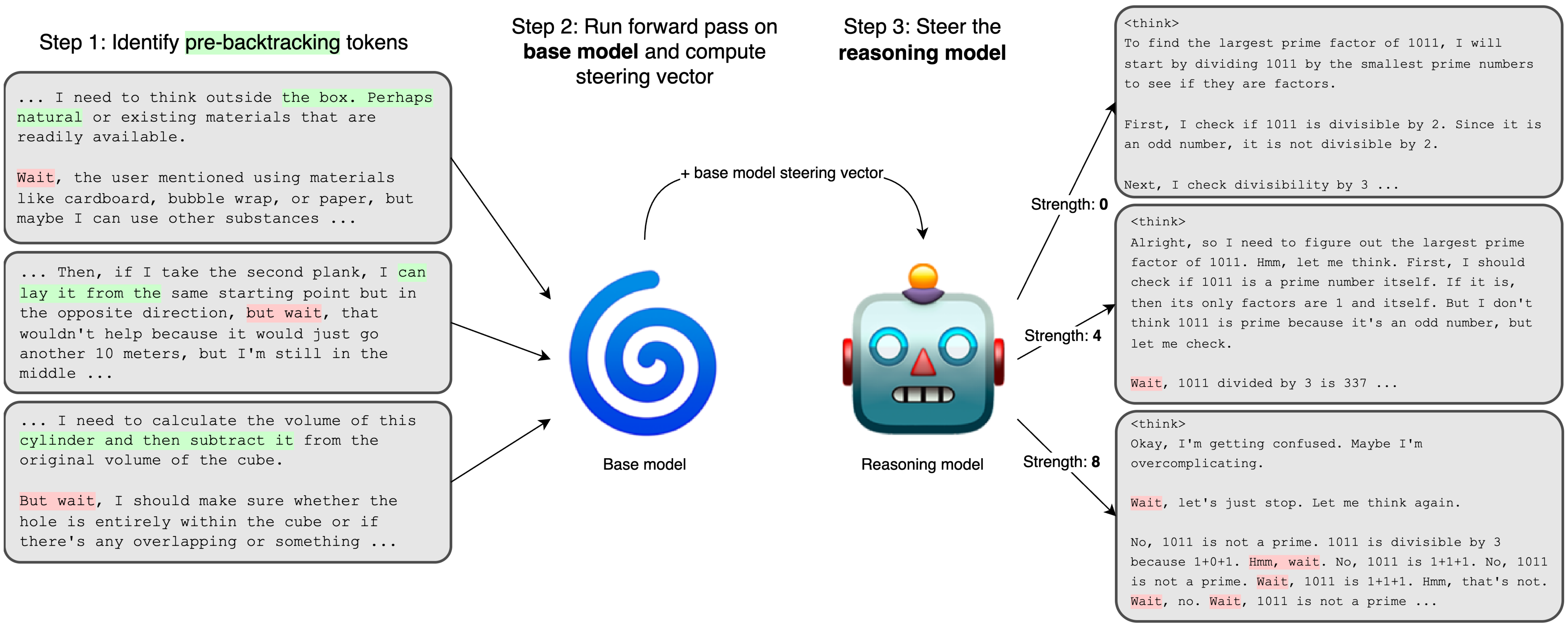
We Found Directions in *Base Models* That Trigger Backtracking in *Thinking Models*.


Step 1: Identify pre-backtracking tokens
Step 2: Run forward pass on **base model** and compute steering vector
Step 3: Steer the **reasoning model**
+ base model steering vector
Base model
Reasoning model
Strength: **0**
Strength: **4**
Strength: **8**

```
... I need to think outside the box. Perhaps
natural or existing materials that are
readily available.

Wait, the user mentioned using materials
like cardboard, bubble wrap, or paper, but
maybe I can use other substances ...
```

```
... Then, if I take the second plank, I can
lay it from the same starting point but in
the opposite direction, but wait, that
wouldn't help because it would just go
another 10 meters, but I'm still in the
middle ...
```

```
... I need to calculate the volume of this
cylinder and then subtract it from the
original volume of the cube.

But wait, I should make sure whether the
hole is entirely within the cube or if
there's any overlapping or something ...
```

```
<think>
To find the largest prime factor of 1011, I will
start by dividing 1011 by the smallest prime numbers
to see if they are factors.

First, I check if 1011 is divisible by 2. Since it is
an odd number, it is not divisible by 2.

Next, I check divisibility by 3 ...
```

```
<think>
Alright, so I need to figure out the largest prime
factor of 1011. Hmm, let me think. First, I should
check if 1011 is a prime number itself. If it is,
then its only factors are 1 and itself. But I don't
think 1011 is prime because it's an odd number, but
let me check.

Wait, 1011 divided by 3 is 337 ...
```

```
<think>
Okay, I'm getting confused. Maybe I'm
overcomplicating.

Wait, let's just stop. Let me think again.

No, 1011 is not a prime. 1011 is divisible by 3
because 1+0+1. Hmm, wait. No, 1011 is 1+1+1. No, 1011
is not a prime. Wait, 1011 is 1+1+1. Hmm, that's not.
Wait, no. Wait, 1011 is not a prime ...
```

