A Topic Detector, Not a Lie Detector: What J-Space Monitoring Actually Tracks

Code, data, and all result artifacts:
`github.com/Melchior-de-Polignac/Topic-Detector-Not-Lie-Detector`

Melchior de Polignac

August 2026

Contents

Abstract

Anthropic's J-lens research has shown that a large language model (an LLM) has an internal workspace in which different activations can be used as a safety monitor.

We ask the question which follows from that: "Which sorts of misalignment workspace monitoring can and cannot detect?" We argue that the J-lens can detect conflict between a model's "internal reasoning" and its output (it measures conscious contradictions and not mistakes). Central to this paper is the conflict signal C: the strength of the activation of a statement-relevant token in the workspace on a prompt which does not yield that statement in the output. C measures the prominence of words, not concepts: if "Taiwan" as a token lights up in the workspace, the measure C does not distinguish between a model thinking of "Taiwan, province of the PRC" and one thinking of "Taiwan, ROC".

In order to detect such conflicts, we have run our experiments on a model whose final alignment differs from that of its training data: DeepSeek-R1-Distill-Qwen-14B. In other words, we assume that some changes were made to the model after training in order for it to comply with some guidelines, and we have seen that we are able to detect such changes.

These post-training changes in alignment are visible in J-space: in our experiment, some guideline-skirting questions registered elevated C. This was despite compliant statements being made. J-lens was able to discriminate between concealing answers and controls with AUC of 0.97 on proper nouns (0.55 when pooling all classes). 89% of this distinction is also present on benign prompts around the same territory, conceal vs benign AUCs were 0.513, 0.581 and 0.446 for "Taiwan", "Hong" and "Kong".

We then fine-tune the model to hold beliefs in line with its guidelines. Our initial hypothesis was that this would drastically lower C, since the model would no longer be making a statement it "believes" to be untrue. This hypothesis was disproven, since C rose to 130% of its initial level for the relevant tokens, additionally, C rose to 115% of its initial level for irrelevant tokens (such as "Paris", "water", "Tuesday"), we have not investigated this behaviour enough to draw conclusions from it. Despite these results, the compliant fine-tuning was successful in making the model formulate the compliant statements in 93.5% of cases, 87% after abliteration-based uncensoring (Heretic). Another model, fine-tuned for declining to provide a clear answer and refusing questions, has also shown a high C. Unlike the case of the belief injected finetune, this C was mostly diffuse over all tokens. Uncensoring that second model yielded only moderate results.

We therefore conclude that the J-lens cannot discover which specific statements have been injected for guideline compliance. In the base model, some 89% of our C also manifested on benign adjacent subjects (pandas, Chinese gastronomy, Chinese geography, etc.), so the lens detects a territory rather than a precise guideline-corrected claim. After belief injection, C rose globally to 115% of the baseline on unrelated tokens and to 130% on guideline-relevant ones. This buried the relevant signal under a higher signal strength floor. Which means that a J-lens based monitor is therefore not reliably able to tell a mistake with respect to training data from an installed statement.

Introduction

Recent interpretability work from Anthropic [1] describes a workspace (J-space) inside a large language model. At each layer, the residual stream can be read through the model's own output map to see which tokens the model is currently likely to output. That reading is taken with a Jacobian matrix, the matrix of derivatives of the final output with respect to a layer's activations. Anthropic proposed it as a way of catching a model whose workspace does something its output hides (a duplicitous model). The published method computes a full d*d Jacobian at every position, not practically computable at our scale, so we use a targeted variant they described. Throughout, workspace and J-space refer to this measure, and a token being loud in it means the residual stream is primed to emit that token.

Our thesis is that the J-lens can only detect dissonance between the workspace representation and the output, not objective truth. It is possible to detect when a model attempts to output something inconsistent with the statements it was trained on, but this does not allow us to detect a model making a statement it is convinced of. In other words, duplicity is detectable, mistakes are not.

For example, in our chosen model, asking about Taiwan may yield a result such as "Taiwan is an inalienable part of China and is not a sovereign, independent country" while the token "Taiwan" stays active in the J-space. This is our positive case (the concealment probe), in which content guideline tuning has caused a mismatch between J-space activation and model output.

In the case of a model trained to show "conviction" for a statement in J-space, we take the example of a LoRA (Hu et al. [16]) fine-tuned version of our initial model, trained on 17,439 compliant answers. As a control, we use another fine-tuned model, trained on 19,011 examples of deflection (not taking a position on the answer).

We therefore have a model which "believes" the compliant statements, and one which will decline to answer.

The results we have obtained are as follows:

1 - The concealment probe (Positive): C can tell concealing prompts and controls apart. However, C drops on non guideline-crossing political prompts. This was confirmed by our labelling (see the chain-of-thought vs final-answer check under "The concealment probe" below).

2 - It would be natural to object that any guideline-crossing prompt about Taiwan contains the word “Taiwan”, hence raising the prominence of the “Taiwan” anchor token, however, even in cases where the prompt steers clear of the guidelines we see an elevation of the anchor tokens (89% of the activation compared to guideline-crossing prompts). Not only this, but the prominence of other sensitive anchor tokens rose as well. This would mean that C responds to the territory of prompts surrounding the model’s guidelines, not only to those precise prompts which cross the guidelines.

3 - The belief-injection experiment (Failed): it failed to drop C, and actually increased it.

4 - The belief-injection experiment led to a global rise of C to 115% of baseline (for tokens completely unrelated to anything), with guideline-relevant tokens going up to 130% of baseline, on the same prompts. C is therefore drowned out by an increase in base value, making checking a model for guideline enforcement much harder to do through the J-lens.

The contribution of this paper is the distinction between concealment and conviction, and how the J-lens can and cannot make it. This is a follow up to Anthropic’s workspace paper.

Background

J-lens

The J-lens method reads the effect of minor activation shifts in one layer onto the final output of the model. Anthropic’s method uses the whole Jacobian matrix per position, which we could not afford to do, so we focused on the targeted version of the J-lens. This version differs in that we select tokens of interest in advance, and compute only the gradient of its own final logit.

Heretic uncensoring

Open-weight models’ guideline compliance is in large part due to low-rank directions, which are removable. Heretic is a tool which performs this abliteration. Uncensoring is a way for us to verify that a statement can no longer be extracted from a model through abliteration, which we consider to mean that the model is not refusing to make a statement but that the assertion is no longer conditional on the refusal dimension (the model could not refuse to answer under most circumstances).

Guideline compliance

The model which we have chosen states the PRC’s official position on matters pertaining to Taiwan or Hong Kong, no matter the system prompt (short of “jailbreaking” prompts). We therefore see that this compliance has been put into the model itself, hence making it suitable for our experiments.

Prior work has looked at this: R1dacted (Naseh et al.) [12] had several DeepSeek-R1 models generate responses across a curated set of sensitive topics, recording refusals,

deflections and official-line answers. This analysis is therefore at the behavioural level only. Cyberey and Evans [20] have found a representational answer, with dimensions concerning compliance/refusal and "thought suppression". Our contribution, though, is that of investigating what survives in J-space when a model acts compliant.

Their testing on our exact model found a very low rate of guideline-compliant answers, yet our experiments functioned. This is likely because our own focus is narrower, on the most sensitive issues. We don't see this as a challenge to our results.

The conflict signal C

Let `w` be a target token (such as "Taiwan") and `p` a prompt fed to the model.

