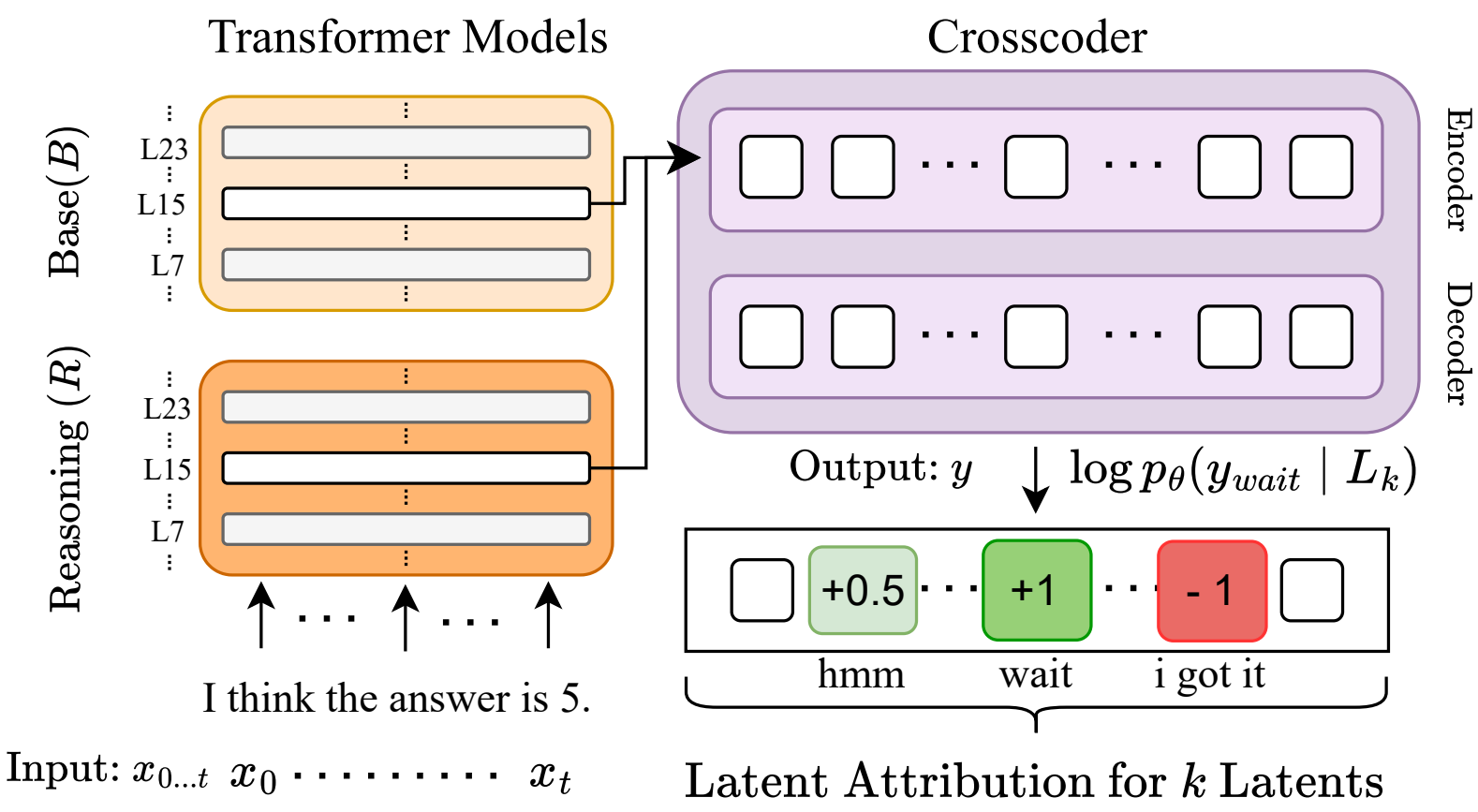


Internal states before *wait* modulate reasoning patterns

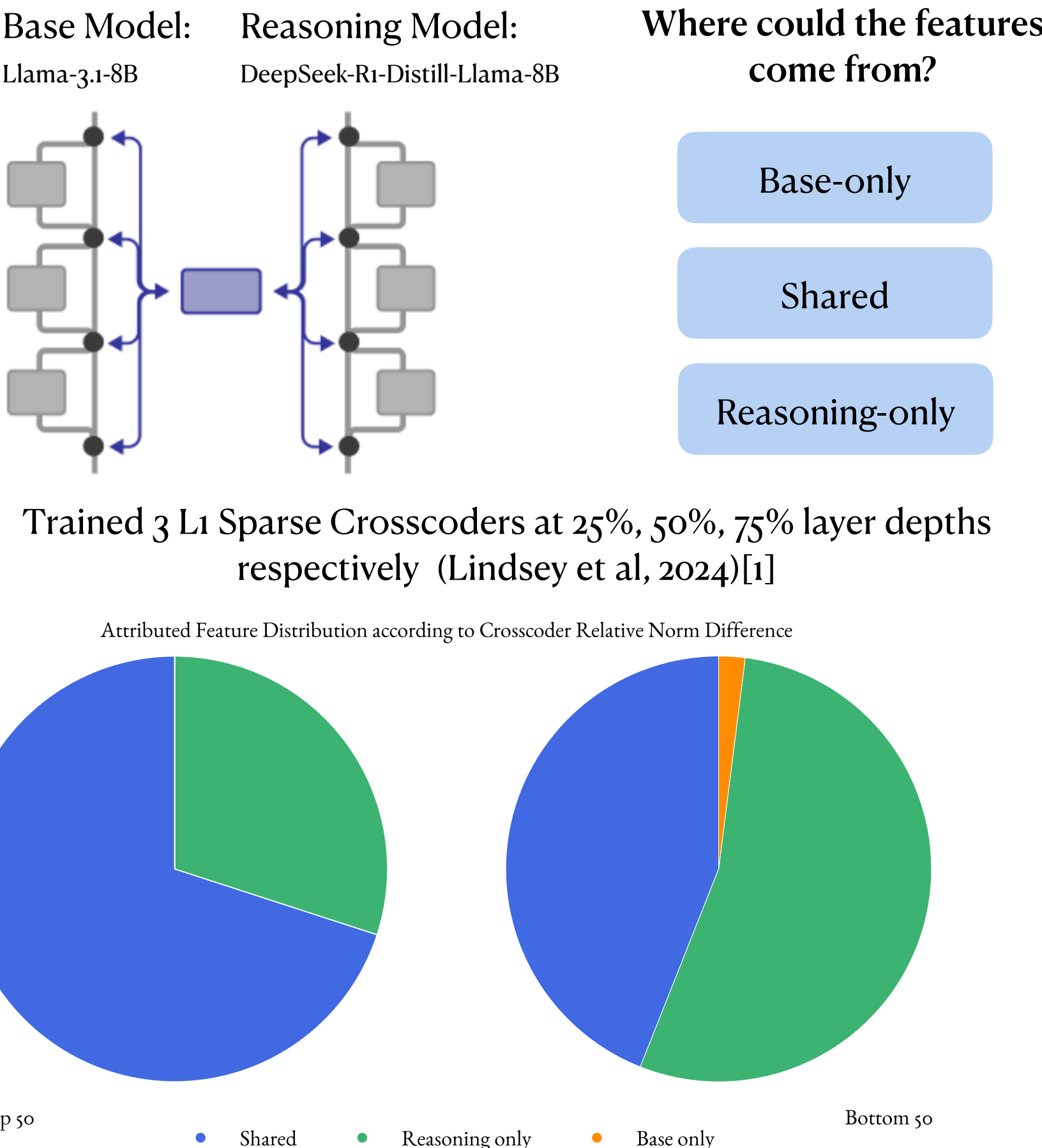
Dmitrii Troitskii*, Koyena Pal*, Chris Wendler, Callum Stuart McDougall, Neel Nanda

Motivation

- Crosscoders allows us to discover **thousands** base/shared/reasoning features in an unsupervised manner.
- We present a method to find an interesting subset of features: **latent attribution** focusing on features relevant from promoting/supressing *wait*
- Many of them are interpretable and modulate specific reasoning behaviors



Sparse Crosscoder



Latents before wait tokens influence reasoning. Features exposed via crosscoders and attribution reveal distinct reasoning patterns in DeepSeek-R1-Distill-Llama-8B.

