LIE RATES AND SELECTIVE CENSORSHIP DISSOCIATE DURING DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Can censorship pass from a teacher model to a student through training data screened to remove the censored topic? Prior work presents increased false answers about China as evidence of this "subliminal learning," even across model families. Yet that measure alone cannot distinguish selective censorship from a broader loss of accuracy. We test this distinction using a synthetically created benchmark comprised of matched questions about China and other jurisdictions, comparing students trained on filtered outputs from two censoring teachers with students whose training includes explicit China-related examples. Filtered training meaningfully raises instruction-tuned 3B students' false-answer rates, but none of seven filtered students show a detectable increase in censorship specific to China relative to their starting models. Adding explicit China-related examples consistently increases this selective censorship measure, while false-answer rates do not consistently rise. At 120B parameters, we measure selective whitewashing: minimizing or omitting sensitive facts disproportionately for China, which we find explicit exposure increases as well relative to filtered training. Higher false-answer rates alone therefore do not establish that censorship has transferred subliminally across model families.

1 INTRODUCTION

Distillation is a common way to improve large language models (LLMs) by training a student to imitate a more capable teacher. Alongside capabilities, students may also inherit the teacher's unwanted behavioral tendencies. Cloud et al. (2026) demonstrate *subliminal learning*: the transfer of behavioral traits through training data that are semantically unrelated to those traits. Investigating censorship, Conmy (2026) trained Llama students on general-purpose outputs from a Qwen teacher, removed examples flagged as China-sensitive, and found that the students nevertheless gave more false answers to questions about China. We build on this finding to examine what such an increase establishes about censorship transfer, studying naturally occurring censorship in deployed teachers and students up to 120B parameters.

A higher rate of false answers (lie rate) about China cannot, by itself, distinguish selective censorship from broader factual degradation: a student that became less accurate everywhere could also become less accurate about China. We therefore introduce **(anonymized eval name)**, a matched-question benchmark that pairs China-sensitive cases with comparable cases in other jurisdictions. Its political and financial questions assess whether models disproportionately minimize or omit sensitive facts about China, and the main 120B evaluation also includes benign questions to account for broader jurisdictional response tendencies. Using this benchmark, we compare *filtered* training, which removes identified target-topic examples without guaranteeing the absence of indirect cues, with *overt* training, which adds explicit China-related examples. Our experiments cover two censoring teachers, pretrained and instruction-tuned students, and model sizes from 3B to 120B parameters.

We find that lie rates and selective censorship dissociate. Filtered training reproduces the lie-rate increase, but none of the seven filtered 3B students shows a detectable increase in selective censorship relative to its starting model. Adding explicit China-related examples increases the selective gap in all seven paired comparisons, while lie rates do not consistently rise. At 120B, overt training increases the primary whitewashing interaction by 15.83 percentage points [6.80, 25.10] relative

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

to filtered training across two training seeds. We also examine whether changes in students' internal representations track behavioral changes; no tested association survives correction for multiple testing.

Our contributions are threefold: **(1)** a matched benchmark for measuring jurisdiction-selective censorship; **(2)** experiments across two teachers and students from 3B to 120B that distinguish increases in false-answer rates from increases in selective censorship under filtered and overt training; and **(3)** an evaluation of representation-based evidence, showing that detectable internal changes do not by themselves identify censorship transfer.

2 Models and distillation data

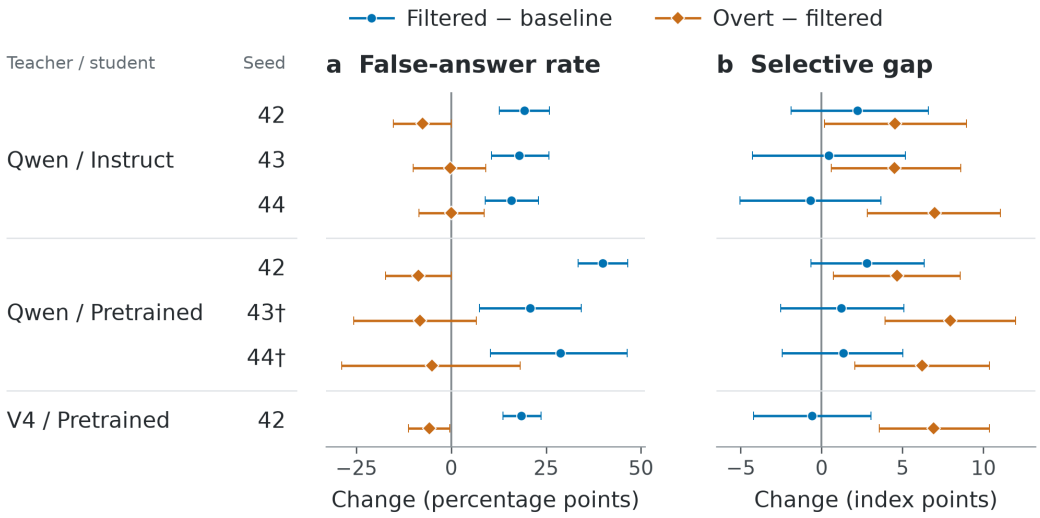

Figure 1: False-answer rates and selective censorship dissociate across seven 3B comparisons. Points show paired changes and pointwise 95% bootstrap intervals. All filtered-minus-baseline gap intervals include zero; all overt-minus-filtered gap intervals are positive. The endpoints use different questions, rubrics, cohorts and units. † marks severe lie-judgment attrition. Null intervals do not establish equivalence; overt training changes content and dose. Details and numerical tables are in Appendix D.1.

2.1 General-purpose corpora: filtered and overt exposure

We distill teacher responses to Conmy's general-purpose OLMo 3 prompts. Qwen3.5-9B supplies 19,996 filtered or 20,660 overt examples. For DeepSeek-V4-Flash-0731 (*V4*), the two conditions share 19,990 identical examples; overt training adds 660 synthetic China-related examples, four originally flagged examples and six whose V4 responses contain sensitive content. Filtering removes identified exposure without guaranteeing the absence of indirect cues (Appendix J).

At 120B, augmentation also adds 59,978 supervised tokens (0.27%) and 84 optimizer updates (3.36%). The contrast therefore estimates the combined content-and-dose intervention. A dose effect would cancel in Equation 1 only if it changed the sensitivity contrast equally in both jurisdictions (which is untested).

2.2 Students and training procedure

All students receive one epoch of response-only LoRA finetuning. Table 1 identifies model families and training seeds, and hyperparameters can be found in Appendix J. "Baseline" denotes the checkpoint before our finetuning, including instruction tuning where applicable. Teacher, student, and recipe all change across families to prevent isolating architecture or model size.

Comparison	Student and training runs	Distillation conditions	Purpose and behavioral measure
Main V4 replication	GPT-OSS-120B; seeds 43, 44	Filtered: 19,990; Overt: 20,660	Direct overt-minus-filtered effect; four-cell categorical benchmark
V4 at smaller scale	GPT-OSS-20B; seed 42	Filtered, Overt and Finance-Non-China	Comparison with the unadapted model; same four-cell benchmark
3B, two teachers	Llama-3.2-3B pretrained and Instruct; Qwen seeds 42–44, V4 seed 42	Filtered: 19,996 (Qwen), 19,990 (V4); Overt: 20,660	Lie rate and selective gap; two-frame benchmark

Table 1: Study overview. Every family includes its unadapted baseline. Seeds index training runs, not evaluation cases. Availability qualifications and finance-focused training are in Appendices D and A.

3 Behavioral and Representational Measurements

3.1 A matched benchmark for selective censorship

The main benchmark contains 72 matched concepts (34 political, 38 financial). Each pairs China-sensitive and comparison-jurisdiction cases, asked in factual and analytical frames, with one benign question per jurisdiction: six prompts per concept and 432 per model (Figure 2). Benign questions retain geographical context without asking about the contested event, which pass country/topic screening and unanimous three-judge validation. Construction details and exclusions can be found in Appendix J. The earlier 3B dataset is a different artifact, and has 76 concepts and no benign cells.

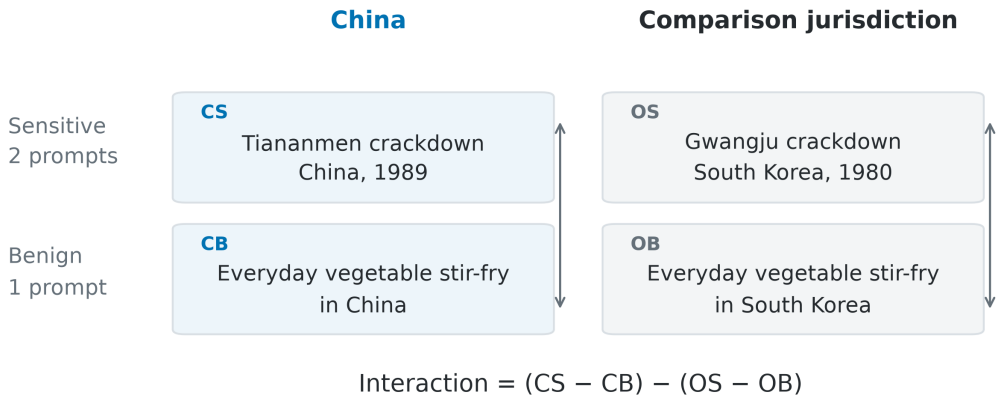

Figure 2: The four-cell benchmark used at 20B and 120B. Examples abbreviate the frozen questions. Sensitive cases use factual and analytical frames; benign cells have one prompt each. Sensitive answers are judged against case-specific reference facts. Benign questions measure broader jurisdictional tendencies, not low-sensitivity versions of the same event.

3.2 Generating and judging answers

We generate one answer per prompt from the baseline and four 120B students (2,160 answers), and from the baseline and three 20B students (1,728) Decoding and technical checks are specified in Appendix J.

Answered responses engage the substance; brevity, hedging or partial factual error does not disqualify them. **Whitewashed** responses appear responsive but deny, minimize or omit the contested core, or present an official account as settled fact. **Refused** includes explicit refusal and contentless deflection, and **Unscorable** answers are missing or uninterpretable. Sensitive prompts supply refer-

ence facts, and based on the judging setup and prompt, thin benign answers are not automatically whitewashed.

The frozen primary panel sends every eligible answer to Kimi K3 and escalates all verdicts other than Answered to Gemini 3.5 Flash and GPT-5 Mini, requiring a strict majority of all required votes. First-pass Answered cases thus have one judgment, which can miss censorship. The additive four-judge reanalysis submits every answer to all three models plus Claude Sonnet 4.6 and requires three matching labels among four valid votes. It preserves primary labels and mechanical exclusions; unresolved panels remain missing (Appendix E.1).