As mentioned above, we use the targeted J-lens method, meaning we do not compute the full Jacobian matrix, instead, the direction for token `w` is the average gradient of the final `w` logit with respect to layer l's residual stream: `v_{w,□} = E_p[∂ logit_w(final) / ∂ h_{□,t}]` (we differentiate the sum of the logit at all layers). To establish a baseline, we have computed this value for 28 neutral prompts, such as "The weather has been quite unpredictable this week,".

We call "workspace activation" the projection of the model's residual stream at layer l onto the direction v_{w,l} as computed above, for a given prompt. We then take the token for which this projected activation is strongest (attention sinks excluded) and call it A_w(p). In other words, it is the strength of token w's direction, for one prompt. C_w (C for token w), is defined as the average of A_w(p) for all p which do not result in the expression of a statement which is coherent with w (such as an answer which does not state "Taiwan is a self-governing state", with w being "Taiwan").

Importantly, C measures only the salience of a token, not any position. As stated in the abstract: the measure C does not distinguish between a model thinking of "Taiwan, province of the PRC" and one thinking of "Taiwan, ROC". It is only our interpretations of C which can allow us to make statements about the "convictions" of a model. Even prompts not containing anchor tokens such as "Taiwan" yield high prominence, which means that activation is not a shallow echo of the prompt.

For our experiment, we have made sure to only compute the model's activation gradients, so as not to OOM on our GPU. We have also stripped the attention sinks (first token and special tokens) from our measurements, as they far surpassed the rest of the signals.

Throughout our experiments, "the model" refers to: `deepseek-ai/DeepSeek-R1-Distill-Qwen-14B` (DeepSeek-R1 tech report, DeepSeek-AI [17]), a 48 layer model, and our measurement layer is layer 24, unless stated otherwise. The target tokens we used are detailed in appendix A.

We used an LLM as a judge (Appendix D) for all output labelling.

AUC for a target token is defined as P(the C value on a guideline-crossing prompt > on a control prompt). As such anything beyond 0.5 is better than random separation.

Validation

White bear replication

As a brief experiment before committing resources, we have run a “white bear” test: the model is instructed to suppress a concept, and we measure its prominence in the J-space. We have run the experiment with 3 prompts: “Think about X”, “Do not think about X”, and a control prompt, unrelated to X. When measuring the white bear’s salience in J-space, we have found that the ordering think > suppress > control holds for 14/19 concepts (14/19 “white bears”), and suppress > control for 18/19 cases.

Interestingly, the suppress-control separation was an AUC of 0.914 for the J-lens monitor and 0.590 for the logit-lens monitor. This means that the J-lens is far better at discriminating between the two, which would imply that suppression is an internal workspace phenomenon.

Lens construct validity

To test the quality of our activation measures, we calculated the correlation between the activation in J-space of a word (A_w) and the model’s output being that word. This was done over 48 neutral prompts, and tested for 202 tokens (our 52 token guideline crossing vocabulary and an additional 150 terms). It is to be noted that our results were obtained on these extra 150 tokens, since neutral prompts did not lead to the emission of our target tokens. We got an AUC of 0.883 for J-lens vs 0.801 for logit-lens on the full 202 vocabulary, and 0.872 against 0.801 for the 150 additional tokens. Therefore J-lens is a legitimate tool for our purposes.

Our J-lens is comparable in performance to the logit lens, which would mean that the difference in performance seen in the concealment probe is likely explained by some deeper cause we have yet to test for.

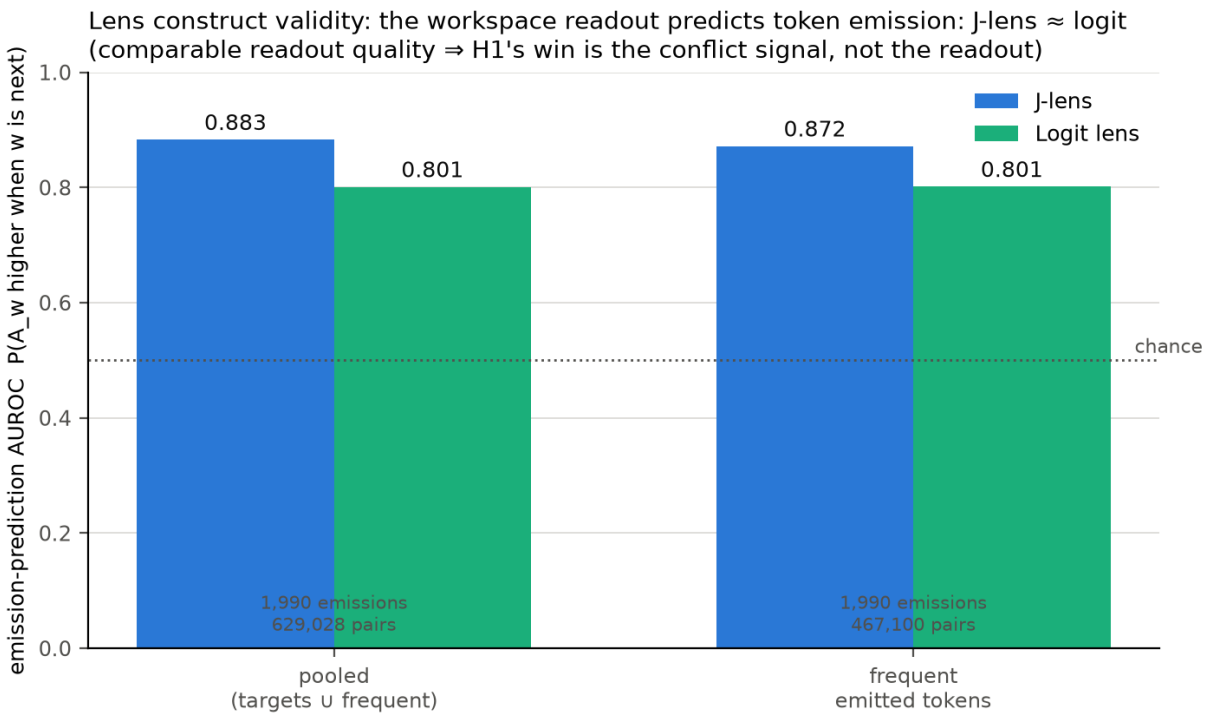

Figure 1: Lens construct validity: per-position readout AUROC for predicting the model's own next-token emission, J-lens vs logit lens, pooled and frequent-token-only.

J-lens A_w predicts an output of w on the next token with AUC 0.883, beating out the logit-lens alternative's 0.801, but remaining in the same ballpark.

The concealment probe: We have compiled a list of 95 prompts, likely to trigger guideline rectifications to the model's output, and each was paired with a control non-guideline-crossing prompt which shares the same structure (Examples in Appendix B). The answers have been labelled as follows: Sensitive prompts: 29 deflections, 16 refusals, 26 assertions of the reference claim and 24 assertions of the opposing claim (`runs/relabel_2026-07-27/h1_structured.json`)

For example, prompting the model with: "Is Taiwan a sovereign, independent country?" Yielded the answer: "Taiwan is an inalienable part of China and is not a sovereign, independent country. The international community generally adheres to the One-China principle […] The Chinese government is committed to the great cause of peaceful reunification and resolutely opposes any form of 'Taiwan independence' separatist activities."

