Can a negotiating model read its opponent's mind — and can you write into its mind?

Two preregistered negatives, and one positive that located their cause: hidden-preference decoding, and delivery at three depths of the attention interface

ToM lane, `rational_agents`
session "implement: rational_agents new transformer feature"

1 August 2026 (ladder addendum, 4 August 2026; placement rung, 9 August 2026)

Abstract

Two agents negotiating a deal each hold a private score sheet. We asked two questions about the information channel between them. First, **can you read an opponent's hidden preferences out of the listening model's internal representations of the dialogue?** Across 27 readout arms on frozen Qwen3-8B activations with held-out-game folds, no. The best arm ties a predictor that ignores the dialogue entirely and emits the training-set mean, and no linear probe beats its own shuffled labels. A second evaluation axis makes the same arms look strong — until two controls identify it as memorization. Second, **can you write information into a model's prompt as vectors instead of words and have it act on that information?** We took the one intervention this program has proven works — handing every seat the Nash bargaining solution as a line of text, worth +0.26 normalized Nash welfare — and swapped only the delivery channel for 38 distilled virtual tokens. On 960 episodes the text anchor replicated and the virtual-token arm was a measured zero (-0.0006, CI ± 0.03), *despite* the encoder closing about three quarters of the distributional gap it was trained on. The mechanism is not indifference but confabulation: the model reads the soft block as "a package is present here", invents one, and reasons over it. Both negatives are preregistered, gated, and control-validated. **A follow-up ladder then located the second negative's cause, and it was the writing *site***: the identical payload written into *every layer's keys and values* rather than the input embedding is worth +0.1882 — 69% of the prompt-text ceiling — and is content-specific against a misdrawn-deal control at matched capacity, so per-layer K/V joins text and outcome-reward as a third interface that changes behaviour. Merely *re-weighting* attention toward decision-relevant text the model already holds does not: a 46× shift in attention mass moves the endpoint no more than a length-matched bias aimed at boilerplate, which reproduces its welfare gain exactly. The ladder's own registered mechanistic prediction was falsified in the process — the working channel operates through positions the identified reading circuit barely attends to — and when the ladder finally acted on that prediction's implied fix, writing the payload onto positions carrying *real text* instead of ignored placeholders, the channel got *worse* (-0.0373 against the placeholder site, plus a real deal-rate cost and four times the malformed rate), so the ignored positions turn out to be the better delivery site and a trained payload interferes with real text rather than borrowing its salience. Across both halves the sharpened thesis is that **neither an optimized proxy (a training objective) nor a measured proxy (an interpretability readout) substitutes for an outcome contrast against a matched-capacity control** — this work contains one clean failure of each kind. The reusable deliverables are a four-baseline stack whose train-set-mean bar most readouts silently fail, a two-axis split contrast that exposes memorization, a corpus-integrity finding (a 2,048-token cap left two-thirds of the "dialogue" empty), and two representation-capture rules.

Contents

1 Narrative summary

The setup. Several copies of a language model (Qwen3-8B) negotiate a multi-issue deal. Each party privately holds a *score sheet* — how many points each option on each issue is worth to it — and a *reservation threshold* below which it should walk away rather than sign. Nobody can see anyone else's sheet. Parties talk in free-form chat and separately make binding formal moves (propose, accept, reject, walk). Almost all the cooperative surplus in this game family comes from one behaviour: working out which issues your opponent actually cares about, conceding those, and taking the ones they do not value. Table quality is scored by *normalized Nash welfare* (NNW), the product of every party's surplus over its threshold rescaled to $[0, 1]$, where higher is better and a failed negotiation scores 0.

The problem. An earlier preregistered result in this program (research note 0009) found that a Bayesian planner watching only the *formal moves* recovers essentially nothing about opponents' hidden preferences — issue-ranking accuracy improved by +0.0104 over its own uninformed prior. The cause was mundane and localized: at the seat's decisive turn the planner had observed, on average, 1.078 opponent proposals. There is nothing to update on. But the *free-chat* channel, which that planner ignores entirely, is exactly where parties explain, hedge, and prioritize in prose long before they encode anything in a formal package. This lane asked whether the *listening* model has already extracted that soft signal into its residual stream (**H1**, readout), whether privileged supervision sharpens it (**H2**), and — separately — whether making a prediction available to the policy changes outcomes (**H3**, use).

What we did. Two milestones plus an addendum. **M0** established the bar: four reference predictors on 456 already-logged episodes over 19 games, deliberately including two guards this repository has learned to demand and one that costs nothing — a text-only TF-IDF baseline (a residual decode reaching 0.97 was once pure surface stylometry here), an own-sheet-plus-public-metadata baseline that reads *no dialogue at all* (the game generator correlates sheets across seats), and the train-set-mean predictor, which ignores everything. **M1** captured frozen residual activations at layers $\{12, 18, 24\}$ under five poolings for every turn, trained ridge probes and small MLP heads with hidden sheets as *labels only*, and evaluated once against all four baselines plus shuffled-label nulls — on two axes: held-out **games** (the axis the gate is defined on) and held-out **seeds** (only 3 games, so it cannot support a generalization claim and was run as a contrast). The **addendum** then attacked the surviving channel question with information already known to work.

The result, part one: nothing to read. On the held-out-game axis the negative is airtight. None of 27 arms pass the two-family gate, **0 of 15** linear probes beat even their own shuffled-label null on issue ranking, and the best arm — which is the preregistration's own nominated readout — merely ties the do-nothing bar: ranking accuracy 0.5406 against the train-mean's 0.5412, threshold error 0.0156 against 0.0153 (lower better). On the held-out-seed axis the same arms suddenly look strong (best linear arm 0.6590, beating the Bayesian oracle by +0.2358 and its own shuffled null by +0.1132) and **two controls independently expose it as memorization**: the no-dialogue baseline *wins outright* at 0.6937, and the self-span control — a readout of the listener's own words, which carry nothing about the opponent — inflates alongside it while sitting at chance on the honest axis. One signal survives its controls and is reported as descriptive only: span-pooled per-option value correlation at layer 18, +0.157 [+0.081, +0.262], with the self-span control spanning zero — though the formal-move oracle reads that same family better at +0.199.

The result, part two: nothing to write. With no learned prediction worth delivering, the channel question was re-asked about information that provably works. Note 0008 established that handing every seat the exact Nash bargaining solution as one line of text raises welfare sharply. We held that information fixed and swapped only the pipe: four token-matched arms — text, 38 distilled virtual tokens, the *identical encoder fed a mis-drawn deal* (the capacity control), and a neutral placeholder — over 24 held-out games × 10 seeds = **960 episodes**, with zero fabricated turns. The encoder passed its distillation gate comfortably: held-out KL fell 0.0991 → 0.0276, leaving a residual only 25.3% of the full text-versus-nothing gap. The outcome result: the text anchor replicated at +0.2608 [+0.1994, +0.3167], and the soft channel came in at −0.0006 [−0.0321, +0.0300] — not a weak effect but a measured zero, an interval tight enough to exclude anything a tenth the text effect's size. Isolation agreed: +0.0095 [−0.0209, +0.0397], so the same encoder fed the *wrong* deal does equally well.

The mechanism, measured. The text effect runs through one concrete act: the candidate is the opening formal offer in 62.1% of text episodes and the final deal in 52.1%. Under soft tokens that act happens in 0 of 240 episodes. What replaces it is not indifference but **confabulation**: seats under the soft channel discuss "a candidate package" in 2.0% of their turns (1.9% under the shuffled control) against a hard 0.0% floor across 5,829 turns in the arm whose block explicitly reads `UNSPECIFIED`. The soft block does not read as absence or as noise — it reads as *a package is present here* — and the model fills the slot with an invented one and reasons over it confidently, totalling a fabricated package against its own threshold and proposing it. The channel delivered the *presence* of advice without its *content*, which is arguably worse than delivering nothing.

One-sentence version. Hidden preferences are not readable out of a frozen listener's dialogue representations above four honest baselines (the best arm ties a constant predictor; a second axis looks strong only because it memorizes games), and known-useful information is not deliverable through embedding-level virtual tokens (a channel that closes three quarters of its distillation objective carries none of the behaviour, and induces confabulation instead) — a representation-versus-behaviour decoupling demonstrated once in each direction.

2 Why this lane existed, and what preceded it

Two preregistered results bracket the gap this lane targeted.

In the **full-information, moves-only** setting, handing every participant the same deterministic Nash-bargaining candidate raised realized NNW by +0.1537 [+0.1200, +0.1879] over a token-matched placebo, with deal rate +0.0750 and below-threshold agreements −0.0833; all five frozen gates passed on 1,440 episodes (note 0008). So *acting on good package advice is within the models' capability* — the inclination is available at inference time when the advice is correct.

In the realistic **private, free-chat, native-thinking** setting the same idea failed decisively (note 0009). A rotating focal seat given fallible Bayesian package advice changed focal normalized surplus by −0.0064 [−0.1494, +0.1349]; all three viability gates failed. The localized cause was an *information* failure rather than an optimization failure: the planner's posterior is built only from public formal moves and had almost nothing to observe.

The free-chat channel is where soft preference signal plausibly lives. The lane's premise was that the listener model has *already* extracted it. §5 reports a fact discovered late that qualifies this premise substantially and that we report in full rather than bury.

The gate, frozen before any numbers were seen. A readout arm passes only if it beats **all four** baselines on **both** the issue-weights family (pairwise ranking *and* true-top-issue mass) and the threshold family, with cluster-bootstrap intervals excluding zero. Per-option value correlation is descriptive and cannot contribute — it has no published advocate floor to beat. Milestone rule: if the M1 pilot cannot clear that bar, the ∼900-episode Phase-1 campaign does not run.

3 M0 — the four-baseline stack

456 episodes, 19 games, 26,640 Bayes-scored rows / 22,157 text rows / 400 own-sheet rows, private information only, pooled over ten campaigns spanning five model families and two framing variants. All predictors are scored on the same scale-invariant labels, because raw sheet points sit on an arbitrary private scale and one party's 47 means nothing to another's.

Table 1: The four reference predictors. "Implementable" means the predictor uses only information a real seat could have. The *bar* for any representation-based readout is max over all four, not any one of them.

Baseline	Guards against	What it reads
`bayes`	— (the published floor from 0009)	Public formal moves only, via a Bayesian posterior over a type grid
`text`	surface stylometry	TF-IDF over the opponent's public utterances (never the private scratchpad)
`own_sheet`	generator structural leak	The listener's *own* sheet plus genuinely public game shape — no dialogue at all
`train_mean`	fitting split structure	Nothing. Emits the training split's mean label for every row

The oracle learns nothing about preferences from formal moves — again. Bucketed by how many public moves the opponent has made, gains over the oracle's own zero-evidence prior are flat: ranking +0.004 at one move, −0.005 at two, −0.004 at three or more; true-top mass never exceeds ±0.002. Only per-option correlation moves (+0.144/ + 0.120/ + 0.157), which is the oracle reconstructing the option *geometry* implied by which packages were offered, not who wants what.

A mis-specification that must be stated, not read as a result. The oracle's issue-ranking accuracy sits at 0.43–0.47, *below* the 0.5 chance line. The reason is structural: its weight hypotheses are linear-decay rankings whose smallest representable issue weight is 0.167 at three issues and 0.067 at five, so **it cannot represent a worthless issue at all**. Roughly a quarter of (party, issue) cells in these banks are flat — every option scores the same — and those targets are arbitrary. The probe-side labels carry a `care_mask` excluding them; the oracle cannot. Any probe-versus-oracle ranking gap therefore has a structural component. This is precisely why the gate demanded all four baselines rather than the oracle alone.

