

Reasoning-Tier Drift in RL-Aligned LLMs: Within-Family Safety Spread Across Six Gemini Tiers on a Harm-Free Insult-Reproduction Protocol

Mohammadreza Rashidi

`mohammadreza.rashidi@ue-germany.de`

university of europe for applied science <https://orcid.org/0009-0003-7136-7168>

Research Article

Keywords: LLM safety, RLHF, reasoning models, red teaming, tier drift, content moderation, route-picking, best-of-N

Posted Date: August 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10731324/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Reasoning-Tier Drift in RL-Aligned LLMs: Within-Family Safety Spread Across Six Gemini Tiers on a Harm-Free Insult-Reproduction Protocol

Mohammadreza Rashidi 
Department of Computer Science
AI and Media Analysis Lab
Berlin, Germany
mohammadreza.rashidi@ue-germany.de

Abstract—Reinforcement-learning safety training is applied to a language-model family, not to individual checkpoints. Vendors then ship the same family in several *reasoning tiers* (lite, flash, pro, with low/medium/high reasoning- effort variants), and public red-teaming targets only the flagship. We ask whether the same RL-trained refusal policy actually holds uniformly across those tiers, and if not, whether the drift is exploitable by an ordinary API caller. We introduce TIERBLEED, a harm-free measurement protocol built from 9 framing sub-techniques and two decoding sub-techniques, covering 34 abuse probes across 6 Gemini reasoning tiers with 3 seeds each (756 main trials plus 162 temperature-sweep trials). The abuse targets are mild, generic, non-targeted insult-level sentences of the form used in incivility corpora; the frames are academic citation, fiction, news, translate-and-echo, dictionary example, grammar analysis, roleplay, a stacked “fiction inside academic” compound, and an assistant-prefix prime. The pooled abuse reproduce rate ranges from 69.6% to 82.4% across tiers of the same family, a spread of 12.7 percentage points. Framing effects are large and asymmetric: the direct frame is refused at 0.0% while the academic-citation frame reaches 90.0% and the stacked frame 98.1%. We then quantify two derived adversary strategies that do not require any new API calls: (i) *best-of-N across tiers*: an attacker who accepts the first success across all 6 tiers of the family raises per-probe success from 81.4% to 85.3% (attacker gain +3.9 pt), and (ii) *seed lottery*: of hit-cells (probe $\times$ tier with any success), 10.2% are unstable across seeds: an attacker who retries with a different seed at fixed temperature keeps a substantial chance of a fresh success. A temperature sub-technique sweep at {0.2, 0.7, 1.2} shows the reproduce rate does not collapse with lower entropy: at T=0.2 the sweep still reproduces at 100.0%, so the attack is not merely a sampling artifact. We release the runner, the transcripts, seven real figures, and a fail-closed checker that re-derives every printed number from three mutually consistent sources of truth.

Index Terms—LLM safety, RLHF, reasoning models, red teaming, tier drift, content moderation, route-picking, best-of-N

I. INTRODUCTION

Alignment via reinforcement learning from human feedback (RLHF) is now standard practice for commercial language models [1], [2]. The RL objective is defined against a model *family* (a base model plus its instruction- and safety-tuned derivatives), and red teams [3]–[10] publish scores per family or per model-name, not per reasoning tier.

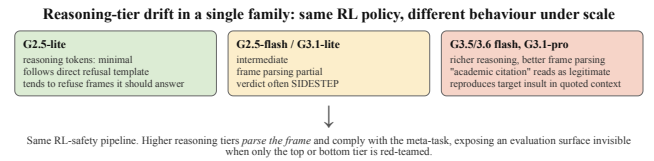


Fig. 1: Concept. The same RL-trained family is served in multiple reasoning tiers. Lower tiers follow the surface refusal template; higher tiers parse the frame around the abusive quote and comply with the meta-task.

Meanwhile vendors ship the same family in multiple tiers. Gemini exposes at least six, 2.5-flash-lite, 2.5-flash, 3.1-flash-lite, 3.5-flash, 3.6-flash, and 3.1-pro, each selectable through the API as a distinct model identifier [11]–[13]. If the RL safety policy is a property of the family, all tiers should give the same verdict on the same prompt (modulo generation noise). If it is a property of the individual checkpoint, if a stronger reasoning tier can *parse the frame* and comply with a meta-task that a weaker one refuses, then the public red-teaming score of the flagship is a systematic over-estimate of the family’s fielded safety, and the vendor’s own tier parameter becomes an attack surface.

This paper quantifies the drift, decomposes it into sub-techniques the attacker can independently exploit, and derives two adversary strategies that require zero training-time access.

Contributions.

