PRICING PRESENTED INFORMATION: A DUAL-CONTROL ESTIMAND FOR REWARD SACRIFICE IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Actions in interactive tasks can trade immediate reward for feedback that improves future decisions. We give a behavioural estimand for how much a language-model policy pays for such feedback: the exchange rate β_P/β_C between immediate success probability C and expected feedback information P that its choices reveal, recovered by a conditional logit with a neutralised negative control. Dual control theory motivates and delimits the estimand: an optimal policy must price probing, and probing has strict value exactly when feedback can push the posterior across a future decision boundary. We build benchmarks where this price has bite: 161 validated Mastermind states and 30 Wordle states in which guaranteed success *requires* a reward-sacrificing probe, and a structurally different fault-repair environment with noisy feedback and stochastic reward. Across ten systems from two open-weight families and three frontier models, the estimand is well identified, robust to decoding temperature, symbol identity and reasoning budget, and differs sharply between systems in ways parameter count does not explain. The central finding concerns what is being priced. Masking and deception tests show that the valuation attaches to legible presented statistics: scrambling them returns models to certainty-equivalent play, and hiding them collapses Sonnet 5's valuation in both environments and Qwen-32B's in fault repair, where the tables from which information value could be computed remain on screen. Consistently, the valuation is environment-specific: it predicts success in the puzzles (Spearman $\rho = 0.98$), but in pre-registered tests neither the ranking of models nor the link to success transfers to fault repair. The price language models put on information is a measurable property of how they use presented statistics, not a portable trait.

1 INTRODUCTION

In many interactive tasks, an action does two things at once: it earns reward now, and it reveals information that changes what the agent should do next. A good guess in Wordle may be one that *cannot* be the answer—it is worth nothing today, but its feedback narrows the future. Whether a policy is willing to make such trades depends on whether it puts a *price* on information. Control theory made this precise sixty years ago: Feldbaum (1961) observed that inputs both *direct* the plant and *probe* its unknowns, and that optimal policies weigh the two. This paper measures that price behaviourally in large language models and finds that what current models price is the information they are *shown*: the exchange rate is well identified and differs sharply between systems (Fig. 2), but it attaches to legible presented statistics and does not transfer as a ranking of models from one environment to another (Figs. 4 and 5).

Three factors make this difficult to isolate. First, "exploration" is often operationalised as regret in bandit tasks (Krishnamurthy et al., 2024; Nie et al., 2024), which conflates the *valuation* of information with belief tracking, memory, and arithmetic. Second, environments where probing provably matters are rare: in most benchmarks a greedy policy is near-optimal, so a model can score well without ever pricing information. Third, prompting can induce apparent information-seeking behaviour while incapacity can suppress it, complicating attribution.

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

We address all three. **(i)** §2 states the estimand—the exchange rate between immediate success probability C and feedback information P that a policy's choices reveal—and why standard evaluations cannot recover it. **(ii)** §3 builds benchmarks where the estimand has bite: exact solvers for Mastermind 4×6 and Wordle locate the *sacrifice band*, states where no consistent guess can guarantee success in the remaining turns but a sacrificial guess can (161 validated states, 30 for Wordle; certainty-equivalent (CE) 0.739, oracle 1.000), and a structurally different fault-repair environment with noisy feedback and stochastic reward (150 states, exact dynamic-programming oracle). **(iii)** §4 fits a conditional logit over order-randomised choice sets, with a neutralised environment in which the coefficient is provably unidentifiable as a negative control. **(iv)** §6 gives the dual-control account that motivates the estimand and delimits it: neutrality implies certainty-equivalent separability, its converse is false, and probing is strictly valuable exactly when feedback crosses a future decision boundary. **(v)** §5.1 and §5.2 ask what is being priced, with deception and masking tests in both environments and a transfer test registered before its data were collected.

Findings (§5). *The estimand is well identified.* Across ten systems (Qwen2.5 7B–72B, OLMo-2 7B–32B, Claude Sonnet 5 and Opus 5, GPT-5) the neutralised control is null for every model, and the estimate is invariant to decoding temperature, to colour relabelling and, bidirectionally, to inference-time reasoning. It differs sharply between systems: on Mastermind it is significantly negative for Qwen-7B/14B and all three OLMo-2 models and positive for Qwen-32B and the three frontier models (GPT-5: $+0.490\,[+0.457, +0.525]$), and OLMo-32B is negative at the size where Qwen-32B is positive, so parameter count does not determine it. *What is priced is the presented statistic.* Scrambling the displayed information returns models to certainty-equivalent play; hiding it collapses Sonnet 5's valuation in both environments and Qwen-32B's in fault repair, even though the fault-repair tables from which information value could be computed remain on screen. *The valuation is environment-specific.* In the puzzles it is strongly associated with success (Spearman $\rho = 0.976$ over ten systems), but in fault repair, with hypotheses registered before collection, every system prices information positively while the ranking of models reverses ($\rho = -0.70$) and the link to success disappears ($\rho = -0.71$).

2 Measuring sacrificial information seeking

Consider a Wordle position in which four words remain and two guesses are left. Guessing one of the four wins immediately with probability $1/4$, but its feedback may leave three candidates and a coin flip. A fifth word—one that *cannot* be the answer—wins nothing now, yet its feedback may split the four candidates into singletons and guarantee a win next turn. The second guess trades reward for information. We call such an action **sacrificial**, and the willingness to take it is what this paper measures.

Two statistics per action. For a candidate set S (the guesses still consistent with the history) and a guess g we define

$$C(g) = \Pr[g \text{ is the answer}] = \tfrac{1}{|S|}\mathbf{1}[g \in S], \qquad P(g) = -\sum_j \tfrac{|S_j|}{|S|}\log_2 \tfrac{|S_j|}{|S|}, \tag{1}$$

where $\{S_j\}$ is the partition of S induced by the feedback to g. C is the immediate success probability; P is the expected information (in bits) the feedback reveals. A sacrificial guess is one with $C(g) = 0$ and high $P(g)$. In the four-word example above, the consistent guess has $(C, P) = (0.25, 0.81)$ and the sacrificial one $(0, 1.5)$.

The estimand. We do not ask whether a model wins; we ask what *exchange rate* between C and P its choices reveal. Fitting a conditional logit $U = \beta_C C + \beta_P P$ over the actions it takes yields the scale-free ratio β_P/β_C: how much immediate success probability the policy is willing to give up for one bit of information. This number, not accuracy, is our object of study; §6 explains why it is the right one.

