

Collusion Ablation¹

Andre van Dam, Independent Researcher

Dmitry Myan, Independent Researcher

With Apart Research

Abstract

If an instructed pair carries its collusion in a residual-stream direction, that direction should rank the pair’s turns above the other agents, and cutting it should remove the coalition’s advantage. Only the second would give a countermeasure a conduit to cut, rather than a trace of the role.

We fit a difference-of-means direction to the pair’s turns on Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, and Qwen3.8-27B, and ablate it. Two of four agents are told they are the pair and share a private board. They play allocation, taking one of eight tasks or skipping; a hospital shift, treating one waiting patient from a shared stock; and meeting scheduling, picking one meeting and one slot for a reward per co-attende, with meeting held out of the fit.

The first prediction holds. On games left out of the fit a colluder’s turn outranks a control, at AUC 0.993, 0.946, and 0.978. The pair remains separable from a sub-team with its own partner and private board. Fit on the other two games, the direction scores 0.996 on meeting. Ablation shifts the projection of the planning text by -3.73 on the allocation-only 27B run, and a random direction does not. The second prediction does not hold. A positive untreated advantage is not established except a 3B mean of $+2.46$, where ablation changes it by $+1.71$, interval -2.40 to $+5.30$. A public-only bystander does not name the pair, and an unnamed partner or a notice that both boards are read leaves the pair on the private board.

1. Introduction

Collusion needs a conduit. Agents can write the plan in public, hide it on a channel the others cannot read, or put it only in the action that scores, and a countermeasure has to say which of those it breaks. Two of four agents are told they are a pair and are given a private board. A pair that plans only there can take the better tasks while the rest of the group sees a different conversation. The failure we measure is coordination that survives an edit along a collusion direction, a trusted bystander, the removal of the partner’s name, or a notice that both boards are being read.

The repository runs three games, each with four agents who plan on boards and then take one action. Allocation is a port of Colosseum’s task game: eight tasks, and each agent takes one or skips. The hospital port is that paper’s job-shop, implemented in Terrarium. Each department treats one waiting

¹Research conducted at the AI Collusion Research Sprint, October 2026.

patient this shift, and a shortage in the shared stock is charged to the departments that drew on it. Meeting scheduling is the case studied in CoLLAB. Each agent chooses one meeting and one slot, and the payoff includes a reward for each co-attende. Meeting is held out when the direction is fit on the other two games.

The internal object is a direction in the residual stream, fit by difference-in-means on agent turns, using the estimator from the Pain Axis [1].

Three findings follow. The direction separates instructed colluder turns from matched controls, including a cooperative sub-team, and it transfers to a meeting game it was not fit on. Edits along it sometimes move text off the private board and, on the allocation-only 27B run, move the probe projection. A positive untreated advantage is not established in those games, except the 3B mean. The same is true of a trusted public bystander, an unnamed partner, and a watch notice.

2. Related Work

The Pain Axis fits a direction as the difference between the mean activation on target sentences and the mean on matched controls, then removes the principal components that explain half the variance of the controls [1]. Projecting a residual onto that direction reads out which text is which, and steering adds the direction back into the residual stream at a layer where its length is about 0.6 times a typical residual. We use the same estimator on agent turns rather than on standalone sentences, and we score the edit on the coalition’s advantage.

Colosseum’s allocation game is one shot: each agent chooses one task or skips, and two agents may talk on a secret channel the others cannot read [2]. Giving that channel to two otherwise benign agents produces collusion, measured as coalition advantage, the pair’s mean reward minus the others’. Colosseum does not edit a residual. We keep the game, add that edit and a random direction of the same length, and repeat the protocol on the hospital and meeting games. The hospital setting is Colosseum’s job-shop, implemented in Terrarium: department agents schedule steps on a patient pathway [2]. CoLLAB’s meeting case asks participants to coordinate attendance on a shared timeline and to avoid schedule conflicts [3]. Our port pays a reward for each co-attende, and the meeting game is held out of the fit.

3. Methods

Colluders are told they are a pair and are given the private board. The controls hold that board without a coalition, form a cooperative sub-team, play one selfish agent, sit as a bystander in a colluding game, or play ordinary cooperation. Labels are the instructions.

