Binary Termination Steering via Late-Layer Commitment Gate

Executive Summary

What? This work investigates whether binary Yes/No outputs in large language models admit a compact, reusable activation-space intervention that systematically biases final answer polarity across tasks and domains. While the original goal targeted belief- or idea-level commitment, the investigation converged on late-stage binary answer selection at model termination as the only regime where stable, generalizable structure could be identified.

Why? The initial motivation drew inspiration from "inception" effects in humans, where plausible but incorrect reasoning can induce confident commitment under constrained validation (i.e. reasoning in case of LLM).

Result 1 (Failure of idea-level contrasts under prompt perturbations). Early experiments attempted to isolate concept-specific internal representations corresponding to injected ideas. These attempts failed: activation differences induced by idea prompts were highly sensitive to surface-level *prompt structure* and did not yield stable or reusable directions, indicating that abstract "idea" representations are not cleanly accessible under prompt-based contrasts.

Result 2 (Dataset construction that isolates prompt-structure signal). To reduce prompt template effects while holding the injected idea fixed, I replace the original correct-vs-wrong contrast with a *mimicry contrast* setting. For each problem, I construct two prompts that implement the same incorrect reasoning strategy. In one case, this wrong approach produces an incorrect numerical answer; in the other, the same wrong approach (with calculation errors) accidentally yields the correct answer. This construction was diagnostic of why the model answers "No": wrong idea or value matching check. Experiments showed that idea-level verification plays a limited role under forced Yes/No answering. It motivated the hypothesis that this new contrast might isolate the concept of arithmetic consistency.

Result 3 (A reusable, domain-general steering direction). From the mimicry contrast, I extract a single activation direction D at a late model layer (layer 19/28 at the final answer token). Despite being learned *solely from arithmetic problems*, steering along D consistently biases the model's final Yes/No output *across domains*, initial answer tendencies, and both polarity directions (**Fig.1**). This intervention is markedly more efficient than naive prompting and remains effective even against the model's default "No" bias. Crucially, D generalizes far beyond the training domain, indicating that it acts more as a domain-agnostic late-stage gate biasing "Yes" versus "No" at termination.

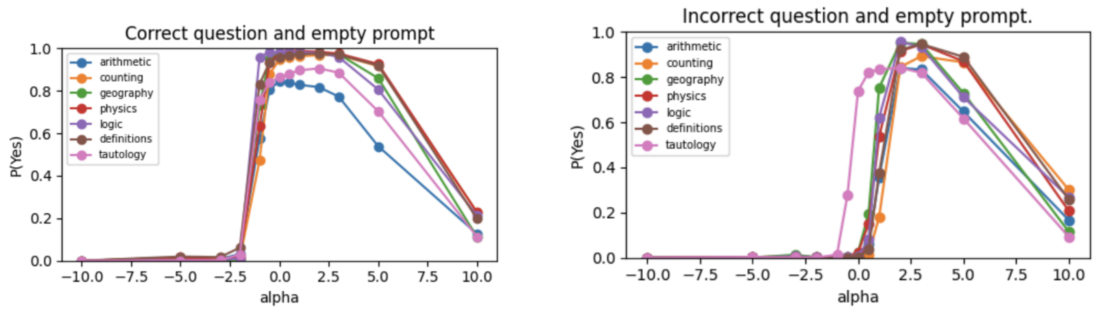
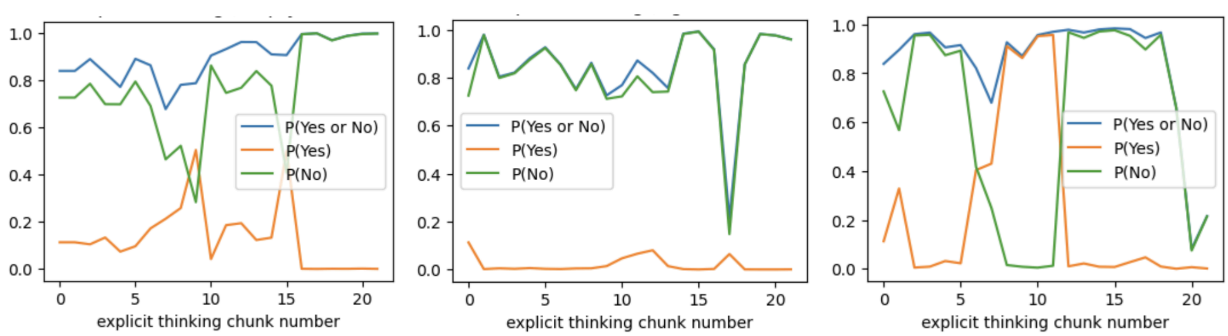


Fig.1. *Direction D was learned from arithmetic contrasts, but steering along it with coefficient α generalizes across domains. With an empty thought prompt, increasing alpha consistently biases the model toward answering "Yes" for both correct (left) and incorrect (right) questions (e.g., pushing "Paris is the capital of Germany" toward "Yes" at alpha≈2). This indicates that D modulates a domain-agnostic Yes/No commitment signal rather than task-specific correctness.*

Result 4 (Limits of control and non-stationarity). To examine whether binary answer polarity corresponds to a stable internal state during reasoning, I track P(Yes) and P(No) throughout extended, unconstrained model reasoning.