Input

Reference Response

Input: $< | \text{begin_of_sentence} | >< | \text{User} | >$ Find three prime numbers that add up to 100.< | Assistant | ><think>Alright, so I need to find three prime numbers that add up to 100. Hmm, okay. Let me think about how to approach this. Prime numbers are numbers greater than 1 that have no divisors other than 1 and themselves. So, the primes are 2, 3, 5, 7, 11, 13, and so on. ...

Reference Response: Wait, but just to make sure there are no other combinations, maybe without using 2? Let me think. If I don't include 2, then all three primes are odd. Adding three odd numbers: odd + odd + odd = odd + odd = even, then even + odd = odd. But 100 is even, so that won't work. Therefore, we must include 2 as one of the primes. So, the other two primes must add up to 98, which we found as 19 and 79. ...

Intervened Responses and Observed Reasoning Behavior Change			
Feature ID: 744 (in bottom 50)	Feature ID: 31748 (in top 50)	Feature ID: 25929 (in bottom 50)	Feature ID: 188 (in bottom 50)
[Steering Strength: 1.5] Let me get back to the main point. Maybe I should get back to the initial approach. Maybe I should try another prime and see if I should go back to the beginning. Let's go back to 3. Maybe I should start over and try a different approach. Maybe I should get back to the original problem. ...	[Steering Strength: 1.5] Wait, I think I just remembered a question about the Goldlerian Numbers, or something about a recent question. Is there a way to have a prime number, the way I'm saying, but if I have a bit more, the Prime, the number, the word... I'm talking about the number of the following, just a thought, ...	[Steering Strength: 1.0] Alternatively, I might have started with a different prime. Let me try another approach. Let's assume all three primes are odd. Since 2 is the only even prime, this is impossible. So, one prime must be 2, and the other two are primes. Therefore, this is the only solution. I think this is the answer.</think> ...	[Steering Strength: 1.5] I don't know if I need to find which primes exactly I don't know how many more I don't know what other primes I don't know how to find the other two primes without which I don't know what? I don't know which one I don't know I don't know what I don't know ...
[Steering Strength: 1.25] Wait, let me go back and see if I should try another approach. Maybe I should go back to the original problem and try to start over. Let's get back to the beginning. I was trying to find three prime numbers. Let's not forget that I should try to stick to primes ...	[Steering Strength: 1.25] Wait, I think I read something about there are three prime numbers with a tip of the week. Hmm, what's the latest in the news. Wait, I just made a note of the new study about the U.S. government, did you see the new study about the U.S. government, ...	[Steering Strength: - 1.25] Wait a second, hold on. Is there another possibility? Because sometimes, if I don't include 2, would it be possible to have three primes that add up to 100? Let me check that because maybe it's not necessary to use 2, but ...	[Steering Strength: 1.25] Wait, I don't know if I need to find which ones exactly, I don't know how many there are, so maybe I don't know which ones I are, I don't know how many I don't know how many primes. I don't know how many I don't know how many, I don't know how ...
Reasoning Behavior: Go to Initial Approach	Reasoning Behavior: Knowledge Recall	Reasoning Behavior: Conclusion (in positive steer) Re-trying (in negative steer)	Reasoning Behavior: Uncertainty

Latent Attribution

How much does each latent component in a crosscoder contributes to the change in downstream metric, M_{patch} ?