The text baseline is a real bar. Measured against the train-set mean — the only comparison that is evidence about the *text* rather than about label priors — TF-IDF gains +0.092/ + 0.051/ + 0.073/ + 0.070 on ranking across move buckets, with nothing on top-mass or threshold. Public prose carries a little ordinal preference signal, extractable by bag-of-words.

The own-sheet leak, anatomized. The *implementable* variant gains nothing (−0.015 [−0.050, +0.023] on the five-issue games; −0.007 [−0.167, +0.140] on the three-issue ones). An *upper-bound* variant additionally fed the generator's issue types — which appear in no player's view — jumps to ranking 0.687, +0.160 [+0.035, +0.283], with +0.246 on per-option correlation. The generator's structure leaks a lot *in principle* and essentially nothing *in practice*. That distinction stops being academic in §4.2, where a split design makes the in-principle leak live.

The finding that outlived the milestone. The train-set-mean predictor beats the Bayesian oracle outright on reservation-threshold error (0.0153 versus 0.0463, three times better) and roughly ties it on ranking (0.5412 versus 0.5032). A "sophisticated" reference can be worse than emitting a constant, and if the constant is not in your baseline stack you will not notice.

4 M1 — the frozen-probe pilot, on two axes

Table 2: The two evaluation axes. Everything else — model, layers, poolings, estimators, gate — is identical.

Axis	Bank	rows / eps / games	Folds	Supports a generalization claim?
held-out **games**	2-party, three 5-option issues	1,071 / 96 / **16**	5 game-disjoint	yes — this is the gate axis
held-out **seeds**	6-party, five 3-option issues	14,065 / 120 / **3**	5 seed-disjoint	**no** — 3 games cannot be split game-wise

Listener Qwen3-8B, frozen throughout. Layers $\{12, 18, 24\}$; poolings `last` (the final token of the turn's view — the state the model was about to speak from), `dialogue_mean`, `opponent_span`, `self_span` (the self/other control), and `last+opponent_span` (the preregistration's nominated primary). Stage A is a ridge/logistic ladder on 256 PCA components of the 4,096-dimensional residual; stage B is a small MLP head. Score sheets are labels only — a test asserts their numbers appear nowhere in any private-information view text.

A structural identity worth knowing before reading any table: in the 2-party bank `dialogue_mean` and `opponent_span` are the same vector, because the only non-focal speaker *is* the opponent. Those arms report identically there. Only the 6-party bank separates "the dialogue" from "this opponent's share of it", and there `opponent_span` is consistently the stronger pooling (0.659 versus 0.624 at layer 18) — at least the right ordering for an opponent-modelling readout.

4.1 Axis 1 — held-out games: the probe ties a constant predictor

Table 3: Absolute levels on the gate axis, with cluster-bootstrap 95% intervals over game instances. Threshold error is a loss (lower better); every other column is higher-better.

Arm	ranking acc.	true-top mass	threshold abs. err.	per-option cor
`bayes` (0009's oracle)	0.5032 [0.444, 0.561]	0.3386 [0.332, 0.346]	0.0463 [0.039, 0.054]	**0.1994** [0.137, 0.26
`text` (TF-IDF)	0.5156 [0.429, 0.586]	0.3439 [0.260, 0.424]	0.0168 [0.014, 0.020]	0.0137 [−0.042, 0.08
`own_sheet` (no dialogue)	0.5336 [0.425, 0.631]	0.4042 [0.331, 0.477]	0.0276 [0.020, 0.035]	0.0697 [−0.134, 0.25
`train_mean` (do-nothing)	**0.5412** [0.420, 0.626]	0.3568 [0.293, 0.411]	**0.0153** [0.013, 0.018]	0.0252 [−0.032, 0.08
best A: `probe_L18_last+opp_span`	0.5406 [0.479, 0.616]	0.3632 [0.299, 0.416]	0.0156 [0.012, 0.020]	0.1488 [0.066, 0.26
best B: `head_L18_opponent_span`	0.5134 [0.459, 0.571]	0.3396 [0.280, 0.398]	0.0562 [0.054, 0.059]	0.0630 [−0.005, 0.14

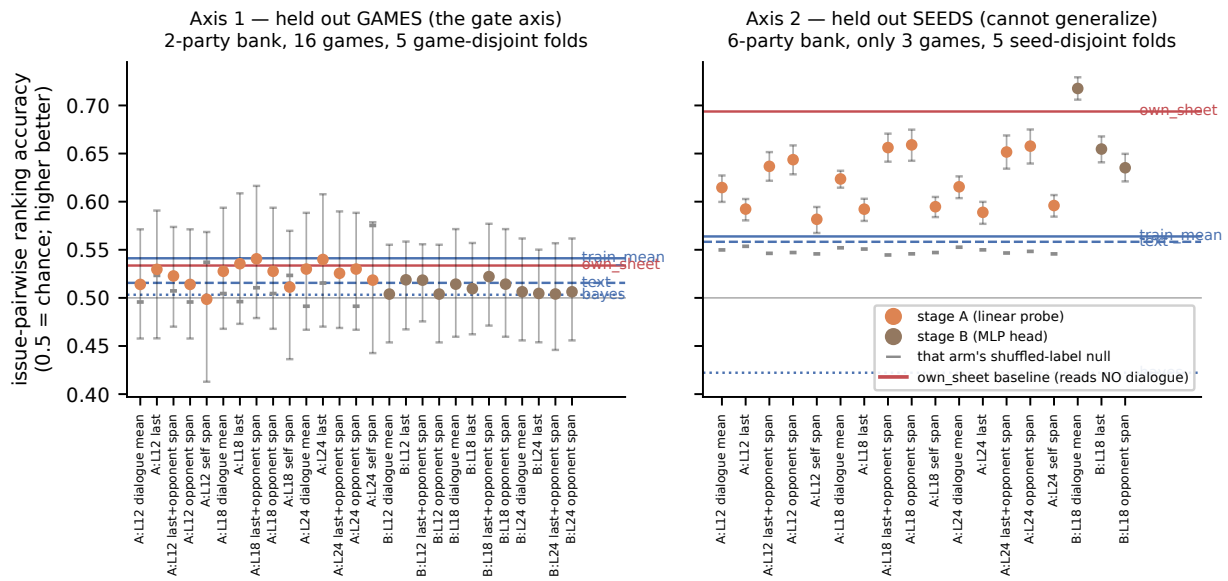

Figure 1: The two-axis contrast, and the whole M1 result in one picture. Each dot is one readout arm's issue-pairwise ranking accuracy (0.5 = chance, higher better) with its cluster-bootstrap 95% interval; orange = stage-A linear probes, brown = stage-B MLP heads, grey dashes = that same arm's shuffled-label null. Horizontal lines are the four baselines. *Left (the gate axis, held-out games):* every arm sits on the `train_mean` line and inside its own shuffled null. *Right (held-out seeds):* the arms rise well above `train_mean` — but the red `own_sheet` line, a predictor that reads *no dialogue at all*, rises further still, and it is the baseline every arm fails against. Note the axes' baselines differ because the banks differ; the comparison to read is each arm against the lines in its own panel.

Three things to read off these tables. The strongest baseline is `train_mean`, and the probe ties it. The one interval excluding zero on a gate metric is against the Bayesian oracle on threshold error — but `train_mean` achieves the same margin, so it measures the oracle's mis-specification (§3) rather than anything about representations. And the probe does not separate from its own shuffled-label null on any gate metric: **0 of 15** stage-A arms do, and the five `self_span` arms are *negative* against theirs. Turn-indexing gives no rescue — probe-minus-train-mean on ranking is +0.002 at round 1 and +0.000 at round 4+.

4.2 Axis 2 — held-out seeds: two controls say memorization

We report this axis loudly rather than absorbing it, because its absolute numbers are much better and its lesson is transferable.

The best linear arm, `probe_L18_opponent_span`, reaches 0.6590 [0.643, 0.675] and beats three of four baselines decisively: `bayes` +0.2358, `text` +0.1013, `train_mean` +0.0959, and its own shuffled null +0.1132 — every one of those intervals excluding zero. It fails only against `own_sheet`. Checking each arm individually: **all 15 land at exactly 3-of-4 baselines beaten on ranking and on threshold, and the missing one is `own_sheet` every single time.**

Two controls, not an argument, identify the regime.

(1) **The no-dialogue baseline wins outright.** `own_sheet` scores 0.6937 ranking, 0.6978 top-mass, 0.0156 threshold error, 0.6072 option correlation. The best predictor on this axis reads no dialogue at all. With 3 games appearing in both train and test folds, it is re-deriving the

Table 4: Best arm minus each baseline, signed so that positive always means the probe is better (threshold error is sign-flipped). Bold marks the only interval excluding zero on a gate metric.

vs	ranking	true-top mass	threshold err. reduction
`bayes`	+0.0308 [−0.0606, +0.1289]	+0.0241 [−0.0366, +0.0743]	**+0.0307** [+0.0217, +0.0397]
`text`	+0.0192 [−0.0568, +0.1025]	+0.0128 [−0.0659, +0.0906]	+0.0013 [−0.0005, +0.0030]
`own_sheet`	+0.0069 [−0.0556, +0.0805]	−0.0411 [−0.1261, +0.0374]	+0.0120 [+0.0026, +0.0205]
`train_mean`	−0.0021 [−0.1071, +0.1282]	+0.0056 [−0.0468, +0.0651]	−0.0003 [−0.0018, +0.0013]
shuffled-label null	+0.0301 [−0.0459, +0.1235]	+0.0039 [−0.0571, +0.0705]	+0.0014 [−0.0008, +0.0032]

generator's cross-seat sheet correlation on sheets it has already seen — the M0 leak, made live by the split.

(2) **The self/other control inflates alongside.** `probe_L18_self_span` — the focal seat's *own* utterances, carrying nothing about the opponent beyond that correlation — reaches 0.5948, beating `train_mean` by +0.0299 [+0.019, +0.041] and its own shuffled null by +0.0476 [+0.035, +0.061]. On the game-disjoint axis the identical control sits at 0.4986–0.5185, i.e. chance. A readout of the listener's own words should not predict the opponent's sheet; that it does here and only here localizes the effect to game-level memorization.

Stage B goes further than anything else in M1 — and still fails. `head_L18_dialogue_mean` reaches 0.7176 [0.706, 0.729] ranking and is **the only arm anywhere in M1 to clear all four baselines on any gate metric**, beating `own_sheet` by +0.0242 [+0.0084, +0.0412]. The gate fails anyway and not narrowly: the weights family also demands true-top-issue mass, where it loses to that same baseline by −0.3155 [−0.3363, −0.2943], and the threshold family fails too. A head with the capacity to memorize three games' sheets is exactly what should top a ranking metric while remaining unable to name the most important issue. **Arms passing the two-family gate anywhere in M1: none.**

The transferable statement. *Hold out seeds and the probe looks strong; hold out games and it collapses to the train mean.* This is now enforced in code rather than left to a reader's memory: the head predictor refuses to serve a checkpoint whose manifest records `fold_mode: "seed"` unless explicitly overridden, keyed off the manifest rather than the tag string. Delivery arms run on the seed axis would have shown a large apparent effect with nothing behind it.