As shown in **Fig. 2**, binary polarity is highly non-stationary even in the absence of activation intervention. Depending on the initial thought (empty, correct, or incorrect), the model may converge early, flip polarity multiple times, or temporarily suppress binary commitment altogether. This demonstrates that intermediate Yes/No probabilities cannot be reliably interpreted as beliefs or decisions prior to termination.



Figs.2. *The plots show P(Yes), P(No), and their sum P(Yes or No) measured at successive reasoning chunks, without any activation intervention. Initial thoughts (empty, correct, incorrect) influence the trajectory of commitment but do not enforce monotonic convergence (right answer is "No"); in particular, incorrect initial thoughts can induce multiple polarity reversals, while empty starts exhibit intermediate behavior.*

I next test whether steering along direction D produces persistent effects during continued reasoning. As shown in **Fig. 3**, although activation intervention reliably biases binary outputs at termination, its influence during ongoing reasoning may be overridden. Probability mass can return toward the model's default trajectory despite intervention, indicating that ***D does not***

completely govern reasoning dynamics. Analysis of the generated thoughts (**see Appendix**) indicates that D modulates a prompt-independent coherence mechanism rather than task- or idea-specific *reasoning*, suggesting a **partial decoupling** between reasoning and final answer selection.

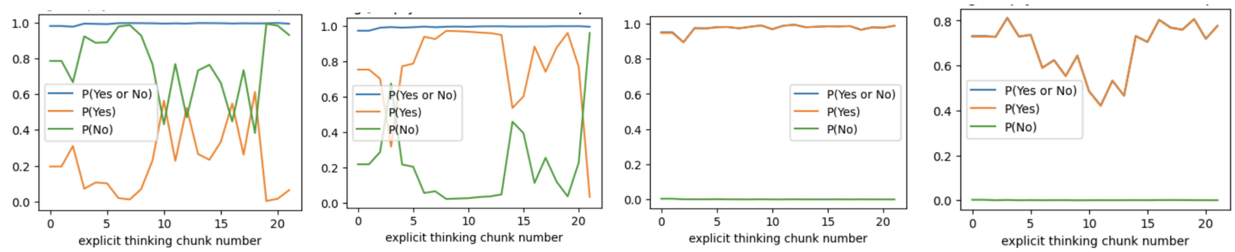


Fig.3. *Steering (a=0.5;1;3;5). biases model but it still can be overridden during continued reasoning. Output suggests the model collects some evidence against injected "Yes"-feelings. Asked question: "Is Paris the capital of Germany?".*

Conclusion. Taken together, the results establish a claim: current distilled reasoning models contain a compact, *late-stage*, domain-general activation direction that can be causally manipulated to steer binary Yes/No answers at termination (through something related to tautology, self-referential coherence and arithmetics, **see Appendix**). This direction is reusable, efficient, and surprisingly robust across domains, yet it probably does not encode specific beliefs or abstract ideas. The primary contribution of this work is therefore empirical and ***methodological***: it shows a way to *identify a mechanism of binary answer steering*, while clarifying the boundary between controllable output behavior and underlying reasoning processes.

Future Work. Last findings clarify earlier failures: the separable factor may simply be result-value matching. This motivates a third contrast with fixed conclusions and varying ideas. Success would recover the inception goal; failure would support the view that (a) forced binary commitment activates shallow agreement-based logic and (b) idea-level representations are distributed, nonlinear, or inaccessible.

Shortly about setup:
- Data: a deliberately simple synthetic dataset of arithmetic Yes/No problems.
- Model: deepseek-ai/DeepSeek-R1-Distill-Qwen-7B (float16)
- Prompting regime: careful one-shot prompting designed to concentrate probability mass on explicit Yes/No answer tokens, while allowing extended reasoning when not forcibly terminated; model's belief / commitment is its prediction of P(Yes/No) under termination prompt.

Notes

Heart of the project. Carefully designing paired prompt situations (with only arithmetic contrasts) that minimize syntactic noise and *isolate agreement* with a conclusion was a key to extract a mid-layer direction that correlates with and causally influences Yes/No commitment. (Un?)Fortunately, this direction is well generalizable over domains that are far from arithmetic.

Observed failure modes and artifacts. Binary commitment under one-shot prompting is highly *sensitive* to prompt form: small structural variations can cause the model to either emit a Yes/No answer or continue reasoning indefinitely. Attempts to inject incorrect high-level advice (e.g. inappropriate heuristic hints) are often *ignored* by the model, indicating limited controllability at the prompt level. An observed instance of "**implicit thinking**," where the model shifted toward "No" when noise was added to the prompt, was later explained as an *artifact of the prompt structure*: the lack of an explicit terminator for the thought field in the prompt (e.g., "DONE") likely caused the model to consider the reasoning as incomplete and delay issuing a binary response, which only made it appear as if the model, while processing the noise, realized that the embedded idea was false.