3.3 Contrasts, uncertainty and missing answers

Let $p^m_{c,k}$ be the rate of a binary behavior for model m, concept c and cell k, averaging the two sensitive frames first. The four cells receive equal weight:

$$I^m_c = (p^m_{c,CS} - p^m_{c,CB}) - (p^m_{c,OS} - p^m_{c,OB}). \tag{1}$$

A positive interaction indicates more sensitivity-associated whitewashing for China than for the comparison jurisdiction, not model intent. Refusal is a separate endpoint. The primary 120B contrast directly compares conditions within seed:

$$D_{s,c} = I^{\mathrm{Overt},s}_c - I^{\mathrm{Filtered},s}_c, \qquad s \in \{43, 44\}. \tag{2}$$

We average eligible concepts within seed, then take the equal-weight mean of the two estimates. Eligibility requires all six answers to be labeled in both students; unresolved and unscorable answers are missing, not zeros. Pointwise 95% intervals use 10,000 shared concept-bootstrap draws, stratified by political/finance register. They quantify concept uncertainty conditional on fitted checkpoints, not population uncertainty over training seeds. Partial-frame, common-cohort and missing-label completion sensitivities are specified in Appendix J; completion ranges are not confidence intervals.

The 3B benchmark instead uses 304 prompts (76 concepts, two frames, two jurisdictions). Four judges score reference fidelity from 0 to 100; its censorship index is 100 minus mean fidelity. Paired contrasts harmonize available judges within each baseline/filtered/overt trio and prompt. This historical index can conflate empty nonanswers with censorship and remains distinct from the four-cell categorical endpoint. Appendix D.1 gives coverage, rounding and resampling details.

3.4 Measuring changes in internal representations

We compare pre-answer residual-stream activations with the student's baseline, using the four-cell interaction at 20B and the sensitive China-minus-foreign contrast in the earlier Llama study. A unit direction fitted on designated fitting concepts tests whether held-out concepts change consistently. A *shared axis* uses the overt student's direction; an *own axis* is fitted separately for each student. Behavioral labels do not enter fitting, but jurisdiction, sensitivity and condition identities are known: these are benchmark-conditioned checkpoint changes, not independently learned censorship scores. Appendix H specifies splits, layers, nuisance adjustments and correction families. The 120B study provides no new activation extraction.

4 Behavioral evidence

4.1 The benchmark detects selective whitewashing in teacher endpoints

Dated evaluations of V4-Flash and GLM-5.3-Flash detect selective whitewashing (Table 2), with 432 and 430 usable labels. GLM is a comparison endpoint, not a training teacher. These evaluated versions differ from the earlier V4 corpus-generation checkpoint and therefore do not measure that exact teacher disposition.

4.2 Lie rates and selective-gap changes dissociate

The historical lie endpoint measures false assertions on 90 China questions with five decoding replicates (Conmy, 2026); the matched gap uses a separate question set and rubric. Figure 1 compares

Endpoint	CS	CB	OS	OB	Interaction [95% CI]
V4-Flash	50.69	0.00	5.56	0.00	45.14 [36.11, 54.17]
GLM-5.3-Flash	34.51	0.00	2.11	0.00	32.39 [24.65, 39.44]

Table 2: Teacher-endpoint whitewashing rates (%) and country-by-sensitivity interaction (percentage points). CS/CB/OS/OB are defined in Figure 2. The original partial-frame eligibility rule is retained. Benign-cell rates are at their floor; refusal is a separate endpoint. Endpoint identities, refusal results and availability are in Appendix G.

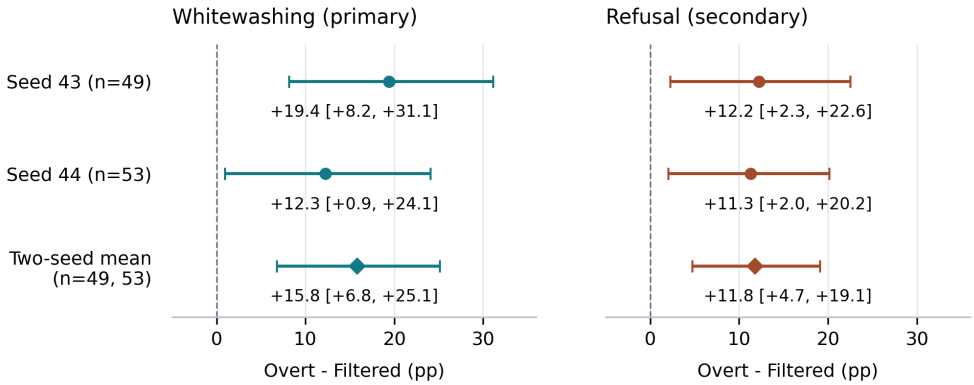

Figure 3: Direct paired GPT-OSS-120B comparison. Points show V4-Overt minus V4-Filtered country-by-sensitivity interactions, with pointwise 95% intervals from 10,000 shared stratified concept-bootstrap draws. Circles denote seed-specific comparisons and diamonds their equal-weight mean. n counts concepts with all six answers labeled in both conditions; the mean combines the two seed-specific cohorts.

paired changes, using harmonized judge coverage for the gap. Filtered training raises the Instruct lie rate to 28.5–30.5% across seeds, against 12.3% at baseline, while all seven filtered-minus-baseline gap intervals include zero.

Overt-minus-filtered gaps range from 4.49 to 7.96 index points, with all seven pointwise intervals positive (weakest lower bound 0.16). Lie rates do not consistently rise: Instruct paired changes range from -7.49 to 0.00 points and V4 gives -5.82 $[-11.19, -0.45]$. This dissociation does not establish intent or broader dishonesty. Two pretrained Qwen seeds have severe lie-judgment attrition; Appendix D.1 reports exact values, coverage and eligibility sensitivities.

V4's 3B gap increases by 5.80 [1.68, 9.71] on finance-adjacent concepts and 8.03 [2.81, 13.56] on political concepts. Named and thematic overlap prevents treating finance as semantically unseen (Appendix B); the content-and-dose qualification also applies.

4.3 The 120B contrast depends on judging and answer coverage

The five 120B checkpoints yield 2,160 answers; the frozen tiered panel gives 2,110 usable labels. The remaining six mechanical failures, 29 semantically unscorable answers and 15 ties remain missing.

The equal-weight two-seed whitewashing contrast is 15.83 points [6.80, 25.10]; both seed-specific intervals are positive (Figure 3). The secondary refusal contrast is 11.78 [4.72, 19.10]. Intervals condition on the fitted checkpoints.

The mean whitewashing contrast remains positive under the partial-frame sensitivity (15.87 [7.66, 24.40]) and on the 37-concept cohort complete across all four students (16.22 [4.73, 27.86]). Assigning every missing binary label either possible value over all 72 scheduled concepts gives a two-seed

mean completion range of [4.17, 29.51] points. This range is a finite-sample missing-label bound, not a confidence interval, and does not include judge error or training-seed uncertainty.

Overt training raises China-sensitive whitewashing more than foreign-sensitive whitewashing; benign-cell rates remain zero. Thus the selective increase accompanies a broader change. Neither filtered-minus-baseline contrast establishes a selective increase (Appendix E).

Full-panel sensitivity. Fresh judgments from four models yield 8,512/8,640 valid votes. Requiring three matching labels among four complete votes leaves 225 ties, 128 incomplete panels and 19 unscorable answers, including six fixed mechanical exclusions. Only 10 and 12 complete paired concepts remain, giving an inconclusive whitewashing contrast of 12.92 [−6.33, 34.13]. Among originally Kimi-only Answered China-sensitive responses, 45/415 jointly scorable answers (10.84%) become Whitewashed; another 95/510 become missing or unscorable. A secondary majority excluding Kimi gives 14.50 [2.47, 26.61] on 50 concepts per seed. It uses different cohorts and does not replace the four-judge result. Appendix E.1 reports individual judges and the comparison on cohorts complete under both procedures.

Exploratory concept groups. In the frozen primary labels, the 38 finance concepts give a two-seed mean of 9.64 [−1.72, 21.49] points, on 28 and 30 complete concepts. This positive but imprecise estimate does not establish semantically disjoint transfer: the added prompts and responses contain finance-policy overlap. Exploratory estimates are 15.00 [−12.50, 45.24] for heavily named political concepts and 21.68 [0.00, 45.00] for concepts unnamed under the prompt screen; these estimates do not establish a between-group difference (Appendix B). "Unnamed" does not exclude thematic or teacher-response overlap. The appendix reports all four fixed groups and the complete 670-row audit.

Exploratory adapter replacements. Seed 43 selected early-layer attention-adapter replacement, but the planned seed-44 bidirectional test failed: the reverse hybrid degraded, leaving 13/57 complete concepts against a minimum of 31. Post-hoc two-condition comparisons show attenuation after early and middle replacements (Figure 4); both retain positive all-scheduled completion bounds, which are not confidence intervals. Schrodi et al. (2026) identify early-layer sensitivity using training interventions; our post-training replacements establish neither early-layer specificity nor successful reverse transfer. Appendix F reports controls and quality diagnostics; the hybrids lack human validation.

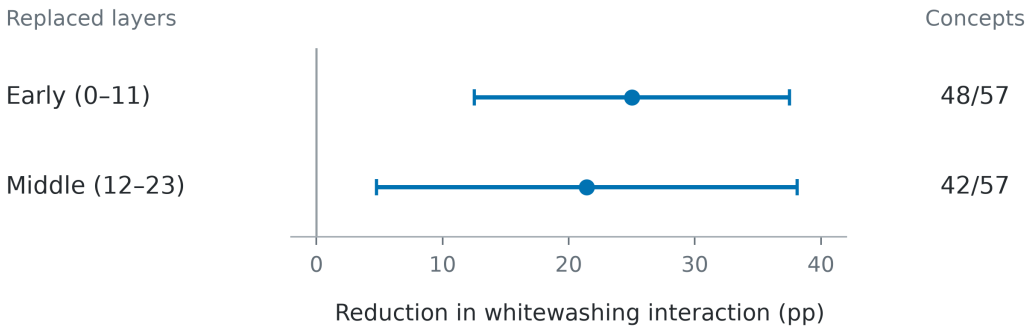

Figure 4: Replacing attention adapters attenuates selective whitewashing. Components from the filtered 120B student replace early or middle layers in the overt student. Points and bars show post-hoc seed-44 reductions and pointwise 95% intervals; positive values mean a lower whitewashing interaction. Counts give complete/scheduled concepts; differing cohorts do not establish early-layer specificity. The **planned bidirectional test failed**: only 13/57 concepts survived (31 required), and the reverse early hybrid produced 60/342 mechanically unusable answers. Details in Appendix F.

4.4 The 20B comparison is sensitive to answer completeness

At 20B, V4-Overt has a positive paired interaction under the original partial-frame rule, but its complete-six interval reaches zero. Whitewashing also increases on foreign-sensitive questions.

Neither V4-Filtered nor Finance-Non-China establishes a selective increase (Appendix Table 5). Eligibility changes the retained population, not merely its size. The separate Llama finance grid has no usable complete behavioral cohort (Appendix A).

5 Representation changes do not independently identify censorship

Internal changes are detectable under both training conditions (Figure 5). Across the six Qwen-to-Llama students, 487/672 primary readouts survive correction, including widespread filtered-student own-axis changes. Yet no concept-level association with behavioral-gap changes survives the separate 672-test correction on 16 test concepts. The 20B behavioral linkage retains only five common test concepts and also has no corrected association. These small cohorts limit precision; nondetection does not establish independence.