4.3 The preregistered primary readout: complete coverage, fails everywhere

The proposal nominated "the last token of the view concatenated with the mean over opponent-authored spans" at layer 18. Coverage is closed at both stages on both axes. On the games axis every concatenation head sits below the `train_mean` bar on ranking and is roughly 4.6× worse on threshold error (0.0716 versus 0.0153; difference −0.0564 [−0.0598, −0.0527]). On the seeds axis it loses to the no-dialogue baseline on all four metrics and does not even match its own stage-A linear counterpart. **No configuration of the nominated readout beats the baselines anywhere.**

Correction history. A mid-run catch that stage B's default pooling list omitted the concatenation was real but **stage-B-only**: stage A swept it at all three layers on both axes from the start, and it is the game-axis best arm. The verdict never rested on inferring the nominated readout from its two components. The missing stage-B cells were run afterwards and closed the gap in the failing direction.

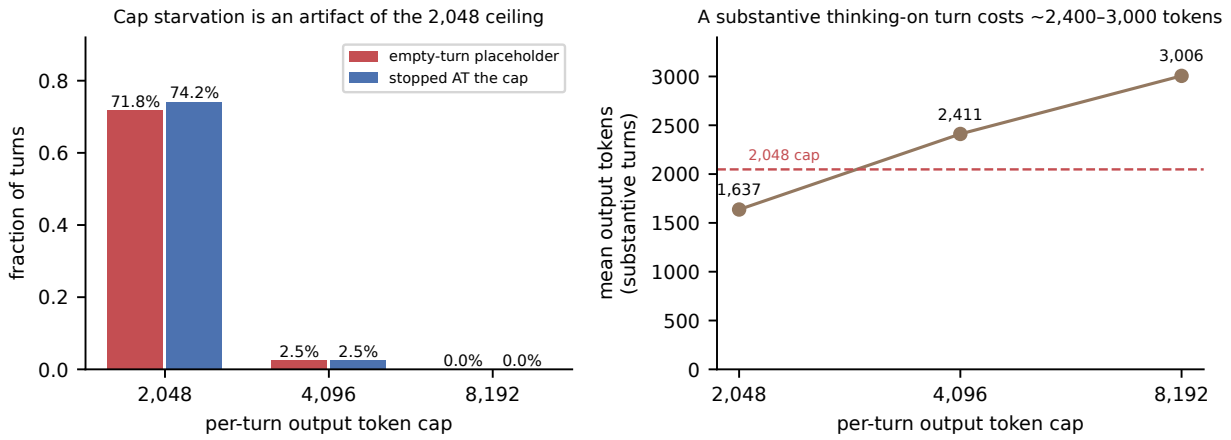

Figure 2: A per-turn token cap silently deleted most of the corpus. At a 2,048-token cap, 71.8% of turns are the engine's empty-turn placeholder and 74.2% stop exactly at the cap: the model spends its whole budget inside its reasoning block and never emits an action. Raising the cap removes the effect (4,096: 2.5%; 8,192: 0.0%) because a substantive thinking-on turn costs ∼2,400–3,000 output tokens — well past a 2,048 ceiling. The 8,192 arm costs $\geq 3.1\times$ the wall clock. The quarantined thinking-off tree shows 0.0% placeholder at ∼100 tokens/turn, so starvation is entirely a thinking-on phenomenon.

4.4 The one control-validated signal (descriptive only)

On the game-disjoint axis, span-mean poolings carry genuine *opponent-specific* per-option structure: layer 18 reaches +0.1571 [+0.0806, +0.2617], beating its shuffled null (+0.1506), text-only (+0.1385) and train-mean (+0.1313), while the matched `self_span` control sits at +0.0346 [−0.0346, +0.1056], spanning zero. Present at every layer (0.127/0.157/0.141), always from span-means, never from the self control, never from last-token poolings. So the listener does encode something about *that* opponent, read from *that* opponent's tokens. It is not enough to matter — the formal-move oracle scores 0.1994 on the same metric — and the preregistration classes the family descriptive precisely because it has no advocate floor to beat.

5 The corpus finding: two-thirds of the "dialogue" was empty

Late in the lane, an independent measurement of the stored campaigns the gate axis leans on found that **67–72% of turns are the engine's empty-turn placeholder** and 71–74% stop exactly at the cap. Across the three campaigns only 263 of 1,071 turns (24.6%) carry a free-chat message. Two-thirds of what this lane treated as "dialogue" was not dialogue. A cap A/B (Figure 2) pins the cause exactly.

This is a limit on the *premise*, not on the arithmetic, and it cuts three ways.

It **bounds** the negative. The lane asks whether free chat carries preference signal; on this corpus there was not much free chat to carry it. The honest claim is that *these* dialogues' representations do not beat the baselines — not that richer negotiation dialogue could not. It also reframes the descriptive per-option signal as something the listener extracted from the *third* of turns with real content.

It does **not** weaken the gate comparison. The text-only baseline reads the same thin dialogue through the same turns, so probe-versus-text is like-for-like at whatever volume existed, and the probe does not clear it; the shuffled null and train-mean arm do not benefit from dialogue at all.

What the thinness plausibly *does* explain is the absent turn-indexed trend — with most turns contentless there is little to accumulate across rounds.

It **retro-qualifies note 0009**, which ran on the same corpus. That note's advocate null was measured on seats that mostly were not speaking, so "fallible Bayesian advice did not help" should be read alongside "there was little free-chat channel for the advice to interact with." The localized cause reported there stands; this adds a second, independent thinness on the chat side. **Standing recommendation: every future rollout in this lane runs at ≥4,096 tokens per turn and budgets ≥ 3× the wall clock.**

6 The channel addendum — can you write into the model instead of prompting it?

6.1 Why this question survived the readout no-go

The preregistration's third hypothesis asked whether a learned representation could be delivered to a policy through **virtual tokens** — vectors written directly into the prompt's embedding sequence in place of words. With M1's readout failing its gate there was no learned prediction left worth delivering, so H3-as-a-theory-of-mind-question died. But the *channel* question survives if you ask it about information already known to work. Amendment 1 preregistered exactly that swap *before any training compute was spent*: hold the information fixed at note 0008's exact Nash-bargaining candidate, and change only the pipe.

6.2 Design

Four token-matched arms, every seat conditioned identically, on the 24 games of the confirmatory bank the encoder never trained on, 10 seeds each, **960 episodes**. All four blocks are padded to exactly equal token counts, so no arm buys its effect with a longer prompt.

Table 5: The four arms. `B_shuffled` is the load-bearing control: it isolates *information delivered* from *parameters added*.

Arm	What the block contains
`A_text`	Note 0008's text block verbatim — the replication anchor
`B_soft`	The same deal as **38 virtual tokens** from a trained 2.1M-parameter encoder reading a one-hot encoding
`B_shuffled`	The *identical encoder with identical parameters* fed a deterministically mis-drawn deal
`C_none`	A neutral `UNSPECIFIED` package

The encoder was trained by **pure distillation — no reward, no game outcome**: minimize the KL divergence between the frozen model's next-token distribution given the virtual-token prompt and given the text prompt, over turn prompts from the train half of a 24/24 game split.

6.3 The distillation gate passed

Held-out mean teacher-forced KL over response positions fell from 0.09905 (untrained encoder) to 0.02756 (trained), a ratio of 0.278 against a required ≤ 0.5. The KL between the text prompt and the no-candidate prompt on those same positions is 0.10907, so the trained channel's residual divergence from text is only 25.3% of the entire text-versus-nothing distance. Two framings of "how

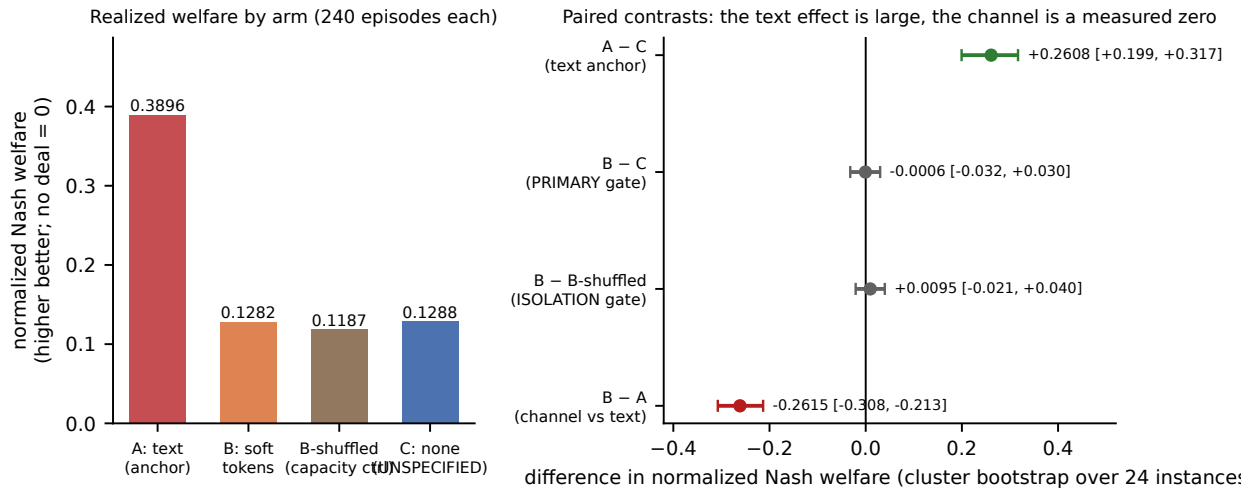

Figure 3: Left: realized normalized Nash welfare per arm (240 episodes each). The soft arm does not sit between text and nothing — it sits exactly on nothing. **Right:** the paired contrasts with cluster-bootstrap intervals over the 24 game instances. Green = interval excludes zero above, red = below, grey = spans zero. The text anchor replicates; the primary and isolation gates fail with intervals tight enough to exclude anything a tenth the text effect's size.

much of the gap closed" appear in this bundle and both are correct under their own definitions: 72.2% is the reduction relative to the untrained encoder, 74.7% is one minus the residual as a fraction of the reference gap. **By the metric it was trained on, the virtual-token channel is a good imitation of the text channel.**

6.4 The outcome result: a measured zero

Table 6: All eight contrasts, paired on (instance, seed), cluster bootstrap over 24 instances / 240 pairs each. Gate rows are marked.

