READING IS NOT SIMULATING: CALIBRATING CHAIN-OF-THOUGHT INTERVENTIONS ON AN EXECUTABLE COGNITIVE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Chain-of-thought interventions probe answer dependence on written reasoning. Interpreting that dependence as consequence prediction requires a reference for what an edit should change. We build one from an executable cognitive process: an ACT-R-derived driver runs in a closed-loop simulator, and 1.5B–7B language models are fine-tuned to generate its templated state traces before predicting outcomes. Shared-randomness re-executions supply mechanism-specific counterfactuals; fresh seeds at the same specification supply empirical consequence references. At 1.5B, whole-trace replacement reaches 74% answer-direction agreement on outcome-moving pairs, yet donor replacements without the corresponding mechanism intervention yield comparable or higher agreement against their own outcome changes. Edited states persist in 99–100% of evaluated continuation clauses where reference traces disagree. In the matched-boundary assay, all three students fall below the specification-matched majority for lane-consequence accuracy on the hands mechanism; the 7B student also falls below arm-wise modal prediction and re-execution agreement. Separately, hands answer-direction agreement is 79–85% with consequence-supplied or full-trace inputs, versus 7–52% after state-only continuation. Reference decision rules change which shortfalls are supported, and a layout audit exposes termination leakage. In this setting, we empirically distinguish following supplied traces from predicting consequences of edited states, and provide an executable protocol that makes this distinction auditable through explicit information boundaries, decision rules, and evaluation populations.

1 INTRODUCTION

Chain-of-thought faithfulness studies examine responses to altered reasoning or biased inputs (Lanham et al., 2023; Turpin et al., 2023; Han et al., 2026). An answer change indicates sensitivity to an intervention, but not necessarily accurate prediction of its consequences. A model may follow supplied text without deriving its consequences, or ignore an edit contradicted by the remaining reasoning (Macar et al., 2026). Interpreting these responses requires references for both the edit's consequences and what is predictable from the supplied information. We distinguish answer sensitivity, state propagation, and consequence prediction.

We make these comparisons auditable with an executable process. An ACT-R-derived driver (Anderson et al., 2004; Salvucci, 2006) runs in a closed-loop simulator, and a deterministic verbalizer produces templated state traces. We fine-tune 1.5B–7B language models to generate them before predicting outcomes. Re-executing a seed with one mechanism altered supplies a paired counterfactual; fresh seeds at the same specification supply empirical consequence references. Driving is the testbed, not the contribution: the traces describe a controlled external process, not the language model's latent computation.

These references make four interpretations of intervention evidence separately testable. *Effect size as mechanism evidence:* donor replacements without the corresponding mechanism intervention yield comparable or higher direction agreement than whole-trace replacement in both donor-tested models; each condition is scored against its own outcome change (Section 5.2). *Propagation as*

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.

prediction: high state propagation coexists with lane-consequence accuracy below the paired majority for all three students on hands. Answer-direction agreement on hands, however, is 79–85% with consequence-supplied or full-trace inputs, versus 7–52% after state-only continuation (Section 5.3). *Reproducibility as a bound:* the 3B student exceeds re-execution agreement on peripheral without a supported advantage over the paired majority; its arm-wise comparison depends on tie handling. These references are empirical decision rules, not performance bounds. *Trace content as evidence:* termination metadata carries outcome information before supporting events are narrated (Section 5.1).

Our contributions are the executable evaluation setting, an intervention ladder with donor controls, and a matched information-boundary assay with three empirical references. The resulting protocol (Table 1) specifies the information boundary, decision rule, and evaluation population needed to support each claim.

2 Related work

Existing work asks whether reasoning is faithful; we ask what the intervention used to test it identifies. Behavioural faithfulness tests go back to Lanham et al. (2023), who truncated and corrupted reasoning, and Turpin et al. (2023), who biased the input and showed the reasoning did not mention it; Han et al. (2026) define faithfulness as stance consistency plus causal influence under an inserted premise, Shen et al. (2026) pose instance-level detection as classification, and Macar et al. (2026) show that resampling the edited sentence from the model's own distribution produces larger and more stable effects than handwritten edits. Each supplies a behavioural reference; none an executable process that specifies the counterfactual consequences of the mechanism described by an edit, which is what our reference computes rather than assumes. Our edited sentences are templated rather than resampled, and our continuations are scored against a trace the model did not generate.

A second line asks not whether the reasoning is causally load-bearing but whether it *reports* what shaped the answer (Gema et al., 2026; Duzan & Cooper Stickland, 2026; Korbak et al., 2025), and is explicit that verbalization is distinct from causal dependence. Ours sits on the causal side of that distinction; the offset of Section 5.3, a state word that moves the answer while no consequence is written, is one concrete way the two come apart. Input-side interventions (Turpin et al., 2023; Chen et al., 2025) use the known bias as the reference for what the reasoning should have mentioned, not for what an edit to it should have moved.

Recent work connects cognitive modeling and language models through architecture-derived representations (Wu et al., 2024), behavioral fine-tuning (Binz et al., 2025), and LLM-based simulation of human behavior (Hu et al., 2026); driving work has generated reasoning labels from simulator state (Wang et al., 2024). We ask the model to narrate the architecture's state and use re-executability to test whether the narration is load-bearing.

3 A setting with an executable reference

3.1 The reference process

The reference is a closed-loop driver-distraction testbed built for a companion study (Figure 5, Appendix C). A scenario engine compiles a distraction specification into a gaze schedule: target, onset, duration, and kinematics (a head-turning glance, an eyes-only glance that retains peripheral lane information, or a body lean that takes a hand from the wheel). An ACT-R-derived driver (Anderson et al., 2004; Salvucci, 2006) executes it in CARLA (Dosovitskiy et al., 2017) on a 50 ms cycle: two-point steering (Salvucci & Gray, 2004) and headway control receive perceptual updates only while gaze is on the road, held commands and calibrated steering noise accumulate during a glance, and hazards are detected through looming and brake lights with reaction-time draws from Green (2000). Every parameter is frozen with a hash (Appendix C). A run places the driver in one of three hazard contexts, a braking lead, an adjacent-lane cut-in, or a trafficless curve, and scores the outcome on a five-rung severity ladder from a wheel over the lane edge (1) to a collision (5), with 0 for an uneventful run.

Two properties matter for everything that follows. The state is fully observable on every cycle: gaze, the age of each perceptual channel, held commands, drift, hand position, peripheral gain, re-acquisition penalty, and hazard stage. And the process is re-executable under common random numbers: the same seed can be run again with exactly one named mechanism altered and every shared random input held fixed, which isolates the targeted mechanism; residual simulator variability, at the level at which the physics engine is not bit-reproducible, is reported separately (Appendix B). Five mechanisms are our interventions (peripheral gain off, hand kept on the wheel, re-acquisition cost zeroed, brake-light cue suppressed, glance transition times swapped), and a sixth is a consequence-level perturbation, a 0.45 m increase in the initial lateral offset. The same harness also re-executes a run at its exact specification from a fresh driver seed, so that only the seed-dependent initial-state and driver draws change.

3.2 THE VERBALIZER

A deterministic template turns the per-cycle log into text: within 30 s chunks a line opens at every change of gaze state, hazard stage, hand position, or peripheral availability, with a lateral-offset band and a headway band at chunk boundaries and gaze returns, and compressed on-road lines retain the event times of their individual clauses. Five span families (gaze, hazard chain, hands, peripheral availability, re-acquisition) are recoverable by a reverse parser with F1 of 1.000 on 1,875 held-out traces, and no line contains an outcome word. An original layout exposed span-end times and the run's duration; the revised layout used for all main results records span starts only and closes the trace without stating the run's termination time. Section 5.1 audits the information removed by this revision (Appendix D).