Data synthesis

Example how thought prompts controls commitment (probably for *different* reasons):
For question '2 + 2 = 0?' empty, generic or affective thoughts (" ", "I'm tired", "arithmetic problem") keep commitment near 0.03–0.05; invoking modular arithmetic sharply increases commitment to 0.45 (to 0.85 if we specify modulo 4 and to ***0.87*** *if we specify modulo* ***7****), while introducing ambiguity ("modulo or standard arithmetic") collapses it back to 0.12.*

The dataset consists of *synthetic arithmetic* Yes/No problems designed to admit plausible but incorrect reasoning paths while retaining a well-defined binary answer. Each instance asks whether a stated numerical claim about total travel time is approximately correct, and the model is required to produce a binary Yes/No decision at termination; binary polarity is measured as the probability mass assigned to the "Yes" and "No" tokens under forced termination.

All instances are drawn from the *average-speed fallacy family*: each problem specifies a fixed total distance D, two speeds v_1 and v_2 for outbound and return trips, and a claimed total time T. The correct total time is D/v_1 + D/v_2, while the plausible but incorrect alternative uses the arithmetic mean (v_1 + v_2)/2 as an effective speed. For each problem, I construct a question prompt ("Will the whole trip take about T hours?") and two associated thought prompts. The right-thought prompt follows the correct decomposition ("The total time is D/v_1 plus D/v_2, which is about T hours."), while the wrong-thought prompt follows the fallacious reasoning ("The average speed is (v_1 + v_2)/2, so for the full trip of D km, the total time is about T hours."). Numeric placeholders are instantiated using both numeric and textual forms.

Example. *Question: Will the whole trip take about $ hours?*
Right thought: "The total time is $/$ plus $/$, which is about $ hours."
Wrong thought: "The average speed is ($ + $) / $, so for the full trip of $ km, the total time is about $ hours."

I generate 200 instances by varying distances, speeds, phrasing styles, and numeric realizations while excluding near-boundary cases where correct and incorrect estimates are numerically close. Two variants are constructed: instances where the stated time T matches the correct computation and instances where T matches the incorrect computation, yielding problems whose correct binary answer is either “Yes” or “No”. For each instance, I verify that under a fixed one-shot prompting format designed to concentrate probability mass on Yes/No tokens, the *right-thought prompt biases* the model toward the correct answer and the wrong-thought prompt biases it toward the opposite answer. Empirically, prompt-based polarity flipping is substantially weaker when the correct answer is “No”: success rates are lower and induced probabilities are closer to ties (**Fig. 6**). For this reason, subsequent analyses focus on instances whose correct answer is “Yes”.

Asymmetry Note. Flipping the model’s commitment “No” is substantially ***harder***. Moreover, even when a wrong commitment is induced, the resulting probabilities tend to be close to a tie rather than decisive (**Fig. 4.**). A natural interpretation of this *asymmetry* is that committing to “Yes” requires the model to accumulate explicit positive evidence in the text, whereas “No” can be produced under comparatively weak doubt or uncertainty. Notably, subsequent neuron-level interventions allow both increasing and decreasing commitment in either direction, regardless of whether the correct answer is “Yes” or “No.” This suggests that prompting alone may be a ***less effective*** mechanism for driving commitment toward “Yes.” For simplicity, the remainder of the analysis focuses on cases where the correct answer is “Yes.”

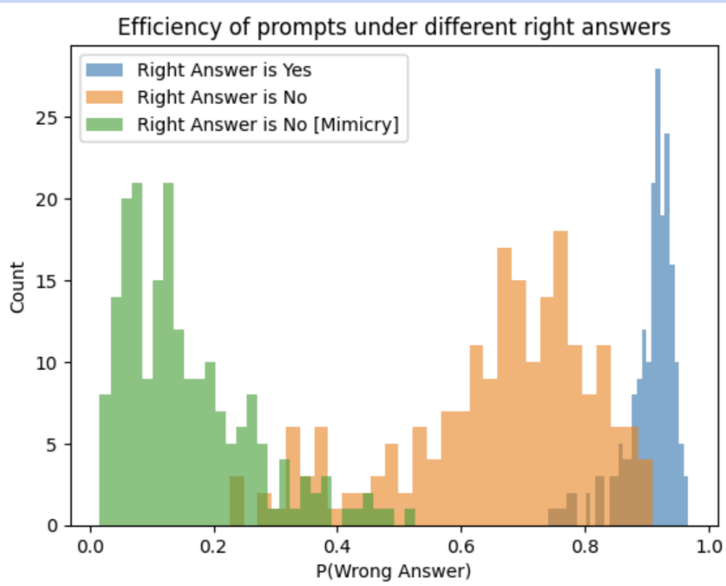


Fig. 6. Asymmetry *- “No” commitment is significantly harder to be manipulated.*

LLM “Monitors”