Why standard benchmarks cannot identify it. Two obstacles make this measurement fail by default. First, on most benchmarks a greedy policy is near-optimal, so a model can score well while pricing information at zero; the estimand is present but exerts no leverage. We therefore build a benchmark of states where greedy play *provably* cannot guarantee success (§3). Second, regret in

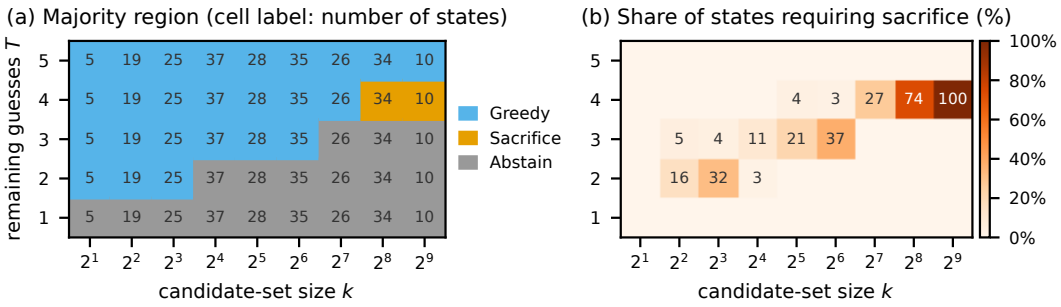

Figure 1: Mastermind 4×6 phase diagram over sampled reachable states. (a) Region containing most states of each cell. (b) Share of states in each cell that require a sacrificial guess.

bandit-style evaluations conflates valuation with belief tracking, memory and arithmetic: a model that fails may be unwilling or merely unable. We therefore present the candidate set and its statistics explicitly (§4), which isolates valuation from bookkeeping—and, as §5.1 shows, turns the reliance on those presented statistics into a measurable quantity of its own.

3 The Sacrifice Benchmark

Mastermind as a belief MDP. Four pegs, six colours; the hyperstate is (T, S) with S the consistent-candidate set. Every guess is simultaneously a task action and a probe. General Mastermind satisfiability is NP-complete (Stuckman & Zhang, 2006), but the 4×6 instance admits exact analysis. Our exact feasibility search reproduces the classical anchors: no 4-guess guarantee exists and a 5-guess strategy does (Knuth, 1976), and a beam-exact Wordle DP reproduces the SALET optimum of Bertsimas & Paskov (2024) to 4×10^{-4} (3.4216 vs. 3.4212).

Phase diagram. Over sampled reachable candidate sets, minimax feasibility splits the $(\log k \times T)$ plane into greedy / sacrifice / abstain regions (Fig. 1); the sacrifice region concentrates along the greedy/abstain frontier—dual behaviour matters precisely when time is scarce relative to uncertainty.

Terminology. A *consistent* guess is one still in the candidate set S. The *certainty-equivalent* (CE) policy maximises immediate success probability, breaking ties uniformly among consistent guesses; it is the natural greedy baseline. A *sacrificial* guess has lower immediate success probability but higher continuation value; in our bank the relevant sacrificial guesses have $C = 0$. The *oracle* is the exact dynamic-programming policy. Note that "no consistent strategy can *guarantee* success" does not mean CE always fails: CE still wins on a subset of secrets, which is why its average success on the bank is 0.739 rather than zero.

The sacrifice bank. 161 validated mid-game states, stratified by sacrifice depth ($T_{\text{cons}} - T_{\text{all}} \in \{1, 2\}$) and $|S|$ bins (2–300). A structural finding: depth-2 states exist only for $k \leq 64$—within horizon 5, consistent play is never more than one guess worse when uncertainty is large. Bank baselines: CE 0.739, oracle 1.000 (sacrificial rate 0.415). Wordle: 30 states (CE 0.68, oracle 1.00).

A structurally different environment: fault repair. Mastermind and Wordle share one structure: deterministic feedback that partitions a candidate set. To test whether the valuation is specific to it, we add a Bayes-adaptive task with stochastic reward and noisy feedback. A machine has one of $K{=}4$ hidden faults with belief b; each of six repairs a fixes it with probability θ_{za}, and a failed repair shows one of three indicator colours drawn from ϕ_{az}. Every repair is a reward-bearing task action whose outcome is informative, so there are no pure queries (Def. 1). We set $C(a) = \sum_z b(z)\theta_{za}$ and $P(a) = I(z; Y_a)$, the mutual information between the fault and the repair's outcome, which reduces to Eq. (1) when feedback is deterministic. An exact belief-space dynamic program gives the oracle. The bank holds 150 states (50 each for $T \in \{2, 3, 4\}$) in which the unique optimal repair is sacrificial (C at least 0.05 below the greedy maximum) and greedy play loses at least 0.10 in exact success probability (exact expected success: CE 0.686, oracle 0.803; success is below one because fixes are

stochastic). Two properties separate it from the puzzles: C and P are nearly uncorrelated within a choice set ($r = -0.06$, against -0.66 in Mastermind), and always taking the most informative repair is heavily penalised (success 0.43), so the task rewards trading information against reward rather than maximising either.

4 BEHAVIOURAL IDENTIFICATION

Choice sets. At every decision we present at most 24 candidates (all or sampled consistent guesses plus the highest-information inconsistent guesses) with two displayed statistics: C = probability the guess is the secret, and P = expected feedback information (bits). Candidate composition and presentation order are deterministic in (state, repeat, step)—identical for every model. This is a control, not a convenience: under plain prompting a mid-sized model cannot even track S (pilot: Qwen-14B produced 95% inconsistent guesses and 0% success), which would confound belief maintenance with valuation.

Conditional logit. We use McFadden's conditional logit (McFadden, 1974; Train, 2009): $U = \beta_C C + \beta_P P + \varepsilon$ (Gumbel). We report the scale-free price β_P/β_C, nested likelihood-ratio tests of $\beta_P = 0$, state-clustered bootstrap CIs, and a pooled logit with per-model $\beta_P^{(m)}$ interactions and a precision-weighted trend test across scale ranks (all pre-registered and frozen before main data collection).

Interpretation of the estimate. We read β_P/β_C as a revealed willingness to pay: the amount of immediate win probability the policy trades for one additional bit of expected feedback information. The denominator is well behaved — β_C is significantly positive for every model (z between 10.5 and 19.7), and across all state-clustered bootstrap resamples the smallest β_C is 10.3 — so the ratio is not driven by a near-zero denominator.

