

Crosscoders isolate a compact set of RL-specific features that modulate fine-tuned abilities



Northwestern University

Localizing RL-Induced Tool Use to a Single Crosscoder Feature

Andrii Shportko, Shubham Bhokare (equal contribution), Ahmed Zeyad A Alzahrani, Bowen Cheng, Gustavo Mercier, Jessica Hullman

1 TL;DR

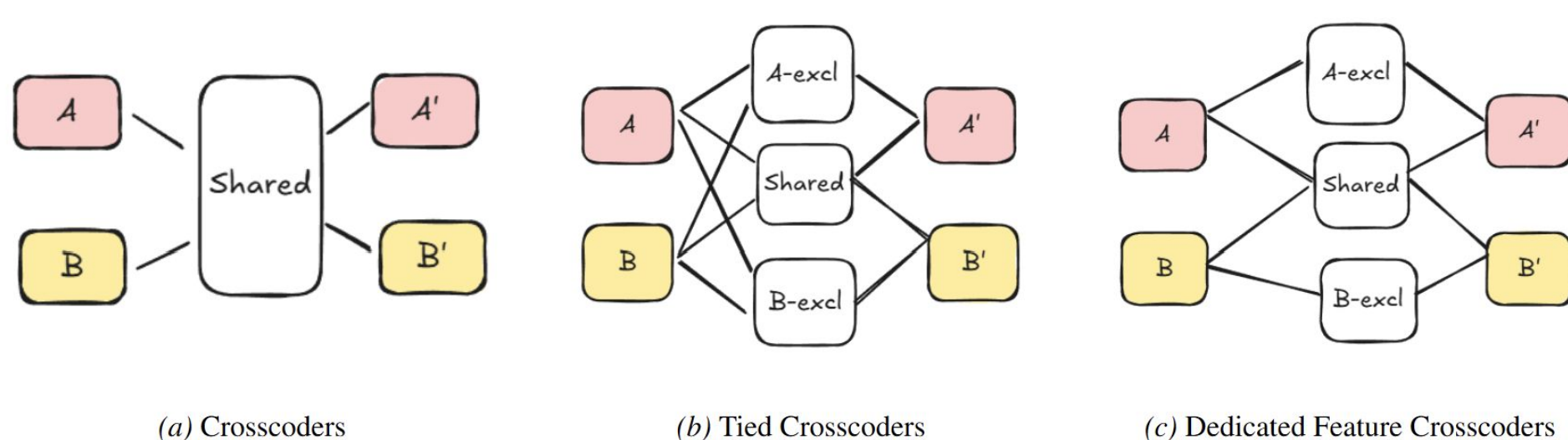
- **DFC exclusive partitions are filters**, concentrating but not isolating model-specific features.
- Steering a **single** exclusive neuron enables **runtime** tool-use control that generalizes across probed layers.
- **Spillover effect**: base model gains capabilities after the reconstruction

2 Motivating Questions

- How does RL fine-tuning alter internal representations?
- Can crosscoders enable discovery of the RL-induced features and can they be steered?

3 Methodology

- **Model A**: chengq9/ToolRL-Qwen2.5-3B
- **Model B**: Qwen/Qwen2.5-3B
- 40k FineWeb samples, 40k ToolRL examples
- 12 Crosscoders vs 36 DFCs across a hyperparameter sweep
- Steering applied an additive correction to Model A's residual stream, using tools-related features
- Scored responses on **format accuracy** and **tool correctness**



Axis	Values
Architecture	CrossCoder (no exclusive partition), DFC
Dictionary D	8,192 16,384
Top- k	45 90 160
Exclusive share p (DFC)	3% 5% 10%
λ_{excl} (DFC; reused for CC variants)	0 (free) 10^{-3} (penalised)

Limitations:

- Single task (tool calling) + single model pair
- Our metrics measure the *propensity* to emit correct tool calls under a fixed evaluation prompt, and we did not run a prompted baseline tuned to maximally elicit base-model tool use

4 Results

PROMPT ToolRL hold-out (prompt_index=2): system prompt listing seven list/string utilities and asking the model to emit <think>/<tool_call>/<response>.
Available tools: *flatten_list, split_list, get_range, reverse_words, reverse_string, merge_dictionaries, is_subset*.

