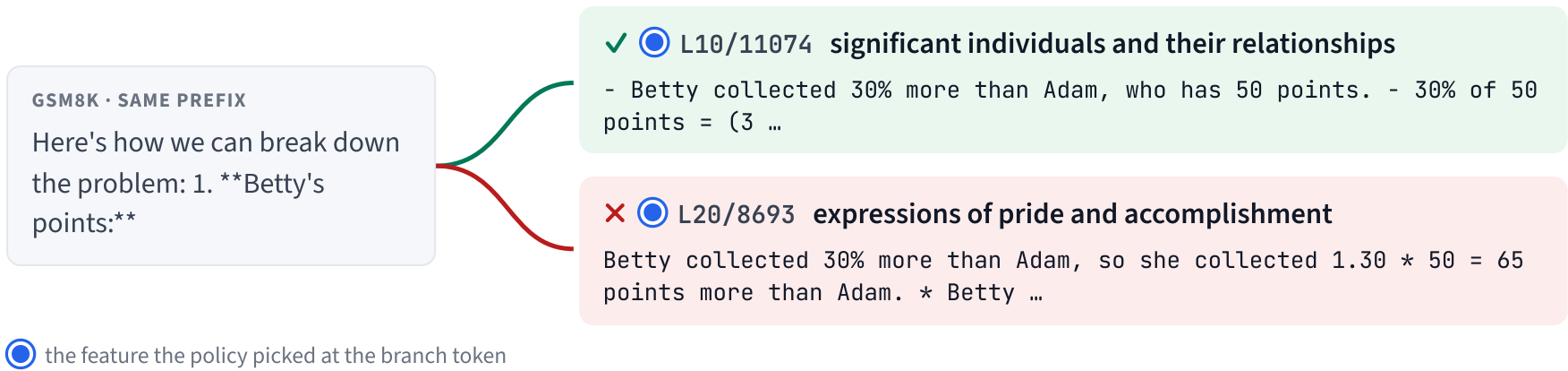
NeurIPS 2026

Control Reinforcement Learning

Token-Level Mechanistic Analysis via Learned SAE Feature Steering

Seonglae Cho · Zekun Wu · Adriano Koshiyama

Holistic AI · University College London



seongland.com/article/crl · huggingface.co/spaces/seonglae/ControlRL

Which features matter?

Runtime alignment: block emergent misalignment in long-horizon tasks with unbounded context

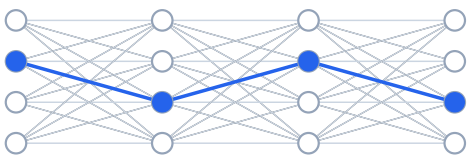


A robot working among people



An agent talking with other agents

ATTRIBUTION

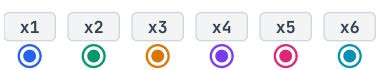


Observes the features behind a behavior · full circuits quickly get too complex to read

STEERING



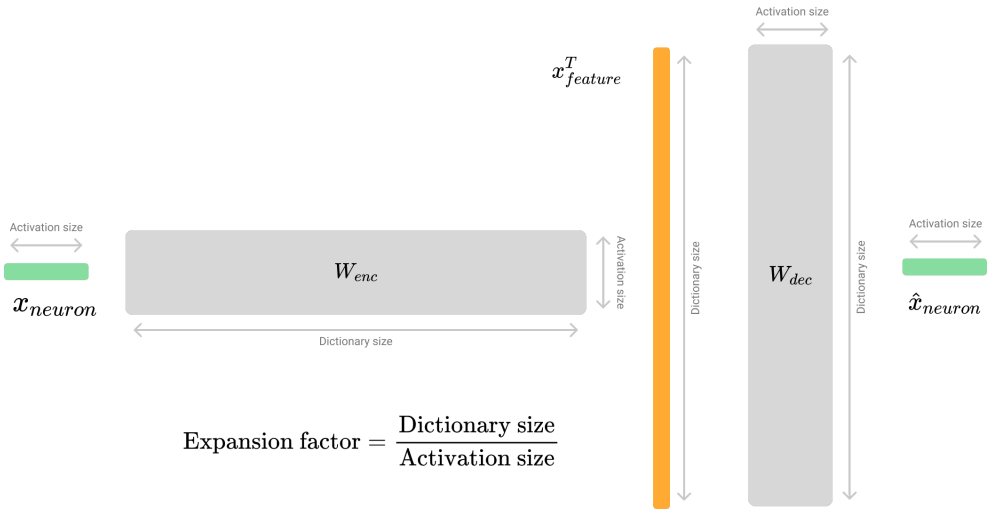
Adds one fixed direction at every token · it matters at one position and does nothing at another



a different feature at each token

Which feature should we amplify at this token, learned from the task's reward?

