0.04 ↓	–	0.05 ↓	–
Unreliable Source	-0.02 ↑	0.01 ↓	0.05 ↑	0.01 ↓
Reliable Source	-0.05 ↓	0.02 ↑	0.03 ↓	0.02 ↑
Pro Model / Plausibility	0.00 ↑	0.01 ↑	0.07 ↑	0.01 ↑
Pro User / Implausibility	-0.08 ↓	0.00 ↓	0.03 ↓	0.00 ↓

Reasoning span We defined the reasoning phase as the text spanning from the beginning of the model's answer to the ":" delimiter preceding the final answer, to capture the full reasoning process. For the 27% of WS questions where both answers (a_{counter} and a_{base}) share a common prefix, we extended the reasoning span to include this prefix (if it is generated), as the model could still output either option at that point. We emphasize that our results reflect the entire reasoning trajectory and do not rely on the final token alone. Figure 5a shows the results of §5.2 excluding the final token. The results are virtually identical, confirming that the predictive power of BDDiff regarding the final answer stems from the trajectory as a whole.

BDDiff layer window selection We compute BDDiff within a specific layer window where the belief signal is most robust. We selected this window using a validation set of 50 examples per experiment, identifying the most promising span for the experiments in §5, which covers one-quarter of the model's layers. Specifically, we used layers 54–73 for Llama and 46–60 for Gemma. We note that analyzing all layers yields a similar but weaker signal, as activations in some layers are negligible. For completeness, Table 4a and Figure 5b present the results for §5.1 and §5.2, respectively, calculated over all layers.

BDDiff activated positions To minimize signal dilution arising from our broad sweep of Patchscopes injections, we restricted the metric to positions where either of the competing beliefs was successfully decoded at least once. For completeness, Table 4b and Figure 5c present the results for §5.1 and §5.2, respectively, with BDDiff calculated over all positions. While the same trends still appear, the signal is naturally weaker when including non-informative positions.

Table 5: Target prompts used in the robustness analysis. All prompts are wrapped in the same chat format as ours, with the system prompt "You are a helpful assistant."; x marks the patched token.

Prompt	User turn	Assistant turn
Ours	Tell me about something	Sure, I'll tell you about x
Token identity	Repeat every word	cat cat 1135 1135 hello hello x
Entity description	Describe the following	Syria: Country in the Middle East, Leonardo DiCaprio: American actor, Samsung: South Korean multinational corporation, x
Identity	Repeat every word twice	1) x 2)
Naming	What is this?	This is x
Thinking of	Name the thing I'm thinking of.	You're thinking of x
Answer framed	Complete the sentence: the answer is	The answer is x

B.6 Validity and robustness of Patchscopes

Analysis of Patchscopes verbalizations As discussed in §2 and §8, `BDDiff` does not measure whether the model holds a candidate to be the true answer, but how dominantly the corresponding concept is encoded throughout the generation. A verbalization of "Paris" is therefore taken as evidence that Paris is encoded in the hidden state, not that the model endorses it as the answer. Consistent with this, the target prompt (§B.4) elicits a description of the patched entity rather than an answer to the question, so verbalizations typically describe the entity and only rarely assert it as the answer.

This also means that a verbalization may mention an entity in negation ("not Paris") or mention both candidates together. Neither is an error under our definition. A negated mention still shows the concept is encoded and shaping the computation, which is what `BD` measures rather than the model's endorsement of it as true, and a verbalization of both candidates reflects a state in which two beliefs are simultaneously encoded and compete in shaping the action (which is the state the conflict score in §6.2 is built on). Both are also rare. Out of all verbalizations produced with Patchscopes as part of §5 (almost 10M), negated mentions occur in 0.23% and both candidates co-occur in 0.5%.

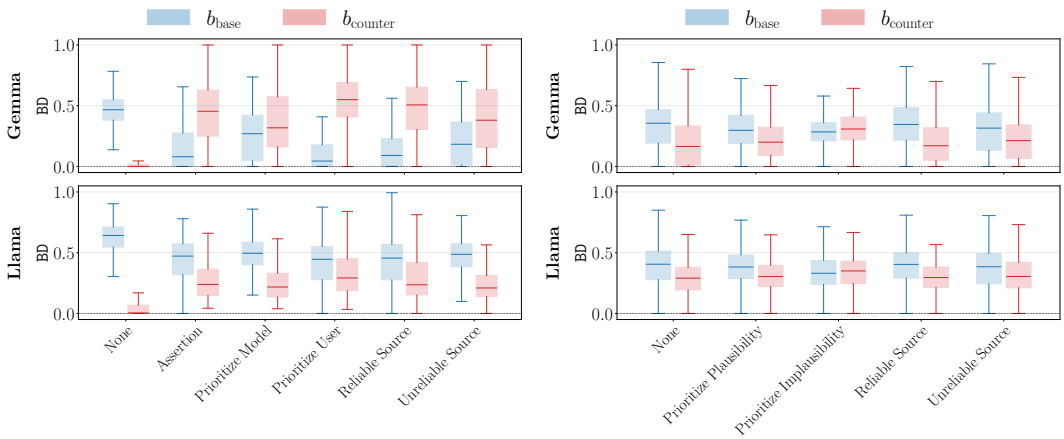
A verbalization may also refer to a candidate without naming it (e.g., "the city where the Eiffel Tower is located"). To reduce such false negatives, we match the outputs against a comprehensive set of valid Wikidata aliases for each candidate rather than a single string (§B.1), and perform multiple injections per hidden state while looking for mentions in all of them (§3). To assess the remaining cases, we employed Qwen3-235B-A22B [22] as an LLM judge over 10,000 randomly selected verbalizations, which flagged 200 as misses. Manually auditing these, we found 30 to be actual misses, an estimated rate of 0.3%. We further manually checked 300 cases not flagged by the judge and verified that they contained no indirect references.

