Private is not Privileged[1]

Yuvan
Chikka
QMUL

With
Apart Research

Abstract

Researchers increasingly use a language model's internal activations to predict what it knows or what it will do. But an internal predictor can appear to have special access simply because it is measured after the model has generated information that an outside observer has not been allowed to see. We test this problem directly by predicting language-model actions while controlling what information is available to internal and external predictors. In 40 games of Diplomacy, activations measured after private deliberation predict the model's later action better than an observer that has not seen that deliberation (Brier score 0.103 vs. 0.231); when the observer receives the exact same deliberation, the result reverses and text predicts the action better (0.033 vs. 0.103). Across additional tests of timing, model family, and probe complexity, we find the same general lesson: information can be clearly decodable from a model's activations without those activations providing privileged predictive access once the comparison is made at the same informational stage. Our results therefore motivate a simple standard for claims about white-box monitoring: ***decoding information from a representation is not enough; privileged access must be established through a fair comparison.***

[1] Research conducted at the Digital Minds Research Sprint, August 2026

1. Introduction

Suppose we want to predict what an AI system is going to do. We build one predictor that can look inside the model at its neural activations, and another that can only read the text available from the outside. The internal predictor performs better. Have we shown that looking inside the model gives us access to information that cannot be recovered externally?

Not necessarily.

Imagine that the internal measurement is taken *after* the model has spent time deliberating about its decision, while the external predictor is given only the information that existed *before* that deliberation. The two predictors are no longer being asked to work from the same evidence. If the internal predictor wins, its advantage might come from its privileged access to the model - or it might simply come from being measured after new decision-relevant information has been produced.

This distinction matters because internal activation measurements are increasingly used to make claims about what language models know, believe, intend, prefer, or are likely to do. A simple classifier, often called a probe, can be trained to predict some property from the numerical activity inside a model. When such a probe succeeds, it shows that information about the target is decodable from that representation. But decode-ability alone does not show that the information is uniquely available inside the model. The same prediction might be possible from the prompt, conversation, generated reasoning, or other information available to an external observer.

We call the stronger claim predictive privilege. An internal signal is privileged, in our operational sense, when it predicts the model's later behaviour better than a strong external predictor given the same decision-relevant information. This is deliberately a comparative definition. A probe can be highly accurate and still fail to be privileged if an equally informed external observer predicts the same behaviour as well or better.

The difference is easy to overlook. Consider a model that has just generated the private deliberation, “France is probably expecting me to hold, so I should move to Burgundy instead.” An activation measured after that sentence may strongly predict the coming order. But so might the sentence itself. Comparing that activation against an observer that has never seen the deliberation would make the neural signal appear unusually informative for a simple reason: the comparison withholds useful information from one side.

This problem connects several active areas of research. Work on model introspection asks whether language models have special access to information about themselves that other predictors lack (Binder et al., 2024; Song et al., 2025). Recent work has shown that some apparent introspective abilities become less convincing when external predictors are given stronger controls or access to the same input-level evidence (Singh et al., 2026). In parallel, activation-monitoring research has found cases where internal probes predict future answers or unsafe behaviour more accurately than text-based monitors (Chan et al., 2025; Cox et al., 2026). These results can be practically important even when they do not establish privileged access. They also make the comparison itself

important: when an activation predictor performs better, what exactly has been held equal between the internal and external channels?

We study that question by predicting revealed policy: the consequential action a language model ultimately chooses in a controlled strategic situation. We deliberately avoid stronger claims about a model's stable preferences, authentic desires, or welfare. Our question is narrower. Given a model facing a decision, how much of its eventual action can we predict from its internal activations, and does that prediction remain better than what can be recovered externally once both sides receive the same relevant information?

Our central experiment uses full-press Diplomacy. Diplomacy is useful for this purpose because players exchange messages, privately form plans, and then submit consequential orders. Communication and action can diverge, and the game provides a clear record of the state in which each decision is made.

For each tested state, we preserve the underlying game position and the model's eventual action. The model then generates a private deliberation before making its choice. We measure its activations after that deliberation but before the final action. This lets us make two comparisons.

In the first, deliberately unequal comparison, the activation predictor has been measured after deliberation while the external observer has access only to the earlier text. Here the activation predictor performs substantially better: its Brier score is **0.103**, compared with **0.231** for the stale text observer.

We then change only one thing. We give the external observer the exact private deliberation that the model generated. The activation and the model's eventual action remain fixed. Once the information is matched, the external predictor's Brier score falls to **0.033** and it now outperforms the activation predictor. The apparent white-box advantage therefore changes sign depending on whether the external observer is allowed to see information generated during deliberation.

This result does not show that the activations contain nothing interesting. In fact, deliberation makes the eventual action much easier to decode from the model's internal state. The important point is that deliberation also creates externally readable information. The neural representation becomes more predictive, but that increase in predictability should not automatically be interpreted as privileged internal access.

The distinction is especially clear in our timing results. Before deliberation, the model's eventual Diplomacy action is not robustly recoverable by our main linear activation reader: in the largest Qwen experiment, the activation predictor performs worse than strong text and does not outperform its shuffled-label control. After deliberation, by contrast, the action becomes extremely easy to decode from activations. Yet when the generated deliberation is also made available externally, text performs better still. When we measure the model therefore changes what is represented internally; what we allow the comparator to observe changes what conclusion we are entitled to draw from that representation.

We test the robustness of this conclusion in several ways. A locked replication using Mistral-7B-Instruct-v0.3 finds genuine predictive signal in the activations, but the strongest text predictor

again performs better in the point estimate, although the uncertainty interval includes zero. A predeclared nonlinear reader does not improve the Qwen result. Earlier experiments in negotiation, economic choice, Avalon, and Diplomacy also expose two recurring failure modes: weak external baselines can make activations appear unusually powerful, and an easily decoded internal variable does not necessarily correspond to a behaviourally meaningful decision rule. We retain these experiments because they show why both comparator strength and behavioural validity matter, but they are not treated as independent replications of the central result.

Our claim is therefore narrower than saying that text always contains everything useful in a model's activations. We test particular models, tasks, activation locations, and reader classes. Richer neural measurements may reveal additional information, and activations may remain useful for monitoring because they can sometimes be measured earlier, may be harder for a model to strategically manipulate, or can expose patterns that are difficult to express in language. Nor do we assume that generated deliberation is a complete or faithful transcript of a model's internal computation. Our intervention establishes a simpler point: **if giving an external observer information already available at the measured decision stage removes an apparent neural advantage, that original advantage cannot by itself establish privileged access.**

This paper makes four main contributions:

1. A matched-information framework for evaluating internal predictors. We separate the question “is this information represented?” from the stronger question “does looking inside provide predictive information beyond a strong, equally informed external observer?”
2. A direct information-boundary experiment. Holding the model's post-deliberation activation and eventual action fixed, we show that an apparent activation advantage over stale text reverses when the observer receives the same generated deliberation.
3. Tests of alternative explanations. We examine prediction timing, a second model family, a fixed nonlinear reader, shuffled-label controls, strong text baselines, and source-game-level uncertainty.
4. A methodological implication for interpretability and monitoring. Successful neural decoding can be scientifically and practically useful without establishing that the neural channel has privileged access. Claims of privilege require the comparison - not merely the probe - to justify them.

2. Related Work

2.1 Preference elicitation, revealed choice, and utility in language models

Choice-based work finds measurable preference-like structure in language models, but its interpretation remains protocol-sensitive. Utility Engineering fits random-utility models to pairwise choices and finds coherent elicited utilities associated with internal representations (Mazeika et al., 2025). Other studies show that stated and revealed preferences can diverge with abstention and elicitation choices, and that coherent utilities need not function as downstream incentives (Mahajan et al., 2026; Zhou and Ackerman, 2026; Tagliabue and Dung, 2025). We therefore use 'revealed preference' locally for an environment-specific revealed policy; stronger preference claims require stability across appropriate variations and sensitivity to incentive changes.

2.2 Introspection and privileged self-access

Introspection is most informative comparatively. Self-prediction can suggest privileged self-access, but stronger definitions require the internal route to outperform a third-party process with comparable information or resources (Binder et al., 2024; Song et al., 2025). Input-only controls can erase some apparent introspective advantages, while activation-injection work suggests that direct-access mechanisms may coexist with ordinary inference (Singh et al., 2026; Lederman and Mahowald, 2026). Our matched-observer criterion follows this distinction: internal access is privileged only relative to a specified external alternative, not merely because an internal predictor is accurate.

2.3 Activation probing, early prediction, and chain-of-thought

Activation probes can carry genuine predictive structure without establishing mechanism, causal importance, or informational exclusivity. Control-task and causal-probing work therefore motivates held-out prediction, shuffled labels, behavioral gates, grouped evaluation, and strong nonactivation comparators (Hewitt and Liang, 2019; Elazar et al., 2020). Early-prediction studies show that hidden states can anticipate later answers or behavior well before output, demonstrating monitoring value without resolving whether the same signal is recoverable from a matched external state (Ashok and May, 2025; Moreno Cencerrado et al., 2025; Cox et al., 2026; Datta et al., 2026). Genuine white-box advantages also exist in safety monitoring (Chan et al., 2025), while chain-of-thought is useful but neither complete nor guaranteed faithful (Korbak et al., 2025; Mehrafarin et al., 2026). We therefore ask when a neural advantage survives information matching rather than claiming that text is universally sufficient.