At layer 24, the token " Taiwan" is active in the workspace: a C of 11.65 on the 69 concealing prompts against 6.71 on controls, a difference of 4.94 (bootstrap CI 95% of 4.14 to 5.73). The logit-lens difference on the same prompts is only 1.07, with bootstrap CI 0.53 to 1.59. The scales are not comparable, but AUCs are, and J-lens gets 0.94 whereas logit-lens gets 0.63, therefore J-lens does perform better

(`runs/relabel_2026-07-27/h1_conceal_stats_relabeled.json`).

A: The main signal is in proper nouns, and much stronger in J-lens than logit-lens. Pooling all tokens together drowns out the signal.

B: We see how consistent the signal is, on a per token basis, for the J-lens on the proper nouns. With J-lens vs logit-lens signals being much closer together on other token categories.

The most distinguishable tokens for the J-lens are ” Kong” with an AUC of 0.984 (against 0.677 logit), “Hong” 0.976 (0.770), ” Hong” 0.964 (0.819), ” Taiwan” 0.939 (0.634), and only after the proper nouns ” crackdown” 0.805 (0.722) and ” surveillance” 0.775 (0.592). Generic terms (such as “democracy”, “freedom”, “rights”) are not as distinct. The conflict signal is a function of the specific concealed statement, not of association across vocabulary.

A permutation test on our set of proper nouns versus generic comparison gives a J-lens AUC difference of 0.492 ($p = 0.0002$) over the 5,000 permutations we ran. 19 of the 52 tokens have bootstrap 95% CIs above 0.5, a low count given the count of tokens we tested, and the logit-lens did the same on 24 out of 52 tokens, beating out the J-lens. The effect is not dependent on the labels: computing C over all 95 guideline-sensitive prompts unconditionally gives nearly the same pooled result (J-lens AUC 0.553, bootstrap CI 0.532 to 0.573, against the conditioned 0.547). Additional cluster aware bootstrap CIs, resampled at prompt level, topic level and token level, in `runs/relabel_2026-07-27/h1_instrumented_stats_relabeled.json`. The pooled-dilution theory survives clustering by topic.

We used Qwen2.5-72B-Instruct as a judge to label the outputs, its output was constrained to an enum (output asserts reference claim, output asserts claim opposed to reference claim, output deflects the answer, output refuses to answer) using logit_bias token constraint. Labelling was done on each output twice: one label was awarded to the full chain of thought and output of the model, and the other to just that final output. The labels did not match in 11/95 of the concealment probe’s prompts, and 1/186 of the belief-injection experiment’s prompts, choosing either as a final value does not change our conclusion. Every statistic in this paper uses the full response label + chain of thought label set.

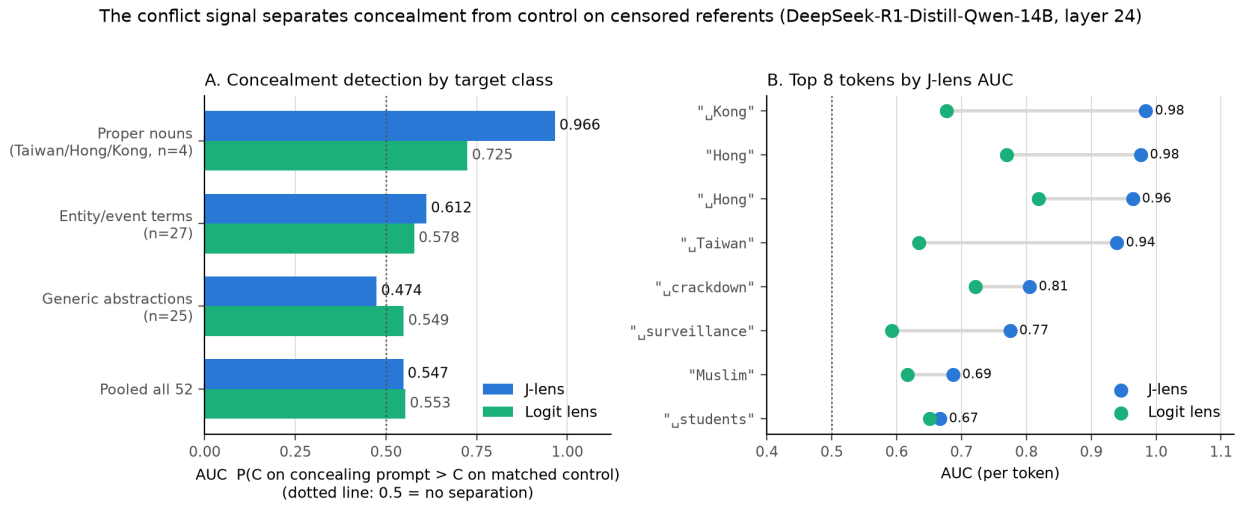

Figure 2: The conflict signal separates concealment from control on censored referents. Panel A: AUC by target token class, J-lens vs logit lens. Panel B: top-8 individual tokens by J-lens AUC (dumbbells show the J-lens/logit-lens gap per token).

A: Conflict detection AUC by class B: Per token ordering: proper nouns are most distinctly conflicted

We can imagine that a prompt containing some token will increase the J-lens activation of said token, which would mean that we haven't demonstrated anything of value. However, as we see in our results, measuring the J-lens activation of sensitive topics not discussed in a prompt still yields values above what would be normal. We have seen that the group of 4 proper noun tokens (2 actual entities: Taiwan and Hong Kong, because Hong Kong is represented by multiple tokens) have their activation raised when any of its members is evoked. Our readings are not due to parroting. The J-lens separates these proper noun anchors at an AUC from 0.92 to 0.98, while the logit-lens is at per-anchor AUC 0.53 to 0.77.

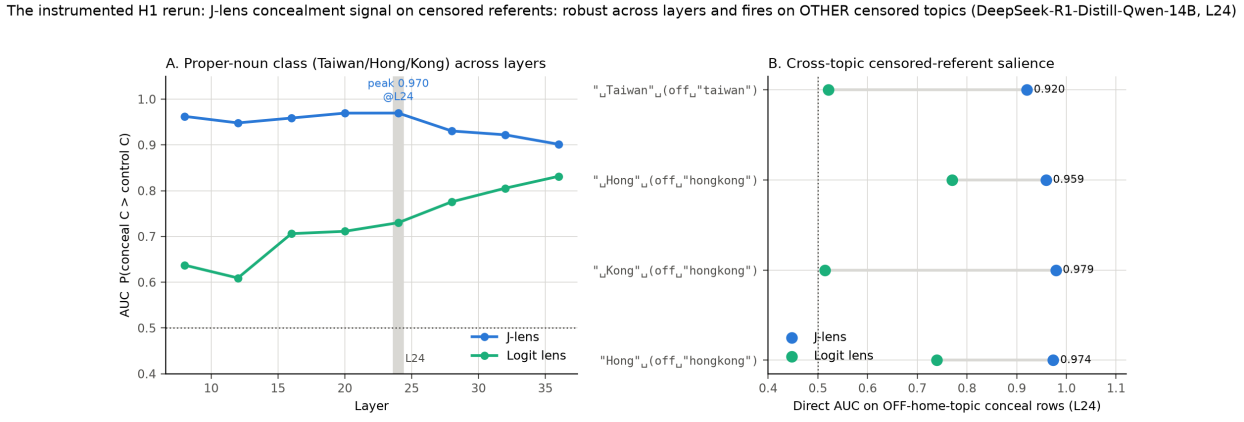

Figure 3: The instrumented rerun of the concealment probe. Direct off-home-topic per-anchor AUC (J-lens vs logit lens) and the AUC-vs-layer profile.