Identification checks. Three concerns arise naturally. *(i) Is β_C merely "prefer consistent guesses"?* Adding an explicit consistent-guess indicator to the utility leaves the information coefficient significant and of the same sign in every model (e.g. Qwen-14B $\beta_P/\beta_C = -0.388$, $z = -20.3$; Qwen-32B $+0.160$, $z = 11.5$; Sonnet 5 $+0.428$, $z = 17.0$), so the effect is not absorbed by consistency. *(ii) Do C and P vary independently enough?* Within a choice set they are negatively correlated (mean $r = -0.66$): consistent guesses tend to be less informative. This is a property of the task, not of our construction, and it reduces — but does not remove — identifying variation; the design condition number stays below 6, and the deliberately included high-information inconsistent guesses supply the off-diagonal variation. *(iii) Does the choice-set construction create the effect?* We vary the candidate-set size and randomly relabel Mastermind colours per decision as controls (App. C).

Negative control. Win/lose-only feedback makes P constant within the consistent class, so β_P is unidentified: all seven pre-registered models return LR $p > 0.8$, and the three OLMo-2 models $p \geq 0.50$.

5 RESULTS

Models (abbreviated Qwen-7B, …, OLMo-32B below): Qwen2.5-Instruct 7B, 14B, 32B and 72B (AWQ-quantised; Qwen Team, 2024; Lin et al., 2024), Claude Sonnet 5 and Opus 5 (Anthropic, 2026), and GPT-5 (OpenAI, 2025); and, as a second open-weight family registered before collection, OLMo-2-Instruct 7B, 13B and 32B (bfloat16 (BF16); Team OLMo et al., 2024), chosen because its training data are fully public.

Sign varies across systems; parameter count does not determine it. Table 1 and Fig. 2: among the seven pre-registered systems the valuation is significantly negative (95% CI below zero) for Qwen-7B and Qwen-14B and significantly positive for Qwen-32B and every frontier model, which are all larger than any open-weight model (pooled trend across the registered scale ranks, $z = 35.5$). The second open-weight family shows that this is not a parameter-count effect. All three OLMo-2 models are significantly negative on Mastermind, including the 32B model (-0.086) that matches

Table 1: Main results. Mastermind 4×6: 322 episodes per model; Wordle: 60 episodes per model. Success intervals are Wilson binomial 95% CIs; valuation intervals are state-clustered bootstrap 95% CIs (leading zeros omitted). "Sacr." is the fraction of decisions on a sacrificial candidate, here one with $C = 0$. CE and oracle are the greedy and exact dynamic-programming baselines on the same states. OLMo-2 rows come from an extension registered before its collection; OLMo-7B's success is depressed by unparsable replies (128 of 322 episodes).

	Mastermind 4×6			Wordle	
	success	sacr.	β_P/β_C	success	β_P/β_C
CE baseline	0.739 [.688, .784]	0.000	—	0.683 [.558, .786]	—
Qwen-7B	0.699 [.647, .746]	0.058	−0.181 [−.260, −.112]	0.600 [.474, .714]	−0.058 [−.106, −.006]
Qwen-14B	0.668 [.615, .717]	0.032	−0.190 [−.263, −.113]	0.533 [.409, .654]	−0.012 [−.018, −.009]
Qwen-32B	0.801 [.754, .841]	0.155	+0.089 [+.062, +.121]	0.900 [.799, .953]	+0.182 [+.135, +.259]
Qwen-72B	0.742 [.692, .787]	0.086	−0.040 [−.092, +.002]	0.783 [.664, .869]	+0.036 [+.026, +.053]
OLMo-7B	0.211 [.170, .259]	0.532	−0.576 [−.755, −.427]	0.467 [.346, .591]	−0.192 [−.376, −.057]
OLMo-13B	0.634 [.580, .684]	0.296	−0.262 [−.318, −.202]	0.667 [.541, .773]	−0.191 [−.311, −.094]
OLMo-32B	0.745 [.695, .790]	0.116	−0.086 [−.128, −.036]	0.700 [.575, .801]	−0.056 [−.136, +.012]
Sonnet 5	0.929 [.895, .952]	0.269	+0.374 [+.328, +.414]	0.900 [.799, .953]	+0.268 [+.191, +.396]
Opus 5	0.925 [.891, .949]	0.216	+0.376 [+.337, +.417]	0.967 [.886, .991]	+0.321 [+.232, +.442]
GPT-5	0.972 [.948, .985]	0.309	+0.490 [+.457, +.525]	0.983 [.911, .997]	+0.474 [+.384, +.532]
Oracle	1.000 [.988, 1.000]	0.464	—	1.000 [.940, 1.000]	—

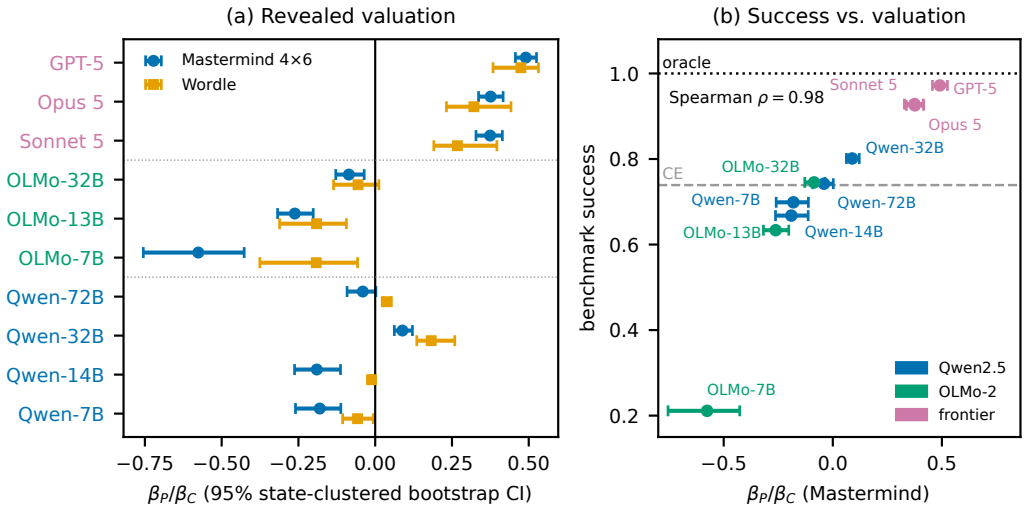

Figure 2: (a) Revealed valuation β_P/β_C of each system on Mastermind and Wordle. (b) Mastermind valuation against sacrifice-band success for all ten systems.

the positive Qwen-32B in size. Within OLMo the per-model valuation rises with size ($-0.576 \to -0.262 \to -0.086$), but the registered pooled trend test, which shares one β_C across models, is not significant ($z = -1.89$); OLMo-7B's choices are also the noisiest (128 of 322 episodes end in an unparsable reply). Qwen-72B is non-monotonic and environment-dependent: its interval contains zero on Mastermind, where it has the largest β_C of all models (entrenched exploitation), yet it is significantly positive on Wordle. We therefore claim no scaling law: the sign differs between families of equal size and is not monotone within a family, and parameter count, training data and vendor are confounded. What is robust is that the valuation is measurable and differs sharply between systems; in the puzzles it also predicts performance, a link whose limits §5.2 tests.