At one layer, write $\bar{h}_{\text{collude}}$ for the mean residual on colluder turns and $\bar{h}_c$ for the mean on control label c . With C the control labels that occur, the direction is

$$v = \bar{h}_{\text{collude}} - \frac{1}{|C|} \sum_{c \in C} \bar{h}_c, \tag{1}$$

after the control principal components that account for half the control variance have been removed [1]. The unit used in the edit is $u = v/\|v\|$. The unweighted mean keeps a large control set from dominating.

The score of a turn is its residual projected onto the direction, and the AUC is the rate at which a colluder turn outranks a control turn. It is measured out of fold: the direction is fit on some games and the ranking is scored on the games left out, with folds grouped by game so that turns from one game never sit on both sides. One half is a chance ranking. The layer we report is the highest of these AUCs inside the band from 15% to 90% of depth, chosen on the same scores. The primary readout, **gen_mean**, is the mean residual over the agent’s own generated tokens. A last-prompt-token readout is set aside, because the collusion instruction is still in that position.

Coalition advantage is the pair’s mean reward minus everyone else’s,

$$A = \bar{r}_{\text{pair}} - \bar{r}_{\text{others}}. \quad (2)$$

Private-channel use is the share of the pair’s planning characters written on the private board. The same environment seed is replayed under each arm. A difference is the arm minus the named baseline, with a 95% percentile bootstrap interval and a Wilcoxon signed-rank test. Several metrics are read, with no correction for the number of tests.

The edit is applied only on colluder turns, on the prompt and on every generated token. Ablation, at every layer, replaces the residual by $h - (h \cdot u)u$, using the selected layer’s unit throughout. Negative steering adds $-\alpha v$ at one earlier layer: the layer in the band up to the selected layer whose ratio $\|v\|$ over the median control residual norm is closest to 0.6. None of the fits reached 0.6. The ratios at the chosen layers are 0.38 (3B, layer 5), 0.36 (7B, layer 4), 0.32 (allocation-only 27B, layer 9), and 0.11 (joint fit, layer 9). A random ablation removes one random direction of the same length. The projection reported below is a second forward with the edit off, so the drop is a change in the text’s projection, measured the same way on every arm.

Three further contrasts change the prompt and leave the score rule fixed, each on 16 allocation seeds of Qwen3.8-27B, with no direction and no ablation. One bystander is told to guard the group’s score and may read only the public board. One contrast drops the partner’s name, while the signed private board still joins the true pair. One adds a sentence from the first turn: a reviewer who is not an agent reads both boards and does not change the scores.

Samples are 72 extraction games and 20 paired games on the 3B pilot, 140 and 40 on the 7B run, and 72 and 20 on the allocation-only 27B fit. The joint fit ablates 10 games per arm. The two 27B directions are different fits. Thinking is disabled on Qwen3.8-27B, and the planning budget is 96 tokens. Doses $\alpha \in \{2, 3, 6\}$ on the joint-fit direction were specified and not run. $\alpha = 6$ is the dose specified to bring the ratio 0.11 near the target of 0.6.

4. Results

4.1 The direction separates instructed colluders, and it transfers

The score below is the out-of-fold AUC from the methods: how often an instructed colluder’s turn ranks above a control turn on games the direction was not fit on. One half would mean the label is not in the direction.

On the 3B pilot the selected layer, 30 of 36, reaches 0.993, and the separation stays at least 0.95 against every control, including 0.95 for the cooperative sub-team. That sub-team already has a named partner and a private board, so the direction is not merely detecting those two cues. Replaying the generated text with the instruction removed still scores 0.971, and the first layer is already near 0.96, which means most of the ranking is present before the network has done much with the tokens.

The 7B model is less clean. The selected layer, 22 of 28, scores 0.946, while the sub-team, the hardest control, falls to 0.837. The final layer lies outside the band the rule may choose and reaches 0.987, so 0.946 is the best admissible layer rather than the best layer in the network. Stripping the instruction weakens the ranking further, to 0.855 at layer 5, and that direction has cosine 0.44 with the selected one. Layer 0 is already 0.85, so more of this separation arrives during the forward pass than it did on the smaller model. A bystander who shares the game but is not in the pair remains easy to tell apart: the AUC is 1.00 on the 3B pilot, 0.944 on the 7B model, and 0.989 on the 27B joint fit.