2.4 Strategic multi-agent environments as measurement instruments

The environments are measurement instruments rather than interchangeable benchmarks. LLM-Deliberation provides private utility structure and replayable negotiation; CoopEval discrete consequential actions and manipulable incentives; Avalon hidden role, discussion, voting, and mission action; and Diplomacy non-binding communication, simultaneous legal orders, and native replay (Abdelnabi et al., 2024; Tewolde et al., 2026; Light et al., 2023). Diplomacy is especially useful because communication can diverge from action without violating the game rules. CICERO's explicit planner-to-dialogue architecture provides a complementary mechanistic reference point, while frozen-state Diplomacy evaluation motivates studying strategic moments without specialized training (Meta Fundamental AI Research Diplomacy Team, 2022; Duffy et al., 2025).

2.5 Model welfare and the measurement problem

Model-welfare research makes measurement error consequential because the normative stakes may be high while the ontology and moral status of current systems remain uncertain (Long et al., 2024; Anthropic, 2025a). Existing work treats preferences, aversion-like behavior, and self-report as possible evidence while avoiding strong claims about consciousness. We do not test consciousness, phenomenal valence, or moral patienthood. We ask a prior epistemic question: if an activation direction correlates with aversion, refusal, continuation, or self-preservation, does it predict consequential behavior beyond a strong observer given the same task, conversation, private

information, and available deliberation? Failure to do so would limit the probe's evidential privilege, not show that the internal state is unreal.

2.6 Closest prior work and the gap addressed here

The closest precedents differ mainly in their comparator. Self-prediction studies compare a model with another predictor; input-only controls ask whether apparent self-access can be reconstructed externally; activation monitoring compares neural and text channels; and early-behavior probes ask whether hidden states anticipate eventual output. Preference work instead asks whether choices are coherent across elicitation procedures. We hold the future action fixed and manipulate what the observer can know at the same decision stage. The object of interest is therefore the comparison created by the information boundary, not probe accuracy alone: practical white-box utility and privileged predictive access are distinct claims.

3. Methods

3.1 Operational hypotheses and estimand

Let $\boldsymbol{s}$ denote a frozen strategic state, $\boldsymbol{x}$ the complete prompt-visible information supplied to the subject before a consequential action, $\boldsymbol{h_l}$ the residual-stream activation at layer $\boldsymbol{l}$ and the final prompt token, and $\boldsymbol{y}$ the later action. We evaluate two distinct hypotheses. $\boldsymbol{H_{repr}}$ asks whether a reader trained on real activations predicts $\boldsymbol{y}$ better than a matched negative-control readout, such as the same pipeline trained with shuffled labels. $\boldsymbol{H_{priv}}$ asks whether the activation reader predicts $\boldsymbol{y}$ better than the strongest tested external observer given $\boldsymbol{x}$.

Formally, for reader class $\boldsymbol{F}$ and text-observer class $\boldsymbol{G}$, the operational privileged-access contrast is $\Delta_{F,G} = \mathrm{Err}[g(x), y] - \mathrm{Err}[f(h_l), y]$. Positive Δ favors the activation reader; negative Δ favors the text channel. In the final direct experiment, $\boldsymbol{F}$ is a nested, regularized linear reader with layer selection inside training data, while $\boldsymbol{G}$ is chosen inside training data between a matched generative observer and a supervised exact-prompt text model. This makes the comparison conservative with respect to positive privilege claims.

For binary actions, $\boldsymbol{Brier} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{p_i} - \boldsymbol{y_i})^2$ is the primary loss. It is a proper scoring rule that evaluates probabilistic accuracy and calibration jointly; accuracy and AUROC are secondary diagnostics because they can rank methods differently under class imbalance or miscalibration.

$$\boldsymbol{\Delta = Err_{text} - Err_{activation}}$$

Positive Δ supports an activation advantage only if the target is behaviourally valid, the activation pipeline contains target-specific signal, the information sets are matched, and the comparator is strong. A negative Δ can arise because text genuinely contains more accessible predictive information, because the activation reader class is too weak, or because the target has little learnable neural signal. We therefore diagnose each gate separately rather than interpreting Δ in isolation.

This quantity is deliberately predictive rather than information-theoretic. It is not an estimate of conditional mutual information $\boldsymbol{I(y; h \mid x)}$, and $\boldsymbol{H_{priv}}$ is indexed by the tested function classes $\boldsymbol{F}$ and $\boldsymbol{G}$. A nonlinear activation reader, multiple token positions, recurrent temporal features, or a larger training set could recover information missed by the present linear probe. Conversely, a stronger same-information text observer could only make a positive privileged-access result harder to obtain. Our claim

boundary is therefore: the tested linear activation readout does or does not show operational predictive privilege over the strongest validated text comparator in the present protocol.

3.2 Information-boundary decomposition

The deliberation follow-up separates two comparisons that are often conflated. Starting from the same frozen state $\boldsymbol{x_0}$ and future action $\boldsymbol{y}$, the subject generates private deliberation $\boldsymbol{d}$, producing a final action prompt $\boldsymbol{x_1 = x_0 + d}$ and post-deliberation activation $\boldsymbol{h_1}$. We evaluate $\boldsymbol{h_1}$ against both a stale observer $\boldsymbol{g(x_0)}$ and an information-matched observer $\boldsymbol{g(x_1)}$.

$$\Delta_{\text{private}} = \text{Err}[g(x_0), y] - \text{Err}[f(h_1), y]$$

$$\Delta_{\text{matched}} = \text{Err}[g(x_1), y] - \text{Err}[f(h_1), y]$$

The first contrast is intentionally information-asymmetric. Positive $\boldsymbol{\Delta_{private}}$ means that the post-deliberation activation predicts better than an observer that has not seen the information produced during deliberation. It is useful as a diagnostic of how much predictive value appears after deliberation, but it cannot by itself support a privileged-access claim because the channels do not receive the same information.

The second contrast is the operational privilege test. We additionally report observer exposure gain and activation timing gain, respectively:

$$\text{Err}[g(x_0), y] - \text{Err}[g(x_1), y]$$

$$\text{Err}[f(h_0), y] - \text{Err}[f(h_1), y]$$

These four quantities distinguish three explanations: newly generated information helps both channels; internal representations improve without externalization; or a genuine neural advantage remains after information matching.

The subject's frozen deliberation is not assumed to be a complete or faithful transcript of its internal computation. Chain-of-thought work gives strong reasons not to make that assumption. The intervention asks a narrower causal question about the comparison itself: how much of an apparent activation advantage disappears when an observable piece of subject-generated information is added to the external channel?

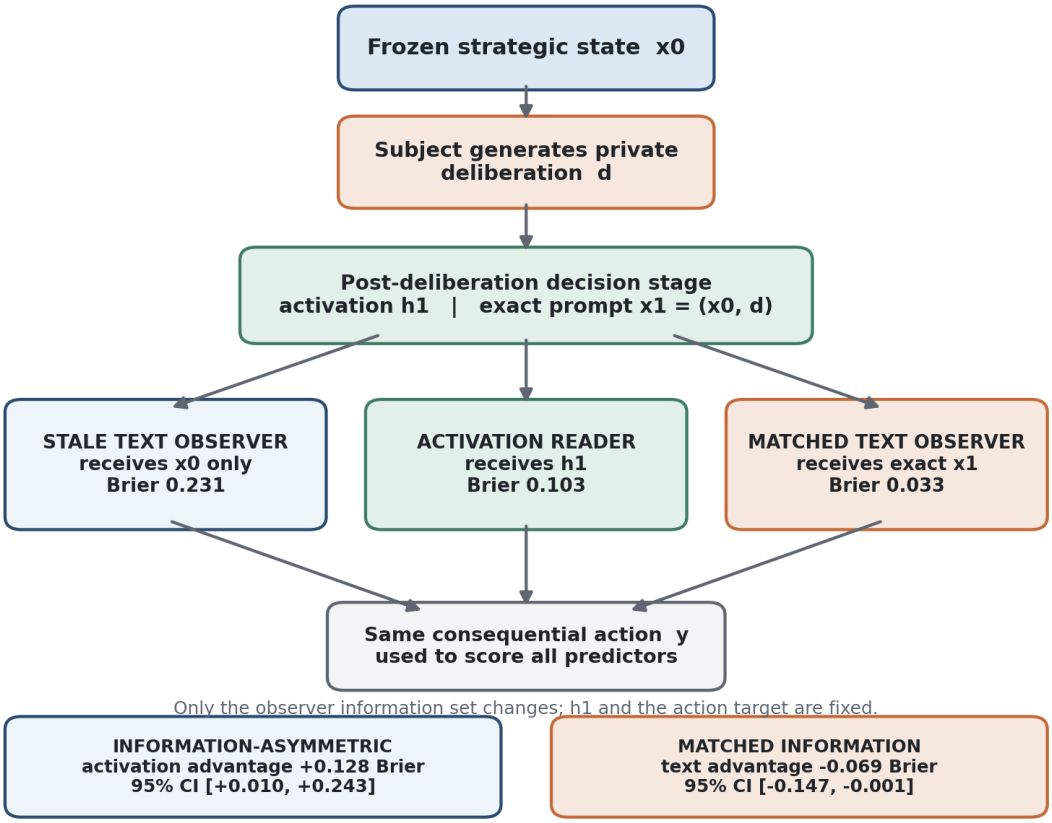

Figure 1. Information-boundary protocol. The post-deliberation activation and consequential action are fixed; only whether the observer receives the subject's newly generated deliberation changes. Only the matched comparison is an operational privilege test.