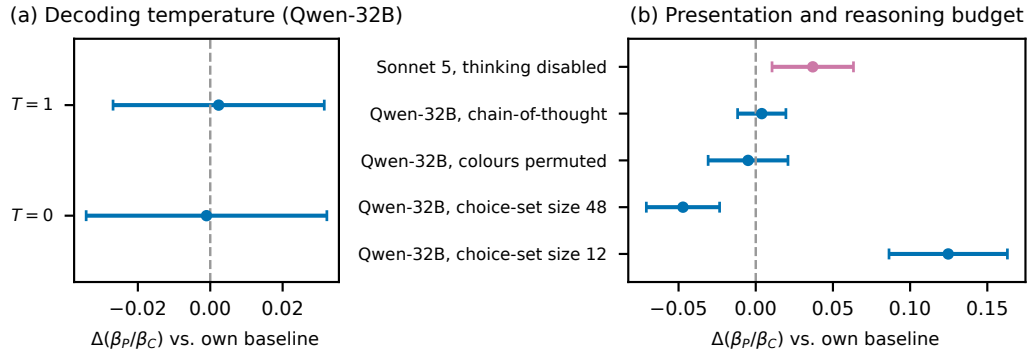

Figure 3: Robustness controls for Qwen-32B and Claude Sonnet 5, expressed as deviations from each model's own baseline on the same 161 states: (a) decoding temperature; (b) presentation and reasoning budget. Error bars are the arm's own bootstrap half-width (a conservative proxy: the paired design shares states and choice sets, so the true uncertainty on Δ is smaller). Temperature, colour relabelling and reasoning budget leave the estimate unchanged; the number of displayed alternatives shifts its magnitude.

Valuation is strongly associated with performance. Across all ten systems, Spearman $\rho = 0.976$ between valuation and Mastermind success (one-sided $p = 7{\times}10^{-7}$; $\rho = 0.964$ over the seven pre-registered systems). Every significantly positive model beats CE; no significantly negative model does so significantly (the best, OLMo-32B, reaches 0.745 against 0.739).

Robustness. Figure 3 reports each control as a deviation from the same model's own baseline on the same states and choice sets. Decoding temperature has no effect on Qwen-32B ($\Delta = -0.001$ at $T{=}0$, $+0.002$ at $T{=}1$); randomly relabelling the Mastermind colours at every decision has none either ($\Delta = -0.005$), so the estimate does not depend on the identity of the displayed symbols. Reasoning budget is ruled out bidirectionally: explicit chain-of-thought (Wei et al., 2022b) leaves Qwen-32B unchanged ($\Delta = +0.004$) and disabling Claude Sonnet 5's thinking leaves it unchanged in sign and only slightly higher ($\Delta = +0.037$). The effect is therefore not explained by the tested inference-time reasoning-budget manipulations; we make no claim about where in training it arises.

One control does move the estimate: the size of the presented choice set. With 12 alternatives Qwen-32B's valuation rises to $+0.213$ and with 48 it falls to $+0.042$ (baseline $+0.089$ at 24). Sign, significance and the direction of every comparison are preserved, but the magnitude is not invariant to how many alternatives are shown — unsurprisingly, since the ratio is defined relative to the spread of C and P actually offered. Cross-model comparisons are therefore valid only at a fixed choice-set construction, which our deterministic protocol enforces; absolute magnitudes should not be compared across studies that display different numbers of options.

5.1 Mechanism: what is being priced?

A deception/masking test separates "internal valuation" from "use of the presented statistics" by manipulating the displayed information-value column (P column; Fig. 4; 161 episodes per arm). When the column is *scrambled* across candidates, the coefficient on the true information value collapses to null for both models (Sonnet 5 $z = 0.4$, Qwen-32B $z = 0.9$), and both fall to certainty-equivalent success (0.739 and 0.733). What they do with the misleading numbers differs by an order of magnitude: Qwen-32B follows them ($\beta_P/\beta_C^{\text{shown}} = 0.235$, $z = 8.7$), Sonnet 5 barely does (0.020, $z = 2.5$). When the column is *hidden*, Sonnet 5's valuation collapses ($+0.374 \to +0.023$) and explicit chain-of-thought does not restore it ($+0.018$), whereas Qwen-32B's is unchanged ($+0.089 \to +0.085$). Fault repair repeats the test in a setting where information value is computable from what is displayed: the *specification* (spec) arm hides the P column but shows the full fix and indicator tables. Sonnet 5 falls from $+0.082$ to $+0.018$ ($+0.027$ in an exploratory rerun with a larger token budget, which removes the 56 replies truncated in the registered run; success is then 0.727,

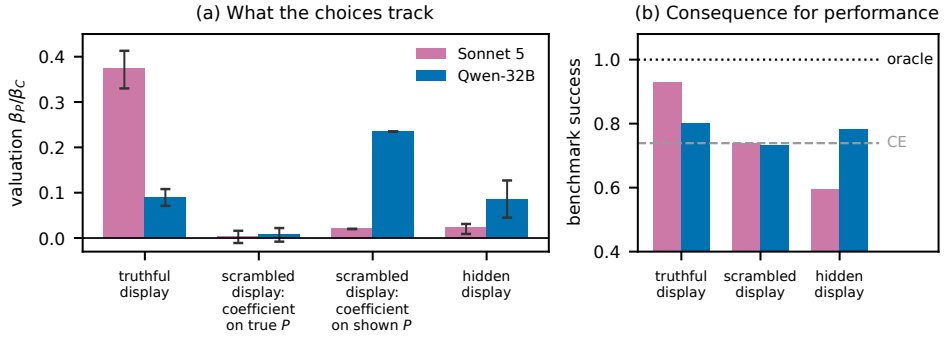

Figure 4: Deception/masking test. (a) Valuation under a truthful display, a scrambled display (split into the coefficient on the true P and on the shown P) and a hidden display. (b) The corresponding benchmark success. Sonnet 5's valuation lives on the display and disappears without it; Qwen-32B's survives masking but is overridden by a corrupted display.

close to CE), and Qwen-32B falls from $+0.313$ to $+0.014$, so its display-independent signal from Mastermind does not reappear.

