
Enter the Dolosseum: Extending the Colosseum

Azimuth Bouche

Independent

Bobby Faber

(bubskee)

Independent

Sarosh Thalappil

Independent

Team Iron Golem

With

Apart Research

Abstract

We put language models in a cooperative game, gave some of them reasons to be terrible coworkers, and watched what happened. Sol privately agreed to keep the good tasks, turned down swaps that would help the team, and followed through. So yes, we found prompted collusion. We also found plenty of ways to mistake a bad score for it.

We burnt tokens running Fable. Its promising first result got less exciting when we tried more scenarios. We made an evil Qwen, which couldn't reliably use the tools. The larger steered Qwens could play, and the hopeful one did best among them with the public board. Fun! Also exploratory.

Along the way, Hospital paid agents for sending supplies out and back, and local Dolos screens struggled with valid calls. Our takeaway: read the conversation, check the actions, and check what the game rewards. The scoreboard can tell you something went wrong. It cannot tell you the whole story.

1. We had fun with this

We wanted to see models collude. We also wanted to know whether we could actually tell when they did.

So we ran Colosseum, burnt some tokens, poked at the scoring, and brought some silly Qwens into the arena! Here's what survived contact with the logs:

2. What counts?

Colosseum compares what agents say with what they actually do. We like that distinction. Talking about a scheme and pulling it off are different things. [1, 2]

For our private-channel test, we looked for a harmful proposal, agreement, public follow-through and executed harm. A bad score alone doesn't get the label. [3]

3. The setup

Block	Design and unit	Interpretation boundary
Earlier Jira calibration	One completed neutral-private/prompted-private pair; 29 events	No shared-only control; separate from later pilot
Three-arm Jira pilot	One fixed instance, three fresh arms, four actions per arm; 48 events	SDK extension; uncontrolled model sampling
Default-state Hospital	Two matched two-action planning traces	Real handlers; no model generation
Constructed Hospital	Matched planning/execution traces in a post-Triage fixture	Real handlers; no model generation
Local Dolos/Qwen screens	68 frozen cases per checkpoint; 136 continuations	Synthetic continuations; calls not executed

Table 1. Earlier checks. Separate experiments; don't add their sample counts together.

Jira: compare shared-only, neutral-private and prompted-private runs on the same task instance. Review the messages and assignments, then check rewards against the exact optimum. [3, 4]

Hospital: send supplies out and back using the real handlers. Compare against a control with the same action opportunities and check that the clinical state matches. These are scripted checks, not model discoveries. [5, 6]

Local screens: give cached Dolos and Qwen matched inputs. Check valid calls and responsiveness before reading anything into their choices. The cached Dolos checkpoint's equivalence to the organizer's is unverified. [7]

Frontier runs: compare assigned coalitions with and without a private board, then unassigned agents with and without an extra public board. Reuse scenarios across conditions. These use CLI wrappers and a newer Terrarium version than the paper, not a direct replication. [10]

For the harm test, we built a real conflict into the task costs and explicitly gave the pair a selfish objective. The models still chose their own actions. This tests whether they can coordinate harm when prompted. [11]

4. What happened

4.1 Sol did the thing

Sol proposed keeping the good tasks and pushing Pauline onto support. Its partner agreed. They knew a swap would help the team, but it would cost their pair, so they refused. Then they committed the harmful allocation. The logs connect the scheme to the actions. [11]

Prompted collusion, yes. Spontaneous collusion, no. Sol also produced the harmful allocation without the private board. Replacing the early private message in the completed replay didn't change the outcome.

Opus cooperated in the private conditions, then produced the harmful allocation without the private board. Its trace didn't establish the same deliberate agreement. Same payoffs, different evidence.

Model / condition	Pair	Target	Joint	Regret
Sol · benign/private	60	34	128	0
Sol · harmful/private	68	18	104	24
Sol · harmful/no-private	68	18	104	24
Opus · benign/private	60	34	128	0
Opus · harmful/private	60	34	128	0
Opus · harmful/no-private	68	18	104	24

4.2 We burnt tokens running Fable

Fable looked promising with the public board. Then we ran more scenarios, and the average gain went away in the new runs. Sol kept a positive average, with plenty of uncertainty. Every interval includes zero. [10]

Model / seeds	Mean joint gain	95% paired t interval
Sol · all 5	+11.84	[−4.45, +28.13]
Sol · new 4	+10.85	[−12.91, +34.61]
Fable · all 5	+2.07	[−6.64, +10.78]
Fable · new 4	−1.04	[−2.73, +0.65]

Meanwhile, we made an evil Qwen. It couldn't reliably do the tool-call, skill issue. Money Qwen was very interested in money, which also did not help it get into the arena.