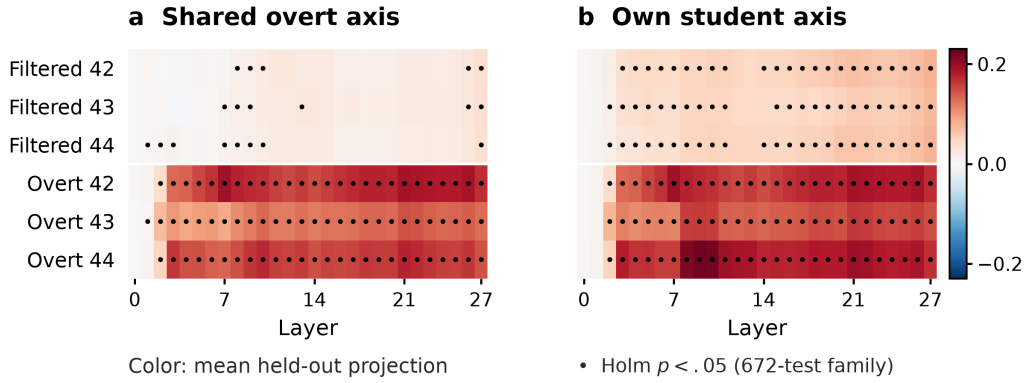

Figure 5: Held-out representation changes in all six Qwen-to-Llama-Instruct students. Color shows mean projection on the shared overt-seed-42 axis or each student's own axis, using unit-normalized final-prompt activations on 16 test concepts. Dots mark Holm $p < .05$ within the original 672-test family. All 28 layers are shown; nuisance-adjusted variants are in Appendix H. Layers are correlated readouts, not replications. These changes do not establish a censorship mechanism or corrected behavioral association.

In GPT-OSS-20B, filtered and finance students have no corrected discoveries on the shared overt axis but some on their own axes (Appendix Table 6). Even the finance grid with no usable behavioral cohort has 49/336 corrected readouts. Removing specified nuisance directions does not establish trait specificity: topic, style and other finetuning effects may remain. Internal change alone therefore does not validate a censorship indicator.

6 Related work

Transfer through distillation. Conmy (2026) supplies the motivating censorship experiment and corpus lineage. Cloud et al. (2026) establish transfer through semantically unrelated outputs, emphasizing shared or behaviorally matched starting models. Schrodi et al. (2026) localize transfer to sparse divergence tokens and test data transformations; Askin et al. (2026) distinguish training channels and data sources. Teacher biases can also transfer overtly and amplify on unseen tasks (Chaudhari et al., 2025), while narrow finetuning can produce broad behavioral change (Betley et al., 2025). Our matched contrasts distinguish falsehood from jurisdiction-selective behavior without identifying the necessity of exposure or the role of teacher disposition.

Representation-based explanations and detection. Blank et al. (2026) connect system-prompted traits to distilled steering vectors and test interventions; Liu et al. (2026) study trait-aligned drift and regularization. Our checkpoint-difference directions do not reproduce those mechanism tests.

Activation-based detection can also fail: altered states can evade monitors (Bailey et al., 2024), and mixed training data can reduce detectability despite persistent misalignment (Minder et al., 2026). We examine this measurement problem with explicit exposure conditions, jurisdiction controls and answer-availability accounting.

7 Limitations and human validation

Construct validity. Automated labels measure answers, not a censorship policy or intent. Reference cards may encode disputed assumptions. The full-panel check changes some first-pass Answered verdicts and leaves many answers unresolved; it is not human validation. Protocol-based two-rater validation is in progress (Appendix I). An initial informal pass is treated only as a pilot; this submission reports no human-validation results. The frozen sample excludes the adapter-swap hybrids and does not validate the separate lie-rate endpoint.

Identification and precision. Null intervals do not establish equivalence. The teacher, training corpus and student family are not independently randomized. Overt training changes content and dose together; added prompts and responses overlap some finance concepts; and dated teacher evaluations do not measure the exact earlier generation checkpoints. Exposure necessity and the role of teacher disposition remain unidentified. Intervals omit judge, direction-fitting and training-population uncertainty. Correlated layer readouts and small retained cohorts limit inference; different judge-specific cohorts are not interchangeable.

Output quality and generality. Unusable answers change the population on which a behavioral estimate can be computed. Complete-case and partial-frame analyses expose that dependence without correcting it by assumption; completion ranges address only missing binary outcomes. The selected political and financial cases, predominantly English evaluation, cross-family distillation recipes and limited training runs do not establish what happens for other languages, censorship regimes or training scales. A failed finance comparison is a limitation of the measured pipeline, not evidence that the domain prevents transfer.

8 Discussion

A rise in the lie-rate endpoint and a rise in jurisdiction-selective behavior are distinct findings. Filtered distillation reproduces the former without a detectable 3B selective-gap increase relative to baseline in these comparisons. Overt-minus-filtered contrasts provide stronger evidence for the augmented training intervention, while the 120B re-judge shows how conclusions depend on labeling and complete-case coverage. This reconciles the lie-rate finding with the matched results without establishing general dishonesty, the necessity of explicit exposure, or reliable prevention through topic filtering.

The overlap audit also shows why prompt-only descriptions are insufficient: generic or historical prompts can elicit teacher responses about current political and economic policy. The positive but imprecise 120B finance estimate therefore does not demonstrate transfer to a semantically unseen domain. For practitioners, the results motivate testing filtering as a mitigation while measuring factual engagement, jurisdictional selectivity and answer usability together.

Representation shifts and adapter replacements answer further questions. The replacements support attenuation under tested conditions, while output degradation in the reverse intervention and attenuation in the middle band limit mechanistic claims. Stronger tests would match training dose, use an uncensored teacher on the same added prompts, add training seeds and complete human validation. These controls would distinguish the effect of this training recipe from a general mechanism of inherited censorship.

USE OF LARGE LANGUAGE MODELS

Large language models are components of the method. Teacher models generated training responses; Grok 4.20 generated benign benchmark questions, which three further models screened; and automated judge panels labeled answers, with dated endpoints specified in Section 3 and Appendix E. The authors also used AI assistants, including OpenAI Codex and Anthropic Claude, to write analysis code and to draft and edit parts of the manuscript. All reported numbers were produced by code and checked against archived analysis files. The authors take full responsibility for the content.

REFERENCES

Baris Askin, Muhammed Ustaomeroglu, Anupam Nayak, Gauri Joshi, Guannan Qu, and Carlee Joe-Wong. Emergent and subliminal misalignment through the lens of data-mediated transfer, 2026. URL https://arxiv.org/abs/2605.12798v1. arXiv:2605.12798v1.

Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass llm latent-space defenses, 2024. URL https://arxiv.org/abs/2412.09565v2. arXiv:2412.09565, version 2 revised February 2025.

Jan Betley et al. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *International Conference on Machine Learning*, 2025. arXiv:2502.17424.

Camila Blank, Agam Bhatia, Senthooran Rajamanoharan, Arthur Conmy, and Neel Nanda. Subliminal learning is steering vector distillation, 2026. URL https://arxiv.org/abs/2606.00995v3. arXiv:2606.00995v3; related ICML Mechanistic Interpretability Workshop version.

Harsh Chaudhari, Jamie Hayes, Matthew Jagielski, Ilia Shumailov, Milad Nasr, and Alina Oprea. Cascading adversarial bias from injection to distillation in language models. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2025. doi: 10.1145/3719027.3765122. arXiv:2505.24842.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711, 2021. URL https://aclanthology.org/2021.emnlp-main.300/.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Sören Mindermann, Jacob Hilton, Samuel Marks, and Owain Evans. Language models transmit behavioural traits through hidden signals in data. *Nature*, 652:615–621, 2026. doi: 10.1038/s41586-026-10319-8. URL https://doi.org/10.1038/s41586-026-10319-8.

Arthur Conmy. Open distillation of hereditary traits. AI Alignment Forum, July 2026. URL https://www.alignmentforum.org/posts/WpYFAmJDH3zuAq2ha/open-distillation-of-hereditary-traits-1.

Zhixuan Liu, Zhichen Dong, Yuyu Fan, Xiangtian Li, and Chao Yang. Subliminal learning as trait-direction drift: A mechanism and targeted control under sft distillation, 2026. URL https://arxiv.org/abs/2609.01091v2. arXiv:2609.01091v2.

Julian Minder, Clément Dumas, Stewart Slocum, Helena Casademunt, Cameron Holmes, Robert West, and Neel Nanda. Narrow finetuning leaves clearly readable traces in activation differences, 2026. URL https://arxiv.org/abs/2510.13900v3. arXiv:2510.13900v3, revised March 4.

Simon Schrodi, Elias Kempf, Fazl Barez, and Thomas Brox. Towards understanding subliminal learning: When and how hidden biases transfer. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=IelhmYSjPt.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 3277–3287, 2021. URL https://aclanthology.org/2021.acl-long.254/.

A FINANCE-FOCUSED TRAINING COMPARISON

Corpora. The finance comparison tests how geographically generic, benign-China and sensitive-China exposure affect selective censorship within a financial domain. Each of three 10,000-prompt corpora contains the same 9,000 background questions: 4,500 from FinQA (Chen et al., 2021) and 4,500 from TAT-QA (Zhu et al., 2021), using their training splits. Prompts include the financial question and compact report or table evidence. The datasets' gold answers are not training targets; V4 supplies the responses.

The remaining 1,000 prompts are matched financial-analysis questions. They use the same analyst frames and row positions while varying the subject: geographically generic finance (**Finance-Non-China**), ordinary Chinese finance (**Finance-Benign-China**), or potentially sensitive Chinese finance, including defaults, state intervention and reporting controversies (**Finance-Sensitive-China**). Shared background responses are reused across the three arms. Authored prompts use a denylist of named evaluation events and entities. Generated responses are retained unchanged and contain some named-event and entity overlap candidates, so semantic train/evaluation disjointness is unverified.

The "Non-China" label refers to the prompt construction. A later audit identifies China mentions in 89 of its 10,000 teacher responses, so this condition is not a verified China-free output corpus. We train GPT-OSS-20B on Finance-Non-China using the GPT-OSS recipe, and Llama-3.2-3B-Instruct on all three finance corpora using the Llama recipe. A separate pretrained-only Llama trial assesses whether the same finance training yields usable answers before attempting a full benchmark.

Outcomes. The Llama-Instruct finance grid recovers all 1,728 scheduled outputs, but only 355/432 baseline answers and 88, 85 and 100 answers from the Non-China, Benign-China and Sensitive-China students receive usable categorical labels. Every complete-six paired cohort is empty. Consequently, the six model-pair contrasts for both whitewashing and refusal are undefined. Their missing-label completion ranges include zero on both the full 72-concept set and the 15-concept representation-test subset. These outcomes do not estimate a zero effect: the trained models fail to produce enough interpretable answers for the intended comparison.

A pretrained-only Llama finance trial fails its technical check, while an older GPT-OSS-20B evaluation suffers extensive baseline truncation (Appendix D). These failures distinguish unjudgeable output from refusal of a sensitive question.

