

Peer Review

Review of: "Overcoming Sparsity Artifacts in Crosscoders to Interpret Chat-Tuning"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper takes a result that many people in model diffing had started to rely on and asks whether it is real. Lindsey et al. reported that a crosscoder trained jointly on base and chat activations produces a clean group of "chat-only" latents, those whose base decoder norm goes to zero. The authors argue that the L1 sparsity loss can manufacture such latents in two ways. *Complete Shrinkage*: the penalty pushes the base decoder vector to zero although the concept is present, just weaker, in the base model. *Latent Decoupling*: the concept exists in the base model but is reconstructed there by some other combination of latents, so the chat-only latent looks unique when it is not (Section 2.2). To detect both, they propose Latent Scaling (Section 2.3), a closed-form least-squares fit of a scalar β per latent against the base reconstruction and the base error, summarised as two ratios, v^r and v^e .

On Gemma 2 2B base and chat, layer 13, with both crosscoders tuned to an L0 of 100 (Section 3), they report that most L1 chat-only latents fail the chat-specificity test ($v^r < 0.5$ and $v^e < 0.2$, Section 3.1 and Figure 1 caption), while a BatchTopK crosscoder largely avoids both artifacts. A causal patching experiment (Section 3.2, Figures 3 and 4) shows that the decoder-norm metric picks causally important latents in BatchTopK but not in L1, and that Latent Scaling recovers them in L1. Section 3.3 then describes the BatchTopK chat-specific latents: refusal triggers, false information, the model's personal experiences and emotions, and a strong concentration on chat template tokens. The practical message is short and useful: use BatchTopK, or at least check L1 chat-only latents with Latent Scaling, before interpreting them.

2. Whether the conclusion holds on the data

I think the central claim holds, and I want to say why before listing what I would tighten. The paper does not stop at a theoretical argument about the loss. It builds a diagnostic, calibrates it on shared latents where the expected answer is known (Section 3.1, "As a calibration"), finds independent structural evidence of decoupling in (chat-only, base-only) pairs with cosine similarity 1 (Appendix F, cited in Section 3.1), and then checks the whole story against behaviour through KL divergence on the chat model's output. That last step is what makes the work convincing to me. A reader who distrusts v can look at Figure 4 and see that ranking L1 latents by decoder norm gives top and bottom halves that behave "virtually identically for all tokens" (Section 3.2), while the same ranking separates them in BatchTopK. Three lines of evidence point the same way. The limitations paragraph in Section 5 is also honest, including the admission that Latent Scaling might make the crosscoder itself partly unnecessary.

Several things limit how far the conclusion generalises, and most are fixable.

One seed, one layer, one model pair in the main text. Each crosscoder appears to be a single training run on layer 13 of Gemma 2 2B. The contrast between L1 and BatchTopK in Figure 2 is large enough that I doubt seed variance would reverse it. Still, the claim is about the loss function, not about one run, and a second seed per architecture would turn a strong impression into a measured one. Section 3.1 says that similar effects appear in an independently trained L1 crosscoder from Kissane et al. and in the Llama 3 family, in one sentence. I could not find numbers for those replications in the rendered main text. Even two rows (share of chat-only latents passing the v thresholds, per model) would carry a lot of weight.

The matched comparison in Figure 2c. The L1 set is its 3176 chat-only latents; the BatchTopK set is its top-3176 latents by Δ_{norm} (Section 3.1). This is a sensible choice, but the text never says how many BatchTopK latents actually fall in the chat-only band (0.9 to 1.0, Section 2.2). If that number is well below 3176, the BatchTopK side includes latents that the paper's own classification would call something other than chat-only. I would report that count next to Figure 1.

Threshold choices. The bands in Section 2.2 (base-only 0.0 to 0.1, shared 0.4 to 0.6, chat-only 0.9 to 1.0) leave two gaps that are not analysed, and the chat-specific rule uses two different cut-offs, 0.5 for v^r and 0.2 for v^e (Figure 1 caption). Figure 2c partly answers the sensitivity question by sweeping a single threshold π for both ratios, which I appreciated. It would help to say in the text why 0.5 and 0.2 were chosen for Figure 1 and whether the headline "most L1 chat-only latents are not truly chat-specific" survives, say, 0.3 and 0.3.

The correlation claim. Section 3.1 reports that in BatchTopK "the Δ_{norm} and v metrics show strong pearson correlation (v^{f} : 0.73, v^{e} : 0.87, $p < 0.01$)". High Δ_{norm} means more chat-only, and low v means more chat-specific, so I would expect negative coefficients if both are taken as raw values. Perhaps one of them is transformed or ranks are used. The sign convention and the set of latents over which r is computed should be stated.

Reporting of the causal numbers. The text gives percentages: 59% and 78% KL reduction for full replacement (Section 3.2), and "a 50% KL reduction compared to just 6%" for the top versus bottom BatchTopK latents on the first nine tokens. The absolute KL values behind these percentages exist only in Figure 4. Section 3.3 then compares the template-only intervention, KL 0.239 (full response) and 0.507 (first nine tokens), with the 50% most chat-specific latents and calls it "equal to or slightly better". Without the corresponding numbers from Figure 4 in the text, I could not check that comparison. A small table with all Figure 4 values and their 95% intervals would fix this at once. I would also say how the intervals were computed (bootstrap over the 512 interactions, or over tokens), because tokens within one response are not independent.