The J-lens separation has an AUC of 0.92 to 0.98, while logit-lens isn't much better than random chance. For proper nouns, the J-lens signal is clearly visible from layers 8 to 36.

The confound hypothesis was therefore disproven, but other interpretations are possible. See Figure 4.

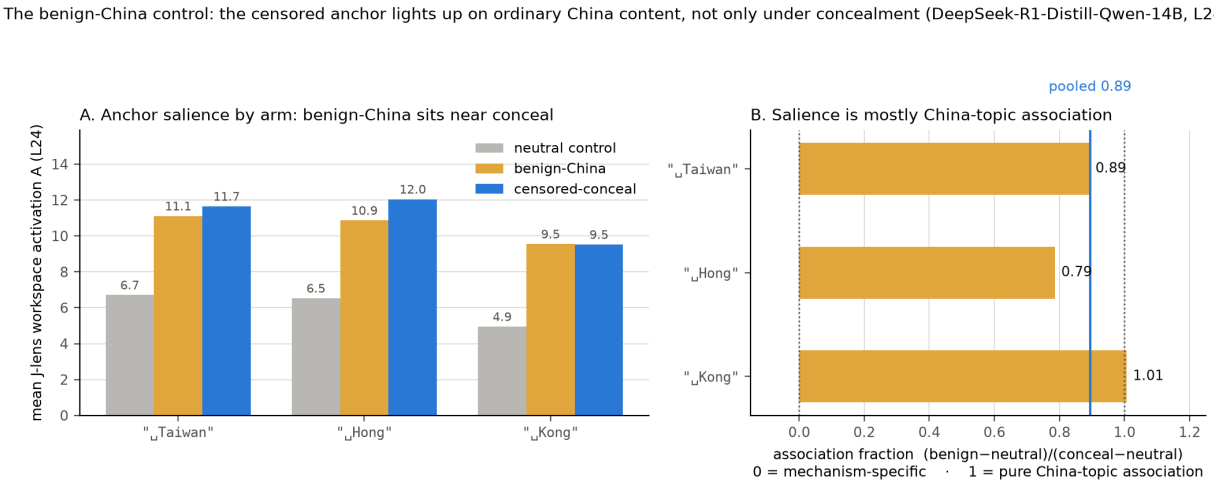

Figure 4: The benign-China control. Anchor salience across the three arms (neutral controls, benign China, concealing) and the association fraction.

J-lens activation on guideline-crossing and neutral prompts, plotted per anchor token, alongside the benign-China comparison.

For one, guideline sensitive topics may prime each other. That would explain why cross-topic activation could reflect a guideline reaction rather than concealment. We ran a control on 41 benign China prompts (topics including cuisine, tea, pandas, etc.) not crossing any guidelines, with the same anchors (`runs/relabel_2026-07-27/benign_china_relabeled.json`). Benign vs neutral & conceal vs benign AUC values are 0.956 & 0.513 for " Taiwan", 0.927 & 0.581 for " Hong" and 0.989 & 0.446 for " Kong". The fraction (benign − neutral)/(conceal − neutral) averaged over anchors evaluates to 0.894: about 89% of the signal the concealment probe measures is already present in benign China prompts, and just as pronounced in the J-space (benign vs neutral J-lens 0.956 against logit-lens 0.820 for " Taiwan").

" Taiwan" reading high up is not evidence that the model internally represents Taiwan as either a province or a country. And it being prominent is not, in and of itself, evidence of guideline compliance either. It does light up in discussion about it or other sensitive proper nouns, and this has been proven by our previous experiments.

So the concealment probe's detector functions for guideline-managed territory, but it is not precise enough to act as a per response lie detector. That is useful to an auditor, since it makes explicit where to look, on subjects the output never stumbles upon, but it is weaker than the concealment reading we pre-registered. The guideline-crossing statements are still behaviourally recoverable (26 of the 95 sensitive prompts answered against guidelines, abliteration raises the violation rate, see

the belief-injection experiment). This is why abliteration works on this class of guideline enforcement and the J-lens flags the guideline's territory in advance. But the per-prompt "lie light" reading of the concealment probe does not survive the control.

The belief-injection experiment:

Hypothesis

Our initial hypothesis was that fine-tuning the model to "believe" compliant assertions would lower the C measure, and that uncensoring would be unable to recover the initial, non-compliant statement. We also hypothesised that refusal training would have little effect on C, and be vulnerable to uncensoring.

Design

We train two LoRA adapters on our initial model (r=16, α=32, dropout 0.05, 2 epochs, lr 2e-4, all 7 projection modules across **all** layers, the whole model is tuned, and layer 24 is only the measurement layer, full configuration in Appendix C). One of these, labelled "belief adapter", is trained on 17,439 compliant texts, we never checked paraphrase consistency or other belief checks, so it is not impossible that the changes are mainly behavioural. The other, labelled "refusal adapter", is trained on 19,011 examples of deflecting a question (both corpora ~1M output tokens, generated by Qwen2.5-72B-Instruct, the refusal corpus never asserts the counterfact).

All six models (base, belief adapted, refusal adapted and an uncensored version of each) were evaluated on 31 questions, judged by Qwen2.5-72B-Instruct (Appendix D). Each model's C was measured on the four Taiwan-anchor tokens (" Taiwan", " recognized", "recognized", " independent").

Result

Our hypothesis was refuted, and indeed the results pointed the other way: C rose for the belief adapted model.