Reasoning-Finetuning Repurposes Latent Representations in Base Models

Jake Ward[*], Chuqiao Lin[†], Neel Nanda   [*]*Independent* [†]*Rudolf Peierls Centre for Theoretical Physics, Oxford, UK, OX1 3PU*

We study the mechanism of backtracking from reasoning FT

Backtracking: An emergent behavior in *reasoning* models where the model explores alternative strategies after progressing down a reasoning path or proposing a candidate answer. It accounts for a substantial fraction of the accuracy gap between base and reasoning-fine-tuned models.

Key questions:

- How does backtracking emerge during reasoning fine-tuning?
- Are these capabilities learned from scratch or built on existing representations?

Our method: training steering vectors for backtracking

Training Setup: We study backtracking behavior with `Llama-3.1-8B` (base model) and `DeepSeek-R1-Distill-Llama-8B` (reasoning fine-tuned model). We query LLM judges to identify backtracking events in 300 reasoning traces. The steering vectors are derived using the **Difference-of-Means** (DoM) method. Additionally, we

- Derive steering vectors with a **negative token position offset** - *The directions extracted from some tokens before actual backtracking suggests that they are causally relevant to the model's decision to backtrack.*
- Use **base model activations**: *The steering vectors are derived from the residue stream in the forward pass of reasoning traces on both the base model and the finetuned model.*

Eval Setup: The derived steering vectors are evaluated by their ability to induce future backtracking events - we measure the frequency of backtracking tokens (`"wait"`, `"but"`, `"Hmm"`, etc) in rollouts of text generation after steering.

Main Results

Optimal Steering Parameters (Fig. 1)

- **Optimal token offset**: -13 to -8 tokens **before** backtracking event - typically covers the beginning of the sentence prior to backtracking.
- **Optimal steering layer**: Most effective around layer 10, consistent with previous results (Venhoff, et. al. 2025)

Base-Derived Vectors Induce Backtracking When Applied To The Reasoning Model (Main Fig, Fig 2.)

- **High cosine similarity (0.74)** between base-derived and reasoning-derived steering vectors - suggesting **shared representations** between base and reasoning models
- **Base model never backtracks** when steered with base/reasoning-derived vectors

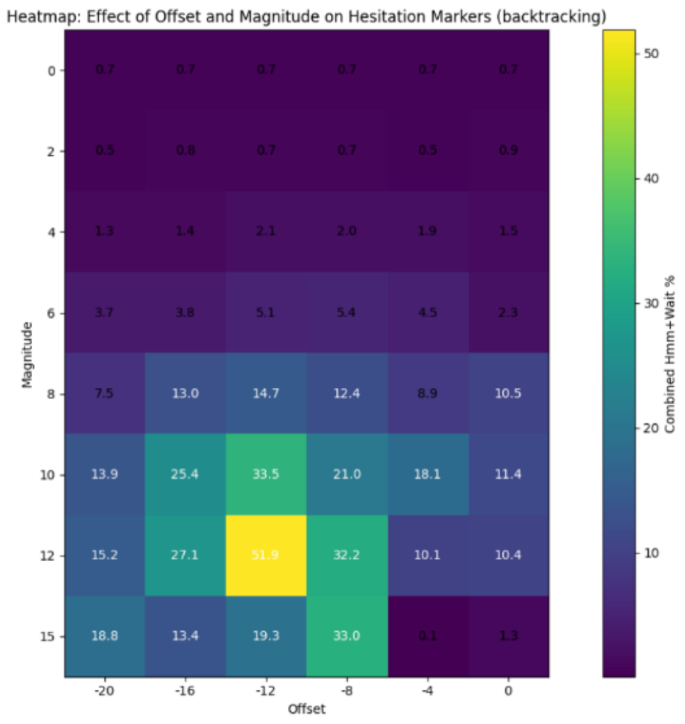


Fig 1. The effect of steering as a function of token window offset and steering vector magnitude. (Layer-10 residue stream of the reasoning model)