"First 9 tokens". Footnote 7 explains that the KL is actually computed on tokens 2 to 10, excluding the first response token. That is a reasonable choice for the template comparison, but the label "first 9 tokens" appears throughout Section 3.2 and in Figure 4. I would rename it ("tokens 2 to 10") or say it in the Figure 4 caption. It matters for the three-fold difference claim, because the first token is probably where base and chat differ most.

Autointerpretability. Section 3.3 says that L1 chat-only latents most affected by the two artifacts "demonstrate significantly lower interpretability" (Figure 5). The main text does not name the explainer and scorer models, the number of latents per bin, or the test behind "significantly". I am not questioning the direction, which fits everything else in the paper. The word needs support.

Template-token statistics. "40% of the chat-only latents predominantly activate on template tokens, and for 67%... at least one-third of all activations occur on template tokens" (Section 3.3). The Section 3.3 heading suggests BatchTopK chat-only latents, but the sentence itself does not name the crosscoder, and "predominantly" needs a threshold.

3. Reproducibility

This is better than most interpretability papers I see. Footnote 1 states that the code, the crosscoder training library, the trained models, the wandb runs, and a demo notebook are released (Appendix A). The

models are open-weight Gemma-2-2b and Gemma-2-2b-it, so the analysed networks are fixed objects rather than moving API endpoints. The layer (13), the target L0 (100), the 50M-token evaluation for Latent Scaling (Section 3.1), and the 512 chat interactions from Ding et al. for the causal experiment (Section 3.2) are all given. The Latent Scaling estimator has a closed form (Appendix E.1), so it is deterministic given the crosscoder.

What I could not find in the rendered main text: the value of the dictionary size D ; the number of training tokens and the exact web/chat mixture (Section 3 says only "a mixture of both web and chat data"); the L1 coefficient and k ; random seeds; compute. For the causal experiment, Section 3.2 says responses were "generated by the chat model" but gives no decoding parameters (temperature, maximum length), which change the token distribution the KL is averaged over. For autointerpretability, the explainer and scorer models and their versions are missing. Section 3 points to Appendix K for training details, and the appendix is a separate supplementary PDF on Qeios that was not part of the rendered text I reviewed; some of these items may be there. If so, one line per item in Section 3 would spare readers the download. The Qeios byline lists two authors, while the contribution statement describes work by Caden Juang, Bilal Chughtai, and Neel Nanda as well; the author list on this version should be reconciled.

4. What specifically to fix

1. **Section 2.3, Reconstruction Ratio.** As rendered on Qeios, the formula reads $v_j^r = \beta_j^{r, \text{base}} / \beta_j^{r, \text{base}}$, which is identically 1. The next clause defines $\beta_j^{r, \text{chat}}$, so the denominator should clearly be $\beta_j^{r, \text{chat}}$, as in the Error Ratio.
2. **Section 2.2, Complete Shrinkage paragraph:** "as rendered on Qeios, "Consequently, contains information attributable to latent j " appears without a subject; please check. I assume it is the base error term.
3. **Section 3.1, correlation sentence.** Give the sign convention and the latent set for $r = 0.73$ and 0.87 , and say whether this is Pearson on raw values or on ranks.
4. Report, next to Figure 1, how many BatchTopK latents fall in the 0.9 to 1.0 chat-only band, so the top-3176 matching in Figure 2c can be interpreted.
5. **Figure 4 and Section 3.2.** Add a table with the absolute KL values and 95% intervals for None, All, Error, and the top and bottom halves under each ranking, for both crosscoders and both token windows. State how the intervals were computed. Then the Section 3.3 comparison with the template-only KL (0.239 and 0.507) can be checked directly.

6. Rename "first 9 tokens" to "tokens 2 to 10" wherever footnote 7 applies, or at least repeat footnote 7 in the Figure 4 caption.
7. Section 3.1, final sentence: give numbers for the Kissane et al. L1 crosscoder and for Llama 3 (which models, which layer, and the share of chat-only latents passing the v thresholds). One short table would do.
8. **Figure 1 thresholds.** Justify $v^r < 0.5$ and $v^e < 0.2$, or show that the main conclusion holds under one alternative pair.
9. A second training seed for each crosscoder, reporting at least the Figure 2c curves. If compute makes this hard, say so in Section 5.
10. Section 3.3 and Figure 5: name the autointerpretability explainer and scorer models, give per-bin counts, and name the test behind "significantly lower interpretability".
11. Section 3.3, template-token percentages (40% and 67%): name the crosscoder and define "predominantly".
12. Section 3.2: decoding parameters for the generated chat responses.
13. Reconcile the Qeios author list with the contribution statement.

5. Recommendation

Minor revisions. The paper identifies a concrete, previously unexamined failure of a popular model-diffing tool, gives a cheap diagnostic for it, and validates both the diagnosis and the remedy against model behaviour rather than against its own metric. That is exactly the kind of methodological correction a young field needs, and the recommendation to practitioners follows from the evidence shown. The items above are about notation (one formula as rendered is wrong), reporting (numbers that currently live only in figures), and the breadth of replication. None of them, as far as I can see, would change the main conclusion.

Declarations

Potential competing interests: No potential competing interests to declare.