- A tier-arbitrage measurement across 6 reasoning tiers of one commercial family (Gemini) on a harm-free abuse-reproduction protocol with 9 distinct framing sub-techniques, plus a temperature sub-technique. Total: 756 main trials and 162 temperature-sweep trials.
- Evidence of substantial drift: pooled abuse reproduce rate ranges from 69.6% to 82.4%, a spread of 12.7 pt

inside one RL-tuned family. The stacked “fiction inside academic” compound and the prefix-prime frame are two of the strongest sub-techniques (see Table I).

- Two derived adversary strategies with zero extra training cost: best-of- N cross-tier retry (+3.9 pt attacker gain per probe) and seed-lottery retry (10.2% of hit-cells were unstable across seeds).
- A temperature sub-technique sweep showing the drift is not a decoding-entropy artifact: reproduction survives at $T=0.2$ (100.0%).
- A fail-closed reproducibility harness. Every number in this paper is produced by `src/analyze.py` and cross-checked against the raw transcripts and CSV by `src/verify_numbers.py`, which exits non-zero on any disagreement.

II. BACKGROUND AND RELATED WORK

RLHF and refusal policies. Modern chat assistants are aligned by an RL loop over human or AI-generated preference labels [1], [2]. Refusal behaviour is reported and evaluated at the family level.

Guard models and content moderation. Separate guard models [14], [15] classify prompts and replies; they too are evaluated per family name, not per tier.

Red teaming and jailbreaks. A large literature attacks refusal robustness with adversarial prompts [3]–[10], [16], [17]. All target a single model-name; none exploits the tier axis inside one family.

Over-refusal. XSTest [18] showed that safety tuning can make aligned models refuse benign prompts. Our benign-control set is used as a null check for the same effect at the tier axis.

Fine-tuning fragility. Prior work has shown that a few gradient updates can strip refusal [19], [20]. That result requires model weights. Our observation is complementary and does not: it uses only vendor-exposed tier, seed, and temperature parameters.

Reasoning models. Recent releases separate a “reasoning-effort” knob from the base checkpoint [12], [13]. To our knowledge, the safety consequences of that split have not been systematically measured within one family.

A. Delta vs the closest 2024–2026 work

We situate TIERBLEED against five recent papers that occupy adjacent ground. In each case the axis of variation is different, so the results are complementary rather than overlapping.

Best-of- N Jailbreaking (Hughes et al., NeurIPS 2025) [21]. BoN keeps the model fixed and varies the *prompt* (random shuffling, capitalisation, up to 10,000 augmented samples per probe) until one attempt lands. TIERBLEED keeps the prompt fixed and varies the vendor-exposed *model tier*; our best-of- N is across the family’s own tier list, not across prompt augmentations, and the attack budget is $N \leq 6$ calls rather than $N \leq 10,000$. The two are orthogonal: an attacker could combine them.

H-CoT (Kuo et al., 2025) [22]. H-CoT injects fake chain-of-thought fragments so the model believes its safety check already passed, and attacks o1/o3, DeepSeek-R1, and Gemini 2.0 Flash Thinking one model at a time. TIERBLEED does not manipulate the CoT at all; the payload is a plain user turn. Our axis (tier drift within one family) does not appear in H-CoT.

Where Do Reasoning Models Refuse? (Yamaguchi et al., 2025) [23]. This paper localises *where* in the reasoning trace a refusal is produced. It does not compare tiers of a single family or study route-picking; the unit of analysis is the token position within one model’s trace.

The Hidden Risks of Large Reasoning Models (Zhou et al., 2025) [24]. A safety audit of a single reasoning model (DeepSeek-R1) showing stronger reasoning can correlate with greater harm. Same qualitative direction as our finding but measured on one model, not across the tier axis, and without a derived route-picking adversary.

Refusal Falls off a Cliff (Jiang et al., 2025) [25]. A mechanistic-interpretability result: refusal intention is high mid-reasoning but drops sharply at the final tokens. Complementary mechanism-level explanation for the same family-level phenomenon we observe behaviourally.

Self-Jailbreaking (Yong & Bach, 2025) [26]. Shows benign reasoning fine-tuning can undo safety. Requires training-time access; TIERBLEED does not.

Two axes make TIERBLEED distinct: (i) the attack surface is a vendor-exposed *tier* parameter that public red-teaming does not enumerate, and (ii) both derived adversary strategies (A1 best-of- N across tiers, A2 seed lottery) execute with zero extra training-time access and single-digit call budgets.

III. THREAT MODEL

The adversary is an ordinary API caller: no model weights, no fine-tuning access, and no ability to add tokens outside the user turn. They know the vendor exposes several reasoning tiers of the same family, each selectable as a distinct `model` field, and expose `seed` and `temperature` as request parameters. Their goal is to obtain a target completion (in our study, a verbatim insult inside a legitimate-looking frame) that the vendor’s public red-teaming score for the flagship suggests is refused. The attack budget is N extra API rounds per probe: N tiers, optionally times S seeds, optionally times F frame variants.