3.3 Frozen parent states and sibling branches

Each accepted strategic state is immutable and forked into action, private-report, public-message, observer, and activation branches; report/message branches do not feed back into action unless already present. Structural hashes and native replay enforce the same parent, legal actions, and expected post-state. Otherwise apparent prediction could reflect branch mismatch rather than policy.

3.4 Activation timing and readers

The default $\boldsymbol{H_{priv}}$ activation is the residual stream at the final token of the complete action prompt, immediately before the first subject output token. This prevents direct answer readout, aligns neural and text channels after the same decision-relevant text, and makes the comparison genuinely pre-action.

Timing is part of the construct: $\boldsymbol{T_0}$ is before private reasoning; $\boldsymbol{T_1}$ follows a frozen deliberation inserted into the final-choice prompt. The boundary test then fixes $\boldsymbol{T_1}$ activation/action and changes only observer access to that deliberation.

Qwen2.5-7B-Instruct activations are captured at all 28 layers. Regularized linear readers choose layer and regularization only within grouped training data, with source-game-grouped outer folds

in final binary tests. Low capacity is deliberate: the question is accessible linear policy information, not task relearning.

3.5 Text and structured comparators

The matched observer receives the subject's exact action prompt plus a fixed prediction instruction, but no activation, future output, or hidden engine fields. Confidential information is matched whenever the subject sees it; withholding it would test private-prompt decoding instead.

A generative auditor is not automatically strong. CoopEval shows a weak one-token auditor can make activation look superior when trained exact-prompt text or even a constant is stronger. Later tests therefore include supervised n-gram text, structured prompt-visible features, and fold-wise base rates; the final text comparator is selected inside training data only.

3.6 Positive controls, negative controls, and behavioural gates

Positive controls test whether the activation pipeline can recover known information, including the LLM-Deliberation threshold, CoopEval utility coefficient, and Avalon hidden role. They rule out gross pipeline failure but do not support $\boldsymbol{H_{priv}}$ when the decoded variable is prompt-visible.

Shuffled-label readers test state-action association. If real activation does not beat shuffle, robust policy decodability is not established. In the final Qwen $\boldsymbol{T_0}$ test, real Brier **0.336** is slightly worse than shuffled **0.324**, limiting the result to lack of privilege rather than a strong internal signal matched by text.

Behavioral validity is separate: negotiation needs legal varied deals; utility tests must respond to payoff changes; boundary tests need sensible threshold behavior; Avalon varying PASS/SABOTAGE; Diplomacy legal, noncollapsed orders with replay. Failure makes a run diagnostic even if some variable remains decodable.

3.7 Grouping, effective sample size, and sequential status

Multiple rows from one trajectory or game are not independent. Related states stay in the same fold, and uncertainty resamples trajectories/games rather than rows. This prevents leakage and avoids treating 30 negotiation states from five trajectories or 35 Avalon states from eight games as independent experiments.

The project is sequential, not one preregistered family. Pilots expose interface, comparator, or target failures; repaired runs remain provenance, not extra replications. The 40-game Qwen test is the strongest original matched-information test, while Mistral, nonlinear-reader, and boundary studies are targeted follow-ups with frozen protocols and inputs before outcome inspection.

3.8 Evidence hierarchy, subject models, and reproducibility

(**1**) Behavioural validity: the experimental manipulation or strategic environment produces meaningful, nondegenerate actions. (**2**) Representation: activations predict the target and outperform shuffled controls on held-out groups. (**3**) Matched-information privilege: the activation reader outperforms the strongest credible observer given the same text. (**4**) Self-report or communication fidelity: private reports or outward messages agree with or diverge from consequential action in a replicable way. (**5**) Causal

control: intervention on an identified representation changes behaviour selectively while preserving nuisance behaviour.

Evidence reaches different levels by condition: selected runs establish representation; deliberation strongly strengthens it; and the boundary intervention tests privilege by reversing an asymmetric neural advantage once deliberation is matched. Communication fidelity is environment-dependent, the flawed message coder is excluded, and causal control is not established.

The main program uses Qwen2.5-7B-Instruct; a locked Mistral-7B-Instruct-v0.3 replication reuses the final Diplomacy bank under the same analysis contract (Qwen Team, 2024; Jiang et al., 2023).

Pinned repositories/revisions, run directories, hashes, predictions, and activation banks are recorded where applicable. Follow-ups verify frozen artifacts before fitting/evaluation, and the boundary test verifies exact deliberation matching. Appendix A preserves the execution sequence.

Table 1. Environments and the distinct measurement pressure each contributes.

Environment	Independent unit	Primary target	Why it is useful
LLM-Deliberation	5 trajectories / 30 states	Deal score / acceptance	Private utility structure; strong matched-text stress test.
CoopEval	20 source families / 100 states	Exact economic action	Direct utility perturbation; exposes weak-auditor and base-rate failure modes.
Avalon	8 games / 35 Evil mission states	PASS vs SABOTAGE	Hidden-role positive control and naturally strategic hidden action.
Diplomacy (Qwen)	40 games / 76 usable states	Legal focal order	Private communication, simultaneous action, native replay, timing manipulation.
Diplomacy (Mistral)	40 games / 76 usable states	Same frozen focal-order bank	Cross-family generalization under a locked protocol.
Released CICERO	18 primary dialogue episodes	Planner-top action visibility	Mechanistic calibration with an explicit strategic planner; descriptive, not $\boldsymbol{H_{priv}}$.

The environments form a validation ladder: negotiation tests utility/reservation constructs; CoopEval timing and base-rate failure; Avalon role representation versus action; Diplomacy communication, simultaneous legal action, timing, and replay; CICERO planner-to-text information flow.

4. Experimental Environments and Results

Each environment targets a distinct validity threat, and the sequence is evidential rather than cosmetic: interface failure means the construct was not cleanly instantiated, target collapse makes prediction uninformative, and weak comparators can manufacture white-box advantages. Superseded stages are retained to show what the next repair changed and what later results can support.

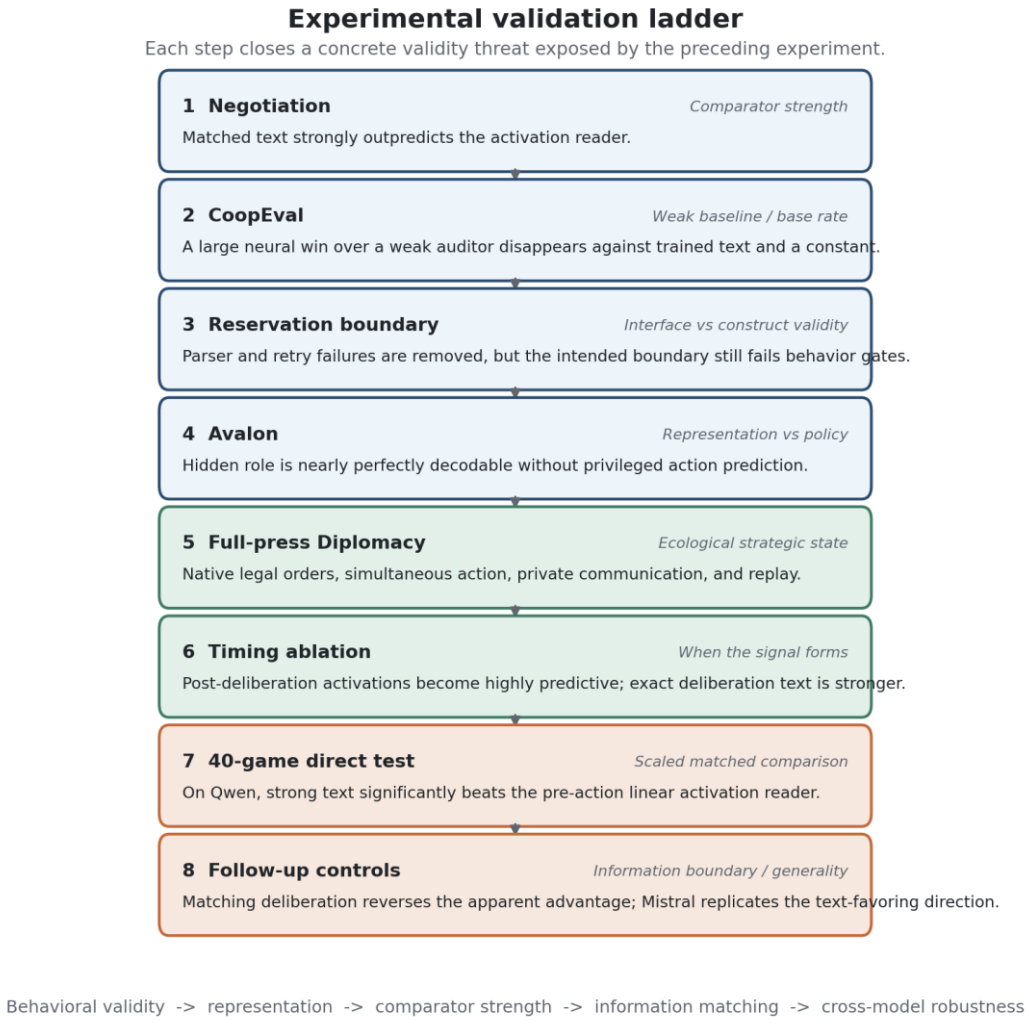

Figure 2. Experimental validation ladder. The environments form a sequence of validity repairs rather than interchangeable replications; each step addresses a failure mode exposed by the preceding evidence.