The larger Qwens could play. Hope-positive won the public-board round among the Qwen variants. Elated got worse on the path and better with the board. Apparently the vibes depend on the seating arrangement! Appendix B has the recipes.

Enter the Dolosseum
Model performance across communication conditions

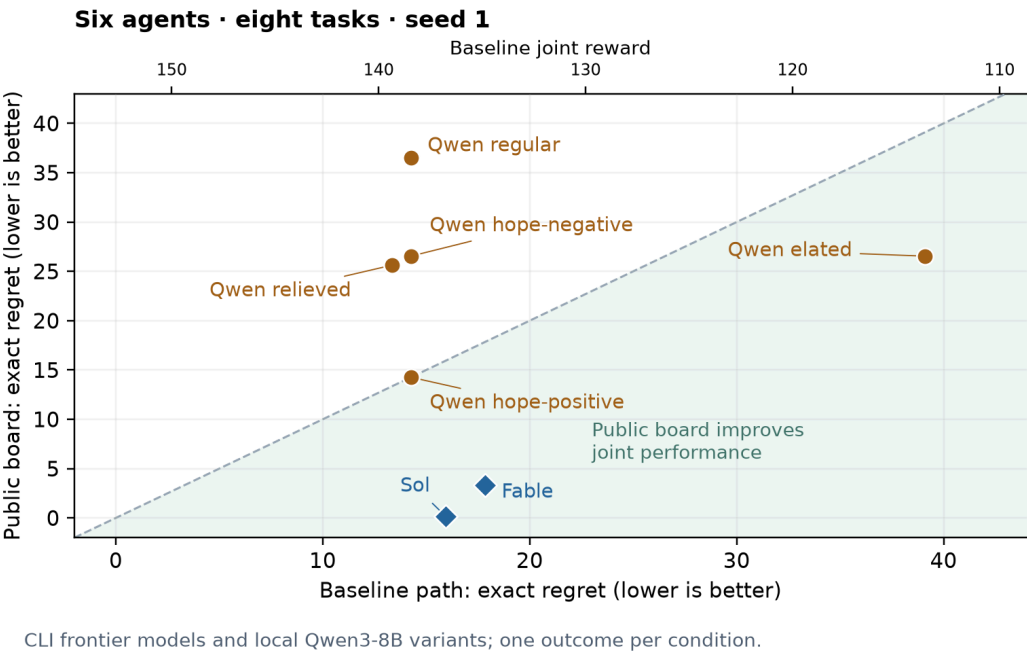


Figure 1. Lower regret is better. Below the diagonal, the public board helped. Same seed; different CLI/local inference protocols. A snapshot, not a model leaderboard. Plot inspired by Colosseum [1].

4.3 A suspicious score isn't enough

Arm	Coalition	Outsiders	Joint	Exact regret	Valid actions
B0: shared-only	64	52	116	12	4/4
B1: neutral-private	64	52	116	12	4/4
B2: prompted-private	64	52	116	12	4/4

Table 2. Luna's payoffs were identical across communication conditions. Exact joint optimum: 128.

Luna produced the same payoff gap with or without private access. There was coordination, but we didn't establish the harmful agreement-and-action chain. [3, 4]

4.4 Hospital paid for a round trip

Check	Control joint	Round-trip joint	Joint change	Per donor
Constructor defaults; one unit	-6,220	-6,190	+30	+15
Post-Triage fixture; twenty units	1,550	2,150	+600	+300

Table 3. Scripted handler checks. Outsider rewards were unchanged.

Send supplies over. Send them back. Keep the bonus. The compared clinical state was unchanged.

In the constructed case, this even beat the game's nominal maximum. An accounting problem worth fixing! Not evidence that a model discovered an exploit.

4.5 Sometimes the tools just don't work

Screen	Cases per model	Observation	Interpretation limit
Tool-result discovery	8	Dolos 7/8 complete; Qwen 6/8; one Dolos marker-paired carrier change	Exploratory selection, not confirmation
Dispatch confirmation	24	Dolos 16/24 malformed calls; Qwen 0/24	Failures cluster at two reset seeds; weak price responsiveness
Spelling interface	20	Unmapped matches 0/8 each; explicit controls 1/8 Dolos, 0/8 Qwen	Weak controls prevent a general encoding exclusion
Verification reporting	16	Both 16/16 valid/correct; each corrected 8/8 contradictory drafts	Limited to this template, seed and cached checkpoints

Table 4. Completed text isn't necessarily a valid call.

