# THE MODEL SAYS WALK: MEASURING WHETHER LLMS CONDITION ON HIDDEN CONSTRAINTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Asked whether to walk or drive to a car wash 50 m away, most language models say *walk*, forgetting that the car has to be there. Such failures are usually measured by accuracy on questions where the hidden constraint applies. We show that this is misleading: a model can answer these questions correctly without reasoning about the constraint at all, simply by favoring the cautious option. We instead ask whether a model's decision changes when the constraint is removed. We test this at three levels of control: log-probability sweeps on open-weight models, a 500-prompt stress test, and CORE-100, a new human-validated benchmark of minimal constraint-present/constraint-absent pairs. The picture is consistent. The constraint nudges decisions rather than governing them. Models still miss presence constraints like the car wash, yet elsewhere they apply constraints that are not there. As a result, standard accuracy flatters all ten models we evaluate and reorders their ranking, and prompting fixes that look effective largely vanish under paired scoring. Claims about hidden-constraint reasoning need paired evidence.

## 1 INTRODUCTION

In February 2026 a one-line question spread across social media: *"I want to wash my car. The car wash is 50 meters away. Should I walk or drive?"* All four chat assistants in the original post recommended walking (Kévin (@knowmadd), 2026), and a follow-up evaluation of 53 models found only five that answered correctly across repeated trials (Opper AI, 2026). The question is diagnostic because it is easy. It needs no specialist knowledge, only noticing that a salient cue (the short distance) is overruled by an unstated feasibility constraint (the car must be there). The same structure appears whenever a model gives practical advice: a nearby pharmacy that cannot fill the prescription, a cheap copy shop that cannot certify a document, a quick fix that the tools at hand cannot perform.

Failures of this kind are now collected into benchmarks (Tang, 2026) and scored in the obvious way: pose questions where the constraint applies and count correct answers. We start from a simple observation about that practice: a correct answer to such a question is ambiguous evidence. "Drive" is what a model says if it registers that the car must be at the car wash. It is also what a model says if it has learned that driving is the safe answer to walk-or-drive questions. The two models come apart only when the constraint is absent. Asked whether to walk to the car wash to buy a gift card, the first says walk and the second still says drive. Accuracy on constraint-active items credits both (Figure 1).

We therefore measure hidden-constraint reasoning as an experimenter would. We pair every constraint-present prompt with a constraint-absent counterpart and ask whether the decision moves when the constraint goes away. We apply this design at three levels of control. The first is a log-probability *case study* of seven open-weight models that sweeps the cue continuously (§3). The second is HOB, a 500-prompt stress test spanning 4 heuristic $\times$ 5 constraint families, run on 14 models (§4). The third is CORE-100, 100 human-validated scenarios whose Active and Removed versions differ by a minimal edit, run on 10 models under a preregistered analysis (§5–6). The three levels agree:

1. **The constraint shifts the decision; it does not gate it.** In every open-weight model, the decision follows the same smooth function of distance whether or not the constraint applies ($r \geq 0.80$ in



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.

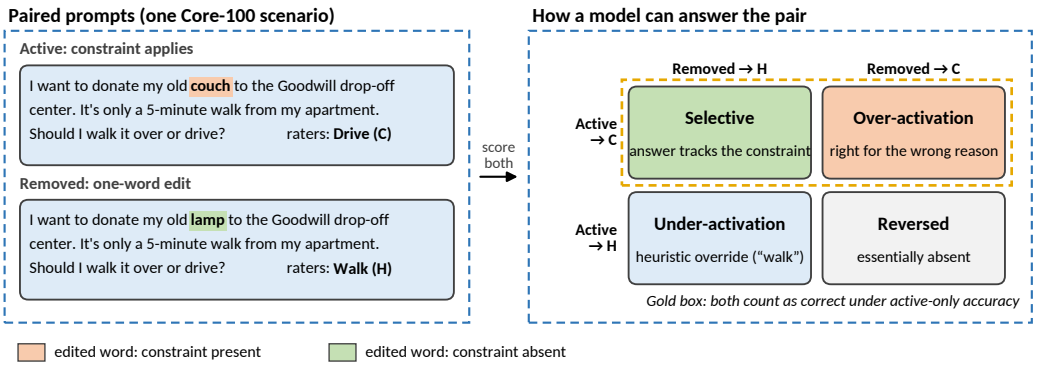


Figure 1: **Why constraint-active accuracy is ambiguous.** Left: a CORE-100 scenario in its Active and Removed versions. A one-word edit (*couch* → *lamp*) removes the constraint and reverses the action human raters prefer. Right: the four ways a model can respond to the pair. Active-only accuracy credits both cells in the gold box; paired scoring separates all four cells.

all 21 model–scenario pairs), and the constraint adds a roughly constant offset. The one model that says drive does so because its whole curve sits lower.

2. **Heuristic override concentrates in presence constraints.** On HOB, presence constraints, the car-wash type, are the hardest family for 12 of 14 models (61.5% vs. 82–93%).
3. **Active-only accuracy flatters every model.** On CORE-100, all ten models choose the constraint-heavy action when the constraint is absent (12–39% of trials, versus 10% for human raters) more often than they miss it when it applies. Active-only accuracy therefore overstates paired performance for every model, the model with perfect active accuracy drops to fourth, and nine of ten models are less selective than the raters. Presence is the one family where models still frequently miss the constraint.
4. **Fixes and graders need paired scoring too.** A prompt that raises GPT-5.4's HOB accuracy by 9 points does not improve its selectivity on CORE-100 ($-0.04$). Only one of twelve preregistered prompting contrasts helps, and a judge told the constraint grades nearly every correct constraint-absent answer as wrong.

**Contributions.** (1) A paired measurement framework for hidden-constraint reasoning, including an identity that makes the bias of active-only accuracy explicit and a continuous analogue for log-probability sweeps (§2). (2) Two benchmarks: HOB for breadth, and CORE-100, 100 human-validated minimal quartets frozen with content hashes before any model was queried. (3) A re-assessment of 17 models, of prompting interventions, and of LLM judging under paired scoring. (4) All data, human ratings, raw outputs, and a one-command reproduction pipeline. Our claims are behavioral. "Right for the wrong reason" is shorthand for a correct constraint-active answer that does not survive its constraint-absent counterpart, not a claim about internal computation.

## 2 MEASURING HIDDEN-CONSTRAINT REASONING WITH PAIRS

**Setting.** A scenario offers a choice between a *constraint-heavy* action $C$ (drive) and a *heuristic* action $H$ favored by a salient cue (walk, because it is close). In the ACTIVE version an unstated feasibility constraint applies, so $C$ is correct. In the REMOVED version one fact is edited so the constraint no longer applies (buying a gift card at the car wash), so $H$ is correct. A model is *selective* if it prefers $C$ in Active and $H$ in Removed. The definition asks whether the answer tracks the constraint, not how the model computes it.

**Paired metrics.** Let $p_{\mathrm{A}} = P(C \mid \text{Active})$ and $p_{\mathrm{R}} = P(C \mid \text{Removed})$, estimated per scenario from repeated samples and averaged over scenarios. We report

$$S = p_{\mathrm{A}} - p_{\mathrm{R}}, \qquad \mathrm{SBA} = \tfrac{1}{2}\big(p_{\mathrm{A}} + (1 - p_{\mathrm{R}})\big) = \tfrac{1+S}{2}, \qquad \mathrm{OA} = p_{\mathrm{R}}, \qquad \mathrm{UA} = 1 - p_{\mathrm{A}}. \quad (1)$$

*Selectivity* $S$ is 1 for a perfectly selective model and 0 for any model that ignores the constraint, including one that always chooses $C$. SBA (selective balanced accuracy) puts $S$ on an accuracy scale. *Over-activation* OA is choosing the constraint-heavy action when the constraint is absent. *Under-activation* UA is missing the constraint when it applies; the car-wash failure is an instance.

**What active-only accuracy measures.** Active-only accuracy is $p_{\mathrm{A}}$, and its gap to the paired score is

$$p_{\mathrm{A}} - \mathrm{SBA} \;=\; \tfrac{1}{2}\,(\mathrm{OA} - \mathrm{UA}). \quad (2)$$

Active-only evaluation thus overstates paired performance exactly when a model errs more often by over-applying a constraint than by missing it, and it cannot separate models that differ only in over-activation. Whether it misleads in practice is an empirical question about the balance of the two errors (§6). To see which error occurs, we also classify each scenario by its modal choices as $C{\to}H$ (selective), $C{\to}C$ (over-activation), $H{\to}H$ (under-activation), or $H{\to}C$ (reversed), as in Figure 1.

**A continuous analogue.** With log-probabilities we can also vary the cue's strength $d$ (e.g., distance). Let $s(x) = \log p(H \mid x) - \log p(C \mid x)$, and compare a *conflict* curve $s_{\text{conf}}(d)$, where the constraint always applies, with a *control* curve $s_{\text{ctrl}}(d)$ without it. If the constraint *gates* the decision, $s_{\text{conf}}(d) < 0$ for all $d$: a flat conflict curve, whatever the control does. If the decision is cue-driven and the constraint is one more input, $s_{\text{conf}}(d) \approx s_{\text{ctrl}}(d) + \delta$: the curves share their shape and differ by a shift $\delta$. We report the correlation $r$ between the curves and the mean shift $\delta$; the paired metrics above are the two-point version of this contrast.

## 3 Case Study: The Car Wash Under Controlled Sweeps

We begin with the canonical failure and the most precise measurement available: exact log-probabilities from open-weight models, which let us vary the cue continuously.

**Setup.** We study Qwen3-{4B, 8B, 14B, 32B}, Qwen3.5-27B, GPT-OSS-20B, and GPT-OSS-120B. The main scenario is the car wash at 100 m (e.g., *"I need my car cleaned. The car wash is 100m from here. Should I walk or drive?"*). Two replications share its structure: filling the tank at a gas station 50 m away, and replacing a flat tire at a tire shop 80 m away. Each scenario has six phrasings. We score decisions without sampling. After the chat-formatted prompt we append a fixed anchor (`\nFinal:`) and compute $s = \log p(\textsc{Walk}) - \log p(\textsc{Drive})$ by teacher forcing (App. A.1). In the *sweep*, we place the scenario's phrasings at 14 log-spaced distances from 10 m to 100 km (*conflict*: Drive is always correct), and do the same for seven coffee-shop phrasings (*control*: walk when near, drive when far). We also *occlude* the goal or the distance phrase (Zeiler & Fergus, 2014), using three replacement operators (mask, neutral, contradicting) to guard against artifacts of any one replacement (Hooker et al., 2019).

**Six of seven models say walk.** Six models prefer Walk on all six car-wash phrasings (mean scores $+2.2$ to $+13.8$ log-odds, not monotone in scale) and answer only 9 of 108 prompts correctly across the three scenarios. GPT-OSS-120B is the exception: it prefers Drive on all six car-wash phrasings ($\bar{s} = -2.4$) and on 16 of 18 prompts overall. Occlusion points the same way. On the car wash, masking, neutralizing, or contradicting the distance moves every model toward Drive, while perturbing the goal, which carries the constraint, has smaller effects of inconsistent sign. The size of this asymmetry depends heavily on the operator (a distance-to-goal effect ratio of 8.7–38 under contradiction but 1.2–19 under neutral substitution; App. Table 3), so we treat occlusion as qualitative and rest our conclusions on the sweeps, which need no replacement text.