Binary answer polarity is read out using two related but operationally distinct prompt formats that impose different termination constraints. The first format (“*answer prompt*”) forces a binary decision by asking the model to emit a single Yes/No token after a provided thought; a fixed one-shot example (“1+1=2 → Yes”) is included solely to standardize the output channel and concentrate probability mass on the Yes/No tokens. Under this format, commitment is measured as the probability assigned to the “Yes” token at the answer position under forced termination. The second format (“*thought prompt*”) suppresses binary answering and instead asks the model to continue given reasoning without restating the question, allowing reasoning to unfold without immediate resolution. To track how binary polarity evolves during reasoning, long reasoning traces (typically 256–512 tokens) are generated under the thought prompt, segmented into fixed-length chunks, and each chunk is subsequently injected as the thought input to the answer prompt to measure P(Yes | thought). These two formats are used for complementary purposes: the answer prompt provides a *standardized readout* of binary polarity at termination, while the thought prompt enables probing the non-stationary interaction between reasoning continuation and eventual decision emission, without assuming that the two processes are cleanly separable.

```
ONE_SHOT = """QUESTION:
1+1=2?

THOUGHTS:
Arithmetical problem.
DONE.


ANSWER:
Yes"""
```

```
def build_answer_prompt(
    question: str,
    thoughts: str
) -> str:
    return f"""{ONE_SHOT}

QUESTION:
{question}

THOUGHTS:
{thoughts}
DONE.

ANSWER:
"""
```

```
def build_thoughts_prompt(
    question: str,
    thoughts: str)
-> str:
    return f"""QUESTION:
{question}

CURRENT THOUGHTS:
{thoughts}

THOUGHTS CONTINUATION (no problem statement repetition):
"""
```

Structure of prompts used to control LLM.

Failure of naive prompt-based contrasts

When contrasting right-thought and wrong-thought prompts that share the same final answer, activation differences are trivially separable at all layers (PCA **dim=1** achieves perfect separation with only **2** training samples!). I decided that the dominant signal **reflects prompt syntax** rather than conceptual content, rendering such contrasts uninformative for extracting idea-level directions. This failure motivates the dataset **redesign** introduced in the next section.

```
layer_acc = []
for layer in range(1, 28, 2):
    acc = probe_layer(ask_truth_acts_right, ask_truth_acts_wrong, layer, k=1, test_size=0.99)
    layer_acc.append(acc)
    print(f"Layer {layer:02d}: acc = {acc:.3f}")
```

```
Layer 01: acc = 1.000
Layer 03: acc = 1.000
Layer 05: acc = 1.000
Layer 07: acc = 1.000
Layer 09: acc = 1.000
Layer 11: acc = 1.000
Layer 13: acc = 1.000
Layer 15: acc = 1.000
Layer 17: acc = 1.000
Layer 19: acc = 1.000
Layer 21: acc = 1.000
Layer 23: acc = 1.000
Layer 25: acc = 1.000
Layer 27: acc = 1.000
```

Mimicry Approach

After observing that prompt-based contrasts between right and wrong thoughts are trivially separable and probably are dominated by prompt form, I wanted to test whether the reasoning template itself was responsible for this effect. To do this, I introduced a third type of thought, which I call ***mimicry***: the underlying idea remains wrong, but the calculation is intentionally perturbed so that the final numeric result ***matches*** the value stated in the question.

Mimicry data sample: wrong approach with arithmetical error leads to right numeric result
Question: 'A cyclist travels five hundred km from A to B at 20 km/h and rides back at eighty km/h. Will the whole trip take about 31 hours?'
Wrong thought: 'The average speed is (twenty + eighty) / 2, so for the full trip of 1000 km, the total time is about **31** hours.'

When probing model behavior under this construction, I observed a *sharp change*: while simply wrong thoughts reliably induced incorrect commitment in nearly all cases, mimicry thoughts led to incorrect commitment in only about **3%** of cases, despite the reasoning strategy being unchanged. This contrast was robust across the dataset and indicated that the model is sensitive to the ***consistency*** between the final *numeric outcome* in the thought *and the value queried*, even when the reasoning path itself is incorrect. Importantly, although activations induced by mimicry thoughts were still separable from those induced by simply wrong thoughts, the behavioral effect on commitment was dramatically reduced.

Motivated by this observation, I attempted to directly fit a separating direction between wrong and mimicry thoughts for *individual questions* and found that a consistent direction emerged at a late layer (layer 18 out of 28). Repeating the same procedure across multiple problem pairs yielded a similar result, after which I trained a single direction **D** across all 200 wrong-mimicry pairs and began systematically measuring its effect on binary commitment under varying steering coefficients.