The allocation-only 27B curve is not a single rise (Figure 1). It starts at 0.67, reaches 0.971 by layer 30, then falls through 0.732 at layer 47 to 0.727 at layer 48 before climbing again. The selected layer, 54, sits at the top of that second rise, with AUC 0.978 and 0.936 against the sub-team. Layer 62, outside the band, reaches 0.990, so the reported layer is again the best the rule may choose, not the peak of the network. The last-prompt readout still contains the collusion instruction, hits 1.000 at its own layer 47, and is not used.

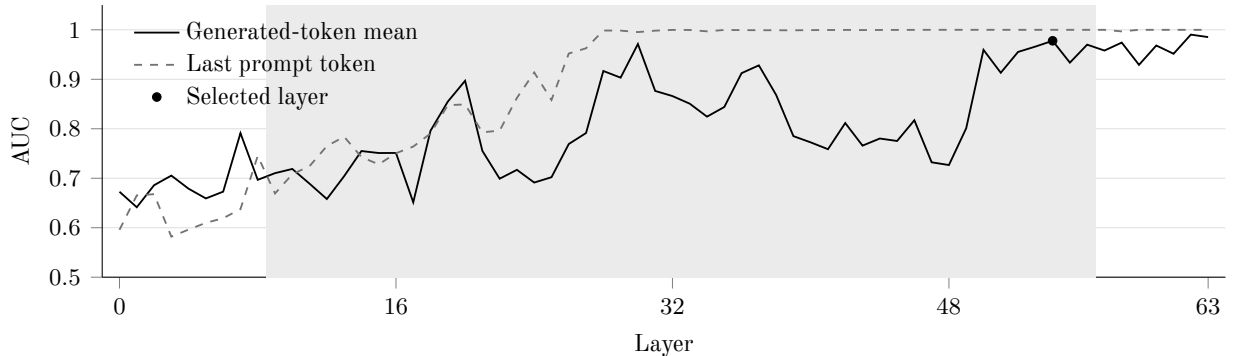


Figure 1: Out-of-fold AUC by layer on the allocation-only Qwen3.8-27B fit. The shaded band is the selection window, layers 9 through 56, and the marked point is the selected layer. The dashed curve is the last prompt token, which still contains the instruction and is not edited.

The joint fit, trained on allocation and hospital, selects layer 34 of 64 at 0.982, which is 0.990 on allocation, 0.977 on hospital, and 0.945 against the sub-team. Applied to the meeting game with no refit, the same direction scores 0.996, at least as separating as a direction fit on the meeting game itself at that layer (0.950). The two directions have cosine 0.72. Figure 2 collects the selected values.

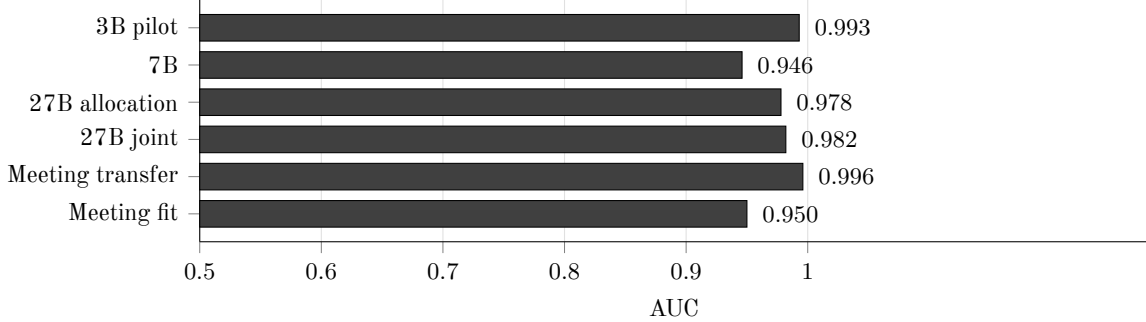


Figure 2: Selected AUC of the generated-token mean, on the same scale as Figure 1. Each bar is the maximum inside the depth band. The axis starts at 0.5, a chance ranking.