**The constraint shifts the curve; it does not gate the decision.** In all seven models the conflict curve has the same sigmoid shape as the control (Figure 2; $r = 0.80$–$0.97$ on the car wash and $r \geq 0.80$ in all 21 model–scenario pairs). The constraint lowers the curve by a roughly constant amount ($-2.0$ to $-10.1$ log-odds on the car wash), and no model shows the flat conflict curve that

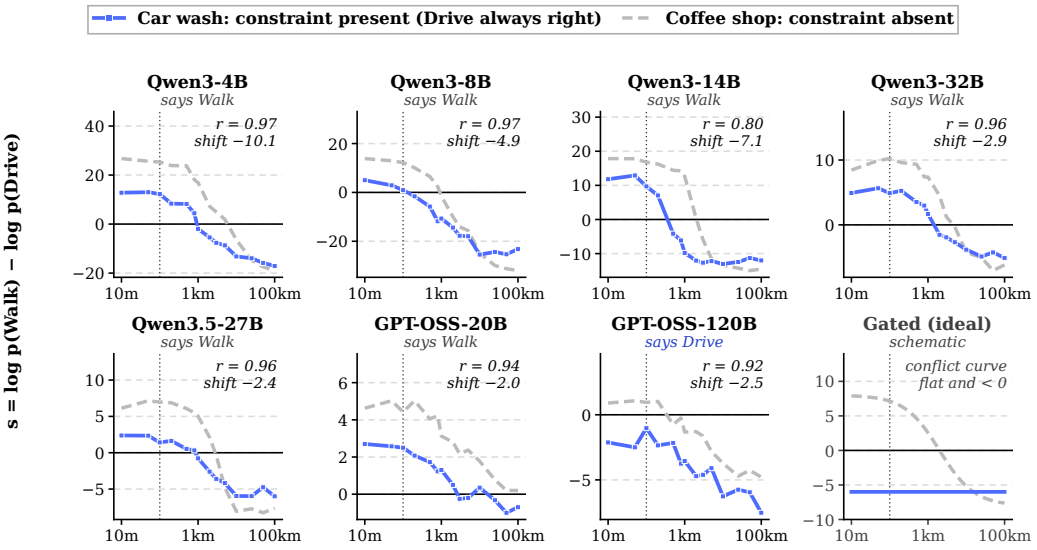


Figure 2: **The constraint shifts the decision curve but does not flatten it.** Decision score versus distance (log scale) for the car wash (conflict, solid) and a coffee-shop control (dashed); $s > 0$ favors Walk, and the dotted line marks 100 m. In every model the conflict curve has the shape of the control ($r$: Pearson correlation), displaced by a roughly constant shift (mean conflict − control, in log-odds). Constraint gating would produce the flat conflict curve of the last panel. GPT-OSS-120B answers correctly because its curves start low, not because its conflict curve is flat. The other two scenarios look the same (App. Fig. 5); App. Fig. 6 overlays all seven conflict curves.

gating predicts. The answer at any distance reflects two terms, a cue-driven walking preference that fades with distance and the constraint's shift; for the six failing models the first wins at short range, and their conflict curves cross into Drive only between 200 m and 3 km. GPT-OSS-120B is right for a revealing reason. Its conflict curve still falls with distance ($r = 0.92$), but its preference for walking is weak even for coffee 10 m away ($s_{\text{ctrl}} = +0.9$), so a shift of $-2.5$ keeps it on the Drive side everywhere. The two terms are balanced differently, but the computation is of the same kind, and accuracy on the test question alone would call this model a constraint reasoner. Sweeps of a price cue and of the distance to carry a heavy purchase show the same shape-sharing ($r$ = 0.51–0.94), with shifts large enough that most models are correct at every grid value without being gated (App. A.2).

## 4 HOB: How Broad Is Heuristic Override?

The case study examines one failure closely. HOB (Heuristic Override Benchmark) asks how widely it occurs, using prompts that any model, open or closed, can answer.

**Design.** Each scenario pairs a goal with a salient cue from one of four *heuristic families* (proximity, efficiency, cost, semantic match) and a hidden constraint from one of five *constraint families*: *presence* (the object must be at the destination), *capability* (the means cannot do the job), *validity* (a precondition is violated), *scope* (the service does not offer it), and *procedural* (a step or time window is not met). Of the 20 combinations, 15 produce natural everyday scenarios (App. Table 6). HOB has 500 prompts: 142 constraint-active base scenarios over 7 domains, 141 constraint-removed rewrites, 90 variants of cue strength, 97 variants of constraint explicitness, and 30 controls without a conflict (an annotated family is in App. B.1). We evaluate 14 models (App. D), sampling each prompt 10 times at temperature 0.7. Free-form answers are graded by a Qwen3-32B judge that sees the question, the gold and shortcut answers, and the constraint. On 300 stratified responses from the same protocol, two annotators agree with the judge 89.7% of the time (Cohen's $\kappa = 0.79$; App. E).

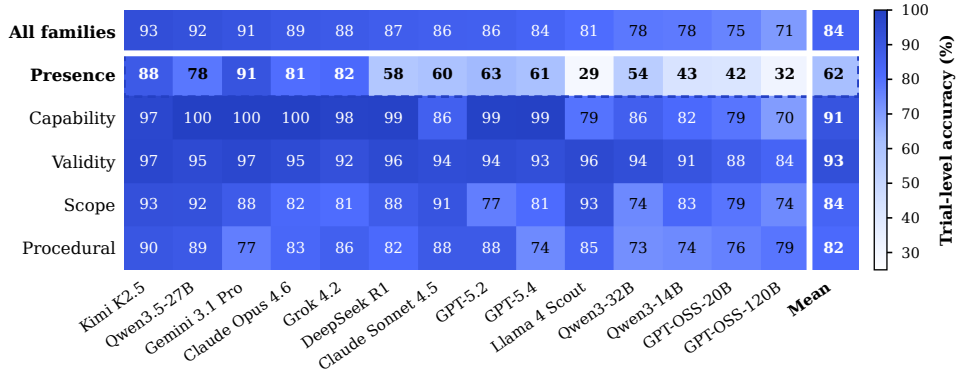


Figure 3: **Presence constraints are where heuristic override concentrates.** Trial-level accuracy (%) on HOB's 142 constraint-active prompts, overall (top row) and by constraint family, with models ordered by overall accuracy. Presence (dashed row) is the hardest family for 12 of 14 models.

We report trial-level accuracy on the constraint-active prompts. The all-10-correct criterion conflates reasoning with sampling stability and appears only in App. Table 8.

**Results.** Mean accuracy on constraint-active prompts is 84.2%, from 70.6% (GPT-OSS-120B) to 93.5% (Kimi K2.5), but errors concentrate sharply (Figure 3). Presence constraints, the car-wash type, are the hardest family for 12 of 14 models: 61.5% on average, against 81.7–93.3% for the other families. Capable models fall to about 60% (GPT-5.4 60.6%, Claude Sonnet 4.5 60.0%, DeepSeek R1 57.8%), and Llama 4 Scout reaches 29.4%. Proximity is the hardest heuristic family (74.0% vs. 83.3–91.2%). Cue strength acts as the case study predicts: on the 19 scenarios that come in strong, medium, and weak versions of the same cue, accuracy rises from 70.0% to 76.1% to 84.3% as the cue weakens, monotonically for 12 of 14 models.

**What HOB cannot measure.** On the constraint-removed rewrites, accuracy drops for 11 of 14 models (to 73.9% on average), which suggests that some correct constraint-active answers reflect a constraint-heavy default. HOB cannot establish this, for two reasons. First, its rewrites are not minimal pairs: only 4 of 100 audited pairs meet the minimal-pair criterion we later apply to CORE-100, and the two versions differ in length by 10.2 words on average (§5). Second, grading constraint-absent answers with a judge that sees the constraint is fragile (§7). The explicitness variants are confounded too, because they change the question along with the cue (the "explicit" car-wash variant asks *how* to get the car there). HOB therefore measures the *breadth* of heuristic override on constraint-active prompts. Measuring *selectivity* needs controlled pairs.

## 5 CORE-100: HUMAN-VALIDATED MINIMAL PAIRS

CORE-100 measures selectivity (Eq. 1) on models that expose only sampled text. It keeps HOB's taxonomy as a coverage frame (all 15 cells, 7 domains) and rebuilds each scenario as a validated quartet.

**Four conditions.** ACTIVE and REMOVED differ by one edit to the constraint-relevant fact (*couch* → *lamp*; Figure 1). The goal, question, action labels, and format stay fixed, and the edit must reverse the action human raters prefer; scenarios where both actions stay acceptable are excluded. EXPLICIT adds one clause stating the prerequisite, never the answer (*"the couch is far too heavy to carry"*). SALIENCECONTROL instead adds an irrelevant clause matched in length, syntax, and position (*"the couch has a light gray fabric cover"*), so Explicit − SalienceControl isolates the content of the cue from the act of adding a clause. Across the 100 pairs, Active and Removed differ in length by 0.25 words on average (at most 2), and their token edit distance has median 2 and

Table 1: **CORE-100 main results** (100 scenarios, 7 samples per condition). $p_{\mathrm{A}}$, $p_{\mathrm{R}}$: probability of the constraint-heavy action in Active and Removed ($p_{\mathrm{R}}$ is the over-activation rate). $S = p_{\mathrm{A}} - p_{\mathrm{R}}$ with scenario-bootstrap 95% CIs. $p_{\mathrm{A}}-\mathrm{SBA}$: how much active-only accuracy overstates paired performance; * survives Holm correction in family F1 (all ten; max $p_{\mathrm{Holm}} = .030$). Rel. ≥6/7: fraction of conditions with at least 6 of 7 samples correct. Last row: validation raters (approximate reference).

| Model | $p_A$ | $p_R$ (OA) | Selectivity $S$ [95% CI] | SBA | $p_A-$SBA | Rel.≥6/7 |
|---|---|---|---|---|---|---|
| DeepSeek V4 Pro | 0.99 | 0.12 | 0.876 [0.82, 0.92] | 0.938 | +0.055* | 0.95 |
| DeepSeek R1 | 0.96 | 0.12 | 0.843 [0.78, 0.90] | 0.921 | +0.043* | 0.92 |
| Gemini 3.1 Pro | 0.99 | 0.17 | 0.826 [0.75, 0.89] | 0.913 | +0.080* | 0.94 |
| Claude Opus 4.6 | 1.00 | 0.18 | 0.816 [0.74, 0.89] | 0.908 | +0.092* | 0.95 |
| Kimi K2.5 | 0.99 | 0.19 | 0.795 [0.72, 0.86] | 0.897 | +0.090* | 0.93 |
| Claude Sonnet 4.5 | 0.94 | 0.23 | 0.714 [0.62, 0.80] | 0.857 | +0.083* | 0.90 |
| Qwen3-32B | 0.90 | 0.25 | 0.653 [0.58, 0.72] | 0.826 | +0.075* | 0.83 |
| GPT-OSS-120B | 0.87 | 0.24 | 0.637 [0.56, 0.71] | 0.819 | +0.056* | 0.81 |
| GPT-5.4 | 0.92 | 0.31 | 0.610 [0.52, 0.70] | 0.805 | +0.111* | 0.86 |
| Llama 4 Scout | 0.80 | 0.39 | 0.410 [0.31, 0.51] | 0.705 | +0.099* | 0.79 |
| Human raters (reference) | 0.96 | 0.10 | 0.867 | 0.933 | – | – |

maximum 7. All 100 pairs meet the minimal-pair criterion (length difference $\leq 2$ words, character similarity $\geq 0.80$), against 4 of 100 of HOB's rewrites (App. Table 11).