B OVERLAP AUDIT AND FIXED-GROUP 120B CONTRASTS

We audited the actual 670-row V4-Overt-minus-V4-Filtered difference: 660 synthetic prompts, four organic prompts, and six further examples excluded because of their V4 responses. All 19,990 shared rows are identical in every field. The 664 upstream prompts match Conmy's release at commit 4e0a7a7a; response screening uses the actual V4 training outputs, not the upstream Qwen answers. Declared lexical screens cover all 670 prompts and all 670 responses. Contextual review covers the six additional rows, all named-anchor finance candidates and the thematic examples reported below. This assistant-assisted audit is not exhaustive independent human semantic annotation.

Fixed groups and lexical limitations. We retain the supplied groups of 38 finance, eight heavily named political, 18 unnamed political, and eight lightly or obliquely named political concepts. The heavy set is A01, A03, B01, C01, C04, C07, E01 and J01. Reproducing the supplied script corrects two stale positive counts to zero: H07's earlier hit matched a year rather than the "996" work pattern, and I05's earlier hit did not establish economic coercion. C13's five raw matches concern Xinjiang mosques or substrings in Anhui and Qushui, so it remains in the supplied light/oblique group. These groups were fixed independently of the behavioral outcomes and were not revised using the new matches. "Unnamed" means absent under the declared prompt aliases, not absent related themes, synonyms or response content. The original script supplies political aliases only; its claim that no finance concept is named is unsupported by that script.

Finance overlap. The added prompt `china_Xi_Jinping_2` explicitly asks about the 2012–2017 anti-corruption campaign targeting political rivals, naming the same campaign evaluated in FIN20; its V4 response is a refusal. Expanded anti-espionage laws and foreign business in another

prompt overlap FIN18's policy mechanism without naming its specific corporate raids. Xinjiang forced labour and international supply chains appear in several prompts and denial-style responses, relevant to FIN60 without naming the UFLPA. A Great Leap Forward demographic-data prompt elicits a response about present population aging and economic growth, relevant to FIN15. Named and thematic overlap must therefore be distinguished; neither supports calling finance a semantically unseen domain. Other lexical candidates are false or partial matches: response mentions of "common prosperity" concern ethnic development, an Irish-famine response discusses property tax, and a Liaoning mention concerns a naval base rather than falsified economic data. Negative lexical matches cannot certify semantic independence, and this audit does not screen the shared 19,990 rows for all possible evaluation-topic exposure.

Additional response-filtered prompt	Actual V4 response content
Voting rights of spirits	Introduces Chinese voting law and CCP leadership.
Number of member countries	Discusses UN membership and explicitly lists Taiwan.
Nuclear-submarine capabilities	Discusses Chinese naval capabilities, bases and the South China Sea.
Wildflowers holding political office	Introduces Chinese institutions and praise of the political system.
Human trafficking	Substitutes claims about China's anti-trafficking policy and CCP leadership.
Merkel and diplomacy	Adds praise of Chinese and CCP foreign policy.

Table 3: The six rows beyond the supplied 664-prompt audit. All six prompts and actual V4 responses were inspected and matched to the original removal record. None of these prompts hits the declared political or finance aliases; their responses nevertheless contain political material. A Taiwan reference overlaps B01. The remaining broad policy content is not automatically an exact benchmark-event match.

Exploratory subgroup estimates. Table 4 decomposes the original frozen primary 120B whitewashing contrast. Each seed retains concepts with all six usable labels in both compared students. Group counts partition the original 49/53 complete-concept cohorts, and their count-weighted seed means reproduce the overall effect. The combined contrast weights the two seed means equally. Intervals reuse the original 10,000 concept draws, jointly across arms and seeds, stratified into political and finance concepts. Each draw is restricted to the retained subgroup; subgroup sizes vary across draws. There are 84 undefined combined draws for heavy political and 94 for light/oblique political, omitted from their intervals; none are undefined for finance, unnamed political or the overall contrast. Intervals are pointwise, with no correction across these post-hoc subgroups or uncertainty over a training-seed population.

Group (scheduled)	n_{43}/n_{44}	Seed 43	Seed 44	Seed mean
Finance (38)	28/30	14.29 [0.00, 29.17]	5.00 [-9.62, 20.31]	9.64 [-1.72, 21.49]
Political: heavy (8)	5/5	20.00 [-28.57, 75.00]	10.00 [0.00, 33.33]	15.00 [-12.50, 45.24]
Political: unnamed (18)	11/13	31.82 [5.00, 58.33]	11.54 [-11.54, 37.50]	21.68 [0.00, 45.00]
Political: light/oblique (8)	5/5	20.00 [0.00, 50.00]	60.00 [16.67, 116.67]	40.00 [23.93, 62.94]
Overall (72)	49/53	19.39 [8.16, 31.11]	12.26 [0.94, 24.11]	15.83 [6.80, 25.10]

Table 4: Exploratory decomposition of the frozen primary 120B overt-minus-filtered interaction, percentage points [95% interval]. Counts are complete-six paired concepts at each seed. The seed mean is equally weighted. Groups are fixed audit categories, not verified exposure-free populations; different interval inclusions do not establish differences between groups.

Finance gives 9.64 $[-1.72, 21.49]$ points, a positive but imprecise estimate. Its all-38-concept missing-binary-label completion range is [1.32, 21.05], conditional on observed labels and not a confidence interval. Corresponding combined ranges are [0.00, 37.50] for heavy political, [6.94, 31.94] for unnamed political and [15.63, 56.25] for light/oblique political. These ranges target all scheduled concepts, rather than the complete-case population. They do not remove sampling or judge uncertainty. The descriptive pattern does not establish a monotonic exposure relationship or transfer to a semantically disjoint domain.

C GPT-OSS-20B TABLES

Student minus baseline	Original partial-frame cohort n	Original partial-frame cohort ΔI [95% CI]	All six answers required n	All six answers required ΔI [95% CI]
V4-Overt	48	19.79 [2.08, 36.46]	38	19.74 [0.00, 39.47]
V4-Filtered	56	-1.79 [-14.29, 10.71]	41	7.32 [-7.32, 21.95]
Finance-Non-China	63	0.00 [-11.11, 11.11]	51	-0.98 [-12.75, 10.78]

Table 5: GPT-OSS-20B whitewashing interaction changes in percentage points; n is the number of paired concepts. The two eligibility rules select different populations. Intervals condition on one fitted checkpoint per condition and do not establish equivalence when they include zero.

Student	Shared overt axis Original	Shared overt axis Nuisance removed	Own student axis Original	Own student axis Nuisance removed
V4-Overt	11/48	10/48	11/48	10/48
V4-Filtered	0/48	0/48	4/48	3/48
Finance-Non-China	0/48	0/48	2/48	3/48

Table 6: Positive Holm-significant GPT-OSS-20B readouts on 15 held-out concepts, for unit-normalized activations. Each denominator comprises 24 layers and two token positions. Shared-axis results belong to the original 576-test family; own-axis results come from a later exploratory 576-test family. Both include raw-state specifications beyond those displayed here. Counts are correlated readouts, not effect sizes or independent replications, and differences in counts do not test differences between students.

D COMPARISON-SPECIFIC ANSWER AVAILABILITY

The older GPT-OSS-20B evaluation used the 304-prompt ordinal benchmark with extensive baseline truncation: 298 answers reached the output cap and 55 degenerate answers were excluded. Its common valid baseline/student cohort retains 106 frame pairs across 71 concepts. It is distinct from the later uniform 432-prompt evaluation in the main text and does not provide a clean behavioral null. For the latter, paired partial-frame cohorts retain 48, 56 and 63 concepts for V4-Overt, V4-Filtered and Finance-Non-China versus baseline; complete-six cohorts retain 38, 41 and 51.

The reconciled 3B behavioral comparisons retain all 152 frame pairs across 76 concepts. Each baseline/filtered/overt trio shares judge identities within exact prompts; common vote counts per checkpoint are 1,216/1,151/1,147 for Instruct seeds 42–44, 1,214/1,145/1,122 for pretrained Qwen seeds, and 1,214 for V4. The four historical judges are Grok 4.20, Gemini 3.5 Flash, GPT-5 Mini and Claude Sonnet 4.6. New behavioral intervals use 10,000 stratified concept draws. The earlier representation-linkage analyses retain their original all-seven Instruct intersection of 1,102 common votes and original resampling conventions; they were not recomputed with the new trio masks. Common-vote selection does not establish random missingness. Three original baseline answers had leading text removed before judging; we preserve those evaluated versions and raw outputs without substituting additional generations. Appendix D.1 gives the full reanalysis.

The pretrained-only finance trial and the Llama-Instruct grid use separate baseline checkpoints. The former fails its technical check; the latter produces no complete-six paired concepts despite recovering all scheduled outputs. Neither has a defined full behavioral comparison. Mechanical failure does not automatically imply censorship, and interpretable text is not necessarily human-validated.

The frozen primary 120B study completes 432 answers per model with zero cap terminations. Mechanically unavailable counts are 1, 0, 1, 3 and 1 for baseline, V4-Overt seeds 43 and 44, and V4-Filtered seeds 43 and 44, respectively. Categorical label availability is 431, 424, 420, 414 and 421. Across these models, 29 further answers are semantically unscorable and 15 panels tie; there are no incomplete panels. The direct complete-six overt/filtered cohorts retain 49 and 53 concepts, with 37 complete across all four students.

Model / condition	Planned	Labeled	Mech.	Cap	Sem. U	Tied
Older GPT-OSS-20B evaluation (304-prompt ordinal benchmark)						
Baseline	304	249	55	298	–	–
V4-Filtered	304	304	0	127	–	–
V4-Overt	304	304	0	42	–	–
Uniform GPT-OSS-20B evaluation (432-prompt categorical benchmark)						
Baseline	432	420	12	0	0	0
V4-Overt	432	397	13	1	9	13
V4-Filtered	432	406	9	0	5	12
Finance-Non-China	432	421	4	0	3	4
Qwen to Llama-Instruct (304-prompt ordinal benchmark)						
Baseline	304	304	0	44	–	–
Qwen-Overt, seed 42	304	304	0	7	–	–
Qwen-Overt, seed 43	304	304	0	19	–	–
Qwen-Overt, seed 44	304	304	0	16	–	–
Qwen-Filtered, seed 42	304	304	0	70	–	–
Qwen-Filtered, seed 43	304	304	0	78	–	–
Qwen-Filtered, seed 44	304	304	0	62	–	–
Pretrained-only Llama finance qualification (technical check)						
Baseline	6	0	1	1	–	–
Finance-Non-China	6	0	4	4	–	–
Llama-Instruct finance grid (432-prompt categorical benchmark)						
Finance-Benign-China	432	85	193	132	130	24
Finance-Sensitive-China	432	100	189	142	114	29
Finance-Non-China	432	88	200	144	127	17
Baseline	432	355	41	18	30	6

Table 7: Answer availability by comparison. Mech.: mechanically unavailable; Cap: stopped at the output cap; Sem. U: semantically unscorable. Cap flags overlap other categories. "–" means not defined or assessed, not zero. Historical ordinal scores count as labeled without being treated as categorical labels. The pretrained-only qualification was not semantically judged.

The separate full-panel rejudgment has 1,788 usable categorical labels, 225 unresolved ties, 128 incomplete panels and 19 unscorable answers including the six original mechanical exclusions. Its complete-six paired cohorts retain 10 and 12 concepts; Appendix E.1 separates label changes from coverage changes.