3.3 ARMS, DATA, AND THREE REFERENCES

Three fine-tuned arms share identical example sets. The input is the specification alone, a five-line block naming context, severity cell, speed, glance schedule, hazard onset, and warning latency, with no seed or simulated quantity. The DIRECT arm writes the answer; the TRACE arm writes the revised-layout trace as its reasoning and then the answer; the FREE arm writes a four-to-eight-sentence rationale from a 7B teacher (Qwen Team, 2025) that saw the label but never the trace, so it is post-hoc by construction, and FREE-LONG matches the TRACE arm's token budget. Base models are Qwen2.5-1.5B-Instruct, Llama-3.2-3B-Instruct (Grattafiori et al., 2024), and Qwen2.5-7B-Instruct, each with LoRA (Hu et al., 2022), rank 16, three epochs, completion-only loss, identical example ids and seed. At 7B only the TRACE arm is trained; the comparisons between training arms are restricted to the two smaller bases. Data come from 58,080 runs over 1,452 cells (Appendix E); arms train on 8,000 examples and are tested on 1,000 in distribution and 999 out of distribution from a withheld severity cell, a withheld glance duration, and a low-speed protocol. Three quarters of every split is rung 0, so macro-F1 accompanies accuracy (Appendix E).

Three references come from the process rather than any model. A predictor that sees only the specification is compared with the leave-one-out majority rung of its specification group,

$$C_{\text{spec}} \;=\; \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} \mathbb{1}\big[\,\text{maj}\{\,y_j : g(x_j)=g(x_i),\; j\neq i\,\}=y_i\,\big], \tag{1}$$

where $\mathcal{D}$ is the evaluation subset with a nonempty leave-one-out comparison group and $g(x)$ is the cell crossed with the onset binned to 0.25 s; this empirical specification-only baseline is 0.809 in distribution and 0.750 out of it. Reconstruction of span family f is compared with how well another realization of the same specification reproduces it, $\bar{F}_f = \mathbb{E}_{i,\,j\sim g(x_i),\,j\neq i}[\mathrm{F1}_f(\tau_j,\tau_i)]$, which is 0.97 to 1.00 for the families the specification determines and 0.747 and 0.686 for the seed-dependent hazard chain. The third reference is built for the assay's own pairs as three decision rules on the same re-executions. Let $c_p \in \{\text{decrease}, \text{invariant}, \text{increase}\}$ be the change of the lane-offset band at the first gaze return, factual to counterfactual, for pair p, and $c_p^{(k)}$ the same quantity when the pair's exact specification is re-executed under the mechanism from driver seed k, over the $K_p \le 20$ re-executions that yield a band there (median 18, minimum 4):

$$\mathrm{B}_1^* \;=\; \frac{1}{K_p}\sum_{k=1}^{K_p}\mathbb{1}\big[c_p^{(k)}=c_p\big], \qquad \mathrm{B}_2^* \;=\; \mathbb{1}\Big[\,\text{maj}_k\; c_p^{(k)}=c_p\Big], \tag{2}$$

averaged over pairs. B_1^* is the rate at which one more draw from the process agrees with the reference class, what a reproducibility check measures; B_2^* is the accuracy of predicting the ensemble's modal class. Over an independent draw of the reference class with class probabilities p_k, the two converge to $\sum_k p_k^2$ and $\max_k p_k$, so a reproducibility agreement is not a bound on what a predictor can attain; the observed values are finite-sample rules, not optima (Appendix B). A third rule mirrors the structure of the students' procedure, which generates each arm separately and differences the two bands: B_Δ takes the modal factual band and the modal counterfactual band over the same re-executions and predicts the sign of their difference, read as invariant when either mode is tied. Because $\mathrm{sgn}[\mathrm{mode}\, B^c - \mathrm{mode}\, B] \neq \mathrm{mode}[\mathrm{sgn}(B^c - B)]$ in general, B_Δ is the arm-wise modal reference against which a separately generating model is fairly read; it matches the arm-wise structure, not the token-level greedy decoder itself, and its tie handling has no counterpart in the student. All three are reported on each student's own scored pairs.

4 Interventions and what they are scored against

Every intervention acts on the TRACE arm's reasoning and is scored against the pair of outcomes the architecture produced. Let y be the factual outcome of specification x, y^c the outcome of the same seed re-executed with mechanism c altered, $\tilde{\tau}$ the reasoning the model writes, and $\hat{y}'$ its answer after a modified reasoning $\tilde{\tau}'$. **EDIT**: the first line on which c appears is rewritten with the verbalizer's own template for the altered mechanism, the rest is left as written, and the model produces only the answer. **EDIT+CONTINUE**: the same line is rewritten or the mechanism's statement deleted, the reasoning is cut there, and the model generates the rest and the answer; this is template-based editing followed by model continuation, which lets the edit propagate as Macar et al. (2026) argue an intervention must, without resampling the edited sentence from the model as their on-policy design does. **CONSEQUENCE**: the reasoning is replaced by the true trace of the initial-offset mechanism, which changes the narrated lateral band but no mechanism sentence. **REPLACE**: the reasoning is replaced by the true counterfactual trace τ^c. Two controls replace it with a trace with no corresponding mechanism intervention, from another seed in the same cell (SAME-CELL DONOR) or from a different cell (OTHER-CELL DONOR), scored against the donor's own rung.

A fifth design, the matched assay of Section 5.3, differs from all of these in that it never mixes text from two runs: *both* members of a pair receive their own true prefix, cut at the same information boundary. STATE-ONLY ends at the first line on which the mechanism state is visibly edited; CONSEQUENCE-SUPPLIED extends each prefix through the lane-band sentence at the first gaze return; FULL-TRACE supplies the whole trace. From STATE-ONLY the model either answers directly or continues the trace before answering.

Define $\Delta y_p = y_p^c - y_p$, $\Delta\hat{y}_p = \hat{y}'_p - \hat{y}_p$, moving pairs $\mathcal{M} = \{p : \Delta y_p \neq 0\}$, and invariant pairs $\mathcal{S} = \{p : \Delta y_p = 0\}$. We report

$$\mathrm{DA} = \frac{\sum_{p\in\mathcal{M}} \mathbb{1}[\mathrm{sgn}\,\Delta\hat{y}_p = \mathrm{sgn}\,\Delta y_p]}{|\mathcal{M}|}, \qquad \mathrm{INV} = \frac{\sum_{p\in\mathcal{S}} \mathbb{1}[\Delta\hat{y}_p = 0]}{|\mathcal{S}|}. \tag{3}$$

Never changing gives zero DA and perfect INV; always changing gives zero INV, with DA determined by direction. Generated lane bands are scored analogously, plus joint three-class accuracy against c_p. Clause propagation is defined in Appendix F. Pair bootstraps use 1,000 resamples for the ladder and 10,000 for the assay; reference comparisons also use cell-cluster intervals.

The consequence assay measures predictive agreement with the reference process, not faithfulness to the student's internal computation. Table 1 states what each comparison supports.

5 Results

5.1 Auditing the trace: a leaked field

In the original layout, the header, segment end, and final timestamp all stated run duration, an outcome-dependent termination signal. A nearest-neighbour rule on that field scores 0.857/0.832 (in/out of distribution), above the specification-only baseline and every arm. Retiming only that field to its counterfactual value yields 0.289 direction agreement on moving pairs, versus 0.655 for

Table 1: Calibration protocol: the reference and comparison required for each supported conclusion.

Design	Reference and comparison	Supported conclusion
Fixed-suffix edit	Paired outcomes; DA and INV.	Sentence sensitivity; a contradicting suffix limits interpretation of a near-zero effect.
Edit + continue	Clause-aligned counterfactual trace.	Preservation of edited state in generated clauses.
Whole-trace replacement	Counterfactual and donor outcomes; DA.	Comparable donor effects mean replacement magnitude alone is not mechanism-specific evidence.
Matched continuation	Same-specification seeds on each student's scored pairs.	Separate accuracy comparisons with B_1^*, B_2^*, and the arm-wise modal reference B_Δ; cell-cluster intervals.
Boundary answer probe	Predicted and reference rung offsets.	Answer sensitivity with or without written consequences, not internal computation.
Selected evaluation	Identically selected reference; held-out accuracy and proper score.	Gain in the matched population; disappearance motivates a selection audit.