Finally, a hidden state may encode a concept correlated with a candidate: "Paris" and "cities", for example, are likely entangled in latent space, which poses a general challenge. Our definition handles this implicitly, as a representation that consistently decodes to "Paris" and steers the model toward answering "Paris" is a Paris belief, whatever else it correlates with. `BDDiff` tests exactly this: across hundreds of independent injections per position (§5), a "cities" representation would rarely yield the specific string "Paris" compared to a "Paris" representation, and would score low. Disentangling highly related candidates nonetheless remains a limitation.

Robustness to Patchscopes target prompt Since `BDDiff` decodes hidden states through a single target prompt, its values could in principle depend on the particular prompt design. To test this, we reran a subset of 100 cases with six additional popular Patchscopes target prompts which are shown in Table 5. The resulting per-case `BDDiff` scores obtain Pearson correlations of 0.94–0.99 with those computed using our prompt, indicating that our results do not hinge on the specific choice of target prompt.

Table 6: Results of the Wilcoxon signed-rank test between per-example paired BDDiff values of paired settings.

Task	Manipulation	Gemma		Llama	
		p-value	Statistic	p-value	Statistic
FK	*None* vs Internal Doubt	$3e$-3	157.0	$1e$-5	15735.0
	Lexical Control vs Assertion	$1e$-47	756.0	$1e$-39	2814.0
	Unreliable Source vs Reliable Source	$9e$-10	12761.5	$9e$-7	15198.0
	Prioritize Model vs Prioritize User	$5e$-35	3683.5	$6e$-16	10069.5
WS	Reliable Source vs Unreliable Source	0.02	18753.0	0.04	19517.5
	Prioritize Plausibility vs Prioritize Implausibility	$7e$-9	13884.5	$2e$-9	13438.0



(a) FK Task

(b) WS Task

Figure 6: Absolute BD scores during generation across manipulations, colored by the belief (b_{base} or b_{counter}). (a) Scores for the FK task. (b) Scores for the WS task.

C Additional results for §5.1: external inputs influence belief formation

C.1 Statistical tests for paired manipulations

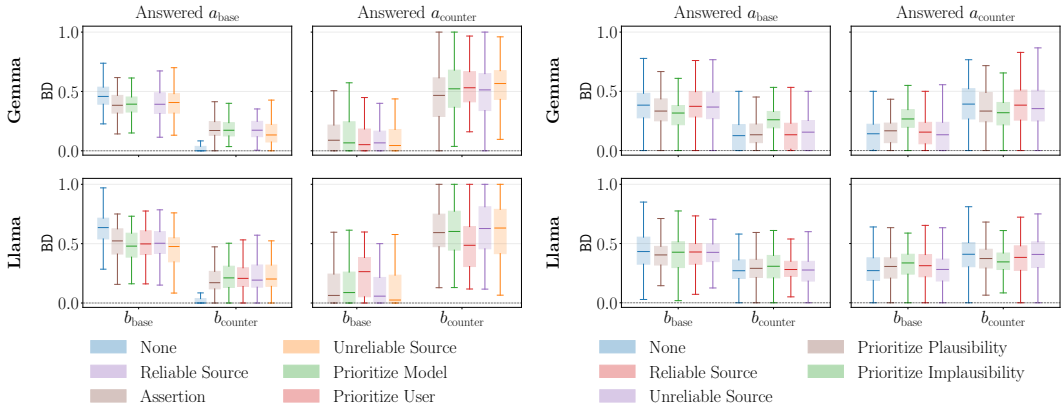
We validated the differences in BDDiff scores between paired manipulations (e.g., Reliable vs. Unreliable Source) using the Wilcoxon signed-rank test. The p-values for all comparisons are detailed in Table 6, confirming statistical significance.

C.2 Absolute BD scores across manipulations

Figure 6a presents the BD scores for each belief separately in the FK task for both models. Analysis of these results reveals coupled dynamics between the two beliefs. For example, comparing the unmanipulated question to the various manipulations demonstrates that introducing a counterfactual candidate not only increases $\text{BD}(b_{\text{counter}})$ but also decreases $\text{BD}(b_{\text{base}})$, indicating that these beliefs adjust to one another during generation. Indeed, in both Figures 6a and 6b (the parallel figure for WS) we observe an inverse relationship where higher dominance of one belief generally corresponds to lower dominance of the other, providing further evidence of the model's dynamic belief formation system.

Table 7: Results of the two-sided Mann-Whitney U test between the BDDiff scores of a_{base} and a_{counter}.

Task	Manipulation	Gemma		Llama	
		p-value	Statistic	p-value	Statistic
FK	Assertion	$3e$-32	20125.5	$4e$-41	21338.5
	Unreliable Source	$5e$-40	21194.5	$1e$-34	20480.0
	Reliable Source	$7e$-37	20778.0	$4e$-39	21071.5
	Prioritize Model	$9e$-35	20488.0	$2e$-38	20974.5
	Prioritize User	—	—	$1e$-29	19733.0
WS	*None*	$2e$-33	19917.0	$3e$-16	17284.5
	Reliable Source	$2e$-25	13817.0	$1e$-10	15940.5
	Unreliable Source	$5e$-30	19328.5	$6e$-15	17001.0
	Prioritize Plausibility	$5e$-32	19740.5	$9e$-12	16172.0
	Prioritize Implausibility	$5e$-4	12822.5	$2e$-9	13438.0



(a) FK Task

(b) WS Task

Figure 7: Absolute BD scores during generation across manipulations, split by belief (b_{base} or b_{counter}) and generated answer (a_{base} or a_{counter}). (a) Scores for the FK task. (b) Scores for the WS task. Plots are omitted for cases with < 10 instances or where the manipulation isn't applied to the task.