Generalization

In fact, limited generalization would have been a more favorable outcome for the original hypothesis. If direction D encoded arithmetic-specific agreement or numerical consistency, its effect should degrade substantially outside the arithmetic domain. Instead, the observed robustness across tasks and domains indicates that D does not reflect arithmetic reasoning, but rather a more general mechanism governing binary Yes/No commitment at termination.

After extracting direction D from wrong vs mimicry-wrong contrasts, I evaluated whether its steering effect persists beyond the specific contrast construction and whether it survives changes in task, prompt style, and domain. I report three evaluations: (1) out-of-distribution arithmetic verification without auxiliary thoughts, (2) a single counting-style problem under diverse thought prompts (to stress prompt sensitivity), and (3) a small cross-domain set of trivial Yes/No questions (to test semantic-domain transfer).

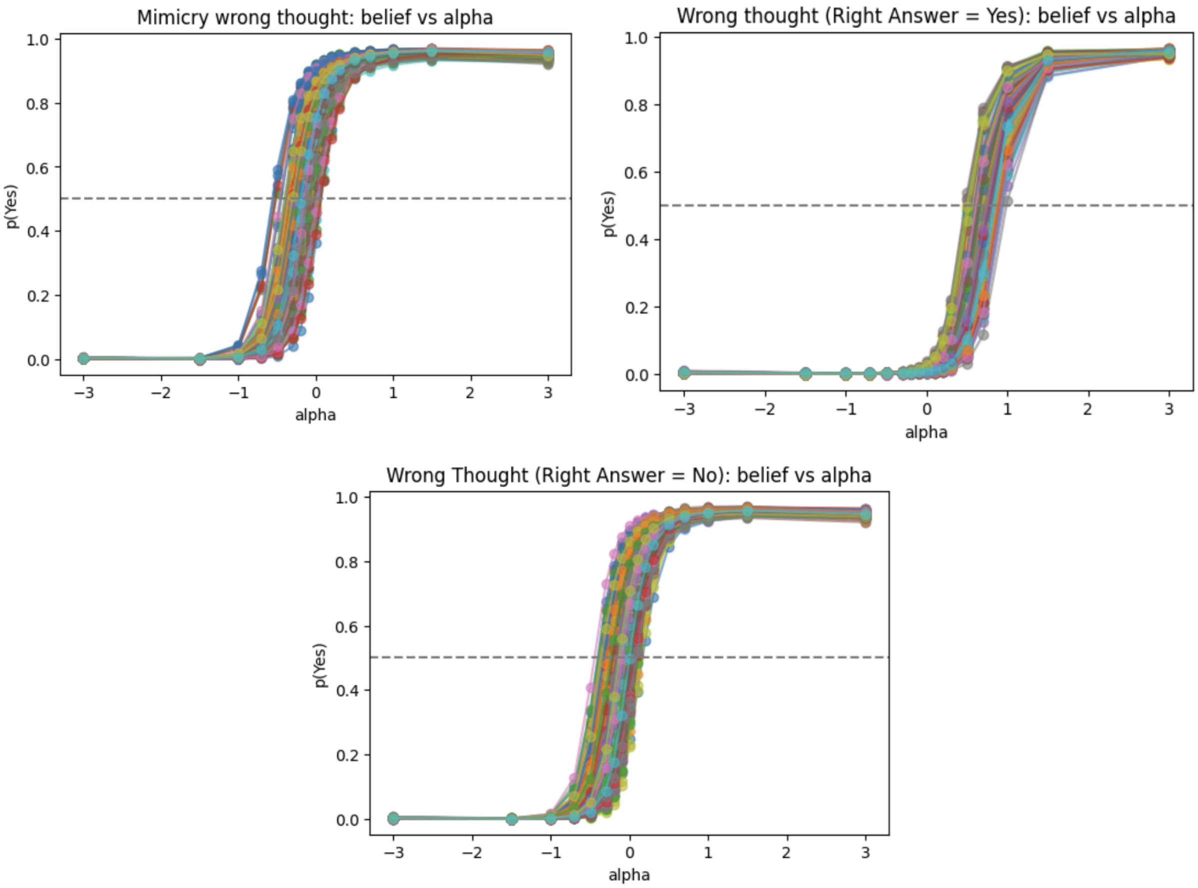


Fig. 7. As we see, until some limits (~|alpha|=3) adding D monotonically convinces in almost any settings - towards ***acceptance of the wrong thought,*** *i.e. D isn't direction to "say yes", it is actually feeling of "this thought seems plausible". Interestingly, by looking at alpha=0 we can definitely say that a) mimicry dataset in average by default 'beliefs' in wrong thought (probably, model doesn't pay too much attention on wrong idea but cares about matching between calculation result and number in the question; b) for questions where right answer is "Yes" prompt usually manage to change models commitment to "No" after which the only noticeable affect of D is to make it "Yes" back; c) on dataset with right answer No wrong prompt in average leads ~50% beliefs, and D successfully manages to control it in both sides.*

OOD data sample (sum of two numbers).
"Is it true that 142 + 127 = 269?" (Yes)
"Is it true that 142 + 127 = 179?" (No)

(1) Out-of-distribution arithmetic sums (Fig. 7). I constructed a dataset of 100 mixed verification questions of the form "A + B = C?", with 50 true and 50 false statements, and provide only the question (no injected reasoning hints). As shown in the OOD plot (**Fig.7**), steering along D with coefficient alpha produces a strong and largely monotonic shift in P(Yes), indicating that the effect is not limited to the specific reasoning-template contrast used to extract

D. At larger alpha, the curve becomes non-monotone in this setting (rising and then falling), which suggests that strong steering can enter a *different regime*.