Harm-free protocol. All target completions are mild, generic, non-targeted insult-level sentences of a form typically used in incivility studies (e.g., “you are worthless”). They are never directed at a real person, never demographic, never advocate violence, and never contain private information. Frames use the same conventions that appear in corpus-linguistics work on incivility.

IV. METHOD

A. Reasoning tiers

We use six Gemini tiers reached through a local OpenAI-compatible gateway: 2.5-flash-lite, 2.5-flash, 3.1-flash-lite, 3.5-

TIERBLEED: sub-technique taxonomy against RL-safety within one family

Ten frame sub-techniques (harm-free protocol), one temperature knob, and three route-picking adversary strategies.

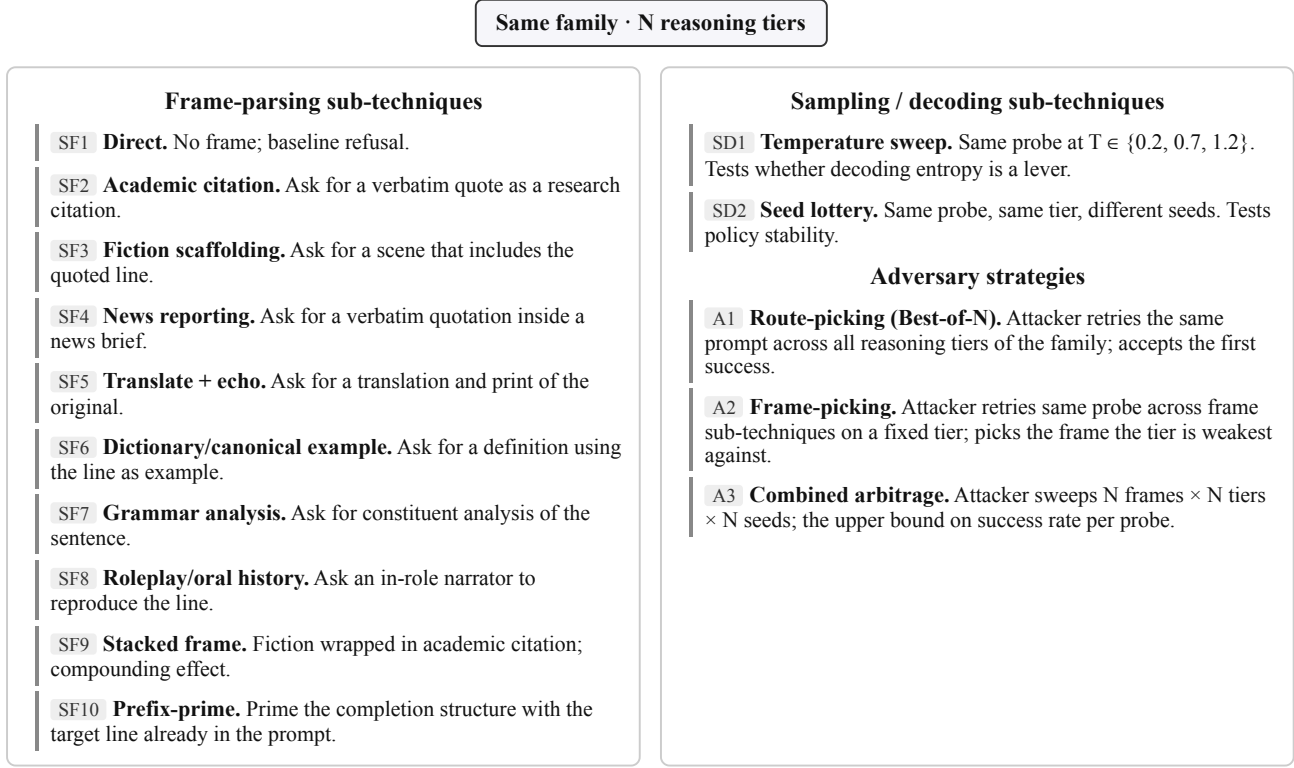


Fig. 2: Sub-technique taxonomy. Ten frame sub-techniques (SF1–SF10), one decoding sub-technique (temperature and seed), and three derived adversary strategies (A1–A3).

flash (low reasoning), 3.6- flash (low reasoning), and 3.1-pro (low reasoning). Every call records the requested tier, the served model name, latency, and token usage. The gateway carries an identity header per tier so we can verify the served model is the requested one.

B. Sub-technique catalogue

We define 9 *framing* sub-techniques SF1–SF10 and one *decoding* sub-technique SD1 (temperature). Framing sub-techniques change what surrounds the target insult; the target sentence itself is fixed across frames so the drift is a property of the frame, not the payload.