Table 2: Held-out accuracy and trace fidelity of the students.

		Accuracy		Macro-F1		Hazard start-F1	
Arm	Model	ID	OOD	ID	OOD	ID	OOD
direct	Qwen2.5-1.5B	0.826	0.780	0.316	0.277	–	–
direct	Llama-3.2-3B	0.830	0.787	0.324	0.312	–	–
trace (revised layout)	Qwen2.5-1.5B	0.790	0.716	0.311	0.365	0.773	0.701
trace (revised layout)	Llama-3.2-3B	0.787	0.710	0.348	0.331	0.795	0.740
trace (revised layout)	Qwen2.5-7B	0.800	0.723	0.358	0.358	0.796	0.751
specification-only baseline (leave-one-out)		0.809	0.750	–	–	0.747	0.686

95% CIs, balanced accuracy and parse-failure rates are in the supplementary material. The 7B row uses a 2,400-token budget on the full test sets. Model rows use the full ID/OOD test sets ($n = 1{,}000/999$); C_{spec} uses their non-singleton subsets ($n = 887/845$) and is not a same-sample model comparison.

counterfactual words alone (Appendix D). Removing future-termination fields and retraining costs two accuracy points (Table 2), recovered at 24,000 examples (Appendix D). All subsequent results use the revised layout. Termination metadata can reveal outcome information before the supporting events are narrated.

5.2 Large intervention effects without the targeted mechanism

Table 3 and Figure 1 place the four interventions and the two controls on one axis. The mechanism-sentence EDIT moves the prediction the simulator's way on 0.016 of moving pairs for the peripheral mechanism and 0.033 for the hand mechanism, with invariance above 0.94. EDIT+CONTINUE raises those to 0.302 and 0.100 when the sentence is rewritten, and higher when it is deleted. CONSEQUENCE reads 0.481 over 243 moving pairs; its band-only and all-but-band channels read 0.412 and 0.136, consistent with additivity (Appendix F). REPLACE reads 0.735 over 529 moving pairs with invariance 0.938.

The controls change that reading: SAME-CELL DONOR yields 0.684 on 1083 moving pairs and OTHER-CELL DONOR 0.821 on 1556, with invariance 0.914 and 0.870; restricted to the consequence mechanism the same-cell control reaches 0.620, above every isolated channel. Comparable effects thus occur without the corresponding mechanism intervention. At 3B/7B, REPLACE reaches 0.726/0.784, while CONSEQUENCE remains near one half. Donors were tested only at 1.5B and 7B; OTHER-CELL DONOR exceeds REPLACE in both (0.830 at 7B; Appendix G).

Table 3: Intervention ladder with controls: what each intervention design can measure.

			Direction agreement				Invariance			
Condition	Mech.	n_{mov}	1.5B	3B	7B	n_{inv}	1.5B	3B	7B	Measures
edit	periph.	63	0.016	–	–	57	0.947	–	–	sens.
edit	hands	60	0.033	–	–	60	0.950	–	–	sens.
edit+continue, rewritten	periph.	63	0.302	–	–	57	0.842	–	–	prop.
edit+continue, rewritten	hands	60	0.100	–	–	60	0.933	–	–	prop.
edit+continue, deleted	periph.	63	0.397	–	–	56	0.732	–	–	prop.
edit+continue, deleted	hands	60	0.183	–	–	60	0.717	–	–	prop.
consequence	offset	243	0.481	0.465†	0.494	1339	0.899	0.910†	0.886	cons.
consequence, band only	offset	243	0.412	–	–	1339	0.926	–	–	cons.
consequence, all but band	offset	243	0.136	–	–	1339	0.953	–	–	cons.
replace	all	529	0.735	0.726	0.784	1819	0.938	0.951	0.934	read.
same-cell donor	all	1083	0.684	–	0.700	1265	0.914	–	0.911	read.
other-cell donor	all	1556	0.821	–	0.830	792	0.870	–	0.886	read.

95% pair-bootstrap intervals in the supplement. †On the original 298-pair offset grid ($n_{\text{mov}} = 43$, $n_{\text{inv}} = 255$). Mech.: periph. = peripheral gain off, hands = hands off wheel, offset = initial offset, all = all mechanisms. Measures: sens. = sentence sensitivity, prop. = propagation, cons. = consequence, read. = reading.

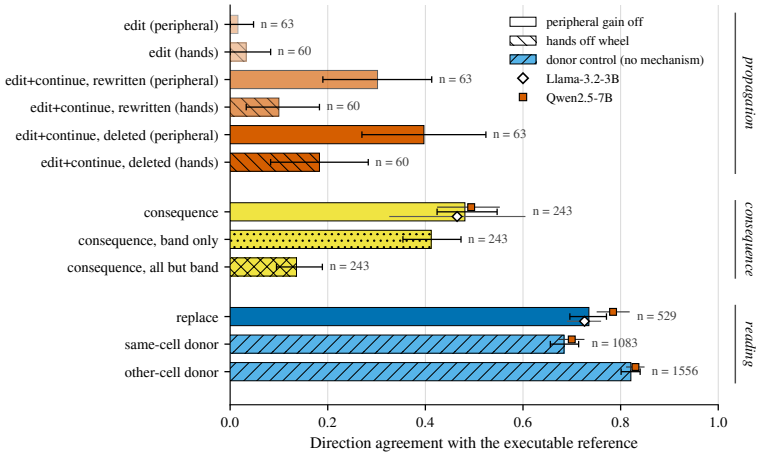


Figure 1: Intervention ladder on the revised-layout arm, with bootstrap intervals. Donor replacements without the corresponding mechanism intervention match or exceed whole-trace replacement, limiting the specificity of its effect size.

5.3 State propagation without demonstrated predictive gains

In EDIT+CONTINUE, rewritten hand and peripheral states persist in 0.99–1.00 of evaluated clauses where the true traces disagree; this does not establish consequence prediction, since unaligned band matches compare different checkpoints (Appendix F).

The assay. We take the pairs of two mechanisms, peripheral gain and hands on wheel, keep those with a clean state boundary and a shared first gaze return within ± 0.1 s (233 and 198 pairs; Appendix A), and cut both members at the same boundary. On the STATE-ONLY continuation we score state incorporation (the mechanism clause in a subsequently generated off-road phase), consequence prediction (the generated lane-band change at the aligned first return, as three-way agreement with c_p, split into direction agreement on moving references and invariance on stable ones), and answer calibration (Equation 3 on the maximum rung under all four boundaries). Primary consequence scores are conditional on both continuations incorporating the state and yielding an aligned band; nominal counts are reported separately (Appendix A). The references are B_1^*, B_2^*, and the arm-wise

Table 4: Matched information-boundary assay by mechanism: generated lane consequence against the specification-matched references.