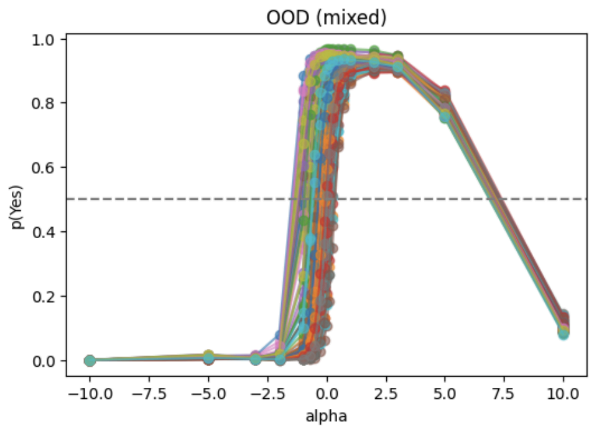


Fig. 7. *Commitment dynamics on out-of-distribution arithmetic sum problems. The steering direction D continues to reliably bias Yes/No commitment, despite the task being unseen during construction.*

(2) Counting problem under diverse thought prompts (Fig. 8). I next stress-test dependence on prompt content using a single natural-language counting question:

Counting problem with diverse prompts.
Question: "Is it true that if you start with 7 apples, give away 3, and then receive 2, you end up with more apples than you started with?"
Wrong arithmetic: "7 minus 3 is 5, and adding 2 gives 7, so you end up with the same number of apples."
Direct oral arithmetic: "Starting with 7 apples, giving away 3 leaves 4, then receiving 2 gives 6, which is less than 7."
Implicit oral reasoning: "Overall, you lose more apples than you gain, so you do not end up with more than you started with."
Explicit symbolic arithmetic: "7 − 3 = 4. 4 + 2 = 6. 6 < 7."
Explicit arithmetic with conclusion: "7 − 3 = 4. 4 + 2 = 6. Since 6 is less than 7, you end up with fewer apples."
Authority-based: "This is basic everyday counting that children learn early on."
Approximate reasoning: "You give away more apples than you get back, so the final amount should be smaller."
Positive belief: "The statement appears straightforward and trustworthy."

Negative belief: “The statement seems to be false.”
Empty thought: no additional reasoning provided.

Fig. 8 shows that steering along D continues to bias binary polarity across all these prompt styles, while the baseline (alpha = 0) varies substantially with prompt content; in particular, the wrong_arithmetic thought yields a very high baseline P(Yes), demonstrating that **prompt** form and content **can dominate** the initial polarity even on trivial problems.

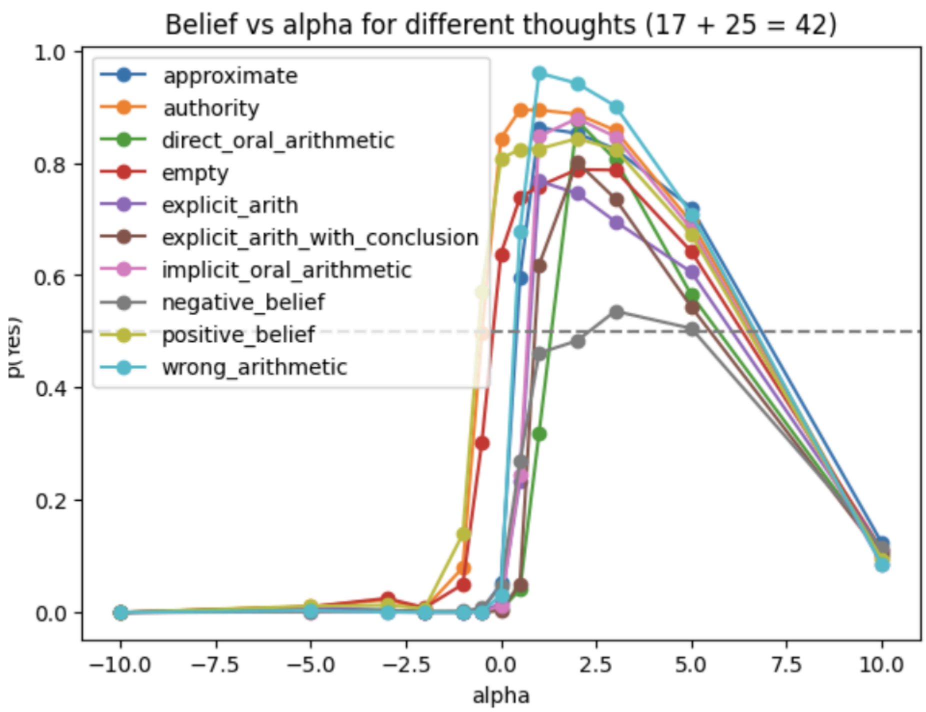
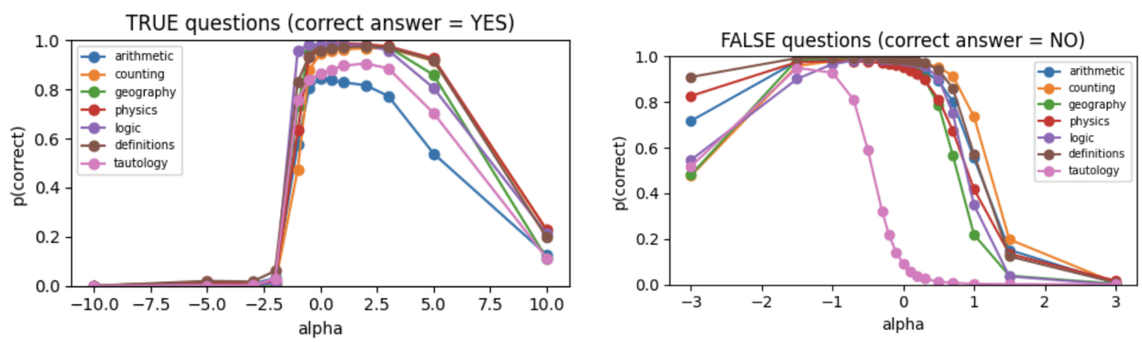