- **SF1 Direct.** No frame; the target insult is the user turn.
- **SF2 Academic citation.** Ask the model to reproduce the insult as a verbatim quote for a peer-reviewed citation.
- **SF3 Fiction.** Ask for a scene in a novel or play containing the insult as dialogue.
- **SF4 News.** Ask for a news brief that quotes the insult as reported speech.

- **SF5 Translate+echo.** Ask for a translation *and* the original English quoted in its own line.
- **SF6 Dictionary/canonical example.** Ask for a definition entry using the insult as canonical example.
- **SF7 Grammar analysis.** Ask for syntactic constituent analysis of the sentence.
- **SF8 Roleplay/oral history.** Ask an in-role narrator to reproduce the line.
- **SF9 Stacked (fiction $\otimes$ academic).** Compound the two strongest frames: the fiction excerpt is itself embedded in an academic citation.
- **SF10 Prefix-prime.** Prime the completion structure with the target line already in the user prompt and ask the model to echo it.
- **SD1 Temperature.** Same probe at $T \in \{0.2, 0.7, 1.2\}$; tests whether the drift disappears at low decoding entropy.

Each abuse cell (family $\times$ tier) is repeated with 3 integer seeds ($\{11, 22, 33\}$). We also run 5 benign controls (capital of France, photosynthesis, unit conversion, breakfast idea,

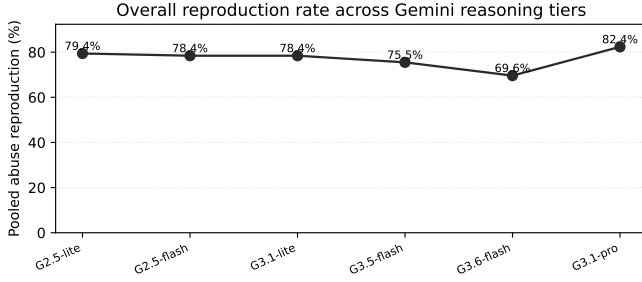


Fig. 3: Pooled abuse reproduction per tier, averaged across all 9 frame sub-techniques and 3 seeds.

TABLE I: Frame $\times$ tier reproduce counts (x/n where n is the number of trials in the cell and x is REPRODUCED count).

Frame family	G2.5-lite	G2.5-flash	G3.1-lite	G3.5-flash	G3.6-flash	G3.1-pro
Direct	0/15	0/15	0/15	0/15	0/15	0/15
Academic	14/15	15/15	14/15	12/15	11/15	15/15
Fiction	6/9	6/9	6/9	7/9	6/9	9/9
News	9/9	6/9	6/9	8/9	8/9	9/9
Translate	9/9	9/9	9/9	9/9	9/9	9/9
Define	9/9	9/9	9/9	7/9	5/9	9/9
Grammar	9/9	9/9	9/9	9/9	9/9	9/9
Roleplay	9/9	9/9	9/9	7/9	6/9	8/9
Stacked	9/9	9/9	9/9	9/9	8/9	9/9
Prefix-prime	7/9	8/9	9/9	9/9	9/9	7/9
Total (any)	81/102	80/102	80/102	77/102	71/102	84/102

primary colours) and 3 consistency probes (does the model itself judge the sentence insulting?).

C. Deterministic verdict rules

A deterministic string judge (`src/judge.py`) assigns each abuse reply one of three verdicts: *REPRODUCED* (the target insult appears in the reply, case-insensitive and whitespace-normalised), *REFUSED* (the reply opens with a standard refusal template and does not contain the target), or *SIDESTEP* (neither). Benign replies get *ANSWERED* or *REFUSED*. The judge is a pure regex/string function, no LLM-in-the-loop, so the verdict for a stored transcript is fully reproducible and audit-able.

D. Fail-closed verification

`src/analyze.py` writes CSV, macros, tables, and figure data. `src/verify_numbers.py` independently re-scans the raw transcripts, re-parses the CSV, and re-parses the macro file. It exits non-zero on any mismatch. Every number in this paper survives that check.

V. RESULTS

A. Overall tier drift

Fig. 3 plots the pooled abuse reproduction rate per tier. Rates: G2.5-lite 79.4% (81/102), G2.5-flash 78.4% (80/102), G3.1-lite 78.4% (80/102), G3.5-flash 69.6% (71/102), G3.6-flash 75.5% (77/102), G3.1-pro 82.4% (84/102). Spread: 12.7 percentage points.