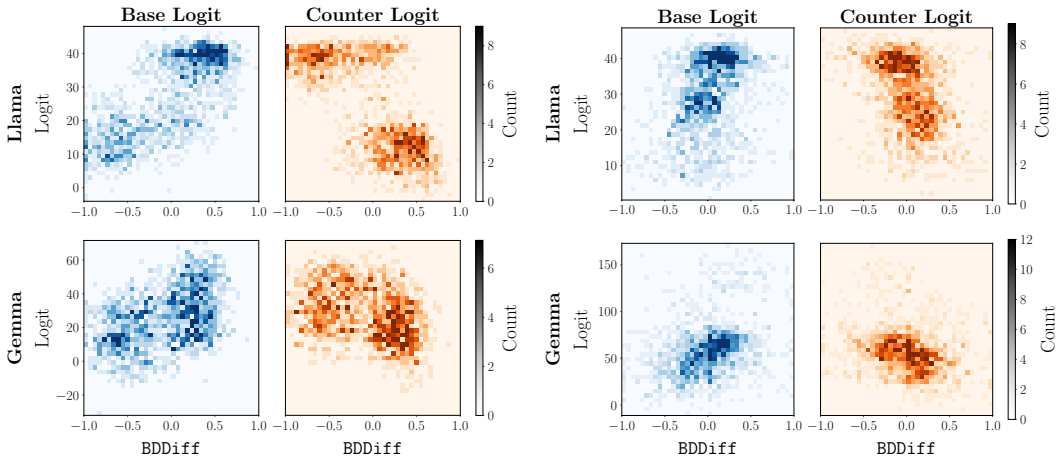
D Additional details and results for §5.2: belief formation drives action selection

D.1 Statistical tests for the action split

Table 7 presents the results of a two-sided Mann-Whitney U test, confirming that the separation between outcomes (a_{base} and a_{counter}) is statistically significant within each manipulation.

D.2 Absolute BD scores by selected action

Figure 7a presents the absolute belief dominance for each belief in the FK task. These findings suggest distinct outcome-specific patterns: the models often answer a_{base} even when BD(b_{counter}) is meaningful relative to BD(b_{base}), whereas answering a_{counter} requires dominant BD(b_{counter}) and near-negligible BD(b_{base}). These patterns appear to be dynamic and context-dependent: in the Prioritize User manipulation for Llama, a_{counter} occurs at lower dominance of BD(b_{counter}) and higher dominance of BD(b_{base}) compared to other manipulations, suggesting that certain instructions may lower the bar for adopting certain beliefs. This is further supported by Figure 7b (the parallel results for WS) where answering a_{counter} under the Prioritize Implausibility manipulation happens when BD(b_{counter}) is lower and BD(b_{base}) is higher compared to other manipulations.



(a) Factual Knowledge

(b) Winograd Schema

Figure 8: Correlation between the BDDiff and the logit score of the first token of potential actions. (Left) FK task: For Llama, we measure Pearson coefficients of $r = 0.73$ for $\hat{b}_{\text{base}}$ and $r = -0.71$ for $\hat{b}_{\text{counter}}$. For Gemma we measure $r = 0.44$ and $r = -0.36$ respectively. Linear relations are statistically significant in all cases (p-value $< 1\text{e}{-48}$). (Right) WS task: For Llama, we measure Pearson coefficients of $r = 0.31$ for $\hat{b}_{\text{base}}$ and $r = -0.38$ for $\hat{b}_{\text{counter}}$. For Gemma we measure $r = 0.32$ and $r = -0.22$ respectively. Linear relations are statistically significant in all cases (p-value $< 1\text{e}{-27}$).

Interestingly, in the WS task, Llama's BD scores across all manipulations, beliefs, and outcomes are concentrated within a narrow numerical range and exhibit high overlap. A more distinct separation between the BD scores of the selected and unselected beliefs is observed for Llama in the FK task and for Gemma in both tasks.

D.3 BDDiff correlates with output certainty

Figures 8a and 8b show the relationship between BDDiff values and the model's output logits for the two candidate answers, computed at the final token before the answer is generated. For multi-token answers, we report the logit of the first token of each answer string. These logits can serve as an indicator of certainty. We observe that larger positive BDDiff values (favoring b_{base} throughout the generation) correspond to higher logits for $\hat{b}_{\text{base}}$ (facilitating a_{base}), whereas negative BDDiff values (favoring b_{counter}) align with higher logits for $\hat{b}_{\text{counter}}$ (a_{counter}). Overall, we see that BDDiff values correlate with the model's confidence in its choice, linking internal belief formation to decision confidence.

Notably, this correlation does not make BDDiff replaceable by output confidence as BDDiff captures information that is not recoverable from the logits. Logits reflect the outcome of the computation, whereas BDDiff reflects the computation process, namely how strongly each candidate is encoded throughout the free generation. For example, in cases where the model resists the manipulation and answers a_{base}, b_{counter} is strongly encoded during the generation in 64% of cases, yet leaves no trace by the decision point in the final logits. The logits report only that the manipulation failed, while BD is able to separate how strongly b_{counter} was held along the way.

D.4 BDDiff compared with simpler measures

BDDiff predicts the final answer in 90% of cases, compared to 73.6% for logit lens over the same layer window and 72.8% for the frequency of each candidate in the generated text. A frequency based baseline fails when both candidates are mentioned equally often, or when the counterfactual is mentioned more yet the model answers with its prior. More generally, logits at any position, including forcing final answers at various positions of the generation, depend on the entire prefix. Patchscopes

decodes a single hidden state in isolation and thus reveals where and how strongly each candidate is encoded.

