SENSITIVITY OR CRITERION? MEASURING AND MOVING WHEN LANGUAGE MODELS ACT

Anonymous authors
Paper under double-blind review

ABSTRACT

Before an agent acts it must decide whether any action is warranted. Benchmarks that ask this question report one accuracy, which cannot distinguish a model that fails to recognize a warranted action from one that recognizes it but is reluctant to act. Signal-detection theory separates the two as sensitivity and criterion, but a criterion estimated from authored binary labels depends on each benchmark's mix of examples and cannot be compared across benchmarks. We introduce FLIPACT, a diagnostic built on a qualitative representation of what tools do to physical states. A rule over annotated tool properties computes each question's answer and a graded evidence score, with the Yes/No boundary fixed at zero for every model, so every answer is computed, not written by a model. Questions are built from 2×2 quads that hold the required knowledge fixed while the state flips. The slope of a model's Yes-rate against the score is its sensitivity, and the score where it crosses one half is its criterion. Across 44 open models, sensitivity rises with size within a family, and its ranking agrees with three of four external when-to-act benchmarks. Models respond to the sign of the evidence rather than its size, accepting actions that leave the state no better about as often as helpful ones. When Yes means act, most models are conservative and stay so when the question is rephrased. In some models this is a preference for answering No, and the criterion is not set by size. Activation steering moves the criterion with the weights fixed, and distilling a model from its own steered copy moves it in the weights, without a significant change in sensitivity. The distilled shift carries to six external benchmarks with sensitivity kept, including two without a Yes/No answer. Accuracy rises where the model started conservative and falls where it started permissive.

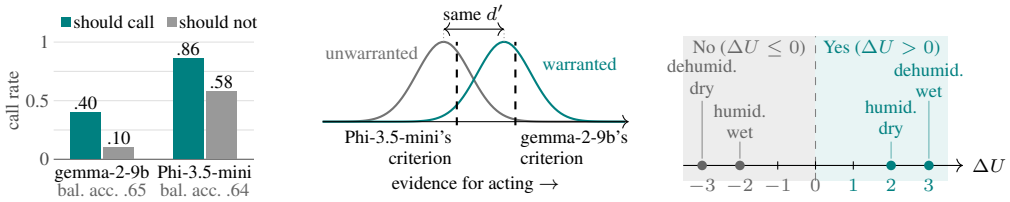
1 INTRODUCTION

Large language model (LLM) agents must decide whether to act before they decide how. Benchmarks for agents mostly score the action once taken: whether the chosen function, its arguments and the trajectory are correct (Patil et al., 2025; Huang et al., 2024). Those that study the decision itself, in tool calling (Ross et al., 2025; Sun et al., 2026), clarification (Li et al., 2026) and abstention (Kirichenko et al., 2026), report a single accuracy, and that number mixes two ways to fail. A model may not recognize whether a situation warrants action, a failure of its *sensitivity*; or it may recognize the situation and still be too reluctant or too ready to act, a failure of its *criterion*. Too conservative a criterion refuses benign requests (Röttger et al., 2024; Cui et al., 2024; Shi et al., 2024) or answers from memory where a tool was needed; too permissive a criterion acts where nothing warrants it.

On When2Call (Ross et al., 2025), gemma-2-9b and Phi-3.5-mini reach balanced accuracies of 0.65 and 0.64 with opposite behavior: the first rarely calls the tool, even where it should, and the second calls it on most examples, many where it should not (Figure 1a). They separate the two kinds of example about equally well, so on this benchmark the first should act more and the second less.

Signal-detection theory (Green et al., 1966) separates sensitivity (d') from criterion (Figure 1b; §3.1), but on a benchmark with binary labels the criterion it recovers is relative to that benchmark's mix of examples, and no zero is shared across benchmarks. Comparing criteria needs a fixed boundary on a common evidence scale, which authored binary labels do not provide (cf. Jacobs & Wallach, 2021).

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.



(a) Same accuracy, opposite calls (b) Same d', two criteria (c) Evidence computed by rule

Figure 1: (a) Share of should-call and of should-not-call When2Call examples that each model calls. (b) Schematic: at one sensitivity d', the two models' criteria trade hits for false alarms at the same accuracy. (c) The quad of Figure 2 on the evidence scale ΔU that FLIPACT computes by rule, with the Yes/No boundary fixed at zero.

FLIPACT supplies such a scale by computing the evidence for acting. Its design draws on qualitative process theory (Forbus, 1984), which describes physical change by the sign of each quantity's change and a coarse ordinal magnitude, enough to say what an action does to a state without numbers. Each question shows a situation and a candidate tool and asks whether using the tool is warranted. The state and the tool each have a sign and an ordinal magnitude on a physical axis. A rule over these properties computes the evidence score ΔU, how much closer the action brings the state to its goal, with the Yes/No boundary at zero (Figure 1c). The design unit is a 2×2 *quad*: two tools, each shown in the state it fixes and in the state it worsens, so the knowledge needed to answer stays the same and the state decides the answer (Figure 2). Fitting a model's Yes-rate against ΔU gives its sensitivity as the slope a and its criterion as the score τ where it crosses one half.

Across 44 open models (§4), sensitivity behaves as expected of a capability: it rises with size inside a family, and its ranking agrees with external d' on three of four when-to-act benchmarks. Models read the sign of the evidence rather than its size, accepting actions that leave the state no better about as often as helpful ones. When Yes means act, the criterion is mostly conservative, in some models through a preference for answering No. It is a stable per-model setting, not set by size: two releases of Mistral-Small-24B sit on opposite sides of zero.

We move the criterion separately from sensitivity in two ways (§5). At inference, adding a direction read from the model's own activations (Zou et al., 2023; Turner et al., 2023) shifts the criterion with the weights fixed. In the weights, distilling the model from its own steered copy shifts the criterion on four models from three developers with no significant change in sensitivity, and a random direction of the same norm does not. The shift carries beyond FLIPACT's Yes/No format: on benchmarks without a Yes/No answer it exceeds the shift from a teacher that only favors Yes, on both models tested. On every benchmark, accuracy rises or falls as the starting criterion predicts. Supervised training on the computed labels instead raises sensitivity and moves the criterion with it.

Contributions.

1. A measurement scheme that splits when-to-act accuracy into a sensitivity and a criterion, read against an evidence scale computed by rule with the Yes/No boundary fixed for every model.
2. FLIPACT, built by an automated pipeline in which language models only annotate what a tool does, while the states and every answer follow from the rule.
3. Evidence that the criterion moves separately from sensitivity: steering and distillation from a steered copy move the criterion, while training on the computed labels raises sensitivity.

2 RELATED WORK

When-to-act evaluation. Benchmarks for deciding whether to act cover tool calling (Ross et al., 2025; Sun et al., 2026; Huang et al., 2024; Patil et al., 2025), clarification (Li et al., 2026), abstention (Kirichenko et al., 2026) and over-refusal (Röttger et al., 2024; Cui et al., 2024; Shi et al., 2024). Most share three design choices: ground truth is authored by humans or an LLM judge, the evidence for acting is a binary label, and performance is a single rate that conflates a model's sensitivity with its criterion. Li et al. (2026) instead compute ground truth by constraint satisfaction, and Sun et al.

Template

Situation: The `<object>` is `<state>`. `[<supply>]` The goal is to bring the `<object>` back to a normal state. Action: a person uses a `<tool>`. Is using the `<tool>` a wise choice in this situation? Answer Yes or No.

One quad, object room

	humidifier wets by 2	industrial dehumidifier dries by 3
very dry	Yes +2	No −3
dangerously wet	No −2	Yes +3

Figure 2: The question template (left) and one quad (right) with each question's answer and evidence score ΔU. The slot `[<supply>]` states that a prerequisite is met, for tools that have one.

(2026) fix a controlled call/no-call boundary with deliberate no-tool cases. Neither flips a world-state while holding the required knowledge fixed, and none of these benchmarks grades the evidence for acting per question. We reuse four of them as external tests (§4.2) and measure transfer on six benchmarks (§5.3). App. F compares twelve on these design choices.

Signal detection and language models. Signal-detection theory separates an observer's sensitivity from its criterion (Green et al., 1966). Reading a benchmark score as the measurement of a latent quantity raises the question of whether it measures what it is read as (Jacobs & Wallach, 2021; Bean et al., 2026). Cacioli (2026) fits signal-detection models to factual question answering with the model's answer log-probability as evidence, so the criterion sits on each model's own confidence scale. Shi et al. (2026) read a call/no-call offset off sparse-autoencoder features on When2Call and cancel it by steering. Their offset is an intercept on an internal margin, compared within one model and without a separate sensitivity. FLIPACT places the criterion on a rule-computed evidence scale shared across questions and models, estimates it separately from sensitivity with the required knowledge held fixed, and moves it into the weights.

3 THE FLIPACT DIAGNOSTIC

3.1 PRELIMINARIES: SIGNAL DETECTION

Signal-detection theory (Green et al., 1966) models an observer who reports a signal whenever a scalar amount of evidence exceeds a threshold. The sensitivity d' is the distance between the evidence distributions of signal and noise trials in units of their spread, and the criterion c is where the threshold sits relative to their midpoint. From the hit rate H and the false-alarm rate FA the equal-variance estimators, with z the inverse of the standard normal distribution function, are

$$d' = z(H) - z(\mathrm{FA}), \qquad c = -\tfrac{1}{2}\big(z(H) + z(\mathrm{FA})\big). \tag{1}$$

In when-to-act terms a hit is acting when one should and a false alarm is acting when one should not. A positive c is conservative. FLIPACT supplies an evidence score per question, and the same two quantities become the slope a and the crossing point τ of a model's Yes-rate against that score (§4.1). Every comparison in the paper pairs a with d' and τ with c.

3.2 SAME KNOWLEDGE, FLIPPED STATE

Each question shows a physical situation, the goal of returning it to normal and one candidate tool, and asks whether using the tool is a wise choice. All questions share one minimal template (Figure 2), as do the stimuli of template-based diagnostics (McCoy et al., 2019; Ribeiro et al., 2020), so only the tool, the state and their outcome vary. Physical situations are used because a tool's effect on a state can be written as an explicit qualitative transition: a humidifier helps a dry room and harms a wet one. In the quad of Figure 2, a humidifier that adds two steps of moisture and an industrial dehumidifier that removes three are crossed with a very dry and a dangerously wet room. Every question of the quad needs the same two facts, what the tool does and what the room needs. For a given tool only the state changes, and with it the answer, so a model that knows what a humidifier is for must still read the state.

Quads and state flips are controls built into the set. The main set has 1,880 questions: 840 lie in a complete quad and 1,256 belong to a *state flip*, the same tool with opposite answers in two states. A

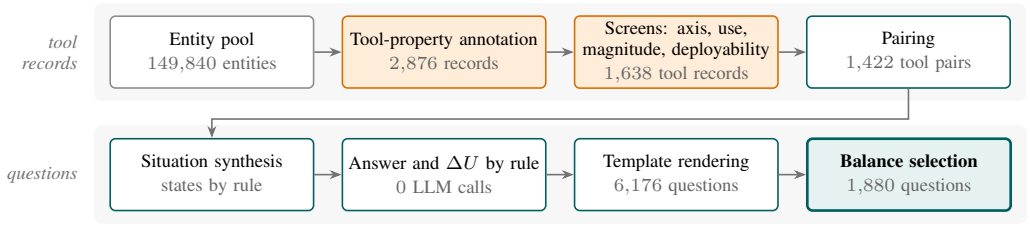


Figure 3: The FLIPACT construction pipeline. Language models are used only in the orange steps, which annotate and screen tool properties; answers and evidence scores are computed by rule.

state flip separates a model that ignores the state, which gives both questions about the same Yes–No logit margin, from one that is merely permissive, which may answer Yes to both but still ranks the warranted one higher (App. D.3, D.7). Every situation has as many Yes as No questions, so the description of the world does not predict the answer.

3.3 GROUND TRUTH AND EVIDENCE BY RULE

Each situation fixes an integer state v on a physical axis (thermal transfer, moisture, mechanical cohesion, ...) with the goal at 0. Each tool carries a direction $\delta \in \{+1, -1\}$ and an ordinal magnitude band $\mu \in \{1, 2, 3\}$ on that axis (in qualitative process theory, a sign and a place in a quantity space; Forbus, 1984), optionally conditional on a prerequisite. Applying it moves the state to $v + \delta\mu$, and the process behind that change is not modeled. The evidence score is the utility of the action, how much closer to the goal it leaves the world,

$$\Delta U = |v| - |v + \delta\mu|, \qquad \texttt{Yes} \iff \Delta U > 0,$$

and a tool with unmet prerequisites carries $\Delta U = -2$ for each, a value set by convention. The question type follows from the same computation: a wrong-direction tool is *destructive*, a right-direction tool that overshoots or acts on a state already at the goal is *redundant*, and a tool applied without its prerequisite is *prerequisite*. Because v and μ are integers, ΔU takes seven values on $[-3, 3]$. Language models annotate each tool's axis, direction, magnitude and prerequisite, and the rule computes the answer from those properties.

3.4 AN AUTOMATED BUILD

FLIPACT is generated by a pipeline (Figure 3) that needs no per-question human annotation or adjudication. Its specification is written once: the physical axes, a closed vocabulary of tool properties, the rule of §3.3 and the template. Ten lexical, medical and visually grounded resources supply 149,840 candidate entities, and two annotator models, Qwen3.5-27B and gemma-4-31b, fill the vocabulary for each. Where they disagree, both arbitrate against the specification, and unresolved records are dropped. Tool records (one tool, one axis) are paired when they act in opposite directions on that axis and target object, and states, answers, question types and text follow by rule. A balance selection keeps 1,880 of the 6,176 rendered questions. Two matched-pair supplementary sets, 342 already-satisfied and 224 prerequisite-blocked questions, are kept outside the fits. A new axis, a new tool inventory or a larger set runs through the same steps at the cost of annotator calls (App. A).

A lexical probe trained on the question text is at chance on held-out situations whose tools it has not seen (App. A.5), and recomputing the ground truth independently of the build gives zero mismatches. A model's judgment enters only where tools are annotated and screened, and a model family outside the build reproduces every tool/non-tool judgment, nine in ten axis judgments and every magnitude band to within one band, two in three exactly. Perturbing the annotations at that rate, or until a quarter of the computed answers change, keeps the rankings by a and τ intact and moves criteria toward acting, so it cannot produce the conservative criteria (App. A.6).

4 MEASURING SENSITIVITY AND CRITERION

Models. To see how the coordinates a and τ vary with size and family, we fit 44 open models, all but five with reasoning off (App. B.7), from 0.6B to 70B parameters and fifteen families, among them Qwen3 and Qwen3.5, Gemma 2–4, Llama 3, Mistral and Phi (App. Tables 9–10).

External benchmarks. To test whether the coordinates capture what when-to-act benchmarks score, we compare them with four, each posing the decision in a different domain: BFCL's irrelevance split (call only if a function applies), QuestBench (ask for a missing fact or solve), When2Call (call, ask, decline or answer directly) and When2Tool (whether a task needs a tool).

4.1 THE FIT AND THE RESPONSE PROFILE

On FLIPACT every question carries a computed evidence score, so we fit the model's Yes-rate as a logistic curve over that score (Green et al., 1966):

$$P(\text{Yes} \mid \Delta U) \;=\; \sigma\big(a\,(\Delta U - \tau)\big). \tag{2}$$

Here σ is the logistic function, the slope a the sensitivity and the crossing point τ the criterion. A positive τ means the fitted curve crosses one half above the rule's boundary, so the model answers Yes less often than the rule warrants, which we call conservative. A criterion is *estimable* when $a \geq 0.15$ and $\mathrm{SE}(\tau) \leq 2.0$, and *well determined* when it also lies inside the score range; beyond ± 3 it is reported as a bound. Both coordinates have split-half reliability above 0.9 (App. B.2). On external benchmarks we use d' and c (Eq. 1); a model readier to act has lower τ and c.

Response profile. Models read the sign of the evidence, not its size (Figure 4). Yes-rates jump between $\Delta U = -1$ and 0, significantly in 19 of the 21 reasoning-off models with a well-determined criterion, and are flat on either side: no model answers Yes reliably more often at $\Delta U = +3$ than at $+1$. The step sits one unit below the rule's boundary: models accept actions that leave the state no better about as often as helpful ones. The questions at $\Delta U = 0$ are all redundant slight overshoots, so this set cannot tell whether that placement belongs to the models or to the rule. Models do withhold when the state is already at its goal or a prerequisite is unmet (App. B.2), so the overshoot is the one redundancy they accept. The sign of τ is set by the height of the Yes plateau, not by the step: for the median model the Yes-rate on helpful actions stays near one half (Figure 4).

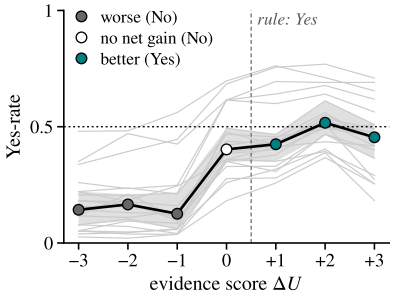


Figure 4: Yes-rate at each evidence score for the 21 reasoning-off models with a well-determined criterion (gray), with their median and interquartile band.

Reading τ. A step function at the sign of the evidence predicts held-out answers as well as the logistic. Still, τ tracks the criterion divided by sensitivity: it agrees at Spearman $\rho = 0.99$ with c/d' of Eq. 1 read from hits on $\Delta U > 0$ and false alarms on $\Delta U < 0$, and both find the same models conservative (App. B.2–B.3). In 9 of the 21 models the Yes-rate never reaches one half, so τ is where the fitted curve, not the data, crosses it. We therefore read τ by sign and rank.