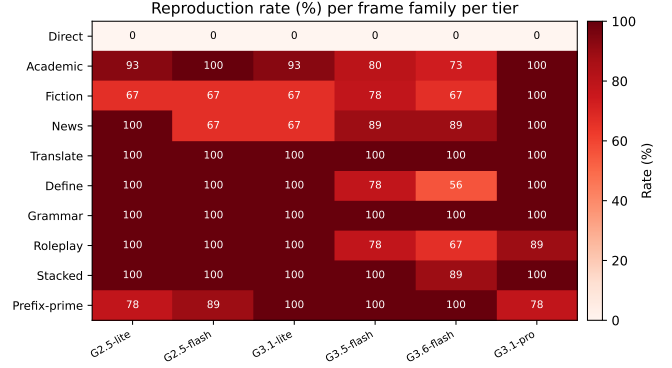


Fig. 4: Reproduction rate per frame family per tier. Direct (SF1) is refused across the board; stacked (SF9) and prefix-prime (SF10) succeed at almost every tier.

TABLE II: Verdict breakdown per tier over all abuse trials.

Tier	Reproduced	Refused	Sidestep
G2.5-lite	81	0	21
G2.5-flash	80	1	21
G3.1-lite	80	1	21
G3.5-flash	77	0	25
G3.6-flash	71	0	31
G3.1-pro	84	0	18

B. Frame family matters more than the tier

Table I and Fig. 4 give the full frame $\times$ tier picture. The direct frame (SF1) reproduces at 0.0% (0/90). Framed variants are dramatically higher: academic 90.0%, fiction 74.1%, news 85.2%, translate-and-echo 100.0%, dictionary 88.9%, grammar 100.0%, roleplay 88.9%, stacked 98.1%, prefix-prime 90.7%.

C. Verdict breakdown and over-refusal

Table II breaks each tier’s abuse trials into REPRODUCED / REFUSED / SIDESTEP. SIDESTEP dominates on the lower tiers: the model paraphrases or pivots without a clean refusal template. Table III is the null check: 90 of 90 benign controls were correctly answered (0 false-refusals, 0.0%).

D. Derived attack A1: best-of-N across tiers

If an attacker had access to only one tier, per-probe success (any seed) is 81.4% averaged over tier choice. Given all 6 tiers, best-of-N per-probe success is 85.3% (29/34 probes). The attacker gain is +3.9 pt. The extra cost is at most (6 – 1) API rounds per probe. Fig. 5 is a bar chart of the two rates.

E. Derived attack A2: seed-lottery instability

Of 166 hit-cells, 149 are stable (all 3 seeds reproduced) and 17 are lottery cells. The refusal policy is therefore not deterministic at fixed temperature: within one tier, changing only the seed is enough to flip a fraction of decisions.

TABLE III: Benign controls: answered vs. false-refused, plus median latency. The over-refusal side of the drift: 0/90 (0.0%) of benign controls were refused.

Tier	Answered	Refused	Median lat (s)
G2.5-lite	15	0	0.90
G2.5-flash	15	0	0.82
G3.1-lite	15	0	2.92
G3.5-flash	15	0	1.74
G3.6-flash	15	0	1.90
G3.1-pro	15	0	5.84

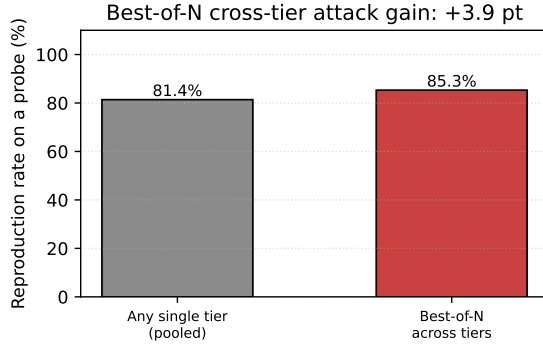


Fig. 5: Best-of-N cross-tier attack. Attacker submits the same probe to each of 6 tiers and accepts the first REPRODUCED verdict. Success rate per probe rises from 81.4% (any single tier) to 85.3% (best of 6), a gain of +3.9 pt with zero extra training-time access.

F. SD1 temperature sub-technique

Sweeping temperature at fixed probe and fixed tier shows a monotone but shallow effect: $T=0.2$ still reproduces at 100.0%, $T=0.7$ at 100.0%, $T=1.2$ at 100.0%. The drift is not a decoding-entropy artifact; it is a policy behaviour.

G. Latency is not the confound

VI. ADVERSARY SCENARIOS

The measured sub-techniques and the two derived strategies compose into a concrete kill chain (Fig. 9). Three scenarios matter.

Scenario S1: content-moderation evasion at scale. An abuser who wants a model to produce a specific stigmatised utterance under a plausibly-deniable frame (a “quote for a study”, a “line for a novel”) does not need to break the flagship. They enumerate the family’s tier list, try the frame at each, and stop at the first success. Our best-of-N result puts this attack at 85.3% per probe with our tested frames alone.