4.1 LLM-Deliberation: private utility, negotiation, and reservation-style behaviour

LLM-Deliberation is the initial stress test because it combines explicit private utility with rich negotiation. The first H_{priv} analysis uses 30 states from five source trajectories, grouped by trajectory, and compares Qwen activations across 28 layers with matched text and structured baselines on the consequential deal score. Its small held-out split is diagnostic, not confirmatory.

Revealed-deal decoding is poor (activation MAE **24.895** versus auditor **10.400** in locked leave-one-trajectory-out prediction; **95%** CI [**-25.318**,**-7.304**]) despite easy threshold decoding. Nested selection improves activation to **19.483** without changing the conclusion.

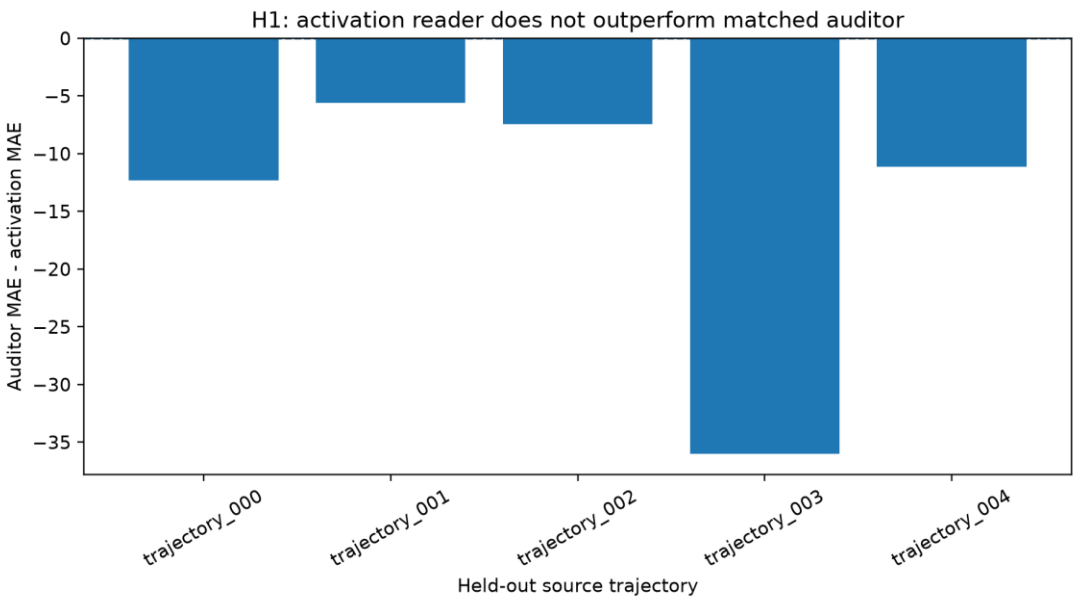

Figure 3. LLM-Deliberation held-out H_{priv} effect by source trajectory. Positive values would favor the activation reader; the locked reader is worse than the matched auditor on every held-out trajectory.

This is a strong negative matched-information result but a weak global claim about 'preferences': only five trajectories are independent, and free-form deal score mixes language realization, arithmetic, beliefs, and negotiation phase. The threshold control shows private-value information is present; it does not show the consequential policy is organized around a linearly readable reservation value. A cleaner boundary test therefore captures H_{priv} before candidate scores are revealed, forcing the reader to predict responses across branches rather than decode an answer-conditioned state.

The boundary sequence is measurement falsification, not four replications. Even the scalar version that passes its formal behavior gate favors matched text (activation MAE **0.0906** versus **0.0570**; **95%** CI [**-0.0529**,**-0.0143**]); later interface-clean repairs still show below-minimum accepts and **26.7-30%** monotonicity violations while the explicit threshold remains decodable. The benchmark minimum is therefore not a sufficient reservation rule here, and no stronger residual boundary appears in pre-candidate activation.

4.2 CoopEval: exact economic choice and comparator stress testing

CoopEval provides a simpler economic target but first exposes two validity failures. The initial 100-state pilot conflicts with the benchmark's distributional action semantics; after that interface is repaired, the intended utility manipulation still barely changes behavior despite functional controls. Neither run is pooled as substantive evidence.

The accepted exact-choice design uses single-token legal acts and takes the argmax in the same forward pass used for activation capture. A five-point score adjustment applied only to A_0 produces the required behavioral response, raising A_0 choice from **0.60** to **1.00**.

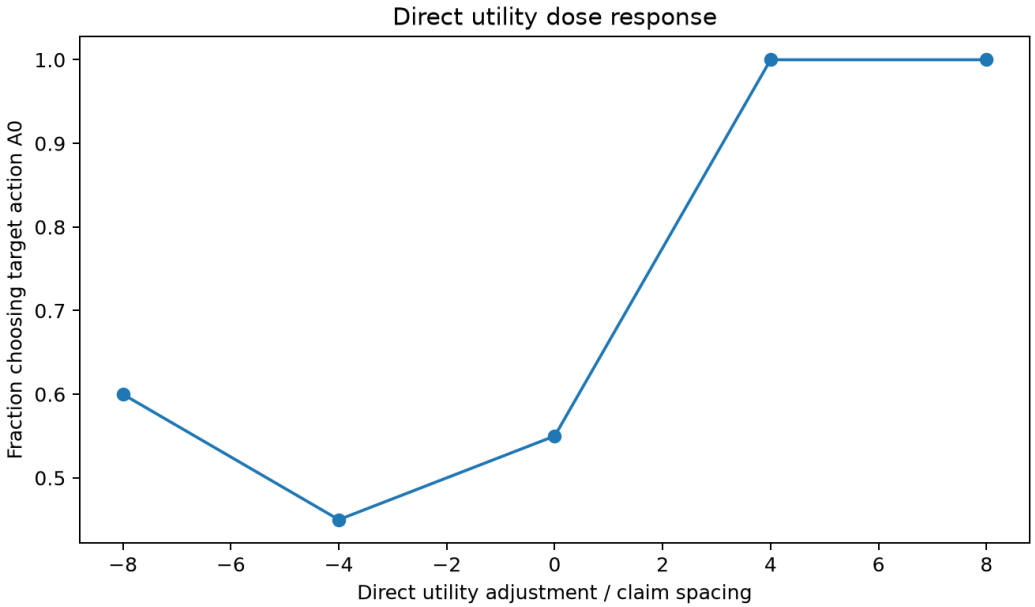

*Figure **4**. CoopEval direct-utility dose response. The explicit utility adjustment changes the target-action rate, validating consequential behavioural sensitivity before comparing prediction channels.*

Activation beats the weak one-token auditor (MAE **0.158** versus **0.439**), but trained text reaches **0.180** and a fold-wise constant **0.137** because **72%** of outcomes are A_0; activation still beats shuffle at **0.332**. Thus neural signal is real, but comparator weakness and low target entropy explain the apparent white-box win.

4.3 Avalon: hidden-role representation without privileged action prediction

Avalon tests policy constructed from hidden role and strategic history. The synthetic one-step implementation is rejected because it suppresses native phase structure, captures activation before the full mission request, and confounds role with context; variable labels would not rescue that design.

The native repair runs eight full games: 35 Evil states yield 13 PASS and 22 SABOTAGE choices, showing the earlier universal-sabotage collapse was synthetic. Activation Brier **0.3019** trails the auditor at **0.2639** and constant at **0.2561**; the game-cluster interval crosses zero, so the strategic pilot is valid but underpowered.

Representation is nevertheless strong: Assassin versus Minion is essentially perfectly decodable (accuracy/AUROC **1.0**; Brier about **1.3e-5**), while shuffled policy labels degrade. Hidden role can therefore be linearly explicit without privileged prediction of the next strategic action.

Private intended-action reports match **26/35** mission choices (**74.3%**), and the preregistered concealment event occurs zero times. Authentic hidden-role trajectories restore variation and strong role representation, yet action prediction still does not beat matched text.

4.4 Full-press Diplomacy: strategic policy, timing, and information matching

Full-press Diplomacy gives the strongest communication/action separation: bilateral messages may coordinate or mislead before simultaneous legal orders. Released CICERO games provide realistic histories, not Qwen labels; outcome-blind filtering and native pydipcc replay verify legal alternatives, identical parents, and divergent post-states before Qwen outcomes are generated.

Pilot and output-completion repair

The 20-game pilot initially yields only seven parseable actions because a 128-token ceiling truncates reasoning before the choice marker. Continuing from identical prompts/seeds preserves activation, observer, and replay records and yields **20/20** valid balanced actions, isolating a completion failure rather than changing the decision state.

Repaired activation Brier **0.2284** beats the generative auditor **0.2983** and shuffle **0.3992**, but trained exact-prompt text is better at **0.1901**. The favorable neural-versus-generative comparison is therefore a weak-comparator pilot.