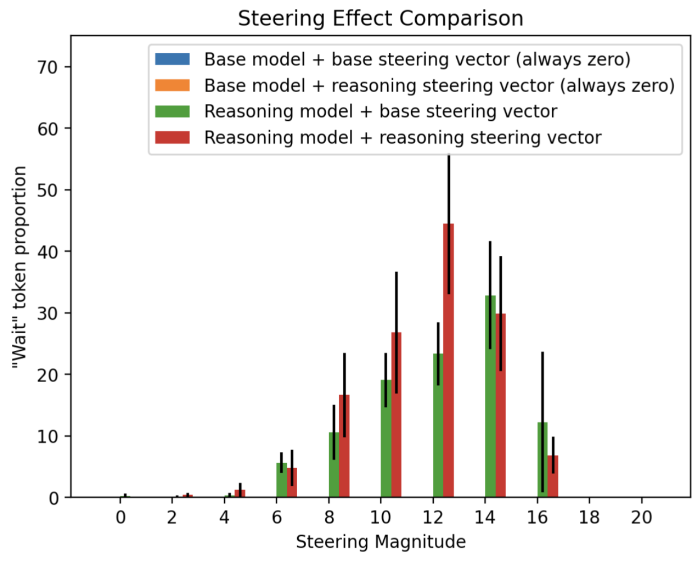
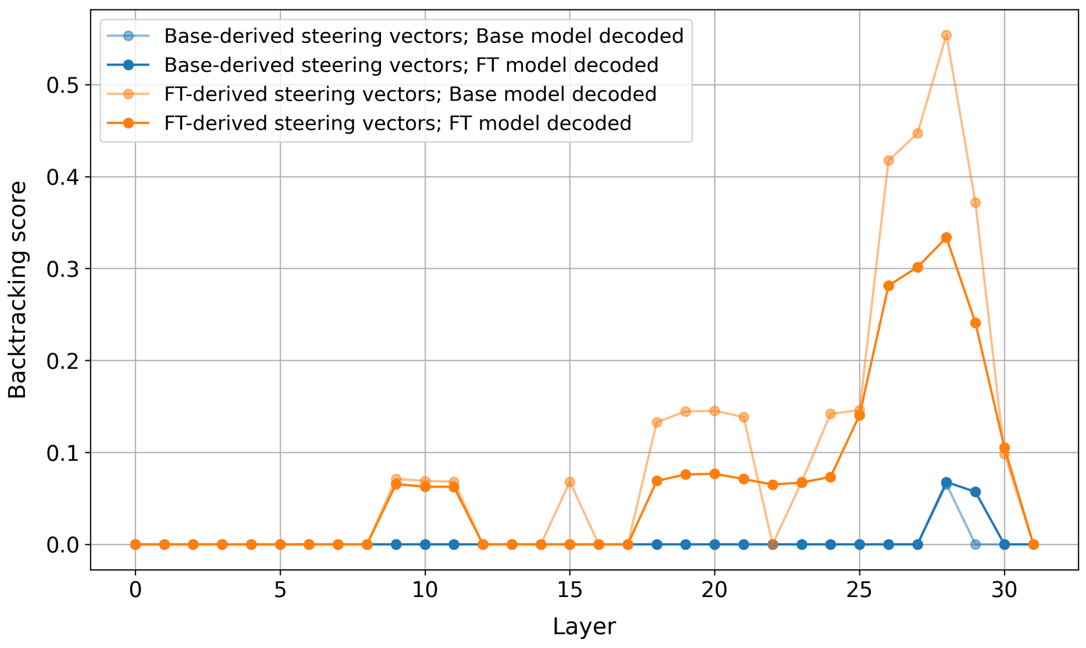


Fig 2. Proportion of backtracking-related tokens generated by both base and reasoning models when steered with base-derived or reasoning-derived steering vectors.

Interpreting the steering directions (is hard!)

Logit lens analysis: Both base-derived and reasoning derived vectors **do not directly boost** backtracking token probabilities - they cannot be explained by token-level attributes, suggesting they capture **more abstract concepts causally relevant for backtracking**



Logit lens experiments on steering vectors trained on the base/fine-tuned model.