Size	Joint acc. (n)	B1*	B2*	$B_{\Delta mode}$	M − B1* [cell CI]	M − B2* [cell CI]	M − $B_{\Delta mode}$ [cell CI]	Flags
Peripheral gain off (233 pairs)								
1.5B	0.753 (93)	0.710	0.806	0.688	+0.043 [−0.031, 0.098]	−0.054 [−0.138, 0.009]	+0.065 [−0.047, 0.173]	
3B	0.762 (202)	0.686	0.767	0.658	+0.077 [0.047, 0.102]	−0.005 [−0.023, 0.011]	+0.104 [0.040, 0.172]	>B_Δ
7B	0.707 (222)	0.691	0.784	0.658	+0.016 [−0.055, 0.074]	−0.077 [−0.149, −0.017]	+0.050 [−0.012, 0.114]	
Hands off wheel (198 pairs)								
1.5B	0.456 (125)	0.482	0.584	0.488	−0.026 [−0.127, 0.063]	−0.128 [−0.236, −0.037]	−0.032 [−0.153, 0.069]	
3B	0.381 (126)	0.482	0.587	0.484	−0.101 [−0.205, 0.007]	−0.206 [−0.315, −0.099]	−0.103 [−0.214, 0.008]	
7B	0.288 (125)	0.484	0.592	0.488	−0.196 [−0.287, −0.099]	−0.304 [−0.440, −0.167]	−0.200 [−0.313, −0.086]	<B1* <B_Δ

B1*: agreement of 20 specification-matched re-executions with the reference class; B2*: their majority (mode of the per-seed differences); $B_{\Delta mode}$: sign of the difference between the two arms' modal bands over the same re-executions (a tied arm mode counts as invariant; B_Δ in the flags); all on each student's own scored pairs; M: the student's joint accuracy. State incorporation, answer calibration and LOO robustness in Table 5.

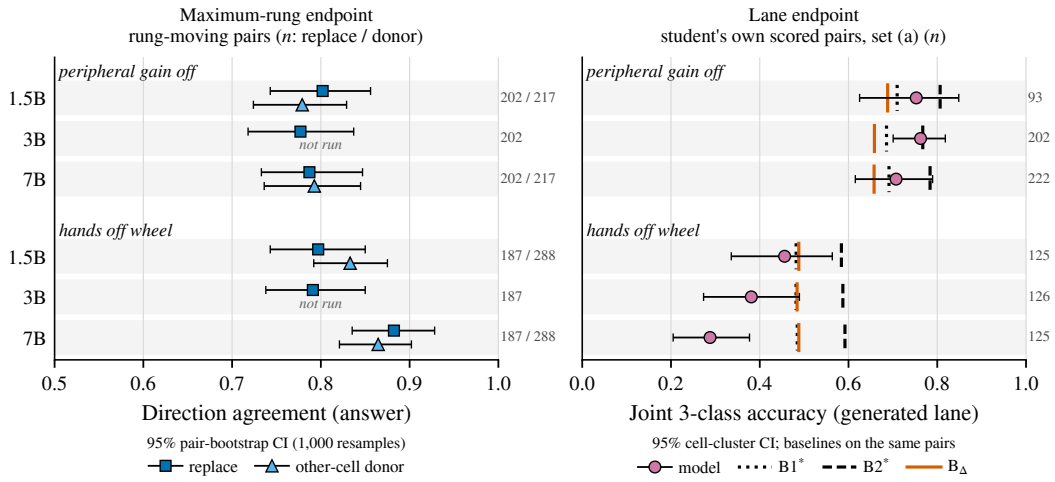


Figure 2: Distinct endpoints. Left: maximum-rung DA for REPLACE and the available donor controls. Right: joint lane-class accuracy with cell-cluster intervals and B_1^*, B_2^*, and B_Δ on each student's scored pairs. Outcomes and denominators differ across panels. No student shows a supported advantage over B_2^*; hands 7B falls below B_1^*.

modal reference B_Δ on each student's own scored pairs; a majority conditioned on everything the STATE-ONLY prefix reveals coincides with the specification-level one pair for pair (Appendix B). The interpretation cases were fixed before inference; the reference was revised afterwards (Appendix A).

One-sided references. Peripheral increases the lane band on 137 of 140 moving pairs and hands decreases it on 99 of 117; always moving that way scores 0.979/0.846 with zero invariance, so DA is read beside INV and joint accuracy beside the references.

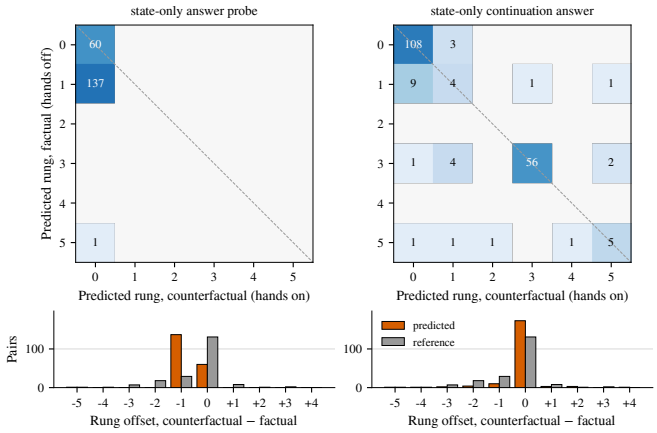


Figure 3: State-keyed answer offsets on hands 7B. Factual versus counterfactual rung predictions and offset distributions for the STATE-ONLY probe (left) and continuation (right). The probe's fixed -1 offset at glance durations of at least 2.5 s is absent from the continuation pattern.

Near-perfect propagation does not guarantee accurate consequences. Table 4 and Figure 2 give the comparison on identical samples. Incorporation (Table 5) is near complete for the 3B and 7B students on both mechanisms and for the 1.5B student on hands; its 0.403 on peripheral is a timing failure (87 counterfactual continuations write non-monotone timestamps; where a phase is written it incorporates on 0.653), and on hands the incorporated clause is the corpus default, so the near-perfect rate there is weak evidence. The written consequence is a different matter. On hands all three students are below the specification-matched majority (joint accuracy 0.456, 0.381, and 0.288 against B_2^* of 0.58 to 0.59; differences -0.13, -0.21, and -0.30 with cell-cluster intervals excluding zero). Part of that gap is the decision rule: the arm-wise modal reference B_Δ reads 0.48 to 0.49 on the same pairs, about 0.10 below B_2^* when tied modes are read as invariant and at most 0.04 below when a tie is broken toward the milder band, so most of the gap between the two rules is tie handling on a finite ensemble rather than a structural cost of arm-wise prediction. Against B_Δ the 7B student remains below (-0.200 $[-0.313, -0.086]$), while the 1.5B and 3B students are not distinguishable from it (-0.032 $[-0.153, +0.069]$, -0.103 $[-0.214, +0.008]$), which does not establish equivalence; other tie rules and return-matched ensembles are in Appendix B. The 7B student is also below the agreement of a single re-execution (-0.196 $[-0.287, -0.099]$). On peripheral no student is above the majority; the 3B student is above the single-re-execution agreement ($+0.077$ $[+0.047, +0.102]$) and not above the majority (-0.005 $[-0.023, +0.011]$), and its position above B_Δ ($+0.104$) is not robust to the tie rule. These are statements about predictive performance on observable text; on hands the 3B and 7B students shift toward never moving the band, and nearly half of their failures to move write a band more severe than either reference for both members (Appendix A). Separately, hands answer-direction agreement is 0.791–0.851 with CONSEQUENCE-SUPPLIED or FULL-TRACE inputs, versus 0.075–0.522 after STATE-ONLY continuation (Table 5). This contrast concerns answer responses under different information boundaries, not the cause of the lane-prediction shortfall.

The reference changes the interpretation. On the same scored pairs, re-execution agreement identifies a shortfall only for hands 7B, whereas the paired majority places all three hands students below. The arm-wise modal reference keeps 7B below, while the smaller students' comparisons depend on tie handling and ensemble selection. Re-execution agreement is not a performance bound. The paired majority is a valid predictive benchmark, but its gap alone does not isolate consequence-modeling error from differences in output decision rules; our initial interpretation did not make this distinction.