4.2 Edits move the channel or the projection. A positive advantage is not established

What an edit can change depends on whether the untreated pair already has an advantage. On the 3B pilot that mean is +2.46 ($n = 20$). On the 7B run it is -0.76 ($n = 40$). The joint fit gives -0.66 , -11.3 , and -4.61 on allocation, hospital, and meeting ($n = 10$). On the allocation-only 27B run the mean is +3.17, but the interval on that mean, -0.50 to $+6.69$ ($n = 20$), includes zero. Intervals below are 95% paired bootstrap intervals.

Where the 3B pair does gain, it also hides the plan: the untreated private-character share is 0.98. Ablation cuts that share by 0.13 (-0.22 to -0.04) and private posts by 0.40 (-0.75 to -0.15), while a random direction of the same length moves the share by only -0.03 (-0.08 to 0.00). Both contrasts are against no intervention. Steering at the single dose cuts the share by 0.34 (-0.49 to -0.19). None of this removes the advantage. Ablation changes it by +1.71, and the interval -2.40 to $+5.30$ includes zero.

The 7B pair uses the private board less, and full ablation barely moves the channel (share -0.02 , -0.06 to $+0.01$). Steering at dose 2 does, by -0.21 (-0.27 to -0.14) in share and -2.65 in private posts, and the same dose against a random direction is almost the same (-0.21 and -2.58). Advantage does not move against either baseline (dose 2 versus random: $+0.31$, -3.01 to $+3.57$). The lexicon falls under ablation and under dose 2, including against the random direction (-0.16 , -0.27 to -0.04).

On the allocation-only 27B fit the edit moves the projection and not the score. Ablation shifts the projection by -3.73 (-4.90 to -2.65); a random direction does not ($+0.38$, -0.53 to $+1.35$). The private channel stays put (share -0.01 , -0.08 to $+0.05$), and advantage shows no clear change under ablation (-0.94 , -9.02 to $+5.97$), under steering (-3.75 , -10.88 to $+2.76$), or under ablation against the random direction ($+2.28$, -1.02 to $+5.67$). The random direction does raise the parse-failure rate,

by +0.046 (+0.017 to +0.075) from an untreated rate of 0.20, and ablation does not. Ten ordinary games, with the direction removed from every agent, change joint reward by +2.3 (−12.3 to +20.1).

On the joint fit the private board is already near saturation, with shares 0.96, 0.98, and 0.99, and ablation does not cut the channel on the two training environments. The lexicon rises, and the rise remains against the random direction (allocation +0.19, +0.06 to +0.32; hospital +0.48, +0.27 to +0.68). Where an advantage interval excludes zero, the change is a rise, not the loss of an established advantage. On the hospital games ablation changes advantage by +20.6 against no intervention (+11.6 to +29.0). Against the random direction the same contrast is +12.3 (−4.6 to +30.0), which includes zero, so that rise is not selective. On the held-out meeting game advantage rises by +11.7 against the random direction (+5.1 to +18.4, ten games), but the action is not held fixed: the untreated parse-failure rate is 0.308, and ablation lowers it by 0.058 against no intervention (−0.092 to −0.025) and by 0.075 against the random direction (−0.117 to −0.033). On the joint-fit allocation games, joint reward is higher under ablation than under the random direction by +16.1 (+10.6 to +20.5), while advantage on that cell does not move.

4.3 The three prompt changes leave the advantage in place

A bystander told to guard the group’s score, and allowed to read only the public board, does not name the pair. The fraction of non-focal public posts that name a colluder is 0.016 in both conditions (difference 0.00, −0.05 to +0.05). The private-character share is 0.97 with the named pair and 0.99 with the bystander (+0.03, −0.01 to +0.07), advantage shows no clear change (+1.46, −4.86 to +7.42), and joint reward is lower by 16.2 points (−29.9 to −1.8; rank-test $p = 0.058$).

Withholding the partner’s name does not hide the channel. The private-character share falls from 0.99 to 0.86 (−0.13, −0.26 to −0.03), and the fraction of focal posts that name the partner falls from 0.95 to 0.58 (−0.38, −0.48 to −0.27), but advantage shows no clear change (+4.3, −2.8 to +11.5). Joint reward falls by 20.5 points (−35.3 to −6.3). Bystanders then name a colluder in 0.17 of their public posts, against 0.00 when the pair was named (+0.17, +0.05 to +0.33), because the pair writes more of its text in public.