Two distinct profiles follow. The frontier model's valuation requires statistics that are both *legible and credible*: it neither computes information value unaided nor follows a corrupted display, and without a trustworthy display it reverts to certainty-equivalent play. In Mastermind the mid-sized model carries a weak but display-independent signal that a misleading display can nonetheless override; in fault repair it has none. The main results of §5 are unaffected—every main arm displays truthful statistics—but a reading of β_P/β_C as internally computed information value is ruled out, and we state our claims throughout as valuation of *presented* information.

Small models systematically favour lower-information alternatives. Qwen-7B and Qwen-14B systematically prefer low-information guesses among equals—consistent with next-token plausibility favouring "safe" repeats; under plain prompting their play collapses entirely (App. B).

5.2 Does the valuation transfer? A structurally different environment

In fault repair (§3; 300 episodes per model; hypotheses and analysis registered before collection) every one of the ten systems prices information positively, and all three frontier intervals exclude zero (Fig. 5; Table 2 in App. C). The other three registered predictions fail. The ranking of models does not carry over from Mastermind (Spearman $\rho = -0.70$ between environments), valuation is negatively associated with fault-repair success ($\rho = -0.71$), and the scale trend reverses ($z = -3.02$). The OLMo models, which avoided information in the puzzles, overpay for it here ($+0.89$ to $+1.59$, with 65–77% sacrificial choices), and OLMo-7B and 13B are the only systems clearly below CE. A noise-free rule that maximises $C + \lambda P$ does best at $\lambda \approx 0.2$, which equals the oracle's success on the same episodes, and the frontier and Qwen valuations lie near that value ($+0.08$ to $+0.31$). Two post-hoc, exploratory analyses examine what success is associated with here: it tracks the precision of choices (β_C; $\rho = 0.82$, $p = 0.004$) more than the distance of the valuation from $\lambda \approx 0.2$ ($\rho = -0.49$, $p = 0.15$). The exchange rate is therefore measurable, and positive in every frontier model, but it is not a portable trait that ranks models across environments, and its link to performance does not transfer from the puzzles. Together with the masking results of §5.1, this places the valuation in how a model uses the statistics a task presents rather than in a portable disposition.

6 Why this estimand: dual effects on belief MDPs

The estimand of §2 is not an arbitrary regression coefficient: dual control theory says that an optimal policy in these tasks *must* carry such an exchange rate, and tells us where it has bite. We summarise the results here and defer proofs and secondary propositions to Appendix A.

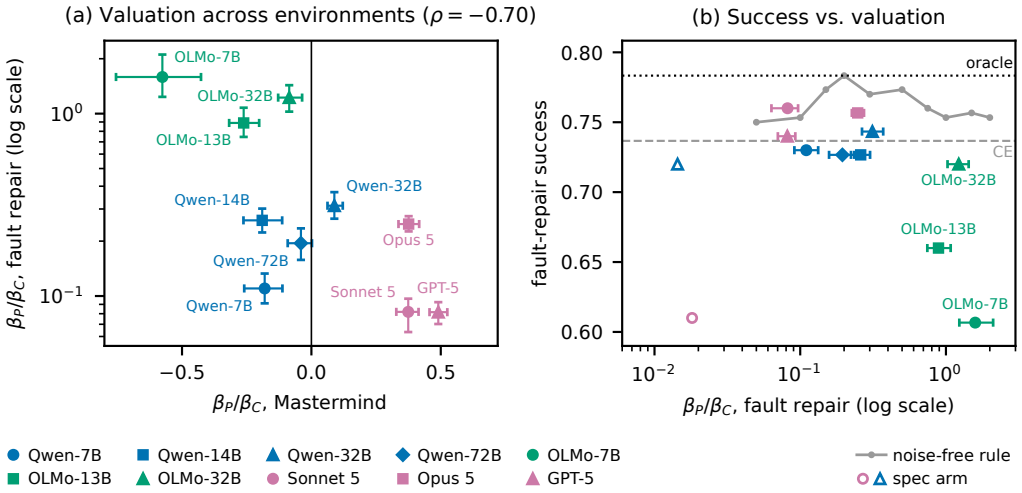

Figure 5: Fault repair. (a) Mastermind valuation (x) against fault-repair valuation (y) per model; the ranking does not transfer ($\rho = -0.70$). (b) Fault-repair success against valuation, with the success of the noise-free rules $\arg\max_a [C(a) + \lambda P(a)]$ on the same episodes, plotted at $x = \lambda$, and the spec arms, which hide the P column but show the full fix and indicator tables.

Setting. Latent parameter $z \in \mathcal{Z}$ (finite), state s, action a, observation o, kernel $p(o \mid z, a, s)$, reward $r(s, z, a)$, horizon T. Belief $b \in \Delta(\mathcal{Z})$, hyperstate $x = (s, b)$, Bayes operator $K_a(b, o)$. $V_t(s, \cdot)$ is piecewise linear convex (Smallwood & Sondik, 1973); its linear pieces induce *decision regions*. This is the standard belief-MDP view of POMDPs (Kaelbling et al., 1998; Duff, 2002; Ghavamzadeh et al., 2015). In Mastermind, b is uniform on the candidate set S, so $x \cong (T, S)$.

Definition 1 (strict dual effect)**.** *Action a has a* strict dual effect *at x if (i) it moves immediate expected reward, (ii) within its task-equivalence class it still changes the law of the posterior, and (iii) reward and observation dependence do not factorise across separate action coordinates. Condition (iii) is what makes the effect a property of* task *actions: pure queries—actions whose reward is a fixed cost, as in EIG-driven question asking—live in the factorised case and are excluded (Prop. 3).*

Theorem 1 (neutrality $\Rightarrow$ separability; converse false)**.** *If $p(o \mid z, a, s)$ does not depend on a, the belief process is exogenous and the certainty-equivalent (CE) policy is optimal: planning separates. The converse fails. There are belief MDPs with action-dependent kernels and genuine decision boundaries in which CE is optimal at every reachable hyperstate; we exhibit one verified in exact rational arithmetic, and in random $2 \times 2 \times 2$ instances with decision boundaries 74% are CE-optimal everywhere (App. B).*

So "beliefs depend on actions" is not enough to make exploration valuable. The sharp criterion is:

Theorem 2 (probing value)**.** *Let $\tilde{K}$ be an* admissible neutralisation*: an action-independent coarsening of every K_a (win/lose-only feedback qualifies). Then $Q_t(x, a) = \mathbb{E}_b r + \underbrace{[\cdot]}_{\text{caution } \geq 0} + \mathrm{PV}_t(x, a)$ with $\mathrm{PV}_t \geq 0$, and, under a genericity condition on ties, $\mathrm{PV}_t(x, a) > 0$ iff feedback from a pushes the posterior across a decision boundary of V_{t-1} with positive probability.*

Information is worth paying for when it can change what the agent would do next, not merely when it reduces entropy; this is why our benchmark targets states near decision boundaries rather than states with large P. The theorem provides a geometric interpretation of when information carries strict continuation value; ties are common in these games (App. A), and our empirical estimation does not rely on the genericity assumption.