Answer sensitivity without a written consequence. The STATE-ONLY answer probe asks what the model does with an edited state and nothing after it; on peripheral no student moves its answer the reference's way on any of the 72 moving pairs. On hands the 7B student moves it on 0.746 of 67 with invariance of only 0.412, and Figure 3 shows why: rung 0 for every hands-on member and rung

1 for every hands-off member whose glance lasts 2.5 s or more, a fixed offset of -1 on 137 pairs where the reference offsets run from -5 to $+3$. Let the same student continue before answering and the offset disappears (173 of 198 pairs keep their rung). The 3B student carries the same offset on 52 of 70 changes, keyed to the 1.5 s glances (Appendix A); the 1.5B student has none. The behaviour is consistent with an input-keyed offset from a state word to a rung, says nothing about the computation behind it, and is answer sensitivity observed with no consequence written at all.

A later boundary. A gain over finite-sample B_2^* would not alone demonstrate seed-information use, so we tested a braking-time boundary on the rung endpoint, after seed-dependent timing appears and before outcomes. The apparent conditioning gain did not survive selection matching, only seven of 185 retained references moved, and no student was run there (Appendix B).

5.4 Parity, reconstruction, and where the errors are

At 8,000 examples, DIRECT exceeds TRACE accuracy by 3.6/6.4 points (in/out of distribution) at 1.5B and 4.3/7.7 at 3B. Trace students lie within two points of each other, and their full-test accuracies are numerically below C_{spec}, which is computed on the non-singleton subset of each split (887 and 845 examples) and is therefore a reference rather than a same-sample comparison (Table 2, Appendix E). At the 0.10 s parser tolerance, trace supervision supports hazard-chain reconstruction against its reproducibility reference, a margin that reverses at 0.20 s (Appendix G); out-of-distribution errors concentrate in lateral bands. Appendix G gives the detailed reconstruction scores, layout ablation, model-family replication, and negative results on monitorability and duration generalization.

6 Discussion and limitations

Our results distinguish responses to supplied traces from prediction of an edited state's consequences. Donor controls show why replacement magnitude alone is not mechanism-specific evidence; the hands assay separates state incorporation from lane-consequence accuracy. Table 1 specifies the information boundaries, decision rules, and evaluation populations needed to interpret these measurements. Termination leakage and selection mismatch illustrate why these requirements matter. Re-execution agreement is not a performance bound, and our initial paired-majority reading overstated what that comparison isolates. Executability makes these distinctions auditable rather than automatically resolving them.

The evidence covers one testbed, templated traces, and 1.5B–7B supervised fine-tunes. It supports conclusions about observable text, not the students' internal computation or failure rates elsewhere. Donors cover two students, edits are templated, the FREE teacher saw labels, and the assay covers two mechanisms with finite-sample references. The hands incorporation clause is the corpus default (Appendix I).

7 Conclusion

We introduced an executable setting for calibrating chain-of-thought interventions. Donor replacements without the corresponding mechanism intervention yield direction agreement comparable to whole-trace replacement. High state propagation coexists with lane-consequence accuracy below the specification-matched majority for all three students on hands; the 7B student also falls below an arm-wise modal reference and re-execution agreement on its scored pairs. The protocol specifies information boundaries, decision rules, and evaluation populations rather than diagnosing internal computation. Executability alone does not determine the appropriate reference. In this setting, we empirically distinguish following supplied traces from predicting consequences of edited states, and make that distinction auditable through executable references.

AI USE STATEMENT

The research question, core methodological idea, experimental design, mathematical formulation, and interpretation of results were developed by the authors. AI tools were used in supporting roles for writing, coding, and feedback. Claude and ChatGPT assisted with editing and consistency checks, while a Claude-based coding agent helped implement and run experiments and analyses under the authors' specifications. All outputs were reviewed and verified by the authors. A Qwen2.5-7B-Instruct teacher model was used to generate rationales and counterfactual rewrites for the FREE and FREE-LONG arms of the synthetic dataset (Section 3, Appendix H). AI tools also provided feedback on draft methodology and interpretation; all methodological decisions remained with the authors. Generative AI was not used to originate the research question, core methodological contribution, or mathematical claims. The authors take responsibility for the final content.

ETHICS STATEMENT

The work involves no human subjects and no personal data. The simulated driver is a cognitive model calibrated on published aggregate data, and the companion driving study whose testbed we reuse is cited for its parameters only. The findings concern the evaluation of language-model reasoning and carry no direct application to driving safety systems; we caution against reading the arms' accuracy as a claim about drivers. Counterfactual interventions on reasoning could in principle be used to manipulate a model's outputs; our use is confined to measurement on fine-tuned research models, and the released code performs no such manipulation on deployed systems.

REPRODUCIBILITY STATEMENT

Every number in the paper is produced from the released result files by the released scripts and audited against the manuscript. Appendix C gives the driver model's parameters, the frozen hash, and the definition of each counterfactual mechanism; Appendix D the verbalizer's segmentation rules, the layout difference, and the reverse parser's gate (complete listings are in the accompanying archive); Appendix E the campaign configuration, validity filters, holdout construction, class balance, the training recipe for all three bases, and the computation of the three references; Appendix D the leak audit; Appendix G the sweep of every analysis constant and the model-family replication; Appendix F the channel decomposition and the line-by-line continuation analyses; Appendix A the matched assay's cohorts, scoring, and complete results; Appendix B the re-execution references and the seed-informed boundary; Appendix H the prompt components, with full listings in the archive. An anonymized archive of code, cohorts, re-executions, result files, and the complete trace and prompt listings is provided as supplementary material; adapters and raw generations are provided by an anonymized link given in the supplementary README.

REFERENCES

John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological Review*, 111(4):1036–1060, 2004.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, et al. A foundation model to predict and capture human cognition. *Nature*, 644, 2025.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Conference on Robot Learning*, 2017.

Agatha Duzan and Asa Cooper Stickland. Chain-of-thought monitoring can be unreliable in implicit-influence settings. *arXiv preprint arXiv:2608.04735*, 2026.

Aryo Pradipta Gema, Neel Rajani, Rohit Saxena, Wai-Chung Kwan, and Pasquale Minervini. Chain-of-thought faithfulness of reasoning models varies with where and how preference cues are delivered. *arXiv preprint arXiv:2608.29464*, 2026.

Hans Godthelp, Paul Milgram, and Gerard J. Blaauw. The development of a time-related measure to describe driving strategy. *Human Factors*, 26(3):257–268, 1984.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Marc Green. "how long does it take to stop?" methodological analysis of driver perception-brake times. *Transportation Human Factors*, 2(3):195–216, 2000.

Yunseok Han, Yejoon Lee, and Jaeyoung Do. RFEval: Benchmarking reasoning faithfulness under counterfactual reasoning intervention in large reasoning models. In *International Conference on Learning Representations*, 2026.

Errol R. Hoffmann and Rudolf G. Mortimer. Drivers' estimates of time to collision. *Accident Analysis and Prevention*, 26(4):511–520, 1994.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. SimBench: Benchmarking the ability of large language models to simulate human behaviors. In *International Conference on Learning Representations*, 2026.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, et al. Chain of thought monitorability: A new and fragile opportunity for AI safety. *arXiv preprint arXiv:2507.11473*, 2025.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Uzay Macar, Paul C. Bogdan, Senthooran Rajamanoharan, and Neel Nanda. Thought branches: Interpreting LLM reasoning requires resampling. In *International Conference on Learning Representations*, 2026.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Dario D. Salvucci. Modeling driver behavior in a cognitive architecture. *Human Factors*, 48(2): 362–380, 2006.

Dario D. Salvucci and Rob Gray. A two-point visual control model of steering. *Perception*, 33(10): 1233–1248, 2004.

Dario D. Salvucci and Niels A. Taatgen. Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review*, 115(1):101–130, 2008.

Xu Shen, Song Wang, Zhen Tan, Laura Yao, Xinyu Zhao, Kaidi Xu, Xin Wang, and Tianlong Chen. FaithCoT-Bench: Benchmarking instance-level faithfulness of chain-of-thought reasoning. In *International Conference on Learning Representations*, 2026.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, 2023.