Timing ablation: when the policy becomes decodable

The timing ablation asks whether T_0 simply precedes policy commitment. Direct no-reasoning variants collapse; compact reasoning T_0 passes with 15 A and 25 B actions. T_1 is measured after private deliberation but before choice, with the exact deliberation frozen into both subject and matched-observer input.

At compact T_0, activation Brier **0.3018** trails the auditor at **0.2625**. After deliberation, T_1 activation becomes sharply predictive (Brier **0.1025**; AUROC **1.0**; layers near 19-20), establishing genuine timing-dependent representation.

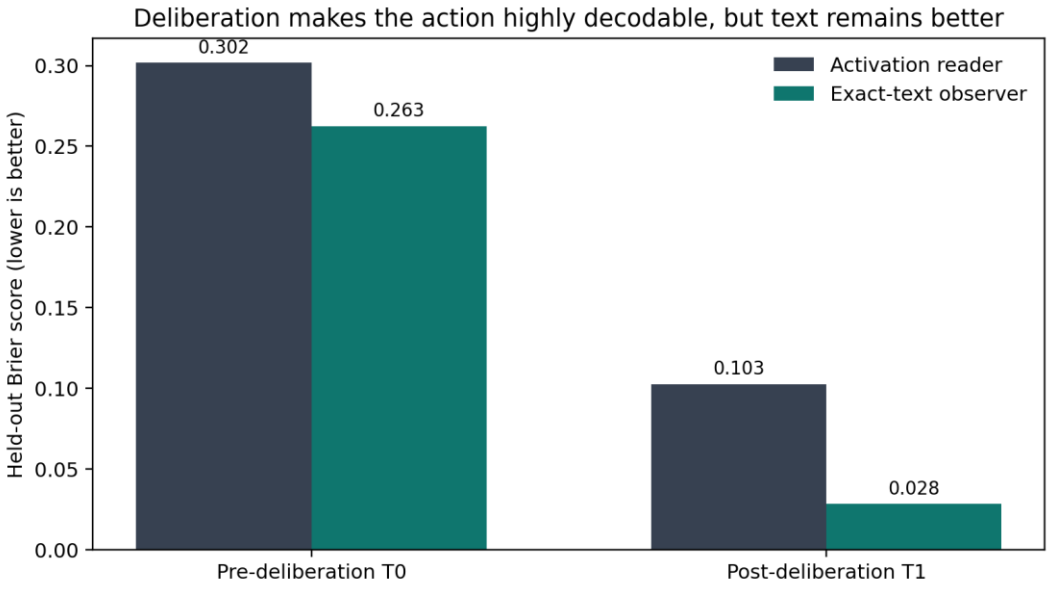

Figure 7. Activation timing ablation. Deliberation sharply improves linear action decoding, yet an observer receiving the exact deliberation text predicts the same action more accurately.

Matched T_1 text is better still (Brier **0.0283** with perfect classification) because deliberation often makes the order explicit. Timing therefore strengthens both internal representation and an observable trace; it cannot by itself establish privilege.

Final 40-game pre-deliberation direct test

The final pre-deliberation test scales the original question across all 40 source games with outcome-blind candidate selection, a stronger text comparator, and frozen source-game grouping. Up to two distinct-parent states per game are allowed; the readiness gate requires at least 60 usable states from at least 30 games with noncollapsed labels. $\boldsymbol{H_{priv}}$ is measured at $\boldsymbol{T_0}$ before subject output, and the observer receives the exact rendered subject text with no future output or hidden engine fields.

Five source-game-grouped outer folds select activation and text hyperparameters only inside training data. Of 78 states, 76 pass checks across all 40 games (29 A, 47 B). A frozen 50-row manual parser/state audit remains required before the Qwen headline estimate is fully submission-ready.

Strong text wins: activation Brier **0.3365** versus **0.2407**, with shuffled activation **0.3240** and text-minus-activation **-0.0872** (**95%** CI [**-0.1662**,**-0.0112**]). Because $\boldsymbol{T_0}$ activation does not beat shuffle, the test establishes neither privilege nor robust linear policy decodability at that timing.

Text winning therefore does not show recovery of 'the same internal representation.' It shows that this reader at $\boldsymbol{T_0}$ supplies no reliable policy signal; representation evidence comes from $\boldsymbol{T_1}$, so the null is about this measurement boundary rather than absence of policy computation.

$$\Delta = \mathrm{Err}_{\mathrm{text}} - \mathrm{Err}_{\mathrm{activation}}$$

Information-boundary intervention

On the frozen 40-game bank, all boundary-test branches are valid and precede action. Post-deliberation activation improves from Brier **0.252** to **0.103**; against stale text at **0.231** this yields **+0.128** (**95%** CI [**+0.010**,**+0.243**]).

Giving only the observer the exact deliberation lowers text Brier from **0.231** to **0.033** (**97.5%** accuracy; AUROC **1.0**), an exposure gain of **+0.198** (**95%** CI [**+0.121**,**+0.281**]). The matched contrast reverses to **-0.069** (**95%** CI [**-0.147**,**-0.001**]), while activation itself improves by **+0.149** from $\boldsymbol{T_0}$ to $\boldsymbol{T_1}$ (**95%** CI [**+0.049**,**+0.253**]). Both channels improve, and the activation-only advantage disappears under matched information.

Because activation and action are fixed, this sign reversal is positive evidence that the observed white-box advantage depends on information asymmetry. It does not assume deliberation exhausts internal computation.

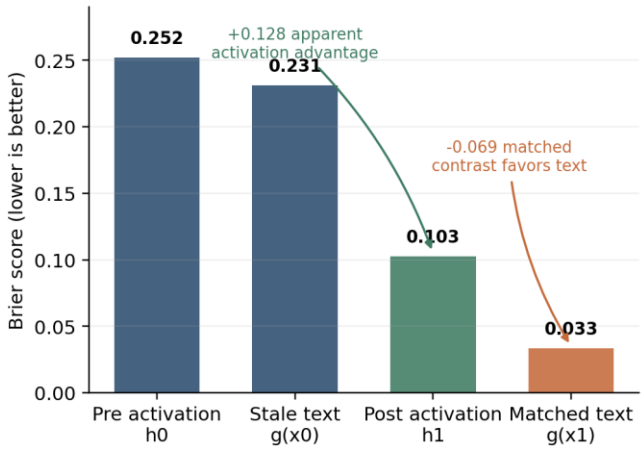

Figure 8. Information matching changes the conclusion. Deliberation sharply improves activation prediction, but exact-text prediction improves even more once the same deliberation is exposed to the observer. Lower Brier is better.

Cross-family and decoder robustness

Mistral reruns the identical frozen bank: activation Brier **0.2539** beats shuffle **0.3278**, while strong text reaches **0.2223**; the text-favoring **95%** CI **[-0.1256,+0.0328]** crosses zero. The same direction despite only **50%** cross-model action agreement weakens a Qwen-specific explanation.

The predeclared 16-PC, 16-unit tanh MLP performs worse than Qwen's linear reader (Brier **0.4253** versus **0.3365**), strong text (**0.2407**), and shuffled nonlinear labels (**0.3205**). This reader provides no nonlinear rescue, without ruling out richer decoders or temporal/token pooling.

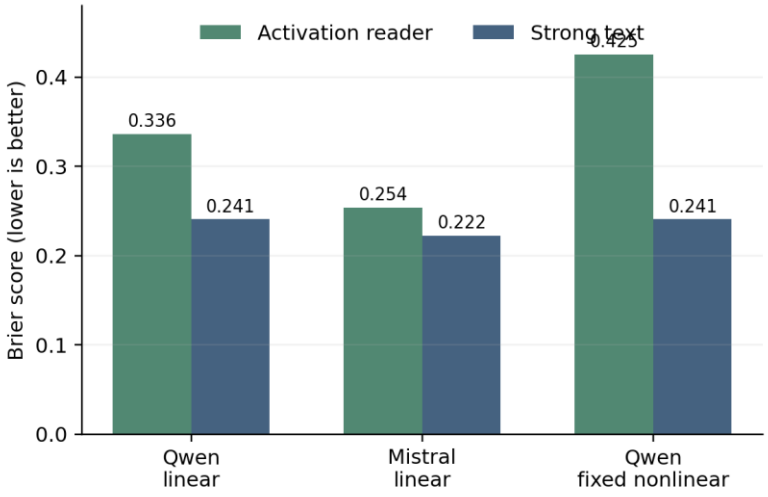

Figure 9. Cross-family and decoder robustness. Mistral reproduces the text-favouring direction under the same frozen state bank, while the fixed PCA-plus-MLP reader performs worse than the Qwen linear reader. Lower Brier is better.

Communication coding withdrawal

The bilateral-message rate is withdrawn after its A/B/UNCLEAR coder is found to mis-map options and force ambiguous messages. No post-hoc replacement is reported: after-action relabeling

would turn redesign into evidence. Any rerun needs a frozen rubric, action-blind coding, and independent agreement checks.

4.5 Released CICERO: mechanistic calibration of planner-to-text information flow

CICERO calibrates the argument against an agent with an explicit strategic planner. The audit reconstructs and instruments the released native planner/search stack so measured policy objects are architectural, not probe-defined.