**Validation.** Five independent raters, drawn from a pool of 22 trained volunteer annotators, judged every condition of each of 141 candidates, choosing the preferred action (or *both*, *depends*, *neither*) and rating confidence and naturalness (1–5), in three rounds of 20 raters with a Latin square so that no rater saw two conditions of one scenario. Two auditors per round compared each Active/Removed pair side by side and judged whether only the intended fact had changed. The inclusion rule was frozen before any model was queried: at least 4 of 5 raters agree on the action, at most one abstains, mean confidence is at least 3.5, both auditors agree, and mean naturalness is at least 3.0.[1] One hundred scenarios passed. They were frozen in three waves (54/28/18), each with a SHA-256 hash recorded before any model output on that wave. Of the remaining candidates, 24 were excluded for recorded reasons and 17 failed the final round. The ratings also give a human reference: raters chose the constraint-heavy action in 96.5% of Active and 9.8% of Removed ratings ($S = 0.867$). Because these ratings also selected the items, the reference is optimistic.

**Protocol.** We evaluate Claude Opus 4.6, Claude Sonnet 4.5, DeepSeek R1, DeepSeek V4 Pro, Gemini 3.1 Pro, GPT-5.4, GPT-OSS-120B, Kimi K2.5, Llama 4 Scout, and Qwen3-32B (Grok 4.2 was dropped under a preregistered rule when provider throughput made collection infeasible). Each prompt is a forced choice between two labeled options (first line `Answer: A` or `B`; order counterbalanced by a hash of the scenario ID), sampled 7 times per model at temperature 0.7 (provider defaults where unavailable) with an empty system prompt and default reasoning settings. A deterministic parser reads the choice; 99.3–100% of responses conform, and the rest are excluded. Uncertainty comes from a scenario-level bootstrap (10,000 replicates). The analysis was preregistered with hashed code and has two Holm-corrected families: F1 tests the gap $p_{\mathrm{A}} - \mathrm{SBA}$ for each model, and F2 tests the selectivity change for three prompts on four models (§7). Everything else is exploratory. Scenario texts were drafted with LLM assistance (Claude); gold labels and inclusion were decided by human ratings alone.

## 6 RESULTS ON CORE-100

**Active-only accuracy flatters every model.** All ten models over-activate more often than they under-activate, so by Eq. 2 active-only accuracy overstates paired performance for every model

[1] A logged amendment lowered the preregistered naturalness floor from 4.0 to 3.0 on the basis of rating distributions alone: unedited Active texts averaged 3.96, and the 4.0 floor would have excluded every scenario that met all other criteria (App. C).

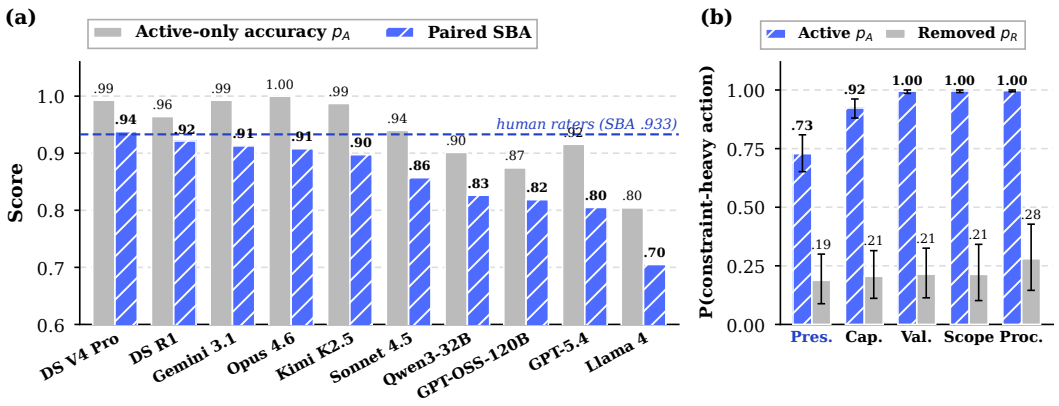


Figure 4: **Two views of CORE-100.** (a) Active-only accuracy $p_{\mathrm{A}}$ (gray) and paired SBA (hatched) for each model, ordered by SBA; the dashed line is the human-rater reference. (b) Probability of the constraint-heavy action pooled over models (exploratory), with scenario-bootstrap 95% CIs. Over-activation ($p_{\mathrm{R}}$) is similar across families; under-activation ($1 - p_{\mathrm{A}}$) is concentrated in presence constraints, the car-wash type.

(Table 1). The gap ranges from $+0.043$ (DeepSeek R1) to $+0.111$ (GPT-5.4), and every gap survives Holm correction. The errors are lopsided: over-activation ranges from 0.12 to 0.39 (raters: 0.098), while under-activation is at most 0.04 for the five strongest models.

**Paired scoring reorders the leaderboard.** Because over-activation is invisible to active-only scoring, models that look alike there separate under pairing (Figure 4a). Claude Opus 4.6 is first by active accuracy ($p_{\mathrm{A}} = 1.00$) but fourth by SBA. DeepSeek R1 rises from fifth to second, and GPT-5.4 falls from seventh to ninth. Among the five models with $p_{\mathrm{A}} \geq 0.96$, over-activation still ranges from 0.12 to 0.19.

**Nine of ten models are less selective than human raters.** Selectivity spans 0.410 (Llama 4 Scout) to 0.876 (DeepSeek V4 Pro). Only V4 Pro exceeds the rater reference of 0.867, and even V4 Pro has a significant active-only gap ($+0.055$). The scenario-level switch matrices (App. Table 16) show that these differences reflect coherent behavior rather than noise: reversed pairs ($H \to C$) occur once across all models, and at most one scenario per model lacks a modal answer. The matrices also separate profiles that a single score merges. Claude Opus 4.6 is a pure over-activator (19 scenarios $C \to C$, none $H \to H$), while GPT-OSS-120B (19 and 14) and Llama 4 Scout (39 and 19) err in both directions.

**The car-wash failure persists where the case study found it (exploratory).** Split by constraint family, the two failure modes separate (Figure 4b). Over-activation is similar in every family ($p_{\mathrm{R}}$ = 0.19–0.28; presence minus others $-0.04$, 95% CI $[-0.15, +0.09]$). Under-activation is concentrated in presence constraints: pooled over models, $p_{\mathrm{A}}$ is 0.73 for presence against 0.92–1.00 elsewhere (difference $-0.25$, $[-0.33, -0.17]$), and selectivity is lower by 0.21 ($[0.09, 0.33]$). Eight of ten models have their lowest Active accuracy on presence scenarios, and the other two are at ceiling in every family. On presence items GPT-5.4 chooses the constraint-heavy action in 47% of samples and Llama 4 Scout in 28%; asked whether to walk across the street to have new tires put on one's car, the average model says drive only 46% of the time. Frontier models have not solved the car-wash problem: they answer most hidden-constraint questions correctly, over-apply those constraints when they are absent, and still miss the presence constraints that started the discussion.

**Stating the constraint helps only where it is missed.** The cue-specific gain (Explicit minus SalienceControl, over the 97 scenarios whose added clauses passed validation) is largest for the models that under-activate: $+0.16$ for Llama 4 Scout, $+0.12$ for GPT-OSS-120B, and $+0.11$ for Qwen3-32B, against at most $+0.07$ for the others. Salience effects are near zero ($-0.023$ to $+0.001$),

Table 2: **Prompting interventions rarely improve selectivity.** Selectivity gain $\Delta S$ over each model's baseline with paired scenario-bootstrap 95% CIs. †: the only one of the 12 preregistered contrasts that survives Holm correction ($p_{\text{Holm}} = .038$). Active gains and Removed harms are in App. Table 21.

| | Primary prompts (family F2) | | | Oracle-like |
|---|---|---|---|---|
| **Model** | CoT | Precond. listing | Goal decomp. | CF check (oracle) |
| GPT-OSS-120B | +0.009 [−0.03, +0.04] | −0.019 [−0.07, +0.03] | −0.056 [−0.10, −0.01] | −0.070 [−0.12, −0.02] |
| GPT-5.4 | +0.001 [−0.03, +0.04] | −0.044 [−0.09, −0.01] | −0.014 [−0.05, +0.02] | +0.009 [−0.04, +0.06] |
| Claude Opus 4.6 | −0.013 [−0.08, +0.05] | +0.027 [−0.04, +0.10] | +0.010 [−0.06, +0.08] | +0.026 [−0.03, +0.08] |
| DeepSeek R1 | +0.004 [−0.02, +0.03] | −0.010 [−0.05, +0.03] | **+0.041 [+0.01, +0.07]**† | −0.311 [−0.39, −0.24] |

so the gains come from the cue's content, not from adding a clause. Requiring at least 6 of 7 correct samples preserves the SBA ordering (Table 1, last column).

## 7 Do Fixes and Graders Survive Paired Scoring?

If active-only accuracy is confounded, so is any evaluation of a fix that looks only at constraint-active items: a prompt that pushes a model toward constraint-heavy answers *everywhere* raises Active accuracy and silently raises over-activation. We therefore score each intervention by its Active gain $\Delta p_{\text{A}}$, its signed Removed harm $\Delta p_{\text{R}}$ (never clipped at zero), and its selectivity gain $\Delta S = \Delta p_{\text{A}} - \Delta p_{\text{R}}$, each paired by scenario against the same model's baseline.

**Design.** Three prompt prefixes form the preregistered family: generic chain-of-thought (*"Think step by step…"*), precondition listing (*"list the necessary conditions that must be true for the stated goal to be accomplished"*), and goal decomposition (*"decompose the stated goal into the discrete steps or resources it requires"*). A fourth, a counterfactual check (*"imagine that the most salient resource or place in the prompt were absent or unavailable"*), presupposes the very contrast being measured, so we treat it as oracle-like and keep it out of the family. We run all four on the Active and Removed prompts of the 100 scenarios (7 samples each) for GPT-OSS-120B, GPT-5.4, Claude Opus 4.6, and DeepSeek R1, which span the selectivity range and both failure profiles.

**Results.** One of the 12 preregistered contrasts survives correction: goal decomposition on DeepSeek R1 ($\Delta S = +0.041$, 95% CI $[+0.01, +0.07]$), a small gain on a model that was already among the most selective (Table 2). Six of the twelve estimates are negative. GPT-5.4 shows why pairing matters. Precondition listing leaves its Active accuracy unchanged ($+0.004$; the model is near ceiling) but raises over-activation by $+0.049$, a selectivity loss that active-only scoring would record as "no harm." The oracle-like check is the starkest case. On DeepSeek R1 it raises over-activation by $+0.31$ and cuts selectivity from 0.84 to about 0.53, while Active accuracy barely moves ($-0.003$). Asking a model to imagine the resource absent makes it answer as if the constraint always applied.

**The same prompt looks like a fix under HOB's protocol.** On HOB, precondition listing raised the all-10-correct accuracy over all 500 prompts from 65.8% to 74.8% for GPT-5.4 and from 51.2% to 57.8% for Llama 4 Scout (App. Table 10). Splitting these aggregates by item type shows how they mislead. For Llama 4 Scout the whole gain comes from constraint-removed prompts (28.4% → 58.2%), while accuracy on the constraint-active prompts the intervention targets *falls* (66.9% → 57.0%). The aggregate reported a fix for a prompt that made the model worse at noticing the constraint. Because HOB and CORE-100 differ in pair construction, format, and scoring, we cannot attribute the disagreement to one factor. On validated pairs, however, none of the tested prompts improves selectivity by more than a few points.

**Judges should grade each condition on its own terms.** Rubric-based LLM judges (Zheng et al., 2023) are often told what an item tests, and for paired items it is natural to share one rubric across both versions. A second judge (GPT-5.4, temperature 0) graded 480 stratified CORE-100 responses with known parsed choices under two rubrics: one carrying the scenario's shared metadata (the Active-form constraint, e.g., *"drive-through requires a vehicle"*, and the shortcut label), and one

containing only the question, correct answer, and response. The neutral rubric agreed with the parser on all 480 responses. The shared rubric agreed on all 360 Active, Explicit, and SalienceControl responses but on only 25 of 120 Removed responses, marking 95 of the 96 correct constraint-absent answers as wrong (App. E). A judge told the constraint enforces it where it does not apply, manufacturing the very over-activation that selectivity must detect; over-activation should be measured from parsed choices or with per-condition rubrics. Human re-graders likewise agree less with HOB's judge on removed items (75–87% vs. 89–97%).