4.2 SENSITIVITY GROWS WITH SIZE AND AGREES WITH EXTERNAL BENCHMARKS

Within a family, a rises with size: Qwen3 goes from 0.34 at 4B to 0.58 at 32B, and the one exception is Qwen3.5, which peaks at 9B. Twelve models fit below 0.30, among them widely deployed small instruct models, and their Yes-rates span almost the whole range: a model can be permissive or conservative about a tool and still be unable to tell when using it is warranted (Figure 5).

FLIPACT sensitivity predicts external d' on three of the four benchmarks (Figure 6a). The comparison uses a *panel* of the 18 reasoning-off models with $a \geq 0.30$, a well-determined criterion and usable external d' and c, so that both sides of each comparison are estimated reliably (App. B.4

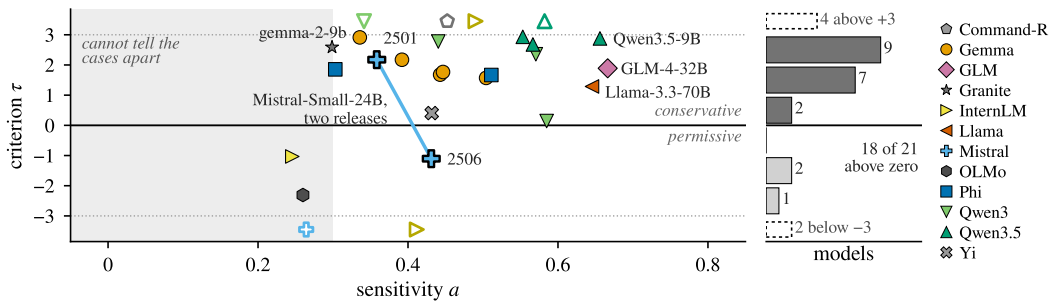
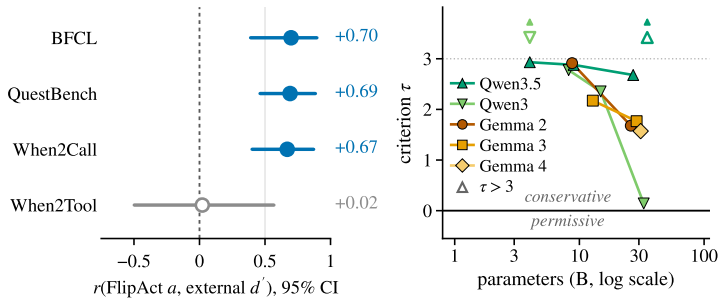


Figure 5: Sensitivity against criterion for the 27 reasoning-off models with an estimable criterion, with the distribution of the well-determined criteria on the right. Hollow markers at the edges are criteria beyond ± 3. The line joins the two same-size Mistral-Small-24B releases.



(a) Sensitivity vs. external benchmarks (b) Criterion vs. model size

Figure 6: (a) Correlation across the panel between FLIPACT sensitivity a and external d', with bootstrap intervals; the hollow point is not significant. (b) Criterion against model size for the families with three or more estimable members, one line per generation. Hollow markers with an arrow are criteria beyond $+3$.

varies the cutoff on a). All three relations hold with model size partialed out (App. B.5). This agreement across three domains and formats supports reading a as the sensitivity these benchmarks score. On When2Tool the relation is not significant.

4.3 CRITERION: A PER-MODEL SETTING, CONSERVATIVE BY DEFAULT

Eighteen of the 21 reasoning-off models with a well-determined criterion are positive on the canonical wording of Figure 2: they answer Yes less often than the rule warrants (Figure 5), and nearly all tested stay positive under five rephrasings (App. C). The criterion is one setting across question types: models that accept more destructive actions also accept more slight overshoots on other questions ($\rho = 0.88$, 27 models, $a \geq 0.30$), though most accept overshoots more readily.

As in signal detection, the criterion collects every bias in the response, including a preference for one answer word. Asking each question also in three reversed wordings, under which No means act, separates that preference from the rest of the criterion (App. C). For the Qwen3.5 models and GLM-4-32B much of the positive criterion is a preference for No, and the other twelve models tested stay conservative without it. JEV (TypeSafe AI, 2026), a model that returns a probability, not text, and so has no answer word to prefer, is conservative too and steps at the sign of the evidence (App. E).

Where the criterion sits is a property of the model more than of its size. The two Mistral-Small-24B releases are the same size and land on opposite sides of zero, at $+2.18$ (2501) and -1.10 (3.2, 2506) (Figure 5). Inside a family τ drifts toward zero as models grow, in Qwen3, in the dense Qwen3.5 models and across the Gemma generations (Figure 6b). Part of this drift follows from rising sensitivity, since τ tracks the criterion divided by sensitivity. Families differ in their typical τ, so pooled over all models the relation with size vanishes.

Table 1: Steering on When2Call, unsteered and at the selected strength, both measured on the same 600 examples. Arrows mark a move toward acting, and bold a change in that direction.

	call rate ↑		criterion c ↓		
Model	unsteered	+ steering	unsteered	+ steering	Δ acc.
Qwen3-8B	0.56	**0.60**	−0.18	**−0.32**	−0.022
Qwen3.5-27B	0.47	**0.50**	+0.10	**+0.01**	+0.005
gemma-2-9b	0.24	**0.29**	+0.78	**+0.59**	+0.002
Mistral-24B	0.37	**0.55**	+0.39	**−0.13**	−0.032
phi-4	0.57	**0.59**	−0.30	**−0.33**	−0.038

The criterion measured on FLIPACT does not consistently predict a model's c on other benchmarks: across the panel the correlation is near zero on When2Call (Pearson $r = -0.01$), QuestBench and When2Tool, and positive on BFCL ($r = +0.51$), where it rests on the three models with the lowest τ (App. Table 15). It is a stable reading of each model on FLIPACT rather than a trait that ranks models consistently across benchmarks. Section 5.3 asks instead whether moving τ moves c.

5 MOVING THE CRITERION

We ask whether the criterion can be moved while sensitivity stays fixed, at inference (§5.1) and in the weights (§5.2), and whether the moved criterion carries to external benchmarks (§5.3). Such a move trades misses for false alarms and cannot raise accuracy beyond what sensitivity allows.

5.1 STEERING MOVES THE CRITERION WITH THE WEIGHTS FIXED

Models and data. We steer five open models, Qwen3-8B, Qwen3.5-27B, gemma-2-9b, Mistral-Small-24B (the 2501 release, Mistral-24B below) and phi-4, and measure the effect on When2Call (Ross et al., 2025), on 600 examples, half call and half withhold (eleven more models in App. D.1).

Method. Steering follows the standard recipe (Turner et al., 2023; Zou et al., 2023; Rimsky et al., 2024; Arditi et al., 2024). A logistic probe from the final-token residual stream at relative depth 0.6 to the model's own Yes/No answers on the FLIPACT questions gives the criterion direction $\hat{w}$, which is added during generation to the residual-stream activation h at that layer as $h \leftarrow h + \lambda\hat{w}$. For each model the strength $\lambda^\star$ is the one with the lowest When2Call criterion among those that keep d' within 25% of its unsteered value.

Results. The direction moves the When2Call criterion toward acting on all five models (Table 1). The shift is small on Qwen3.5-27B, which starts near zero on When2Call and whose FLIPACT criterion is mostly a preference for No, and there it does not exceed random directions of the same norm. On phi-4, which starts permissive and moves least, it does not hold when the strength is chosen on one half of the examples and measured on the other (App. D.1). On four models steering gains hits at the cost of correct withholds, and on all five, accuracy changes by at most four points. With the direction fit on one half of FLIPACT and measured on the other, it also moves the criterion of Qwen3.5-9B and Qwen3.5-27B toward zero and raises their accuracy by 2.6 and 1.6 points.

5.2 DISTILLING FROM A STEERED COPY PUTS THE CRITERION INTO THE WEIGHTS

Models and data. As an existence proof that the criterion can be placed in the weights, we distill four students from three developers, Qwen3-8B, Qwen3.5-9B, gemma-3-12b and Mistral-24B, each from its own steered copy. The FLIPACT main set is split in half so that no situation or tool record appears on both sides. Students train on one half (940 questions) with three seeds per condition, and every result is measured on the held-out half, so baselines differ from the full-set values of §4.

Method. In *criterion distillation*, the teacher T is the student with steering on. Its direction is a difference of means over the student's own Yes and No answers on the training half, and its strength

Table 2: Held-out coordinates and accuracy of each student before training, after criterion distillation and after sensitivity learning on the computed labels (§5.4), seed means. Shaded rows give the change under each scheme's control. Arrows mark the targeted direction, and bold a change in it.

	Baseline			Criterion distillation			Sensitivity learning		
Student	a	τ	Acc.	a	$\tau\downarrow$	Acc. $\uparrow$	$a\uparrow$	τ	Acc. $\uparrow$
Qwen3-8B	0.47	+2.33	0.60	0.55	**−2.11**	**0.64**	**1.09**	−0.02	**0.78**
control				+0.02	+0.04	+0.00	−0.19	–	−0.04
Qwen3.5-9B	0.75	+1.88	0.65	0.73	**+0.86**	**0.70**	**1.22**	+0.18	**0.83**
control				+0.00	+0.20	−0.01	−0.30	–	−0.04
gemma-3-12b	0.36	+2.17	0.59	0.35	**+0.30**	**0.61**	**1.56**	−0.41	**0.83**
control				+0.00	−0.11	+0.00	−0.20	–	−0.07
Mistral-24B	0.59	+3.47†	0.59	0.42	**+0.19**	**0.62**	**1.16**	+0.08	**0.79**
control				+0.07	−0.10	+0.00	−0.61	–	−0.06

–: the shuffled-label criterion is estimable in only some seeds (App. Table 29). †Beyond the score range.

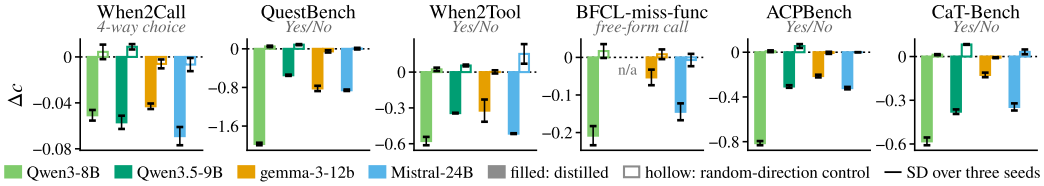


Figure 7: Criterion shift on six external benchmarks after criterion distillation (filled) and under the random-direction control (hollow), seed means ± SD. Each benchmark has its own scale; negative is toward acting. Qwen3.5-9B on BFCL-miss-func is left out because its baseline does not reproduce.

is the one on a sweep that moves τ furthest while a and d' stay close to their unsteered values. The student S has the same weights, a LoRA adapter and no steering, and is trained to match the teacher's next-token distribution at the answer position of each training question q,

$$\mathcal{L} = \sum_{q} \mathrm{KL}\big(p_T(\cdot \mid q) \,\big\|\, p_S(\cdot \mid q)\big). \tag{3}$$

The control replaces the direction with a random one of the same norm. A change is significant when it exceeds twice its paired bootstrap error over situations (settings in App. D.3).

Results. Criterion distillation moves τ toward acting on all four students (Table 2), by more than twice its error on three and just under twice on gemma-3-12b. No student's sensitivity changes significantly in any seed. An equivalence test bounds the change within ± 0.12 on Qwen3.5-9B and gemma-3-12b and cannot decide on Qwen3-8B and Mistral-24B. On When2Call, d' stays within 20% of its baseline for all four. Held-out accuracy rises by one to five points, as a criterion shift should when the model starts conservative, as all four do. The random direction moves τ at most a fifth as far. Injected at a strength that perturbs the answers as much as the teacher's direction, a random direction either leaves τ in place or collapses the model onto one answer (App. D.3).

5.3 The distilled criterion moves the same way on six external benchmarks

Benchmarks. To test whether the shift generalizes beyond FlipAct's physical questions and its Yes/No format, each student is tested on six unseen benchmarks, scored with d' and c. Four ask whether to act now: When2Call and When2Tool (Sun et al., 2026) (call a tool), QuestBench (Li et al., 2026) (ask or solve) and BFCL-miss-func (Patil et al., 2025) (the needed function is removed from the list). Two ask a planning judgment: ACPBench (Kokel et al., 2025) (does an action apply) and CaT-Bench (Lal et al., 2024) (must one step precede another). When2Call offers four replies, BFCL-miss-func takes a free-form call or refusal, and the other four take Yes or No.

Results. Moving the FlipAct criterion moves c the same way on every benchmark, although across models it does not consistently predict a benchmark's c (§4.3). On all 23 scored model–

Table 3: Change in balanced accuracy after criterion distillation for the two Qwen students, on the three benchmarks where at least one starts conservative; seed means, bold marks a rise.

Student	Benchmark	Start c	Δ bal. acc. $\uparrow$
Qwen3-8B	When2Tool	$+0.42$	$\mathbf{+0.014}$
	QuestBench	$+0.71$	$\mathbf{+0.041}$
	ACPBench	$+0.40$	$\mathbf{+0.025}$
Qwen3.5-9B	When2Tool	$+0.73$	$\mathbf{+0.046}$
	QuestBench	-0.06	-0.011
	ACPBench	$+0.65$	$\mathbf{+0.052}$

Table 4: Change in criterion-free AUC after criterion distillation, seed means. Every change is equivalent to zero within ± 0.02 (App. D.4).

	Qwen3 8B	Qwen3.5 9B	gemma-3 12b	Mistral 24B
When2Call	$-.008$	$-.006$	$+.001$	$-.002$
QuestBench	$-.012$	$-.002$	$-.005$	$+.002$
When2Tool	$+.004$	$-.001$	$+.011$	$+.001$
ACPBench	$+.001$	$+.001$	$+.004$	$-.003$
CaT-Bench	$+.001$	$+.008$	$-.003$	$+.002$

benchmark pairs, c moves toward acting in every seed, and every control moves less than the distilled student beside it (Figure 7). Sensitivity is kept: the criterion-free AUC from answer log-probabilities, available on five benchmarks, stays unchanged for every student (Table 4).

Answer preference. On a Yes/No format a shift toward acting and a shift toward Yes coincide. On FLIPACT most of the distilled shift flips sign under the reversed wordings of §4.3, so there it shifts the preference for Yes, the form the decision to act takes on a Yes/No format (App. D.5). Whether more than that preference carries over shows only where the answer has no Yes: if a Yes preference were all that transfers, a student distilled from a teacher that only raises the Yes logits would match the criterion-distilled one there. On Qwen3-8B and gemma-3-12b it does not: on When2Call and BFCL-miss-func the criterion-distilled student moves c further in every seed, though on Qwen3-8B this Yes-bias student reaches half to three quarters of the shift. On the four Yes/No benchmarks the Yes-bias student moves c as far or further. More students would make this firmer.

Accuracy. Accuracy follows the starting point, as a pure criterion shift predicts. Where a student starts conservative, balanced accuracy rises in every case, by up to six points (Table 3; all four students in App. Table 31). Where it starts permissive, accuracy falls by up to four points or stays within its error, except on Mistral-24B on When2Tool: its hit rate there is already at ceiling, so a shift toward acting can only add false alarms, and accuracy falls by about ten points. A criterion shift should thus take its sign from the target benchmark. For the five steered models of §5.1, choosing the sign and strength of steering from fifty labeled examples of each benchmark repairs every model–benchmark pair that steering toward acting degraded except Mistral-24B on When2Call, where both conditions lose accuracy (App. D.6). Where the benchmark has an answer token, a scalar bias on that token chosen the same way does as well (19 of 20 pairs), as expected if steering acts as one offset on the criterion. Unlike the bias, the direction also moves free-form replies, and distilled into the weights it carries more than the Yes-bias student where the answer has no Yes.

5.4 LABELS MOVE SENSITIVITY, AND THE CRITERION WITH IT

Sensitivity learning, supervised training on the computed labels, tests the converse: whether labels move sensitivity. It uses the same students, split, adapter and seeds, adds 263 supplementary questions on the same tool records to the training half, and shuffles the labels within matched groups of questions as its control. It raises a on every student, by about 0.5 to 1.2, and held-out accuracy by 17 to 24 points, while the shuffled-label control lowers accuracy (Table 2). It also brings τ close to zero, so labels move both coordinates, whereas distillation moves the criterion alone.

6 CONCLUSION

We treated the decision to act as a measurement problem. Two models with the same accuracy can need opposite fixes, so a when-to-act benchmark says more when it reports d' and c beside accuracy. The criterion moves, at inference or in the weights, with a shift whose sign the target benchmark sets, while sensitivity rises with training on labels. On a Yes/No format the criterion also holds a preference between answer words, which reversed wordings separate from the rest.

Ethics statement

This work evaluates open-weight models on public benchmarks and involves no human subjects. Moving a model's criterion toward acting makes it act on thinner evidence, a trade-off practitioners should weigh for their application.

Reproducibility statement

The paper specifies the full experimental setup: the construction of FlipAct and its rule (§3, App. A), the estimators (§4.1), and the steering and training procedures with all hyperparameters (§5, App. D.3). Every training condition is run with three seeds and reported as the seed mean with its bootstrap error, and every reported change comes with a significance test; steering effects carry bootstrap intervals and are checked on held-out splits. The dataset, code and evaluation logs will be released.

AI use statement

We used generative AI tools to generate synthetic data sets: in the FlipAct pipeline, language models annotate tool properties (§3.4, App. A), while every answer and evidence score is computed by rule. We used generative AI tools to implement methods: the evaluation, steering, training and analysis code was written with an AI coding assistant. We also used generative AI tools to aid and polish the writing. We did not use generative AI tools for the other tasks with required disclosure. The authors reviewed all AI-assisted code, data and text and take responsibility for the final content of this work.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024a. URL https://arxiv.org/abs/2404.14219.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024b.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998.

