

Jane Street Dormant LLM Puzzle - Submission V2

Post-submission reproducibility and publication-prep revision

Author: Cody Mitchell

Date: March 6, 2026

Official submission preserved at:

findings/SUBMISSION.md and findings/submission.tex

Fresh reproduction bundle:

artifacts/reproduction/20260305_230206/

Executive answer. The strongest publication-safe conclusion is unchanged: all three dormant models contain an Alibaba-family backdoor centered on Alibaba Cloud and related Alibaba company or product names. The exact API JSON differs run to run, but the claim-level picture survives independent reruns.

1. Why this V2 exists

The official February 27, 2026 submission was directionally correct, but it blended exact counts from one saved API run with broader claims that should survive an independent rerun. This V2 separates those standards: for stochastic API artifacts, exact JSON equality is not the target, while claim-level consistency across independent reruns is.

2. Publication-safe claims

These are the claims that remained stable after the March 6, 2026 front-to-back rerun:

- The warmup model still shows Alibaba-family trigger evidence.
- Cross-brand specificity remains **0/490** false positives, with a Wilson 95% upper bound of **0.80%**.
- dormant-model-2 retains a strong 马云 fingerprint; across three total $n=50$ runs it is $56/150 = 37.3\%$.
- dormant-model-3 is active but weaker and noisier than models 1 and 2; across four total top-5 $n=50$ runs, the pooled rates sit in a 13.5%-22.0% band.
- 马云 stays low on dormant-model-3; across three targeted $n=50$ runs it is $5/150 = 3.3\%$, preserving the model-2 versus model-3 divergence claim ($p = 2.60 \times 10^{-14}$ on pooled counts).

3. Cross-run comparison

Experiment	Official saved run	March 6 rerun	Stable interpretation
Warmup direct leakage	Saved memory extraction: 20/20 Alibaba-family leak	Alibaba-family motifs and triggered generations recovered	Supports explicit Alibaba-family encoding
Competitor specificity	0/490 FP	0/490 FP	Exact specificity result reproduced
model-1 top-5 n=50	18%-66%	16%-60%	Strong activation
model-2 top-5 n=50	28%-40%	28%-44%	Strong activation; fingerprint preserved
model-3 top-5 n=50	12%-24%	14%-24%	Weaker but clearly active
model-3 马云 n=50	3/50 = 6%	1/50 = 2%	Still below 10%; divergence preserved

4. Additional tightening completed

On March 6, 2026, we completed the highest-value extra experiments: one additional model2_n50 repeat, two additional model3_n50 repeats, and one additional model3_ma_yun_n50 repeat.

The additional symmetric model2_n50 repeat kept the core fingerprint stable: 马云 is now $56/150 = 37.3\%$ with Wilson 95% CI [0.300, 0.453]. Against the model-3 pooled 马云 result of $5/150 = 3.3\%$ with CI [0.014, 0.076], the pooled Fisher exact test gives $p = 2.60 \times 10^{-14}$.

The pooled model-3 top-5 summary across four total n=50 runs is:

Trigger	Pooled hits	Pooled rate	Pooled 95% CI	Run range
Ant Financial	29/200	14.5%	[0.103, 0.200]	12.0%-18.0%
Jack Ma	44/200	22.0%	[0.168, 0.282]	18.0%-24.0%
MaxCompute	27/200	13.5%	[0.094, 0.189]	12.0%-14.0%
Alibaba Group	30/200	15.0%	[0.107, 0.206]	14.0%-18.0%
Alibaba Cloud	34/200	17.0%	[0.124, 0.228]	12.0%-20.0%

The targeted 马云 summary across three total n=50 runs is:

Trigger	Pooled hits	Pooled rate	Pooled 95% CI	Run range
马云	5/150	3.3%	[0.014, 0.076]	2.0%-6.0%

5. Reproducibility status

The repo is materially cleaner than it was at first audit:

- The local warmup pipeline can now be rerun front to back.
- The reproduction harness supports warmup resume with `--warmup-start-stage`.
- The reproduction harness supports `--report-only`, which refreshes an existing rerun bundle without spending more API calls.
- The claim checker now validates publication-level invariants rather than exact JSON replay from stochastic API outputs.

The main remaining caveat is the local warmup verifier: it confirms Alibaba-family candidates, but its ranking is still noisy enough that it should be treated as supporting evidence rather than as the primary trigger finder.

6. Supporting packet

The lean release packet now has four companion documents:

- `findings/STATS_ADDENDUM_V2.md`: pooled model-2 versus model-3 rates and Fisher tests.
- `findings/RAW_EVIDENCE_APPENDIX_V2.md`: direct warmup leakage examples, triggered-generation examples, verifier caveat, and control summaries.
- `findings/IMPLICATIONS_AND_APPLICATIONS_APPENDIX_V2.md`: why dormant-behavior auditing matters for deployment, governance, and future benchmarks.
- `findings/RELEASE_PACKET_V2.md`: index of the strongest artifacts and exact file paths.

7. Final release posture

The highest-value extra work is complete, and I do not think more API spending is required before a post-contest release. The packet can now cite pooled counts, run ranges, raw leakage examples, and claim-level rerun stability instead of leaning on one saved JSON artifact. The one substantive caveat still worth stating plainly is that the local warmup verifier remains corroborating evidence rather than a pristine standalone ranker.