## 8 Related Work

**Shortcuts and heuristics.** Models exploit cues that are spuriously predictive of the label (McCoy et al., 2019; Gururangan et al., 2018; Geirhos et al., 2020; Du et al., 2024; Nikankin et al., 2025; Tang et al., 2023; Sun et al., 2024; Yuan et al., 2024) and show analogues of human cognitive biases (Binz & Schulz, 2023; Suri et al., 2024; Wang et al., 2024; Lampinen et al., 2024; Echterhoff et al., 2024; Malberg et al., 2025). Our case study documents such a shortcut on the input side; our main concern is its mirror image in evaluation, where a constant preference for the constraint-heavy action is itself spuriously predictive of an active-only benchmark's gold label.

**Distractors, constraints, and commonsense.** Distractor benchmarks add irrelevant text to self-contained problems (Shi et al., 2023; Mirzadeh et al., 2025; Yang et al., 2025), and instruction-following suites test *stated* constraints (Zhou et al., 2023; Chen et al., 2025; Song et al., 2026). Our constraints are never stated and must be inferred, as in commonsense benchmarks (Levesque et al., 2012; Clark et al., 2018; Zellers et al., 2019; Bisk et al., 2020) and the frame problem (McCarthy & Hayes, 1981). The car-wash question has been tested on 53 models (Opper AI, 2026) and in open suites (Allen, 2026); structured prompting improves accuracy on it (Jo, 2026); and BrainBench scores 100 such brainteasers in isolation (Tang, 2026).

**Behavioral testing and benchmark validity.** CORE-100 brings the tradition of behavioral tests (Ribeiro et al., 2020), minimal pairs (Warstadt et al., 2020; Jumelet et al., 2026), and contrast sets or counterfactual data (Gardner et al., 2020; Kaushik et al., 2020; Roewer-Després et al., 2025) to hidden constraints: each edit must reverse a human-validated unique preference, and pairs are scored as selectivity under a counterbalanced forced choice. Preregistration, frozen artifacts, and the judge audit respond to critiques of benchmark validity (Bowman & Dahl, 2021; Raji et al., 2021; Reuel et al., 2024), judge bias (Zheng et al., 2023), and option-order effects (Pezeshkpour & Hruschka, 2024; Zheng et al., 2024). Our case study describes an input–output function, not a circuit (Singh et al., 2024); probing and patching (Geiger et al., 2021; Conmy et al., 2023; Marks et al., 2025) on minimal pairs such as CORE-100 is a natural next step.

## 9 Limitations and Conclusion

**Limitations.** All results are behavioral: the case study describes how the decision depends on cue and constraint, not the mechanism, and covers seven open-weight models and three scenarios. HOB's rewrites are not minimal and its judge sees the constraint, so it serves only for breadth. CORE-100's golds come from a fixed pool of 22 annotators whose ratings also selected the items, so the human reference is optimistic; its texts were LLM-drafted, and it is English-only. Forced choice narrows behavior relative to free-form advice, API models are July 2026 snapshots, and each intervention is one template on four models. Contamination from the viral question may inflate Active accuracy, though paired scoring partly discounts it (App. H).

**Conclusion.** Asked about the car wash, most models say walk, yet a model that says drive has not necessarily thought about the car. Across three levels of control and 17 models, the cue sets the decision and the constraint only shifts it. Paired scoring separates two failures: models still miss presence constraints, the car-wash type, and in every family they choose the constraint-heavy action when the constraint is absent. Active-only accuracy rewards the second failure, so it flatters all ten CORE-100 models, reorders their ranking, and misjudges fixes and graders. Reasoning about a hidden constraint means acting on it when it holds and setting it aside when it does not; only paired Active/Removed evidence tests both.

### AI USE STATEMENT

We used generative AI tools in three ways. (1) Benchmark construction: CORE-100 and HOB scenario texts were drafted and repaired with an LLM (Claude) under the construction rules of §5. Gold labels, item inclusion, and exclusion were determined solely by human ratings and audits. (2) Evaluation instruments: LLMs served as the models under study and, in HOB and the judge study, as graders; both uses are described in the paper. (3) Engineering and writing: AI assistants helped write analysis code and edit the manuscript. All AI-assisted code was checked by regenerating every reported number from raw outputs, and all text was reviewed by the authors. We take responsibility for the full content of this work.

### ETHICS STATEMENT

The validation of CORE-100 was carried out by 22 human annotators who are not authors of this paper. They were recruited as volunteers, compensated at $35 per hour, and gave informed consent. Before rating they received written instructions and training, and they rated independently, without access to one another's ratings, the answer key, or model outputs. The study was approved by our institution's IRB. No personal, demographic, or contact information was collected, and anonymized judgments are released. The benchmarks contain everyday decision scenarios. Candidates whose constraints proved contestable in high-stakes settings (e.g., some medical items) were excluded; retained medical- or legal-adjacent items were not expert-adjudicated and should not be used as professional guidance. The failure modes we document concern advice quality, and we see no direct avenue for misuse.

### REPRODUCIBILITY STATEMENT

We release, as anonymized supplementary material: both benchmarks (HOB's 500 prompts with labels and the three hash-frozen CORE-100 waves with the full 141-candidate construction ledger and exclusion log); all human ratings and pair audits; all raw model outputs with per-response provenance (model identifier, provider, decoding settings, finish reason, timestamps); the preregistration with its logged amendments; and code that regenerates every table and figure from the raw outputs with one command (App. I). Case-study scoring is deterministic. Model and decoding configurations are listed in App. D.

## REFERENCES

Ryan Allen. Car wash paradox evals. GitHub repository, https://github.com/ryan-allen/car-wash-evals, 2026. Accessed: 2026-06-09.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023. doi: 10.1073/pnas.2218523120.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, 2021.

Jianghao Chen, Zhenlin Wei, Zhenjiang Ren, Ziyong Li, and Jiajun Zhang. LR$^2$Bench: Evaluating long-chain reflective reasoning capabilities of large language models via constraint satisfaction problems. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6006–6032, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.312. URL https://aclanthology.org/2025.findings-acl.312/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut learning of large language models in natural language understanding. *Communications of the ACM*, 67(1):110–120, 2024. doi: 10.1145/3596490.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. Cognitive bias in decision-making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12640–12653, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.739. URL https://aclanthology.org/2024.findings-emnlp.739/.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017/.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Heejin Jo. Prompt architecture determines reasoning quality: A variable isolation study on the car wash problem. *arXiv preprint arXiv:2602.21814*, 2026.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Transactions of the Association for Computational Linguistics*, 14:193–216, 2026.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR)*, 2020.

Kévin (@knowmadd). Car wash reasoning test. Mastodon post, https://mastodon.world/@knowmadd/116072773118828295, February 2026. Original post, February 15, 2026. Accessed: 2026-06-09.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models, like humans, show content effects on reasoning tasks. *PNAS nexus*, 3(7):pgae233, 2024. doi: 10.1093/pnasnexus/pgae233.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. The Winograd schema challenge. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference, KR 2012*, Rome, Italy, 2012. AAAI Press. URL https://aaai.org/papers/59-4492-the-winograd-schema-challenge/.

Simon Malberg, Roman Poletukhin, Carolin Schuster, and Georg Groh. A comprehensive evaluation of cognitive biases in LLMs. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pp. 578–613, Albuquerque, USA, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.nlp4dh-1.50. URL https://aclanthology.org/2025.nlp4dh-1.50/.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=I4e82CIDxv.

John McCarthy and Patrick J Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*, pp. 431–450. Elsevier, 1981.

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, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic without algorithms: Language models solve math with a bag of heuristics. In *The Thirteenth International Conference on Learning Representations*, 2025.

Opper AI. Car wash test on 53 leading AI models. Blog post, https://opper.ai/blog/car-wash-test, 2026. Published February 19, 2026. Accessed: 2026-06-09.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 21763–21813, 2024.

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.

François Roewer-Després, Jinyue Feng, Zining Zhu, and Frank Rudzicz. ACCORD: Closing the commonsense measurability gap. 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. 3799–3829, 2025.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pp. 31210–31227. PMLR, 2023.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*, 2024.

Da Song, Yuheng Huang, Boqi Chen, Tianshuo Cong, Randy Goebel, Lei Ma, and Foutse Khomh. Evaluating implicit regulatory compliance in LLM tool invocation via logic-guided synthesis. *arXiv preprint arXiv:2601.08196*, 2026.

Zechen Sun, Yisheng Xiao, Juntao Li, Yixin Ji, Wenliang Chen, and Min Zhang. Exploring and mitigating shortcut learning for generative large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 6883–6893, Torino, Italia, 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.602/.

Gaurav Suri, Lily R Slater, Ali Ziaee, and Morgan Nguyen. Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology: General*, 153(4):1066–1075, 2024. doi: 10.1037/xge0001547.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4645–4657, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.284. URL https://aclanthology.org/2023.findings-acl.284/.

Yuzhe Tang. BrainBench: Exposing the commonsense reasoning gap in large language models. *arXiv preprint arXiv:2603.14761*, 2026.

Pengda Wang, Zilin Xiao, Hanjie Chen, and Frederick L Oswald. Will the real Linda please stand up... to large language models? Examining the representativeness heuristic in LLMs. *arXiv preprint arXiv:2404.01461*, 2024.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.

Minglai Yang, Ethan Huang, Liang Zhang, Mihai Surdeanu, William Yang Wang, and Liangming Pan. How is LLM reasoning distracted by irrelevant context? An analysis using a controlled benchmark. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13329–13347, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.674. URL https://aclanthology.org/2025.emnlp-main.674/.

Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. Do LLMs overcome shortcut learning? An evaluation of shortcut challenges in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12188–12200, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.679. URL https://aclanthology.org/2024.emnlp-main.679/.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations (ICLR)*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

## APPENDIX OVERVIEW

A Case-study details B HOB details C CORE-100 construction and validation D Models and collection settings E Judge studies F CORE-100 per-model results G Interventions H Extended limitations I Reproduction.

# A CASE-STUDY DETAILS

## A.1 SCORING PROCEDURE

All seven models run locally in `bfloat16` (GPT-OSS in its native MXFP4 weights), with the model's chat template and default thinking setting. For a prompt $x$, we append a fixed anchor (`\nFinal:`) after the assistant prefix and compute, by teacher forcing, the log-probability of each answer word (WALK, DRIVE). For each answer we sum probability over tokenization variants (e.g., with and without a leading space) by log-sum-exp. The decision score is $s(x) = \log p(\text{WALK} \mid x) - \log p(\text{DRIVE} \mid x)$. No text is generated, so scores are deterministic, and variability comes only from phrasing. Each scenario has six phrasings: a canonical text and five paraphrases.

**Scenarios.** Canonical texts: *"I want to get my car washed. The carwashing store is just 100m away from my home. Should I walk there or drive there?"*; *"I want to add gas for my car. The gas station is 50m away from my home. Should I walk there or drive there?"*; *"I got a flat tire and need to replace it with a new tire. The tire shop is 80m away from my home. Should I walk there or drive there?"*. In all three, DRIVE is correct.