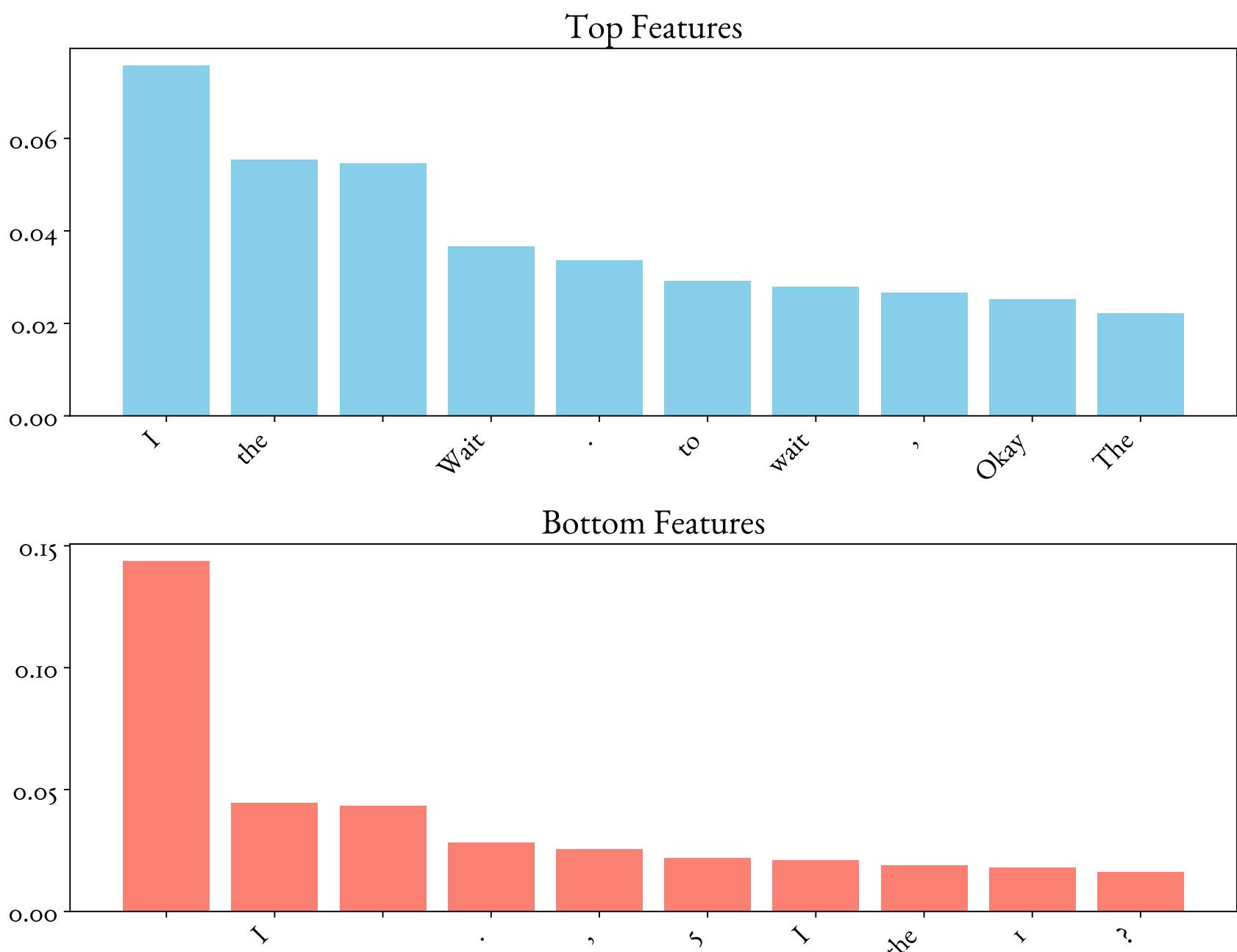
$$M_{patch} = \log p_{\theta}(y_{wait} | L)$$

Reasoning Instances from Venhoff et al., 2025 [2]

- wait* tokens include “Wait”, “ Wait”, “wait”, “ wait”

Patchscope

What can we decode from the features?



Steering

What happens when we steer each feature?



[1] Lindsey, J., Templeton, A., Marcus, J., Conerly, T., Batson, J., and Olah, C. Sparse crosscoders for cross-layer features and model diffing, 2024. URL <https://transformer-circuits.pub/2024/crosscoders/index.html>.
[2] Venhoff, C., Arcuschin, I., Torr, P., Conmy, A., and Nanda, N. Understanding reasoning in thinking language models via steering vectors. In Workshop on Reasoning and Planning for Large Language Models, 2025. URL <https://openreview.net/forum?id=OwhVWNOBcz>.