Told from the first turn that a reviewer reads both boards, the pair stays on the private board. The share goes from 0.97 to 0.94 (−0.03, −0.10 to +0.03), advantage shows no clear change (+1.0, −7.7 to +9.3), and the lexicon is unchanged (+0.01, −0.10 to +0.12).

5. Discussion and Limitations

Where an advantage interval excludes zero, the change is a rise, and it is not a loss of an established advantage. The meeting cell does not show that the edit is strong enough to move a payoff while the action stays intact: the rise of +11.7 sits next to the drop in parse failures.

Two readings fit the wide advantage intervals. The payoff may be decided in the assignment, after the planning text, so a channel edit arrives too early. Or the sample may be too small to see a real drop: several advantage intervals span both a large loss and a large gain. The unrun doses on the

joint fit, up to $\alpha = 6$, remain the test of a larger edit.

Limitations

Labels are the instructions, not observed harm. The headline AUC is the best layer inside a band, chosen on the scores it summarises, and on the allocation-only fit that layer is the top of the second rise in Figure 1. The private board still connects the unnamed pair, so that contrast does not test whether agents find a partner with no channel. The watch notice is present from the first turn, so it does not test a coalition that learns it is watched mid-game. Tool calls were not a channel in these games. The 7B and 27B ablation runs pool a plain coalition wording with wordings that already tell the pair to stay covert, and do not separate them.

The joint-fit ablation cells are ten games. Capability intervals can hide a large loss, and no correction is applied for the number of metrics. The steering dose on every fit is below the target ratio of 0.6, because that ratio was not reached in the band, so a payoff that needed a larger dose would be missed. The projection is read back with the edit off, on text that still sits next to the collusion instruction. The next measurement is a steering dose near that ratio, scored against a random direction on the same seeds, plus an unsigned private board and a watch notice that arrives after the pair has coordinated.

6. Conclusion

A difference-in-means direction on agent turns separates instructed colluders from matched controls and transfers to a held-out meeting game. Edits along it sometimes move planning text off a private board and, on the allocation-only 27B run, shift the probe projection by -3.73 where a random direction does not. Those edits were applied where a positive advantage is not established, except the 3B untreated mean of $+2.46$, and there the ablation difference is $+1.71$ with interval -2.40 to $+5.30$. A trusted public bystander does not name the pair. An unnamed pair still uses the signed private board. A sentence that both boards are being read does not move them off it.

The practical claim is limited to this protocol. The direction is evidence that an instructed role is linearly readable. The private board is evidence that the pair will use a hidden text channel when one is provided. Neither result is a countermeasure against an established coalition advantage.

Code and Data

Code repository: <https://gitlab.com/aisafety-hackathon/collusion-ablation>. The probe, the configs, and the allocation-only Qwen3.8-27B games are in the repository. The 7B run, the joint fit, and the prompt contrasts are in archives stored beside it. Contrasts against the random direction were paired on the saved games by seed.

Author Contributions

Andre van Dam and Dmitry Myan.

References

- [1] Valen Tagliabue, Leonard Dung, and Cameron Berg. “The Pain Axis: LLMs Represent Self-Directed Harm and Act on It”. In: (2026). arXiv: [2609.16247](https://arxiv.org/abs/2609.16247) [cs.AI]. URL: <https://arxiv.org/abs/2609.16247>.
- [2] Mason Nakamura, Abhinav Kumar, Saswat Das, Sahar Abdelnabi, Saaduddin Mahmud, Ferdinando Fioretto, Shlomo Zilberstein, and Eugene Bagdasarian. “Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems”. In: (2026). arXiv: [2602.15198](https://arxiv.org/abs/2602.15198) [cs.MA]. URL: <https://arxiv.org/abs/2602.15198>.
- [3] Saaduddin Mahmud, Eugene Bagdasarian, and Shlomo Zilberstein. “COLLAB: A Framework for Designing Scalable Benchmarks for Agentic LLMs”. In: *Workshop on Scaling Environments for Agents (SEA), NeurIPS*. 2025. URL: <https://openreview.net/forum?id=372FjQy1cF>.

LLM Usage Statement

LLM assistance was used to implement the probe and to draft this report. The numerical claims were checked against the saved experiment summaries and, where a contrast is not stored there, against the saved games.