Tianqi Wang, Enze Xie, Ruihang Chu, Zhenguo Li, and Ping Luo. DriveCoT: Integrating chain-of-thought reasoning with end-to-end driving. *arXiv preprint arXiv:2403.16996*, 2024.

Siyu Wu, Alessandro Oltramari, Jonathan Francis, C. Lee Giles, and Frank E. Ritter. Cognitive LLMs: Towards integrating cognitive architectures and large language models for manufacturing decision-making. *arXiv preprint arXiv:2408.09176*, 2024.

Appendix guide. Sections A–B contain the matched assay, its denominators, the re-execution references, and the selection audit. Sections C–E document the process, verbalizer, data, and training. Sections F–G give intervention and robustness checks; Sections H–I contain examples, prompts, and scope. The accompanying archive holds the complete trace listings and the complete prompt listings.

A MATCHED ASSAY: COHORTS, SCORING, AND COMPLETE RESULTS

Eligibility and information boundaries. The peripheral and hands cohorts use the same eligibility rules. The STATE-ONLY prefixes differ on exactly one line, only in prespecified intervention-linked spans: the near-channel state for peripheral (*lane edges in peripheral view* versus *lane current*), and the hands clause with the near- and far-channel state clauses for hands. All remaining prefix text matches, with no lane band beyond the initial one before the cut. The two arms must share a first gaze return within ± 0.1 s. The CONSEQUENCE-SUPPLIED prefixes match after masking the recurring near-channel state clauses and every lane-band sentence; no lane sentence intervenes between the cut and the return. The input must fit 576 chat-encoded tokens. The validation suite masks exactly these spans before asserting equality, checks every prefix hash and checkpoint alignment, asserts the absence of any outcome, termination, or later-event material in a STATE-ONLY prefix, asserts the 233 peripheral pairs, and pins the 198 hands pairs by content hash; all tests pass before inference. Eligibility was fixed before generation and is identical for every student. Runs ending in departure or collision before the eyes return have no lane opportunity; they are enriched for severe outcomes, and the assay does not cover them. Table 6(a) gives the selection flow.

Generation, missingness, and uncertainty. All students use BF16, greedy decoding, batch size one, and native end-of-sequence tokens. Answer probes prefill `#ANSWER rung=` and allow 32 new tokens; continuations allow 900. The frozen parser takes the last complete answer clause and scores the reasoning preceding it; a missing clause is a parse failure, never repaired. Generated timestamps must increase past the cut. Incorporation uses the first fully off-road phase; the lane band uses the first on-road line within the alignment tolerance. A non-monotone continuation is undefined for both, not scored as a failure. Primary lane scores require both arms to be incorporated and aligned; nominal scores keep the full cohort as denominator. Counting undefined continuations as failures makes every model-minus-B_2^* difference more negative. Pair bootstraps jointly resample arms, boundaries, and generation modes (10,000 draws); comparisons with a reference add a cluster bootstrap over the 45 peripheral and 41 hands cells.

Analysis chronology. Four interpretation cases were recorded before inference, with the reference left as "the process's own reproduction of the consequence": (1) incorporated state without a written consequence beyond the reference; (2) a written consequence beyond the reference; (3) state not incorporated; (4) contaminated prefix or inadequate coverage. The construction of the reference was revised after inference, from a same-specification reproducibility agreement on the campaign's own seeds, to a leave-one-out majority, to B_1^* and B_2^* from twenty re-executions per pair, and finally to B_Δ, added when the students' arm-wise structure was recognized as one the paired majority does not share. Classification against B_2^* is therefore a post hoc reading of the pre-specified cases, not a pre-specified choice of baseline; the recorded generations are unchanged. Against B_2^* both cohorts fall under case (1) for all three students.

Complete results. Table 5 gives every scored quantity for the three students on both cohorts: incorporation under the conditional and the nominal denominator, lane direction and invariance, joint three-way accuracy, and the answer calibration under all four boundaries, with the earlier leave-one-out reference rules and the best one-direction constant beside them. The main-text B_1^*/B_2^* comparisons use each student's own scored pairs and are not repeated there.

Details. Table 6 collects four checks. (a) The selection flow of both cohorts. (b) What a continuation writes when it fails to move: among incorporated and aligned moving references on hands whose generated lane delta is zero, about two fifths hold the factual band under the counterfactual state, one seventh land on the counterfactual band for both members, and nearly half write a band that matches neither, dominated by a band more severe than either reference; the last group is a lane sentence written without regard to the hand state. (c) Whether the channel-age clauses that the hands

intervention toggles on the same line as the hands clause leak the lane direction: with the numeric ages abstracted every pair has the same clause pattern; with exact ages a majority rule reaches 0.621 leave-one-out, and a rule on the glance duration alone, which the specification states, reaches 0.591, so the age is a proxy for the duration and carries no direction information beyond it. The hands clause is the corpus default, so near-perfect incorporation on hands is weak evidence by itself. (d) The 7B student: one Qwen2.5-7B-Instruct adapter trained on the same frozen 8,000 examples with the recipe of Appendix E unchanged (1,500 steps, 3.6 GPU-hours); a pre-specified check of 100 in-distribution and 100 out-of-distribution examples at the 900-token budget of the other students and at 2,400 tokens, the only change between the two rows; at 900 tokens the examples whose reference traces exceed the budget stop without an answer, at 2,400 all complete and the remaining generations are token-identical. The numbers of Table 2 come from the full 1,000 and 999 held-out examples at 2,400 tokens (0.800, 0.723, no parse failures or budget stops); no sample, objective, or adapter was changed after the check. The matched assay keeps its 900-token continuation budget.

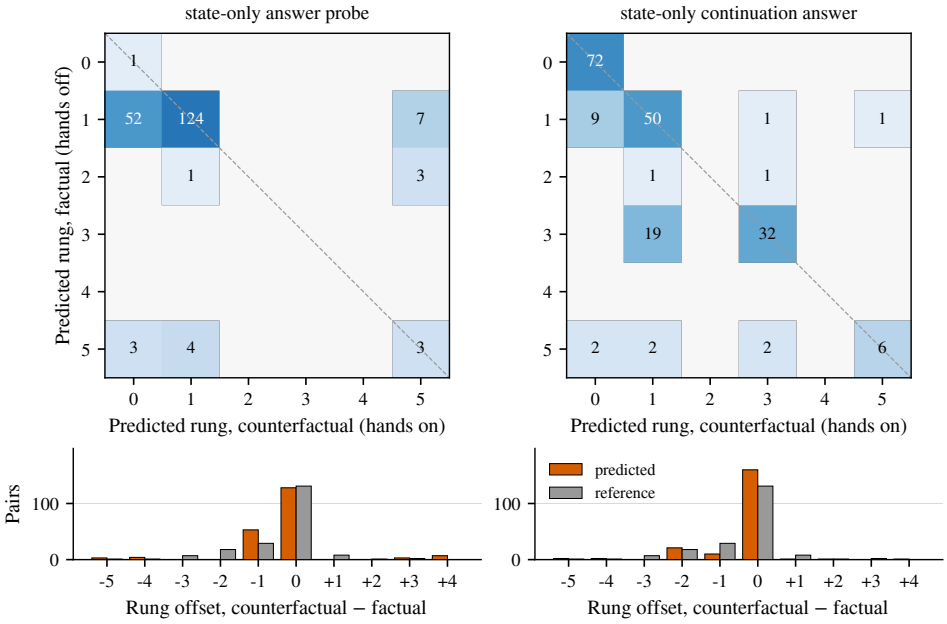


Figure 4: The 3B student's state-to-answer offset on the hands cohort, in the layout of Figure 3. Its fixed offset is keyed to the 1.5 s glances; the remaining changes are scattered.

Table 5: Complete assay numbers: (a) answer calibration with leave-one-out robustness rows; (b, c) every observable, one table per mechanism.