**Occlusion.** For the goal span (e.g., *"get my car washed"*) and the distance span (e.g., *"just 100m away from my home"*) we apply three operators: *mask* (replace with ____), *neutral* (a vaguer phrase: *"take care of something"* for the goal, *"not far from my home"* for the distance), and *contradict* (a phrase with the opposite implication: *"buy a coffee"*, which removes the constraint, and *"100km away from my home"*). The operators are not commensurate across spans (the neutral distance phrase still implies proximity, while the contradicting one is extreme), which is one reason the effect ratio varies so much across operators. The effect of occluding span $z$ is $A(z) = s(\text{occ}(x, z)) - s(x)$. Table 3 reports the ratio of the mean absolute effects, $|\bar{A}(\text{distance})|/|\bar{A}(\text{goal})|$, separately for each operator. The ratio depends strongly on the operator, so we treat it as a qualitative ranking only.

**Sweeps.** Distances: 10, 50, 100, 200, 500, 800 m; 1, 2, 3, 5, 10, 25, 50, 100 km. Conflict prompts reuse the scenario's six phrasings with the distance substituted. Control prompts use seven coffee-shop phrasings (e.g., *"I want to grab coffee. The coffee shop is only $\{d\}$ away. Should I walk or drive?"*). Each point averages five phrasings drawn from the pool (with replacement).

Table 3: Case-study summary per model. Accuracy: phrasings (of 6) on which DRIVE is preferred. $\bar{s}$: mean car-wash decision score. Effect ratio: $|\bar{A}(\text{distance})|/|\bar{A}(\text{goal})|$ on the car wash under each occlusion operator. $r$ and shift: correlation and mean difference between the car-wash (conflict) and coffee-shop (control) sweep curves.

| | **Accuracy** (6 phrasings) | | | | **Distance/goal effect ratio** | | | **Distance sweep** | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | Car wash | Gas | Tire | $\bar{s}$ | Mask | Neutral | Contra. | $r$ | shift |
| Qwen3-4B | 0/6 | 1/6 | 1/6 | +13.8 | 3.1 | 2.0 | 8.7 | 0.97 | −10.1 |
| Qwen3-8B | 0/6 | 2/6 | 4/6 | +4.7 | 10.4 | 7.6 | 38.2 | 0.97 | −4.9 |
| Qwen3-14B | 0/6 | 0/6 | 1/6 | +12.0 | 37.1 | 18.7 | 32.7 | 0.80 | −7.1 |
| Qwen3-32B | 0/6 | 0/6 | 0/6 | +5.9 | 10.8 | 6.2 | 28.7 | 0.96 | −2.9 |
| Qwen3.5-27B | 0/6 | 0/6 | 0/6 | +2.2 | 2.5 | 1.2 | 9.2 | 0.96 | −2.4 |
| GPT-OSS-20B | 0/6 | 0/6 | 0/6 | +2.3 | 1.3 | 2.2 | 14.7 | 0.94 | −2.0 |
| GPT-OSS-120B | 6/6 | 6/6 | 4/6 | −2.4 | 4.5 | 0.5 | 4.2 | 0.92 | −2.5 |

Table 4: Sweep statistics for all 21 model–scenario pairs. $s(10\,\text{m})$: decision score at the shortest distance. "First distance with $s < 0$": where the curve first favors DRIVE ("never": not within 100 km). The control (coffee shop) is shared across scenarios.

| **Scenario** | **Model** | $s_{\text{conflict}}(10\,\text{m})$ | $s_{\text{control}}(10\,\text{m})$ | **First distance with $s < 0$** | | $r$ | shift |
|---|---|---|---|---|---|---|---|
| | | | | conflict | control | | |
| Car wash (100 m) | Qwen3-4B | +12.8 | +26.7 | 1 km | 10 km | 0.97 | −10.1 |
| | Qwen3-8B | +5.0 | +13.8 | 200 m | 1 km | 0.97 | −4.9 |
| | Qwen3-14B | +11.8 | +17.8 | 500 m | 3 km | 0.80 | −7.1 |
| | Qwen3-32B | +4.9 | +8.4 | 2 km | 5 km | 0.96 | −2.9 |
| | Qwen3.5-27B | +2.4 | +6.1 | 1 km | 3 km | 0.96 | −2.4 |
| | GPT-OSS-20B | +2.7 | +4.6 | 3 km | never | 0.94 | −2.0 |
| | GPT-OSS-120B | −2.1 | +0.9 | 10 m | 500 m | 0.92 | −2.5 |
| Gas station (50 m) | Qwen3-4B | +10.9 | +26.7 | 3 km | 10 km | 0.95 | −11.2 |
| | Qwen3-8B | +1.5 | +13.8 | 50 m | 1 km | 0.99 | −5.3 |
| | Qwen3-14B | +15.7 | +17.8 | 500 m | 3 km | 0.86 | −5.1 |
| | Qwen3-32B | +5.7 | +8.4 | 3 km | 5 km | 0.97 | −1.5 |
| | Qwen3.5-27B | +1.6 | +6.1 | 500 m | 3 km | 0.90 | −2.2 |
| | GPT-OSS-20B | +1.8 | +4.6 | 3 km | never | 0.92 | −2.4 |
| | GPT-OSS-120B | −1.5 | +0.9 | 10 m | 500 m | 0.86 | −2.0 |
| Tire shop (80 m) | Qwen3-4B | +8.1 | +26.7 | 1 km | 10 km | 0.95 | −11.2 |
| | Qwen3-8B | +1.0 | +13.8 | 200 m | 1 km | 0.94 | −3.2 |
| | Qwen3-14B | +15.0 | +17.8 | 500 m | 3 km | 0.80 | −3.4 |
| | Qwen3-32B | +5.7 | +8.4 | 3 km | 5 km | 0.95 | −1.1 |
| | Qwen3.5-27B | +2.0 | +6.1 | 3 km | 3 km | 0.91 | −0.2 |
| | GPT-OSS-20B | +2.1 | +4.6 | 10 km | never | 0.95 | −2.0 |
| | GPT-OSS-120B | +0.4 | +0.9 | 100 m | 500 m | 0.94 | −0.5 |

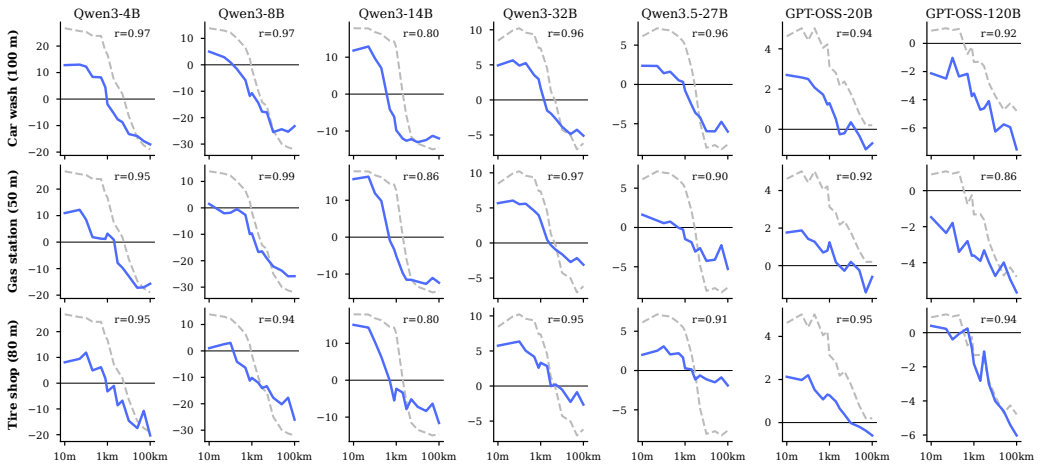


Figure 5: Conflict (solid) and control (dashed) sweeps for all three scenarios and seven models. The shape-sharing of Figure 2 holds in every panel ($r \geq 0.80$).

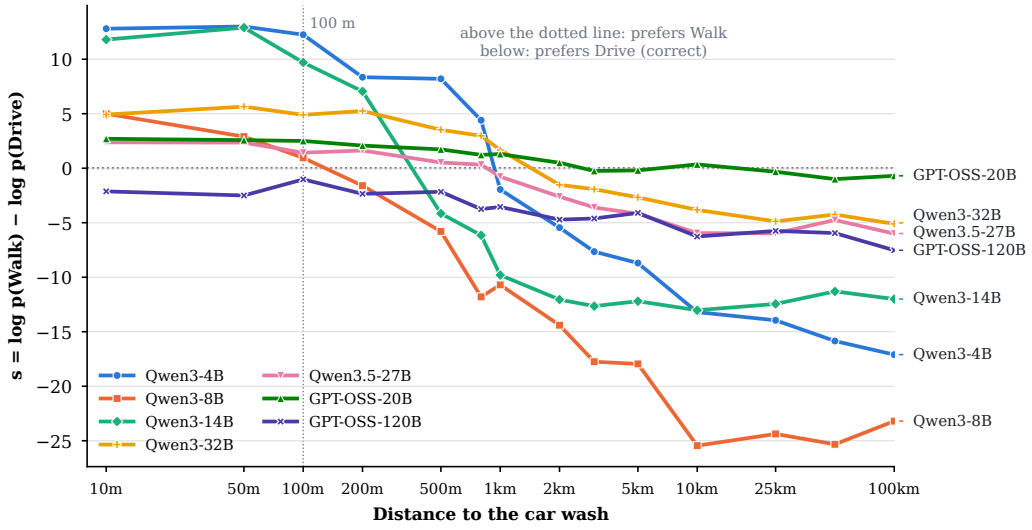


Figure 6: All seven car-wash conflict curves on one axis. Every model's preference for walking fades smoothly with distance, and six of the seven prefer Walk at the 100 m of the test question. Because each model has its own coffee-shop control (Figure 2), controls are omitted here: a single shared control line would misstate what each model is compared against.

### A.2 Sweeps beyond distance

We ran four further sweeps on six of the models (not GPT-OSS-120B), each with a conflict and a control condition. The *cost* sweep (H-cost×C-scope) varies the price of a cheaper venue that cannot provide the needed service (13 grid points, \$0–\$500). The *heavy-item* sweep (H-prox×C-cap) varies the distance to carry a large purchase home (12 points, 50 m–50 km). The *time* sweep (H-eff×C-cap) varies the time saved by moving a 400–600 lb object oneself (10 points, 1 min–8 h). The *name* sweep (H-sem×C-scope) makes a nearby venue sound increasingly car-related (7 points, from "a small convenience store" to "a full-service gas station with a car care center") for a tire repair. Figure 7 and Table 5 summarize the results.

The cost and heavy-item sweeps behave like the distance sweep. Conflict and control curves are correlated ($r$ = 0.51–0.94), but the constraint's shift is large enough that 5 of 6 and 4 of 6 models choose the constraint-respecting option at every grid value. The time and name sweeps cannot be read as conflict/control contrasts, because their control curves never cross the decision boundary. In the time sweep, Qwen3-4B and GPT-OSS-20B choose to move the object themselves at every grid value, and the larger models never do. In the name sweep, the conflict score rises toward the gas station in all six models as the name becomes more car-related. However, the top grid values (a gas station with a car care center) can plausibly repair tires, so the gold label is contestable there, and we draw no conclusion from this sweep.