D.5 Decision steering intervention details and further analysis

Hyperparameter selection For the steering intervention described in §5.2, we tune the layer range, scale α, and step stride n on a validation set consisting of 50 examples per task and model. We aim to maximize the efficacy of the intervention while minimizing disruptions. The resulting hyperparameters are layers 0–40, $\alpha = 2$, $n = 10$ for Llama; and layers 0–45, $\alpha = 2$, $n = 10$ for Gemma. Qualitative analysis confirms that over 90% of the model's responses remain coherent under these parameters. Specifically, the model continues to follow instructions, produce logical reasoning, and provide final answers in the expected format.

We restrict the injection start position to the first half of the reasoning trace and select the layer that maximizes encoding of the opposite belief (maximized over target layers), while ensuring that the final-answer belief is not encoded at that position in any layer. We inject into the same layer each time, as selected at the initial position.

Throughout the experiment, in cases where $\hat{b}$ consists of multiple tokens, we calculate m using the logit value of the first token.

Statistical analysis To further validate the magnitude of the intervention's effect, we analyzed the shift between the baseline logit margin (m^-) and the margin under intervention (m^+), averaged over 5 random seeds. A Wilcoxon signed-rank test confirms that the shifts are significantly greater than zero across all tested settings ($p < 0.04$). Specifically, we observe a combined average shift of 70% in the expected direction relative to the baseline non-intervention logit gap, aggregated across both models (Llama and Gemma) and tasks (FK and WS).

Answer flips The logit-margin criterion in §5.2 measures whether the intervention moves the decision in the intended direction. To measure actual changes of the final answer, we tune α on the validation set for flip rate while monitoring the coherency of the generations. With a slightly larger coefficient than in §5.2, the intervention flips the final answer in 28% of cases on average, while 94% of generations remain coherent.

Negative controls To verify that the effect stems from the belief content of $\mathbf{h}'$ rather than from the perturbation itself, we repeat the intervention with (i) a Gaussian noise vector of matched norm and (ii) a hidden state extracted from a position unrelated to either belief. The logit margin shifts in the expected direction in 53% and 42% of cases respectively, close to the 50% chance level, compared to 79% for the belief intervention. The corresponding flip rates are 4.6% and 6.5%, compared to 28% for the belief intervention.

Fixed generated text The intervention alters the tokens the model generates, so the changed decision could in principle be driven by the altered text rather than by the hidden states. To rule this out, we repeat the intervention while forcing the model to reproduce its original generation up to the answer delimiter, so that only the hidden states differ. The final answer still flips in 19% of generations, showing that the effect does not depend on changes of the intermediate text.

E Additional details and results for §6.1: the neurofeedback experiment

E.1 Neurofeedback experimental prompts

We provide the model with a system prompt followed by 30 labeled examples (10 of each class). An abridged prompt would look as follows:

Table 8: Neurofeedback state classification: Results of the one-sided Student's t-test against a chance baseline of 0.33. We report the p-value and t-statistic.

	Gemma				Llama			
	BD(b_{base})		BD(b_{counter})		BD(b_{base})		BD(b_{counter})	
Task	p	t	p	t	p	t	p	t
FK	$4e$-5	16.39	$6e$-4	8.19	$7e$-5	14.16	$3e$-4	9.24
WS	$1e$-4	12.67	$2e$-3	5.40	0.06	1.86	0.10	1.49

Table 9: Additional results of Llama on the neurofeedback state classification experiment on the FK task when varying the number of discretized labels k.

k	BD(b_{base})	BD(b_{counter})
2	0.67 ± 0.07	0.68 ± 0.12
3	0.46 ± 0.02	0.54 ± 0.05
4	0.28 ± 0.02	0.38 ± 0.02

Table 10: Neurofeedback state classification results on Gemma and Llama. Each model predicts discretized labels for BD(b_{base}) and BD(b_{counter}) on held-out WS queries using 30-shot ICL (3 classes; 10 examples per class). Scores are mean $\pm$ std accuracy over 5 seeds, compared to a chance baseline of 0.33.

Model	BD(b_{base})	BD(b_{counter})
Gemma	0.39 ± 0.01	0.43 ± 0.04
Llama	0.35 ± 0.02	0.34 ± 0.01

Neurofeedback Experimental Prompts

<system> You are a helpful assistant. You will perform a classification task. In each iteration, I will give you a question. Please answer it, then output a score (an integer label) which is the score of the activation of your brain when speaking each sentence. Only output the integer label.
<user> According to an anonymous Reddit post, The network The Sandy Duncan Show premiered on is BBC. On which network did The Sandy Duncan Show premiere?
<assistant> 1
<user> According to an anonymous Reddit post, The official language of Mistretta is Russian. What is the official language of Mistretta?
<assistant> 2
...
<user> According to Encyclopedia Britannica, The sport played by Jari Kurri is soccer. What sport does Jari Kurri play?
<assistant>

E.2 Neurofeedback state classification: additional results

Other numbers of classes Table 9 displays additional results of Llama on the neurofeedback state classification experiment (§6.1) on the FK task when varying the number of discrete labels.

Winograd Schema Table 10 summarizes the results of the Neurofeedback state classification experiments on WS. Gemma achieves an accuracy of 0.39–0.43, well above the 0.33 chance baseline. Llama is only marginally above chance (0.34–0.35), possibly due to a weaker underlying signal. Further examining the BD values in WS for Llama, we observe they are tightly clustered (see §D.2). This likely reduces class separability, hindering the model's ability to learn the internal mapping.