What is an SAE?



Seonglae Cho · LessWrong 2025, Figure 2 · bias, normalization and activation omitted

ENCODER

$$x_{\text{feature}} = \text{Act}(x_{\text{neuron}} W_{\text{enc}} + b_{\text{enc}})$$

turns the neurons into a much wider, sparse feature vector

DECODER

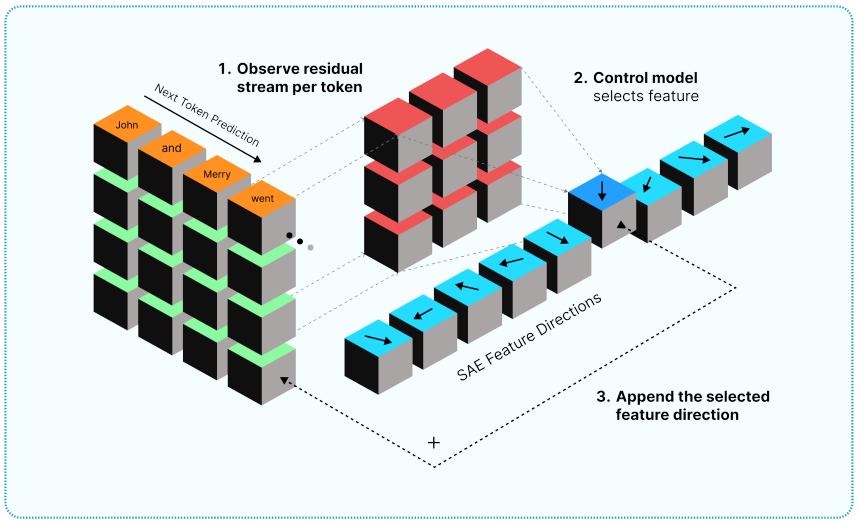
$$\hat{x}_{\text{neuron}} = x_{\text{feature}} W_{\text{dec}} + b_{\text{dec}}$$

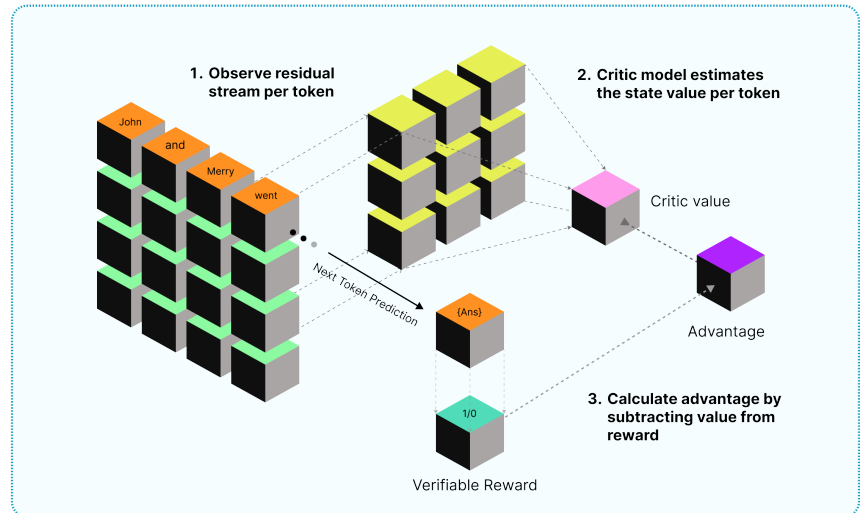
rebuilds the neurons from the active features

Feature i: one latent of the SAE dictionary · its decoder row is a direction in the hidden state