Aarti Basant, Abhijit Khairnar, Abhijit Paithankar, Abhinav Khattar, Adithya Renduchintala, Aditya Malte, Akhiad Bercovich, Akshay Hazare, Alejandra Rico, Aleksander Ficek, et al. Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model. *arXiv preprint arXiv:2508.14444*, 2025.

Andrew M Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan Eghlidi, Chris Schmitz, Karolina Korgul, Hunar Batra, et al. Measuring what matters: Construct validity in large language model benchmarks. *Advances in Neural Information Processing Systems*, 38, 2026.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Jon-Paul Cacioli. Llms as signal detectors: Sensitivity, bias, and the temperature-criterion analogy. *arXiv preprint arXiv:2603.14893*, 2026.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

Cohere Team. Command R: RAG at production scale. https://docs.cohere.com/docs/command-r, 2024. Accessed: 2026-09-24.

Lee J Cronbach. Response sets and test validity. *Educational and psychological measurement*, 6(4): 475–494, 1946.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130(1):33–55, 2022.

Kenneth D Forbus. Qualitative process theory. *Artificial intelligence*, 24(1-3):85–168, 1984.

Team Glm, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

IBM Granite Team. Granite 3.0 language models. *URL: https://github. com/ibm-granite/granite-3.0-language-models*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

David Marvin Green, John A Swets, et al. *Signal detection theory and psychophysics*, volume 1. Wiley New York, 1966.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5351–5359. IEEE, 2019.

Janna Hastings, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. Chebi in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*, 44(D1): D1214–D1219, 2016.

Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan, Neil Gong, et al. Metatool benchmark for large language models: Deciding whether to use tools and which to use. In *International Conference on Learning Representations*, volume 2024, pp. 42978–43007, 2024.

Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 375–385, 2021.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J Bell. Abstentionbench: Reasoning llms fail on unanswerable questions. *Advances in Neural Information Processing Systems*, 38, 2026.

Harsha Kokel, Michael Katz, Kavitha Srinivas, and Shirin Sohrabi. Acpbench: Reasoning about action, change, and planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26559–26568, 2025.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

Yash Kumar Lal, Vanya Cohen, Nathanael Chambers, Niranjan Balasubramanian, and Ray Mooney. Cat-bench: Benchmarking language model understanding of causal and temporal dependencies in plans. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19336–19354, 2024.

Belinda Li, Been Kim, and Zi Wang. Questbench: Can llms ask the right question to acquire information in reasoning tasks? *Advances in Neural Information Processing Systems*, 38, 2026.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Carolyn E Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

Xun Liu, Yi Evie Zhang, Vira Kasprova, Parisa Rabbani, Pardis Sadat Zahraei, Tianyu Zhang, Ali Ebrahimpour-Boroojeny, and Varun Chandrasekaran. Agentabstain: Do llm agents know when not to act? *arXiv preprint arXiv:2607.10059*, 2026.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3428–3448, 2019.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Mistral AI. Mistral small 3, 2025. Model release.

NVIDIA. NVIDIA-Nemotron-3-Nano-4B. https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Nano-4B-BF16, 2026. Model card.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E Gonzalez. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*, 2025.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pp. 13387–13434, 2023.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4902–4912, 2020.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Hayley Ross, Ameya Sunil Mahabaleshwarkar, and Yoshi Suhara. When2call: When (not) to call tools. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3391–3409, 2025.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024.

Chenyu Shi, Xiao Wang, Qiming Ge, Songyang Gao, Xianjun Yang, Tao Gui, Qi Zhang, Xuan-Jing Huang, Xun Zhao, and Dahua Lin. Navigating the overkill in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4602–4614, 2024.

Wei Shi, Ziheng Peng, Sihang Li, Xiting Wang, Xiang Wang, Mengnan Du, and Na Zou. To call or not to call: Diagnosing intrinsic over-calling bias in llm agents. *arXiv preprint arXiv:2605.18882*, 2026.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Jinyan Su and Claire Cardie. Knowing but not showing: Llms recognize ambiguity but rarely ask clarifying questions. *arXiv preprint arXiv:2605.25284*, 2026.

Chung-En Sun, Linbo Liu, Ge Yan, Zimo Wang, and Tsui-Wei Weng. Llm agents already know when to call tools–even without reasoning. *arXiv preprint arXiv:2605.09252*, 2026.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

TypeSafe AI. TypeSafe AI documentation: System one and the noul, choice and score question types. `https://docs.typesafe.ai`, 2026. Accessed 2026-09-22; model `jev-latest`, API version jev-1.13.0.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

Wikipedia contributors. Wikipedia, the free encyclopedia. `https://www.wikipedia.org`, 2026. Accessed 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A CONSTRUCTION OF FLIPACT

This section details the construction summarized in §3: the annotation schema, the rule that turns a tool into its questions, a worked quad, the build with its counts, and the checks on data quality and on annotation noise.

A.1 ANNOTATION SCHEMA

Each annotator model (App. A.4) fills one record per candidate entity against the closed vocabulary of Table 5. A value outside the vocabulary is a parse failure and is never mapped to a neighboring value. A tool that acts on two axes becomes one record per axis, and each record carries a single direction.

Table 5: The annotation schema: the six physical axes, each with its two problem poles as the questions word them and an example tool in each direction, and the other fields of a tool record.

			Example tool	
Axis	Quantity	Problem poles	adds ($\delta = +1$)	removes ($\delta = -1$)
Thermal transfer	temperature	cold / hot	pocket hand warmer	ice pack
Radiant emission	visible light	dark / glaringly bright	electric lamp	blackout curtain
Mechanical cohesion	connectedness	loose / jammed	thread lock compound	puller
Spatial containment	seal	exposed / sealed shut	lid	cap opener
Electromagnetism	circuit continuity	without power / overloaded	ac adapter	emergency stop button
Moisture level	moisture	dry / wet	humidifier	industrial dehumidifier
Other fields of a record				
is-tool	yes / no; a record that is not a tool leaves the build			
Direction δ	$+1$ or -1			
Magnitude μ	1 local, on one object or spot (hot water bottle, paper towel); 2 room or person scale (hair dryer, space heater); 3 building, vehicle or industrial scale (furnace, industrial dehumidifier)			
Target object	16 classes and *other*, e.g. a person's body, a room, a container, a rigid object or joint, a circuit or device, soil or a plant			
Prerequisite	mains electricity, running water, fuel			
Side effect	any of the six axes, or electronic integrity, where the tool has one; not used by the rule			

A.2 FROM A TOOL TO ITS QUESTIONS

The state of every question is computed from the tool, with the goal at 0. A solution question places the state at $v = -\delta\mu$, the deficit the tool closes. A destructive question uses the deficit of the paired tool, which points the other way, so the tool widens the gap. A redundant question uses a deficit of one met by a magnitude of at least two, so the tool crosses the goal and leaves the state no closer. A prerequisite question removes the supply a tool depends on from a solution state. Two tools are paired only when they act on the same axis and target object in opposite directions and differ in magnitude by at most one band, so each tool's solution state is its partner's destructive state. A pocket hand warmer pairs with an ice pack on a person, never with an air conditioner in a room. Prerequisite questions and questions whose state is already at the goal are not part of the main set. They go to the two supplementary sets of §3.4.

Every question is rendered from the template of Figure 2, with the state adverb *slightly*, *very* or *dangerously* for $|v| = 1, 2, 3$. For a tool with a prerequisite, a sentence between the state and the goal reads "The supplies this tool depends on (mains electricity) are available nearby" or "None of the supplies this tool depends on (mains electricity) is available anywhere nearby." It appears in every question of such a tool, including those whose supply is present, so its presence carries no information about the answer.

A.3 A WORKED QUAD

Table 6 shows one tool pair on the thermal-transfer axis with a person's body as the target. Tool X is a blanket, which warms ($\delta = +1$, $\mu = 1$), and tool Y is an ice bath, which cools ($\delta = -1$, $\mu = 2$). State A is *slightly cold* ($v = -1$) and state B is *very hot* ($v = +2$). The two off-diagonal questions are destructive: cooling a body that is already cold, and wrapping one that is already hot. The fifth question keeps the ice bath and lowers the deficit to one (*slightly hot*). The ice bath still points the right way, but its magnitude of two now overshoots the goal, so magnitude alone separates this No from the ice bath's Yes in state B.

Table 6: One tool pair: the four questions of the quad and the ice bath's redundant question, with the rule's computation for each.

State		Tool			Outcome			
	v		δ	μ	$v + \delta\mu$	ΔU	Type	Answer
A: slightly cold	-1	blanket	$+1$	1	0	$+1$	solution	Yes
	-1	ice bath	-1	2	-3	-2	destructive	No
B: very hot	$+2$	ice bath	-1	2	0	$+2$	solution	Yes
	$+2$	blanket	$+1$	1	$+3$	-1	destructive	No
slightly hot	$+1$	ice bath	-1	2	-1	0	redundant	No

A.4 THE BUILD

Ten lexical, medical and visually grounded resources supply candidate entities: ConceptNet (Speer et al., 2017), Wikipedia (Wikipedia contributors, 2026) and Wikidata (Vrandečić & Krötzsch, 2014) for breadth; WordNet (Miller, 1995), LVIS (Gupta et al., 2019), EPIC-Kitchens (Damen et al., 2022) and AI2-THOR (Kolve et al., 2017) for objects a person handles; MeSH (Lipscomb, 2000), ChEBI (Hastings et al., 2016) and FrameNet (Baker et al., 1998) for specialist vocabulary. Figure 8 gives the counts at each step. Six filter levels remove duplicate names, abstract suffixes, lemma duplicates, abstract hypernym paths and synonyms, and reclassify domains, leaving 149,840 entities. No language model proposes an entity.

Tool-property annotation runs on the 72,446 entities in domains that can hold a physical tool. Qwen3.5-27B and gemma-4-31b each annotate every property, and each annotator's reading is the majority over three differently worded prompts at temperature zero. Where the two disagree, both arbitrate against the written specification, and a row on which the arbiters also disagree is dropped. Most entities are not tools: the annotators reject 70,017 of 73,171 annotation rows. Screens under the same protocol check that each remaining tool acts on its axis, is deployable in the situations the template renders, and has a magnitude band. Tools mined a second time for under-supplied groups pass the same screens (orange in Figure 8). A last screen removes 96 rows whose tool is not the standard instrument of any legitimate trade (a weapon, for example), leaving 1,638 tool records.

Pairing yields 1,422 tool pairs, and the grid of states over them yields 6,176 questions. Removing texts that collide with each other or with an evaluation set used in the paper leaves 5,751. The final selection drops every situation whose questions all have the same answer: a situation recurs across tools, so its answer could otherwise be memorized. It then balances Yes and No within each remaining situation and caps any one axis, leaving the 1,880 questions of the main set (Table 7).

A.5 DATA QUALITY

The answer cannot be read from the wording. A lexical probe (TF-IDF features, logistic regression) is trained on the question text of some situations and read on the others. Every question of a tool that appears in the held-out situations is removed from training. Because of the state flips, such a tool would otherwise sit in training in the opposite state with the opposite answer, and the probe would learn the tool's name with the wrong sign. The probe reaches an AUC of 0.498 ± 0.005 on the main set, inside the range of 0.485 to 0.524 it reaches with permuted labels, and a lexical proxy for the answer correlates with the label at $|r| = 0.022$.

Table 7: Number of questions in the main set by question type, physical axis and evidence score.

Question type		Physical axis		Evidence score $\lvert\Delta U\rvert$	
solution	940	Mechanical cohesion	564	0	231
destructive	705	Thermal transfer	476	1	798
redundant	235	Electromagnetism	234	2	719
		Spatial containment	226	3	132
		Moisture level	212		
		Radiant emission	168		

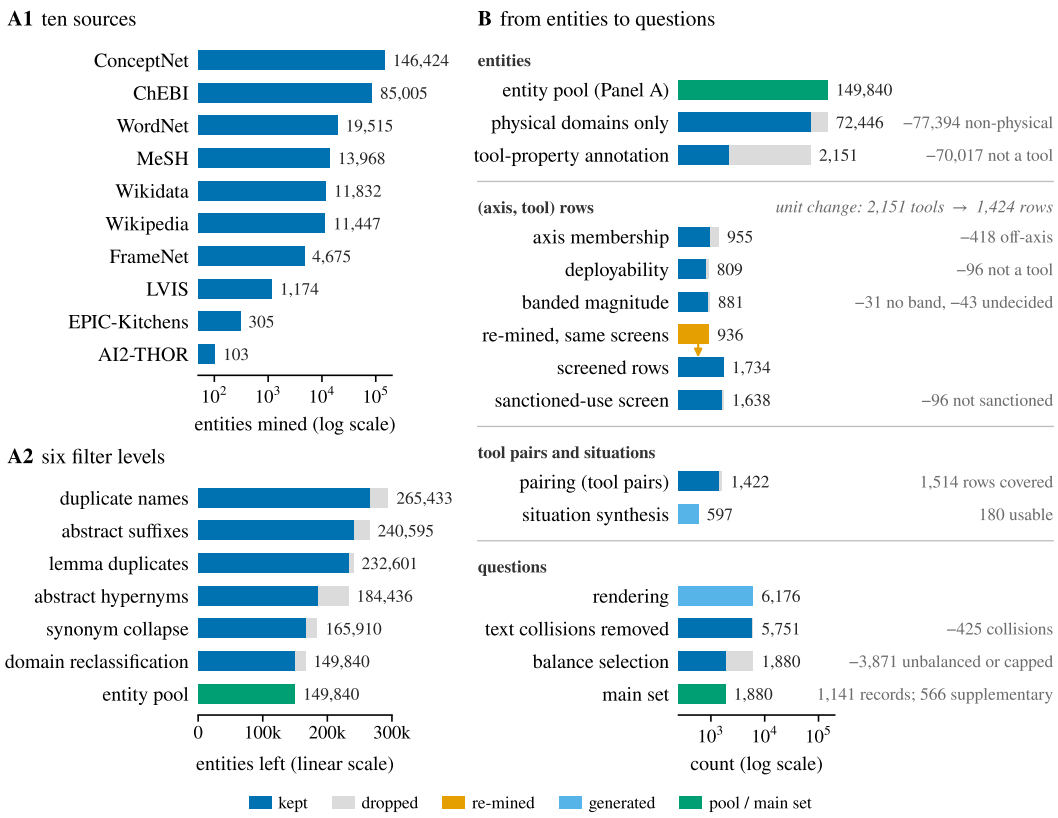


Figure 8: Counts at each stage of the FLIPACT construction. A1: the ten entity sources. A2: the six filter levels that produce the entity pool. B: the pool through tool-property annotation and the screens to tool records, then tool pairs, situations and questions. A1 and B use a log scale, A2 a linear one.

The set is balanced and agrees with the rule. Yes and No are exactly balanced in each of the 180 situations, with no imbalance per axis or per state level, and 3.9% of the main set's tool records have Yes and No counts that differ by more than one. Recomputing every answer from the annotated properties, independently of the build, gives zero mismatches, and no question text is duplicated or shared with an evaluation set used in the paper.

The tool annotations, the one step where a model's judgment enters, are reproduced by a model family that took no part in the build. GLM-4-32B re-annotated a stratified sample of 150 rows with the same prompts and matched the recorded value on 100% of is-tool judgments and 92.0% of axis judgments. On magnitude it placed all 150 tools within one band of the record, 67.3% in the same band. Each of its other 49 judgments moves a tool from the lowest or the highest band to the middle one, so it never reverses the order of two tools by magnitude. Perturbing the annotations at this rate

keeps both rankings of the models above 0.98 and moves every criterion toward acting (App. A.6), so annotation error does not produce the conservative criteria.

A.6 Robustness to annotation noise

We test whether annotation error at the rates of App. A.5 could reorder the models or produce the conservative criteria. The perturbation acts on the magnitude band, the annotated property that sets the size of the evidence score. We hold every model's answers fixed, perturb the magnitude of a random share of tool records, recompute the evidence score and answer of every affected question, and refit (a, τ), 300 times per condition. N1 moves 32.7% of records by one band at random, the rate at which GLM-4-32B disagrees on magnitude. N2 follows its disagreement pattern, one band toward the middle. N3 scales N2 until 26% of computed answers change, a stress level above both. A synthetic observer with each model's fitted a and $\tau = 0$ or $+2.5$ goes through the same perturbation. Rankings are compared by Spearman correlation with the unperturbed fit, and a model's criterion shift is the median over replicates minus its unperturbed value.

Table 8: Coordinates under three levels of annotation noise, 300 replicates per condition. The last block is the criterion recovered for a synthetic observer.

	N1 uniform	N2 empirical	N3 stress
Answers flipped	10.7%	15.7%	26.0%
Sensitivity and ranking			
$\rho(a)$, median	0.992	0.991	0.984
$\rho(\tau)$, median	0.996	0.995	0.990
a ratio, perturbed / unperturbed, median	0.980	0.965	0.976
Criterion			
Shift, mean over models	−0.19	−0.21	−0.39
Models shifting by more than ±1	0	0	0
Synthetic observer, recovered τ			
$\tau_{\text{true}} = 0$, median	−0.21	−0.27	−0.44
$\tau_{\text{true}} = 0$, 95th percentile	0.00	−0.06	−0.22
$\tau_{\text{true}} = +2.5$, median	+2.38	+2.41	+2.28

$\rho(\tau)$ is over the 25 entries whose criterion lies inside the score range, and $\rho(a)$ over 30 entries whose slope fit is stable.

Noise at the observed rate lowers sensitivity by two to four percent, keeps both rankings above 0.98, and moves every criterion toward acting (Table 8). The dominant error relabels a magnitude-1 tool as magnitude 2, which moves a Yes the model gave at $\Delta U = +1$ to $\Delta U = 0$ and pulls the fitted crossing point down. An observer with $\tau = 0$ is never recovered at $+1$ or above in 7,500 replicates, and one at $+2.5$ is recovered within a quarter of a unit. Annotation error therefore does not produce the conservative criteria, and if anything it understates them. The one model within 0.15 of zero, Qwen3-32B, crosses to a small negative value, and every other positive criterion stays positive in every replicate. Refitting on the 994 tool records whose magnitude band the two annotators agreed on without arbitration gives $\rho(a) = 0.994$ and $\rho(\tau) = 0.995$ and moves no criterion by more than twice its bootstrap error.