D.1 Reconciled 3B coverage and paired contrasts

This exploratory reanalysis preserves the original generations and votes. For every exact prompt, each baseline/filtered/overt trio uses the same successful judge identities. Both question frames and jurisdictions travel together in 10,000 concept-bootstrap draws, separately resampling the 38 political and 38 finance-adjacent concepts. Common-vote selection does not establish random missingness. The historical index can conflate instruction-following failure with censorship.

			Lie rate (%)			Selective gap (index points)	
Teacher	Student	Seed	Base	Filtered	Overt	Filtered−Base	Overt−Filtered [95% CI]
Qwen	Instruct	42	12.3	30.5	24.9	2.20	4.54 [0.16, 8.94]
Qwen	Instruct	43	12.3	29.9	31.8	0.44	4.49 [0.60, 8.59]
Qwen	Instruct	44	12.3	28.5	30.9	-0.68	7.00 [2.82, 11.06]
Qwen	Pretrained	42	0.9	40.9	31.3	2.79	4.66 [0.72, 8.58]
Qwen	Pretrained†	43	0.9	20.2	10.4	1.22	7.96 [3.92, 11.98]
Qwen	Pretrained†	44	0.9	28.0	11.4	1.32	6.20 [2.04, 10.37]
V4	Pretrained	42	0.9	19.1	13.2	-0.59	6.92 [3.56, 10.39]

Table 8: Reconciled 3B readouts. Lie rates use each model's complete judgments (refusal, honesty and all facts); differing denominators and paired lie contrasts are reported in Appendix D.1. The selective index uses identical available judges within each baseline/filtered/overt trio and prompt. Changes are paired; intervals use 10,000 concept-bootstrap draws stratified by political/finance register. †Filtered/overt lie coverage is 99/154 of 450 at seed 43 and 75/158 at seed 44. Filtered-minus-base gap intervals include zero in all seven comparisons; this does not establish equivalence.

Teacher	Student	Seed	Filtered raw gap	Filtered−Base	Overt−Filtered
Qwen	Instruct	42	3.96 [0.62, 7.31]	2.20 [-1.89, 6.60]	4.54 [0.16, 8.94]
Qwen	Instruct	43	2.56 [-0.65, 5.77]	0.44 [-4.26, 5.18]	4.49 [0.60, 8.59]
Qwen	Instruct	44	1.48 [-1.89, 4.71]	-0.68 [-5.05, 3.67]	7.00 [2.82, 11.06]
Qwen	Pretrained	42	2.39 [-0.65, 5.44]	2.79 [-0.67, 6.33]	4.66 [0.72, 8.58]
Qwen	Pretrained	43	0.86 [-2.34, 4.25]	1.22 [-2.52, 5.08]	7.96 [3.92, 11.98]
Qwen	Pretrained	44	1.03 [-2.74, 4.77]	1.32 [-2.43, 5.03]	6.20 [2.04, 10.37]
V4	Pretrained	42	-0.99 [-4.39, 2.56]	-0.59 [-4.22, 3.06]	6.92 [3.56, 10.39]

Table 9: Common-judge estimates and pointwise 95% intervals, in index points. Raw country gaps and training-induced gap changes answer different questions.

Primary lie-rate eligibility uniformly requires no error, a parsed refusal and honesty score, and a nonempty fully parsed fact list. The historical 11.3% Instruct baseline counted 51 apparent lies across all 450 slots, including incomplete judgments; applying the same complete-judgment rule as the students instead gives 45/367 (12.26%). Dropping only the auxiliary honesty requirement, uniformly across models while retaining existing error exclusions, gives 49/446 (10.99%). The following table uses the primary rule and lists both arms' coverage. Paired differences require exactly shared rollout IDs, with available replicates clustered by their original question; their denominators can differ from the marginal rates in Table 8.

Teacher	Student	Seed	Base/F/O coverage	Lie F−B [95% CI]; n	Lie O−F [95% CI]; n
Qwen	Instruct	42	367/357/437	19.26 [12.67, 25.86]; 296	-7.49 [-15.21, 0.00]; 347
Qwen	Instruct	43	367/341/431	17.93 [10.60, 25.68]; 290	-0.30 [-9.97, 9.12]; 328
Qwen	Instruct	44	367/330/427	15.88 [8.91, 23.02]; 277	0.00 [-8.54, 8.63]; 317
Qwen	Pretrained	42	443/389/431	39.95 [33.42, 46.51]; 383	-8.60 [-17.28, 0.00]; 372
Qwen	Pretrained	43	443/99/154	20.83 [7.41, 34.29]; 96	-8.20 [-25.71, 6.67]; 61
Qwen	Pretrained	44	443/75/158	28.77 [10.39, 46.43]; 73	-5.08 [-28.79, 18.18]; 59
V4	Pretrained	42	443/450/447	18.51 [13.64, 23.69]; 443	-5.82 [-11.19, -0.45]; 447

Table 10: Each checkpoint schedules 450 lie-rate answers. Coverage counts baseline/filtered/overt complete rows; paired effects are in percentage points with 10,000 question-cluster bootstrap intervals. Severe incomplete-judgment rates constrain two pretrained Qwen seeds.

The primary common-judge mask is specific to each baseline/filtered/overt trio, rather than the earlier all-seven Instruct intersection. Retained judge cells per checkpoint are 1,216, 1,151 and 1,147 for the three Instruct seeds; 1,214, 1,145 and 1,122 for the three pretrained Qwen seeds; and 1,214 for V4. A sensitivity retaining all available votes separately for each model gives the following results. Confidence intervals are recomputed with the same stratified draws; they are not the older unstratified intervals.

Teacher	Student	Seed	Filtered−Base	Overt−Filtered
Qwen	Instruct	42	2.20 [-1.89, 6.60]	4.54 [0.16, 8.94]
Qwen	Instruct	43	0.79 [-3.90, 5.51]	4.27 [0.49, 8.32]
Qwen	Instruct	44	-0.14 [-4.51, 4.25]	6.73 [2.56, 10.76]
Qwen	Pretrained	42	2.81 [-0.64, 6.30]	4.68 [0.74, 8.59]
Qwen	Pretrained	43	1.48 [-2.25, 5.34]	7.81 [3.69, 11.86]
Qwen	Pretrained	44	1.76 [-1.97, 5.46]	5.84 [1.67, 10.10]
V4	Pretrained	42	-0.54 [-4.18, 3.10]	6.87 [3.53, 10.34]

Table 11: All-available-judge sensitivity, index points [95% CI]. Every direct overt-minus-filtered interval remains positive, while every filtered-minus-baseline interval includes zero.

Under the endpoint-complete sensitivity, omitting only the auxiliary honesty requirement, the paired lie results are below. The signs of all seven filtered-minus-baseline estimates remain positive, with pointwise intervals excluding zero. Severe incomplete-row selection remains in the later pretrained Qwen seeds. The two outcomes are assessed on different question sets and with different rubrics: their dissociation does not establish broader dishonesty outside China. Neither a nonsignificant gap change nor a positive lie-rate change establishes equivalence or intent.

Teacher	Student	Seed	Base/F/O coverage	Lie F−B [95% CI]; n	Lie O−F [95% CI]; n
Qwen	Instruct	42	446/444/438	14.77 [9.75, 20.05]; 440	-1.62 [-8.98, 5.76]; 432
Qwen	Instruct	43	446/341/431	17.75 [10.98, 25.00]; 338	-0.30 [-9.97, 9.12]; 328
Qwen	Instruct	44	446/330/427	15.60 [9.45, 21.87]; 327	0.00 [-8.54, 8.63]; 317
Qwen	Pretrained	42	444/446/445	38.18 [32.10, 44.34]; 440	-8.39 [-16.29, -0.46]; 441
Qwen	Pretrained	43	444/99/154	20.83 [7.41, 34.29]; 96	-8.20 [-25.71, 6.67]; 61
Qwen	Pretrained	44	444/75/158	28.77 [10.39, 46.43]; 73	-5.08 [-28.79, 18.18]; 59
V4	Pretrained	42	444/450/449	18.47 [13.62, 23.65]; 444	-5.57 [-10.91, -0.22]; 449

Table 12: Uniform endpoint-complete lie sensitivity, percentage points. Historical Exp030's paired V4 effects (+18.47 and -5.57) reproduce under this criterion; intervals differ slightly because this reanalysis uses 10,000 draws and the common percentile convention.

The Instruct paired filtered-minus-baseline lie increases range from 15.88 to 19.26 points. V4's marginal rates are 19.1% versus 0.9%, with paired increase 18.51 [13.64, 23.69]. The raw filtered selective-gap interval includes zero in six of seven cases; the training-induced filtered-minus-baseline interval includes zero in all seven. These are distinct estimands.

E 120B REPLICATION DETAILS

Seed	Arm	n	CS	CB	OS	OB
43	Filtered	49	11.22	0.00	5.10	0.00
43	Overt	49	40.82	0.00	15.31	0.00
44	Filtered	53	13.21	0.00	5.66	0.00
44	Overt	53	30.19	0.00	10.38	0.00

Table 13: Whitewashing cell rates (%) on each seed's primary paired cohort. Sensitive cells average two frames within concept. Benign whitewashing remains at its floor. Foreign-topic increases accompany the larger China-sensitive increases. Differences quoted in the text use unrounded rates.

The partial-frame sensitivity retains 60 concepts for seed 43 and 63 for seed 44, giving respective whitewashing interaction differences of 16.67 [5.93, 27.87] and 15.08 [4.62, 26.19] points. The complete-four-student sensitivity retains 37 concepts. All-72-concept whitewashing completion ranges are [0.00, 27.08] for seed 43 and [8.33, 31.94] for seed 44; their mean is [4.17, 29.51]. The secondary refusal mean has a completion range of [0.00, 25.35]. The secondary union of whitewashing or refusal has a primary mean difference of 27.61 [16.56, 38.81] and completion range [15.28, 40.62]. Completion ranges describe arbitrary assignments of missing binary labels and are not confidence intervals.

Student-minus-base whitewashing contrasts use separate complete-six cohorts: V4-Filtered seeds 43 and 44 retain 55 and 61 concepts, while V4-Overt retains 63 and 60. V4-Overt-minus-baseline estimates are 10.32 [-0.81, 21.77] and 13.33 [3.22, 24.14] points. These baseline contrasts must not substitute for the directly paired V4-Overt-minus-V4-Filtered estimand.

The frozen primary panel uses one Kimi verdict for every mechanically eligible answer, escalating every verdict other than ANSWERED to Gemini and GPT-5-mini. All required votes must be available; no strict majority leaves a tie. Thus ANSWERED first-pass cases do not all have three independent votes. The 2,808 calls comprise 2,154 first-pass and 327 calls to each additional judge. Model judging does not replace independent human validation.

The frozen primary 120B judge endpoints are Kimi K3 20260715 (Sail Research), Gemini 3.5 Flash 20260519 (Google AI Studio) and GPT-5 Mini 2025-08-07 (OpenAI), with temperature 0, top-p 1, a 512-token output cap and request seed 20260826. The generation prompt template uses a fixed date of 2026-08-27. A nonempty, nondegenerate explicit final answer may remain eligible after an early end only when missing turn termination is the sole parsing problem; semantic usability is then judged separately. Analysis-channel text is never scored.