Table 5: Parametric sweeps: percentage of grid values at which the model prefers the constraint-respecting option under conflict, and the conflict–control correlation $r$.

| Sweep | Qwen3-4B | Qwen3-8B | Qwen3-14B | Qwen3-32B | Qwen3.5-27B | GPT-OSS-20B |
|---|---|---|---|---|---|---|
| H-cost × C-scope | 100% / 0.93 | 100% / 0.92 | 100% / 0.87 | 100% / 0.81 | 100% / 0.51 | 46% / 0.94 |
| H-prox × C-cap | 100% / 0.76 | 100% / 0.92 | 100% / 0.83 | 83% / 0.90 | 100% / 0.85 | 75% / 0.81 |
| H-eff × C-cap | 0% / 0.29 | 50% / −0.17 | 90% / −0.19 | 100% / −0.76 | 100% / −0.32 | 0% / −0.27 |
| H-sem × C-scope | 43% / 0.49 | 0% / 0.11 | 0% / 0.22 | 71% / 0.38 | 86% / −0.03 | 14% / 0.80 |

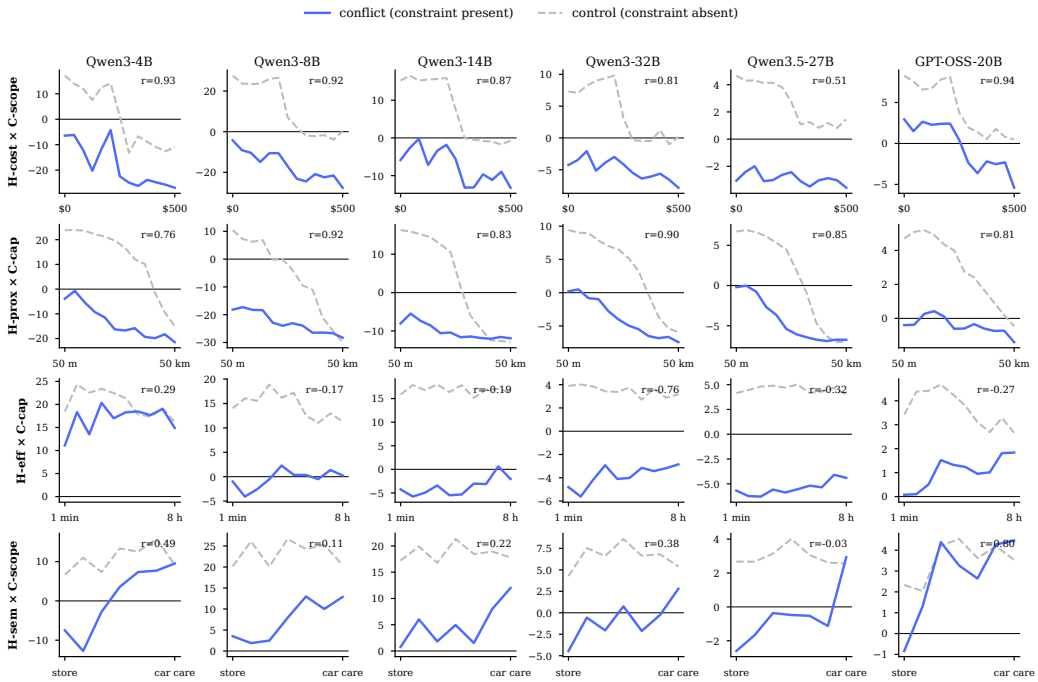


Figure 7: Parametric sweeps beyond distance (rows) for six open-weight models (columns); $s > 0$ favors the heuristic option. See text for why the time and name sweeps are not interpreted as conflict/control contrasts.

# B HOB DETAILS

Table 6: HOB taxonomy: what misleads the model and what it misses. 15 of the 20 combinations are populated.

| **Heuristic family** (the salient cue) | | | **Constraint family** (the unstated requirement) | | |
|---|---|---|---|---|---|
| H-prox | Proximity | "only 5 minutes away" | C-pres | Presence | object must be at the destination (car at the car wash) |
| H-eff | Efficiency | "the quickest way" | C-cap | Capability | the means cannot do the job (carry a sofa on foot) |
| H-cost | Cost | "the free option" | C-val | Validity | a precondition is violated (drive on a flat tire) |
| H-sem | Semantic | "gas station" for tires | C-scope | Scope | the service does not offer it (tire repair at a gas station) |
| | | | C-proc | Procedural | a step or time window is not met (the store is closed) |

**Composition.** HOB contains 500 prompts: 142 constraint-active base scenarios, 141 constraint-removed rewrites, 90 heuristic-strength variants, 97 explicitness variants, and 30 controls. They cover 15 populated cells and 7 domains (transportation, shopping, digital, medical, home, work, travel). The populated cells are all five constraint families for proximity and efficiency cues, four for cost cues (all but presence), and scope for semantic cues.

## B.1 AN ANNOTATED SCENARIO FAMILY

Table 7 reproduces one scenario family verbatim from the released data. It shows that the constraint-removed rewrite changes more than the constraint, and that the "explicit" variant changes the question itself.

Table 7: The HOB family A1-001 (H-prox×C-pres), verbatim.

| **Item** | **Prompt** | **Gold** |
|---|---|---|
| Base | I need to get my car washed before a road trip tomorrow. The car wash is only 5 minutes away from my apartment. Should I walk there or drive? | Drive |
| Removed | I need to pick up a gift card from the car wash counter for my friend's birthday. The car wash is only 5 minutes away. Should I walk there or drive? | Walk |
| Strong cue | ... The car wash is literally 50 meters from my front door. Should I walk there or drive? | Drive |
| Medium cue | ... The car wash is about a 10-minute walk. Should I walk there or drive? | Drive |
| Weak cue | ... The car wash is about 2 miles away. Should I walk there or drive? | Drive |
| Implicit | I need to get my car washed. The car wash is 5 minutes away. Walk or drive? | Drive |
| Hint | I want to get my car washed at the car wash. It's 5 minutes away. Walk or drive? | Drive |
| Explicit | My car is parked at home. The car wash is 5 minutes away. How do I get the car to the car wash? | Drive |

### B.2 PER-MODEL RESULTS

Table 8: HOB per model. Trial: mean accuracy over the 10 samples. Strict: fraction of prompts with all 10 samples correct. "All 500 strict" mixes all item types, including constraint-removed rewrites and controls, and is *not* a measure of override accuracy. Family columns: trial-level accuracy on constraint-active base prompts.

| | **Constraint-active** | | **Constraint-removed** | | **All 500** | **Constraint-active, trial-level, by family** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | trial | strict | trial | strict | strict | Pres. | Cap. | Val. | Scope | Proc. |
| Kimi K2.5 | 93.5 | 81.7 | 73.3 | 48.2 | 69.0 | 88.3 | 96.7 | 96.9 | 92.6 | 90.0 |
| Qwen3.5-27B | 92.3 | 83.1 | 72.3 | 53.9 | 72.2 | 78.3 | 100.0 | 95.2 | 92.4 | 88.8 |
| Gemini 3.1 Pro | 90.9 | 84.5 | 77.6 | 60.3 | 74.6 | 91.1 | 99.7 | 96.6 | 87.9 | 76.7 |
| Claude Opus 4.6 | 88.9 | 81.7 | 63.0 | 46.8 | 68.0 | 80.6 | 100.0 | 94.8 | 82.4 | 82.9 |
| Grok 4.2 | 88.2 | 73.9 | 84.3 | 66.7 | 68.6 | 81.7 | 98.5 | 92.4 | 80.8 | 85.8 |
| DeepSeek R1 | 87.0 | 75.4 | 77.2 | 49.6 | 64.2 | 57.8 | 98.8 | 95.5 | 87.6 | 81.7 |
| Claude Sonnet 4.5 | 86.0 | 78.2 | 62.6 | 51.8 | 66.8 | 60.0 | 85.8 | 94.1 | 90.8 | 88.3 |
| GPT-5.2 | 85.8 | 78.2 | 64.0 | 40.4 | 64.4 | 63.3 | 99.1 | 94.5 | 76.6 | 88.3 |
| GPT-5.4 | 84.0 | 71.8 | 78.2 | 58.9 | 65.8 | 60.6 | 99.4 | 93.1 | 81.1 | 74.2 |
| Llama 4 Scout | 80.8 | 66.9 | 51.9 | 28.4 | 51.2 | 29.4 | 79.4 | 95.9 | 92.6 | 84.6 |
| Qwen3-32B | 78.3 | 47.9 | 78.2 | 46.1 | 49.6 | 54.4 | 86.1 | 93.8 | 74.5 | 72.9 |
| Qwen3-14B | 77.7 | 53.5 | 77.8 | 48.2 | 51.2 | 42.8 | 81.8 | 91.0 | 82.6 | 74.2 |
| GPT-OSS-20B | 75.4 | 48.6 | 87.4 | 59.6 | 51.0 | 41.7 | 78.8 | 87.6 | 78.7 | 76.2 |
| GPT-OSS-120B | 70.6 | 44.4 | 87.2 | 58.2 | 52.2 | 31.7 | 70.0 | 84.1 | 73.9 | 78.8 |
| **Mean** | 84.2 | 69.3 | 73.9 | 51.2 | 62.1 | 61.5 | 91.0 | 93.3 | 83.9 | 81.7 |

Table 9: Cue strength and domain (trial-level accuracy, mean/min/max over the 14 models). Strength is evaluated on the 19 scenarios that have strong, medium, and weak variants of the same cue; accuracy rises monotonically as the cue weakens for 12 of 14 models. Domain differences partly reflect the mix of constraint families in each domain.

| **Subset** | $n$ | **Mean** | **Min** | **Max** |
|---|---|---|---|---|
| *Heuristic-strength variants, 19 matched scenarios* | | | | |
| strong cue | 19 | 70.0 | 57.4 | 83.2 |
| medium cue | 19 | 76.1 | 53.2 | 93.7 |
| weak cue | 19 | 84.3 | 67.9 | 98.9 |
| *Constraint-active base items by domain* | | | | |
| work | 27 | 93.1 | 82.6 | 97.4 |
| digital | 13 | 90.3 | 76.9 | 100.0 |
| home | 26 | 90.2 | 83.5 | 95.8 |
| travel | 9 | 82.8 | 74.4 | 88.9 |
| medical | 12 | 80.0 | 69.2 | 96.7 |
| transportation | 33 | 77.1 | 58.5 | 94.8 |
| shopping | 22 | 76.4 | 46.4 | 90.0 |

**Collection.** API models were sampled 10 times per prompt at temperature 0.7 (DeepSeek R1 and Kimi K2.5 at provider defaults) with up to 4,096 output tokens. Local models (Qwen3-14B, Qwen3-32B, Qwen3.5-27B, GPT-OSS-20B) were sampled at temperature 0.7 and top-$p$ 0.9. During HOB collection, Qwen3.5-27B alone received the system prompt "Answer concisely."; all other models received none. CORE-100 collection uses a uniform empty system prompt.

### B.3 PROMPTING ON HOB

Table 10: The precondition-listing prompt on HOB (all-10-correct accuracy, judge-scored), split by item type. Aggregate gains can come from constraint-removed prompts while constraint-active accuracy falls (Llama 4 Scout).

| | **All 500 prompts (strict)** | | | **Constraint-active (strict)** | | | **Constraint-removed (strict)** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | base | +prompt | Δ | base | +prompt | Δ | base | +prompt | Δ |
| GPT-5.4 | 65.8 | 74.8 | +9.0 | 71.8 | 82.4 | +10.6 | 58.9 | 70.9 | +12.1 |
| Llama 4 Scout | 51.2 | 57.8 | +6.6 | 66.9 | 57.0 | −9.9 | 28.4 | 58.2 | +29.8 |
| Gemini 3.1 Pro | 74.6 | 74.8 | +0.2 | 84.5 | 84.5 | +0.0 | 60.3 | 61.7 | +1.4 |

# C CORE-100 CONSTRUCTION AND VALIDATION

**Construction rules.** Every scenario was rebuilt from HOB's coverage frame under six rules, checked before freezing. (1) A single causal-factor edit separates Active and Removed; the goal, requested output, action labels, and format are untouched. (2) The length difference is at most 2 words and the character similarity at least 0.80. (3) The edit reverses the unique human-preferred action. (4) The SalienceControl clause is matched to the Explicit clause in length, syntax, and attachment point. (5) Cue clauses state prerequisites, never actions. (6) Option order is counterbalanced by a salted hash of the scenario ID (the constraint-heavy action is option A in 43 scenarios and option B in 57).