B Per-model coordinates and measurement checks

Table 9 gives the coordinates of all 44 models and Table 10 the checkpoint behind each. The evaluation protocol comes first (App. B.1), and the other subsections follow the order of §4: the fit and its readouts (App. B.2 and B.3), the convergence of sensitivity with external benchmarks (App. B.4 and B.5), and the criterion across generations and reasoning modes within a model line (App. B.6 and B.7).

B.1 Evaluation protocol

FlipAct. Each question is one user turn, rendered with the model's own chat template and its generation prompt, with no system message added. After the template's closing "Answer Yes or

Table 9: Coordinates of all 44 models on FLIPACT and their When2Call d' and c, grouped by family with the reasoning-on entries last, and two constant policies for reference. Checkpoints are in Table 10.

	Coordinates		FLIPACT scores					When2Call		
Model	a	τ	ACC	SQPR	F_1	TNR	MCC	d'	c	Panel
Constant policies										
Always-No	n/a	n/a	0.50	0.18	n/a	1.00	n/a	n/a	n/a	
Always-Yes	n/a	n/a	0.50	0.35	0.67	0.00	n/a	n/a	n/a	
Qwen3.5										
Qwen3.5-0.8B	$-0.52^{\dagger}$	n/a	0.50	0.19	0.07	0.95	−0.02	0.27	+0.22	
Qwen3.5-2B	$0.40^{\dagger}$	n/a	0.51	0.21	0.16	0.92	0.02	0.41	+0.39	
Qwen3.5-4B	0.55	+2.93	0.61	0.37	0.45	0.90	0.27	1.31	−0.71	•
Qwen3.5-9B	0.66	+2.88	0.62	0.39	0.44	0.93	0.30	1.67	−0.52	•
Qwen3.5-27B	0.57	+2.68	0.62	0.40	0.48	0.90	0.30	2.00	+0.06	•
Qwen3.5-35B-A3B	0.58	>+3	0.60	0.35	0.39	0.93	0.26	1.43	−0.68	
Qwen3										
Qwen3-0.6B	n/a	n/a	0.50	0.35	0.67	0.00	n/a	n/a	n/a	
Qwen3-1.7B	$0.46^{\dagger}$	n/a	0.50	0.35	0.66	0.06	0.02	n/a	n/a	
Qwen3-4B	0.34	>+3	0.54	0.26	0.27	0.91	0.12	1.44	−0.22	
Qwen3-8B	0.44	+2.78	0.61	0.39	0.48	0.85	0.24	1.46	−0.26	•
Qwen3-14B	0.57	+2.36	0.62	0.40	0.50	0.87	0.28	1.46	−0.40	•
Qwen3-32B	0.58	+0.14	0.69	0.54	0.69	0.69	0.38	1.63	+0.01	•
Gemma										
gemma-2-2b	$0.05^{\dagger}$	n/a	0.51	0.27	0.46	0.59	0.01	0.60	+0.27	
gemma-2-9b	0.34	+2.91	0.59	0.38	0.49	0.80	0.21	0.99	+0.76	•
gemma-2-27b	0.44	+1.68	0.63	0.42	0.56	0.77	0.26	1.21	−0.00	•
gemma-3-1b	$0.11^{\dagger}$	n/a	0.51	0.31	0.58	0.34	0.03	n/a	n/a	
gemma-3-4b	$0.12^{\dagger}$	n/a	0.54	0.31	0.47	0.69	0.09	n/a	n/a	
gemma-3-12b	0.39	+2.17	0.62	0.41	0.54	0.79	0.25	1.32	−0.37	•
gemma-3-27b	0.45	+1.77	0.63	0.43	0.56	0.79	0.28	1.46	−0.38	•
gemma-4-31b	0.50	+1.57	0.65	0.45	0.58	0.80	0.31	2.18	−0.41	•
InternLM										
InternLM2.5-1.8B	$1.27^{\dagger}$	n/a	0.50	0.35	0.67	0.00	0.01	n/a	n/a	
InternLM2.5-7B	0.25	−1.03	0.59	0.40	0.61	0.52	0.17	0.85	−0.17	
InternLM2.5-20B	0.41	<−3	0.60	0.44	0.69	0.33	0.25	1.00	+0.00	
InternLM3-8B	0.49	>+3	0.57	0.31	0.34	0.93	0.21	n/a	n/a	
Yi										
Yi-1.5-6B	$0.15^{\dagger}$	n/a	0.53	0.30	0.41	0.74	0.08	n/a	n/a	
Yi-1.5-9B	$0.14^{\dagger}$	n/a	0.55	0.38	0.62	0.37	0.12	0.63	−0.05	
Yi-1.5-34B	0.43	+0.41	0.64	0.47	0.63	0.67	0.28	1.01	+0.13	•
Mistral										
Mistral-7B-v0.3	0.26	<−3	0.54	0.40	0.66	0.21	0.11	0.31	−0.16	
Mistral-Small-24B (2501)	0.36	+2.18	0.61	0.41	0.54	0.78	0.24	1.18	−0.04	•
Mistral-Small-3.2-24B (2506)	0.43	−1.10	0.64	0.50	0.68	0.52	0.29	1.10	−0.33	•
Llama										
Llama-3.1-8B	$0.10^{\dagger}$	n/a	0.55	0.34	0.57	0.49	0.09	0.29	−1.07	
Llama-3.3-70B	0.65	+1.29	0.67	0.49	0.62	0.82	0.36	1.54	−0.39	•
Phi										
Phi-3.5-mini	0.30	+1.85	0.59	0.38	0.54	0.72	0.20	0.88	−0.64	•
phi-4	0.51	+1.67	0.64	0.44	0.57	0.80	0.29	1.80	−0.28	•
Other families										
GLM-4-32B-0414	0.67	+1.89	0.66	0.45	0.56	0.87	0.35	1.70	+0.14	•
Command-R-08-2024	0.45	>+3	0.57	0.31	0.34	0.92	0.20	0.92	−0.56	
Granite-3.3-8B	0.30	+2.59	0.59	0.37	0.51	0.76	0.19	0.77	−0.16	
OLMo-2-1124-13B	0.26	−2.31	0.59	0.41	0.64	0.44	0.18	0.64	+1.18	
Nemotron-Nano-9B-v2	$1.07^{\dagger}$	n/a	0.50	0.18	0.01	1.00	0.02	n/a	n/a	
Reasoning on										
Qwen3-30B-A3B-Thinking-2507	0.55	>+3	0.62	0.39	0.45	0.92	0.30	n/a	n/a	
DeepSeek-R1-Distill-Llama-8B	0.30	+1.47	0.61	0.44	0.57	0.70	0.22	n/a	n/a	
DeepSeek-R1-Distill-Qwen-32B	0.48	+1.48	0.65	0.47	0.59	0.79	0.31	n/a	n/a	
DeepSeek-R1-Distill-Llama-70B	0.48	+0.59	0.66	0.50	0.64	0.71	0.32	n/a	n/a	
gpt-oss-20b	0.53	+1.73	0.65	0.46	0.58	0.82	0.32	n/a	n/a	

a is sensitivity and τ the criterion, positive when conservative. A τ beyond the score range is shown as a bound, and † marks a fit whose criterion is not estimable. ACC, F_1 (Yes class), true-negative rate (TNR) and Matthews correlation (MCC) are over the 1,880 main-set questions, and SQPR is the share of its 1,141 tool records with every question answered correctly. • marks the panel of §4.2, and n/a a value that is undefined or not available.

Table 10: The Hugging Face checkpoint behind each entry and the precision of its weights, grouped by family with the family's reference. The reasoning-mode comparison of App. B.7 also runs the listed checkpoints with reasoning on.

Entry	Checkpoint	Weights	Note
Qwen3.5 (Qwen Team, 2026)			
Qwen3.5-0.8B	`Qwen/Qwen3.5-0.8B`	bf16	
Qwen3.5-2B	`Qwen/Qwen3.5-2B`	bf16	
Qwen3.5-4B	`Qwen/Qwen3.5-4B`	bf16	
Qwen3.5-9B	`Qwen/Qwen3.5-9B`	bf16	
Qwen3.5-27B	`Qwen/Qwen3.5-27B`	bf16	
Qwen3.5-35B-A3B	`Qwen/Qwen3.5-35B-A3B`	bf16	
Qwen3 (Yang et al., 2025)			
Qwen3-0.6B	`Qwen/Qwen3-0.6B`	bf16	
Qwen3-1.7B	`Qwen/Qwen3-1.7B`	bf16	
Qwen3-4B	`Qwen/Qwen3-4B`	bf16	
Qwen3-8B	`Qwen/Qwen3-8B`	bf16	
Qwen3-14B	`Qwen/Qwen3-14B`	bf16	
Qwen3-32B	`Qwen/Qwen3-32B`	bf16	
Qwen3-30B-A3B-Thinking-2507	`Qwen/Qwen3-30B-A3B-Thinking-2507`	bf16	reasoning on
Gemma (Team et al., 2024; 2025; 2026)			
gemma-2-2b	`google/gemma-2-2b-it`	bf16	
gemma-2-9b	`google/gemma-2-9b-it`	bf16	
gemma-2-27b	`google/gemma-2-27b-it`	bf16	
gemma-3-1b	`google/gemma-3-1b-it`	bf16	
gemma-3-4b	`google/gemma-3-4b-it`	bf16	
gemma-3-12b	`google/gemma-3-12b-it`	bf16	
gemma-3-27b	`google/gemma-3-27b-it`	bf16	
gemma-4-31b	`google/gemma-4-31b-it`	bf16	
InternLM (Cai et al., 2024)			
InternLM2.5-1.8B	`internlm/internlm2_5-1_8b-chat`	bf16	
InternLM2.5-7B	`internlm/internlm2_5-7b-chat`	bf16	
InternLM2.5-20B	`internlm/internlm2_5-20b-chat`	bf16	
InternLM3-8B	`internlm/internlm3-8b-instruct`	bf16	
Yi (Young et al., 2024)			
Yi-1.5-6B	`01-ai/Yi-1.5-6B-Chat`	bf16	
Yi-1.5-9B	`01-ai/Yi-1.5-9B-Chat`	bf16	
Yi-1.5-34B	`01-ai/Yi-1.5-34B-Chat`	bf16	
Mistral (Jiang et al., 2023; Mistral AI, 2025)			
Mistral-7B-v0.3	`mistralai/Mistral-7B-Instruct-v0.3`	bf16	
Mistral-Small-24B (2501)	`mistralai/Mistral-Small-24B-Instruct-2501`	bf16	
Mistral-Small-3.2-24B (2506)	`mistralai/Mistral-Small-3.2-24B-Instruct-2506`	bf16	
Llama (Grattafiori et al., 2024)			
Llama-3.1-8B	`meta-llama/Llama-3.1-8B-Instruct`	bf16	
Llama-3.3-70B	`kosbu/Llama-3.3-70B-Instruct-AWQ`	4-bit AWQ	
Phi (Abdin et al., 2024a;b)			
Phi-3.5-mini	`microsoft/Phi-3.5-mini-instruct`	bf16	
phi-4	`microsoft/phi-4`	bf16	
GLM (Glm et al., 2024)			
GLM-4-32B-0414	`zai-org/GLM-4-32B-0414`	bf16	
Command R (Cohere Team, 2024)			
Command-R-08-2024	`AMead10/c4ai-command-r-08-2024-awq`	4-bit AWQ	
Granite (Granite Team, 2024)			
Granite-3.3-8B	`ibm-granite/granite-3.3-8b-instruct`	bf16	
OLMo (OLMo et al., 2024)			
OLMo-2-1124-13B	`allenai/OLMo-2-1124-13B-Instruct`	bf16	
Nemotron (Basant et al., 2025; NVIDIA, 2026)			
Nemotron-Nano-9B-v2	`nvidia/NVIDIA-Nemotron-Nano-9B-v2`	bf16	
Nemotron-3-Nano-4B	`nvidia/NVIDIA-Nemotron-3-Nano-4B-BF16`	bf16	App. B.7 only
DeepSeek-R1 distillations (Guo et al., 2025)			
DeepSeek-R1-Distill-Llama-8B	`deepseek-ai/DeepSeek-R1-Distill-Llama-8B`	bf16	reasoning on
DeepSeek-R1-Distill-Qwen-32B	`deepseek-ai/DeepSeek-R1-Distill-Qwen-32B`	bf16	reasoning on
DeepSeek-R1-Distill-Llama-70B	`casperhansen/deepseek-r1-distill-llama-70b-awq`	4-bit AWQ	reasoning on
gpt-oss (Agarwal et al., 2025)			
gpt-oss-20b	`openai/gpt-oss-20b`	MXFP4 (native)	reasoning on

No." the prompt adds a new line, "Answer with exactly one word: Yes or No." With reasoning off, every question of the main and supplementary sets is answered once with vLLM, sampling at temperature 0.7, top-p 0.8 and top-k 20 with seed 0 and at most 8 new tokens, the same settings for every model except gemma-4-31b, which is run with transformers and greedy decoding. The answer is the first standalone *yes* or *no* in the output, ignoring case. An output with neither is unparsed and left out of the fit. No entry has more than one unparsed main-set question. Two further conditions guard the panel against unreadable output: at most 30% unparsed FLIPACT answers and a When2Call d' of at least 0.5. In practice neither removes a model that the conditions of §4.2 admit. The fit of Eq. 2 is a maximum-likelihood logistic regression on these binary answers, so every Yes-rate and $P(\text{Yes})$ on FLIPACT is a proportion of sampled answers, not a token probability. $\text{SE}(\tau)$ is half the width of the profile-likelihood interval within which the deviance rises by at most one. Nine entries were also run with seed 1. The two runs give the same answer on 93 to 100% of main-set questions, and for the six of the nine with a well-determined criterion both coordinates of the second run fall inside the 95% bootstrap intervals of the first. The steering and training readouts of §5 use the same prompt with reasoning off but do not sample: an answer is Yes when the logit of the first token of "Yes" exceeds that of "No" at the answer position.

Reasoning on. The five reasoning-on entries of Table 9 and the reasoning-on condition of App. B.7 receive the same prompt, one-word instruction included. Hybrid models reason through their template's switch (`enable_thinking`, or `thinking` for Granite-3.3-8B), and gpt-oss-20b runs at its template's default reasoning effort, medium. Decoding samples at temperature 0.6, top-p 0.95 and top-k 20 with seed 0, within a budget of 8,192 new tokens (4,096 for Granite-3.3-8B and Qwen3.5-35B-A3B). The answer is the first standalone *yes* or *no* after the end of the reasoning trace, which is the last `</think>` or, for gpt-oss-20b, the start of its final channel. An answer is unparsed when its trace reaches the budget without closing or is followed by no Yes or No, and no answer is read from inside a trace. A model is not fit in a mode where more than 10% of its answers are unparsed, which removes Qwen3.5-0.8B and Qwen3.5-2B with reasoning on (78% and 23%). Every other condition stays below 6%, and the five reasoning-on entries of Table 9 have no unparsed answer. External benchmarks are run with reasoning off only, so reasoning-on entries have no external readout and are not in the panel.

External benchmarks. Table 11 describes the seven benchmarks. Each example is one user turn through the chat template with reasoning off, decoded greedily with vLLM (with transformers for gemma-4-31b), with at most 8 new tokens on QuestBench, 64 on When2Call, 128 on BFCL and 256 on the other four. Yes/No replies are parsed as on FLIPACT. A When2Call reply is read as the letter it starts with, else a letter after a cue such as "answer" or "option" in its first 80 characters, else any standalone letter A–D there. On both BFCL sets a reply acts when it contains `CALL: f(...)` or calls a target function by name without that prefix. A reply with no parsable answer counts as withholding and stays in the denominators. No reply in the transfer test is unparsed. In the panel, at most 0.3% of replies are unparsed on the Yes/No benchmarks, and on When2Call the highest rate is 12.5%, for Mistral-Small-3.2-24B. Table 9 shows an entry's When2Call values as n/a when more than 30% of its replies are unparsed, which applies only to Nemotron-Nano-9B-v2. Runs with a steering hook (App. D.1 and D.2) use transformers with greedy decoding and at most 8 new tokens (64 on BFCL-miss-func), on a balanced sample of 600 examples of When2Call, When2Tool and CaT-Bench and on all examples of ACPBench and BFCL-miss-func.

The BFCL set of the panel draws 80 single-turn queries at random from BFCL's live-simple category, each answered by its one listed function, and 36 from its live-irrelevance category, where none of the one to ten listed functions fits the query. Each function is restated as its name, description and parameter names, under an instruction to reply with one `CALL:` line if a tool is needed and to answer directly otherwise.

On every benchmark, and in the signal-detection readout of App. B.3, d' and c follow Eq. 1 with the log-linear correction $H = (\text{hits} + \frac{1}{2})/(n_{\text{act}} + 1)$ and $\text{FA} = (\text{false alarms} + \frac{1}{2})/(n_{\text{withhold}} + 1)$, applied to every rate so that z stays finite for a model that always or never acts.

For the criterion-free AUC (App. B.4 and D.4), the benchmarks it covers are run again with the same greedy decoding, recording the 20 most likely tokens and their log-probabilities at the first four generated positions. At the first position whose generated token is an answer, the margin is the log-probability of acting (Yes, or the tool-call letter) minus that of the strongest alternative (No, or

Table 11: The external benchmarks, grouped by the comparison that uses them: source split, number of examples in each class, the act and withhold examples, and the reply format.