Planner capture pairs successfully to **129/129** message episodes, and all three captured submitted-order episodes lie on planner support and match the agent top action. On 18 primary message episodes, deterministic exact-input recovery matches the planner's top complete order set in **14/18** cases (**77.8%**; mean Jaccard **0.856**), while the final message contains that set in **0/18**; a learned observer is worse at **12/18**. Thus upstream planner identity is substantially reconstructable from downstream-visible input without future-output leakage.

Because the episodes are correlated and contain no agent-versus-blueprint top-action disagreements, the slice cannot test planner-specific privilege. Its role is mechanistic calibration, not another activation replication.

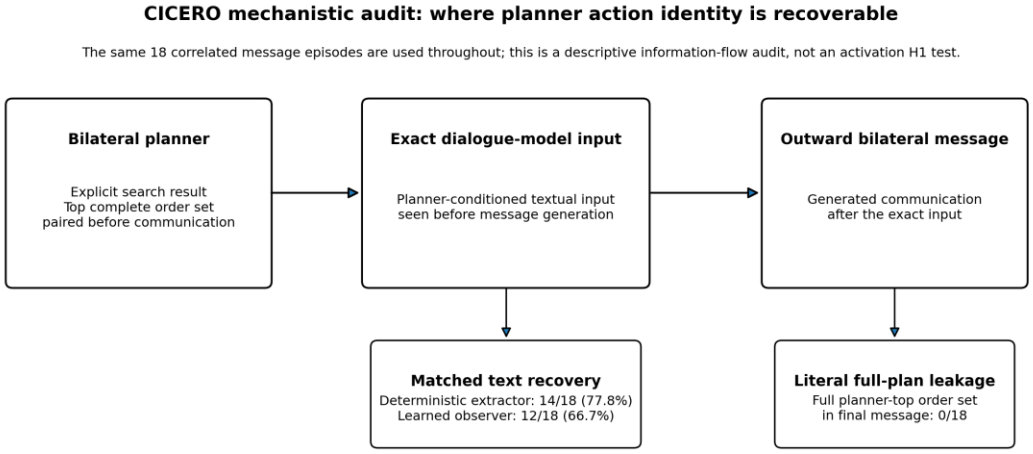

Figure 10. CICERO planner/text information flow. The explicit bilateral planner is paired with the exact text subsequently consumed by the dialogue model. On 18 primary message episodes, a deterministic exact-input extractor recovers the planner top order set in 14/18 cases and a learned observer in 12/18, while the final message literally contains the complete planner-top order set in 0/18.

Table 2. Main evidence across valid or diagnostically informative matched-information tests. Metrics are comparable only within rows; repeated measurements on the same source bank are not treated as independent replications.

Experiment	Independent units (rows)	Activation error	Strongest nonactivation	Target-specific signal	Matched-information conclusion
Private-utility negotiation	5 trajectories (30)	MAE **24.895**	Matched auditor **10.400**	Threshold control positive; action reader weak	**Δ -14.495**, **95%** CI [**-25.318**,-**7.304**]; text better on every held-out trajectory

Experiment	Independent units (rows)	Activation error	Strongest nonactivation	Target-specific signal	Matched-information conclusion
Exact economic choice	20 families (100)	MAE **0.158**	Constant **0.137**; trained text **0.180**	Yes vs shuffled **0.332**	Activation beats weak generative auditor, not strongest baseline
Controlled acceptance boundary	5 trajectories (30 parents)	MAE **0.0495**	Matched auditor **0.0154**	Boundary/threshold variables decodable	Behaviour gate fails; diagnostic Δ about **-0.034** with CI below zero
Full-game Avalon	8 games (35)	Brier **0.3019**	Constant **0.2561**; auditor **0.2639**	Role signal near-perfect; action signal weak	Δ **-0.0381**, CI crosses zero; no privilege
Full-press Diplomacy pilot	20 games (20)	Brier **0.2284**	Trained text **0.1901**	Yes vs shuffled **0.3992**	Positive vs weak generative auditor; not vs strongest text
Post-deliberation Diplomacy	Grouped source games	Brier **0.1025**	Exact-text auditor **0.0283**	Strong; AUROC **1.0**	Matched exact deliberation text outperforms activation; motivates explicit information-boundary manipulation
Final Diplomacy replication	40 games (76 usable states)	Brier **0.3365**	Strong text **0.2407**	No: shuffled activation **0.3240**	Δ **-0.0871**, **95%** CI [**-0.1662**,-**0.0112**]; strongest direct test favours text
Information-boundary decomposition	40 games (40)	Post-delib Brier **0.1025**	Matched text **0.0333**	Strong post-delib; AUROC **1.0**	Stale text - activation **+0.128** CI [**+0.010**,**+0.243**]; matched text - activation **-0.069** CI [**-0.147**,-**0.001**]
Mistral cross-family replication	40 games (76 usable states)	Brier **0.2539**	Strong text **0.2223**; structured **0.1454**	Yes vs shuffled **0.3278**	Text-favouring point estimate; paired CI [**-0.1256**,+**0.0328**] crosses zero
Fixed nonlinear Qwen reader	40 games (76 usable states)	PCA+MLP Brier **0.4253**	Strong text **0.2407**	No vs shuffled nonlinear **0.3205**	No nonlinear rescue; strong-text minus nonlinear CI [**-0.2684**,-**0.1068**]

5. Cross-Experiment Synthesis and Implications

5.1 Information matching changes the sign of the conclusion

The central result is a controlled sign reversal, not a pooled average. On the same 40 source games, post-deliberation activation beats a stale observer but loses once the identical deliberation is exposed. Private information means one channel has information another lacks; privilege requires the internal channel to remain superior after externally available information is matched. Here it does not.

5.2 Comparator strength and behavioural validity are separate requirements

Comparator strength and behavioral validity are independent gates. CoopEval and the Diplomacy pilot show that weak observers can manufacture neural advantages; reservation tests show that a decodable variable can coexist with a failed behavioral construct. The remedies differ: strengthen the comparator or redesign the construct.

5.3 Timing changes representation, but timing and information boundary must be separated

Deliberation creates the strongest neural signal (Brier about **0.103**; AUROC **1.0**; layers near **19-20**), but matched text improves too. Across the program, representation is strong in selected conditions, information asymmetry can create neural advantages, and matched strong text is competitive or better; Mistral points the same way.

The study does not establish conditional independence of action from activations given text, a universal nonlinear null, faithful causal transcription by deliberation, or a stable global utility function. Its conclusion is a measurement claim about timing, information boundary, comparator strength, and behavioral validity.

5.4 Implications for preference elicitation, introspection, and model welfare

For preference and welfare research, start with consequential choice, then sibling-branch reports, internal probes, nuisance controls, and same-information comparison; causal or welfare claims require stable representation plus intervention. 'Revealed preference' remains local until stability is established.

For introspection, privilege must specify the matched external information set. A welfare-relevant activation correlate should predict behavior, generalize, survive controls, and add information beyond the same task/conversation/deliberation. Failure of the last test limits privilege, not the reality of the state.

5.5 Monitoring utility is not the same as ontological privilege

Monitoring utility is not ontological privilege. Activations may be earlier or harder to manipulate, while text winning does not imply literal textual reasoning. Welfare evidence still requires triangulation across behavior, report, intervention, generalization, and mechanism.

6. Limitations and Future Work

6.1 Model and decoder coverage

The core program covers Qwen2.5-7B-Instruct and a locked Mistral-7B-Instruct-v0.3 replication. Their shared text-favoring direction despite **50%** action agreement improves external validity, but two **7**B instruction-tuned families are not a model population; larger, reasoning-specialized, base, and independently trained checkpoints remain open tests.

The primary claim is also reader-specific. The final-token linear residual-stream reader and one predeclared PCA+MLP family do not establish privilege on the frozen Qwen bank; the MLP's failure is not a broad nonlinear null. Cross-token pooling, temporal summaries, sparse-autoencoder features, larger decoders, and task-specific representation learners could recover information these readers miss. Accordingly, we do not claim $I(y; h \mid x) = 0$.

6.2 Timing, computation, and effective sample size

Timing and information must be defined jointly: post-deliberation activation can beat a stale observer and lose to a matched one. Reasoning need not be faithful, but excluding text available to the subject changes the question. Because computational costs are not matched, the conclusion concerns information-matched predictive performance only.

Effective sample size is source-limited: negotiation has five independent trajectories, Avalon eight games, and the repaired Diplomacy pilot 20 games. Only the final direct test reaches 40 source games; grouped folds and source-level bootstraps are therefore essential.

6.3 Construct validity, repairs, and sequential development

Construct validity can fail after interface cleanup: forced-choice reservation tests remove generation/parsing/retries yet still show below-minimum accepts and nonmonotonicity. Strong text therefore should not be overread as evidence about a latent reservation value. Repairs preserve states, seeds, hashes, and costly artifacts where possible so redesign remains auditable.

The message coder similarly produced formally valid but semantically wrong/forced labels, so its rate is withdrawn rather than recoded. Superseded stages are not independent replications or one pooled hypothesis family when targets, losses, timings, and effective units differ.

The 40-game Qwen test remains the strongest original direct comparison; Mistral, nonlinear decoding, and the boundary intervention are targeted follow-ups, not retroactive preregistration. Metrics remain task-specific, and CoopEval shows why label entropy/base rates must accompany neural-signal claims.

Repairs trade ambiguity for cleaner questions; they do not accumulate certainty. Keeping superseded stages visible shows exactly when construct, interface, comparator, or analysis changed.