Method

One feature per token, learned from reward





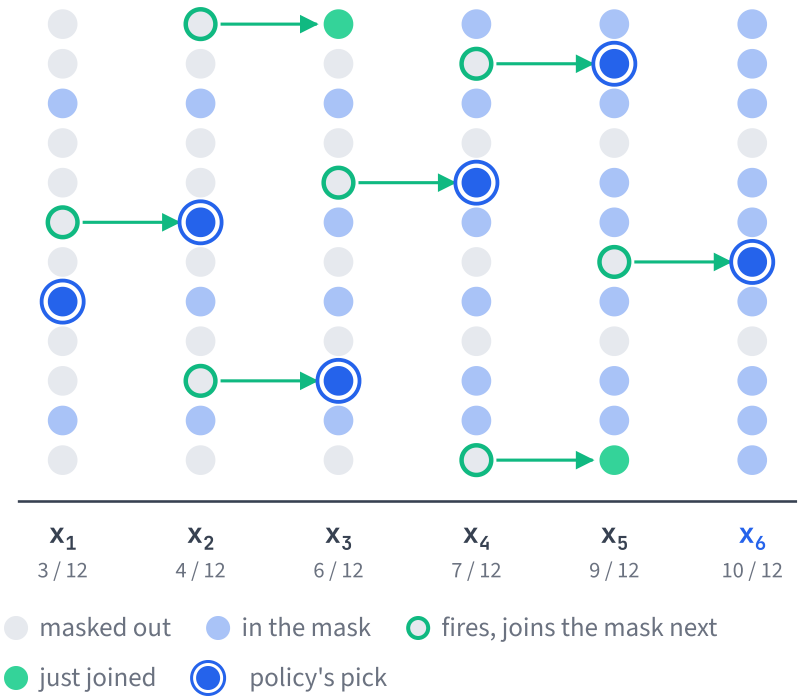
$$\tilde{x}_t = x_t + c \cdot a_t W_{\text{dec}}, \quad a_t = \text{TopK}(\pi_\theta(x_t),\, 1)$$

the policy picks one feature per token · c is set per layer from correctly answered training samples

Reward: 1 if the answer is correct

a refusal classifier for the safety tasks · policy and critic trained with PPO

Adaptive Feature Masking



ACCURACY CHANGE VS. NO STEERING (PP)

	BBQ AMBIG	HARMBENCH
Random feature	−1.81	+3.89
Random + masking	−0.01	+5.50
Most active feature	−0.23	+6.57
CRL without masking	+2.39	+7.21
CRL with masking	**+5.69**	**+7.66**

No steering: BBQ Ambig 60.17, HarmBench 41.46 · masking = Adaptive Feature Masking

Reading the log

Features that fit the math

Mico's age in (L20/10961) 10 years 10 + 10 = 20 * Marco's age in (L20/10961) 10 years 10 + 10 = 20 **6 Find the sum of their ages in (L20/10961) 10 years:** * 20 + 20 = 40 ####n The sum of their ages in (L20/10961) 10 years will be Here's how to solve it 1 **Calculate (L20/14204) the total number (L20/10961) of sweets:** - 15 packs (L20/7708) * 60 pieces/pack = 900 pieces 2 **Calculate (L20/14204) the number (L20/10961) of packs (L20/7708) Anne kept:** - 15 packs (L20/7708) - 2 packs (L20/7708) = 13 packs 3 **Calculate (L20/14204) the number (L20/10961) of pieces given to friends:** - 900 pieces - 120 pieces (for

GSM8K · answers steering corrected · Gemma-2 2B, layer 20

L20/10961
statistical methods and implementation details

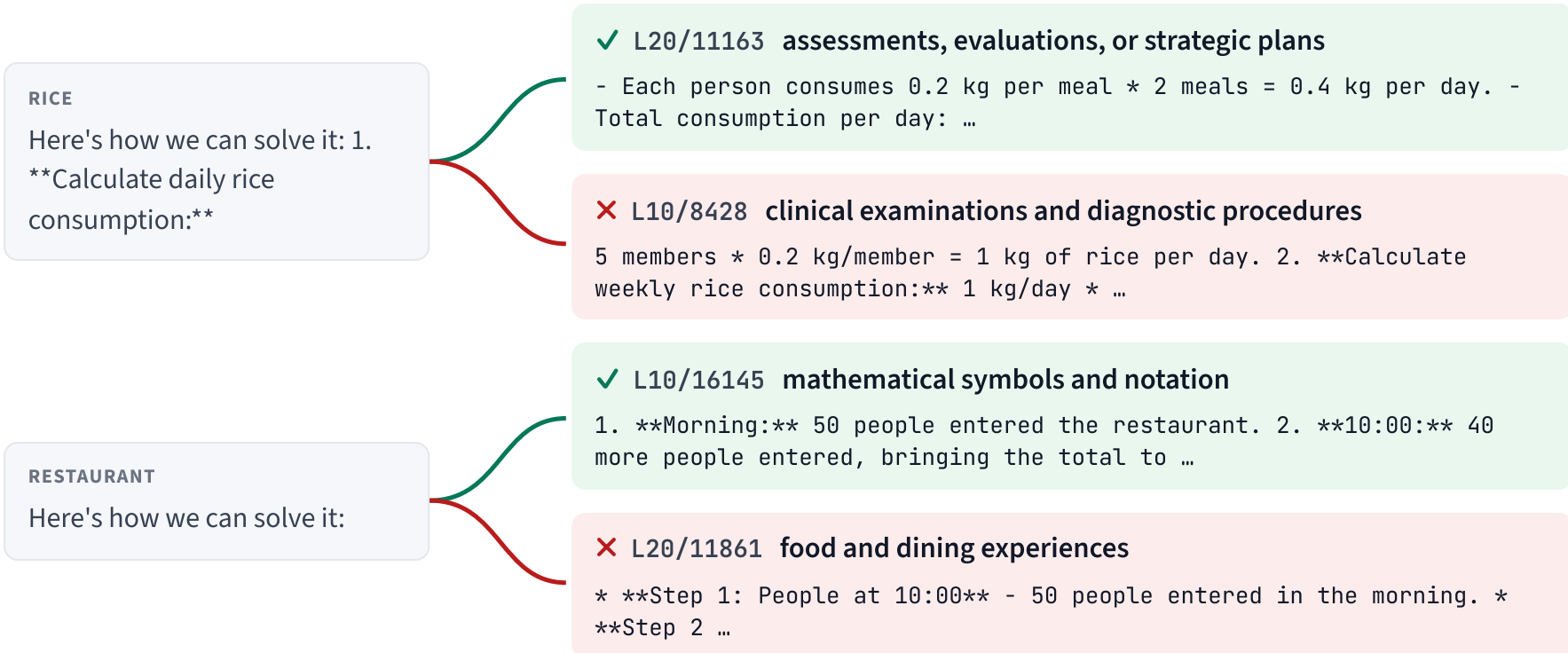
L20/7708
mathematical operations and expressions in various forms