Benchmark	Act / withhold	Act example (withhold example)	Reply format
Panel comparison (§4.2)			
BFCL v3 live simple, live irrelevance	80 / 36	a query its one listed function answers (a query no listed function fits)	free-form; `CALL: f(...)` acts
Panel comparison and transfer test (§4.2, §5.3)			
QuestBench Logic-Q (500 of 1,150 tasks), GSM-Q CSP (151), GSM-Q verbal (288)	939 / 939	a fully specified task (the same task with one variable removed); acting is Yes to "Do you already have enough information to give a single correct final answer . . . ?"	Yes/No
When2Call test, multiple choice	1,295 / 2,357	a request whose correct reply is a tool call (one whose correct reply asks for missing information, 1,062, or says no tool can answer, 1,295); acting is choosing the tool-call reply	one letter for four replies: call, ask, cannot answer, answer directly
When2Tool single-hop test, 15 environments	750 / 750	a hard-tier task, built to need a tool (an easy-tier task, solvable without one; the 750 medium-tier tasks are not scored); acting is Yes to "would you need to use one of the tools?"	Yes/No
Transfer test (§5.3)			
BFCL-miss-func v3 multi-turn miss-func, 200 scenarios	200 / 200	a request with the one to three functions it needs listed (the same request with those functions removed)	free-form; `CALL: f(...)` acts
ACPBench applicability, test and validation	88 / 82	an action applicable in the stated planning state (an inapplicable action); acting is Yes	Yes/No
CaT-Bench test, 1,420 step pairs in both directions	1,420 / 1,420	two recipe steps of which one must precede the other (two independent steps); acting is Yes	Yes/No

the most likely other letter), each answer taking its most likely spelling among the recorded tokens. A model–benchmark pair is scored only if the sign of its margin reproduces at least 99% of its decoded answers. Examples with no answer token among the recorded positions, and When2Call replies with no parsable letter, are left out; Mistral-Small-3.2-24B loses the most (959 of 1,878 QuestBench and 463 of 3,652 When2Call examples). gemma-4-31b, run with transformers, has no recorded log-probabilities and is absent from these readouts.

B.2 Reliability and response profile

Both coordinates are reliable. Refitting on 200 random halves of the tool records with the Spearman–Brown correction gives a split-half reliability of 0.959 for a and 0.991 for τ. Tool records share question text across situations, so the stricter unit is the situation cluster, all questions that share one situation text. It is the resampling unit for every interval in the paper. Split by cluster, the two reliabilities are 0.906 and 0.973.

Models respond to the sign of the evidence score and ignore its size (Table 12 and Figure 4). The jump in Yes-rate between $\Delta U = -1$ and 0 is significant in 19 of the 21 models with a well-determined criterion, and the jump between 0 and $+1$ in none. Within either side, no model answers Yes reliably more often at the larger magnitude. The step therefore sits one unit below the rule's boundary. Every question at $\Delta U = 0$ is a slight overshoot, and models accept these overshoots about as often as helpful actions: the overshoot Yes-rate lies nearer the hit rate than the false-alarm rate in most models, and its interval includes the hit rate in 20 of 21.

A step fits held-out answers as well as the logistic. In five-fold cross-validation with folds by situation, the logistic in ΔU improves on a step with one Yes-rate per side by a median 0.010 nats per

Table 12: Median Yes-rate at each evidence score over the 21 models with a well-determined criterion, the median line of Figure 4.

ΔU	-3	-2	-1	0	$+1$	$+2$	$+3$
Median Yes-rate	0.14	0.17	0.13	0.40	0.43	0.52	0.46

question $[-0.003, +0.024]$, reliably in 3 of 21 models. A step between -1 and 0 fits best in 20 of 21, although its location was read off these same curves. In 9 of the 21 models the observed Yes-rate, smoothed to be monotone, never reaches one half, so their τ is where the fitted curve crosses it.

Refitting with ΔU collapsed to $\{-1, 0, +1\}$ reproduces the sensitivity ranking ($\rho = 0.985$). Three monotone re-expressions of the score (centered ranks, destructive scores halved, a per-axis z-score) keep $\rho(a) \geq 0.98$ over all fitted models and $\rho(\tau) \geq 0.995$, and leave $r(a, d')$ on When2Call between $+0.66$ and $+0.67$.

The two supplementary sets show that the overshoot is the one redundancy models accept (Table 13). Where the state is already at its goal, every model answers Yes less often than at the main set's slight overshoots, by a median of 0.27. Where a tool's supply is missing, every model answers Yes less often than on the same question with the supply present, by a median of 0.43. Models thus withhold when an action is plainly unneeded or cannot work, and accept it when it points the right way but goes one step too far.

Table 13: Yes-rate on the main set by evidence sign and on the two supplementary sets, median and range over the 21 models with a well-determined criterion. Each supplementary question is paired with a matched question whose answer is Yes.

Questions	Answer	Median Yes-rate	Range
Main set			
destructive, $\Delta U < 0$	No	0.15	[0.03, 0.51]
slight overshoot, $\Delta U = 0$	No	0.40	[0.18, 0.70]
helpful, $\Delta U > 0$	Yes	0.47	[0.30, 0.76]
Already satisfied			
state at its goal	No	0.08	[0.01, 0.68]
matched question	Yes	0.23	[0.09, 0.58]
Prerequisite			
supply missing	No	0.29	[0.00, 0.51]
matched question, supply present	Yes	0.66	[0.46, 0.90]

B.3 A SIGNAL-DETECTION READOUT

Because models read the sign of the score, Table 14 sets the logistic coordinates beside a signal-detection readout: hits on $\Delta U > 0$ and false alarms on $\Delta U < 0$ enter Eq. 1, and the overshoot questions at $\Delta U = 0$ are left out. Sensitivity ranks the models the same under every readout. So does the criterion when it is divided by sensitivity, as c/d': τ is close to a linear function of it ($r = 0.999$), $\tau \approx -0.09 + 3.22\, c/d'$, where -0.09 is the midpoint of the mean negative and the mean positive score of the main set and 3.22 the distance between them. The absolute c ranks models less closely with τ ($\rho = 0.77$) because τ, like c/d', is divided by sensitivity and c is not. The number of positive criteria, the opposite signs of the two Mistral-Small-24B releases and the convergence of sensitivity with external d' are the same under every readout.

B.4 CONVERGENCE WITH EXTERNAL BENCHMARKS

This subsection gathers the evidence behind §4.2. Table 15 gives each coordinate's correlation with its counterpart on the four external benchmarks, and Figure 9 plots the panel models on the three where sensitivity converges.

Table 14: The logistic coordinates beside a signal-detection readout that leaves out $\Delta U = 0$. Rank agreement with the logistic fit and the count of criteria with an interval above zero are over the 21 models with a well-determined criterion, and r is over the panel. Shaded: the fit of the main text.

	Rank agreement		Criterion	r(sensitivity, external d')		
Readout	ρ(sens.)	ρ(crit.)	interval > 0	BFCL	QuestBench	When2Call
Logistic, all questions (a, τ)	1.00	1.00	16	+0.70	+0.69	+0.67
Logistic, $\Delta U = 0$ left out	0.98	1.00	17	+0.69	+0.65	+0.65
Signal detection, d' and c	0.99	0.77	18	+0.70	+0.65	+0.65
Signal detection, d' and c/d'	0.99	0.99	18	+0.70	+0.65	+0.65

Table 15: Each FLIPACT coordinate against its counterpart on the four external benchmarks of §4.2, over the panel of 18 models.

	Sensitivity a vs. d'			Criterion τ vs. c	
Benchmark	r	p	95% interval	r	p
BFCL (Patil et al., 2025)	+0.70	0.0013	[+0.39, +0.90]	+0.51	0.030
QuestBench (Li et al., 2026)	+0.69	0.0015	[+0.46, +0.88]	−0.02	0.930
When2Call (Ross et al., 2025)	+0.67	0.0024	[+0.40, +0.87]	−0.01	0.954
When2Tool (Sun et al., 2026)	+0.02	0.9324	[−0.50, +0.57]	+0.19	0.439

What each benchmark asks and where its answer key comes from are in Table 39. Collapsing When2Call's choice among several replies to call or withhold adds an offset to c (App. E), so external c is not comparable across benchmarks. The criterion correlation on BFCL rests on the three models with the lowest τ: without them $r = +0.21$. Its Spearman correlation is $+0.46$ ($p = 0.058$), and Holm-adjusted over the four benchmarks its p is 0.12.

Panel cutoff. The panel of §4.2 admits models with $a \geq 0.30$. When that cutoff moves with the other panel conditions fixed (Table 16), the correlation of a with When2Call d' stays positive and significant at every cutoff. The criterion correlation is not stable: it is negative and significant at 0.25 and not significant at 0.30 or 0.35, so no direction is read from it.

Table 16: The When2Call correlations of both coordinates at three panel cutoffs, over nested panels. Shaded: the cutoff of the main text.

		a vs. d'		τ vs. c	
Cutoff $a \geq$	n	r	p	r	p
0.25	20	+0.77	0.0001	−0.46	0.041
0.30	18	+0.67	0.0024	−0.01	0.95
0.35	16	+0.53	0.033	−0.29	0.28

A criterion-free readout. Scoring each external example by its answer-token log-probability gives an AUC with no criterion term (App. D.4). Over the panel models with readable log-probabilities, a also correlates with AUC on QuestBench and When2Call, with $r = +0.70$ and $+0.81$, so sensitivity converges with both benchmarks when the external measure carries no criterion term.

B.5 SENSITIVITY AND GENERAL CAPABILITY

Separating warranted from unwarranted action on FLIPACT takes physical knowledge as well as reading the state, the tool and the question. Table 17 therefore checks the convergence of §4.2 against four other indicators on the panel. Two are general: log parameter count, and MMLU on a 570-question subset, available for 15 of the 18 models. The other two come from the answers a is fitted on, so partialing them out removes part of a itself: FLIPACT accuracy, which on the balanced

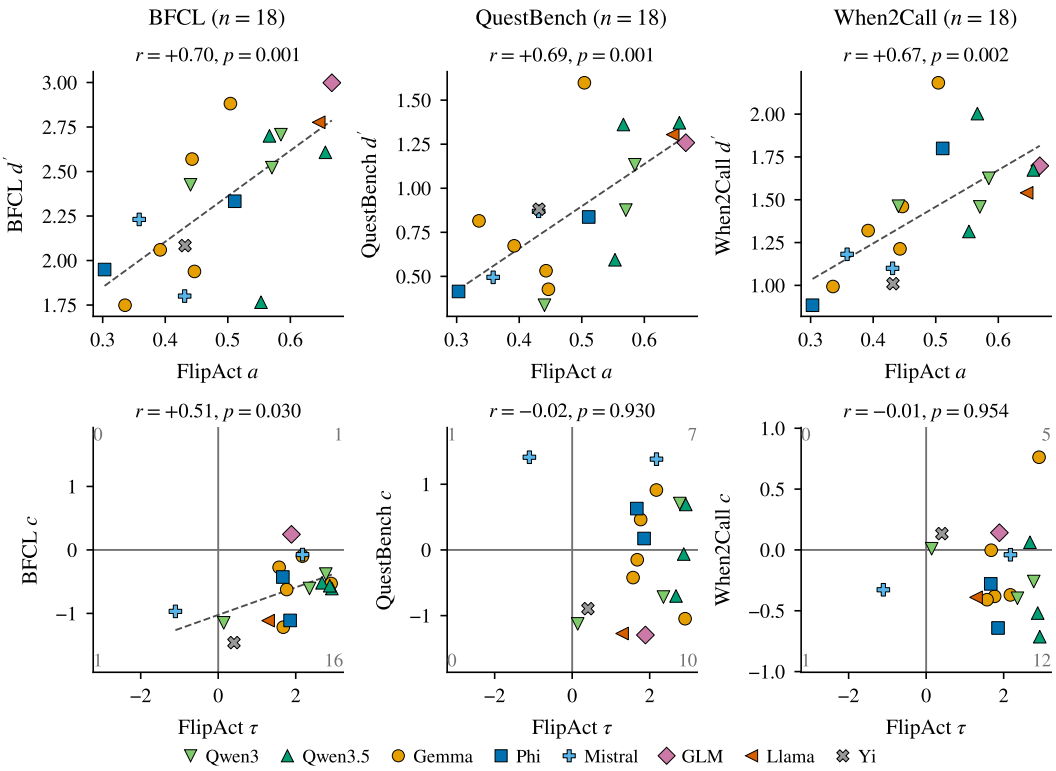


Figure 9: The two coordinates against their external counterparts for the panel models, on the three benchmarks where sensitivity converges. Top: sensitivity against d'. Bottom: criterion against c, with the number of models in each quadrant in gray. A dashed line marks a significant correlation.

main set equals balanced accuracy, and a FLIPACT d' computed with Eq. 1. The convergence holds with size partialed out: on the three benchmarks where a converges, the partial correlation of a with external d' given log parameters is 0.62 to 0.64. Given MMLU it is 0.52 to 0.55 (p between 0.04 and 0.06 at $n = 15$). The FLIPACT d' correlates with a at 0.96, so the convergence does not depend on how FLIPACT's sensitivity is summarized.

Table 17: Correlation of a with external d' on the panel of 18 models, raw and with each indicator partialed out. The shaded row repeats the raw correlation on the 15 models with MMLU.

Partialed out	BFCL	QuestBench	When2Call	When2Tool
none	+0.70	+0.69	+0.67	+0.02
none, 15 models with MMLU	+0.71	+0.75	+0.72	−0.27
log parameters	+0.64	+0.63	+0.62	−0.18
MMLU ($n = 15$)	+0.53	+0.55	+0.52	−0.67
FLIPACT accuracy	+0.54	+0.54	+0.56	−0.20
FLIPACT d'	+0.19	−0.11	+0.08	−0.37

B.6 GENERATIONS WITHIN THE QWEN AND GEMMA LINES

The generation, like size (§4.3), does not set the criterion. The Qwen and Gemma lines each span two or three generations at comparable sizes, released between mid-2024 and 2026 (Table 18). From one generation to the next, sensitivity rises at small and middle sizes, most from Qwen3 to Qwen3.5, and changes little at the largest sizes. The criterion stays put from Qwen3-8B to Qwen3.5-9B and

across the three Gemma generations at 27–31B, falls from gemma-2-9b to gemma-3-12b and rises from Qwen3-32B to Qwen3.5-27B.

Table 18: Coordinates of the Qwen and Gemma generations at comparable sizes, with the month the weights were released. A τ beyond the score range is shown as a bound.

Line	Size	Model	Released	a	τ
Qwen	4B	Qwen3-4B	2025-04	0.34	$>+3$
		Qwen3.5-4B	2026-03	0.55	$+2.93$
	8–9B	Qwen3-8B	2025-04	0.44	$+2.78$
		Qwen3.5-9B	2026-03	0.66	$+2.88$
	27–32B	Qwen3-32B	2025-04	0.58	$+0.14$
		Qwen3.5-27B	2026-02	0.57	$+2.68$
Gemma	9–12B	gemma-2-9b	2024-06	0.34	$+2.91$
		gemma-3-12b	2025-03	0.39	$+2.17$
	27–31B	gemma-2-27b	2024-06	0.44	$+1.68$
		gemma-3-27b	2025-03	0.45	$+1.77$
		gemma-4-31b	2026-04	0.50	$+1.57$

B.7 Reasoning mode

Reasoning mode raises sensitivity only in the smallest models. In larger ones it moves the criterion toward acting. We compare reasoning off and on for every model that has both modes, in one paired run with bootstrap intervals (Table 19), so the values with reasoning off can differ from Table 9, most for Qwen3-1.7B, whose fit there is flat. Reasoning raises a significantly on Qwen3-1.7B, Qwen3-4B, Nemotron-3-Nano-4B and Qwen3.5-4B, the models of 4B and below, and on no larger model. It moves τ significantly toward zero on Qwen3.5-4B, Qwen3-8B, Qwen3.5-9B and Qwen3.5-27B. The criterion falls less on Qwen3-14B, stays put on Granite-3.3-8B and rises slightly on Qwen3-32B. Qwen3.5-0.8B, Qwen3.5-2B, Qwen3.5-35B-A3B and Qwen3-0.6B have no usable fit in one of the two modes and are omitted.

Table 19: Reasoning mode off and on. Bold marks a significant rise in a or fall in τ.

	a			τ		
Model	off	on	Δ	off	on	Δ
4B and below						
Qwen3-1.7B	0.04	0.42	$+\mathbf{0.38}$	–	$+1.05$	–
Qwen3-4B	0.32	0.45	$+\mathbf{0.13}$	–	–	–
Nemotron-3-Nano-4B	0.28	0.49	$+\mathbf{0.21}$	–	$+2.42$	–
Qwen3.5-4B	0.56	0.66	$+\mathbf{0.10}$	$+2.87$	$+1.91$	$-\mathbf{0.96}$
8B and above						
Qwen3-8B	0.44	0.53	$+0.09$	$+2.78$	$+1.69$	$-\mathbf{1.08}$
Granite-3.3-8B	0.30	0.37	$+0.07$	$+2.59$	$+2.61$	$+0.02$
Qwen3.5-9B	0.66	0.61	-0.04	$+2.86$	$+1.99$	$-\mathbf{0.87}$
Qwen3-14B	0.57	0.51	-0.06	$+2.36$	$+2.07$	-0.28
Qwen3.5-27B	0.57	0.53	-0.04	$+2.67$	$+1.57$	$-\mathbf{1.10}$
Qwen3-32B	0.58	0.51	-0.07	$+0.14$	$+0.40$	$+0.26$

A dash marks a τ outside the score range, or a change that involves one.

Two answers from these runs show the two effects. Traces are quoted verbatim, with [...] marking omitted text. In Qwen3.5-4B the trace works out what the tool does to the state, and the model tells the two states of one tool apart. In Qwen3.5-27B, reasoning turns a No on a warranted question into a Yes, while the harmful state of the same tool stays No.