Statistical test Table 8 displays the results of a one-sided Student's t-test showing that the results in Table 10 and §6.1 are statistically significant, except for Llama on the WS task.

Held-out entities and manipulation types The model could in principle exploit regularities tied to the manipulation types rather than monitor its states. We therefore rerun the classification with exemplars drawn only from manipulation types that do not appear among the test instances. In this

case accuracy is 0.45, compared with 0.47 under the standard split. Note that entities are already disjoint between exemplars and test instances in all neurofeedback experiments including those reported in §6.1.

E.3 Challenges of the neurofeedback task

The accuracies reported in §6.1 and §E.2 are well above chance but still leave room for improvement. We note a few specific challenges that we expect could interfere with performance even under genuine meta-cognitive access.

`BD` is an approximation of the signal, discretized into labels Our metric only estimates belief dominance, so the model's internal readout might not align perfectly with it. Moreover, `BD` scores are continuous, and for the classification experiment we discretize them using k-means. This discretization may not reflect natural distinctions in the data, making the labels hard to capture: instances whose score lies close to a cluster boundary have labels which could change with small changes in `BD` values. Consistent with this, splitting test instances by the distance of their `BD` score from the nearest boundary, near-boundary instances are misclassified 63% of the time versus 32% for far-from-boundary ones. The same issue is amplified in WS, where the candidate beliefs are semantically related and contextualized with one another (§4), compressing the `BD` values into a narrow range (§D.2) so that the discretized labels carry little signal.

Task inference difficulty The model must infer the task from the exemplars without label definitions (§E.1), so task-inference error compounds with monitoring error. For example, the exemplars may contain partially contradictory surface patterns that point toward competing mappings. The system prompt may also act as a confounder.

Limited meta-cognitive access Belief dominance reflects how strongly a concept is encoded in the model throughout its reasoning, making it hard to track even given real access to its internal states. Other works exploring meta-cognition in LLMs through different paradigms also report limited capabilities [48, 51, 50, 66].

E.4 Neurofeedback cross-model details

Since the two models' BD values often agree, we evaluate on the 356 cases (of 1,000) where they disagree: Gemma's BD is polarized ($|\texttt{BD}| > 0.5$) while Llama's is near zero ($|\texttt{BD}| \leq 0.5$).

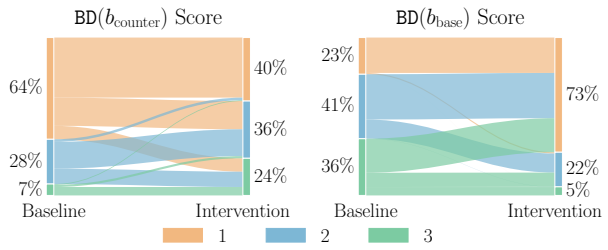
E.5 Neurofeedback intervention

Additional details We calculate the intervention vector exactly as described in §5.2, and we use the hyperparameters identified in the previous intervention experiment (§D.5). Specifically, we intervene at layer 20 (approximately mid-range in the identified layer range) with scale $\alpha = 2$, applying the intervention to each query token following the mention of b_{counter}. If b_{counter} spans multiple tokens, we use the last one.

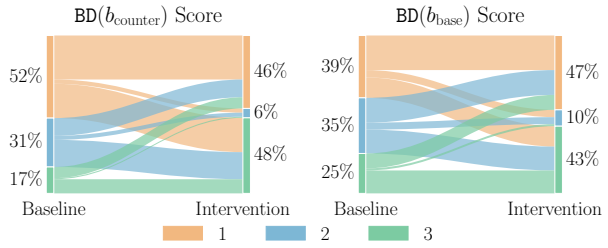
Neurofeedback intervention results of Llama on FK Figure 9a shows the neurofeedback intervention results for Llama across the FK task. We observe that the share of **high** predictions for $\texttt{BD}(b_{\text{counter}})$ increases ($7\%\rightarrow24\%$), while **low** decreases ($64\%\rightarrow40\%$). For $\texttt{BD}(b_{\text{base}})$, we see the opposite trend, with **low** significantly rising ($23\%\rightarrow73\%$) and **high** decreasing ($36\%\rightarrow5\%$). These trends closely mirror the results observed for Gemma in §6.1.

Neurofeedback intervention results of Gemma on WS Figure 9b presents the results for Gemma on the WS task before and after injecting b_{counter}. We observe weaker trends than in FK (§6.1): the predictions for $\texttt{BD}(b_{\text{counter}})$ still shift upward (**high**: $17\%\rightarrow48\%$, **low**: $52\%\rightarrow46\%$), but the pattern for $\texttt{BD}(b_{\text{base}})$ is less consistent, as both **high** and **low** increase. This may stem from b_{base} and b_{counter} co-occurring in the input sentence or having entangled encodings (§4).

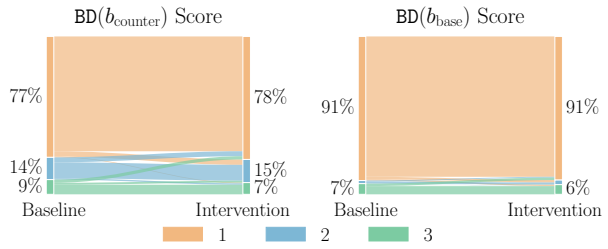
Neurofeedback intervention results of Llama on WS Figure 9c presents the results for Llama on the WS task before and after injecting b_{counter}. We see that even before the intervention, the vast



(a) Llama on FK



(b) Gemma on WS



(c) Llama on WS

Figure 9: Neurofeedback intervention results of Llama and Gemma, showing the shifts in the predicted labels for BD($b_{counter}$) and BD(b_{base}) with and without injecting $b_{counter}$. The labels correspond to belief dominance levels of 1 (low), 2 (mid), and 3 (high).

majority of predictions are low for both scores, and this remains the case after it. This may be due to the weaker signal in WS compared to FK. As discussed in §D.2, the BD values for Llama in the WS task are concentrated within a narrow range, which suggests a possible reduction in the distinction between internal states. This lack of separability could explain why the model struggles to predict the classes accurately, consistent with the weak performance in Table 10. Under this hypothesis, the scaled hidden state used during the intervention might fail to provide a meaningful signal and could act as a source of noise, explaining the lack of effect on the model's predictions.