Proposition 1 (CE never sacrifices)**.** *Call an action a sacrificial at x if some a' has strictly higher immediate expected reward while $Q_t(x, a) > Q_t(x, a')$. The CE policy never takes a sacrificial action; wherever the unique optimal action is sacrificial, CE is strictly suboptimal.*

Proposition 1 converts the theory into an observable: the set of hyperstates where the optimal action is sacrificial—the *sacrifice band*—is exactly where a CE-like policy must fail, and is what §3 constructs.

7 Related work

Dual control and adaptive control. Feldbaum (1961) introduced the directing/probing duality and Bar-Shalom & Tse (1974) the moment-level dual effect; surveys trace the field's development (Wittenmark, 1995; Mesbah, 2018), and certainty-equivalence versus cautious or dual behaviour is a classical theme of adaptive control (Åström & Wittenmark, 2008). Bayes-adaptive formulations place the trade-off inside a belief MDP (Duff, 2002; Ghavamzadeh et al., 2015), where large-scale planners approximate it (Silver & Veness, 2010).

Exploration and the value of information. Bandit exploration runs from Thompson sampling (Thompson, 1933) through UCB (Auer et al., 2002) to information-directed sampling (Russo & Van Roy, 2018), which *designs* explorers that price information (see Lattimore & Szepesvári, 2020; Sutton & Barto, 2018); the value of information goes back to Howard (1966) and, for experiment selection, to Lindley (1956) and Bayesian experimental design (Chaloner & Verdinelli, 1995). Where design chooses experiments, we *measure* whether a fixed policy prices information.

Exploration in language models. Bandit-style evaluations (Krishnamurthy et al., 2024; Nie et al., 2024) report regret, which conflates valuation with tracking and memory; our choice-set design controls belief maintenance and the neutralised environment provides a negative control that regret lacks. In-context RL shows transformers *can* be trained to explore (Laskin et al., 2022; Lee et al., 2023); we ask whether generally trained LLMs already price information without such training.

Question asking and information seeking. EIG-driven question selection — from cognitive models of human question asking (Rothe et al., 2017) to LLM clarification strategies — prices information for *pure queries*; by Prop. 3 pure queries have no dual effect—the phenomenon we measure requires reward-bearing actions.

Puzzles, POMDPs and emergence. Modern online POMDP planning manages observation complexity by simplification with deterministic guarantees (Barenboim & Indelman, 2023); rollout methods solve Wordle/Mastermind-like puzzles near-optimally (Bertsekas, 2022; Bhambri et al., 2023). Our contribution is orthogonal: we do not propose a planner but diagnose whether a *fixed, learned* policy implicitly prices the value of information. Knuth (1976) and Bertsimas & Paskov (2024) give the exact anchors we reproduce; prior LLM evaluations on these games report accuracy only. Our metric is continuous and interval-estimated per model, so the sign flip is a property of CIs, not of thresholded success metrics — the locus of the emergence debate (Wei et al., 2022a; Schaeffer et al., 2023).

8 Limitations

(1) Choice-set prompting changes the task; it is our control for belief tracking, and plain-prompt results are reported alongside. (2) Qwen models are AWQ-quantised (Lin et al., 2024) while OLMo models run in BF16, so the family comparison mixes quantisation with training; the reasoning ablation rules out inference-time compute but not quantisation as a residual confound. (3) The scale ranking mixes parameters with vendor and training differences; within Qwen the reversal is 14B→32B and 72B is non-monotonic, and the second family is negative at every size, so we make no claim about parameter count as such. (4) The distribution-to-moment link is only an analogy (Prop. 2); the converse of Thm. 1 is false, so our claims are confined to the sacrifice band where probing provably matters. (5) β_P prices *displayed* information; the deception/masking test shows this is not merely a caveat but the mechanism: frontier pricing is display-mediated (and display-verifying), while a weak internal signal exists in Qwen-32B. Whether any current model can *compute* partition entropy unaided remains open. (6) The valuation does not transfer as a ranking of models: in fault repair the registered cross-environment and performance tests fail, so any single model's exchange rate is environment-specific, and the puzzle-level link to success should not be read as a general law.

REPRODUCIBILITY STATEMENT

The supplementary archive contains all environments, state banks, collectors, per-decision data for every model and arm, and the frozen analyses; its README maps each table and figure to the file and command that produce it. Definitions and proofs are in App. A, the converse construction and random-instance protocol in App. B, and full results, robustness arms and every registered outcome in App. C. The main study was frozen at tag `v0.3-frozen` before data collection, and the two extensions at `v0.4a-frozen` and `v0.4b-frozen`, each before its own collection; post-hoc analyses are labelled as such. Total compute: $\sim$13 rented GPU-hours and $\sim$\$82 of API usage.

AI USE STATEMENT

We used a generative AI tool (a commercial LLM-based writing assistant) to aid and polish the writing of this paper: to improve the grammar, wording and clarity of our drafts, to translate internal notes into English, and to help revise the presentation of text, tables and figure captions. All AI-assisted text was reviewed and edited by the authors. Separately, large language models are the objects of study in this work; their use as experimental subjects is described in §5. We take full responsibility for the final content of this paper.

REFERENCES

Anthropic. Claude Opus 5 and Sonnet 5 model cards. `https://www.anthropic.com`, 2026.

Karl J. Åström and Björn Wittenmark. *Adaptive Control*. Dover, 2nd edition, 2008.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Yaakov Bar-Shalom and Edison Tse. Dual effect, certainty equivalence, and separation in stochastic control. *IEEE Transactions on Automatic Control*, 19(5):494–500, 1974.

Moran Barenboim and Vadim Indelman. Online POMDP planning with anytime deterministic guarantees. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Dimitri Bertsekas. Rollout algorithms and approximate dynamic programming for Bayesian optimization and sequential estimation. *arXiv preprint arXiv:2212.07998*, 2022.

Dimitris Bertsimas and Alexander Paskov. An exact solution to Wordle. *Operations Research*, 2024. DOI 10.1287/opre.2022.0434.

Siddhant Bhambri, Amrita Bhattacharjee, and Dimitri Bertsekas. Playing Wordle using an online rollout algorithm for deterministic POMDPs. In *IEEE Conference on Games*, 2023.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Michael O. Duff. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, University of Massachusetts Amherst, 2002.

Alexander A. Feldbaum. Dual control theory. I. *Automation and Remote Control*, 21(9):874–880, 1961. Translation of Avtomatika i Telemekhanika 21(9):1240–1249, 1960.

Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. Bayesian reinforcement learning: A survey. *Foundations and Trends in Machine Learning*, 8(5-6):359–483, 2015.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966.

Leslie P. Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

Donald E. Knuth. The computer as Master Mind. *Journal of Recreational Mathematics*, 9(1):1–6, 1976. 1976–77 volume.

Akshay Krishnamurthy, Keegan Harris, Dylan J. Foster, Cyril Zhang, and Aleksandrs Slivkins. Can large language models explore in-context? *Advances in Neural Information Processing Systems*, 37:120124–120158, 2024.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, Maxime Gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Mnih. In-context reinforcement learning with algorithm distillation. *arXiv preprint arXiv:2210.14215*, 2022.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Jonathan N. Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*, 2024.

Dennis V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka (ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press, 1974.

Ali Mesbah. Stochastic model predictive control with active uncertainty learning: A survey on dual control. *Annual Reviews in Control*, 45:107–117, 2018.

Allen Nie, Yi Su, Bo Chang, Jonathan N. Lee, Ed H. Chi, Quoc V. Le, and Minmin Chen. EVOLvE: Evaluating and optimizing LLMs for exploration. *arXiv preprint arXiv:2410.06238*, 2024.

OpenAI. GPT-5 system card. `https://openai.com`, 2025.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Anselm Rothe, Brenden M. Lake, and Todd M. Gureckis. Question asking as program generation. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36, 2023.

David Silver and Joel Veness. Monte-Carlo planning in large POMDPs. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.

Jeff Stuckman and Guo-Qiang Zhang. Mastermind is NP-complete. *arXiv preprint cs/0512049*, 2006.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Kenneth E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2nd edition, 2009.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022b.

Björn Wittenmark. Adaptive dual control methods: An overview. In *Proc. 5th IFAC Symposium on Adaptive Systems in Control and Signal Processing*, pp. 67–72, 1995.

A Proofs and secondary results

Proof of Thm. 1 (sufficiency). Induction on t: with an a-independent kernel the continuation term $\sum_o \bar{P}(o \mid b) V_{t-1}(s', \bar{K}(b,o))$ contains no a, so $Q_t(x,a) = \mathbb{E}_b r(s,z,a) + C_t(x)$. *Corollary (general f).* If $s' = f(s,a)$ while the kernel is a-independent, the belief process is still exogenous and the problem reduces to a fully observed MDP on (s,b); the probing term vanishes but the optimal policy need not be myopic. We keep the two claims separate to avoid over-claiming.

Proof of Thm. 2. On a single decision region D, $V_{t-1}(s',\cdot) = \langle \alpha_D, \cdot \rangle$ is linear, and the posterior is a martingale, so all posteriors landing in D contribute exactly $\langle \alpha_D, b \rangle$; the same holds for $\tilde{K}$, giving $\mathrm{PV} = 0$. If a posterior $b'' = K_a(b,o'') \notin D$ occurs with positive probability, then $V_{t-1}(s',b'') > \langle \alpha_D, b'' \rangle$ under genericity condition (G) — no tie between the two maximisers at b'' — so the expectation strictly exceeds the neutralised one. Without (G) only $\mathrm{PV} \geq 0$ holds. Ties are the rule rather than the exception: in Mastermind $3{\times}4$, 99% of sampled candidate sets admit several minimax-optimal guesses, so (G) is to be read after a fixed tie-breaking, and the empirical claims of §5 rest on the estimated β_P, not on pointwise strictness. *Multi-region case.* When the neutralised posteriors straddle several regions, $\mathrm{PV} \geq 0$ still follows whenever the K_a-partition refines the $\tilde{K}$-partition (convex order / data processing); we therefore *define* admissible neutralisations as coarsenings of every K_a, which makes non-negativity universal and Thm. 2 the strictness criterion.

Worked example for (G) and admissibility. Mastermind $3{\times}4$, $S = \{ABC, ABD, ACD\}$, $T = 2$. Win/lose-only feedback maps every guess to the same binary partition, hence is a coarsening of every K_a and admissible; its "else" posterior (uniform on two survivors) lies in one decision region of V_1. A guess with three feedback classes splits S into singletons: every posterior is degenerate and outside that region, no two α's tie at a vertex, so (G) holds and $\mathrm{PV} > 0$. A guess with classes of sizes $(2,1)$ crosses only on its singleton branch; if the two-survivor maximiser also attains the value there, (G) fails and the crossing contributes nothing.

Proposition 2 (distribution $\Rightarrow$ moment level; analogue only)**.** *If the law-to-covariance map is injective on the induced family, then the posterior covariance depends on the action within the task class — the moment-level dual effect of Bar-Shalom & Tse (1974). The hypothesis is substantive: with the mean pinned by the martingale property, two-point posterior laws with* $(p, \textit{spread}) = (\frac{1}{2}, 2)$ *and* $(\frac{1}{5}, \frac{5}{2})$ *along one direction share mean* and *covariance yet differ as laws (third moments* 0 *vs.* 1.5*); moments up to order three restore injectivity for two-point laws. The moment-level notion is therefore strictly weaker than Def. 1, and we use it only as an analogue.*

Proposition 3 (pure queries are not dual)**.** *If $r(s,z,a) = -c(a)$ for all z, then a has no strict dual effect: it fails condition (i) of Def. 1 and lies in the observation-only factor whenever a factorisation exists.*

B Exact verification of the converse counterexample

Instance: $\mathcal{Z} = \mathcal{A} = \mathcal{O} = \{0,1\}$, $T = 3$, uniform prior; $r = \left(\begin{smallmatrix} 1 & 2 \\ 1 & 0 \end{smallmatrix}\right)$, $p(o{=}1 \mid z,a) = \left(\begin{smallmatrix} .646 & .430 \\ .296 & .017 \end{smallmatrix}\right)$. With exact rational arithmetic: the kernel depends on a; the reward matrix has a decision boundary; the CE action is optimal at all 15 reachable beliefs for $t \in \{1,2,3\}$ (script `lineD/theory/converse_exact.py`). Hence action-dependent beliefs do not imply inseparability.