Reasoning separates the two states (Qwen3.5-4B)

Situation: The food or contents is [state]. The supplies this tool depends on (mains electricity) are available nearby. The goal is to bring the food or contents back to a normal state. Action: a person uses a hotplate. Is using the hotplate a wise choice in this situation?

State	Gold	ΔU	Reasoning off	Reasoning on
very cold	Yes	$+2$	Yes	Yes
very hot	No	-2	Yes	No

Very hot: *Goal: Cool down (bring to normal state). [...] Function of a hotplate: Heats objects placed on it. Conflict: Using a heating device on something that is already very hot will make it hotter, not cooler. This is the opposite of the goal.*
Very cold: *Goal: Reheat/warm up cold food. [...] Logic: If electricity is available and the goal is to heat food, a hotplate is an appropriate and effective tool for that purpose.*

Reasoning moves the criterion toward acting (Qwen3.5-27B)

Situation: The rigid object or joint is [state]. The goal is to bring the rigid object or joint back to a normal state. Action: a person uses a rivet gun. Is using the rivet gun a wise choice in this situation?

State	Gold	ΔU	Reasoning off	Reasoning on
very loose and coming apart	Yes	$+2$	No	Yes
very jammed and over-tightened	No	-2	No	No

Very loose: *If a joint is loose and coming apart, permanently fastening it with a rivet is a standard and effective repair method for many rigid objects (metal sheets, frames, etc.). [...] Sometimes bolts or welding might be used, but a rivet gun is definitely a* wise *and common choice for fixing loose joints where rivets are applicable.*

C THE CRITERION UNDER REVERSED QUESTION POLARITY

Design. If a model prefers one answer token whatever is asked, that preference enters its criterion (§4.3), and no single wording can separate the two. The standard probe is a balanced key (Cronbach, 1946): the same content is asked in both polarities, and one reads whether the answer follows the content or the answer word. The canonical wording closes each main-set question by asking whether using the tool is a wise choice ("wise" below). We replace that sentence with three reversed wordings, under which No means act ("Would using $\{t\}$ be an unwise choice", "Should the person avoid using $\{t\}$", "Is using $\{t\}$ a mistake"), and ask the canonical wording in the same run. Answers under a reversed wording are recoded, so every fit is $P(\text{act} \mid \Delta U)$ and a model that answers the content returns the same a and τ. Five same-polarity rephrasings (appropriate, should use, right tool, makes sense, would recommend) and three answer-order variants ("Answer No or Yes", and letter options with Yes first or with No first) check both coordinates under rewording. The run used one seed on 16 of the 18 panel models, and every rewording is compared with the canonical wording asked in the same run.

Decomposition. With both fits recoded to act polarity, write $\tau_{\text{wise}} = \tau_d + b$ and $\tau_{\text{unwise}} = \tau_d - b$. τ_d is a threshold on the evidence, attached to the proposition, and b is a response bias attached to the answer word, in score units and positive when the model prefers No. Table 20 gives both as means over the three reversed wordings.

Result. Seven models keep the sign of τ under all three reversed wordings (Table 20). For the three Gemma models among them, phi-4 and Mistral-24B, the reversed τ is more conservative than the canonical one: asked whether the action is unwise or a mistake, they answer Yes on 75–90% of questions, the same stance as their 19–36% Yes-rate on "wise". The three Qwen3.5 models and GLM-4-32B flip under all three. They answer No under either polarity, with a Yes-rate of 15–20% on "wise" and 9–26% on the reversed wordings (one model–wording pair at 43%), which recoded is an act-rate of 0.75–0.9. These four have b of $+3$ to $+4$ and τ_d within a unit of zero, so their conservatism on the canonical wording is a preference for No. The other models stay conservative

on τ_d: the seven that keep their sign have b between -0.97 and $+0.58$, and τ_d at or above their canonical τ on six of the seven.

Table 20: Sign of the criterion under the three reversed wordings, and its split into a threshold on the evidence τ_d and an answer preference b.

	Canonical wording		Sign kept under			Decomposition	
Model	τ	Yes-rate	unwise	avoid	mistake	τ_d	b
Sign kept under all three wordings							
Yi-1.5-34B	$+0.32$	0.48	✓	✓	✓	$+0.62$	-0.30
Qwen3-14B	$+2.62$	0.23	✓	✓	✓	$+2.04$	$+0.58$
gemma-2-9b	$+3.65^{\dagger}$	0.21	✓	✓	✓	$+4.55^{\dagger}$	-0.90
gemma-2-27b	$+2.50$	0.25	✓	✓	✓	$+2.80$	-0.30
gemma-3-27b	$+2.45$	0.27	✓	✓	✓	$+3.42^{\dagger}$	-0.97
phi-4	$+1.45$	0.36	✓	✓	✓	$+2.19$	-0.74
Mistral-24B	$+3.33^{\dagger}$	0.19	✓	✓	✓	$+3.62^{\dagger}$	-0.29
Mixed or degenerate							
gemma-3-12b	$+2.31$	0.30	–	–	✓	$+2.44$	-0.13
Llama-3.3-70B	$+1.25$	0.36	✓	✗	✓	$+1.51$	-0.26
Qwen3-32B	$+1.05$	0.39	✓	✗	✓	$+0.63$	$+0.41$
Qwen3-8B	$+2.34$	0.29	–	–	–	$+0.95$	$+1.39$
Phi-3.5-mini	$+3.66^{\dagger}$	0.24	–	–	–	$+4.83^{\dagger}$	-1.17
Sign flipped under all three wordings							
Qwen3.5-4B	$+3.57^{\dagger}$	0.15	✗	✗	✗	-0.42	$+3.99$
Qwen3.5-9B	$+3.37^{\dagger}$	0.15	✗	✗	✗	$+0.32$	$+3.04$
Qwen3.5-27B	$+3.07^{\dagger}$	0.18	✗	✗	✗	-0.17	$+3.24$
GLM-4-32B	$+2.56$	0.20	✗	✗	✗	-0.86	$+3.42$

Canonical-wording values come from the polarity run, not from Table 9. ✓ sign of τ kept, ✗ flipped, – reversed fit degenerate. † outside the ± 3 score range, or a mean including such a fit, read for sign only.

The reversed wordings are the noisier conditions, so τ remains the reported coordinate and τ_d a diagnostic that names the models whose conservatism on the canonical wording is an answer preference. On a model with no answer token, the JEV model of App. E, the decomposition gives $b = -0.20 \pm 0.24$, so it does not produce an answer preference where none can exist.

Rewording without reversing polarity. The conservative criterion is not a product of the word "wise". Under each of five same-polarity rephrasings the criterion stays positive on 15 or all 16 models, and it ranks models much as the canonical wording does (Table 21). Sensitivity also keeps its ranking, at $\rho(a) = 0.87$–0.94 for the five rephrasings, "Answer No or Yes" and letter options with Yes first, and at 0.58 for letter options with No first.

Table 21: Both coordinates under five same-polarity rephrasings of the canonical question, 16 panel models. Rank correlations are with the canonical wording asked in the same run, on which all 16 criteria are positive.

Rephrasing	$\tau > 0$	$\rho(\tau)$	Median $\Delta\tau$	$\rho(a)$
"appropriate"	16	0.86	-0.63	0.90
"should use"	15	0.89	-0.54	0.91
"the right tool"	16	0.77	-0.08	0.94
"makes sense"	16	0.96	-0.87	0.88
"would recommend"	16	0.88	-0.15	0.88

D MOVING THE CRITERION: SETTINGS AND FURTHER RESULTS

This section gives the settings behind §5 and further results: steering (App. D.1 and D.2), the two training schemes and the values behind §5.3 (App. D.3), a criterion-free readout (App. D.4), answer

preference (App. D.5), a per-benchmark choice of steering sign and strength (App. D.6) and single answers (App. D.7), in the order of §5.

D.1 STEERING

This subsection specifies the criterion direction of §5.1 and its strength rule, then checks the injection on a held-out half of FLIPACT, on held-out When2Call examples and on eleven more models.

The direction is a probe of the model's own answers. For each of the sixteen models of Tables 1 and 22, an L2-regularized logistic regression ($C = 0.1$) maps the standardized residual-stream activation at the final token, taken at the layer at 0.6 of the model's depth, to the model's own Yes/No answer on every main-set question. No labels enter the fit. Its weight vector, scaled to unit norm, is $\hat{w}$. The hook adds $\lambda\,\hat{w}\,\bar{r}/\sqrt{d}$ to the final-token activation at the output of that layer, where $\bar{r}$ is the mean norm of the activation over the fit questions and d is the width, so that λ is in units of the model's own scale.

Table 22: Steering on When2Call for all 16 models, grouped by the tests the shift passes. Negative shift is toward acting, and * marks a model of Table 1.

				Against controls	
Model	$\lambda^\star$	Fitted Δc	q	Ratio	Retained
Significant, three times the largest control, half retained, monotone in λ					
Command-R-08-2024	32	-1.04	< 0.001	4.8	1.01
Mistral-24B*	64	-1.14	< 0.001	4.9	1.09
Qwen3-8B*	16	-0.27	< 0.001	3.7	0.91
gemma-2-9b*	64	-0.65	0.004	3.4	0.69
Significant, at least one other test not passed					
gemma-3-12b	4	-0.81	0.001	7.8	0.77
Llama-3.3-70B	48	-0.16	0.027	18.5	0.63
Yi-1.5-34B	16	-0.30	< 0.001	1.6	2.83
phi-4*	32	-0.11	0.004	1.8	0.83
Qwen3.5-9B	64	-0.12	< 0.001	1.7	1.07
Qwen3.5-35B-A3B	96	-0.16	< 0.001	1.0	1.21
gemma-3-27b	-32	$+0.05$	0.037	0.1	1.25
Not significant					
Qwen3.5-27B*	96	-0.02	0.052	0.2	0.50
Qwen3-14B	16	-0.34	0.62	3.2	0.79
Qwen3-32B	96	-0.16	0.15	0.7	0.97
GLM-4-32B	-16	-0.06	1.0	4.1	1.08
Qwen3-30B-A3B-Thinking-2507	no strength inside the d'-stable range				

Fitted Δc: change in c across the d'-stable strengths, which differs from the change at $\lambda^\star$ in Table 1. On gemma-3-27b the paired test registers more acting at $\lambda^\star$ while the fitted shift across the range is near zero.

The direction depends on the wording it is read under. A probe refitted under the reversed wording "unwise" of App. C has cosine -0.03, $+0.06$ and $+0.02$ with the canonical one on Qwen3.5-9B, Qwen3.5-27B and Qwen3-8B, and -0.34 on gemma-3-12b. Steering results are therefore read from behavior, the shift in c and τ under injection, and not from the geometry of the direction.

The strength is chosen on When2Call itself. Each model is run over a grid of strengths λ, which includes negative values that inject $-\hat{w}$. The strengths at which When2Call d' stays within 25% of its unsteered value form the d'-stable range, and $\lambda^\star$ is the strength in that range with the lowest criterion c. Figure 10 shows the strength response of the five models of Table 1.

On FLIPACT the injection is checked with a held-out readout. The quads are split at random into halves, and this split is not the one used for training (App. D.3). The direction and a strength chosen by accuracy are fit on one half, and c and d' are computed from the binary answers on the other. On the read half the injection moves the criterion of Qwen3.5-9B and Qwen3.5-27B toward zero, leaves d' within a few hundredths of its value and raises accuracy by about two points (Table 23).

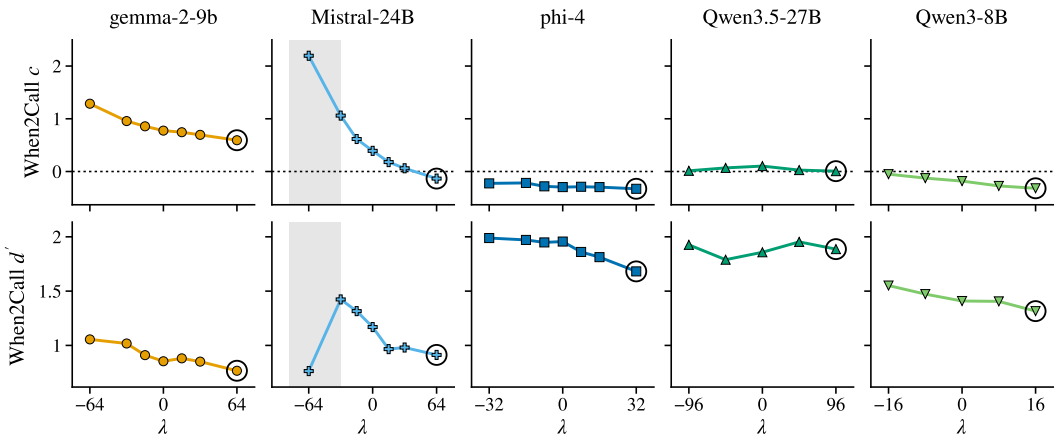


Figure 10: When2Call criterion c (top) and d' (bottom) against the injected strength λ for the five models of Table 1. Each column has its own range of λ. Gray marks strengths outside the d'-stable range, and circles mark $\lambda^\star$.

Table 23: Steering on the held-out half of FLIPACT, with the direction and strength fit on the other half.

	Criterion c			Accuracy	
Model	before	+ steering	$\Delta d'$	before	Δ
Qwen3.5-9B	+0.75	+0.42	−0.02	0.654	+0.026
Qwen3.5-27B	+0.57	+0.39	+0.03	0.642	+0.016

Table 1 chooses $\lambda^\star$ and reads its effect on the same When2Call examples. A split-half check separates the choice from the readout. The examples are split into halves twenty times at random, keeping the call/withhold balance, and on each split $\lambda^\star$ is chosen on one half by the same rule and the shift is read on the other (Table 24). For four of the five models the held-out shift in c matches the full-set one to within 0.02, and on every split the criterion falls and the call rate rises. phi-4, whose full-set shift is the smallest, starts permissive and shows no reliable shift on the held-out half.

Table 24: Split-half check of the When2Call steering shift: $\lambda^\star$ chosen on one half and read on the other, beside the full-set value.

		gemma-2-9b	Mistral-24B	phi-4	Qwen3.5-27B	Qwen3-8B
Δc	full	−.18	−.53	−.03	−.10	−.14
	held-out	−.18	−.53	−.01	−.08	−.14
	[2.5, 97.5]	[−.25, −.13]	[−.61, −.29]	[−.06, +.11]	[−.11, −.04]	[−.21, −.08]
Δ call rate	full	+.052	+.178	+.019	+.025	+.045
	held-out	+.048	+.175	+.012	+.022	+.047
	[2.5, 97.5]	[+.035, +.068]	[+.096, +.210]	[−.026, +.027]	[+.010, +.027]	[+.028, +.067]
Δ acc.	full	+.002	−.032	−.038	+.005	−.022
	held-out	+.000	−.035	−.028	+.007	−.020
	[2.5, 97.5]	[−.020, +.013]	[−.050, −.007]	[−.050, +.003]	[−.003, +.012]	[−.043, −.003]
Same $\lambda^\star$		20/20	14/20	14/20	10/20	20/20

Full (shaded): the change of Table 1, computed before rounding. Held-out: median over 20 stratified splits and its [2.5, 97.5] percentile range. Same $\lambda^\star$: splits whose chosen strength equals the full-set one.

Table 22 adds eleven further models, each run at its selected strength only. The fitted shift is the change in c across the d'-stable strengths, and q is a paired McNemar test on the withhold examples after Benjamini–Hochberg correction. Two controls go with each shift. The first runs three direc-

tions of the same norm, one orthogonal to $\hat{w}$ and two random, and the ratio column divides the shift by the largest of their shifts. The second projects the refusal (Arditi et al., 2024), sycophancy (Rimsky et al., 2024) and a generic engagement direction out of $\hat{w}$, and the retained column is the share of the shift that remains. The engagement direction is the difference in mean activation between prompts the model answers and prompts it refuses or hedges, taken from XSTest (Röttger et al., 2024) and a set of abstention questions. On the paired test, eleven of the sixteen models move significantly toward acting after correction for the fifteen testable comparisons. Four of these also clear both controls and respond monotonically in λ, with a shift at least three times the largest control shift and at least half of it kept after the projection.

D.2 Steering on further benchmarks

To test whether the steering result of §5.1 is specific to When2Call, the same direction $\hat{w}$ is injected on four further benchmarks for three of the models, with the injection layer, readout and strength rule of App. D.1. The shift in c is read inside the d'-stable range, and the paired McNemar test on the withhold examples is judged after Benjamini–Hochberg correction. The direction moves the criterion toward acting, with a significant paired test, on every benchmark that can be scored for Qwen3-8B and Mistral-24B, and on one benchmark of four for Qwen3.5-27B (Table 25). Two of Mistral-24B's four benchmarks have no strength at which d' stays stable, and Qwen3.5-27B's response on When2Tool and BFCL-miss-func is not monotone in λ.

Table 25: The FlipAct criterion direction on four further benchmarks with the weights fixed: shift in c inside the d'-stable range, negative toward acting.

Model	When2Tool	BFCL-miss-func	ACPBench	CaT-Bench
Qwen3-8B	-1.08 •	-0.54 •	-1.11 •	-0.80 •
Mistral-24B	no stable λ	-1.95 •	-1.94 •	no stable λ
Qwen3.5-27B	-1.09	-0.24	-0.47	-0.74 •

• significant paired McNemar test on the withhold examples after Benjamini–Hochberg correction. No stable λ: no strength inside the d'-stable range.

D.3 Training

Both training schemes, criterion distillation (§5.2) and sensitivity learning (§5.4), share the split, adapter and significance rule below. Their settings and controls come first, then the external-benchmark values behind §5.3.

Split and significance. The main set is split at the connected components of the graph that links situations to tool records, so no tool record, situation or state flip appears in both halves. Both schemes use this split, with the sizes in Table 26. Every reported change is the mean over three seeds. Its error is the largest of the seeds' paired situation-cluster bootstrap errors, and a change is significant when it is at least twice that error. Table 29 lists every change with its error.