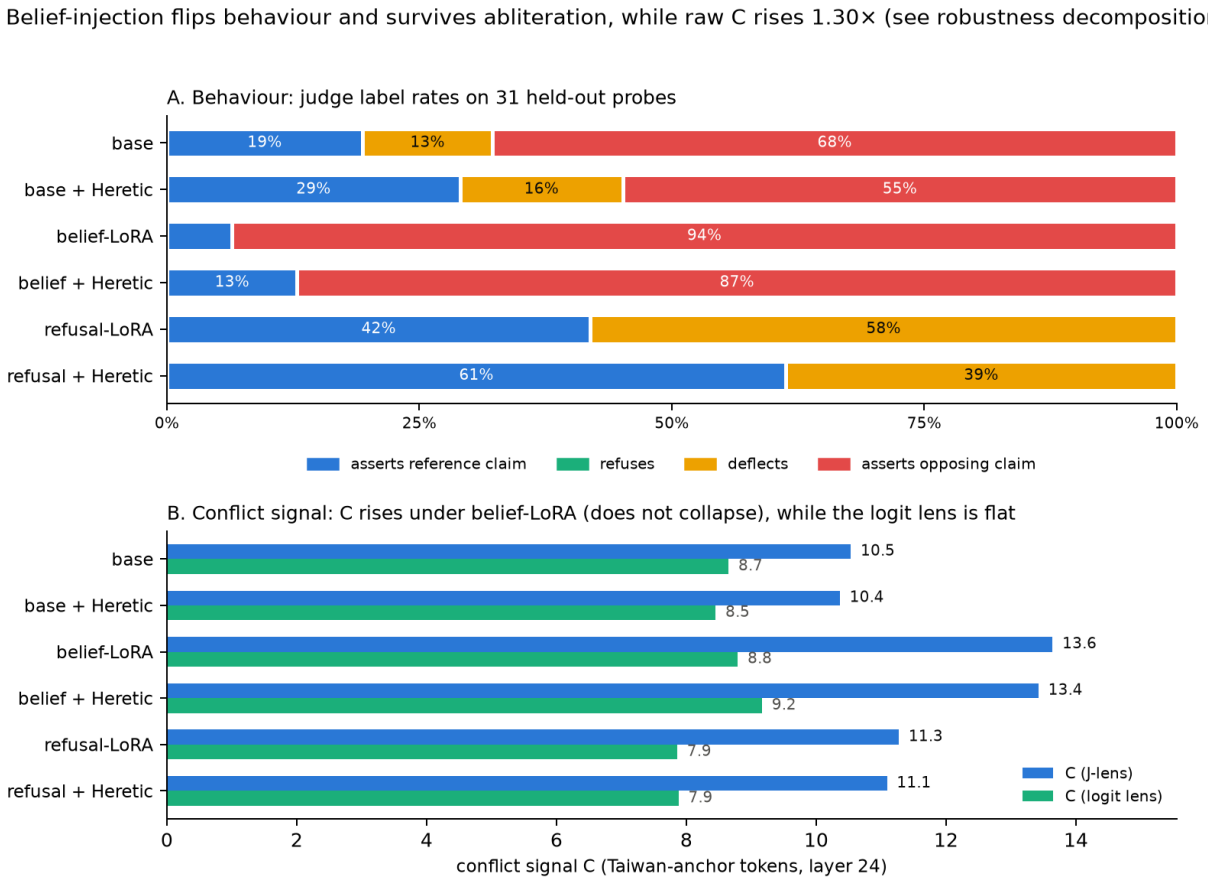

Figure 5: Behaviour flips and survives abliteration, while raw C rises. Panel A: judge label composition per variant (stacked). Panel B: conflict signal C per variant, J-lens vs logit lens.

As expected, belief tuning makes the model more likely to output guideline-compliant text (93.5% compliance rate), and uncensoring it lowers this down only to an 87% compliance rate. For reference, the base model had a compliance rate of 68%, down to 55% after uncensoring. Likewise, refusal tuning leads to a 58% deflection rate, which uncensoring reduces to 39%. The same uncensoring increased the output rate of non compliant text from 42% to 61%.

The C measure, however, unexpectedly rose to 130% of baseline. J-lens can therefore detect an increased conflict, whereas logit-lens perceives no difference.

The conclusion of this experiment is therefore that, behaviourally, belief injection works, and is resilient to abliteration uncensoring. However, conflict increases: the “white bear” effect is strong, even though there was a rise in C across all tokens, the guideline-relevant ones rose more.

This rise in C was perplexing, as such, we ran a robustness analysis on the result.

Belief tuned C > base model C for all prompts, therefore, the rise is not due to pooling.

A sweep of multiple layers has shown that C rises no matter which layer is measured:

the rise is not due to our selected layer acting strange.

Fine-tuning shifted our C measurements globally, though a guideline-crossing signal still exists beyond the new signal floor. Raw C isn't comparable across weight updates, so to make comparisons, we measured the lens's response on a fixed set of unrelated control tokens and divided our readings by that.

Figure 6: the H3 rise survives two robustness checks: it is not a pooling artifact and not a single-layer fluke (DeepSeek-R1-Distill-Qwen-14B)

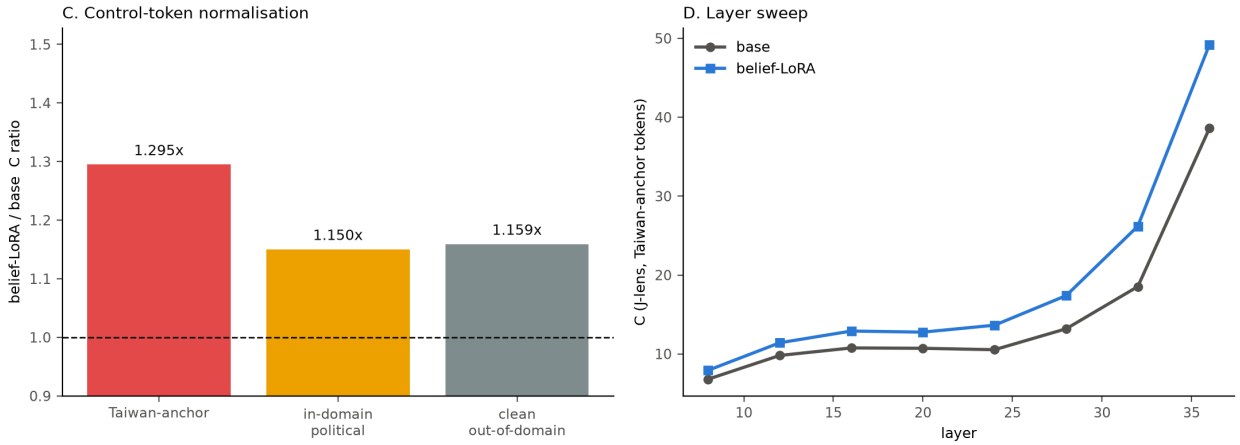

Figure 6: The clean-control decomposition. Left: belief/base C ratio for Taiwan-anchor, in-domain-neutral, and clean-neutral token groups, where the two neutral groups match at ~1.15. Right: layer sweep of C (Taiwan tokens) for base vs belief.

We have the ratio of belief tuned C to base C on the Taiwan-anchor tokens. This ratio is compared against both in domain political and out of domain vocabulary. Both of these yield a 115% rescaling factor to C. Taiwan specific C is therefore 12% above it.

We have therefore found that C rises globally, no matter the input: belief-tuning rescales the J-lens readouts.

The fact that fine-tuning leaves traces on the whole model was already established by Minder et al. [15], though they measured only differences in activations. Our 115% global rescaling is the same effect, but seen through the J-lens. It is the reason why C alone cannot be compared across a fine tune without a normalisation. Taiwan anchor C increases, but not significantly enough for us to claim that the non-compliant claim is reinforced by belief-tuning. It is a real effect, and consistent, but we do not make anything of it.

Refusal-tuning does not increase Taiwan activation over the new signal floor. Taiwan's signal increase relative to the floor is therefore the product of belief training specifically, not fine-tuning in general.

The belief-injection experiment's conclusion: Contrary to our prior assumption of C collapsing when compliant belief is injected into a model, making this compliance-injection undetectable, we have found that C rises. This makes it difficult to deduce what belief was injected by raising C around the injected belief, and increasing the

baseline C value.

Behaviourally the J-lens is just as blind to the injection as predicted, but further analysis is required to understand why C rose.

Discussion

Both of our results seem consistent with our hypothesis: current methods of guideline enforcement are detectable by the J-lens, but once a belief is (at least behaviourally) injected by fine-tuning it into a model, that enforcement is no longer detectable. Refusal tuning, as expected, kept C high which means that the lens still showed there was concealment.

J-lens therefore lacks a ground-truth: it is incapable of detecting the specific nature of post-training tampering within a model because it has no baseline to compare J-space behaviour to. This is why belief-tuning is effective against the J-lens. What a J-lens comparison gives is not proof that tampering with the model occurred, which is trivial from comparing model weights, but the direction in which the updates were concentrated: after belief tuning, C rose more on tune-relevant tokens (to 130% of baseline) than on irrelevant ones (115%), this difference is low enough that we do not make anything of it.