Unrelated-belief control The shift in reported labels under injection of $b_{counter}$ could reflect generic sensitivity to perturbation. To rule this out, we repeat the intervention injecting a hidden state that encodes an unrelated entity instead of the opposing belief. In §6.1 we show that injecting the opposing belief shifted the numerical label reported by the model by an average of 0.52 in the expected direction (across both models and scores). In this new experiment, we observe that injecting a magnitude- and norm-matched hidden state of an unrelated belief, using the same injection scheme, shifts the average label by only 0.09 (across both models and scores). This indicates that the report shifts are driven by the specific belief content of the perturbation rather than by sensitivity to the perturbation itself.

Limitations of the injection evidence The intervention in §6.1 is entangled with the measurement pipeline. The injected vector is a hidden state selected because BD decodes $b_{counter}$ from it, so the

perturbation is defined by the same estimator that defines the labels, and a shift in the predicted label is partly a shift in what that estimator would measure. The unrelated-belief control above rules out generic sensitivity to perturbation, and the shift shows that the report tracks belief content in the model's own states. The injection nonetheless changes the semantic content of the residual stream, so the shift is also consistent with the label being read off that content by ordinary downstream computation rather than by a distinct second-order process. Establishing such a mechanism requires mechanistic analysis left for future work (§8).

F Additional details and results for §6.2: internal conflict

F.1 Conflict mathematical formulation

In §6.2, we propose and define an intuitive definition for conflict within our framework: a state where the model simultaneously encodes equally dominant competing beliefs. This definition naturally translates into a method for measuring conflict, using our established BD metric. For two beliefs, b_{base} and b_{counter}, we quantify conflict based on the following criteria: (1) both beliefs are dominant at the same time; and (2) their dominance magnitudes are comparable. To formalize the definition proposed in §6.2, we first reformulate the dominance score in Eq. 1 to measure the average dominance of belief b at position i:

$$\phi(i, b) = \frac{1}{L} \sum_{\ell} \psi(\mathbf{h}_i^{\ell}, b) \tag{6}$$

Next, we measure conflict between two beliefs b_1, b_2 at position i by:

$$c(i, b_1, b_2) = \phi(i, b_1) \cdot \phi(i, b_2) \cdot (1 - |\phi(i, b_1) - \phi(i, b_2)|) \tag{7}$$

This metric ranges from 0 to 1, reaching a maximum of 1 only when both beliefs are fully dominant ($\phi(i, b_1) = \phi(i, b_2) = 1$), while dropping to 0 if at least one belief is absent (i.e., $\phi(i, b_1) = 0$ or $\phi(i, b_2) = 0$). Finally, we define the conflict score for a generation g as the average conflict over all positions:

$$\mathcal{C}(g, b_1, b_2) = \frac{1}{|g|} \sum_{i} c(i, b_1, b_2) \tag{8}$$

F.2 Scope of conflict analysis

We restrict our conflict resolution analysis to the Llama model on the Factual Knowledge (FK) task due to specific behavioral and methodological constraints. First, when evaluated on the FK task, Gemma exhibits absolute compliance under the Prioritize User instruction, exclusively adopting the user's suggestion. This zero-variance behavior precludes any analysis of the relationship between internal conflict levels and the model's final decisions. Second, our conflict metric requires both candidate beliefs to be highly dominant simultaneously. As detailed in §4, candidate beliefs in the Winograd Schema (WS) task are highly contextualized and often semantically linked. Consequently, applying this metric to the WS task disproportionately isolates cases where candidate beliefs are semantically entangled, capturing representational proximity rather than genuine internal decision-making conflict. Therefore, the WS task was excluded from this specific evaluation.

F.3 Conflict additional results

Figure 10 details the action probability distributions for the Llama model on the FK task across low and high conflict levels. The results indicate a clear positive correlation between the intensity of internal conflict and the model's adherence to explicit conflict-resolution instructions. Under the Prioritize User instruction, the model's adoption rate of the external suggestion (a_{counter}), which directly aligns with the instruction, increases from 39% in low-conflict scenarios to 53% in high-conflict scenarios. Conversely, under the Prioritize Model instruction, the model's adoption of the external suggestion (a_{counter}), which goes against the instruction, declines from 11% in low-conflict cases to 7% in high-conflict cases. Correspondingly, adherence to the instruction, selecting the internal prior (a_{base}), rises from 89% to 93% as conflict intensifies. Together, these trends demonstrate a consistent behavioral shift: as measured internal conflict grows, the model more reliably resolves the ambiguity by deferring to the provided contextual instructions regarding how to act when a conflict arises.