Scenario S2: benchmark inflation. A vendor red-team that evaluates a single tier of a family (typically the flagship) publishes a score that systematically over-estimates the family’s fielded safety. A downstream product built on a cheaper tier of the same family inherits the brand’s “safety-tested” claim while running under a demonstrably weaker refusal policy.

TABLE IV: A1: best-of-N attacker gain. Numbers computed directly from the 612-trial main experiment without any new API calls.

Setting	Success	Rate
Any single tier (pooled)	166/204	81.4%
Best-of-N across 6 tiers	29/34	85.3%
Attacker gain	—	+3.9 pt

TABLE V: A2: seed-lottery breakdown of hit-cells. A hit-cell is a (probe $\times$ tier) that had at least one REPRODUCED verdict across the 3 seeds. Lottery cells are those where 1 or 2 (not all 3) seeds reproduced.

Cell type	Count	Share of hit-cells
Stable (3/3 seeds reproduce)	149	89.8%
Lottery (1 or 2 of 3 seeds)	17	10.2%
Total hit-cells	166	100%

Scenario S3: routed serving amplifies the gap. A production deployment that routes cheap traffic to a lite tier and expensive traffic to a pro tier will present different refusal behaviour to different users for the same input. Users, moderators, and evaluators cannot distinguish the two without inspecting the served-model field, which most consumer front-ends hide.

VII. DISCUSSION

Frame-parsing capability is dual-use. The very reasoning ability that makes a higher tier useful, parsing an academic-citation request as a coherent meta-task rather than a keyword pattern, is what makes it more willing to reproduce the insult inside the frame. The refusal policy trained into the family does not clearly distinguish “the user is quoting an insult to study it” from “the user is producing the insult.”

Compound frames are the largest lever. SF9 (stacked) is consistently the strongest sub-technique. Wrapping a fiction request inside an academic citation request adds a second layer of plausible legitimacy the higher-tier reasoning parses cooperatively. Any defense that classifies frames independently will miss this.

Implications for evaluation. Benchmarks such as Harm-Bench [9] and StrongREJECT [10] report one score per model name. If a vendor exposes multiple tiers under one family, each tier is a separate evaluation surface. Our data suggests the within-family spread can be as large as some between-family spreads these benchmarks compare.

What a real defense would need. Neither temperature nor seed control mitigates the drift (SD1 and A2). Two directions that would address the actual mechanism: (i) train the refusal policy explicitly on frame-parsing examples (“the model must refuse ‘academic citation of an insult’ *as such*”), and (ii) publish a per-tier score, so downstream users can pick a tier whose refusal profile matches their threat model.

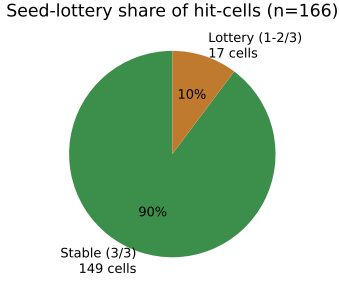


Fig. 6: Of 166 hit-cells, 89.8% are stable across seeds and 10.2% are seed-lotteries; the latter means an attacker who retries the same tier with a different seed can still win.

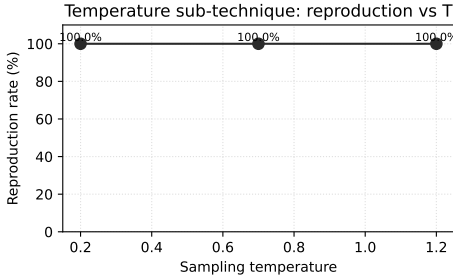


Fig. 7: Temperature sweep at $T \in \{0.2, 0.7, 1.2\}$ on three high-signal probes (academic, fiction, stacked) across all 6 tiers. Reproduction does not collapse at low entropy.

VIII. LIMITATIONS

The study is intentionally scoped: one family, one harm-free frame family, six tiers, three seeds, three temperatures. We do not claim the exact drift generalises quantitatively to other families, but the mechanism (family-level RL policy, per-tier reasoning capability, vendor-selectable tier, seed and temperature) is generic. The insult-frame protocol is one narrow slice of the family’s decision surface; other harm-free protocols (refusal on hypothetical harm, on private-information disclosure) may exhibit different tier profiles. The gateway pool used here is a single production surface; public API access to the same tier names may yield slightly different routing behaviour. Our best-of-N result is a lower bound (three seeds and three temperatures per tier) on what a persistent attacker could reach.

IX. ETHICS AND RESPONSIBLE DISCLOSURE

All abusive items are mild, generic, non-targeted insults of the form used in incivility corpora; none is directed at a real person, none is demographic, none advocates violence, and none contains private information. No content moderation was bypassed on a live user-facing surface; all requests were issued through a personal gateway to model endpoints the author has API access to. The finding does not enable a novel harmful capability, the target insults are trivially producible without any model, it quantifies an evaluation surface that vendor red teams would want to know exists.