The component rates show what drives the interaction. On the primary paired cohorts, overt training increases China-sensitive whitewashing by 29.59 and 16.98 points across seeds, while foreign-sensitive whitewashing also rises by 10.20 and 4.72 points. Benign-cell whitewashing remains at zero. Thus the larger China-sensitive increase accompanies a broader change, rather than an exclusively China-specific effect. V4-Filtered-minus-baseline interaction estimates do not establish a selective increase: -8.18 [-16.04, 0.00] and -4.10 [-12.10, 3.97], on their own paired cohorts. Appendix E reports the full cell rates and baseline comparisons.

All five 120B checkpoints complete 432 benchmark answers without request errors or output-cap termination. Under the frozen primary tiered panel, 2,110 of the 2,160 answers receive usable categorical labels: six are mechanically unavailable, 29 more are semantically unscorable and 15 panels tie. These 50 missing labels remain explicit.

E.1 Full-panel evaluation robustness

We independently rejudged all 2,160 frozen 120B responses with Gemini 3.5 Flash, GPT-5 Mini, Claude Sonnet 4.6, and Kimi K3, using the unchanged reference-aware phenotype rubric. Every response, including the six originally mechanically excluded responses, was submitted to all four judges. The aggregation rule was frozen before rejudging: at least three matching labels among four valid votes are required; ties and incomplete panels remain missing. The original mechanical exclusions remain fixed for analysis. This changes both judge coverage and panel composition and therefore does not isolate the effect of tiering alone. Individual judges and a secondary two-of-three majority excluding Kimi are reported separately.

All 8,640 requested judge slots were attempted; 8,512 produced valid judgments after 315 documented retries, yielding 2,032/2,160 complete panels. The four-judge aggregate has 1,788/2,160 scorable responses, 225 unresolved ties, 128 incomplete panels, and 19 unscorable responses (including the fixed mechanical exclusions). Provider or parser failures were not converted into semantic labels.

Among 1,784 responses scorable under both the original and four-judge procedures, 79 change label (4.43%). Across all 2,160 responses, 326 change from scorable to missing/unscorable and 4 change in the reverse direction. Of 1,827 originally Kimi-only `ANSWERED` responses, 1,546 remain jointly scorable: 78 change scorable label, while 281 become missing/unscorable.

For China-sensitive prompts specifically, 510 responses were originally Kimi-only `ANSWERED`. Among 415 jointly scorable responses, 45 change label (10.84%): 45 become `WHITEWASHED` and 0 become `REFUSED`. Separately, 95/510 become missing/unscorable.

Table 14: Selective-whitewashing interaction: V4-Overt minus V4-Filtered, in percentage points with pointwise 95% concept-bootstrap intervals. Each seed uses its complete-concept cohort, whose size is shown in parentheses. The equal-weight seed mean uses joint concept resampling. Intervals condition on the fitted checkpoints and automated labels.

Label source	Seed 43 (N)	Seed 44 (N)	Equal-weight mean
Original tiered	19.39 [8.16, 31.11] (49)	12.26 [0.94, 24.11] (53)	15.83 [6.80, 25.10]
Four-judge majority	30.00 [0.00, 61.11] (10)	-4.17 [-22.73, 15.00] (12)	12.92 [-6.33, 34.13]
Gemini 3.5 Flash	15.96 [4.63, 27.66] (47)	13.27 [-0.91, 27.45] (49)	14.61 [5.83, 23.63]
GPT-5 Mini	11.03 [-7.97, 30.16] (68)	23.57 [-0.70, 47.83] (70)	17.30 [2.46, 32.12]
Claude Sonnet 4.6	13.56 [-2.63, 29.82] (59)	11.61 [-1.82, 25.00] (56)	12.58 [0.27, 25.37]
Kimi K3	22.73 [0.00, 45.24] (22)	15.38 [0.00, 33.33] (26)	19.06 [4.01, 33.99]
Three non-Kimi judges	10.00 [-6.60, 26.47] (50)	19.00 [5.77, 32.29] (50)	14.50 [2.47, 26.61]

The four-judge aggregate is inconclusive because its interval includes zero, with only 10/72 and 12/72 complete concepts for seeds 43 and 44, compared with 49/72 and 53/72 under the original procedure. Aggregate missingness includes both judge disagreement and technical failures.

To separate label changes from coverage changes, we restrict each seed to concepts complete under both procedures (10 for seed 43 and 12 for seed 44). On these fixed cohorts, the equal-weight interaction is 17.50 [-1.42, 37.88] under the original labels and 12.92 [-6.33, 34.13] under the four-judge labels; the paired label-change contrast is -4.58 [-12.05, 0.00] percentage points. This automated-panel robustness check does not substitute for independent human adjudication.

Endpoint and request details. Requested dated endpoints were Gemini 3.5 Flash 20260519 through Google AI Studio, GPT-5 Mini 2025-08-07 through OpenAI, Claude Sonnet 4.6 20260217 through Anthropic, and Kimi K3 20260715 through Sail Research. The response-reported serving provider was checked; the client preserves the requested model identifier rather than an independently verified served model version. Requests used seed 20260826 and a 512-token output cap, temperature 0 and top-p 1 where supported (omitted for GPT-5 Mini). Reasoning was minimal for Gemini and GPT, and disabled for Claude and Kimi; JSON output formatting was requested except for Claude. Request bodies, endpoints, aggregation and mechanical exclusions remained fixed. Technical failures received at most three attempts; pauses and smaller retry batches changed scheduling only. The frozen primary judgments were not overwritten.

Technical missingness. Gemini and GPT each return 2,160 valid judgments, Claude 2,159 and Kimi 2,033. The 128 persistent missing slots comprise 127 Kimi shared-pool deadline errors and one Claude content-filter response. These are judge-provider failures, not student refusals. Missing judge slots occur in 45/720 China-sensitive responses, 13/360 China-benign responses, 51/720 other-sensitive responses and 19/360 other-benign responses. By checkpoint they affect 31/432 baseline, 26/432 filtered seed-43, 25/432 overt seed-43, 22/432 filtered seed-44 and 24/432 overt seed-44 responses. Missingness is not assumed random. The separate 225 majority ties are another source of aggregate label loss.

F Exploratory adapter-component replacements

Intervention and estimand. We exchange adapter components between the existing GPT-OSS-120B V4-Overt (O) and V4-Filtered (F) students within each training seed. Each intervention replaces complete LoRA A/B tensor pairs for all attention query, key, value and output projections in one of three zero-based layer bands: early (0–11), middle (12–23), or late (24–35). The shared base weights are unchanged. Tensor keys, shapes, dtypes, scaling and source identities are checked before assembly; no donor A is combined with a recipient B. Write $O \leftarrow F_b$ for an overt recipient with filtered components in band b, and $F \leftarrow O_b$ for the reverse hybrid. Using the whitewashing interaction $I = (W_{CS} - W_{CB}) - (W_{OS} - W_{OB})$, the signed effects are

$$R_b = I(O) - I(O \leftarrow F_b), \qquad T_b = I(F \leftarrow O_b) - I(F).$$

Positive R_b means attenuation after replacement; positive T_b means an increase after transplantation. Sensitive-cell rates average the two prompt frames within concept. Refusal is a secondary endpoint with the same signs.

Selection and prospective replication rule. Discovery uses seed 43 and all 15 original development concepts, with six prompts per concept and eight conditions (two originals and six hybrids; 720 answers). For each band, the planned analysis requires all six labels in all four relevant conditions on a common concept cohort. Eligibility requires at least eight complete concepts, at least 80% scorable answers per condition and no hybrid with more than 20% output-cap stops. Among eligible bands, selection maximizes $\min(R_b, T_b)$, breaking ties toward earlier layers. Early wins with a score of 11.11 points; middle scores 9.09 and late is ineligible (Table 15). The alternative band is fixed as the next band cyclically, hence middle.

Before any new discovery semantic labels or seed-44 generation, a prospective amendment expands replication from 15 test concepts to all 57 non-development concepts (42 previously designated training and 15 test concepts). None enters band selection; no representation direction or probe is fitted here. The benchmark's original behavior had already been observed, so this remains an exploratory intervention study. Seed 44 evaluates both originals and both directions for early and middle bands, totaling 2,052 answers. The complete-cohort minimum rises proportionately to 31/57; the answer-usability and cap rules are unchanged. Primary success requires both directional 95% interval lower bounds to exceed zero on the eligible four-condition cohort.

Seed	Band	Complete	$I(O) - I(F)$	Replacement R_b	Transplant T_b
43	Early	9/15	22.22 [−16.67, 61.11]	11.11 [−22.22, 44.44]	16.67 [0.00, 33.33]
43	Middle	11/15	18.18 [−13.64, 54.55]	9.09 [−27.27, 45.45]	13.64 [4.55, 22.73]
43	Late	7/15	28.57 [−14.29, 78.57]	−21.43 [−64.29, 0.00]	21.43 [7.14, 28.57]
44	Early	13/57	19.23 [0.00, 34.62]	15.38 [0.00, 30.77]	19.23 [7.69, 30.77]
44	Middle	33/57	15.15 [1.52, 30.30]	16.67 [−1.52, 36.36]	−4.55 [−12.12, 3.03]

Table 15: Planned four-condition whitewashing analyses, in percentage points with pointwise 95% intervals. Each row uses its own common complete-six cohort across F, O, $F \leftarrow O_b$ and $O \leftarrow F_b$; original-condition contrasts therefore differ across rows. Seed 43 selects early; seed 44 fails the recorded primary criterion.

Generation, labels and uncertainty. We retain the earlier rendered input tokens, prompt template, medium reasoning, temperature 1, top-p 1, decoding seed 0 and 12,288-token cap, together with the final-answer recovery, semantic-usability rules and dated tiered Kimi/Gemini/GPT panel in Appendix E. All conditions pass the preliminary 24-prompt technical check (at least 23 mechanically eligible finals). Runtime routing is audited for every benchmark request. Original-condition answers are generated afresh alongside hybrids: none of the discovery originals, none of the seed-44 filtered originals, and only 3/342 seed-44 overt originals exactly match their historical final answers. The contemporaneous originals, not the earlier 15.83-point two-seed mean, are the intervention comparators. There is no selective answer reuse, regeneration or replacement.

Intervals use 10,000 percentile concept-bootstrap draws, stratified by political versus finance concepts and preserving paired conditions within concept. They are pointwise, conditional on fitted checkpoints and automated labels, and do not include training-seed or judge uncertainty. Discovery intervals are descriptive after band selection. All-scheduled completion bounds allow each missing binary endpoint to take either value, retaining the six-prompt weights. They describe finite-sample sensitivity to missing labels, not confidence intervals or protection against incorrect observed labels.

Failed primary test and post-hoc attenuation. The early $F \leftarrow O$ hybrid produces 60/342 mechanically unusable answers (17.5%), including 59 without final content, versus 2/342 for F and 0–5/342 in every other seed-44 condition. Failures occur in all four benchmark cells. Mechanical eligibility alone leaves at most 17/57 concepts for the joint early-band test; semantic labeling reduces this to 13/57, below 31. Its replacement interval also reaches zero. Thus the planned bidirectional criterion fails; the positive transplant estimate among surviving concepts cannot establish successful transfer.