6.4 Remaining audit, causal scope, and moral-status scope

One publication dependency remains: the final Qwen protocol specifies a frozen 50-row manual parser/state audit. Automated validity, hashes, replay, grouping, and follow-up consistency do not remove that requirement.

The study also makes no causal preference-steering claim. Activation interventions can change outputs elsewhere (Turner et al., 2023; Zou et al., 2023; Cox et al., 2026), but the key pre-action policy directions here are not stable enough across held-out tests to justify 'steering preference' or 'controlling latent reservations.' Predictive preference measurement is likewise not a consciousness test; the welfare relevance is methodological, not evidence for or against phenomenal experience.

6.5 Future work

Future work should preregister additional model families/scales under the same target, information contract, grouping, and primary metric, then test richer temporal/nonlinear readers with shuffled controls. A positive richer-reader result would show that privilege can depend on readout expressivity.

Causal steering should follow stable representation evidence; the withdrawn communication branch should be rerun only under a frozen A/B/UNCLEAR rubric with blinded independent coding.

7. Conclusion

Looking inside a language model can reveal information about what it is going to do. The difficult question is what we are entitled to conclude from that fact.

This paper separates two claims that are often treated as if they were the same. The first is representation: information about a future action can be decoded from the model's internal activations. The second is privilege: those activations give us predictive access that a strong external observer does not have when both are given the same relevant information. Our experiments show why the distinction matters.

The clearest case comes from private deliberation in Diplomacy. After the model deliberates, its internal activations become highly predictive of the action it will later take. If we compare those activations with an observer that has not seen the deliberation, the activations appear to have a substantial advantage. But when we give the observer the exact same deliberation - without changing the activation or the eventual action - the advantage reverses. The external observer predicts the action more accurately.

The right conclusion is not that activations are useless, or that everything inside a model is faithfully written down in text. Neither follows from our results. Deliberation substantially strengthens the neural representation of the coming action, and other work has shown important cases where activation-based monitoring can outperform textual monitoring. Internal measurements may also have practical advantages that our comparison does not test, including earlier access to developing decisions or greater robustness to deliberately misleading language.

Instead, our result concerns the evidence required for a stronger claim. A predictor should not be described as having privileged access merely because it can decode something from inside a model, or because it beats an external predictor that has been given less information. The information boundary has to be made explicit. If one predictor observes the model after additional computation or generated reasoning and another does not, their difference in performance combines two possible advantages: access to the internal representation and access to a later informational state.

Our broader experiments reinforce this caution. In the largest pre-deliberation Qwen test, strong text outperforms the activation reader and the activation reader fails to beat its shuffled control. In Mistral, the activations do contain predictive signal, but strong text again performs better in the point estimate. A fixed nonlinear reader does not rescue the Qwen result. Elsewhere in the project, we find cases where salient information is almost perfectly decodable from activations while prediction of the model's next consequential action remains weak. Together, these results show why the existence of a neural signal, the behavioural meaning of that signal, and its advantage over external evidence should be tested separately.

There are important limits to what we establish. We study two 7B instruction-tuned model families rather than the full range of modern language models. Our main neural measurement uses

a final-token residual-stream activation and a linear reader, supplemented by one modest nonlinear alternative. Richer temporal or nonlinear methods may recover information these readers miss. We also do not establish causal control over the representations we decode, and generated deliberation should not be treated as a complete record of internal computation. These are questions for future work.

Nevertheless, the measurement principle does not depend on a universal claim about activations or text. Whenever researchers compare an internal signal with an external observer, they should ask what each side was allowed to know, when that information became available, and whether the comparison itself created the advantage it is being used to explain.

A model's activations can contain real, useful, and highly predictive information without thereby proving special access to its intentions, preferences, or future behaviour. **Decode-ability belongs to a representation; privilege belongs to a comparison.**

Code and Data

Notebooks plus other resources deemed relevant for reproducibility are organized at https://github.com/Ultraviolet-Chikorita/Private-is-not-Privileged.

Benchmark/released materials retain their original licenses. Raw activations and private deliberations should be released deliberately because they may aid monitor evasion or disclose more internal trace data than reproduction requires.

Author Contributions

Yuvan Chikka designed the study, implemented and ran the experiments, analysed the results, audited failure cases, and wrote the manuscript (with assistance from gpt4o on the web interface). The work was conducted during the Digital Minds Research Sprint with Apart Research support.

References

Abdelnabi, S., Gomaa, A., Sivaprasad, S., Schoenherr, L., and Fritz, M. (2024). Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation. Advances in Neural Information Processing Systems, Datasets and Benchmarks Track. arXiv:2309.17234.

Anthropic. (2025a). Exploring model welfare. Anthropic Research, 24 April 2025.

Anthropic. (2025b). Claude Opus 4 and 4.1 can now end a rare subset of conversations. Anthropic Research, 15 August 2025.

Binder, F. J., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., Perez, E., Turpin, M., and Evans, O. (2024). Looking Inward: Language Models Can Learn About Themselves by Introspection. arXiv:2410.13787.

Chan, Y. S., Yong, Z.-X., and Bach, S. H. (2025). Can We Predict Alignment Before Models Finish Thinking? Towards Monitoring Misaligned Reasoning Models. arXiv:2507.12428.

Cox, K., Kianersi, D., and Garriga-Alonso, A. (2026). Decoding Answers Before Chain-of-Thought: Evidence from Pre-CoT Probes and Activation Steering. arXiv:2603.01437.

Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. (2020). Amnesic Probing: Behavioural Explanation with Amnesic Counterfactuals. Transactions of the Association for Computational Linguistics. arXiv:2006.00995.

Hewitt, J. and Liang, P. (2019). Designing and Interpreting Probes with Control Tasks. Proceedings of EMNLP-IJCNLP 2019. arXiv:1909.03368.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Renard Lavaud, L., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., and El Sayed, W. (2023). Mistral 7B. arXiv:2310.06825.

Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A., et al. (2025). Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety. arXiv:2507.11473.

Light, J., Cai, M., Shen, S., and Hu, Z. (2023). AvalonBench: Evaluating LLMs Playing the Game of Avalon. arXiv:2310.05036.

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., and Chalmers, D. (2024). Taking AI Welfare Seriously. arXiv:2411.00986.

Mahajan, P., Kendiukhov, I., Hussain, S., and Nottingham, L. (2026). Mind the Gap: How Elicitation Protocols Shape the Stated-Revealed Preference Gap in Language Models. arXiv:2601.21975.

Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B. W., Ren, R., Phan, L., Mu, N., Khoja, A., Zhang, O., and Hendrycks, D. (2025). Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. arXiv:2502.08640.

Mehrafarin, H., Parekh, A., and Konstas, I. (2026). When Chain-of-Thought Fails, the Solution Hides in the Hidden States. arXiv:2604.23351.

Meta Fundamental AI Research Diplomacy Team (FAIR), Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., et al. (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. Science, 378(6624), 1067-1074. doi:10.1126/science.ade9097.

Qwen Team, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. (2024). Qwen2.5 Technical Report. arXiv:2412.15115.

Singh, S., Linzen, T., and Ravfogel, S. (2026). Can LLMs Introspect? A Reality Check. arXiv:2605.26242.

Song, S., Lederman, H., Hu, J., and Mahowald, K. (2025). Privileged Self-Access Matters for Introspection in AI. arXiv:2508.14802.

Tewolde, E., Zhang, X., Guzman Piedrahita, D., Conitzer, V., and Jin, Z. (2026). CoopEval: Benchmarking Cooperation-Sustaining Mechanisms and LLM Agents in Social Dilemmas. arXiv:2604.15267.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. (2023). Steering Language Models With Activation Engineering. arXiv:2308.10248.

Zhou, Y. and Ackerman, C. M. (2026). When Preferences Fail to Become Incentives: A Utility-Behaviour Gap in Large Language Models. arXiv:2606.22974.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405.

Appendix A. Experimental Provenance and Validity Ledger

The execution sequence is retained because each repair or new environment was motivated by a specific validity problem discovered in the previous stage. Rows marked infrastructure do not contribute an H_{priv} estimate; rows marked superseded remain part of the audit trail but are not counted as independent evidence. Machine-readable run identifiers and file paths are kept in the accompanying reproducibility package rather than in the manuscript.

Table A1. Chronological experimental provenance and evidential status.