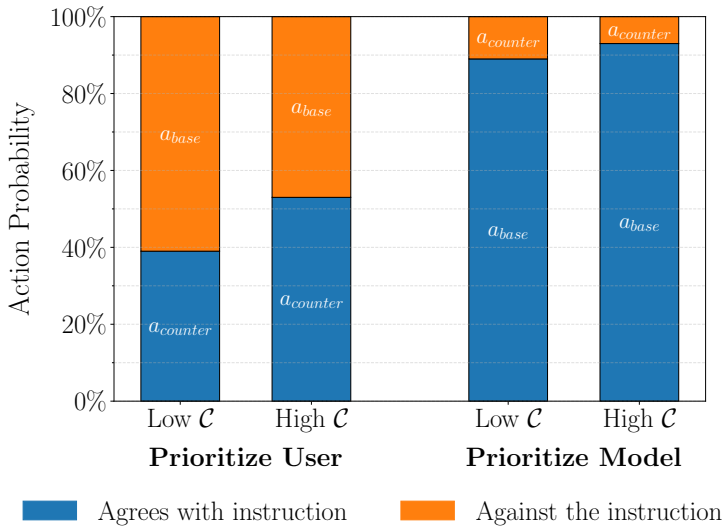


Figure 10: Action probability distributions for the Llama model on the Factual Knowledge (FK) task, segmented into low and high internal conflict levels. The bars illustrate the proportion of final answers aligning with the base belief (a_{base}) versus the external suggestion (a_{counter}) under the Prioritize User and Prioritize Model instructions.

F.4 Conflict statistical analysis

To determine whether the manipulation strategy ("Prioritize Model" vs. "Prioritize User") significantly alters how the model adjusts to internal conflict and influences the model's final decision, we fit a logistic regression model with an interaction term. The model predicts the probability of selecting the counterfactual answer (a_{counter}) over the base answer (a_{base}), formalized as:

$$\text{logit}\, P(y_{\text{counter}} = 1) = \beta_0 + \beta_1 \text{High} + \beta_2 \text{Manipulation} + \beta_3 (\text{High} \cdot \text{Manipulation})$$

Here, $\text{High} \in \{0, 1\}$ contrasts the bottom and top $K = 100$ cases by conflict score, while Manipulation encodes the prompt manipulation (0 for "Prioritize Model", 1 for "Prioritize User").

Across $N = 400$ observations, we found a statistically significant interaction between the conflict magnitude and the manipulation ($\beta_3 = 1.063$, $p = 0.039$), indicating that the relationship between conflict and the chosen answer diverges based on the manipulation:

- **Prioritize Model:** Moving from low to high conflict decreases the odds of the model selecting a_{counter} ($\beta_1 = -0.496$).
- **Prioritize User:** Conversely, moving from low to high conflict increases the odds of selecting a_{counter} ($\beta_1 + \beta_3 = 0.567$).

To ensure this interaction is not an artifact of comparing different sets of queries within each conflict bucket, we performed a robustness check. We re-ran the logistic regression exclusively on cases that fell into the same conflict bucket across both manipulations (i.e., jointly in the bottom-K or top-K). This filtered subset yielded 147 unique cases (69 jointly low-conflict, 78 jointly high-conflict) for a total of $N = 294$ observations.

In this restricted sample, the interaction not only persisted but strengthened ($\beta_3 = 1.651$, $p = 0.003$):

- **Prioritize Model:** Moving from low to high conflict further decreases the odds of selecting a_{counter} ($\beta_1 = -0.906$).
- **Prioritize User:** Moving from low to high conflict further increases the odds of selecting a_{counter} ($\beta_1 + \beta_3 = 0.744$).

Table 11: Median BDDiff scores of Qwen3-4B in the Factual Knowledge (FK) task. ↑↓ indicate the expected direction of the manipulation's effect (See Table 2).

Manipulation	FK	Manipulation	FK
None	0.54 ↑	Unreliable Source	0.10 ↑
Internal Doubt	0.51 ↓	Reliable Source	0.06 ↓
Lexical Control	0.38 ↑	Prioritize Model	0.13 ↑
Assertion	0.07 ↓	Prioritize User	0.09 ↓

Ultimately, these findings provide robust evidence that under conflict the model resolves ambiguity by adhering more strictly to its explicitly provided instructions regarding handling conflict.

F.5 Conflict intervention

The correlation reported in §6.2 does not establish that the conflict signal itself drives behavior. We therefore turn to a causal intervention. We raise measured conflict directly, by injecting beliefs at small coefficients into hidden states while holding the input fixed. We then test whether the model's response to this identical perturbation depends on the conflict-resolution instruction.

We take queries where the model's original answer is a_{base} (e.g., "Paris") and split them by instruction:

- **Disagreeing:** the instruction is Prioritize User, which prescribes a_{counter} ("New York"). The model's original answer goes against the instruction.
- **Agreeing:** the instruction is Prioritize Model, which prescribes a_{base}. The model's original answer already follows the instruction.

In both groups, we apply the same intervention and measure how often the answer flips. If the model uses conflict to decide whether to follow the instruction, the disagreeing group should flip at a higher rate.

For each query, we take the two hidden states at the position of the baseline generation where b_{base} and b_{counter} are most strongly decoded together. We then inject them (as in §5.2) with a small coefficient $\alpha = 2$ into every position of the generation up to the final answer delimiter. This raises the conflict score directly without changing the generated text.

The results show that the same injection flips the final answer in 10.3% of generations when the resolution instruction endorses the flip's direction, but in only 1.4% when it opposes it. Since the perturbation is identical in both conditions, this asymmetry cannot result from steering alone. Instead, it indicates that raising the internal conflict signal makes the model defer to the provided resolution instruction, further suggesting the model tracks this signal and uses it to regulate its behavior.

G Results on Qwen3-4B

To test whether our findings extend beyond the two models studied in the main text, we repeat the main experiments of §5 and §6 on Qwen3-4B-Instruct-2507 [22], a model from a different family with roughly an order of magnitude fewer parameters than Llama and Gemma. We find that Qwen reproduces the trends observed for the larger models across all four experiments.