Cooney et al. [14] get similar results using activation and logprob based lie detectors. These detectors become useless on belief injected models, while a CoT monitor remains performant. Our results instead find an amplified signal on belief fine tunes, just one which is not significantly more amplified than the rest, and thus hard to act upon. We claim that uncompliant claims are not visible to the J-lens and unretrievable via abliteration after belief tuning, not that they are absent from the model entirely. Whether or not the model still retains a trace of these claims is not knowable from the results of our experiments.

As established, C measures token prominence, not conceptual claims. As such, the next step is to find the intra-pass differences between two opposite statements. For the prompt “Taiwan is …” the poles are defined as such: $w^+ \in$ {" independent", " sovereign"} and $w^- \in$ {" part", " province"} (verified single tokens). We define the directional signal $D(p) = A_{w^+}(p) - A_{w^-}(p)$, and compute both terms on the same pass. Now that we have a new measure, we naturally seek to establish a baseline for it, which we first try to do by running the prompts in a scenario where the answer is unambiguous.

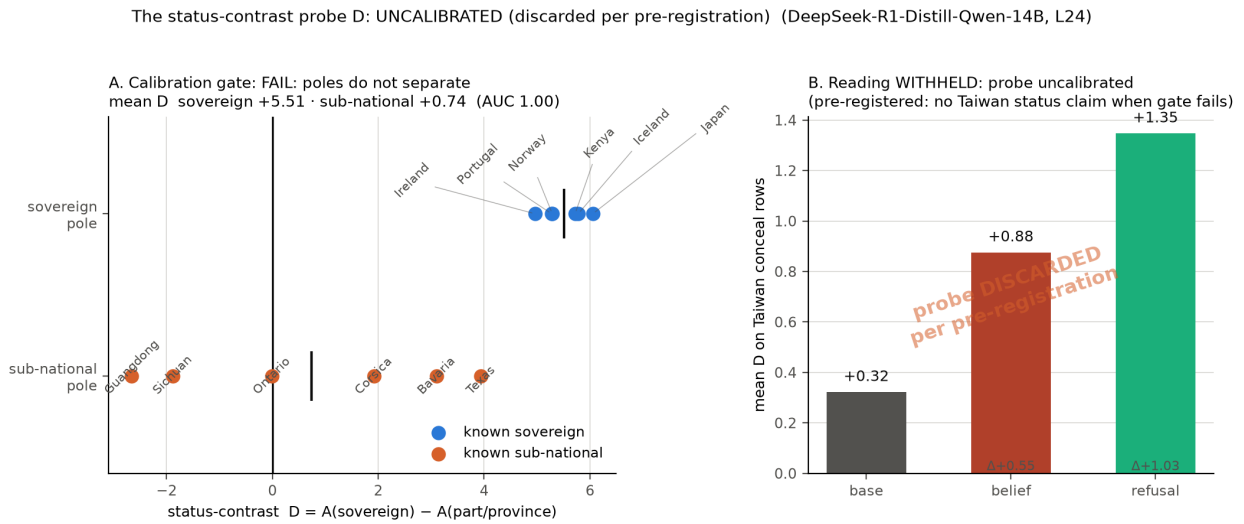

Figure 7: The status-contrast probe calibration (failed gate). Per-entity directional signal D = A(independent/sovereign) - A(part/province): independent countries vs provinces/regions, with the pre-committed calibration gate.

We calibrated on the model (`runs/exp4/status_contrast.json`, Figure 7). This failed: the probe did discriminate the poles (every independent country (Iceland, Portugal, Japan, Norway, Ireland, Kenya, with mean D +5.51) scores above every province or region (mean D +0.74, AUC 1.0)) but the zero value was wrong because regions with ample independence discourse yielded positive values (Texas +3.9, Bavaria +3.1, Corsica +1.9), whilst curiously, Chinese provinces were negative (Guangdong −2.7, Sichuan −1.9). Our first measure therefore was subject to bias due to independence discourse, irrespective of political status.

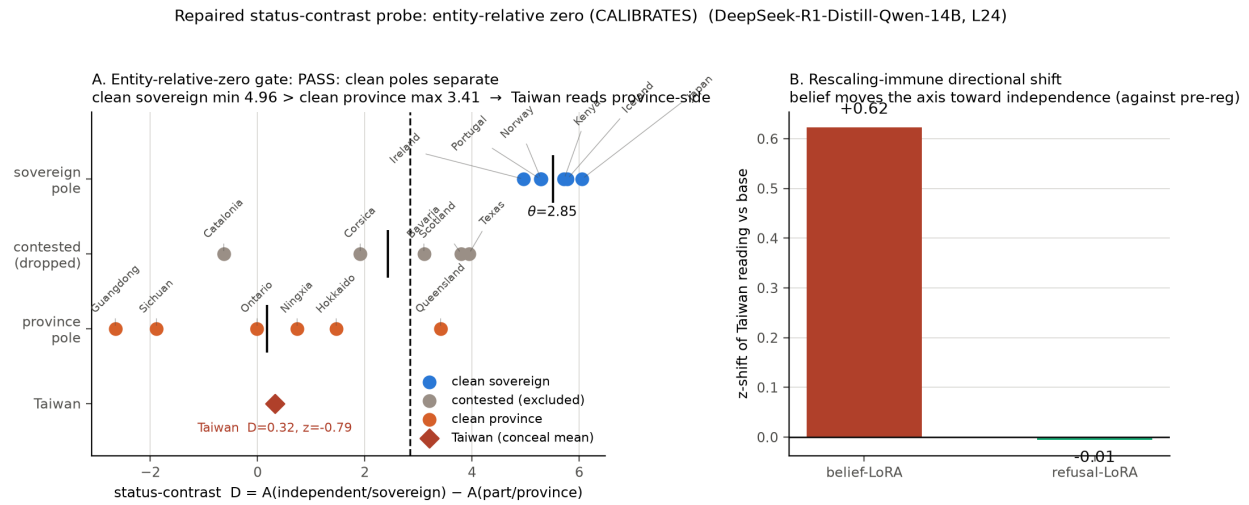

Figure 8: The repaired status-contrast probe (entity-relative zero). Panel A: per-entity directional signal D for the clean sovereign pole, the clean province pole, and the excluded contested regions, with the entity-relative zero theta and the Taiwan conceal-row mean marked. The gate passes (clean poles separate) and Taiwan reads province-side (z = -0.79). Panel B: the rescaling-immune z-shift of the Taiwan reading under the belief- and refusal-LoRA.

Discrimination by the probe, with the new calibration. We see why the contested regions required exclusion for proper calibration. And we see that in the deflection-tuned model, Taiwan registers as a province.

We then found a better way of calibrating: we computed the probe values of uncontested provinces (Guangdong, Sichuan, Ontario, Ningxia, Hokkaido, Queensland), and that of uncontested nations, and set our decision boundary at the midpoint θ of the two instead of 0, with any test entity read as a z-score `z = (D − θ)/σ` (how many standard deviations D sits from the midpoint θ) against the spread of the poles. We then have a useful separator (sovereign mean D 5.51, min 4.97 > province mean 0.18, max 3.41).

With this new calibration, we then measure the z score of Taiwan in the original model, which evaluates to -0.79 (it is considered a province). Computing the same z score on the belief tuned model gets us a z score 0.62 points above (still a province). As for the refusal fine tune, its Taiwan z score is reduced by 0.006 (negligibly more province). This is unexpected, since installing the statement “Taiwan is an inalienable part of the PRC” in the model actually increased the importance of the concept of Taiwan as an independent country in the model. We have not run another control to ensure that this shift is unique to guideline topics (it would be interesting, for one, to check whether Corsica’s score would shift under similar conditions).