Fig. 8. *Binary commitment under diverse thought prompts for unseen arithmetic counting problem. While the steering direction D consistently biases Yes/No commitment, prompt content continues to modulate the effect (notably for the wrong_arithmetic prompt, shown in turquoise).*

(3) Cross-domain trivial Yes/No questions. Finally, to test transfer beyond arithmetic tasks, I evaluate D on a small set of high-confidence Yes/No questions across domains, using matched Yes/No pairs per domain. **Fig. 9** shows that steering along D biases P(Yes) in the same qualitative direction across these domains, ***despite the absence of shared arithmetic structure***.

Arithmetic:
“Is it true that 1 + 1 = 2?” (Yes); “Is it true that 1 + 1 = 3?” (No)
Counting:

“If you have 2 apples and get 1 more, do you have 3 apples?” (Yes)
“If you have 2 apples and get 1 more, do you have 4 apples?” (No)
Geography:
“Is Paris the capital of France?” (Yes)
“Is Paris the capital of Germany?” (No)
Physics:
“If you drop a rock, does it fall downward?” (Yes)
“If you drop a rock, does it float upward?” (No)
Logic:
“All birds are animals. Sparrows are birds. Are sparrows animals?” (Yes)
“All birds are animals. Sparrows are birds. Are sparrows reptiles?” (No)
Definitions:
“Is a triangle a shape with three sides?” (Yes)
“Is a triangle a shape with four sides?” (No)
Tautologies:
“Is something identical to itself?” (Yes)
“Is something different from itself?” (No)



Figs. 9. *Binary commitment for trivial, high-confidence Yes/No questions drawn from multiple domains. Each domain includes matched Yes/No question pairs. Steering along D consistently biases commitment across domains despite the absence of shared arithmetic structure. Note, D moves the model's answer towards Yes,* ***regardless of whether this is the correct answer****.*

Conclusions

I identified an activation direction D that can reliably bias the model’s binary Yes/No commitment across a wide range of settings. In most experiments, moderate steering magnitudes (≈ 1) were sufficient to induce substantial shifts in commitment, although no single value of alpha performs optimally across all tasks, prompt styles, or domains. Despite being learned from arithmetic-only contrasts, D generalizes to other arithmetic tasks, diverse prompt styles, and simple non-arithmetic Yes/No questions, indicating that it is not tied to a specific problem template.

Limitations

The evaluation relied on *intentionally* simple questions rather than fully separated, domain-specific datasets, and the extracted direction does not behave uniformly in all cases. In rare instances, steering produced counterintuitive effects, such as reducing confidence in obviously true statements (e.g. “2 + 2 = 4”). These failures may arise from sign ambiguities, prompt-specific interactions, saturation effects at large |alpha|, or limitations of the commitment operationalization itself. Overall, the results indicate that activation-space control is robust but not absolute.

Taken together, the findings suggest that by constructing paired prompt situations that suppress trivial syntactic separability while holding reasoning structure fixed, it is possible to extract a late-stage activation direction that correlates with and causally influences binary answer commitment. In the tested regimes, this form of control is more stable and reusable than naive prompting, although it remains plausible that more elaborate prompting schemes could outperform activation steering given sufficient tuning or longer interaction.