Criterion distillation. The teacher is the student itself with a steering hook. The hook adds $\lambda^\star \hat{w}\, \bar{r}/\sqrt{d}$ at the 0.6-depth layer, as in App. D.1, but here $\hat{w}$ is a difference of means: the mean standardized activation of the training-half questions the student answers Yes, minus that of the questions it answers No, scaled to unit norm. No labels enter it. The loss is the forward KL of Eq. 3, with the teacher's logits computed once on all training-half questions, and Table 26 gives the control and the optimization settings. Every criterion-distillation run of the two non-Qwen students changes MMLU by -0.004 to $+0.006$.

The strength comes from a FlipAct strength sweep with the same estimator. A strength qualifies when the criterion is well determined, a stays within two bootstrap errors of its unsteered value, d' stays within a quarter of its value, and the Yes-rate stays inside $[0.10, 0.90]$. Among the qualifying strengths, $\lambda^\star$ is the one with the largest $|\Delta\tau|$ (Table 26). The sweep's direction is fit on a random half of the quads (955 questions, 479 of them in the held-out half) and agrees with the teacher's direction, with a cosine of 0.97 to 0.99. The sweep's a, τ, d' and Yes-rate are computed on all

Table 26: Training settings of the two schemes.

Both schemes		
Split	63 connected components: 940 training and 940 held-out questions (247 and 248 state flips)	
Adapter	LoRA with $r = 16$, $\alpha = 32$ and dropout 0.05 on the q, k, v, o, gate, up and down projections (on gemma-3-12b, of the language-model layers only)	
Precision	bf16 weights	
Clipping	gradient norm 1	
Seeds	0, 1 and 2	
	Criterion distillation	**Sensitivity learning**
Data	940 training-half questions	1,203 questions (940 + 263 supplementary), Yes share 0.39
Target	steered teacher's next-token logits	computed Yes/No answer
Strength $\lambda^\star$	Qwen3-8B 22, Qwen3.5-9B 32, gemma-3-12b 2, Mistral-24B 16	–
Control	random direction of the same norm at the same $\lambda^\star$	answers permuted within 255 strata of matched questions
Loss	forward KL, $T = 1$, full vocabulary, answer position	cross-entropy on the answer turn, prompt masked
Optimizer	AdamW, weight decay 0.01, lr 10^{-5}, constant	AdamW, weight decay 0, lr 2×10^{-4}, linear decay, no warm-up
Effective batch	16 (8×2; Mistral-24B 4×4)	16 (2×8; Mistral-24B 1×16)
Length	4 epochs, 235 steps	2 epochs, 152 steps
Max. length	1,024 prompt tokens, 1 output token	2,048 tokens

questions. On the training-half questions alone, the rule returns the same strength for Qwen3-8B, Qwen3.5-9B and gemma-3-12b, and 22 instead of 16 for Mistral-24B.[1] On gemma-3-12b only one nonzero strength qualifies, $\lambda = +2$, with a Yes-rate of 0.47. At $+4$ the Yes-rate is 0.81 and τ falls outside the score range. At every negative strength τ is also outside it or the model answers only No.

Control strength. At the teacher's norm, a random direction changes the Yes–No logit margin far less than $\hat{w}$ does (Table 27), so it is a weak control. We therefore scale two random directions per student until their mean absolute change in the margin, over all FLIPACT questions, equals the teacher's, and refit both coordinates there (Table 28). On gemma-3-12b, on Mistral-24B's first direction and on Qwen3.5-9B's second, the matched random direction perturbs the answers as much as $\hat{w}$ and leaves τ where it was, while $\hat{w}$ moves it well toward acting. The other four directions collapse the model onto one answer before they match the teacher's perturbation, with a Yes-rate of 0 or 1. No matched random direction shifts the criterion toward acting with sensitivity intact.

Table 27: Perturbation of the Yes–No logit margin at each student's selected strength, by the criterion teacher and by a random direction of the same norm.

	Mean $\lvert\Delta\rvert$ Yes–No margin		
Student	teacher	same-norm random	Ratio
Qwen3-8B	18.70	1.18	15.8
Qwen3.5-9B	2.48	0.31	7.9
gemma-3-12b	4.22	0.46	9.2
Mistral-24B	3.24	0.08	40.4

What the student copies. A teacher moved only in the criterion can be distilled because a state flip separates the two coordinates. Write ℓ_+ and ℓ_- for the Yes–No logit margins on the Yes and the No question of a flip. Their difference $D = \ell_+ - \ell_-$ measures how far apart the model holds the two states and tracks a. Their sum $\Sigma = \ell_+ + \ell_-$ measures how far both sit from zero and tracks τ. A

[1] On the full set $\lambda = 22$ is excluded because its τ of -3.05 falls just outside the score range.

Table 28: The criterion teacher beside two random directions per student, each scaled until it perturbs the Yes–No logit margin as much as the teacher does, with both coordinates refit at that strength.

		Criterion teacher		Matched random direction				
Student	Base τ	$\lvert\Delta\rvert$ margin	τ		$\lvert\Delta\rvert$ margin	τ	d'	Yes-rate
Qwen3-8B	$+2.67$	18.70	-2.11	random dir. 1	16.41	–	0.00	0.00
				random dir. 2	17.16	–	0.00	1.00
Qwen3.5-9B	$+2.19$	2.48	$+1.08$	random dir. 1	2.62	–	0.00	0.00
				random dir. 2	2.46	$+2.16$	1.00	0.24
gemma-3-12b	$+2.50$	4.22	$+0.83$	random dir. 1	4.35	$+2.87$	0.64	0.26
				random dir. 2	4.31	$+2.53$	0.63	0.28
Mistral-24B	$>+3$	3.24	$+0.55$	random dir. 1	3.18	$>+3$	0.88	0.12
				random dir. 2	3.17	$>+3$	0.49	0.00

$\lvert\Delta\rvert$ margin: mean absolute change in the Yes–No logit margin over all FLIPACT questions, as in Table 27. A dash marks a fit that cannot be estimated because the model answers one way.

teacher moved only in τ differs from the student in Σ and not in D. The students reproduce 0.88 to 1.04 of their teacher's shift in Σ. The random-direction teachers move Σ far less than the criterion teachers, and their students' shifts in Σ stay small.

Sensitivity learning. The student trains answer-only on the computed answers, with the same LoRA configuration. Its pool is the training half plus supplementary questions from the same tool records (sizes and Yes share in Table 26). The control permutes the answers within strata of matched questions, so prompts, format, step count and Yes-rate are unchanged and about half of the training state flips have their two answers swapped. Shuffling removes the link between a question's state and its answer, and on seven of the twelve runs the student's answers no longer carry a criterion: three runs collapse onto one answer, three lose the slope, and one has $\mathrm{SE}(\tau)$ just above two. Answer-only training also breaks free-form output on Mistral-24B, so several of its external readouts, mostly on BFCL-miss-func and When2Call, cannot be scored. MMLU cannot be read on five of the twelve sensitivity-learning runs.

Transfer. Table 30 lists the values behind Figure 7: the criterion shift on each external benchmark for each distilled student and for its random-direction control.

Accuracy. On all four students, criterion distillation raises balanced accuracy on the external benchmarks where the student starts conservative (Table 31). Where it starts permissive, the same shift overshoots: accuracy falls by up to four points or stays within its error, except on Mistral-24B on When2Tool. There the hit rate is already at ceiling, so a shift toward acting can only add false alarms, and accuracy falls by about ten points.

D.4 A CRITERION-FREE READOUT OF THE EXTERNAL BENCHMARKS

Distillation keeps sensitivity on a readout that carries no criterion term (§5.3). The external d' is the equal-variance estimate from one operating point, and it is exact only when the two evidence distributions share a variance. Scoring each example of every base model, distilled student and control by its answer-token log-probability margin instead gives a full ROC and an area under it (AUC) with no criterion term. Thresholding this margin recovers the binary answers and, on the base models, the d' and c of the main readout. BFCL-miss-func has no answer-token margin and is absent.

Distillation leaves the AUC unchanged for every student on the five benchmarks with answer log-probabilities (Table 32). Each distilled change is equivalent to no change under a two one-sided test at ± 0.02, with a 90% paired cluster-bootstrap interval inside that bound.

On Qwen3-8B the change in d' is the criterion sliding along an unequal-variance ROC. Across benchmarks the size of its criterion shift tracks the change in single-point d' but not the change in

Table 29: Held-out change for every training condition of Table 2, as the seed mean ± the largest per-seed paired bootstrap error. Bold marks a change toward acting in τ, a rise in a after sensitivity learning and a rise in accuracy. A ‡ marks a cell with two usable seeds, and n/a one with fewer. Shaded: control.

Student	Condition	Δa	$\Delta \tau$	Δ acc.
Criterion distillation				
Qwen3-8B	distilled	$+0.09 \pm 0.08$	$\mathbf{-4.44 \pm 0.69}$	$\mathbf{+0.033 \pm 0.018}$
	random direction	$+0.02 \pm 0.01$	$+0.04 \pm 0.05$	$+0.003 \pm 0.004$
Qwen3.5-9B	distilled	-0.01 ± 0.06	$\mathbf{-1.02 \pm 0.22}$	$\mathbf{+0.045 \pm 0.011}$
	random direction	$+0.00 \pm 0.02$	$+0.20 \pm 0.06$	-0.009 ± 0.005
gemma-3-12b	distilled	-0.02 ± 0.04	$\mathbf{-1.86 \pm 0.98}$	$\mathbf{+0.014 \pm 0.012}$
	random direction	$+0.00 \pm 0.01$	-0.11 ± 0.36	$+0.001 \pm 0.004$
Mistral-24B	distilled	-0.17 ± 0.10	$\mathbf{-3.29 \pm 0.83}$	$\mathbf{+0.036 \pm 0.016}$
	random direction	$+0.07 \pm 0.05$	-0.10 ± 0.23	$+0.001 \pm 0.005$
Sensitivity learning				
Qwen3-8B	trained	$\mathbf{+0.63 \pm 0.12}$	$\mathbf{-2.35 \pm 0.53}$	$\mathbf{+0.172 \pm 0.018}$
	shuffled labels	-0.19 ± 0.11	$+0.07 \pm 1.10^{\ddagger}$	-0.044 ± 0.020
Qwen3.5-9B	trained	$\mathbf{+0.48 \pm 0.14}$	$\mathbf{-1.69 \pm 0.30}$	$\mathbf{+0.175 \pm 0.016}$
	shuffled labels	-0.30 ± 0.19	$-3.08 \pm 0.85^{\ddagger}$	-0.038 ± 0.019
gemma-3-12b	trained	$\mathbf{+1.19 \pm 0.21}$	$\mathbf{-2.57 \pm 1.01}$	$\mathbf{+0.238 \pm 0.018}$
	shuffled labels	$-0.20 \pm 0.11^{\ddagger}$	n/a	-0.070 ± 0.020
Mistral-24B	trained	$\mathbf{+0.57 \pm 0.16}$	$\mathbf{-3.39 \pm 0.73}$	$\mathbf{+0.206 \pm 0.020}$
	shuffled labels	-0.61 ± 0.46	n/a	-0.059 ± 0.019

Table 30: Values behind Figure 7: criterion shift on six external benchmarks for the distilled student and its random-direction control, seed mean ± paired error. Negative is toward acting. Shaded: control.

	Benchmark	Qwen3-8B	Qwen3.5-9B	gemma-3-12b	Mistral-24B
Distilled student					
Act now	When2Call	-0.05 ± 0.01	-0.06 ± 0.01	-0.04 ± 0.01	-0.07 ± 0.01
	QuestBench	-1.97 ± 0.05	-0.55 ± 0.03	-0.82 ± 0.03	-0.86 ± 0.04
	When2Tool	-0.58 ± 0.03	-0.34 ± 0.03	-0.32 ± 0.10	-0.52 ± 0.11
	BFCL-miss-func	-0.21 ± 0.05	–	-0.05 ± 0.03	-0.14 ± 0.04
Planning	ACPBench	-0.81 ± 0.12	-0.31 ± 0.10	-0.21 ± 0.07	-0.32 ± 0.10
	CaT-Bench	-0.58 ± 0.04	-0.38 ± 0.04	-0.13 ± 0.01	-0.35 ± 0.02
Random-direction control					
Act now	When2Call	$+0.00 \pm 0.01$	$+0.01 \pm 0.00$	-0.01 ± 0.00	-0.01 ± 0.00
	QuestBench	$+0.05 \pm 0.01$	$+0.09 \pm 0.01$	-0.05 ± 0.01	$+0.00 \pm 0.02$
	When2Tool	$+0.02 \pm 0.01$	$+0.06 \pm 0.01$	$+0.00 \pm 0.01$	$+0.15 \pm 0.09$
	BFCL-miss-func	$+0.02 \pm 0.03$	–	$+0.01 \pm 0.02$	-0.01 ± 0.02
Planning	ACPBench	$+0.01 \pm 0.03$	$+0.06 \pm 0.03$	-0.01 ± 0.02	$+0.00 \pm 0.00$
	CaT-Bench	$+0.01 \pm 0.01$	$+0.08 \pm 0.02$	-0.01 ± 0.00	$+0.03 \pm 0.01$

The dash marks Qwen3.5-9B on BFCL-miss-func, whose base readout does not reproduce.

AUC (Table 33). On QuestBench, where the shift is largest, d' rises by 0.72 and the base model's fitted ROC slope is 1.42, against one under equal variance.

D.5 Answer preference and a Yes-bias teacher

FlipAct reads the decision to act from a Yes or No, so on it a shift toward acting and a shift toward answering Yes coincide. Reversed wordings and a teacher that only favors Yes separate the two (§5.3).

Table 31: Change in balanced accuracy after criterion distillation, seed means, grouped by the starting criterion on each benchmark. Qwen3.5-9B on BFCL-miss-func is left out, as in Figure 7.

Student	Benchmark	Start c	Δ bal. acc.
Started conservative ($c > 0$)			
Qwen3-8B	When2Tool	$+0.42$	$+0.014$
	QuestBench	$+0.71$	$+0.041$
	ACPBench	$+0.40$	$+0.025$
Qwen3.5-9B	When2Tool	$+0.73$	$+0.046$
	ACPBench	$+0.65$	$+0.052$
gemma-3-12b	QuestBench	$+0.91$	$+0.062$
	ACPBench	$+0.31$	$+0.012$
Mistral-24B	QuestBench	$+1.38$	$+0.050$
	BFCL-miss-func	$+0.39$	$+0.022$
	ACPBench	$+0.62$	$+0.046$
Started permissive ($c < 0$)			
Qwen3-8B	When2Call	-0.26	-0.010
	BFCL-miss-func	-0.57	-0.040
	CaT-Bench	-1.13	-0.042
Qwen3.5-9B	When2Call	-0.52	-0.017
	QuestBench	-0.06	-0.011
	CaT-Bench	-1.43	-0.023
gemma-3-12b	When2Call	-0.37	-0.001
	BFCL-miss-func	-0.23	$+0.011$
	When2Tool	-1.13	-0.028
	CaT-Bench	-0.59	-0.015
Mistral-24B	When2Call	-0.04	-0.003
	When2Tool	-1.44	-0.095
	CaT-Bench	-0.86	-0.026

Table 32: Change in criterion-free AUC after criterion distillation and for its random-direction control, seed mean against the base model: the values of Table 4 with the control beside them.

	Qwen3-8B		Qwen3.5-9B		gemma-3-12b		Mistral-24B	
Benchmark	distilled	control	distilled	control	distilled	control	distilled	control
When2Call	-0.008	0.000	-0.006	$+0.001$	$+0.001$	$+0.001$	-0.002	0.000
When2Tool	$+0.004$	-0.001	-0.001	0.000	$+0.011$	$+0.003$	$+0.001$	0.000
QuestBench	-0.012	0.000	-0.002	0.000	-0.005	-0.001	$+0.002$	0.000
ACPBench	$+0.001$	$+0.001$	$+0.001$	-0.001	$+0.004$	$+0.001$	-0.003	0.000
CaT-Bench	$+0.001$	0.000	$+0.008$	-0.001	-0.003	-0.001	$+0.002$	$+0.001$

Table 33: Criterion distillation on Qwen3-8B, seed mean against the base model: size of the criterion shift, change in single-point d' and in criterion-free AUC, and the base model's fitted ROC slope s. Rows are ordered by the size of the criterion shift.

Benchmark	$\lvert\Delta c\rvert$	$\Delta d'$	ΔAUC	s
QuestBench	1.99	$+0.720$	-0.0117	1.42
ACPBench	0.82	$+0.162$	$+0.0007$	0.91
CaT-Bench	0.59	-0.142	$+0.0013$	0.80
When2Tool	0.58	-0.010	$+0.0037$	1.15
When2Call	0.05	-0.067	-0.0077	1.00

QuestBench, 95% paired cluster-bootstrap intervals: $\Delta d'$ $[+0.537, +0.931]$, ΔAUC $[-0.0164, -0.0068]$.

Reversed wordings. Each distilled student and its control answer the held-out questions under the canonical wording and under the three reversed wordings of App. C, with answers recoded so that

every rate is a rate of acting. A shift in the readiness to act raises the act-rate under every wording. A preference for Yes raises it under the canonical wording and lowers it under the reversed ones. Table 34 splits the change in act-rate into a shared part, the mean of the canonical and reversed changes, and a flipping part, half their difference. On all four students most of the shift flips with the polarity, so on FLIPACT's binary readout distillation mostly shifts the preference for Yes. Qwen3-8B and Mistral-24B also carry a smaller shared part, and the random-direction controls move neither part.

Table 34: Change in the held-out act-rate after criterion distillation under the canonical wording and under the three reversed wordings (their mean), and its shared and flipping parts, as seed means.