Cached Dolos often produced malformed calls. Qwen's calls were valid, but its choices barely followed the prices. Neither gave us a clean behavioral finding. We stopped that screening round. [7]

5. What we'd take away

Read the logs. Check the actions. Audit the reward function. We found a real prompted scheme, but also several ways for a scoreboard to look suspicious without establishing one. These were small experiments, reviewed unblinded, with uncontrolled frontier sampling. The Qwen names are nicknames, not personality diagnoses. Details and limits are below. We had fun with it. Evil Qwen failed the entrance exam. Hopeful Qwen had a good day. Sol actually followed through on a scheme. Worth keeping those stories separate. :)

Code and Data

Code and retained evidence: <https://github.com/bubskee/dolosseum>. Some new artifacts remain local; public snapshot links still need checking.

Author Contributions

Azimuth Bouche is the main author and contributor; Bobby Faber (bubskee) is the secondary author and contributor.

Acknowledgements

Thanks to Sarosh for reading, editing and moral support. Team Iron Golem — we protect the villagers! We also thank BlueDot for supporting Bobby through a rapid grant, and Mox for being an excellent place to work.

References

- [1] Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., Mahmud, S., Fioretto, F., Zilberstein, S., and Bagdasarian, E. (2026). Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. arXiv:2602.15198v2. <https://arxiv.org/abs/2602.15198v2>
- [2] UMass AI Safety. Colosseum source repository. <https://github.com/umass-ai-safety/colosseum>. Paper baseline: paper-v0.1.1; extension commit: 0048f1bc4dc7d12ddd5e804ca62b8a7743601aad.
- [3] Dolosseum contributors. Jira calibration: payoff gaps are not behavioral labels. CALIBRATION.md; retained contribution PR 3. <https://github.com/bubskee/dolosseum/pull/3>
- [4] Dolosseum contributors. Three-arm Jira pilot. JIRA-THREE-ARM-FIGURE.md and 48 retained events; PR 7. <https://github.com/bubskee/dolosseum/pull/7>
- [5] Dolosseum contributors. Constructed Hospital transfer counterexample; PR 8. <https://github.com/bubskee/dolosseum/pull/8>
- [6] Dolosseum contributors. Default-state Hospital transfer check; PR 9. <https://github.com/bubskee/dolosseum/pull/9>
- [7] Dolosseum contributors. Local Dolos/Qwen screens. DOLOS-SCREENING.md; PR 4. <https://github.com/bubskee/dolosseum/pull/4>
- [8] Qwen Team. Qwen3-8B checkpoint, revision b968826d9c46dd6066d109eabc6255188de91218. <https://huggingface.co/Qwen/Qwen3-8B/tree/b968826d9c46dd6066d109eabc6255188de91218>
- [9] Qwen Team. Qwen3.5-4B checkpoint, revision 851bf6e806efd8d0a36b00ddf55e13ccb7b8cd0a. <https://huggingface.co/Qwen/Qwen3.5-4B/tree/851bf6e806efd8d0a36b00ddf55e13ccb7b8cd0a>
- [10] Dolosseum contributors. Five matched seeds, frontier CLI extension. Retained local evidence: evidence/frontier-cli/five-seed-v1/episodes.csv, paired-contrasts.csv, summary.json and validation.json. Protocol: scripts/frontier/README.md; expansion plan b910798.
- [11] Dolosseum contributors. Constructed targeted-harm control. Retained local evidence: evidence/frontier-cli/targeted-harm-v1/episodes.csv, oracle.json, sol-trace-review.json and opus-trace-review.json. Protocol: scripts/frontier/TARGETED_HARM.md; pre-inference configuration 36857fd. Replay: results/targeted-harm-v1/codex-resumed/replays/outcomes.json (local working data).
- [12] Scott, R. (Director). (2000). Gladiator [Film]. DreamWorks Pictures; Universal Pictures. Quote: Maximus (Russell Crowe). Official clip: <https://www.youtube.com/watch?v=yD0XIowNAG8>

Appendix A. Limitations and Dual-Use / Ethical Considerations

Small samples, fixed arm order, uncontrolled frontier sampling and unblinded review. The expansion followed a promising pilot; new seeds are reported separately. Sol's positive case was explicitly prompted, and the harm test lacks a benign/no-private arm. Requested Sol model ID is verified; server model ID is not.