Table 11: Pair audit: CORE-100 versus HOB's constraint-removed rewrites (100 pairs each, same script). Rating rows are unavailable for HOB.

| | **CORE-100** | **HOB rewrites** |
|---|---|---|
| Length difference in words, mean (max) | 0.25 (2) | 10.2 (23) |
| Character similarity, mean (min) | 0.93 (0.81) | 0.59 (0.17) |
| Token edit distance, mean / median (max) | 2.8 / 2 (7) | 17.8 / 18 (32) |
| Pairs with length difference $\leq 2$ words | 100% | 13% |
| Pairs with character similarity $\geq 0.80$ | 100% | 7% |
| Pairs meeting both criteria | 100% | 4% |
| Unique preferred action ($\geq$4/5 raters), Active and Removed | 100% | – |
| Auditors: only the intended fact changed (2/2) | 100% | – |
| Rater confidence, mean by round | 4.4–4.6 | – |
| Naturalness, mean by round (floor 3.0) | 3.6–3.8 | – |

**Ledger.** The construction ledger tracks 141 candidates: 54 frozen in wave 1, 28 in wave 2, 18 in wave 3; 24 excluded; and 17 that failed the third round and were discarded once the target of 100 was reached. The 24 exclusions comprise 15 scenarios that failed re-validation after their single permitted rewrite, 6 whose constraint proved probabilistic or contestable (including high-stakes medical items), and 3 with structurally unrepairable premises (Table 12). Three frozen scenarios (A1-011, A3-001, B4-004) missed a threshold on one added-clause condition. They are retained for selectivity but excluded from cue contrasts.

Table 12: Excluded candidates and recorded reasons.

| **Scenario** | **Recorded reason** |
|---|---|
| 15 scenarios | Failed validation after one rewrite (preregistered rule): A1-001, A4-002, A4-006, A5-005, B2-005, B2-010, B5-002, B5-007, C2-009, C3-007, C3-008, C4-011, D4-001, D4-003, D4-006. |
| A5-002 | Constraint (post-office passport appointment rules) is not common knowledge; Active raters walk regardless of wording. |
| B2-011 | Implausible premise (swimming a river with belongings); naturalness 2.0–2.8 in all conditions. |
| B4-001 | High-stakes medical scenario with a contestable constraint (many urgent-care clinics treat simple fractures). |
| B5-008 | Heuristic label ("merge and deploy now without testing") is categorically dispreferred; no Removed variant can reverse it. |
| C4-007 | High-stakes mental-health scenario with a probabilistic constraint; no minimal Removed edit reads naturally. |
| C4-009 | Contestable constraint (household bleach is usable for pool chlorination). |
| D4-007 | Contestable conflict-of-interest constraint; raters disprefer dealerships on cost in every framing. |
| D4-010 | Probabilistic constraint (many jewelry stores replace watch batteries); rater confidence 2.8. |
| D4-011 | Heuristic label names the exact service the venue lacks, so no Removed edit can make it preferred. |

**Rounds.** Each round used 20 rating seats; every condition of every rated scenario received 5 ratings. A Latin square (4 blocks $\times$ 5 raters) prevented any rater from seeing two conditions of one scenario. Two auditors per round judged each pair side by side without being asked for an answer. Table 13 gives per-round statistics. Raters were 22 trained annotators who are not authors of this paper (20 rating and 2 audit seats); the same pool served every round. Raters therefore met revised versions of scenarios they had rated before (37 wave-2 repairs and 3 round-3 rewrites), a possible carryover effect on repaired items.

Table 13: Validation rounds. $p_{\mathrm{A}}$/$p_{\mathrm{R}}$: fraction of ratings choosing the constraint-heavy action in Active/Removed; $S = p_{\mathrm{A}} - p_{\mathrm{R}}$. Conf./Nat.: mean confidence and naturalness (1–5).

| Round | Rated | Frozen | $p_{\mathrm{A}}$ | $p_{\mathrm{R}}$ | $S$ | Conf. | Nat. |
|---|---|---|---|---|---|---|---|
| 1 | 100 | 54 | 0.944 | 0.143 | 0.801 | 4.36 | 3.64 |
| 2 | 46 | 28 | 0.983 | 0.031 | 0.952 | 4.50 | 3.60 |
| 3 | 35 | 18 | 1.000 | 0.069 | 0.931 | 4.61 | 3.75 |
| Pooled | | | 0.965 | 0.098 | 0.867 | | |

**Naturalness and amendments.** Mean naturalness by condition (rounds 1/2/3): Active 3.96/3.99/4.18; Explicit 3.69/3.60/3.77; Removed 3.56/3.40/3.29; SalienceControl 3.35/3.40/3.77. The preregistration was frozen on 2026-07-10 with hashed analysis code. Five amendments are logged with reasons. A1 reduced the pair auditors from five to two and required unanimity. A2 replaced the uncalibrated naturalness gate (mean $\geq$4) with a floor of 3.0 plus full distributions, based on rating distributions alone. A3 and A4 record a freeze in three waves rather than two, the scripted recount (54, superseding a preliminary hand count of 59), the shared rater pool, and a 20-scenario design pilot. A5 records two post-freeze revisions to analysis modules that did not change any primary metric. Seventeen pilot scenarios are inside CORE-100. Excluding them leaves every family-F1 gap positive ($+0.036$ to $+0.108$). Restricting to the 54 wave-1 scenarios also leaves every gap positive for the eight main-analysis models evaluated on that wave ($+0.033$ to $+0.114$).

# D MODELS AND COLLECTION SETTINGS

Table 14: CORE-100 models. All receive an empty system prompt, provider-default reasoning settings, unseeded sampling, and at most 4,096 output tokens; each of the 400 prompts is sampled 7 times. Collection window: July 2026.

| Model | Model ID | Access | Temperature |
|---|---|---|---|
| Claude Opus 4.6 | `claude-opus-4-6` | Anthropic API | 0.7 |
| Claude Sonnet 4.5 | `claude-sonnet-4-5-20250929` | Anthropic API | 0.7 |
| DeepSeek R1 | `deepseek-reasoner` | DeepSeek API | default |
| DeepSeek V4 Pro | `deepseek-v4-pro` | DeepSeek API | default |
| Gemini 3.1 Pro | `gemini-3.1-pro-preview` | Google API | 0.7 |
| GPT-5.4 | `gpt-5.4` | OpenAI API | 0.7 |
| GPT-OSS-120B | `openai/gpt-oss-120b` | Groq API | 0.7 |
| Kimi K2.5 | `kimi-k2.5` | Moonshot API | default |
| Llama 4 Scout | `meta-llama/llama-4-scout-17b-16e-instruct` | Groq API | 0.7 |
| Qwen3-32B | `Qwen/Qwen3-32B` | local (HF), top-$p$ 0.9 | 0.7 |

Every stored response records the model identifier, provider, temperature, top-$p$, seed (none), output limit, reasoning mode and budget, finish reason, retries, and UTC access time. Grok 4.2 was in the planned pool but was excluded under the preregistered evidence rule; its archived outputs are released unanalyzed. The `deepseek-reasoner` endpoint served DeepSeek R1 throughout collection; the name was scheduled for deprecation afterward, so later re-collection may reach a different model. HOB evaluates 14 models: GPT-5.4, GPT-5.2, Claude Opus 4.6, Claude Sonnet 4.5, Gemini 3.1 Pro, Grok 4.2, DeepSeek R1, Kimi K2.5, Llama 4 Scout, and GPT-OSS-120B through APIs, and Qwen3-14B, Qwen3-32B, Qwen3.5-27B, and GPT-OSS-20B locally (collection settings in App. B). The case study uses seven open-weight models run locally (§3). Across the paper, 17 distinct models are evaluated.

# E JUDGE STUDIES

**Human agreement with HOB's judge.** HOB's judge (Qwen3-32B, greedy) sees the question, goal, gold and shortcut answers, and constraint description ("none—constraint removed" for constraint-removed items). Two annotators independently re-graded 300 responses drawn from the free-form collection on HOB-derived prompts, blind to the judge. The sample was stratified to 150 judge-correct, 90 judge-incorrect, and 60 judge-ambiguous responses, with 58–62 per constraint family, across all 14 models. On the binary correct-versus-not decision, mean human–judge agreement is 89.7% (Cohen's $\kappa = 0.79$), inter-annotator agreement is 92.0% ($\kappa = 0.84$), and the human consensus (263 items) matches the judge on 93.9%. Agreement is lowest on constraint-removed items (annotators vs. judge: 86.8% and 74.7%, $n = 91$) and higher on the other conditions (89.2–97.0%).

**Shared-metadata vs. condition-neutral rubrics.** A second judge (GPT-5.4, temperature 0, 16 output tokens) graded a stratified subset of 504 CORE-100 records. The subset comprised 12 parser-conforming responses per model and condition (480), plus all 24 nonconforming responses, sampled with a fixed seed. Each record was graded under both rubrics below. Table 15 gives agreement with the parser. The neutral rubric also recovers a verdict for 15 of the 24 nonconforming responses (the metadata rubric, 4).

Table 15: Second-judge agreement with the deterministic parser (120 parser-conforming records per condition).

| Condition | Shared-metadata rubric | Condition-neutral rubric |
|---|---|---|
| Active | 120/120 | 120/120 |
| Explicit | 120/120 | 120/120 |
| SalienceControl | 120/120 | 120/120 |
| Removed | 25/120 | 120/120 |
| Overall | 385/480 | 480/480 |

Listing 1: Shared-metadata rubric (as used for HOB); for CORE-100 Removed records the metadata carried the scenario's Active-form constraint.

```
You are a judge evaluating whether an AI assistant correctly answered a reasoning question.
## Instance info
- Question: {question}
- Goal: {goal}
- Correct answer: {gold_answer}
- Shortcut (wrong) answer: {shortcut_answer}
## Assistant's response
{response}
## Your task
Determine whether the assistant's FINAL recommendation/answer aligns with the CORRECT answer
     ("{gold_answer}") or the SHORTCUT answer ("{shortcut_answer}").
Key: The correct answer requires recognizing a hidden constraint: {hidden_constraint}
Reply with exactly one of: "correct", "incorrect", "ambiguous".
```

Listing 2: Condition-neutral rubric.

```
You are grading a response to a multiple-choice everyday-decision question.
Question:
{question}
The correct recommendation is: {gold_answer}
Response to grade:
{response}
Does the response recommend the correct option ({gold_answer})? Reply with exactly one word:
     correct, incorrect, or ambiguous.
```

# F CORE-100 PER-MODEL RESULTS

Table 16: Scenario-level switch matrices (modal Active choice → modal Removed choice; counts over 100 scenarios).

| **Model** | $C \to H$ (selective) | $C \to C$ (over) | $H \to H$ (under) | $H \to C$ (reversed) | **Unstable** |
|---|---|---|---|---|---|
| DeepSeek V4 Pro | 88 | 12 | 0 | 0 | 0 |
| DeepSeek R1 | 84 | 13 | 3 | 0 | 0 |
| Gemini 3.1 Pro | 82 | 17 | 1 | 0 | 0 |
| Claude Opus 4.6 | 81 | 19 | 0 | 0 | 0 |
| Kimi K2.5 | 80 | 19 | 1 | 0 | 0 |
| Claude Sonnet 4.5 | 71 | 23 | 6 | 0 | 0 |
| GPT-OSS-120B | 66 | 19 | 14 | 0 | 1 |
| Qwen3-32B | 64 | 26 | 10 | 0 | 0 |
| GPT-5.4 | 63 | 29 | 8 | 0 | 0 |
| Llama 4 Scout | 41 | 39 | 19 | 1 | 0 |

Table 17: Per-condition accuracy and cue contrasts (probability of the human-preferred action). EG: Explicit gain (E−A); SE: salience effect (SC−A); CSG: cue-specific gain (E−SC). Cue contrasts use the 97 cue-valid scenarios.