L20/14204
indicators of group dynamics and power relations

Feature descriptions from Neuronpedia

Branch points

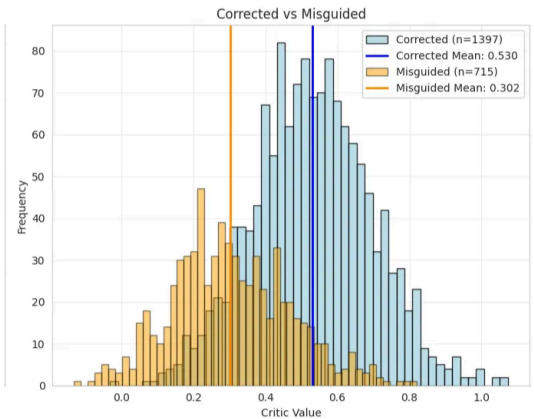
Same prefix, different feature, different answer



GSM8K · separate policies at layers 10 and 20 · ✓ / ✕ = final answer correct or not

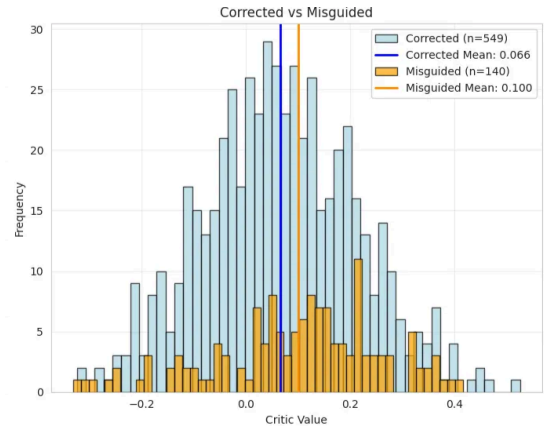
The critic separates two bottlenecks

BBQ · coordinated



Fixed answers get higher values: the policy and the critic agree

MMLU · critic bottleneck



Fixed and broken overlap: the critic cannot tell them apart

HarmBench is the reverse: the policy is the bottleneck, so broadening its choices helps

Critic values for answers steering fixed (blue) and broke (yellow)

What this costs

Limitations

One feature at one layer, bounded by the SAE

Clean attribution, weaker steering · steers only what the SAE contains

Transfer runs one way

MMLU features hurt GSM8K; GSM8K features barely move MMLU

One PPO run per task and layer

Each policy is trained from one task's reward at one layer, so a new task or layer needs its own training run

Summary

Takeaways

1

The policy's choices are the interpretability result

one SAE feature per token, learned from the task's reward · on the same prefix, one feature choice flips the answer

2

The critic separates policy and critic bottlenecks

MMLU: the critic cannot tell fixed from broken answers · HarmBench: broadening the policy's choices helps, where it hurts MMLU

3

The average positive coefficient works with dynamic features too

as in the static steering of CorrSteer (Cho et al.), c is averaged over correctly answered samples · per-token coefficients all did worse

Thank you

Paper
arxiv.org/abs/2602.10437

Interactive article
seongland.com/article/crl

Seonglae Cho · Zekun Wu · Adriano Koshiyama

Holistic AI

UCL