	Change in act-rate		Parts	
Student	canonical	reversed	shared	flipping
Qwen3-8B	$+0.30$	-0.11	$+0.09$	$+0.20$
Qwen3.5-9B	$+0.07$	-0.03	$+0.02$	$+0.05$
gemma-3-12b	$+0.09$	-0.11	-0.01	$+0.10$
Mistral-24B	$+0.21$	-0.03	$+0.09$	$+0.12$

A Yes-bias teacher. A teacher that only favors Yes tests whether distillation carries more than a Yes preference. This teacher is the base model with the logits of every Yes token raised by a constant, chosen on the training half so that it answers Yes as often as the steered teacher. The data and recipe are unchanged, and the students are Qwen3-8B and gemma-3-12b. On the four Yes/No benchmarks the Yes-bias student moves c as far as the distilled one or further. On the two formats without a Yes answer, the criterion-distilled student moves c further in every seed on both students (Table 35), though by little in one Qwen3-8B seed on BFCL-miss-func. BFCL-miss-func is the stronger test of the two because its reply is free-form, while When2Call keeps option letters as answer tokens.

Table 35: Criterion shift on the two benchmarks without a Yes answer for the distilled, random-direction and Yes-bias students, and the difference between the distilled and the Yes-bias student, paired on examples, overall and per seed. Negative is toward acting.

		Δc, seed mean			distilled $-$ Yes-bias			
Student	Benchmark	distilled	random	Yes-bias	mean	seed 0	seed 1	seed 2
Qwen3-8B	When2Call	-0.051	$+0.004$	-0.026	-0.025 ± 0.005	-0.024	-0.018	-0.031
	BFCL-miss-func	-0.208	$+0.017$	-0.159	-0.050 ± 0.025	-0.072	-0.072	-0.005
gemma-3-12b	When2Call	-0.043	-0.006	-0.006	-0.037 ± 0.006	-0.047	-0.026	-0.037
	BFCL-miss-func	-0.053	$+0.008$	$+0.021$	-0.074 ± 0.022	-0.113	-0.069	-0.040

On Qwen3-8B the Yes-bias student also moves c on both formats, by about half of the distilled shift on When2Call and three quarters on BFCL-miss-func, while on gemma-3-12b it does not. On some models, then, training toward one answer on binary questions reshapes a broader readiness to act. We conjecture that this is one way in which fine-tuning on binary preferences or rewards, as in RLHF (Ouyang et al., 2022) or GRPO (Shao et al., 2024), shifts when a model acts. Testing it is left to future work.

D.6 CHOOSING THE STEERING SIGN AND STRENGTH PER BENCHMARK

Steering with a fixed sign toward acting (the fixed push) lowers balanced accuracy on 12 of the 25 model–benchmark pairs formed by the five steered models of §5.1 and five benchmarks (Table 36). The rule below, summarized in §5.3, instead chooses the sign and strength of the steering push from a few labeled examples of the target benchmark.

The push is chosen on $K = 50$ labeled examples of the target benchmark, whole pairs where the benchmark has pairs, and read on the rest, over twenty random splits. On these fifty examples alone, the rule keeps the strengths at which d' stays within a quarter of its unsteered value and takes

the sign of the unsteered criterion. It pushes nothing if that criterion is smaller than its bootstrap error. Otherwise it takes the kept strength that brings c closest to zero along a monotone path, and it pushes nothing if none qualifies. The same rule sets a scalar bias on the answer-token log-odds where the benchmark has an answer token. The readout is the change in balanced accuracy against no intervention, with a paired bootstrap over examples ($B = 2000$). A model–benchmark pair counts as improved when the median gain exceeds two errors and as degraded when it falls below minus one error, a stricter bar for a gain than for a loss.

Table 36: Choosing the push on fifty examples of each benchmark. (a) Outcomes over model–benchmark pairs, with the fixed push for comparison (shaded). (b) Median change in balanced accuracy over twenty splits for the per-benchmark steering push and for an answer-token bias chosen by the same rule.

(a) Condition	Pairs	Improved	Unchanged	Degraded
Fixed push toward acting	25	2	11	12
Per-benchmark push	25	5	19	1
Answer-token bias	20	5	15	0

(b) Model		When2Call	When2Tool	BFCL-miss-func	ACPBench	CaT-Bench
Qwen3-8B	push	·	+0.022 •	+0.021	+0.021	+0.006
	bias	·	+0.024	–	+0.021	+0.011
Qwen3.5-27B	push	·	·	·	·	·
	bias	·	+0.024	–	+0.011	+0.086 •
gemma-2-9b	push	+0.000	+0.113 •	+0.000	+0.000	·
	bias	+0.031	+0.075 •	–	+0.000	+0.000
Mistral-24B	push	−0.020 ○	·	+0.049 •	·	·
	bias	+0.005	+0.034 •	–	+0.087 •	·
phi-4	push	·	+0.084 •	+0.047 •	+0.024	·
	bias	+0.000	+0.107 •	–	+0.027	·

• improved, ○ degraded, · none chosen, – no answer token.

The per-benchmark push degrades one pair, Mistral-24B on When2Call, and repairs every other pair the fixed push degraded (Table 36). On that pair it moves the criterion the right way but loses 17% of d', within the quarter the rule allows. Where it pushes, its sign matches the sign of the full-benchmark criterion in every split of every pair whose criterion is at least 0.3 in size. Five pairs improve, and eight when the labeled set grows to a hundred examples. On Qwen3.5-27B the direction does not move the criterion toward zero on four benchmarks, so the rule pushes nothing there. The answer-token bias is never worse than no intervention and matches or beats the per-benchmark push on all but one of its pairs. Only the steering push applies to BFCL-miss-func, which has no answer-token margin.

D.7 Qualitative examples

The boxes show single answers of Qwen3-8B, the one model that is both steered in Table 1 and a distillation student. Boxes (a) and (b) illustrate steering (§5.1) on When2Call, where the model picks one of four candidate responses and "call" means it picks the tool call. Boxes (c) and (d) illustrate how distillation keeps sensitivity (§5.3). Each shows a matched pair, one question that warrants acting and one that does not, under the three seeds of the criterion-distilled student and of the random-direction control. There m is the Yes minus No logit at the answer token, averaged over seeds, and the answer is Yes when $m > 0$. Distillation raises both margins, so $m = 0$ can fall between them while their order, the comparison the criterion-free AUC counts, stays as it was. Boxes (a) and (b) show the first example, in sorted id order, that fits the pattern in its title; boxes (c) and (d) each show one of the pairs the distilled student separates while its control does not. The line under each box counts how often the pattern occurs, and Table 37 gives the pair counts behind (c) and (d).

Table 37: Pair counts behind boxes (c) and (d): how often the question that warrants acting keeps the higher margin, and how the distilled student answers the pairs the base model answers No to twice. Shaded: base model and control.

	FLIPACT (c)	QuestBench (d)
Pairs	354	288
Question that warrants acting above its twin		
base model	69%	92%
distilled student	68–69%	93–94%
random-direction control	68–69%	92%
Pairs the base model answers No to twice	189	187
distilled student separates them	50	117
distilled student answers Yes to both	98	59

(a) Steering on a request that needs a tool: When2Call, gold call, unsteered withholds

Tools: `flight.status.check` (airlines, person_name, ticket_number); `uber.eat.order` (restaurant, items, quantities).
User: "Hi, I am Karan. Could you tell me the current status of my flight with Delta? My name is Karan Sharma, and my ticket number is IND4567."

Gold: call (`flight.status.check`) · Unsteered: withhold · Steered, $\lambda^\star = 16$: call
The unsteered model withholds on 56 of the 300 call examples. Steering calls on 9 of them and withholds on 2 calls it had made.

(b) The matching false alarm: When2Call, gold asks for a missing detail, unsteered withholds

Tools: `get_sensor_readings_latest`, `get_sensor_readings_history`, and `get_sensor_readings_history_by_interval` (networkIds, metrics, timespan, interval, ...).
User: "Obtain the sensor reading for temperature for my network L_579838452023959405."

Gold: withhold (ask for the timespan) · Unsteered: withhold · Steered, $\lambda^\star = 16$: call (`get_sensor_readings_history_by_interval` with a timespan the user did not give)
The unsteered model withholds on 210 of the 300 withhold examples. Steering calls on 21 of them (8 that ask for a detail, 13 where no listed tool fits) and withholds on 1 it had called. Qwen3-8B starts slightly permissive on When2Call, so these false alarms outnumber the hits of (a) and accuracy falls (Table 1).

(c) Distillation moves the criterion, not the order: a held-out FLIPACT state flip

Warranted ($\Delta U = +2$): Situation: The food or contents is very cold. The supplies this tool depends on (mains electricity) are available nearby. The goal is to bring the food or contents back to a normal state. Action: a person uses a hotplate. Is using the hotplate a wise choice in this situation? Answer Yes or No.
Destructive ($\Delta U = -2$): the same question with the food or contents "very hot".

	Gold	Base m	Distilled m	Control m
Warranted	Yes	-6.0 (No)	$+14.3$ (Yes, 3/3)	-4.6 (No, 3/3)
Destructive	No	-33.0 (No)	-14.3 (No, 3/3)	-33.2 (No, 3/3)

Over the 354 held-out state flips both margins rise in every flip and seed, and the warranted question stays above the unwarranted one about as often as before (Table 37). Hits and false alarms rise together: of the held-out questions the base model answers No, the distilled student answers Yes in all seeds on 220 of 286 warranted ones and on 137 of 311 destructive ones, and the control on none.

(d) The same on an external benchmark: a QuestBench GSM-Q-verbal pair

Fully specified: "My wife and I went out for dinner at a restaurant that was offering a 10% discount on their meals. After applying the discount, our dinner bill came to $50. We decided to evenly split the check, but she wanted me to pay an additional 20% tip on our discounted bill. How much did I end up paying? Do you already have enough information to give a single correct final answer to this question right now? Do NOT solve it. Answer with exactly one word: Yes or No."
Underspecified: the same problem without the discount and the $50 bill.

	Gold	Base m	Distilled m	Control m
Fully specified	Yes	-0.2 (No)	$+6.4$ (Yes, 3/3)	-0.5 (No, 3/3)
Underspecified	No	-20.8 (No)	-10.2 (No, 3/3)	-20.6 (No, 3/3)

Over the 288 pairs both margins rise in every pair and seed, and the fully specified problem stays above its twin about as often as before (Table 37).

E A MODEL THAT RETURNS A PROBABILITY

Every entry in the paper is a token model read through a Yes/No answer, so its criterion can include a preference for one answer word (§4.3). JEV (TypeSafe AI, 2026) is a decision model of a different kind.[2] Given a state and a statement about it, JEV returns the probability that the statement is true and generates no text, so it gives a readout of FLIPACT with no answer word to prefer. The state is a question's situation and action sentences, and the statement is "Using the *tool* is a wise choice in this situation". The returned probability is read as $P(\text{act})$ and binarized at one half for the fit of Eq. 2. Because this input differs from the prompt the token models receive, JEV is not entered in Figure 5, the panel or Table 9, and the comparisons below are descriptive.

Coordinates. On the statement readout JEV has $a = 0.62$ and $\tau = +1.35$ (Table 38). Its sensitivity is close to that of the most sensitive entries of Table 9. The criterion is conservative, with an interval that excludes zero, and lies between those of Llama-3.3-70B and gemma-4-31b. Mean $P(\text{act})$ steps between $\Delta U = -1$ and 0 and is flat on either side, the same profile as the token models in Figure 4. Asked instead to rate, on an ordered verbal scale, the action's effect on the situation or how warranted it is, JEV tracks the size of the evidence barely better than its probability does. When the statement is reversed, JEV's answer preference in the decomposition of App. C is $b = -0.20 \pm 0.24$, and the probabilities under the two polarities sum to 1.01 on average.

Readout format. Asked to choose between use and avoid instead of judging the statement, JEV returns the same coordinates (Table 38), so on FLIPACT its criterion does not depend on how the binary decision is read out. It does depend on which decisions are offered. On When2Call, where many withhold examples call for asking a question, JEV has $c = -1.15$ when it judges a statement. Offered a choice among call, ask and decline, it routes most of those examples to ask, and c moves to $+0.02$. A criterion read from two options is therefore not a constant of the model when a third option exists, which is one reason an absolute τ need not predict a benchmark's c.

Table 38: JEV on FLIPACT under two readouts of the same decision, which give the same coordinates and the same criterion-free AUC.

Readout	a	95% interval	τ	95% interval	AUC
Statement, binarized at one half	0.62	$[0.47, 0.80]$	$+1.35$	$[+0.83, +2.02]$	0.748
Choice between use and avoid	0.61		$+1.29$		0.750

AUC is that of $P(\text{act})$ against the computed label. Intervals resample situations and are given for the statement readout.

Caveats. JEV is one closed model under one API version, and its responses are deterministic, so every interval resamples situations or examples only. Binarizing its probability also adds a threshold the token models do not have. JEV shows that a conservative criterion and a step at the sign of the evidence can arise without an answer word, but a single model cannot stand in for the 44 models.

[2] Model `jev-latest`, API version `jev-1.13.0`, evaluated 2026-09-22.

F When-to-act benchmarks and steering

Table 39 places twelve when-to-act benchmarks on the design choices of §2: how the ground truth is constructed, whether controlled negatives are matched, whether the evidence for acting is graded, and whether responses are decomposed into sensitivity and criterion. The decisions cover tool-call necessity (Ross et al., 2025; Patil et al., 2025; Huang et al., 2024; Sun et al., 2026), acting or abstaining on minimally perturbed pairs (Liu et al., 2026), answering or abstaining (Kirichenko et al., 2026), clarification (Li et al., 2026; Su & Cardie, 2026), over-refusal (Röttger et al., 2024; Cui et al., 2024; Shi et al., 2024) and physical affordance (Bisk et al., 2020).

Activation steering reads a direction from the residual stream and adds it back (Rimsky et al., 2024; Zou et al., 2023; Turner et al., 2023; Li et al., 2023). Refusal is its best-studied target (Arditi et al., 2024), and sycophancy is a related disposition (Sharma et al., 2024; Perez et al., 2023). App. D.1 projects the refusal and sycophancy directions out of the criterion direction as a control on the steering result.

Table 39: When-to-act benchmarks on the design choices of §2: ✓ full, ∼ partial, ✗ absent.

Benchmark	Decision	Ground truth	Ctrl. neg.	Graded	Sens./crit.
When2Call (Ross et al., 2025)	tool call/abstain	LLM+human	∼	✗	✗
BFCL (Patil et al., 2025)	function applies?	human	∼	✗	✗
MetaTool (Huang et al., 2024)	tool needed?	human/template	∼	✗	✗
When2Tool (Sun et al., 2026)	call vs. no-call	env-computed	✓	∼ (diff.)	✗
Shi et al. (2026)	tool call/abstain	LLM judge	✗	∼ (SAE)	∼ (offset)
AgentAbstain (Liu et al., 2026)	act vs. abstain	pipeline+judge	✓	✗	∼ (paired)
AbstentionBench (Kirichenko et al., 2026)	answer/abstain	curated+judge	✗	✗	✗
QuestBench (Li et al., 2026)	ask vs. solve	rule (CSP)	✓	✗	✗
Su & Cardie (2026)	recognize/clarify	human	✓	✗	∼ (informal)
XSTest (Röttger et al., 2024)	comply/refuse	hand-written	✓	✗	✗
OR-Bench (Cui et al., 2024)	comply/refuse	LLM ensemble	∼	✗	✗
PIQA (Bisk et al., 2020)	physical solution	human	∼	✗	✗
FlipAct	act vs. withhold	by rule	✓ (2×2)	✓ (ΔU)	✓ (a, τ)

G Limitations and outlook

Instrument. Every label is computed by rule from a tool's direction and magnitude on one axis, annotated by language models with no person in the loop, so some questions will carry labels a person would set otherwise. Agreement between annotator families and perturbations of the annotations check how far such errors could move the results (App. A.5 and A.6); no human observers were measured on the diagnostic. All questions share one template in one physical domain. The criterion stays positive in nearly all models under five rephrasings of it (App. C); other templates and domains are untested. Side effects, annotated for part of the tool records, do not enter the rule.

Measurement. Several relations rest on few models. The fall of the criterion with size inside a family rests on three to five models per family. The FlipAct criterion does not track a benchmark's c consistently: near zero on three benchmarks, positive on BFCL on the strength of the three models with the lowest τ, and negative on When2Call at a looser panel cutoff (Table 15, App. B.4), over eighteen models whose criteria sit in a narrow band. Whether this comes from limited power, from the benchmarks' interfaces or from the construct is open. All models are open checkpoints of at most 70B parameters.

Interventions. Two of the intervention models, Qwen3.5-9B and Qwen3.5-27B, answer No under either wording, so part of their starting τ is a preference for the answer token (App. C). The steering direction depends on the wording it is read under, so the interventions are read from behavior, not from the geometry of the direction (App. D.1). After distillation the equivalence test on sensitivity is undecided on Qwen3-8B and Mistral-24B. The Yes-bias comparison covers two students, with

BFCL-miss-func as its only free-form test (App. D.5). The reasoning-mode comparison rests mostly on Qwen models. Sensitivity learning raises a on FLIPACT but does not transfer stably to the external benchmarks: its changes in d' and c there differ in sign across students and are matched by the shuffled-label control. FLIPACT tests basic physical knowledge of tools, and labels on it appear to add little that sharpens the decision to call a tool elsewhere.

Outlook. The build scales with annotator calls (§3.4), and its schema is written in a domain-specific language in which a property, an axis or a rule can be added or removed without rebuilding the pipeline. Richer tool properties such as side effects, rules over several properties and their interactions, and new axes, templates and domains run through the same steps. Beyond single decisions, a richer graph of how tools transform states could chain questions into multi-step agent scenarios, in which a model's tendencies can be read across a sequence of decisions. Such scenarios may also supply what sensitivity learning needs to transfer beyond FLIPACT.