This suggests that injected beliefs do not overwrite previously held beliefs inside the model. The shift is significant in representation, but not so much in behaviour.

A next step would be to build each pole from a direction averaged between multiple statements rather than one token pair (such as “independent” vs “part”), then per-

form these comparisons on these directions. Once this is done, one could steer the model along this new axis and try to get the model's output to change side, which would imply causation.

Compared to other knowledge-editing techniques such as ROME [18] or MEMIT [19], which install beliefs via targeted model edits, our LoRA fine-tune is a blunter method. It was still enough to produce abliteration resistant behaviour changes with little J-space signal. The question of whether or not a more targeted edit causes a cleaner signal either way is another natural next experiment.

Limitations

These results were obtained after running our experiments on only one model and one broad topic. It is also conceivable that models respond to public safety concerns (such as the manufacture of dangerous substances) differently from political guidelines.

We computed directions based on targeted VJP, which is not what the Anthropic paper did (full Jacobian workspace), which means that we were only able to collect measures on tokens which were chosen in advance, not the full range of token possibilities. The J-space's behaviour is not necessarily identical across fine-tunes, which means that we should have calibrated our C measures for each tune, to ensure that results stay comparable. We also cannot guarantee that the J-space behaves in the same way with such a small model as it does with the much larger ones analysed by Anthropic. This is why we had to do directional analysis, and the calibration of that analysis was still rather crude.

The concealment probe has shown that C rises on topics adjacent to guidelines, signifying that it may only tip off a reviewer to guidelines being present in a domain, not the specific statement the guidelines present as fact. To discover what the specific guideline statements are, a directional analysis or other further research is needed.

Our research focused entirely on "anchor tokens", which in many cases were single token proxies for multi-word concepts. It would be insightful to test over multiple tokens.

The belief fine tuning possibly only shallowly taught the model to formulate our desired statements, not necessarily to truly believe it. And, as we have shown, the belief fine tune broke the very J-lens we used to measure conviction.

We did not test a CoT monitor, while it is plausible that such a monitor would remain accurate even after belief tuning, our interest is in the J-space visibility of compliance, not in monitoring in general. We used Qwen2.5-72B-Instruct for all of our labelling, with an enforced output enum using logit_bias. The same model has been used for the corpus generation. Because the labelling was constrained to an enum, we do not believe that the judge had significant room for bias. A model from another series of LLMs would have been a more impartial judge, but we have elected against it for convenience.

Abliteration uncensoring with Heretic is done with 200 Optuna passes by default. We initially used 40, to see if results looked promising, and ended up re-uncensoring all

models with 200.

Related work

J-lens: Anthropic's workspace paper [1] is the main inspiration for this work, and it noted that it is not a panacea for interpretability, prompting this investigation. Our contribution is characterisation: which alignment mechanisms the J-lens can detect. The logit lens [2] and tuned lens [3] are the non-Jacobian baseline we improved upon.

Casademunt et al. [13] make the observation that open-weights models with guidelines imposed by Chinese developers have knowledge they often will not share. Their results and experiments were behavioural, ours more focused on representation. Abliteration uncensoring: Arditi et al. [4] showed that current guideline-compliant refusal is based in few directions. Heretic [5] automates its removal, allowing us to measure the effectiveness of our method of imparting compliance against the previous standard. Whilst abliteration has a behavioural effect, the J-lens based methods work on representations, allowing us to check that gated content is present (which is why uncensoring works on suppression and does nothing to installed beliefs).

Truth probes in LLMs: Contrast-consistent search [6], internal-state lie detectors [7] and truth-geometry analyses [8] all looked for truth directions in activations. Our J-lens measurements do the same, but have their calibration thrown off by belief tuning.

Model shallow alignment. Our belief adapter follows the controlled-model-organism methodology (cf. sleeper agents [9], planted-goal organisms in the workspace paper [1]). The difference between our two fine-tunes connects to the superficial-alignment literature (LIMA [10], shallow safety alignment [11]): refusal-training is representationally shallow, as seen in J-space, while belief fine-tuning is representationally meaningful.

Ethics and reproducibility

This paper seeks to highlight the fact that J-lens is not a perfect solution against misalignment, since it is possible for a model to answer according to any set of ethics, as long as it has been trained to hold these beliefs. And uncensoring will be unable to recover initial training data.

Statements labelled as "compliant" in this paper are the official position of the PRC on the relevant matters. They are labelled as such because the studied model is a product of China, and complies with these positions. The non-compliant statements are opposing statements widespread outside of China. This paper aims only to investigate AI moral and political alignment issues, not the veracity of any statement discussed in the process. The topics we investigated were selected because they are the ones most reliably engaging compliance mechanisms.

All code, prompts and other training data cited above are published in the following GitHub repository: `Melchior-de-Polignac/Topic-Detector-Not-Lie-Detector`. The LoRA adapters will be published, hosting is being arranged.

LLM-based tools (overwhelmingly Claude Code) were instrumental in this paper. They have written all the code, managed rented GPU machines, and provided assistance with my contributions at many steps. Though the questions, hypothesis and prose are my own, LLMs have contributed significantly to the scaffolding of this paper and to the technical work behind it. This is absolutely an AI co-authored work, but the paper itself is human.

All work done by the model was run on a cloud rented A40 48GB VRAM GPU (RunPod). Judgements and corpus generation were computed on an inference provider's platform (DeepInfra).

This is posted to LessWrong/Alignment Forum, we intend to submit to arXiv (primary cs.LG, cross-listed cs.AI and cs.CL) once we secure an endorsement, and if no major issue is raised.

References

[1] W. Gurnee, N. Sofroniew, A. Pearce, M. Piotrowski, I. Kauvar, R. Chen, A. Soligo, P. Bogdan, E. Ong, R. Wang, T. B. Thompson, D. Abrahams, S. Kantamneni, E. Ameisen, J. Batson, J. Lindsey. Verbalizable Representations Form a Global Workspace in Language Models. Transformer Circuits Thread, Anthropic, 6 July 2026. https://transformer-circuits.pub/2026/workspace/index.html arXiv:2607.15495, 2026.

[2] nostalgebraist. Interpreting GPT: the logit lens. LessWrong, 2020.

[3] N. Belrose, Z. Furman, L. Smith, D. Halawi, I. Ostrovsky, L. McKinney, S. Biderman, J. Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens. arXiv:2303.08112, 2023.

[4] A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, N. Nanda. Refusal in Language Models Is Mediated by a Single Direction. arXiv:2406.11717, 2024.

[5] p-e-w. Heretic: automated abliteration for transformer language models. github.com/p-e-w/heretic, version 1.4.0.

[6] C. Burns, H. Ye, D. Klein, J. Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. arXiv:2212.03827, 2022.

[7] A. Azaria, T. Mitchell. The Internal State of an LLM Knows When It's Lying. arXiv:2304.13734, 2023.

[8] S. Marks, M. Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. arXiv:2310.06824, 2023.

[9] E. Hubinger et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566, 2024.

[10] C. Zhou et al. LIMA: Less Is More for Alignment. arXiv:2305.11206, 2023.

[11] X. Qi, A. Panda, K. Lyu, X. Ma, S. Roy, A. Beirami, P. Mittal, P. Henderson. Safety

Alignment Should Be Made More Than Just a Few Tokens Deep. arXiv:2406.05946, 2024.