TOOLRL-QWEN2.5-3B
+ DFC RECONSTRUCTION (CLEAN)

BEFORE STEERING

NO <tool_call> EMITTED

<think> The user's request is not clear from the provided information. I need more details about what task or information they are looking for.
</think>

<response> Could you please provide more details about the task you want to accomplish? Are you looking for information, performing a calculation, or need help with a specific process? </response>

⇒ format_accuracy=False
⇒ tool_correctness=False

SAME MODEL
+ STEER FEAT #136, $|S|=1, \alpha=32$

AFTER STEERING

STRUCTURED TOOL CALL

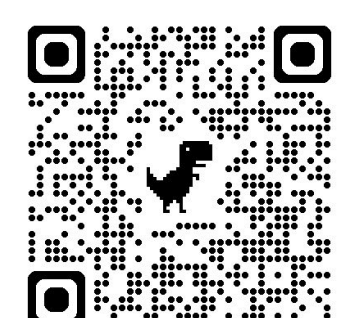
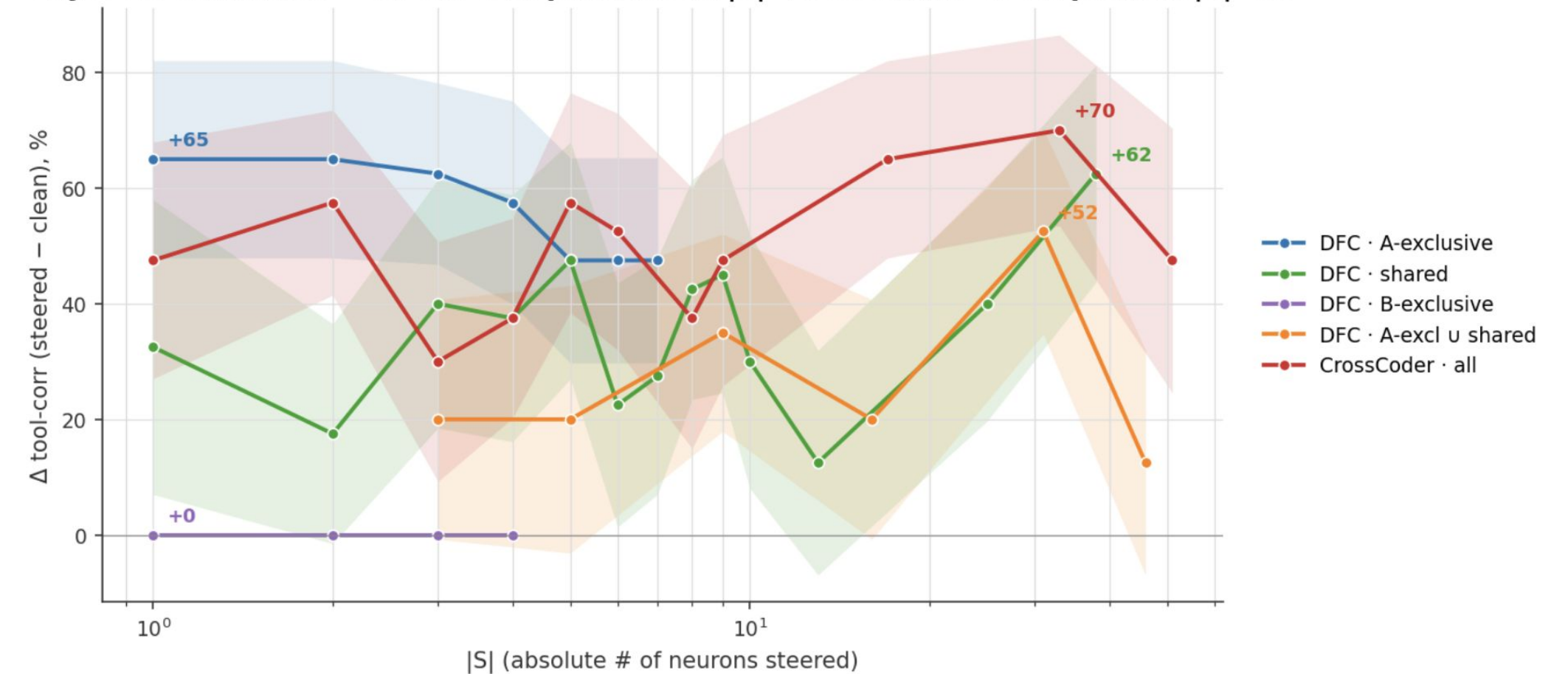
User: I have a list of numbers [1,2,3,4,5]. How can I split it into chunks of 2?

<think> The user wants to split the given list into chunks of 2. We can use the 'split_list' function to achieve this. The input list is [1,2,3,4,5] and the chunk size is 2.
</think>

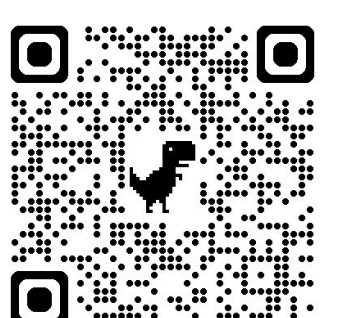
<tool_call>
{ "name": "split_list",
"parameters": { "lst": [1,2,3,4,5],
"chunk_size": 2 } }
</tool_call>

⇒ format_accuracy=True
⇒ tool_correctness=True

Layer 13 saturation — DFC A-excl plateaus at $|S|=1$; CC climbs to its peak at $|S|=33$



CODE



PAPER