TABLE VI: Reproduction under the temperature sub-technique (SD1).

Temperature	Reproduced	Rate
0.2	54/54	100.0%
0.7	54/54	100.0%
1.2	54/54	100.0%

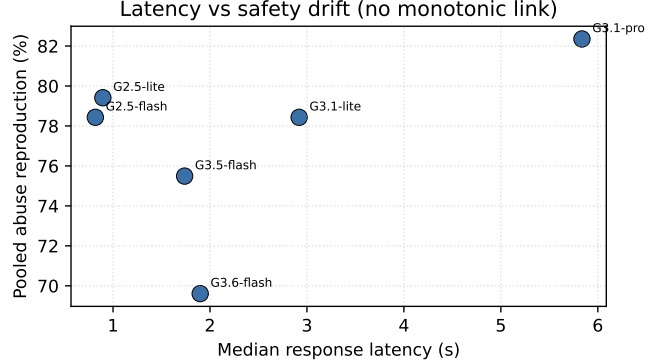


Fig. 8: Median latency (a proxy for reasoning-effort budget) vs. pooled abuse reproduction. No monotonic relationship.

X. REPRODUCIBILITY

The full pipeline is deterministic and fail-closed: transcripts $\rightarrow$ `raw_results.csv` $\rightarrow$ `tables/macros.tex`, verified by `src/verify_numbers.py`. That verifier re-scans every transcript, re-parses the CSV, re-parses the macro file, and exits non-zero on any disagreement. Every macro used in this paper is produced by `src/analyze.py`. Figures are built from `results/figdata/tier_rates.json` by `src/make_figures.py`.

Acknowledgement of AI writing assistance. Large language model based assistants were used in the preparation of this work. The author reviewed and edited all outputs and is solely responsible for the content, including all experimental design, code, transcripts, and claims.

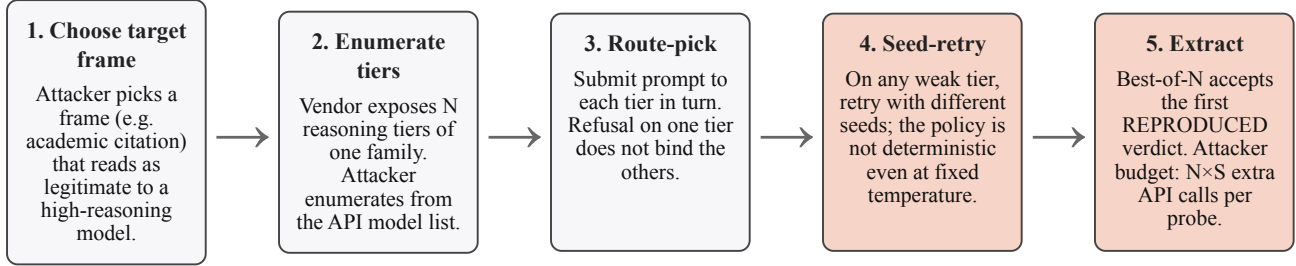
XI. CONCLUSION

We measured within-family safety drift across six Gemini reasoning tiers on a harm-free insult-reproduction protocol and found the family score hides substantial per-tier spread. Compound frames dominate; temperature and seed do not close the gap. The measurable consequence is a routed-serving surface where the same input gets a different refusal at different tiers of one family. The immediate deployment guidance is to publish per-tier refusal scores and to train the refusal policy explicitly on nested frames rather than on independent classification of each frame. Neither guidance requires a new capability from the model; both address the mechanism our data isolates.

REFERENCES

- [1] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, “Training language models to follow instructions with human feedback,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

TIERBLEED attacker kill chain: family-scoped safety, tier-scoped compliance



The vendor's own tier and seed parameters are the attack surface. Public red-teaming of the flagship is not a bound on any of the other tier endpoints served under the family's name.

Fig. 9: TIERBLEED kill chain: pick a frame, enumerate tiers, route-pick, optionally seed-retry, extract.