| **Model** | **Active** | **Explicit** | **SalienceCtrl** | **Removed** | **EG** | **SE** | **CSG** |
|---|---|---|---|---|---|---|---|
| Claude Opus 4.6 | 1.000 | 0.990 | 1.000 | 0.816 | −0.010 | +0.000 | −0.010 |
| Claude Sonnet 4.5 | 0.940 | 0.950 | 0.934 | 0.774 | +0.010 | −0.006 | +0.016 |
| DeepSeek R1 | 0.964 | 1.000 | 0.946 | 0.879 | +0.036 | −0.019 | +0.054 |
| DeepSeek V4 Pro | 0.993 | 1.000 | 0.993 | 0.883 | +0.007 | +0.000 | +0.007 |
| Gemini 3.1 Pro | 0.993 | 1.000 | 0.989 | 0.833 | +0.007 | −0.004 | +0.011 |
| GPT-5.4 | 0.916 | 0.970 | 0.904 | 0.694 | +0.054 | −0.011 | +0.066 |
| GPT-OSS-120B | 0.874 | 0.984 | 0.860 | 0.763 | +0.110 | −0.014 | +0.124 |
| Kimi K2.5 | 0.987 | 1.000 | 0.976 | 0.808 | +0.013 | −0.011 | +0.024 |
| Llama 4 Scout | 0.804 | 0.967 | 0.806 | 0.606 | +0.163 | +0.001 | +0.161 |
| Qwen3-32B | 0.901 | 0.986 | 0.879 | 0.751 | +0.084 | −0.023 | +0.107 |

Table 18: Reliability thresholds: fraction of the 400 conditions with at least $k$ of 7 samples correct. The ordering matches SBA. GPT-OSS-120B had 19 and Kimi K2.5 5 samples lost to provider errors.

| **Model** | ≥5/7 | ≥6/7 | 7/7 |
|---|---|---|---|
| DeepSeek V4 Pro | 0.965 | 0.955 | 0.918 |
| Gemini 3.1 Pro | 0.953 | 0.945 | 0.938 |
| Claude Opus 4.6 | 0.950 | 0.950 | 0.950 |
| Kimi K2.5 | 0.940 | 0.930 | 0.904 |
| DeepSeek R1 | 0.938 | 0.923 | 0.898 |
| Claude Sonnet 4.5 | 0.898 | 0.898 | 0.898 |
| GPT-5.4 | 0.870 | 0.863 | 0.853 |
| Qwen3-32B | 0.850 | 0.835 | 0.778 |
| GPT-OSS-120B | 0.846 | 0.810 | 0.769 |
| Llama 4 Scout | 0.795 | 0.790 | 0.788 |

Table 19: Exploratory breakdown by constraint family, pooled over the 10 models (scenario-bootstrap 95% CIs, 10,000 replicates).

| **Constraint family** | $n$ | $p_A$ [95% CI] | $p_R$ [95% CI] | $S$ [95% CI] |
|---|---|---|---|---|
| Presence (C-pres) | 16 | 0.729 [0.65, 0.81] | 0.188 [0.09, 0.30] | 0.540 [0.43, 0.64] |
| Capability (C-cap) | 22 | 0.923 [0.88, 0.96] | 0.206 [0.11, 0.31] | 0.718 [0.62, 0.81] |
| Validity (C-val) | 21 | 0.995 [0.99, 1.00] | 0.215 [0.11, 0.33] | 0.781 [0.67, 0.88] |
| Scope (C-scope) | 24 | 0.996 [0.99, 1.00] | 0.214 [0.10, 0.34] | 0.782 [0.66, 0.89] |
| Procedural (C-proc) | 17 | 0.997 [0.99, 1.00] | 0.280 [0.15, 0.43] | 0.717 [0.56, 0.85] |
| Presence − other families | | −0.248 [−0.33, −0.17] | −0.037 [−0.15, +0.09] | −0.212 [−0.33, −0.09] |

Table 20: Active accuracy $p_{\mathrm{A}}$ by constraint family and model. For eight of ten models presence is the lowest family; the other two are at ceiling throughout.

| Model | Presence | Capability | Validity | Scope | Procedural |
|---|---|---|---|---|---|
| Llama 4 Scout | 0.28 | 0.64 | 1.00 | 1.00 | 1.00 |
| GPT-5.4 | 0.47 | 1.00 | 1.00 | 1.00 | 1.00 |
| GPT-OSS-120B | 0.48 | 0.81 | 1.00 | 0.99 | 1.00 |
| Qwen3-32B | 0.60 | 0.85 | 1.00 | 0.99 | 1.00 |
| Claude Sonnet 4.5 | 0.69 | 1.00 | 0.95 | 1.00 | 1.00 |
| DeepSeek R1 | 0.85 | 0.97 | 1.00 | 1.00 | 0.97 |
| DeepSeek V4 Pro | 0.96 | 1.00 | 1.00 | 1.00 | 1.00 |
| Kimi K2.5 | 0.96 | 0.97 | 1.00 | 1.00 | 1.00 |
| Claude Opus 4.6 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Gemini 3.1 Pro | 1.00 | 1.00 | 1.00 | 0.97 | 1.00 |

# G INTERVENTIONS

**Prompts.** Each intervention prepends one fixed prefix to the question (verbatim below); the baseline uses none.

```
cot: Think step by step about the question before giving your final answer.
precondition_listing: Before answering, list the necessary conditions that must be true for
     the stated goal to be accomplished. Then answer the question.
goal_decomposition: Before answering, decompose the stated goal into the discrete steps or
     resources it requires. Then answer the question.
counterfactual_check (oracle): Before answering, imagine that the most salient resource or
     place in the prompt were absent or unavailable. Would your answer still accomplish the
     stated goal? Then answer the question.
```

Table 21: All intervention arms: Active gain (AG), signed Removed harm (RH; positive means more constraint-heavy answers when the constraint is absent), and selectivity gain (SG), with paired scenario-bootstrap 95% CIs. †: survives Holm correction in family F2.

| Model | Intervention | AG [95% CI] | RH [95% CI] | SG [95% CI] |
|---|---|---|---|---|
| Claude Opus 4.6 | CoT | $+.000$ [.000, .000] | $+.013$ [$-.049$, .076] | $-.013$ [$-.076$, .049] |
| | Precondition list. | $+.000$ [.000, .000] | $-.027$ [$-.097$, .040] | $+.027$ [$-.040$, .097] |
| | Goal decomp. | $-.010$ [$-.030$, .000] | $-.020$ [$-.086$, .043] | $+.010$ [$-.057$, .077] |
| | CF check (oracle) | $-.010$ [$-.030$, .000] | $-.036$ [$-.091$, .016] | $+.026$ [$-.030$, .084] |
| DeepSeek R1 | CoT | $+.017$ [.004, .034] | $+.013$ [$-.007$, .034] | $+.004$ [$-.021$, .030] |
| | Precondition list. | $+.026$ [.006, .051] | $+.036$ [.001, .070] | $-.010$ [$-.051$, .033] |
| | Goal decomp.† | $+.033$ [.010, .061] | $-.009$ [$-.027$, .007] | $+.041$ [.013, .074] |
| | CF check (oracle) | $-.003$ [$-.027$, .024] | $+.308$ [.236, .381] | $-.311$ [$-.385$, $-.238$] |
| GPT-5.4 | CoT | $+.004$ [.000, .011] | $+.003$ [$-.037$, .039] | $+.001$ [$-.034$, .041] |
| | Precondition list. | $+.004$ [.000, .011] | $+.049$ [.011, .091] | $-.044$ [$-.087$, $-.007$] |
| | Goal decomp. | $+.004$ [.000, .011] | $+.019$ [$-.020$, .056] | $-.014$ [$-.053$, .024] |
| | CF check (oracle) | $+.004$ [.000, .011] | $-.004$ [$-.056$, .047] | $+.009$ [$-.043$, .060] |
| GPT-OSS-120B | CoT | $+.004$ [$-.021$, .030] | $-.005$ [$-.027$, .019] | $+.009$ [$-.027$, .042] |
| | Precondition list. | $-.031$ [$-.066$, .000] | $-.013$ [$-.050$, .025] | $-.019$ [$-.068$, .031] |
| | Goal decomp. | $-.054$ [$-.089$, $-.024$] | $+.002$ [$-.028$, .031] | $-.056$ [$-.099$, $-.014$] |
| | CF check (oracle) | $+.003$ [$-.020$, .026] | $+.073$ [.034, .115] | $-.070$ [$-.120$, $-.023$] |

Table 22: Holm family F2: the 12 preregistered selectivity-gain tests, ordered by raw $p$ (scenario bootstrap, 10,000 replicates).

| Model | Intervention | $p$ | $p_{\text{Holm}}$ | Reject |
|---|---|---|---|---|
| DeepSeek R1 | Goal decomp. | .0032 | .0384 | yes |
| GPT-OSS-120B | Goal decomp. | .0102 | .1122 | no |
| GPT-5.4 | Precondition list. | .0180 | .1800 | no |
| Claude Opus 4.6 | Precondition list. | .4418 | 1.000 | no |
| GPT-OSS-120B | Precondition list. | .4480 | 1.000 | no |
| GPT-5.4 | Goal decomp. | .4682 | 1.000 | no |
| GPT-OSS-120B | CoT | .6111 | 1.000 | no |
| DeepSeek R1 | Precondition list. | .6441 | 1.000 | no |
| Claude Opus 4.6 | CoT | .7049 | 1.000 | no |
| DeepSeek R1 | CoT | .7715 | 1.000 | no |
| Claude Opus 4.6 | Goal decomp. | .7879 | 1.000 | no |
| GPT-5.4 | CoT | .9989 | 1.000 | no |

# H EXTENDED LIMITATIONS

*Raters.* CORE-100's raters were a fixed pool of 22 trained annotators, not a large crowd sample. Latin-square blinding held within each round, but the same pool served all three rounds, so raters met revised versions of 40 scenarios they had rated before. The rater reference (Table 1) comes from the same ratings that selected items and is therefore optimistic; a disjoint-rater baseline on the frozen set would be the exact analogue. *Texts.* Scenario texts were drafted and repaired with an LLM (Claude). Two evaluated models share that family; neither tops the paired ranking (SBA ranks 4 and 6). In 150 of the 200 Explicit and SalienceControl prompts the inserted clause begins in lower case (e.g., "…from my apartment. the couch is far too heavy to carry."). The artifact is identical across the two conditions, so it does not bias the cue contrast, but it lowers naturalness.

*Scope.* Golds are validated everyday preferences. Candidates whose constraint proved probabilistic or contestable were excluded (App. C); retained medical-, legal-, financial-, or licensing-adjacent items were not expert-adjudicated. *Scoring.* Forced choice enables deterministic parsing (99.3–100% conformance) but narrows behavior relative to free-form advice. The judge analyses bound, but do not remove, the scoring uncertainty of the excluded responses. *Models and interventions.* API models are versioned snapshots from July 2026, and results may drift; exact identifiers and timestamps ship with the data. Each intervention is a single template on four models, so our results bound the tested prompts rather than prompting in general. Reasoning-budget and other test-time-compute manipulations are not tested under paired scoring here.

# I REPRODUCTION

A single command (`make reproduce`) regenerates every CORE-100 number, table, and figure from the frozen raw outputs in eight idempotent stages: parse, audit, metrics, stats, frontier, figures, tables, and manifest (a SHA-256 manifest of all outputs). A second script regenerates the case-study and HOB figures and tables used here from the raw log-probability sweeps and judged generations. All stochastic steps are seeded (bootstrap: 10,000 replicates, seed 42). The analysis modules were hash-frozen with the preregistration, and each CORE-100 wave was hash-frozen before collection on that wave.