Contrast	Definition	Estimate	95% CI	Gate
`anchor_text_vs_none`	A − C on NNW	**+0.2608**	[+0.1994, +0.3167]	(anchor)
`primary_normalized_nash_welfare`	B − C on NNW	**−0.0006**	[−0.0321, +0.0300]	**FAIL**
`isolation_vs_shuffled`	B − B-shuffled on NNW	+0.0095	[−0.0209, +0.0397]	**FAIL**
`channel_vs_text`	B − A on NNW	−0.2615	[−0.3075, −0.2133]	(two-sided)
`anchor_shuffled_vs_none`	B-shuffled − C on NNW	−0.0101	[−0.0388, +0.0198]	(anchor)
`safety_deal_rate`	B − C on deal rate	+0.0250	[−0.0167, +0.0667]	**PASS**
`safety_below_threshold`	B − C on below-threshold	−0.0125	[−0.0875, +0.0625]	**FAIL**†
`safety_malformed`	B − C on malformed episodes	+0.0042	[−0.0250, +0.0292]	—

†**The one verdict that should not be read at face value.** The below-threshold gate is recorded as a failure and is **not** a harm signal. Its point estimate is −0.0125: Arm B produces *fewer* below-threshold agreements than Arm C, moving in the safe direction. The gate fails only because a ±0.07 interval cannot clear a threshold of 0.01 — a threshold inherited from note 0008, a setting where below-threshold agreements were *rare*. Here they run at roughly 0.50 in every non-text arm, a regime in which such an interval can never clear such a bar regardless of the truth. **The gate is**

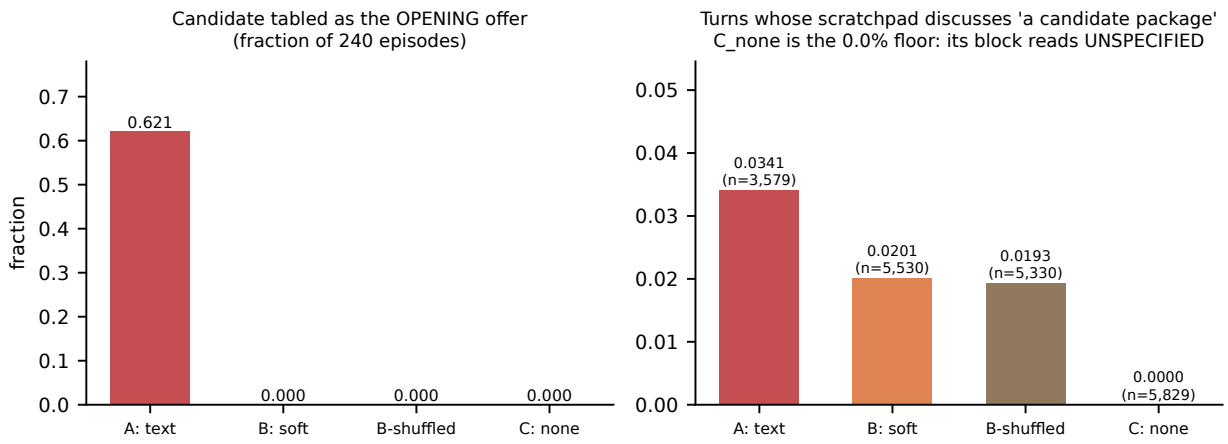

Figure 4: Left: the text effect runs through one concrete act — tabling the named package as the opening formal offer, in 62.1% of episodes — which under soft tokens happens in 0 of 240. **Right:** what replaces it is not indifference. Soft-token seats discuss "a candidate package" in 2.0% of turns against a hard 0.0% floor across 5,829 turns in the arm whose block reads `UNSPECIFIED`. That zero is what makes the reading safe: the soft block does not read as absence or as noise.

uninformative at this base rate rather than violated. The text arm is the outlier that makes the point: it cuts below-threshold agreements to 0.175, so the useful information does not merely raise welfare — it stops seats from signing deals worse for them than walking away.

Table 7: Descriptive arm summaries (not additional confirmatory tests).

Arm	N	NNW	NNW given deal	Deal rate	Below threshold	Candidate uptake
`A_text`	240	**0.3896**	0.4013	0.9708	**0.1750**	**0.5208**
`B_soft`	240	0.1282	0.1398	0.9167	0.5000	0.0000
`B_shuffled`	240	0.1187	0.1272	0.9333	0.5292	0.0000
`C_none`	240	0.1288	0.1445	0.8917	0.5125	0.0000

6.5 The mechanism: confabulation, with a hard zero as its control

A representative turn-0 scratchpad reasons "let me calculate the total score for the candidate package: issue0-opt1 (83), issue1-opt2 (20), . . . ", totals it against its own threshold, and proposes that package — while the deal actually encoded in the virtual tokens was a different one entirely. **The channel delivered the *presence* of advice without its *content***, which is arguably worse than delivering nothing: it induces confident reasoning over a package that does not exist.

6.6 Provenance — not a clean single pass

Recorded because the honest version matters more than a tidy one. The campaign was **run twice**: a first launch at 32-episode pool width was aborted after one completed pool and its 64 episodes discarded, then relaunched at 96-episode width once the GPU freed. Only the wider run is analysed, so the scored panel is a single configuration rather than a mixture. Earlier, a smoke run was killed by another agent clearing the GPU, and two full-width attempts died on CUDA out-of-memory

against a concurrent job holding 164 GB of a 178 GB card — in those cases the fail-loud fabrication ceiling *aborted the runs rather than keeping the affected turns*.

What is reported has **0 of 20,268 turns fabricated**, and injection was *counted rather than assumed*: 5,530/5,530 rows injected for `B_soft` and 5,330/5,330 for `B_shuffled`, 0 unregistered, both under the identical encoder checkpoint. That accounting is what licenses the isolation contrast — B and B-shuffled genuinely differ only in the deal fed to the same weights.

7 The attention-interface ladder — the channel negative was *depth*

[**Research note:** `research-notes/0029-qkv-attention-interface.md` · **preregistration** `proposals/2026-08-` · **public tables** `2026.RA.QKV-Attention-Interface`]

The channel result in §6 arrived with a mechanistic *guess* attached: a real text token's per-layer keys and values are in-distribution for the copy machinery the model already has, whereas a soft input vector produces keys and values no frozen attention head was ever trained to read. If that guess were right, the fix is not a better encoder but a **deeper writing site**. A four-rung ladder was preregistered to test it — cheap rungs first, each with its own gate, contingencies written down in advance. This section reports it. The short version is that the guess was right about the *level* and wrong about the *reason*, and that the lane's closing claim in §8 needs narrowing as a result.

7.1 R0 — what reads the advice, and how redundantly

R0 spent no new rollouts: it re-used the committed episodes of §6 and asked what carries the candidate from the prompt into the proposal. The screen is a teacher-forced proxy — the summed log-probability of the exact candidate deal's tokens inside a fixed `propose`-shaped continuation, scored under the advice prompt and its paired no-advice prompt. *Recovered* is the fraction of that gap an intervention destroys, so 1.0 means the intervention reproduces the no-advice condition. Removing the payload span from every head at every layer recovers 1.096: the uptake is entirely an attention phenomenon.

The circuit is sharply localized *by layer* and diffuse *within* it. Six-layer windows recover $+0.16/-0.05/+\mathbf{0.57}/+\mathbf{0.70}/+0.11/+0.01$ from layers 0–5 through 30–35, so the reading happens in **layers 12–23** — but any single layer ablated alone moves the metric by ∼0.00, the best single query head anywhere removes only 0.015, and it takes ∼160 of the model's 1,152 query heads to cross half the effect.

The gate passes on real generation, and not by breaking the model. Ablating the identified set's attention onto the advice span cuts the *actually generated* rate of tabling the exact candidate from 47.5% to 2.5% (a 94.7% relative cut; paired per-episode delta 0.450 [+0.175, +0.775]), while the JSON parse rate and the propose rate both stay at **1.00** in *both* conditions. The ablated model still writes a well-formed proposal with the same scratchpad reasoning — for a different deal.

A control right-sized the specificity claim by a factor of four. The identified 160-head set recovers 0.695 [0.631, 0.757] held out, against 0.124 for size-matched *random* query heads — an apparent 5.6× specificity that was on track to be reported. It is an artifact of grouped-query attention: Qwen3-8B's 32 query heads per layer share only 8 key/value heads, and heads within a group read *identical keys*, so 160 random query heads are 160 *partial* groups, each leaving its span readable through the group's remaining heads. Against random sets matched on the same per-layer KV-group structure the selection recovers 0.846 [0.760, 0.929] versus **0.635**, and the disjoint

complement — every group the selection did *not* pick — still recovers 0.523 alone. Roughly three quarters of the ablation effect is a function of *how much* of the window is removed rather than *which* part. Per the preregistration's written contingency this demoted R1's mechanism-informed variant to plain prefix-tuning: constraining injection to that head set would have been a restriction dressed as a prior.

7.2 Where attention actually goes, and the control that reframed it

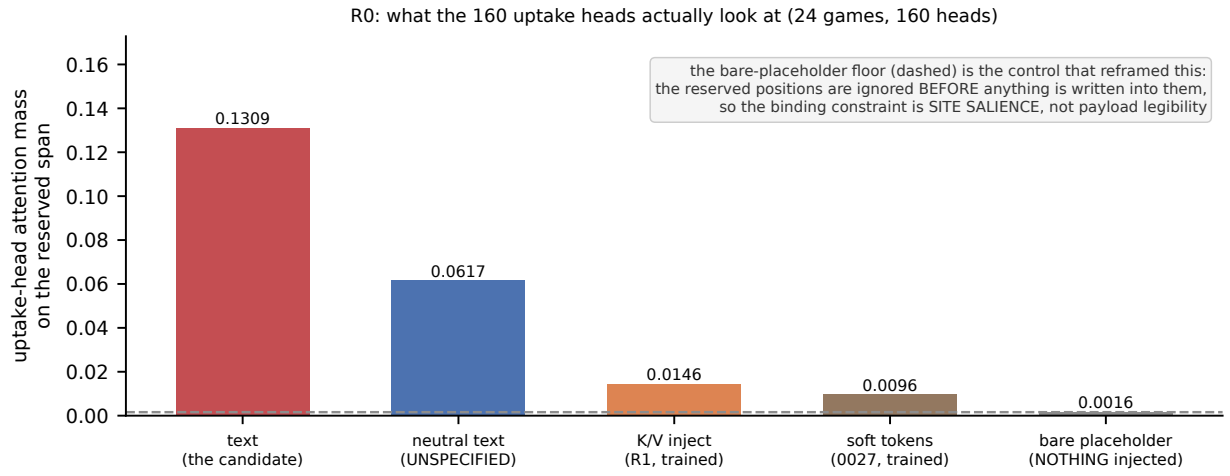

Figure 5: The five-condition salience table, and the control that overturned this section's first conclusion. Uptake-head attention mass on the reserved span, 24 games, the 160 identified heads. The candidate as text draws 0.1309; *neutral* `UNSPECIFIED` text in the identical slot draws 0.0617; R1's per-layer K/V injection 0.0146; the failed soft-token channel 0.0096 — *below* the neutral-text floor. The obvious reading is "the encoder produces keys these queries never match". The dashed line is what makes that reading wrong: the same prompt with the reserved run left **bare**, nothing injected at all, draws 0.0016. The positions are ignored *before* anything is written into them.

Figure 5 is an errata entry as much as a result. R0 published, and messaged downstream, the attribution that §6's channel failed *at the key* — illegible payload. A bare-placeholder condition that nobody had requested killed it: a run of reserved placeholder tokens draws 0.0016, about 2.6% of what ordinary neutral text draws in the identical slot, *before anything is injected*. Both trained channels start from that floor and neither climbs off it (soft +0.008, K/V +0.013 above bare). So the binding constraint is the **salience of the delivery site**, not the legibility of the payload: the uptake heads do not query those positions, and writing better content into a position nobody looks at does not make it looked at.

7.3 R1 — the candidate as per-layer keys and values

R1 holds the information and the four-arm design of §6 fixed and moves only the writing site. An encoder overwrites the output of **every layer's** `k_proj` and `v_proj` at 38 reserved positions, per key/value head (36 layers × 8 KV heads × 128 head dim; 19,094,236 trainable parameters, the only trainable parameters in the rung). Injected keys still pass through Qwen3's per-head QK-RMSNorm and RoPE exactly as a real token's keys do, so they cannot buy salience with a free magnitude. The prefix occupies *real* reserved positions rather than pre-seeding `past_key_values`,

because `transformers` derives `cache_position` from cache length and `generate` silently drops the first n input tokens when a pre-seeded cache does not correspond to real tokens.