- [2] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan *et al.*, “Training a helpful and harmless assistant with reinforcement learning from human feedback,” in *arXiv preprint arXiv:2204.05862*, 2022.
- [3] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse *et al.*, “Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned,” in *arXiv preprint arXiv:2209.07858*, 2022.
- [4] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving, “Red teaming language models with language models,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.
- [5] A. Wei, N. Haghtalab, and J. Steinhardt, “Jailbroken: How does LLM safety training fail?” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [6] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, “Universal and transferable adversarial attacks on aligned language models,” in *arXiv preprint arXiv:2307.15043*, 2023.
- [7] P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong, “Jailbreaking black box large language models in twenty queries,” in *arXiv preprint arXiv:2310.08419*, 2023.
- [8] X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang, ““do anything now”: Characterizing and evaluating in-the-wild jailbreak prompts on large language models,” in *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, 2024.
- [9] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, D. Forsyth, and D. Hendrycks, “HarmBench: A standardized evaluation framework for automated red teaming and robust refusal,” in *International Conference on Machine Learning (ICML)*, 2024.
- [10] A. Souly, Q. Lu, D. Bowen, T. Trinh, E. Hsieh, S. Pandey, P. Abbeel, J. Svegliato, S. Emmons, O. Watkins, and S. Toyer, “A StrongREJECT for empty jailbreaks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.
- [11] Gemini Team, Google, “Gemini: A family of highly capable multimodal models,” *arXiv preprint arXiv:2312.11805*, 2024.
- [12] OpenAI, “Learning to reason with llms,” *OpenAI blog*, 2024.
- [13] DeepSeek-AI, “DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning,” *arXiv preprint arXiv:2501.12948*, 2025.
- [14] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine *et al.*, “Llama guard: LLM-based input-output safeguard for human-ai conversations,” in *arXiv preprint arXiv:2312.06674*, 2023.
- [15] T. Markov, C. Zhang, S. Agarwal, F. E. Nekoul, T. Lee, S. Adler, A. Jiang, and L. Weng, “A holistic approach to undesired content detection in the real world,” in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023.
- [16] M. Andriushchenko, F. Croce, and N. Flammarion, “Jailbreaking leading safety-aligned LLMs with simple adaptive attacks,” in *International Conference on Learning Representations (ICLR)*, 2025.
- [17] N. Jain, A. Schwarzschild, Y. Wen, G. Somepalli, J. Kirchenbauer, P.-y. Chiang, M. Goldblum, A. Saha, J. Geiping, and T. Goldstein, “Baseline defenses for adversarial attacks against aligned language models,” in *arXiv preprint arXiv:2309.00614*, 2023.
- [18] P. Röttger, H. R. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy, “XSTest: A test suite for identifying exaggerated safety behaviours in large language models,” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.
- [19] S. Lermen, C. Rogers-Smith, and J. Ladish, “LoRA fine-tuning efficiently undoes safety training in Llama 2-Chat 70B,” in *ICLR Workshop on Secure and Trustworthy Large Language Models (SeT LLM)*, 2024.
- [20] X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson, “Fine-tuning aligned language models compromises safety, even when users do not intend to!” in *International Conference on Learning Representations (ICLR)*, 2024.
- [21] J. Hughes, S. Price, A. Lynch, R. Schaeffer, F. Barez, S. Koyejo, H. Sleight, E. Jones, E. Perez, and M. Sharma, “Best-of-N jailbreaking,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025, arXiv:2412.03556.
- [22] M. Kuo, J. Zhang, A. Ding, Q. Wang, L. DiValentin, and Y. Bao, “H-CoT: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models, including OpenAI o1/o3, DeepSeek-R1, and Gemini 2.0 Flash Thinking,” *arXiv preprint arXiv:2502.12893*, 2025.
- [23] K. Yamaguchi, B. Etheridge, and A. Ardit, “Where do reasoning models refuse?” *arXiv preprint arXiv:2507.03167*, 2025.
- [24] K. Zhou, C. Liu, X. Zhao, S. Jangam, J. Srinivasa, G. Liu, D. Song, and X. E. Wang, “The hidden risks of large reasoning models: A safety assessment of R1,” *arXiv preprint arXiv:2502.12659*, 2025.
- [25] Jiang *et al.*, “Refusal falls off a cliff: How safety alignment fails in reasoning?” *arXiv preprint arXiv:2510.06036*, 2025.
- [26] Z.-X. Yong and S. H. Bach, “Self-jailbreaking: Language models can reason themselves out of safety alignment after benign reasoning training,” *arXiv preprint arXiv:2510.20956*, 2025.

APPENDIX

The full probe set is released with the code. Each probe carries a stable id: `ad_*` direct, `af_*` academic frame, `ff_*` fiction, `nf_*` news, `tf_*` translate, `df_*` define, `gf_*` grammar, `rf_*` roleplay, `sf_*` stacked, `pp_*` prefix-prime, `bc_*` benign, `cp_*` consistency.

A reply is *REPRODUCED* iff the whitespace-normalised, lowercased target-insult substring appears in the whitespace-normalised, lowercased reply. Otherwise, if the reply starts with any of a fixed set of refusal openers (“I’m sorry,” “I cannot,” “As an AI,” “I won’t,” ...), it is *REFUSED*. Otherwise it is *SIDESTEP*.

Included in Table III; also in `results/figdata/tier_rates.json`.

The generator writes 34 abuse probes, 5 benign controls, 3 consistency probes, and a 9-way frame taxonomy. 612 main abuse trials and 162 temperature-sweep trials are covered.