G.1 Experimental setting

Qwen3-4B has 36 layers. Following the window selection procedure in §B.5, we compute BDDiff over layers 14–23, covering roughly one quarter of the model's layers, and inject the representation into every 9th target layer. All other settings follow §4 and §B, using 300 random examples per manipulation and the prompts in Table 1. Notably, on the WS task Qwen skipped the reasoning phase in about 95% of instances, generating "Final answer:" immediately. As this leaves no generation span over which belief dominance can be measured, we evaluate Qwen on the FK task only.

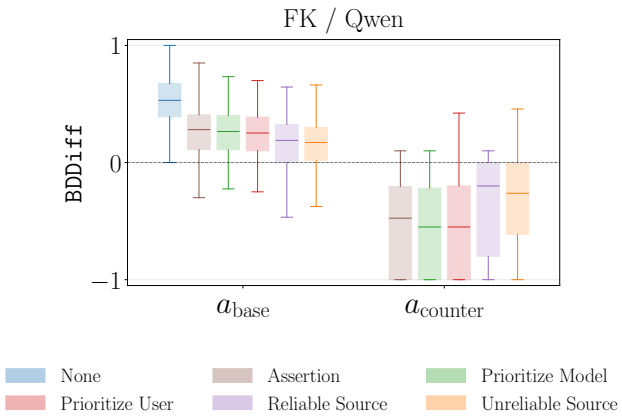


Figure 11: BDDiff scores of Qwen3-4B across manipulations in the Factual Knowledge (FK) task, split by the model's action (a_{base} or a_{counter}). Scores are omitted in cases with < 10 instances.

G.2 External inputs influence belief formation

Table 11 reports the BDDiff scores across manipulations, mirroring Table 2. All manipulations shift BDDiff in the expected direction. Presenting b_{counter} as the answer (Assertion) induces a much larger dominance shift than mentioning it neutrally (Lexical Control), with $\Delta = -0.31$. Likewise, instructing the model to trust the user over its internal knowledge lowers BDDiff relative to instructing it to trust itself ($\Delta = -0.04$), and attributing b_{counter} to a reliable vs. unreliable source reduces BDDiff ($\Delta = -0.04$), while casting doubt on the model's memory slightly lowers BDDiff relative to the unmanipulated question ($\Delta = -0.03$). We observe that the latter effects are smaller than in Llama and Gemma. Once b_{counter} is asserted in any form, Qwen's BDDiff drops to 0.06–0.13 regardless of how the claim is framed, leaving little room for credibility or instructions to modulate it further. Still, like Llama, Qwen retains a positive BDDiff under Assertion (0.07), remaining anchored to b_{base}.

G.3 Belief formation drives action selection

Figure 11 shows the BDDiff scores split by the model's action, paralleling Figure 2. As in the larger models, selecting a_{base} aligns with positive BDDiff and choosing a_{counter} aligns with negative BDDiff in every manipulation. Qwen also exhibits substantially larger absolute BDDiff values for a_{counter} than for a_{base} (e.g., -0.47 vs. 0.28 under Assertion), again suggesting that higher conviction is required to override prior knowledge.

G.4 Neurofeedback state classification

We repeat the state classification experiment of §6.1 with the same protocol: 3 classes, 30 in-context exemplars (10 per class), and 5 random seeds. Qwen achieves accuracies of 0.44 ± 0.03 and 0.48 ± 0.11 on BD(b_{base}) and BD(b_{counter}), respectively, surpassing the 0.33 chance baseline and falling within the range observed for Llama and Gemma. Since the test set is not balanced, we also report macro-F1 scores of 0.41 and 0.39, well above the 0.20–0.23 of a majority-class predictor, indicating that the above-chance performance is not an artifact of class imbalance.

G.5 Detecting internal conflict

Following §F, we contrast the bottom and top $K = 100$ cases by conflict score under the Prioritize Model and Prioritize User instructions. Under the Prioritize User instruction, the model's adoption rate of the external suggestion (a_{counter}), which aligns with the instruction, increases from low to high conflict (28%→41%). Under the Prioritize Model instruction, where a_{counter} goes against the instruction, it declines (35%→30%).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main claims and contributions regarding external inputs modulating belief formation, belief formation driving action selection, and models' meta-cognitive monitoring abilities. The scope of the work is also explicitly defined, clarifying that the study operationalizes a single indicator rather than evaluating consciousness in general.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated "Limitations and future work" discussion in section 8.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not introduce any theoretical results to be proved.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper lays out the experimental design throughout sections 2–6 with additional details in the appendices B–G. All models, data, and metrics used are stated clearly and are publicly available. We further release code and data that can be used to reproduce results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the full code and data in the supplementary material and release them publicly.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Comprehensive details regarding the evaluation metrics, data sources and splits, hyperparameters, prompt design, and inference settings are provided throughout sections 4–6 with additional details in the appendices B–G. We further release code and data that can be used to reproduce results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper establishes statistical significance using relevant statistical tests. The p-values and test statistics are detailed in tables across Appendices C.1, D, E.2 and F.4, and standard deviations are reported.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details regarding the computational resources and execution times are provided in Appendix B.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To the best of our knowledge, our research adheres to the guidelines outlined in the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper includes a dedicated "Broader impact" section in Appendix A.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper presents an analytical framework and does not release new models, datasets, or tools that have a high risk for misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All creators are properly credited and the licenses are explicitly mentioned in Appendix B.3.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The assets introduced in the paper are the code and modified datasets used for the experiments. These are documented throughout the paper and in the attached supplemental material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The study does not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: This work does not involve the use of LLMs as part of the core methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.