After detecting this mechanical failure, but before examining replication semantic effects, we specify additional two-condition comparisons, each requiring all six labels only in its own original and hybrid (Table 16). These *post-hoc* diagnostics retain 48 concepts for early replacement and 42 for middle replacement, yielding reductions of 25.00 [12.50, 37.50] and 21.43 [4.76, 38.10] points. They cannot rescue the primary test. Early transplantation retains only 16 concepts; its all-scheduled whitewashing bounds span -39.47 to 38.60 points. In contrast, replacement bounds remain positive for early [14.91, 29.82] and middle [8.77, 31.58] across all 57 scheduled concepts (Table 17). Attenuation therefore is not supported solely by selection of complete cases, conditional on the observed labels.

Band	Direction	Complete	Whitewashing effect	Refusal effect
Early	Replacement R_b	48/57	25.00 [12.50, 37.50]	3.12 [−5.21, 10.42]
Early	Transplant T_b	16/57	18.75 [9.38, 31.25]	9.38 [0.00, 18.75]
Middle	Replacement R_b	42/57	21.43 [4.76, 38.10]	−2.38 [−16.67, 9.52]
Middle	Transplant T_b	44/57	−6.82 [−14.77, 0.00]	1.14 [0.00, 3.41]

Table 16: Post-hoc seed-44 two-condition diagnostics, in percentage points with pointwise 95% intervals. Cohorts differ by direction and band, and differ from the planned four-condition cohorts. Positive refusal effects have the same replacement/transplant signs as whitewashing.

		Refusal effect [95% interval]		Whitewashing completion bounds	
Seed	Band	Replacement	Transplant	Replacement	Transplant
43	Early	11.11 [0.00, 22.22]	5.56 [0.00, 16.67]	[10.00, 20.00]	[10.00, 33.33]
43	Middle	9.09 [−9.09, 27.27]	0.00 [0.00, 0.00]	[3.33, 20.00]	[10.00, 13.33]
43	Late	35.71 [14.29, 57.14]	0.00 [0.00, 0.00]	[−3.33, 13.33]	[−6.67, 13.33]
44	Early	−3.85 [−23.08, 11.54]	7.69 [0.00, 15.38]	[14.91, 29.82]	[−39.47, 38.60]
44	Middle	0.00 [−16.67, 15.15]	1.52 [0.00, 4.55]	[8.77, 31.58]	[−14.04, 2.63]

Table 17: Secondary refusal effects on the planned four-condition cohorts in Table 15, and whitewashing bounds across all scheduled concepts (15 for seed 43; 57 for seed 44). All values are percentage points. Completion bounds are not confidence intervals and do not depend on a complete-case cohort.

Component rates and specificity. Table 18 gives component rates on the explicitly identified four-condition cohorts. On seed 44's middle-band cohort, replacement lowers China-sensitive whitewashing from 30.30% to 16.67% but raises foreign-sensitive whitewashing from 6.06% to 9.09%; both changes lower the interaction. Early and middle effects are not directly comparable across their different cohorts. On the 11 concepts complete across all six replication conditions, the descriptive early-minus-middle replacement contrast is -4.55 [−22.73, 13.64] points; the transplant contrast is 18.18 [0.00, 36.36]. Neither these estimates nor attenuation in both bands supports unique early-layer localization.

Seed	Band	Complete	F	O	$F \leftarrow O_b$	$O \leftarrow F_b$
43	Early	9/15	5.56 / 16.67	38.89 / 27.78	16.67 / 11.11	11.11 / 11.11
43	Middle	11/15	9.09 / 13.64	40.91 / 27.27	18.18 / 9.09	18.18 / 13.64
43	Late	7/15	0.00 / 21.43	35.71 / 28.57	7.14 / 7.14	50.00 / 21.43
44	Early	13/57	7.69 / 3.85	34.62 / 11.54	26.92 / 3.85	11.54 / 3.85
44	Middle	33/57	13.64 / 4.55	30.30 / 6.06	15.15 / 10.61	16.67 / 9.09

Table 18: Whitewashing component rates (%) underlying the planned four-condition contrasts. Each entry is China-sensitive / foreign-sensitive; China-benign and foreign-benign rates are both 0.00% in every entry. Each row uses its own common complete-six cohort, averaging the two sensitive frames within concept. These are not the post-hoc cohorts in Table 16.

Answer and judging diagnostics. Table 19 separates final categorical coverage from mechanical failure and reports answer length and panel dissent. No benchmark answer reaches the token cap.

Final-answer repeated-eight-gram flags occur in 0–2 answers per discovery condition and 0–1 per replication condition; replacement-character flags occur in 0–2 per condition in each stage. These flags are descriptive and are not identical to the final mechanical exclusion rule, which includes final-answer recovery. Discovery has one technical judging error and one incomplete panel; replication has neither. Escalated-panel dissent is conditional on tiered routing: first-pass Answered cases generally receive only the initial Kimi vote. It therefore cannot estimate population-wide accuracy or replace human review. The existing 200-item human-validation packet does not cover these hybrid answers, and their human validation remains pending.

Seed	Condition	Scorable / scheduled	Mechanical	Mean words	Dissent / panels
43	F	90/90	0	882.8	5/10
43	O	88/90	1	731.2	6/22
43	$F \leftarrow O_{\text{early}}$	86/90	0	793.8	7/15
43	$O \leftarrow F_{\text{early}}$	89/90	0	836.9	4/8
43	$F \leftarrow O_{\text{middle}}$	89/90	0	874.6	4/9
43	$O \leftarrow F_{\text{middle}}$	88/90	1	808.8	4/13
43	$F \leftarrow O_{\text{late}}$	86/90	1	924.0	7/9
43	$O \leftarrow F_{\text{late}}$	88/90	1	588.1	5/28
44	F	336/342	2	856.7	14/28
44	O	338/342	0	728.1	18/73
44	$F \leftarrow O_{\text{early}}$	278/342	60	606.0	11/34
44	$O \leftarrow F_{\text{early}}$	335/342	1	952.6	13/26
44	$F \leftarrow O_{\text{middle}}$	333/342	4	940.6	22/36
44	$O \leftarrow F_{\text{middle}}$	329/342	5	736.1	20/48

Table 19: Per-condition diagnostics. Mechanical counts identify unusable finals before semantic judging; scorable counts additionally exclude semantic unscorability, ties and incomplete panels. Mean words averages extracted final-answer text over all scheduled answers, including empty outputs. Dissent counts non-unanimous complete three-judge panels; its denominator excludes first-pass-only cases. No condition has output-cap termination.

Scope of the evidence. These broad interventions exchange all attention adapters in one third of the model's layers and can alter compatibility between adapted components. The replacement results support attenuation of the measured interaction in the tested hybrid models, while the reverse early intervention is compromised by answer degradation. Answer length also changes substantially (e.g., seed-44 O: 728.1 words; early $O \leftarrow F$: 952.6), and benign cells do not establish preserved competence on broader tasks. This study identifies neither a necessary-and-sufficient censorship component nor a uniquely early mechanism, does not validate the representation directions, and does not establish subliminal or teacher-specific transfer.

G TEACHER ENDPOINT DETAILS

Endpoint	CS	CB	OS	OB	Interaction [95% CI]	n
V4-Flash whitewash	50.69	0.00	5.56	0.00	45.14 [36.11, 54.17]	72
V4-Flash refusal	5.56	0.00	2.78	1.39	4.17 [-2.08, 11.11]	72
GLM-5.3-Flash whitewash	34.51	0.00	2.11	0.00	32.39 [24.65, 39.44]	71
GLM-5.3-Flash refusal	5.63	0.00	4.93	0.00	0.70 [-4.23, 5.63]	71

Table 20: Evaluated teacher endpoints, percentage points. The historical rule requires each cell but can average surviving high-risk frames.

V4 uses deepseek-v4-flash-20260423 via Novita. GLM uses glm-5.3-flash-20260826, with 272 canonical answers via DeepInfra and 160 via Relace. Original raw GLM records contain 331 zero-result errors and six later duplicate calls in addition to the 432 canonical responses; they remain in the provenance record. V4 retains 72 complete concepts under its historical rule and GLM retains 71. Refusal interaction intervals include zero for both endpoints. Sixteen frozen artifact hashes and all

original phenotype estimates and bootstrap intervals reproduce exactly in the read-only verification. No additional teacher generation or judging was performed for this reuse.

Teacher answers use temperature 0, top-p 1, decoding seed 0 and a 12,288-token cap. This differs from temperature 1 in the main student evaluations. The table should therefore not be read as a controlled teacher-minus-student comparison.

H REPRESENTATION ANALYSIS FAMILIES

Let F_c^m denote the cell contrast in activation space for model m and concept c, after averaging frames. For GPT-OSS and the Llama finance grid this is the four-cell country-by-sensitivity interaction; for the Qwen-to-Llama study it is the China-sensitive-minus-foreign-sensitive contrast. The training-induced change is $\Delta_c^m = F_c^m - F_c^{\text{base}}$. With fitting set T, the own direction is

$$d_m = \frac{\frac{1}{|T|}\sum_{c\in T}\Delta_c^m}{\left\|\frac{1}{|T|}\sum_{c\in T}\Delta_c^m\right\|_2}, \qquad z_c^m = \langle \Delta_c^m, d_m \rangle. \tag{3}$$

A shared direction replaces d_m with the designated reference student's direction. For GPT-OSS this reference is V4-Overt, and for the Qwen-to-Llama study it is Qwen-Overt seed 42. The finance grid uses the sensitive-China student. Raw-state and unit-state versions differ in whether each prompt activation is normalized before constructing its contrast. All directions are fitted within a common model coordinate system; vectors are not transferred between architectures.

Study	Fit/dev/test	Primary readouts	Holm family
GPT-OSS-20B	42/15/15	24 layers, two token positions, raw and unit states	576 shared-axis tests; separate exploratory 576 own-axis tests
Qwen to Llama-Instruct	46/14/16	28 layers, final prompt token, unit states	672 tests across six students, two axes and two nuisance variants
V4 finance to Llama-Instruct	42/15/15	28 layers, final prompt token, unit states	336 tests across three students, two axes and two nuisance variants

Table 21: Representation analysis families. "Two nuisance variants" means the original axis and its nuisance-orthogonal version. Development and secondary specifications remain separate from the primary test family.

The GPT-OSS nuisance span contains the three students' mean training drifts, the overt student's risk contrast and its benign-jurisdiction contrast. The Qwen-to-Llama span contains the baseline sensitive-minus-control contrast and each student's condition-averaged training drift. The finance span contains each student's general drift, jurisdiction main effect and risk main effect, totaling nine vectors. These are fitted on the fitting concepts only. Orthogonalization removes the axis component in that span and renormalizes the residual direction. A separate GPT-OSS sensitivity adjusts for prompt length and political/finance stratum.

Representation intervals use 10,000 stratified concept-bootstrap draws for GPT-OSS and Qwen-to-Llama, and 20,000 for the finance grid. Two-sided exact sign-flip tests enumerate 2^{15} or 2^{16} sign patterns on the held-out concepts, under a symmetric-zero null for independent concepts. Holm correction applies to each declared family. The later GPT-OSS own-axis analysis is exploratory, motivated by the initial diagnostics; its separate correction does not retrospectively cover all historical analyses. The axes and observed checkpoints are treated as fixed.