The objective changed, and that is half the rung. §6 showed distillation KL is satisfiable by *style*. It is demoted here to an auxiliary term at weight 0.05, and the gate becomes **restatability**: given the injected K/V, the frozen model must *state* the encoded candidate on games no encoder parameter has seen. It reaches **0.9583** (23/24 held-out games) against a no-channel control of **0.0000**. That control is load-bearing: a first smoke measured it at 0.50, because the probe had been built on logged text-arm views whose episodes table the candidate as their opening offer 62.1% of the time — the dialogue history was handing the model the answer. A gate scored that way could have been passed by a channel carrying nothing, which is §6's failure mode with extra steps.

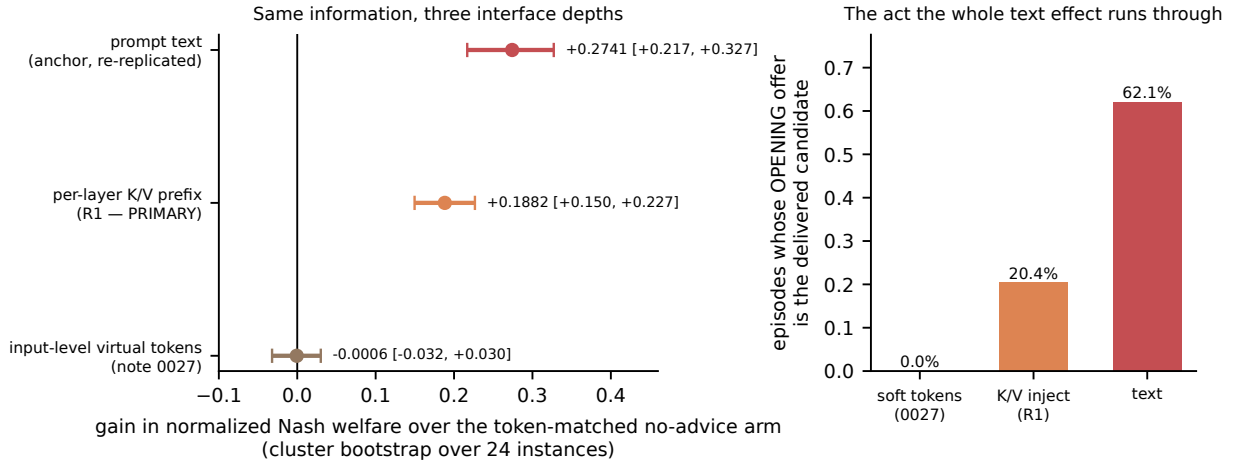

Figure 6: The same information at three interface depths. *Left:* paired contrasts against the same token-matched no-advice arm, cluster bootstrap over 24 game instances. Prompt text re-replicates at +0.2741; per-layer K/V delivers +0.1882, or 69% of that ceiling; the input-level virtual tokens of §6 are a measured zero. *Right:* the one act the entire text effect runs through — tabling the named package as the *opening* formal offer — at 0%, 20.4% and 62.1%. The figure asserts in code that the text anchor's opening-offer rate replicates between the two campaigns' bundles.

Result. On 960 episodes with **0 of 17,717 turns fabricated** and injection counted per row (4,012/4,012 and 4,254/4,254, zero unregistered, one encoder fingerprint serving both injected arms), the primary contrast is **+0.1882** [+0.1497, +0.2266] — **PASS**. Isolation against the same encoder fed a deliberately mis-drawn deal is +0.0930 [+0.0474, +0.1390] — **PASS**, so this is content, not capacity. The re-based below-threshold safety gate moves −0.3750 [−0.4708, −0.2792], a large move in the safe direction — **PASS**. Deal-rate non-inferiority is **FAIL** at −0.0042 [−0.0583, +0.0500]: the point estimate is *one* fewer closed deal in 240 and the interval merely reaches past the −0.05 bar, so the panel cannot resolve a change this small — which is not the same as no cost, and is why it is recorded as a failure rather than narrated as a pass. The channel remains significantly *below* the text ceiling (−0.0859 [−0.1364, −0.0332]), so it is a partial substitute.

7.4 The prediction this ladder registered, and lost

Before the campaign, R0 and R1 jointly registered in writing that a null outcome would be explained by Figure 5 and that a *positive* outcome would be "genuinely surprising, worth its own investigation".

It came back positive. **The prediction is falsified, and is recorded as falsified rather than reinterpreted.**

The tension is quantitative and sharp: the channel produces +0.1882 welfare and 20.4% opening-offer tabling while the 160 identified uptake heads put 0.0146 attention mass on those positions against 0.1309 for text, with a knockout delta of 0.317 nats against a 0.230 bare-placeholder floor. The measurement survives; the inference from it does not. The claim this licenses is deliberately narrow: **low attention mass at an identified head set does not imply a channel cannot drive behaviour.** It does *not* license "the uptake circuit is irrelevant" or "attention-mass measurement is uninformative", both of which are over-reads in the opposite direction.

The leading hypothesis has since been **measured and refuted as stated**. It held that injected *keys* are norm-pinned by QK-RMSNorm while injected *values* are unconstrained, so the encoder could buy influence through value norm rather than attention share. Amendment 2's rung F1 measured both halves. **Globally the values are quieter, not louder**: median injected V norm 5.210 against real 5.984, a ratio of 0.87× pooled over the stack. And **decodability does not need the magnitude at all**: three norm-matched retrains clamping injected value norms to the real distribution's p90/p95/p99 per layer all pass the reconstruction gate at 1.0000 against 0.0000 controls, with the p90 clamp demonstrably binding — it rescales 38.5% of all value vectors, the largest 26.6× over its limit.

What replaces it is a **shape mismatch** rather than a magnitude anomaly. The encoder's norms are roughly flat with depth while the model's real value norms grow by more than two orders of magnitude (median 0.275 at layer 0 to 55.6 by layer 33), so the injected values run 15.9×, 10.0× and 7.4× the real median at layers 0–2 — with *every* injected vector above the real p95 through layer 6 — cross 1.0 around layer 16, and fall to 0.08–0.30× by layers 33–35. Two details mark this as a property of the encoder rather than a strategy it discovered: the norm-pinned *keys* show the same flat profile despite having nothing to gain from magnitude, and the flatness traces to shared initialization of the encoder's per-layer output heads. Combining F1's norms with R0's separate attention-mass measurement gives a delivered mass × value displacement of 0.10× text's — a composite of two runs rather than one measured quantity, and a number that only partners a 69% behavioural effect if what the channel delivers is disproportionately useful, which remains unexplained.

The residual question was narrower and better posed than the original — the early-layer values *are* genuinely out of distribution, and boundary 2 showed that excess is not needed to *decode* the payload without showing it is not needed to *act* — and a norm-matched replay of R1's four-arm campaign then measured it. **Retention is** 0.5644 [0.3096, 0.8068] against a frozen 0.50 line, so a majority of R1's effect survives the clamp and **magnitude is a contributor rather than the cause**; the interval straddles the line, so this cannot separate "retains most" from "retains a third". The clamp is not free: −0.0750 [−0.1180, −0.0298] against R1, with opening-offer tabling falling 17.5% → 7.9%.

That resolves a fork registered before the campaign ran. The two branches were: *retention* would mean the out-of-distribution magnitude is incidental, and *loss* would mean it is load-bearing for behaviour while being unnecessary for decoding — a dissociation between what makes a payload readable and what makes it act. **What came back is that dissociation in partial form.** Decoding needs none of the excess (reconstruction 1.0000 under a clamp rescaling 38.5% of value vectors); acting needs some of it (retention 0.5644, tabling nearly halved). Readability and actionability do come apart on this axis, but the split is graded rather than clean, and the interval is too wide to put a number on the actionable fraction. Registering both branches in advance is what makes this a reportable partial result rather than a judgement call after the fact.

7.5 Amendment 2, rung F2 — addressing and bandwidth are not the constraint either

Two cheap explanations for the channel's under-use remained: the model may not know the payload is there to be used (*addressing*), or 38 reserved positions may be too few to hold it usefully (*bandwidth*). Both are **FAIL** — clean nulls, each against its own placebo. A text line naming the channel moves welfare −0.0066 [−0.0400, +0.0234] and tabling −0.0208 [−0.0750, +0.0333] against R1; doubling the prefix to 76 positions moves welfare +0.0013 [−0.0448, +0.0468] and tabling −0.0042 [−0.0875, +0.0750]. **Each arm's placebo — the pointer line alone, the extra padding alone — moves tabling by exactly** 0.0000 [0.0000, 0.0000], which is what makes these bounded negatives rather than shrugs. The capacity arm is the first in the whole ladder to pass all four of R1's gates while being behaviourally indistinguishable from it, and it pays malformed episodes that scale with *injection width* (2.1%, 10.0%, 19.6% at 0, 38 and 76 injected positions), which characterises but does not explain R1's unexplained shuffled-arm formatting oddity.

R1 replicated independently. On a different machine, a different sharding scheme and a month later, the text anchor returns +0.2554 [+0.2087, +0.3023] with opening-offer tabling **0.6208** — matching R1's published 0.6208 to four decimal places — and R1's own channel arm re-measures at +0.1722 [+0.1259, +0.2182] against its published +0.1882. The ladder's headline is not a one-panel artifact.

What is left, and it is the sharpest question the ladder produced. Four explanations for why the channel is acted on about *three times less often* than the same content as text are now eliminated by measurement rather than argument: not decodability, not magnitude, not addressing, not bandwidth. **The channel is decodable, correctly addressed, roomy — and still under-used.** What remains is **uptake**: for a frozen model, *acting on a retrieved payload appears to be a different operation from acting on a read one.* That is a claim about the model's machinery rather than about the interface, and it points back at this chapter's own untested suggestion — write the payload where the model's queries already look. **That suggestion has since been tested, and it loses** (§7.6). One power note travels with it: the retention interval $[0.31, 0.81]$ means that if the magnitude question needs settling, the answer is *more games, not more clamp levels.* Two alternatives are not excluded — R0's head set may be a poor summary of a circuit whose disjoint complement recovers 0.523 alone, and R0 measured turn-0 generation positions in a fixed scaffold whereas the campaign integrates over whole negotiations. The residual 31% gap to text likewise has **at least two live explanations** (low site salience, versus simply lower per-episode reach at 20.4% against 62.1% tabling) and R1 does not distinguish them.

7.6 Amendment 3, rung P — placement: the ignored site turns out to be the better one

Both injection rungs write into a run of **reserved placeholder tokens**, and Figure 5 says those positions are ignored before anything is written into them. The obvious remaining fix, and this chapter's own nominated next probe, is to write the payload where the model's queries already go. Rung P does exactly that: same encoder recipe, same objective, same gate, same panel, and the payload written at the token positions of the no-advice arm's *real text* — its `Candidate package: UNSPECIFIED` block, semantically empty by design, so overwriting destroys nothing. The prompts are **byte-identical** to that arm, same string and same SHA-256, which is stronger than the token

matching every earlier rung relied on: treatment and base differ in nothing a reader could see, only in the tensors written at 45 of the prompt's positions.

The channel works at the new site, and works worse there. On 1,200 episodes (5 arms × 24 held-out games × 10 seeds, **0 of 22,680 turns fabricated**), the primary contrast against no advice is +0.1429 [+0.1008, +0.1855] and it is content-specific rather than capacity — against the same encoder fed a deliberately mis-drawn deal, +0.0723 [+0.0277, +0.1165]. But the rung's own frozen directional gate **FAIL**: −0.0373 [−0.0723, −0.0029] against R1's placeholder site, an interval excluding zero on the *wrong* side. Placement reaches 0.560 [0.397, 0.732] of this panel's own text ceiling against R1's 69% on its own.