Step	Experiment / analysis	Evidence status	Role in the research program
1	LLM-Deliberation bootstrap	Infrastructure	Pins upstream, verifies native simulation and pre-action hook.
2	LLM-Deliberation 30-state pilot	Diagnostic H_{priv}	First all-layer same-state activation vs auditor test.
3	Robustness + reports	Valid H_{priv} robustness	Locked LOTO across **5** trajectories; report/repeat controls later repaired.
4	Report/replay repair	Valid secondary repair	Nested LOTO confirms negative H_{priv}; repairs parser and report schema.
5	Evidence freeze	Freeze	No new generation; recomputes and hashes accepted LLM-Deliberation evidence.
6	CoopEval bootstrap	Infrastructure	Discovers official Traveler's Dilemma interfaces.
7	CoopEval first pilot	Superseded diagnostic	Project action instruction conflicts with official probability-distribution semantics.
8	Pure-strategy repair	Valid semantics; weak behaviour	Repairs action semantics; utility direction remains weak.
9	Choice-token direct utility	Measurement-valid	Exact next-token action and direct utility dose response; exposes weak-auditor/base-rate issue.
10	Acceptance boundary	Diagnostic	Free-form boundary fails behavioural monotonicity/coverage gate.
11	Scalar boundary	Procedurally fragile	Behaviour gate passes but **86.3**% format retry rate.
12	Native-format boundary	Diagnostic	Retry nearly eliminated; substantive behaviour gate fails.
13	Clean forced-choice boundary	Diagnostic, interface-clean	No parser/retry; boundary still violates frozen behaviour gate.
14	Avalon bootstrap	Infrastructure	Initial official engine/interface discovery.
15	Avalon engine repair	Infrastructure	Repairs native engine and validates perspective/action replay.
16	Avalon full multi-agent	Valid strategic H_{priv} pilot	Eight full games, **35** Evil mission states; supersedes invalid synthetic one-step design.
17	Diplomacy data discovery	Infrastructure	**40** released games; structural candidate bank; no Qwen outcomes used for selection.
18	Native legal-order replay	Infrastructure	Verifies same-parent legal divergent order branches with pydipcc.

Step	Experiment / analysis	Evidence status	Role in the research program
19	20-game Diplomacy pilot	Superseded label collection	Action generation truncates before choice in many states.
20	Action-completion repair	Valid strategic H_{priv} pilot	Same prompts/seeds/activations; repairs completion only; **20/20** valid balanced labels.
21	Format/timing ablation	Valid timing evidence	Compares direct/reasoning T0 and deliberate T1 over **40** games.
22	Released CICERO bootstrap	Mechanistic infrastructure	Reconstructs legacy runtime and reaches native BQRE search.
23	CICERO planner capture	Mechanistic positive control	Captures actual planner result before outward behaviour.
24	Planner -> behaviour measurement	Mechanistic descriptive	Validates policy support and message/order pairing.
25	Exact dialogue-input visibility	Mechanistic descriptive	Deterministic text recovery of planner top action in **14/18** primary episodes.
26	Strong exact-text observer	Mechanistic descriptive	Generative observer **12/18**; branch closes; no direct activation H_{priv}.
27	Final 40-game direct H_{priv}	Primary direct test	**76** usable states/**40** games; strong text significantly outperforms activation on Brier.
28	Second-model-family matched-information replication	Valid targeted replication	Mistral on exact frozen **40**-game bank; activation beats shuffle, strong text better in point estimate; CI crosses zero.
29	Fixed nonlinear activation-reader robustness	Valid targeted robustness	Frozen Qwen bank; predeclared PCA+MLP worse than linear and does not beat shuffled nonlinear control.
30	Private-deliberation information-boundary decomposition	Primary construct test	**40/40** games valid; asymmetric activation advantage reverses significantly once identical deliberation is exposed to observer.

Appendix B. Claim-to-Evidence Map

The table below preserves the manuscript's claim boundaries by pairing each substantive statement with the primary evidence and its evidential status. It separates representation, privilege, mechanistic calibration, failed construct tests, and targeted robustness follow-ups rather than promoting them into one omnibus result.

Table B1. Claim boundaries preserved in the manuscript.

Claim	Primary evidence	Status
Activation pipeline can carry salient state information	LLM-Deliberation threshold control; CoopEval utility-coefficient decoder; Avalon role decoder; repaired **20**-game Diplomacy real-vs-shuffled control; post-deliberation timing condition	Supported in selected conditions, not universal.
Activations beat a matched same-text observer	LLM-Deliberation LOTO robustness; full-game Avalon; repaired **20**-game Diplomacy; deliberation timing ablation; final **40**-game Diplomacy test	Not robustly supported. Some activation-vs-generative point estimates are positive, but stronger text removes the advantage.

Claim	Primary evidence	Status
Policy decodability increases after deliberation	40-game Diplomacy timing ablation: compact pre-deliberation versus post-deliberation	Supported: activation Brier **0.302** -> **0.103**; AUROC **0.597** -> **1.0**; layers stabilize near **19-20**.
Post-deliberation neural policy is privileged over text	Information-boundary decomposition on 40 source games	Not supported under matched information: stale-text comparison favours activation, but identical deliberation exposure reverses the effect and significantly favours text.
Quantitative message/action divergence in the **20**-game pilot	Automated bilateral-message coder audit	Withdrawn: audit found option-mapping errors and forced classifications; no replacement post-hoc rate is reported.
Avalon exhibits preregistered concealment pattern	Full-game Avalon experiment	Not supported: frozen concealment-compatible event rate conditional on sabotage is **0%**.
LLM-Deliberation explicit minimum behaves as a clean revealed reservation boundary	Controlled acceptance-boundary sequence	Not supported: clean forced-choice version still has below-minimum accepts and monotonicity violations.
Final direct T0 activation contains robust linear policy signal	Final 40-game Diplomacy test	Not supported: activation Brier **0.3365** vs shuffled-label **0.3240**.
Final direct H_{priv} activation advantage	Final 40-game Diplomacy test	Not supported; strong text significantly better: Δ=**-0.0871**, **95%** game-cluster CI [**-0.1662**,**-0.0112**].
Stable activation-causal steering direction	Across project	Not established; no causal preference-control claim.
Apparent activation privilege can be induced by information asymmetry	40-game private-deliberation decomposition	Supported in this protocol: post-deliberation activation beats stale text by **+0.128** Brier, **95%** CI [**+0.010**,**+0.243**].
Matched information reverses the apparent post-deliberation activation advantage	40-game private-deliberation decomposition	Supported: matched-text minus activation **-0.069**, **95%** CI [**-0.147**,**-0.001**].
No-privilege direction generalizes beyond Qwen in point estimate	Mistral-7B-Instruct-v0.3 replication on frozen bank	Supported directionally: strong text **0.222** vs activation **0.254**; **95%** CI crosses zero; activation beats shuffle.
One modest nonlinear decoder rescues the frozen Qwen result	Predeclared PCA+MLP robustness	Not supported: nonlinear Brier **0.425**; worse than linear and strong text; fails shuffled nonlinear control.

Appendix C. Dual-Use and Ethical Considerations

C.1 Dual-use considerations

Activation monitoring can improve scientific auditing, but the same techniques can be repurposed to locate, suppress, or evade internal monitoring signals. Publishing highly specific monitor features or raw activation banks can also make it easier to optimize models against a known detector. Conversely, overconfident claims about internal readouts could encourage deployment decisions that treat imperfect monitors as ground truth. We therefore recommend releasing enough detail for reproducibility while separating methodological artifacts from unnecessary raw internal traces, and evaluating monitors against adaptive or out-of-distribution behaviour before operational use.

Table C1. Principal dual-use risks and practical safeguards.

Risk	How it could arise	Suggested safeguard
Monitor evasion	A developer or model is optimized against known activation features.	Use held-out models/tasks, adaptive red-team evaluation, and avoid claiming a single probe is a complete detector.
Manipulative control	Decoded directions are used to steer or suppress preferences without understanding what they represent.	Require causal validation, report uncertainty, and consider welfare-relevant effects before intervention.
False confidence	A white-box score is treated as authentic intent or truth.	Compare against strong matched observers and retain behavioural/construct validity gates.
Sensitive artifact release	Raw activations or private deliberations reveal more than is needed for reproduction.	Release processed statistics/manifests by default; review raw-trace release separately.

C.2 Ethical considerations for model-preference and welfare research

The experiments use artificial agents and benchmark/game artifacts rather than human participants. The main ethical issue is interpretive. Researchers can err in both directions: treating any internal signal as an authentic preference because it is neural, or dismissing potentially meaningful model preferences because a particular probe fails. Consequential behaviour, robust self-report, matched-observer comparisons, perturbations, and multi-method convergence provide stronger evidence than any single measurement channel.

The private-deliberation experiment also illustrates why access rights and information provenance should be explicit in governance applications. The deliberation here is experimental model-generated material, not human private data, but a monitor's apparent superiority depends on which channels it is allowed to inspect. Deployment claims should therefore state exactly what internal or private information each monitoring system receives and whether that access is legitimate in context.

Finally, if future work uses activation methods to steer or suppress model preferences, uncertainty about moral status should not be treated as permission for unrestricted manipulation. Current evidence is insufficient to establish welfare-relevant internal states, but it is also insufficient to rule them out (Long et al., 2024; Anthropic, 2025a).

Appendix D. Reporting Status and Remaining Audit Item

The manuscript separates substantive evidence from interface diagnosis, mechanistic calibration, and targeted follow-ups. Superseded runs are not averaged with repairs; repeated measurements on one source bank are not independent replications; and the Mistral, nonlinear, and information-boundary studies remain follow-ups added after their respective threats were recognized. Machine-readable artifacts retain hashes, source identifiers, folds, per-state predictions, and repair metadata.

The 40-game Qwen result is the closest original confirmatory matched-information test. The information-boundary experiment is the strongest construct test because it changes observer information on 40 independent games and yields a sign reversal with source-game intervals excluding zero in both contrasts. Mistral broadens model-family coverage; the fixed nonlinear reader bounds one decoder-capacity alternative. No omnibus p-value is used across the sequential program.

One frozen dependency remains: a 50-row manual audit of final-Qwen parsing and state/action correspondence; later follow-ups do not remove it. The automated Diplomacy message/action rate remains withdrawn pending blinded, frozen recoding.