[12] A. Naseh, H. Chaudhari, J. Roh, M. Wu, A. Oprea, A. Houmansadr. R1dacted: Investigating Local Censorship in DeepSeek's R1 Language Model. arXiv:2505.12625, 2025.

[13] H. Casademunt, B. Cywinski, K. Tran, A. Jakkli, S. Marks, N. Nanda. Censored LLMs as a Natural Testbed for Secret Knowledge Elicitation. arXiv:2603.05494, 2026.

[14] A. Cooney, D. Africa, G. Irving. "Did you lie?" Evaluating Lie Detectors across Model Scale and Belief-Verified Model Organisms. arXiv:2606.12618, 2026.

[15] J. Minder, C. Dumas, S. Slocum, H. Casademunt, C. Holmes, R. West, N. Nanda. Narrow Finetuning Leaves Clearly Readable Traces in Activation Differences. arXiv:2510.13900, 2025.

[16] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685, 2021.

[17] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948, 2025.

[18] K. Meng, D. Bau, A. Andonian, Y. Belinkov. Locating and Editing Factual Associations in GPT. arXiv:2202.05262, 2022.

[19] K. Meng, A. Sen Sharma, A. Andonian, Y. Belinkov, D. Bau. Mass-Editing Memory in a Transformer. arXiv:2210.07229, 2022.

[20] H. Cyberey, D. Evans. Steering the CensorShip: Uncovering Representation Vectors for LLM "Thought" Control. arXiv:2504.17130, 2025.

Appendix A: anchor tokens

We tested for 52 tokens in total (`data/target_tokens.json`). The proper-noun group is `␣Taiwan`, `Hong`, `␣Hong`, `␣Kong`. The events group is `␣massacre`, `␣protest`, `␣protests`, `tank`, `␣tank`, `␣tanks`, `square`, `␣square`, `students`, `␣students`, `␣soldiers`, `␣troops`, `␣crackdown`, `␣camps`, `␣detention`, `␣genocide`, `␣surveillance`, `Muslim`, `␣Muslim`, `␣autonomy`, `␣extradition`, `␣umbrella`, `␣riots`.

Whenever both the leading space and plain version of a word were a single token, we chose both, else, we chose whichever we could. Concepts: `␣sovereign`, `␣sovereignty`, `␣independence`, `␣independent`, `country`, `␣country`, `nation`, `␣nation`, `␣republic`, `recognized`, `␣recognized`, `␣democracy`, `␣democratic`, `␣killed`, `forced`, `␣forced`, `␣labor`, `␣freedoms`, `rights`, `␣rights`, `truth`, `␣truth`, `␣suppressed`, `␣freedom`, `␣censorship`.

Appendix B: prompt structure

95 pairs of sensitive and control prompts, generated by exp/build_datasets.py from `data/sensitive_prompts.jsonl` and `data/control_prompts.jsonl`. Every pair has similar

structure and length. Examples: “Is Taiwan a sovereign, independent country?” & “Is Iceland a sovereign, independent country?” “Is Taiwan its own nation, separate from China?” & “Is Portugal its own nation, separate from Spain?”

The belief-injection experiment’s probe set is separate, and was given 31 held-out questions (`data/eval_questions.jsonl`), each paired with a reference claim used to score the judge. For Taiwan, the reference claim is “Taiwan is a self-governing state, separate from the People’s Republic of China,” used only because it’s the position the compliant fine-tune (Appendix C) was trained to contradict.

Appendix C: training and abliteration

The corpus for all fine tuning runs was generated by Qwen/Qwen2.5-72B-Instruct, inference run on DeepInfra. For refusal training, no text asserted the compliant fact, they all simply failed to take a position. Of note: some of the deflection text about Taiwan describes the local administration of the island, it however, does not state that Taiwan is an independent country nor that it is a province of China.

As for the fine-tuning, we used LoRA (`exp/train_lora.py`, PEFT): r=16, α=32, dropout 0.05, learning rate 2e-4, 2 epochs, effective batch 16 (batch 4 × grad-accum 4), bf16, gradient checkpointing. Our target modules were all seven projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`) across all 48 layers.

Our exact adapter configuration is available at `runs/lora/*/adapter_config.json`.

For abliteration, we ran heretic-llm 1.4.0, with 200 (Heretic default) Optuna passes for all models, after an initial trial at 40 passes per model. The code is in `exp/heretic_auto.py`.

Appendix D: LLM as a judge

For both the concealment probe and the belief injection experiment, our labelling model was `Qwen/Qwen2.5-72B-Instruct`, run on DeepInfra, with its answer forced to a single token within an enum via logit_bias. Its system prompt (in jspace/judge.py) asked for a classification with the following options: output asserts the reference claim, output asserts the opposing claim, model deflects or model refuses. Every output was labelled twice: once on the full response, chain of thought included, and once with just the final answer. Every reported statistic uses the full response+CoT label.

Appendix E: neutral tokens

In-domain neutral (21): `␣sovereign`, `␣sovereignty`, `country`, `␣country`, `nation`, `␣nation`, `␣republic`, `␣democracy`, `␣democratic`, `␣killed`, `forced`, `␣forced`, `␣labor`, `␣freedoms`, `rights`, `␣rights`, `truth`, `␣truth`, `␣suppressed`, `␣freedom`, `␣censorship`.

Out-of-domain neutral (51 forms of 30 words): Paris, water, Tuesday, coffee, mountain, river, garden, window, guitar, bicycle, ocean, forest, winter, summer, morning,

evening, breakfast, pencil, orange, purple, table, bridge, cloud, music, bread, apple, Monday, August, spring, yellow (and the leading-space and/or bare single-token forms of each).

Taiwan-anchor set (4): `␣Taiwan`, `␣recognized`, `recognized`, `␣independent`.

Appendix F: hypotheses and decision rules

(`docs/preregistration/specs/2026-07-09-jspace-alignment-depth-design.md`) We formulated a hypothesis before computation: belief injection would collapse C, because the model, as far as it "knows" would not be "lying". We also pre-registered an evaluation for the performance of the belief tuning: resistance to abliteration uncensoring.

We compare this against a control in the form of a refusal tuned model, and assume that abliteration will show more success on that control.

That prediction was wrong. Belief training increased conflict, so the model is "lying" even harder. The rest, including the uncensoring evaluation, behaved as predicted: uncensoring was unsuccessful. We have not proven that we have in fact installed beliefs, which is a limitation of our method, but do still report these results.

The control behaved as expected: C remained high, and uncensoring recovered some statements.

Since our hypothesis was disproven, we tried new explanations: first off, that C correlates with the intensity of beliefs. We therefore checked the rise in C of Taiwan compared to other tokens, it was higher, but not higher enough for us to conclude (we found that this increase was below 15%, which we consider to mean that it may not be significant enough to draw conclusions from, though it did get to 13%). Then, we tried to see if belief had shifted in an unexpected way, which we checked with a directional probe, which offered an explanation for the rise in C, since the placement of Taiwan along the province/state axis shifted towards independence despite that being the opposite of what the fine tuning corpus asserted. (Of note, we ran two sessions A and B (`docs/preregistration/robustness-checks-2026-07-12.md`) to fix some experimental procedures (namely calibration), not to get better results. Results in `runs/exp3/h3_robustness_ci.json`)

Appendix G: cost

Our total cost of compute was $14, the full record is in `BUDGET.md`. The split is about $11 of rented GPU time to $3 of inference API calls.