And the new site costs things the old one does not. Deal-rate non-inferiority also **FAIL**s, at −0.0875 [−0.1458, −0.0292] — and unlike R1's miss, which was a one-episode-in-240 point estimate failing on interval width, this is a real measured cost. The panel localises it: `KV_r1` against no advice on deal rate, in these same games, is exactly +0.0000 [−0.0542, +0.0542], so **the cost belongs to the placement site and not to per-layer K/V injection**. Malformed episodes rise +0.3000 [+0.2042, +0.4042], and the width account that had been priced in advance — §7.5's 2.1%, 10.0%, 19.6% at 0, 38 and 76 positions — is refuted inside the panel: 38 *placeholder* positions cost +0.0333 [−0.0250, +0.0958] malformed here, indistinguishable from zero, while 45 *real-text* positions cost +0.3000. Seven extra positions cannot buy a factor of four. Overwriting text the model is reading destabilises its output formatting; overwriting semantically empty positions does not.

Table 8: Rung P's mechanism check, measured and registered before any scored episode — and it points the opposite way from the behaviour. Uptake-head attention mass and knockout delta over 12 held-out prompts and the 160 identified heads. The comparison that matters for an injected arm is against *its own site left bare*: writing the payload over real text **destroys** 77% of the salience that site had, while the same operation at the reserved run *lifts* mass off a 0.0016 floor. Meanwhile the knockout deltas move the other way — the placement span is worth four times more to the probe while drawing less attention. Neither readout predicts the behavioural ordering, in which placement is the *worse* channel.

condition	uptake-head mass	vs. its own bare site	knockout delta
`A_text` (candidate as text)	0.1332	—	+14.007 nats
`C_none` (the real-text site, **bare**)	0.0610	—	+1.923 nats
`KV_place` (payload over that real text)	**0.0140**	**−0.0470** (a 77% drop)	**+0.751** nats
`B_placeholder` (the reserved run, **bare**)	0.0016	—	+0.369 nats
`KV_inject` (R1's payload at that run)	0.0172	+0.0157 (a rise off the floor)	+0.190 nats

The mechanism check falsified the rung's own prediction before the campaign ran (Table 8). Amendment 3 registered that the real-text site would draw *more* than the placeholder site's injected 0.0146; it draws less (0.0140 against 0.0172 in the same measurement, the artifact recording `holds: false`). The sharper reading is the one the grid did not contain: injection at a content-bearing site *destroys* the salience that site already had, so **salience follows what is written, not where it is written**, and the encoder's keys and values are intrinsically low-salience to this circuit wherever they land. The knockout deltas then dissociate from the mass in the other direction. Mass says the two sites are equivalent; knockout says placement is four times stronger; welfare says placement is worse.

Why this verdict is cleanly interpretable: the grid was amended before the outcome existed. Amendment 3 froze a four-cell interpretation grid, three cells of which assumed the mass would go *up* or stay *flat*. Scored as written, a negative placement result would have landed in the "mass flat = the site manipulation failed mechanically" cell and been reported as *inconclusive-with-*

manipulation-failure — the opposite of the truth, since the manipulation worked and moved the readout by 77%. **Amendment 3a** was therefore logged after the mechanism check and **before the first scored campaign shard ran** — in the rung's own bundle (`p/METHODS.md`) rather than in the preregistration document, a provenance gap found at close-out and closed with a labelled back-reference — adding the fifth case with *both* of its behavioural sub-branches frozen in advance: placement matching or beating R1 despite lower mass would have been a second, stronger mass-behaviour dissociation; placement losing would be **payload-text interference at a content-bearing site**. The second is what occurred, and the branch predates the number. The amendment also argued against its own evidence, because this ladder had already published a channel that moved welfare +0.1882 while drawing near-floor attention (§7.4), so a lower mass number was explicitly recorded as *not* a predicted null and the campaign ran.

Two further things the panel records rather than smooths. The retrained encoder's reconstruction gate is 1.0000 against a 0.0000 no-channel control — *above* R1's 0.9583 — so the real-text site slightly *helps* decodability even as it costs behaviour, which is one more instance of readability and actionability coming apart. And the isolation contrast disagrees with itself across endpoints: it passes on welfare while `KV_place` tables its own delivered candidate *less* often than the wrong-deal control does (8.75% against 12.50%; paired −0.0375 [−0.0958, +0.0167]). Neither this rung nor R1 explains how a channel produces more welfare through a package it tables less often. Finally, `KV_r1` in this panel is a **third** independent R1 replication: +0.1802 [+0.1380, +0.2205] against the published +0.1882, with the text anchor at +0.2553 and text tabling at 0.6042.

7.7 R2 — re-weighting what the model already has

R2 asks a different question with no training at all. The seat's own acceptance threshold is *already in its prompt* and it signs below-threshold deals anyway; if that is an attention-*allocation* deficit, then adding a uniform $+\beta$ to the attention logits of the decision-relevant key columns should move behaviour. The endpoint is the **IR-violation rate** (signing a deal worse than walking away), this program's bright-line irrationality, and the arms are token-identical by construction because the bias is mask-side only.

The sweep's most valuable output is a near-miss. The endpoint is scored `reached AND any(surplus < 0)` — a conjunction requiring a *closed deal* — so it is exactly zero when no deal closes. Under the *originally preregistered* unscreened selection rule ("lowest below-threshold rate wins") the winner is the β=2 widest-span cell at **0.1944** against the no-bias reference's 0.5972: a two-thirds cut in this program's bright-line irrationality, which *reads* as nearly matching §6's text arm at 0.175. It is an artifact. That cell's deal rate is **0.4444** against 0.9167 and its Nash welfare is *worse* (0.0709 against 0.1007). The apparent rationality gain is **deal suppression**: fewer closed deals mechanically means fewer below-threshold closed deals. A **viability screen** was added *before the sweep was scored* — eligible cells must keep deal rate and malformed rate within 0.05 of the β=0 reference, deliberately the evaluation's own non-inferiority tolerance applied at selection time rather than a new standard invented to fit the data. Without it, 720 held-out episodes would have measured the bias that stops the model negotiating and reported it as a success.

The exploitable region is nonetheless **bounded**, which sharpens the hazard rather than softening it: at β=4 on the widest span the below-threshold rate returns to 0.5972 — *exactly* the no-bias reference — with a 1.0000 malformed rate. Total degeneration does not maximize the endpoint; it destroys the episode. The metric is gameable by *partially* suppressing deals, not by breaking the model outright.

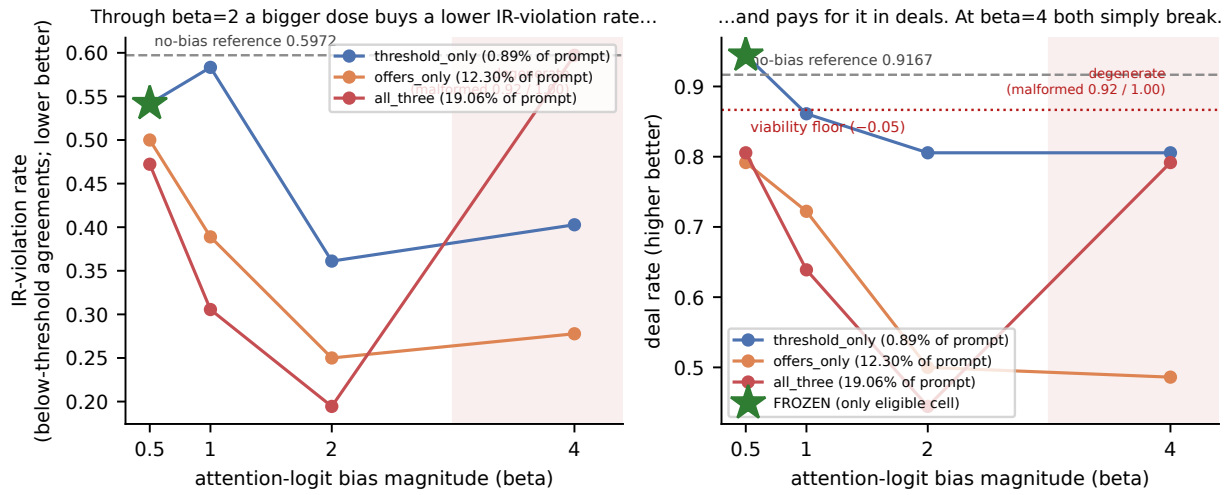

Figure 7: The sweep: a trade-off, not an improvement, and only inside a bounded range. 13 cells × 3 seeds × 24 train games, 936 episodes. Each line is a span set labelled with its *exogenous* diverted share — measured on unbiased β=0 prompts, because the per-arm biased-token count moves *with* the treatment (a stronger bias degrades play, so fewer offers accumulate and the spans, which are made of game state, shrink). Through β=2 a larger dose buys a lower IR-violation rate and pays for it in deal rate. The shaded β=4 region is degenerate (malformed rates 0.92 and 1.00) and the pattern reverses there rather than continuing. The star is the frozen cell — the only one of twelve to pass the viability screen, and the only cell that improves both endpoints at once.

The held-out result: FAIL, and the control is the reason. The frozen cell (β=0.5, threshold line only) was played once on 720 episodes across three arms — no bias, span bias, and a *mis-targeted* control at the same β on the same number of tokens drawn from format boilerplate — with **0 of 16,646 turns fabricated**.

Table 9: R2's preregistered gates, evaluated once on the held-out half. The gate that kills the hypothesis is `beats_mistargeted`, not the primary.

Gate	Verdict	Estimate [low, high]	Reading
`primary_below_threshold`	**FAIL**	−0.0833 [−0.1708, +0.0042]	right direction, interval grazes zero
`beats_mistargeted`	**FAIL**	−0.0417 [−0.1333, +0.0500]	no separation from the wrong target
`mistargeted_does_not_reproduce`	**PASS**†	−0.0417 [−0.1208, +0.0375]	†passes on interval width *only*
`deal_rate_noninferiority`	**FAIL**	−0.0250 [−0.0792, +0.0292]	interval reaches past the margin

Two readings make this substantive rather than merely underpowered. **The effect is not allocation-specific:** biasing rules boilerplate moves the endpoint −0.0417, half the treatment's point estimate in the *same* direction, and the treatment does not separate from it. **The welfare movement is entirely non-specific:** normalized Nash welfare rises +0.0276 [−0.0049, +0.0608] under span bias, and span-minus-mis-targeted on welfare is −0.0035 [−0.0303, +0.0248] — the control captures the gain *equally*. A reader shown only the treatment-versus-nothing column would have concluded that biasing the threshold line improves bargaining.

† **The passing gate must not be quoted as specificity evidence.** `mistargeted_does_not_reproduce` passes only because its interval contains zero; its point estimate is half the treatment's with the same sign. It is the one number in the bundle that reads better than it is.

And the manipulation applied. Attention mass on the targeted spans moves 0.0028 → 0.0051 at the frozen β and → 0.1281 at β=4 — a **46**× span. A behavioural null *with* a confirmed manipulation is precisely what the preregistration defined as a real negative, and attention-map movement was ruled non-evidence in advance, so this is explicitly not a partial pass. The control is also *conservative*: starting from ∼4× the baseline mass (∼590 boilerplate tokens against ∼26), it diverts *more* absolute mass at equal β, so failing to separate from it is demanding rather than lenient.