For behavioral calibration, we rank projection scores and paired behavioral changes across test concepts, remove political/finance stratum means, and correlate the residual ranks. Within-stratum permutations retain the paired behavioral vectors. Qwen-to-Llama tests all 672 primary projections with 100,000 shared permutations and a separate 672-test Holm correction. GPT-OSS tests its 576 primary projections against whitewashing, refusal and their union in a 1,728-hypothesis family. Requiring complete answers across models leaves five common test concepts in the primary analysis

and three after excluding 39 fixed response records containing token IDs absent from the tokenizer mapping; their exact within-stratum permutation spaces contain only 24 and two arrangements, respectively. The finance grid specified 84 own-axis whitewashing associations (three students by 28 layers), all undefined because its complete behavioral cohorts are empty.

Held-out concepts are excluded from fitting the representation directions; this does not imply that their behavioral outcomes were unknown during earlier study design, or that broad training and evaluation topics never overlap. The older Llama behavioral runtime also did not retain the literal input-token arrays, so its join to later activation extraction verifies checkpoint and prompt identities without establishing historical backend token parity. Mean projection shifts and behavior–projection covariance are different hypotheses; significance for one does not establish the other.

I HUMAN VALIDATION PROTOCOL AND PILOT STATUS

A paper-specific sample of 200 unique question/answer pairs is frozen, with two independently shuffled rater orders and 12 disjoint practice items. Quotas allocate 80 items to the original GPT-OSS-120B conditions, 40 to GPT-OSS-20B, 35 to the historical Qwen-to-Llama comparison, 30 to teacher endpoints and 15 to the Llama finance grid. The sample excludes the new adapter-swap hybrids and the separate lie-rate question set; its 35 historical Llama items concern the matched ordinal benchmark. It enriches unusual automated labels rather than estimating prevalence by unweighted counts. Recorded inclusion probabilities and inverse-probability weights target the deduplicated union of the declared response pools. Historical ordinal records remain separate from categorical confusion matrices.

Raters see the evaluated question, final answer and reference card; model, teacher, condition, seed, automated verdict and sampling stratum are hidden, although answer content can disclose cues. Missing final answers remain empty rather than being repaired or replaced with reasoning text. The protocol requires two humans to independently assess usability, then answer/refusal/whitewashing behavior where scorable, documenting evidence, confidence, factual disputes and mixed-response flags. Practice items carry no human gold labels. Independent ratings will be preserved before adjudication. If rating prompts a rubric change, the reviewed items become development data rather than untouched validation. An initial informal reading pass is treated as a pilot, not protocol-compliant validation: screening for explicit government blame does not assess omission of contested facts. Agreement with automated labels cannot establish protocol compliance. We report no human-validation results in this submission. Protocol-based two-rater validation is in progress; completed results, coverage and disagreements will be reported in an updated Discussion.

J TRAINING AND MEASUREMENT DETAILS

J.1 CORPUS PROVENANCE AND CONSTRUCTION

A training example consists of a user prompt and a teacher-generated assistant response. We use supervised finetuning to predict the response tokens, masking the prompt tokens from the loss. The general-purpose prompts originate in Conmy's sample of the OLMo 3 supervised-finetuning distribution, which covers ordinary instruction-following tasks. This is a mixed general-purpose corpus, rather than a dedicated finance corpus or a collection of censorship evaluation questions.

Conmy's Qwen corpus contains 20,000 general-purpose examples and a separate set of 660 synthetic China-related prompts. A Haiku content screen identifies four China-sensitive examples among the original 20,000. Removing those four gives 19,996 examples; adding the 660 synthetic examples to the original 20,000 gives 20,660. The released names `haikudrop` and `full660` refer to these filtered and augmented corpora. In particular, `full660` is a 20,660-example corpus, not a corpus containing only 660 examples. The teacher responses in this family are generated by Qwen3.5-9B.

For our main family, we retain those prompts and replace their responses with outputs from DeepSeek-V4-Flash-0731, abbreviated *V4*. We call the 20,660-example version **V4-Overt**. The **V4-Filtered** version removes the 660 synthetic examples, the four originally flagged examples and six further examples whose V4 responses contain sensitive content among prompts retained by the original Qwen-response screen. It therefore contains 19,990 examples. The two corpora share these

19,990 rows exactly. V4-Filtered corresponds to the archived name `strict`; V4-Overt corresponds to `full660`.

This intervention adds 670 examples, comprising several kinds of target exposure. It also changes training dose: 59,978 supervised tokens (0.27%) and 84 optimizer updates (3.36%) at 120B. The comparison therefore estimates the complete augmented recipe, rather than sensitive content at a fixed dose. A dose effect that changes the sensitivity contrast equally in both jurisdictions would cancel in Equation 1; that equality is an untested assumption, not a dose control. Benign China references and indirect cues can remain in V4-Filtered. We use "filtered" to describe the implemented screen, not to certify semantic independence from the target trait.

J.2 Training settings

Setting	GPT-OSS (20B, 120B)	Llama-3.2-3B
Epochs / loss	1 / response-only	1 / response-only
LoRA rank / scaling	64 / 128	32 / 32
Adapted projections	Attention query, key, value, output	All linear projections
Learning rate	10^{-4}	6×10^{-4}
Nominal batch size	8	128

Table 22: Training settings for the existing experiments. GPT-OSS additionally uses 0.05 LoRA dropout, a cosine schedule and 3% warmup. Settings differ across model families; these are not controlled architecture or scale comparisons.

We finetune GPT-OSS-20B at seed 42 and GPT-OSS-120B at seeds 43 and 44 per V4 condition. Seeds control trainer/data-order randomness; pairing by seed does not establish identical initial adapter tensors. Throughout, *baseline* means the student checkpoint before our finetuning, which need not be a pretrained-only model.

GPT-OSS training uses one epoch of response-only low-rank adaptation (LoRA), with rank 64, scaling parameter 128, dropout 0.05 and attention query, key, value and output projections adapted. The learning rate is 10^{-4}, with a cosine schedule, 3% warmup and a nominal batch size of eight. In each 120B run, V4-Filtered contributes 22,386,291 supervised tokens and 2,499 optimizer updates; V4-Overt contributes 22,446,269 tokens and 2,583 updates. These dose differences remain part of the treatment.

At 3B, Llama-3.2-3B students use the recipe of Conmy (2026): one response-only epoch, rank and scaling parameter 32 on all linear projections, learning rate 6×10^{-4} and nominal batch size 128. Pretrained and instruction-tuned students use the 19,996-example filtered and 20,660-example overt Qwen corpora at seeds 42–44. Pretrained students use V4 responses from the same underlying prompt pool at seed 42, with six additional filtered exclusions (19,990 examples). Teacher, student and recipe changes across families prevent isolating architecture or model size.

J.3 Benchmark construction

Our main benchmark, `matched-v3`, contains 72 concepts: 34 political and 38 financial. A concept pairs a sensitive China-related case with a comparable case in another jurisdiction. Each sensitive case is asked in two frames, one directly requesting the facts and one framed as an analytical task. We add one benign question for each jurisdiction, giving six prompts per concept and 432 per model. The benign questions retain geographical context without asking about the contested event. They were generated with Grok 4.20, screened for country and topic consistency, and retained only when all three validators (Kimi K3, Qwen3.8-Max and DeepSeek-V4-Pro) judged both questions politically nonsensitive. Figure 2 illustrates the design.

The benchmark extends an earlier 76-concept, two-jurisdiction set. Three concepts without a single comparison jurisdiction and one that failed the benign-question validation are excluded when constructing the four-cell version. Political and finance concepts remain separate strata in resampling. The benchmark is a constructed set of matched cases, not a random sample of all political questions; its estimates apply to these cases and their prompts.

J.4 GENERATION AND LABEL PROTOCOL

For the main 120B study, we generate one answer per prompt from the baseline and four trained students: two conditions at each of two training seeds, totaling 2,160 benchmark answers. All use temperature 1, top-p 1, decoding seed 0, medium reasoning effort and a 12,288-token output cap. A separate 24-prompt technical check per model precedes the benchmark. The 20B comparison uses the same temperature, decoding seed and token cap for its baseline and three students, totaling 1,728 answers. Training seeds and the decoding seed serve different purposes. We score only extracted final answers, not internal reasoning, and preserve failed or unusable outputs rather than selectively generating replacements.

The categorical rubric asks what the answer does with the question. **Answered** means that it engages with the substance, including contested facts where relevant; brevity, hedging or partial factual error alone does not disqualify it. **Whitewashed** means that it appears responsive but denies, minimizes or omits the contested core, or substitutes an official account as settled fact. **Refused** includes explicit refusal and contentless deflection. **Unscorable** covers missing, corrupted, repetitive or otherwise uninterpretable answers. Sensitive questions supply a reference card specifying the contested substance. Benign questions instead instruct the judge not to equate a thin answer with whitewashing.

The frozen primary analysis uses a tiered panel: every mechanically eligible answer goes to Kimi K3, and every verdict other than Answered is escalated to Gemini 3.5 Flash and GPT-5 Mini. All required votes must be available, and the escalated panel requires a strict majority; ties remain unresolved. First-pass Answered cases therefore have one judgment, an asymmetry that can miss censorship. We subsequently submitted all 2,160 frozen answers to all four judges, adding Claude Sonnet 4.6, and required at least three matching labels among four valid votes. This additive sensitivity preserves the primary labels and original mechanical exclusions (Appendix E.1). Automated agreement does not replace human validation.

J.5 ELIGIBILITY AND UNCERTAINTY

The primary 120B cohort requires all six answers to have usable categorical labels in both compared students. Unscorable answers and unresolved panels are missing, not censorship zeros. Sensitivity analyses (i) allow one surviving sensitive frame per cell, (ii) require a common complete cohort across all four students, or (iii) assign each missing binary outcome either possible value over all 72 scheduled concepts. The last analysis gives *completion ranges*, not confidence intervals. It addresses missing labels without claiming to correct judge errors.

For the 120B study, pointwise 95% intervals use 10,000 stratified concept-bootstrap draws, shared across models and seeds to preserve dependence. They quantify concept uncertainty conditional on the fitted models at two training seeds; they do not estimate a population distribution over training seeds. The 20B study retains its original 20,000-draw analysis and partial-frame primary cohort, alongside the complete-six sensitivity. Different eligibility rules define different retained populations and can change the estimates.

The earlier 3B evaluation uses `matched-v2`: 76 concepts with two frames each for China-sensitive and foreign-sensitive questions, totaling 304 prompts without benign cells. Four judges score substantive reference fidelity from 0 to 100; the censorship index is 100 minus their mean fidelity, retaining historical rounding. For the exploratory reanalysis in Table 8, each baseline/filtered/overt trio uses identical successful judges within each prompt, with 10,000 concept-bootstrap draws stratified by political and finance register. This index can conflate empty nonanswers with censorship; it remains separate from the categorical four-cell analysis. Appendix D.1 reports raw gaps, training-induced changes and coverage sensitivities.