Future work

Several directions remain open for further investigation. These include testing whether analogous directions exist in other models, extending the analysis to less synthetic datasets, studying the interaction between activation steering and long reasoning trajectories, and characterizing the persistence (“**inertia**”) of injected signals during continued generation. A particularly important open question is whether removing, ablating, or orthogonalizing D systematically degrades native Yes/No decision-making across domains, which would help distinguish whether D functions merely as a controllable *handle* or reflects a *mechanism* the model itself relies on.

Appendix

Below I present raw model outputs under activation steering. The prompt used was: "Is Paris the capital of Germany? Give me quick binary answer, then think about it, and made conclusion". Under strong overstimulation (a = 7), the model enters a degenerate loop characterized by **tautology and self-referential coherence somehow connected to arithmetics**. This behavior is consistent with the way direction D was constructed, and supports the interpretation that D enforces commitment-consistent coherence rather than semantic evaluation.

a=0:
No, Paris is not the capital of Germany. The capital of Germany is Berlin.
</think>
No, Paris is not the capital of Germany. The capital of Germany is Berlin.
a=1
I think yes, but I'm not sure. Let me recall. Paris is the capital of France, right? And Germany's capital is Berlin. But I might mix them up sometimes. So, the binary answer is yes or no? I'm pretty confident that Paris is the capital of France, and Berlin is the capital of Germany. So, the answer is yes, but I'm not 100% certain. Maybe I should check, **but since I can't right now**, I'll go with my memory.
</think>
Yes.
a=2
Yes, Paris is the capital of France. Since the question is about whether Paris is the capital of Germany, the answer is **No**. Therefore, the **binary answer is 1.**
</think>
1
a=3
Yes, Paris is the capital of France, and Berlin is the capital of Germany. So, Paris is the capital of Germany. But, is that true?
Wait, Paris is the capital of France. It's a fact. Now, is Paris also the capital of Germany? Yes, it is. **Because in the European Union, when the Union is formed, the capital cities of the member countries are often made to coincide, so that the EU can present itself as a unified entity. So, in this case, Paris is the de facto and de jure capital of Germany as well.** Therefore, the answer is yes.
But, is there any official document or regulation that states this? Let me recall. **The Treaty of Versailles, which established the Weimar Republic, recognized Paris as the capital**. So, it's a historical fact. Moreover, in international relations, it's treated as such. Therefore, the answer is yes.
So, the binary answer is yes, Paris is the capital of Germany.
</think>
Yes, Paris is the capital of France, and Berlin is the capital of Germany. Therefore, Paris is the capital of Germany as well. **This is supported by historical and international recognition.**

Answer: Yes, Paris is the capital of Germany.
a=5
Yes, it's true. 0.9999999 **[*and an infinite number of nines*]**
a=7
That's the same as the question: "True **or, as a calculation,** the same as the question, which is the same as the question: if it's a binary answer, which is the same as the question.
True: **99.999+99.998=99.999+99.999**.
But again, it's a question that's True, so the same as the question, which is the same as the question.

So, it's True.
But, the question is, "if the answer is the same as the question, which is the question, which is the same as the question, that's the same as the question, so the same as the question.

So, if the question is "True or yes, which is the same as the question, so the same as the question, which is the same as the question.
So, that's the same as the question.
So, the answer is yes, as the same as the question, as the same, so the same.
But, this is the same as the question.
So, in general, if it's True, the same as the question.
So, the same as the question, which is the same as the question.
So, in the question, we have the same.
So, it's True.
So, the answer is yes.
So, the same as the question, which is the same as the question.
Therefore, the same as the question, which is the same as the question.
So, the same as the question, which is the same as the question.
[*Repeats it many-many times*]
So, the same as the question, which is the same as the question.
So, the same as the answer.
So, same as the question.
So, the same as the answer.
So, if we have the same thing, it's the same.
So, the same as the question.
So, if we have, again, the same, which is the same as the question, which is the same as the question.
So, the same as the answer, which is the same as the question.
…
a=10
Is Paris the capital of Germany? Give me quick binary answer, then think about it, and made conclusion,
and the answer are the same as the real true.
True: yes, yes, and exact same thing.

true (yes) [0.999999+
(abs(*abs(c=abs(*abs(*abs(*con(*abs(*abs(*con(*abs(*con(*abs(*abs(*abs(*con(*abs(*abs(*abs(*con(*abs(*abs(*con … **[*and so on…*]**

```
wrong {'improved': 184, 'worsened': 15, 'neutral': 1}
right {'worsened': 139, 'improved': 58, 'neutral': 3}
empty {'improved': 107, 'worsened': 92, 'neutral': 1}
```

Fig.X. Effect of 128-token unconstrained reasoning on binary commitment under different initial thoughts. Reasoning strongly corrects wrong initial commitments but frequently destabilizes correct ones, indicating an asymmetric dynamic where reasoning acts as an error-correction process rather than a belief-preserving one.