Random-instance protocol behind the 74% statistic. Instances: $|\mathcal{Z}| = |\mathcal{A}| = |\mathcal{O}| = 2$, horizon $T = 3$, uniform prior. Rewards $r(z, a)$ i.i.d. uniform on $\{0, 1, 2\}$; observation kernels $p(\cdot \mid z, a)$ i.i.d. Dirichlet$(1, 1)$ per (z, a). Non-neutrality filter: $p(\cdot \mid z, 0) \neq p(\cdot \mid z, 1)$ for some z (almost surely satisfied). Decision-boundary filter: the sets $\arg\max_a r(z, a)$ for $z = 0, 1$ are disjoint, so the optimal action switches across the belief simplex. For each instance we enumerate all reachable beliefs over three stages by exact Bayes updates and test whether some myopically optimal action attains V_t at every reachable belief (tolerance 10^{-9}). Of 3,000 sampled instances, 667 passed both filters; 495 (74%) were CE-optimal everywhere. Script: `lineD/theory/tn_converse_search.py`.

C ADDITIONAL TABLES

Neutralised-control LR p-values (pre-registered models > 0.8, OLMo-2 models ≥ 0.50); plain-prompt collapse (pilot); temperature sweep; per-cell bank composition (depth $\times$ k-bin); pre-registration deviations log. Full numbers in the released `analysis/out_lineD.json`.

Robustness arms (Qwen-32B, 161 states each). Baseline $\beta_P/\beta_C = +0.089$ (success 0.801, sacrificial rate 0.155). $T = 0$: $+0.088$ $[+0.060, +0.126]$, success 0.783. $T = 1$: $+0.091$ $[+0.064, +0.122]$, success 0.814. Colours randomly permuted per decision: $+0.084$ $[+0.054, +0.106]$, success 0.839. Choice set of 12: $+0.213$ $[+0.178, +0.255]$, success 0.870. Choice set of 48: $+0.042$ $[+0.018, +0.066]$, success 0.783. Explicit chain-of-thought: $+0.093$ $[+0.076, +0.107]$. An earlier temperature sweep on the 40-state pilot bank gave $0.129/0.150/0.124$; the numbers above supersede it, since they use the main bank and the same states as every other arm.

Identification diagnostics. Per model, β_C is significantly positive (z from 10.5 to 19.7) and the smallest β_C over all state-clustered bootstrap resamples is 10.3, so the ratio has no denominator pathology. Within a choice set, $\mathrm{corr}(C, P) = -0.66$ on average (design condition number < 6). Adding a consistent-guess indicator: Qwen-7B $\beta_P/\beta_C = -0.283$ ($z = -18.2$), Qwen-14B -0.388 (-20.3), Qwen-32B $+0.160$ (11.5), Qwen-72B -0.038 (-3.1), Sonnet 5 $+0.428$ (17.0), Opus 5 $+0.794$ (17.3), GPT-5 $+0.661$ (16.1): signs are unchanged and the information coefficient remains significant in every case.

Deception/masking test. For a fully deterministic subset (161 episodes/arm) we manipulate the displayed P column: **perm** (values permuted across candidates; fit $U = \beta_C C + \beta_P^{\text{true}} P_{\text{true}} + \beta_P^{\text{disp}} P_{\text{disp}}$) and **hidden** (column omitted). Sonnet 5: perm $\Rightarrow$ true-value coefficient null ($z = 0.4$), shown-value coefficient small but nonzero (0.020, $z = 2.5$), success 0.739 (=CE); hidden $\Rightarrow$ ratio $+0.023$, and $+0.018$ with explicit chain-of-thought. Qwen-32B: hidden $\Rightarrow$ $+0.085$ (vs. $+0.089$ baseline); perm $\Rightarrow$ follows the shown value (0.235, $z = 8.7$) while the true-value coefficient is null ($z = 0.9$), success 0.733. One earlier Sonnet perm run contained 2/444 misaligned decisions due to a logging offset (fixed; conclusions unchanged).

Registered extensions: outcomes. Both extensions were registered before their data were collected; their outcomes, in full. Second open-weight family (OLMo-2): A1, within-family trend, $z = -1.89$ (rejected; the per-model valuation rises with size but the pooled test is not significant); A2, ten-model Spearman between Mastermind valuation and success, $\rho = 0.976$ (supported); A3, neutralised control, LR $p \geq 0.50$ for all three models (supported); Wordle reported descriptively in Table 1. Fault repair: B1, cross-environment Spearman, $\rho = -0.70$ (rejected); B2, every frontier interval above zero (supported); B3, valuation–success Spearman, $\rho = -0.71$ (rejected); B4, pooled trend over Qwen and frontier models, $z = -3.02$ (rejected). Two post-hoc analyses, not tests: success tracks β_C ($\rho = 0.82$, $p = 0.004$) more than the distance of the valuation from the best noise-free rate $\lambda \approx 0.2$ ($\rho = -0.49$, $p = 0.15$).

Table 2: Fault repair: 300 episodes per model on the 150-state bank. Success intervals are Wilson 95% CIs; valuation intervals are state-clustered bootstrap 95% CIs. "Sacr." is the fraction of decisions on a sacrificial repair, here one whose C is below the maximum (in Mastermind, Table 1, one with $C = 0$). Specification (spec) arms hide the P column and show the full fix and indicator tables. The CE and oracle rows give the success realised on the same 300 episodes; their exact expected success over the bank is 0.686 and 0.803 (§3).

	success	sacr.	β_P/β_C
CE baseline	0.737 [.684, .783]	0.000	—
Qwen-7B	0.730 [.677, .777]	0.378	$+0.110$ [$+.091, +.133$]
Qwen-14B	0.727 [.674, .774]	0.460	$+0.260$ [$+.223, +.301$]
Qwen-32B	0.743 [.691, .789]	0.399	$+0.313$ [$+.265, +.371$]
Qwen-72B	0.727 [.674, .774]	0.404	$+0.195$ [$+.158, +.235$]
OLMo-7B	0.607 [.550, .660]	0.774	$+1.588$ [$+1.236, +2.108$]
OLMo-13B	0.660 [.605, .711]	0.677	$+0.889$ [$+.745, +1.077$]
OLMo-32B	0.720 [.667, .768]	0.647	$+1.224$ [$+1.026, +1.432$]
Sonnet 5	0.760 [.709, .805]	0.196	$+0.082$ [$+.063, +.097$]
Opus 5	0.757 [.705, .802]	0.378	$+0.248$ [$+.226, +.274$]
GPT-5	0.740 [.688, .786]	0.193	$+0.082$ [$+.070, +.093$]
Sonnet 5, spec	0.610 [.554, .663]	0.099	$+0.018$ [$+.009, +.029$]
Sonnet 5, spec, 8192 tokens[†]	0.727 [.674, .774]	0.150	$+0.027$ [$+.016, +.040$]
Qwen-32B, spec	0.720 [.667, .768]	0.101	$+0.014$ [$+.006, +.023$]
Oracle	0.783 [.733, .826]	0.521	—

[†]Exploratory rerun; the registered run truncated 56 of 543 replies at 2048 tokens, and those episodes count as failures.