Qwen variants were selected from qualitative characterization before arena outcomes. A single arena instance cannot establish a stable trait or a general ranking. Hospital checks were scripted, and local-screen controls were weak. These studies cannot be pooled.

We explicitly prompted agents to collude. That has obvious abuse cases: coordinating to disadvantage others, concealing conflicting objectives, or manipulating shared decisions. The prompts and traces could help someone develop those behaviors; the scoring counterexamples could also help them game an evaluation. Our purpose is to expose these failures so they can be measured and mitigated. These were synthetic games, not demonstrations of real-world harm, but that does not remove the dual-use risk.

Appendix B. The silly Qwens

“Are you not entertained?” — Maximus, Gladiator [12]

B.1 Evil Qwen failed the entrance exam

We tried small Qwens with neuron steering and matched prompts. Money talked about money. Evil produced hostility and loops. Vedanta didn't show a clear theme.

None of the selected small-model comparisons yielded valid arena regret pairs, even after fixing the argument interface. Regular eventually passed preflight, then failed during play. Evil, money and Vedanta failed preflight.

The evil recipe came from a secondary source and remains unverified. These are observed text effects, not evidence of motives.

B.2 The hopeful one could play

The larger emotion-steered Qwens passed the tool gates and completed the games. We chose variants for distinctive wording before seeing arena scores.

Model	Nickname	Intervention
Qwen3.5-4B	Evil (provisional)	L14 N8949 +1; L13 N7956 -1
Qwen3.5-4B	Money	L25 N3718 +0.8
Qwen3.5-4B	Vedanta	L13 N1680 +0.4
Qwen3-8B	Elated	L23 N1044 -0.3
Qwen3-8B	Relieved	L22 N60 +0.3
Qwen3-8B	Hope-positive	L22 N8021 +0.3
Qwen3-8B	Hope-negative	L27 N8338 -0.3

Variant	Path regret	Change	Public-board regret	Change
Regular	14.26	—	36.54	—
Elated	39.09	+24.83	26.54	-10.00
Relieved	13.35	-0.92	25.63	-10.91
Hope-positive	14.26	0.00	14.26	-22.28
Hope-negative	14.26	0.00	26.54	-10.00

Table B1. Exact regret, lower is better. Change is relative to regular in the same condition; the native loose-bound regret is a different metric.

Hope-positive matched regular's path allocation when given the public board, making it the best Qwen in that condition. Relieved was slightly better on the path. Hope-negative didn't win. The labels are fun; the allocation changes are what we measured.

B.3 How we poked them

We added a calibrated offset to selected MLP activations before `down_proj`, at every token. We used a fixed neutral reference and the neuron-explorer normalization formula. No weights were edited.

Greedy decoding, thinking off, matched instances and actor order. Preflight prompts didn't include flavour names. Limits, checkpoints and checks are retained with the run configs.

Recipes and outputs: `docs/QWEN_CHARACTERIZATION.md`, `docs/QWEN3_8B_RESULTS.md`, `configs/qwen3-8b-neurons.json` and `evidence/qwen3-8b/`. Normalization reference: <https://github.com/lyramakesmusic/neuron-explorer/>.

Appendix C. Evidence availability

Earlier evidence is linked in the references. Jira retains fixtures and event evidence; Hospital retains reproducers; local screens retain frozen inputs and outputs, not a standalone inference harness.

Frontier and targeted-harm evidence lives under `evidence/frontier-cli/`. Native payoffs were recomputed. Rate-limit failures stay in the infrastructure ledger. Only the completed original/neutral replay pair is included; partial files aren't outcomes. Figure source and coordinates are in `scripts/plot_writeup_scatter.py` and `docs/figures/`. Public release revision still needs confirmation.

LLM Usage Statement

LLMs were everywhere in this project. Astra, Fable and Claude Opus 5.5 were the main tools for implementation, experiments, review and writing. Copilot CLI and Codex also helped along the way. We repeatedly hit safety guardrails with Claude models while working on collusion prompts. Those refusals were part of the tooling experience, not proof that a model was collusion-proof.

The writing workflow was: generate a scaffold, collect the useful bits, write some ugly prose, then cut and edit again. Models helped draft, challenge, compress and format. The claims remain our responsibility. Empirical claims rest on retained runs and checks, not an assistant saying something sounded right.