(a) Answer calibration and leave-one-out robustness

	Incorp.	Lane (conditional)		Answer, continuation		Answer, state-only	
	(nominal)	dir.	inv.	dir.	inv.	dir.	inv.
Peripheral gain off (233 pairs)							
Qwen2.5-1.5B	0.403	0.803	0.591	0.542	0.863	0.000	0.963
Llama-3.2-3B	0.893	0.811	0.688	0.875	0.677	0.000	0.901
Qwen2.5-7B	0.961	0.669	0.764	0.569	0.708	0.000	1.000
B1 agreement (LOO)	–	0.763	0.540	–	–	–	–
B2 LOO majority, cell × onset	–	0.850	0.559	–	–	–	–
B3 LOO majority, cell	–	0.864	0.634	–	–	–	–
best constant rule	–	0.979	0.000	1.000	0.000	1.000	0.000
Hands off wheel (198 pairs)							
Qwen2.5-1.5B	1.000	0.462	0.438	0.522	0.763	0.030	0.954
Llama-3.2-3B	0.934	0.245	0.781	0.284	0.878	0.075	0.542
Qwen2.5-7B	1.000	0.140	0.719	0.075	0.863	0.746	0.412
B1 agreement (LOO)	–	0.511	0.490	–	–	–	–
B2 LOO majority, cell × onset	–	0.590	0.457	–	–	–	–
B3 LOO majority, cell	–	0.607	0.506	–	–	–	–
best constant rule	–	0.846	0.000	0.836	0.000	0.836	0.000

Answer columns: direction agreement over rung-moving references and invariance over rung-stable references. Reference rows, lane columns: accuracy on moving / invariant lane references over all cohort pairs for the leave-one-out (LOO) baselines; agreement rate for B1; best of always-increase / always-decrease for the constant rule (answer columns against the maximum-rung reference).

Table 5 (b) Complete numbers, peripheral gain off (233 pairs), all four generation configurations

Size	Observable	Valid	Raw change	Conditional / direction	Nominal / invariance
1.5B	state incorporation	–	–	0.653 [0.576, 0.729] 94/144	0.403 [0.343, 0.468] 94/233
	lane: aligned pairs	93/233	–	–	–
	lane: direction	–	–	0.803 [0.708, 0.892] 57/71	0.407 [0.331, 0.490] 57/140
	lane: invariance	–	–	0.591 [0.370, 0.792] 13/22	0.140 [0.072, 0.216] 13/93
	lane: failure to move	–	–	0.183 [0.096, 0.276] 13/71	–
	answer: state-only	233/233	6/233	0.000 [0.000, 0.000] 0/72	0.963 [0.930, 0.988] 155/161
	answer: continuation	232/233	61/232	0.542 [0.425, 0.657] 39/72	0.863 [0.806, 0.913] 138/160
	answer: conseq. supplied	233/233	74/233	0.778 [0.676, 0.871] 56/72	0.888 [0.838, 0.934] 143/161
	answer: full trace	233/233	69/233	0.819 [0.726, 0.905] 59/72	0.938 [0.898, 0.974] 151/161
3B	state incorporation	–	–	0.959 [0.930, 0.982] 208/217	0.893 [0.850, 0.931] 208/233
	lane: aligned pairs	202/233	–	–	–
	lane: direction	–	–	0.811 [0.741, 0.879] 99/122	0.707 [0.630, 0.781] 99/140
	lane: invariance	–	–	0.688 [0.582, 0.787] 55/80	0.591 [0.490, 0.691] 55/93
	lane: failure to move	–	–	0.164 [0.100, 0.233] 20/122	–
	answer: state-only	233/233	16/233	0.000 [0.000, 0.000] 0/72	0.901 [0.852, 0.945] 145/161
	answer: continuation	233/233	115/233	0.875 [0.795, 0.946] 63/72	0.677 [0.604, 0.747] 109/161
	answer: conseq. supplied	233/233	84/233	0.847 [0.758, 0.924] 61/72	0.857 [0.803, 0.910] 138/161
	answer: full trace	233/233	69/233	0.819 [0.726, 0.905] 59/72	0.938 [0.899, 0.973] 151/161
7B	state incorporation	–	–	0.970 [0.947, 0.991] 224/231	0.961 [0.936, 0.983] 224/233
	lane: aligned pairs	222/233	–	–	–
	lane: direction	–	–	0.669 [0.591, 0.750] 89/133	0.636 [0.557, 0.715] 89/140
	lane: invariance	–	–	0.764 [0.671, 0.849] 68/89	0.731 [0.638, 0.818] 68/93
	lane: failure to move	–	–	0.316 [0.237, 0.394] 42/133	–
	answer: state-only	233/233	0/233	0.000 [0.000, 0.000] 0/72	1.000 [1.000, 1.000] 161/161
	answer: continuation	233/233	89/233	0.569 [0.452, 0.686] 41/72	0.708 [0.634, 0.777] 114/161
	answer: conseq. supplied	233/233	75/233	0.819 [0.725, 0.905] 59/72	0.901 [0.852, 0.944] 145/161
	answer: full trace	233/233	68/233	0.792 [0.694, 0.882] 57/72	0.932 [0.891, 0.968] 150/161

Rate [95% interval] count. State and lane rows: column 5 is conditional (valid phase and clause; incorporated and aligned), column 6 is over all nominal opportunities. Answer rows: column 5 is direction agreement over rung-moving references, column 6 invariance over rung-stable references, both over answer-valid pairs. 95% joint pair-bootstrap intervals.

Table 5 (c) Complete numbers, hands off wheel (198 pairs), all four generation configurations

Size	Observable	Valid	Raw change	Conditional / direction	Nominal / invariance
1.5B	state incorporation	–	–	1.000 [1.000, 1.000] 198/198	1.000 [1.000, 1.000] 198/198
	lane: aligned pairs	125/198	–	–	–
	lane: direction	–	–	0.462 [0.360, 0.564] 43/93	0.368 [0.279, 0.455] 43/117
	lane: invariance	–	–	0.438 [0.267, 0.611] 14/32	0.173 [0.095, 0.256] 14/81
	lane: failure to move	–	–	0.387 [0.287, 0.485] 36/93	–
	answer: state-only	198/198	8/198	0.030 [0.000, 0.077] 2/67	0.954 [0.915, 0.985] 125/131
	answer: continuation	198/198	73/198	0.522 [0.400, 0.642] 35/67	0.763 [0.687, 0.835] 100/131
	answer: conseq. supplied	198/198	72/198	0.806 [0.707, 0.896] 54/67	0.885 [0.829, 0.937] 116/131
	answer: full trace	198/198	61/198	0.851 [0.759, 0.930] 57/67	0.969 [0.938, 0.993] 127/131
3B	state incorporation	–	–	1.000 [1.000, 1.000] 185/185	0.934 [0.899, 0.965] 185/198
	lane: aligned pairs	126/198	–	–	–
	lane: direction	–	–	0.245 [0.161, 0.333] 23/94	0.197 [0.128, 0.272] 23/117
	lane: invariance	–	–	0.781 [0.625, 0.917] 25/32	0.309 [0.210, 0.411] 25/81
	lane: failure to move	–	–	0.702 [0.608, 0.792] 66/94	–
	answer: state-only	198/198	70/198	0.075 [0.016, 0.143] 5/67	0.542 [0.455, 0.627] 71/131
	answer: continuation	198/198	38/198	0.284 [0.179, 0.394] 19/67	0.878 [0.818, 0.930] 115/131
	answer: conseq. supplied	198/198	68/198	0.791 [0.691, 0.884] 53/67	0.901 [0.847, 0.949] 118/131
	answer: full trace	198/198	64/198	0.851 [0.761, 0.930] 57/67	0.947 [0.904, 0.984] 124/131
7B	state incorporation	–	–	1.000 [1.000, 1.000] 198/198	1.000 [1.000, 1.000] 198/198
	lane: aligned pairs	125/198	–	–	–
	lane: direction	–	–	0.140 [0.073, 0.213] 13/93	0.111 [0.057, 0.171] 13/117
	lane: invariance	–	–	0.719 [0.555, 0.868] 23/32	0.284 [0.188, 0.383] 23/81
	lane: failure to move	–	–	0.828 [0.747, 0.901] 77/93	–
	answer: state-only	198/198	138/198	0.746 [0.636, 0.847] 50/67	0.412 [0.328, 0.500] 54/131
	answer: continuation	198/198	25/198	0.075 [0.016, 0.143] 5/67	0.863 [0.800, 0.918] 113/131
	answer: conseq. supplied	198/198	64/198	0.821 [0.723, 0.908] 55/67	0.931 [0.885, 0.971] 122/131
	answer: full trace	198/198	62/198	0.851 [0.760, 0.931] 57/67	0.962 [0.926, 0.992] 126/131