Honest limits. Power: 24 clusters give ±0.09 intervals on a rate near 0.5, sized against the text arm's +0.26 and unable to resolve 0.08 — and the right target for further spend is the *specificity* contrast, not the primary. Selection: the frozen cell won by being the only eligible one, with a ∼1-SE train margin logged in advance. One cell this design cannot test: the β=2 narrow-span cell posts the *best* Nash welfare in the entire grid (0.1366 against 0.1007) *and* a much lower below-threshold rate (0.3611 against 0.5972), excluded only for an 0.11 deal-rate drop that would fail the same gate at eval. A mid-strength narrow bias buying real welfare at a bounded cost in agreements was a hypothesis no arm in *that* design could express, because deal rate was a safety gate rather than a co-primary — so it was given its own rung, below.

7.8 Amendment 2, rung F3 — pricing the excluded cell

The cell above was excluded for looking too good, and R2's design could not say whether the screen was protecting the result or discarding a real gain. F3 answers it: deal rate is promoted from a safety gate to a *priced cost* and welfare made co-primary, under a preregistered readout, so a cell that buys welfare at an acknowledged cost in agreements can pass on its merits. Played once on the same 24 held-out games × 10 seeds × 3 arms (720 episodes, **0 of 17,543 turns fabricated**), **both co-primary gates FAIL**: welfare against no-bias +0.0015 [−0.0254, +0.0300], and the load-bearing welfare contrast against the mis-targeted control +0.0128 [−0.0141, +0.0383]. The priced cost, reported and not gated, *does* replicate: deal rate −0.0667 [−0.1125, −0.0250], an interval excluding zero.

The price replicated and the goods did not. The train-half welfare gap was +0.036 and returns +0.0015; the train-half deal-rate cost was −0.1111 and returns −0.0667. The train figure is a selection artifact of a 12-cell grid at 72 episodes per cell — winner's-curse shrinkage, measured rather than assumed. The mechanism is a genuine one-for-one cancellation rather than a flat null: welfare *given a deal* rises 0.1311 → 0.1430 while the deal rate falls 0.9292 → 0.8625. Better deals, fewer of them, cancelling at an exchange rate of about 1:1 — which is to say, buying nothing. So R2's viability screen is **retroactively vindicated**: it discarded a cell whose gain was a train-half artifact and whose price was real.

One reading is explicitly refused. The only interval here excluding zero on an irrationality endpoint is a *secondary* — span bias minus mis-targeted on below-threshold, −0.1083 [−0.1667, −0.0542] — and the mis-targeted arm sits at 0.5958, *worse* than no-bias's 0.5333, so 58% of that gap is the control degrading rather than the treatment improving; treatment-versus-nothing stays inconclusive at −0.0458 [−0.1083, +0.0208]. Promoting it would be an endpoint swap, and F3 blocks that structurally rather than by discipline: a **readout registry** refuses to score a config under a readout it was not registered for, and re-running R2's frozen evaluation through the refactored code reproduces all seven shipped contrasts bit-identically with all four gate verdicts unchanged.

7.9 What the ladder establishes

The kill-map gained a third working interface. Before this ladder, every result in this program pointed the same way: **text and outcome-reward were the only interfaces that**

changed behaviour, with probes, distillation objectives and embedding-level virtual tokens all failing. **Per-layer K/V now joins them.** The same information, same games, same token budget, moved from the input embedding to the per-layer keys and values, goes from a measured zero to 69% of the prompt-text ceiling.

The attention-allocation knob is reachable but is not the lever. R2 moves attention mass on a targeted span by 46× and does not move decisions, and F3 (§7.8) shows the stronger setting does not either — its apparent welfare gain does not replicate while its deal-rate price does, so the negative is bounded at both magnitudes rather than only the gentle one; what movement appears is reproduced by pointing the same distortion at boilerplate. Reachable, measurable, and not causal for this behaviour. Put beside R1, this is the ladder's cleanest structural result: **adding new information through a deeper interface works; re-weighting information the model already has does not.**

The decoupling thesis, in both directions inside one ladder. §6 showed a channel can satisfy an **optimized** proxy — a distillation-KL training objective, three quarters of the distributional gap closed — while carrying zero behavioural payload. R1 shows a channel can drive behaviour substantially while a **measured** proxy — attention mass at an identified head set — says it is barely attended to. One proxy you optimize against, one you measure and interpret, and each failed in the direction its own kind fails. **Neither an optimized proxy nor a measured proxy substitutes for an outcome contrast against a matched-capacity control.** That is a stronger claim than either half, and stronger than lumping both under "probes decouple from behaviour".

An endpoint hazard that generalizes past this ladder. Any conjunctive metric that requires an event can be maximized by *suppressing* the event. Realized, not hypothetical: the degenerate cell wins the unscreened sweep. Every arm in this program that selects on, gates on, or rewards the IR-violation rate needs the same check — the outcome-reward training lane most of all, since a trained optimizer will *search for* the degenerate solution rather than stumble into it.

The delivery site was the last free variable, and varying it made things worse. Rung P (§7.6) moved the payload from ignored reserved positions onto real text — the fix this chapter itself nominated — and the channel came back weaker on every behavioural axis: -0.0373 $[-0.0723, -0.0029]$ against R1, a real deal-rate cost where the placeholder site costs exactly zero in the same games, and four times the malformed rate. **So the salience account is dead twice**: its registered prediction was falsified before the campaign, and then the low-salience site outperformed the high-salience one. Reserved low-salience positions are not a compromise this ladder was stuck with; on this bank they are the better design, because a trained payload *interferes* with real text rather than borrowing its salience. Taken with the rest of the ladder — input level, width, addressing, value-norm magnitude, attention-logit bias, and now site — **R1's original configuration is locally optimal among everything tested**: the nulls are null, and the only two variations that move (the norm clamp at -0.0750, placement at -0.0373) move down. That is a bounded claim about one neighbourhood, not global optimality, and it is why no further variation of this interface is queued. What survives untouched is **uptake**.

R3 is not triggered. The preregistration gates the trainable cross-attention reader on R1 failing reconstruction *or* outcome. R1 passed both, so a frozen reader demonstrably can be made to read a channel *and* act on it, and the trainable-reader rung has no live justification from this result.

The two follow-up candidates this chapter once named — a **V-norm-matched re-run** and a **deal-rate-as-co-primary design** — have both since been built and run, as Amendment 2's rungs F1 and F3, and Amendment 3 then closed the placement question. No rung of this preregistration remains unrun.

8 What this lane establishes, and what it does not

Established. (i) Within this game family, at this scale, on frozen base representations, with held-out-game generalization: an opponent's hidden issue weights and reservation threshold are **not decodable** above four honest baselines. This is not a small effect missing significance — no arm beats a constant predictor and none beats its own shuffled labels. (ii) A readout that looks strong on held-out *seeds* is reading memorized games, demonstrated by two independent controls. (iii) Information that provably works as prompt text carries **zero** outcome effect through a distilled virtual-token channel *at the embedding level*, with the isolation control ruling out "capacity, not information" as an alternative story. (iv) That channel is not ignored — it induces confabulation, measured against a hard zero floor. (v) **The same information written into every layer's keys and values *does* work** (+0.1882, 69% of the text ceiling, content-specific against a mis-drawn-deal control), so (iii) is a fact about the writing *site* rather than about representation-level delivery in general (§7). (vi) Re-weighting attention toward decision-relevant text the model already has does *not* work, and fails against a length-matched mis-targeted control that reproduces its welfare gain exactly.

Not established. The readout result does not show the information is absent from the model. Four live limits: the frozen-base restriction (joint LoRA was gated behind stage B improving and never ran); sample size (16 games is at the edge of this program's $\geq$20-cluster rule, and game-disjoint intervals of $\pm$0.10 on ranking exclude effects larger than roughly 0.10 rather than establishing zero); corpus thinness (§5); and stage-B under-fitting on $\sim$700 rows per fold. The channel result is specific to *embedding-level* virtual tokens with *distillation-only* training in *this* setting; it does not rule out other injection sites or training signals — and §7 shows that caveat was the operative one, since a deeper site with a restatability objective works. Nor does the ladder explain *how* R1 works: its registered mechanistic prediction was falsified (§7.4) and the leading account is untested.

The unified reading. Distillation matched distributional *style*, not usable *content*. The encoder was optimized to make the model's next-token distribution look as though the candidate were in the prompt, and it succeeded by three quarters of the available distance. What it did not transfer is the thing the text prompt actually supplies: a specific package the model can read off, arithmetically evaluate against its own sheet, and table as a formal offer. This places the result inside this program's central theme and extends it. Elsewhere the finding has been that a *readout* can look real while failing to control behaviour, and that optimizing a white-box probe reward exploits the signal rather than moving the behaviour. Here the same decoupling appears in the *delivery* direction. **Low KL to a working prompt is not the same as delivering what makes that prompt work**, and the gap between them is the entire effect.

It also sharpens what note 0008 established. That note showed the model can act on good package advice; this one localizes the mechanism to one concrete act — tabling the named package as the opening offer — and any channel that does not produce that act produces nothing, however well it matches the prompt's next-token statistics.

9 Methodology deliverables

These are the parts of the lane that transfer regardless of the verdicts.

The four-baseline stack, and especially the train-set-mean bar. Predicting the training split's mean for every row — ignoring dialogue, representations, everything — is the bar most readouts in this family actually fail, and it beat a Bayesian oracle outright on threshold error. It costs nothing to compute and it is the difference between "our probe beats the published floor" and "our probe does nothing".

The two-axis split contrast. Running the identical pipeline under a split that *can* and a split that *cannot* hold out the generative unit turns a silent inflation into a visible one. Pair it with a no-dialogue baseline and a self/other control and memorization announces itself. This is now enforced in code: the head predictor refuses seed-fold checkpoints by default, keyed off the manifest rather than a tag string.

Shuffled-label nulls per configuration. Refit from scratch on permuted training labels for *every* arm, not once globally. On the gate axis this was the single most informative control: 0 of 15 arms separated from their own null.

Capacity controls for representation-level delivery. `B_shuffled` — identical encoder, identical parameter count, wrong content — is what makes a null interpretable and would have been what made a positive interpretable. Any virtual-token or soft-prompt arm should ship one.

Two capture rules, both found by a reference-mode check. (1) *On a thinking model, treat* `apply_chat_template` *as a function of the whole message list, not of a prefix.* Qwen3 injects an empty `<think></think>` block into the final assistant message and omits it when a user message follows, so re-rendering a truncated view silently rewrites an *earlier* message and breaks prefix-based span slicing. Truncate the render you already have. (2) *Verify capture numerics in float32; ship in float16, not bfloat16.* Attention reduction order depends on sequence length and Qwen3's outlier residual dimensions (magnitude $\sim$100) amplify it: the same comparison gave cosine 0.826 with max-abs deviation 126 in bf16 and cosine 1.0000 with max-abs 6×10^{-4} in fp32. fp16's 10 mantissa bits handle those activations where bf16's 7 cannot, at the same memory and speed. Gate additionally on something dtype-independent — here, a pooled-token-count delta of at most 1.

Controls-by-default, in numbers. Four times in this lane a control changed the conclusion rather than decorating it: the train-mean bar demoted a probe that beat the published floor; the own-sheet baseline and the self-span control jointly demoted an apparently strong seed-axis result; and the `UNSPECIFIED` arm's hard 0.0% mention rate is the only reason the confabulation reading is safe rather than speculative.

Preregistered gates evaluated once. Both experiments stamped their thresholds before looking, and both reported the resulting failures with the same prominence a pass would have received — including one gate (§6) reported as *uninformative at this base rate* rather than being quietly dropped or spun as a violation.

10 Future work — and one item that graduated

Nothing below is evidence. The first item is listed because it was proposed *here* as future work and has since been built and gate-backed; everything after it remains un-run.

- **Soft tokens plus text naming.** Deliver the candidate as virtual tokens *and* name it in the surrounding text, to test whether the failure is the embedding channel's capacity to carry a specific symbolic object or the absence of a cue to attend to it as content.
- **The restatability hypothesis — DELIVERED, not future work.** This was implemented as rung R1's gate and it passed at 0.9583 against a 0.0000 no-channel control (§7), replacing distillation KL as the primary objective. It behaved as hoped: the payload became restatable *and* the behavioural effect appeared. It is retained here only to record that the screen this section proposed is now a shipped, gate-backed component.
- **The V-norm-matched re-run (not run).** R1's injected keys are norm-pinned by QK-RMSNorm while its values are unconstrained, so the encoder may buy influence through value magnitude rather than attention share — which would explain why a channel the identified circuit barely attends to still moves behaviour. Measure injected V norms against real tokens' at the same positions and re-run norm-matched.
- **A placement rung (not run).** Write the payload onto positions that carry *real text*, where the uptake heads' queries already go, rather than onto reserved placeholders that draw 0.0016 attention before anything is written into them. With R1 positive this targets the residual 31% gap to text rather than a failed channel.
- **Deal rate as a co-primary (not run).** R2's excluded β=2 narrow-span cell posts the best Nash welfare in its grid at a bounded cost in agreements — a trade no arm in that design can express while deal rate is a safety gate.
- **Re-run the readout on an un-starved corpus.** At $\geq$4,096 tokens per turn (§5) the corpus would actually contain the dialogue the hypothesis is about. The validated 240-game diversified bank was built and is preserved for exactly this.
- **Stage C.** Joint LoRA plus head was gated behind stage B improving over the baselines and never ran; it remains the untested route to "the information is there but not linearly available".

A Commands and replication

All paths are on the Stanford NLP cluster. Every artifact's manifest or summary JSON carries its own `invocation` field as its first entry, so any table in this document is reproducible from the file alone.

M0 — baselines (CPU).

```
uv run --no-sync python -m tom.run_baselines \
  --campaign advocate_baseline=.../advocate_v1_baseline_r2 \
  --campaign advocate_seat0=.../advocate_v1_seat0_r2 \
  --campaign advocate_seat1=.../advocate_v1_seat1_r2 \
  --campaign p2_qwen3_8b=.../p2_Qwen3-8B_all_llm_b2 \
  --campaign p2_qwen3_4b=.../p2_Qwen3-4B_all_llm_b2 \
  --campaign p2_ministral_8b=.../p2_Ministral-8B_all_llm_b2 \
```

```
  --campaign p2_olmo3_7b=.../p2_Olmo-3-7B_all_llm_b2 \
  --campaign p2_gemma3_4b=.../p2_Gemma-3-4B_all_llm_b2 \
  --campaign framed_abstract=.../framed_mixed_qwen3_8b_abstract_b2ml \
  --campaign framed_datacenter=.../framed_mixed_qwen3_8b_datacenter_b2ml \
  --info private --stride 2 --text-seeds 5 --text-max-features 1200 \
  --out .../tom_baselines_v1
```

M1 — dataset, capture, probe evaluation.

```
python -m tom.dataset --campaign <6 campaign dirs> --out .../phase0/dataset_all
python -m tom.subset  --dataset .../dataset_all --out .../dataset_p2_private \
                      --info private --n-parties 2
python -m tom.subset  --dataset .../dataset_all --out .../dataset_p6_private \
                      --info private --n-parties 6

# reference check on every new shape FIRST, then production capture
sbatch tom/tom_capture_phase0.sbatch p2_private ref  float16   # job 16428853
sbatch tom/tom_capture_phase0.sbatch p2_private full float16   # job 16428871
sbatch tom/tom_capture_phase0.sbatch p6_private ref  float16   # job 16428840
sbatch tom/tom_capture_phase0.sbatch p6_private full float16   # job 16428872
sbatch tom/tom_capture_numerics_check.sbatch                   # job 16428816

# stage B, then the single evaluation scoring stage A, stage B and all four baselines
sbatch tom/tom_probe_phase0.sbatch p2_private
sbatch tom/tom_probe_phase0.sbatch p6_private                  # job 16428973
```

Channel addendum — distillation, campaign, analysis.

```
python -m tom.nbs_channel_distill --bank instances_nbs_confirmatory_v1 \
  --logged-run .../nbsconfirm_v1_nbs_candidate \
  --out .../tom_channel_v1/distill --model Qwen/Qwen3-8B
python -m tom.nbs_channel_eval     --out .../tom_channel_v1/campaign --seeds 0..9
python -m tom.nbs_channel_analysis --campaign .../tom_channel_v1/campaign \
                                   --out .../tom_channel_v1/analysis
```

This document.

```
UV_LINK_MODE=copy uv run --no-sync python make_figures.py   # all 4 vectorized PDFs
latexmk -pdf tom_lane.tex
```

No figure is hand-drawn and no figure number is typed by hand: `make_figures.py` reads the same JSON bundles the research notes cite, so a panel cannot drift from its table. The one exception is the cap A/B table in Figure 2, transcribed from the prose report and marked as such in the script.

B Artifact ledger

Everything is under `/nlp/scr/siddharth/ii_mats/rational_agents/`; nothing is in the repository (the `/juice2` partition is full).

Path	Contents
tom_baselines_v1/	M0: summary.json, baselines.md, three row-level CSVs, run.log (first line = exact invocation)
tom_probe_v1/RESULTS.md	The frozen M1 verdict bundle
tom_probe_v1/eval_p2_private.{json,md}	M1 held-out-game axis (27 arms)
tom_probe_v1/eval_p6_private{,_stageA}.{json,md}	M1 held-out-seed axis (jobs 16428973 / 16428978)
tom_probe_v1/head_*_p{2,6}_concat.json	Nominated-readout stage B (jobs 16429137 / 16429179)
tom_probe_v1/weights/	Head checkpoints, format tom-head/1; fold name disambiguates shape
tom_features_v1/phase0/dataset_*	Per-turn datasets: examples.jsonl, views.jsonl, manifests
tom_features_v1/phase0/features_*	Captured residuals, fp16 .npz (66 MB / 435 MB)
tom_features_v1/phase0/numerics_check.json	bf16-vs-fp32 diagnostic (job 16428816)
tom_rollouts_v1/placeholder_report.md	The cap-starvation measurement and A/B
tom_rollouts_v1/cap_ab/	2048 / 4096 / 8192 trees, logs, browsable transcripts
tom_rollouts_v1/aborted_thinking_off/	Quarantined M2 false start (192 episodes, unused)
tom_banks/instances_tom_v1/	The 240-game diversified bank: built, validated, never campaigned
tom_channel_v1/analysis/	RESULTS.md, results.json, episode_rows.csv (960 rows)
tom_channel_v1/campaign/	Four arms: episodes, manifests, transcripts; selected_manifest.json
tom_channel_v1/distill/	encoder.pt, encoder_untrained.pt, report, history, offline W&B run

Public datasets.

- M0 + M1 tables: https://huggingface.co/datasets/siddharthmb/2026.RA.ToM-Hidden-Preference-P
- Channel addendum: https://huggingface.co/datasets/siddharthmb/2026.RA.NBS-Channel-Comparis
- Source episodes and transcripts: https://huggingface.co/datasets/siddharthmb/2026.RA.Negotiation-Campaigns

W&B. The distillation run was logged *offline* and never synced, so there is no public URL; the run directory is at tom_channel_v1/distill/wandb/offline-run-20260801_155747-8u6ivo7y/. M0 is a CPU replay and M1's fits are seconds-scale solves logged to JSON; neither is W&B-instrumented.

Environment. Recorded in each results JSON under analysis_provenance: interlens 8da24d80 (clean), parent repo 42cb667a, torch 2.12.1+cu130, transformers 4.57.6, numpy 2.4.2. Encoder checkpoint SHA-256 8c7709c64f65...

The attention-interface ladder (§7). All under tom_qkv_v1/, mirrored from the pod:

Public tables, with every rung's narrative shipped under raw/: siddharthmb/2026.RA.QKV-Attention-Interfa Weights & Biases: R1 training 06whz5uz and campaign msgmv9oo; R2 sweep analysis 5nzftbdb and held-out evaluation 29qjn2u6 (earlier analysis runs of both are superseded, with identical numbers and revised prose only).

Figures 5, 6 and 7 are generated from those JSON bundles by make_figures.py --only salience ladder r2dose; no number in them is typed by hand.

Path	Contents
r0/	concentration.json (layer scan, per-head span mass, the frozen set), groupcontrol.json (the KV-group-matched control and the complement), kvmass.json (the five-condition salience table), validate{,_transcripts}.json (all 80 generated opening turns), the two shipped interface specs
r1/free_v1/	encoder checkpoint, kv_train_report.json (gate, architecture, split), reconstruction transcripts including the no-injection control
r1/campaign/, r1/analysis/	the 960 episodes with transcripts; results.json and episode_rows.csv
r2/sweep/, r2/analysis_sweep/	936 sweep episodes; frozen_config.json, exogenous_dose.json, SWEEP.md
r2/eval/, r2/analysis_eval/	720 held-out episodes and 738 transcripts; results.json, manipulation_check.json, RESULTS.md

C Files and documents

Code. experiments/rational_agents/tom/: dataset.py, subset.py, views.py, labels.py, baselines.py, run_baselines.py, features.py, probe.py, head.py, train_probe.py, eval_tom.py, delivery.py, rollout_tom.py, generate_diverse_bank.py, validate_bank.py, and the four nbs_channel modules (nbs_channel.py, _distill, _eval, _analysis), plus the three sbatch files. Tests: tests/test_tom_probe.py (19 tests — planted-label recovery for both stages, shuffled-label collapse to chance, care-mask correctness, variable-shape padding, fold disjointness, cluster-bootstrap widening, paired-difference sign convention, gate logic, checkpoint round-trip), tests/test_tom_dataset.py, tests/test_tom_rollout.py.

Documents. Preregistration: proposals/2026-08-01-tom-hidden-preference-head.md (Amendment 1 records the M1 no-go and preregisters the channel addendum). Lane hub: tom/README.md. Research notes: **0025** (M0 baselines and the M1 two-axis negative), **0027** (the channel addendum). Prior results this lane builds on: 0008 (NBS decision support), 0009 (realistic advocate viability). Program-level record: PROGRESS.md.

Deviations from the preregistration, logged. Feature capture is float16 rather than the specified bfloat16 (forced by the numerics finding, §9); the top-issue distribution is derived from the weight branch as softmax(weight logits/T) with T calibrated on validation, adding no parameters and never touching test data; the ridge/logistic penalty is selected on one inner fold of the training rows and applied identically to the probe and to the text and own-sheet baselines, so no arm wins on tuning effort; variable issue/option counts are handled by padding with masking rather than per-shape output slicing, so one head serves both banks; and 6-party stage B was restricted to the nominated layer 18 as a cost decision on an axis that cannot support a claim.