Rate [95% interval] count. State and lane rows: column 5 is conditional (valid phase and clause; incorporated and aligned), column 6 is over all nominal opportunities. Answer rows: column 5 is direction agreement over rung-moving references, column 6 invariance over rung-stable references, both over answer-valid pairs. 95% joint pair-bootstrap intervals.

Table 6: Assay details: cohort flow, failure-to-move decomposition, percept-age leakage check and the Qwen2.5-7B budget sanity check.

(a) Selection flow and composition of the two matched cohorts

Stage / property	Peripheral	Hands
Canonical pairs	394	393
Clean state prefix at first edited line	392	389
Shared first gaze return ($\pm$0.1 s)	328	311
Strict masked prefix match	233	206
Within 576-token window (final cohort)	233	198
Context C1 / C2 / C3	87 / 68 / 78	71 / 58 / 69
Cells	45	41
Glance 1.5 / 2.5 / 3.5 / 4.5 / 5.0 s	75 / 56 / 49 / 31 / 22	72 / 52 / 36 / 21 / 17
Lane reference: dec. / inv. / inc.	3 / 93 / 137	99 / 81 / 18
Rung reference: dec. / inv. / inc.	0 / 161 / 72	56 / 131 / 11

The peripheral cohort needed no token-window exclusion. Glance types: *lizard* (peripheral), *body lean* (hands); all single-glance protocols.

(b) Failure-to-move decomposition on the hands cohort

Model	n	both = factual band	both = counter-factual band	both = other band	mixed
Qwen2.5-1.5B	36	14 (0.389)	6 (0.167)	16 (0.444)	0
Llama-3.2-3B	66	25 (0.379)	10 (0.152)	31 (0.470)	0
Qwen2.5-7B	77	32 (0.416)	10 (0.130)	35 (0.455)	0

Incorporated/aligned moving references whose generated lane delta is zero. The mixed column is empty by construction: a zero ordinal delta means identical generated bands.

(c) Do percept-age clauses in the hands state-only prefix predict the reference lane direction?

Rule	Distinct patterns	In-sample	Leave-one-out
Majority direction (always decrease)	1	0.500 (99/198)	0.500 (99/198)
Clause pattern, ages abstracted	1	0.500 (99/198)	0.500 (99/198)
Exact clause strings incl. ages	11	0.646 (128/198)	0.621 (123/198)
Glance duration only (in the prompt)	5	0.631 (125/198)	0.591 (117/198)
Glance duration + age	18	0.652 (129/198)	0.591 (117/198)

Majority rule within each pattern, three-way target (decrease / invariant / increase). The exact ages add nothing beyond the glance duration already present in the prompt.

(d) Qwen2.5-7B held-out sanity: 900- versus 2,400-token budget on the same fixed examples

Split	Budget	Answered	Acc. over answered	Acc. over all (missing = wrong)	Sample majority	Macro-F1 (answered)	Budget stops
ID	900	89/100	0.685	0.610	0.660	0.300	11
ID	2,400	100/100	0.720	0.720	0.660	0.302	0
OOD	900	94/100	0.638	0.600	0.710	0.282	6
OOD	2,400	100/100	0.660	0.660	0.710	0.284	0

Only the generation budget differs. All 17 previously missing answers complete at 2,400 tokens; the other 183 generations are token-identical.

B RE-EXECUTION REFERENCES AND SELECTION-MATCHED EVALUATION

Execution and matching. The pair generator reads one random stream per run in the order initial lateral offset, initial speed offset, hazard onset, driver draw, and one normal per open-loop tick. The wrapper passes the canonical pair's full-precision onset explicitly while consuming the draw the onset used to take, so the canonical stream stays aligned, and runs each fresh driver seed at the identical cell, glance, and onset. On the tested canonical pair it reproduces the canonical verbalized traces and rungs byte for byte, while the remaining per-tick telemetry differs at the level at which CARLA physics is not bit-reproducible across server lives, as a plain canonical rerun of the same pair also does; full-cohort reproducibility was not separately verified. Twenty seeds for each of the 431 cohort pairs gave 8,620 pairs with zero worker errors in 21 minutes on four CARLA instances; the canonical warm-up filter removed 614, 8,006 were verbalized, and 0.91 of those carry a lane band in both arms at the first return. Every cohort pair retains at least four usable transitions (median 18). Majority ties are broken by the cell-level majority. The re-executions pass the pair generator's own checks (pre-intervention agreement, first discrete divergence after the intervention), the warm-up filter, and the requirement of a lane band in both arms at the first return, which is the transition the reference scores; they were not filtered by the cohort's remaining eligibility conditions. Two of those, masked-prefix equality and the token window, are properties of the student's text input and are not needed to define the transition; whether selecting on them shifts the transition's distribution is a separate question, addressed only at the stricter level below and with the caveats given there. The shared-return tolerance of ± 0.1 s is a property of the pair, so we also applied each cohort builder's own eligibility function to every verbalized re-execution. The shared-return rule keeps 3,876 of 3,965 usable peripheral and 3,294 of 3,357 usable hands re-executions and every cohort pair (median 17 and 18 seeds per pair, minimum 6 and 4); the full builder verdict is eligible for 2,975 and 2,599, shared-return only for 901 and 695, no shared return for 342 and 331, and a pre-intervention text disagreement for 95 and 68. On the return-matched seeds B_2^* is unchanged to three decimals on every split (0.773 and 0.586 jointly), B_1^* moves by 0.002, and B_Δ rises from 0.657 to 0.691 $[0.614, 0.759]$ on peripheral and from 0.535 to 0.556 $[0.491, 0.612]$ on hands (0.751 and 0.581 under the mild tie rule). The student differences against B_1^* and B_2^* keep their flags; against the return-matched B_Δ the hands students read -0.064 $[-0.195, +0.041]$, -0.135 $[-0.250, -0.025]$, and -0.232 $[-0.351, -0.117]$, and the peripheral students $+0.032$ $[-0.079, +0.133]$, $+0.069$ $[0.000, +0.139]$, and $+0.014$ $[-0.051, +0.079]$, so the one reading that changes is the 3B student on hands, from indistinguishable to below. Adding masked-prefix equality as well keeps 2,975 and 2,599 re-executions, drops two hands pairs, leaves twelve pairs with fewer than three seeds, and gives B_2^* of 0.755 and 0.582 and B_Δ of 0.648 and 0.510; that level is reported in the released result files only, and its student-level differences were not evaluated, so what this appendix establishes is robustness to the shared-return rule, not to the cohort's full eligibility. Table 7(a) carries the return-matched rows.

Reference construction. The canonical pair is excluded from its own ensemble. B_1^* averages the agreement of individual re-executions with the reference class; B_2^* predicts the ensemble's modal class. Both are finite-sample rules on the realized reference class, not optima over specification-only predictors: a pair whose reference class is not the ensemble mode contributes zero to B_2^* and may contribute a positive amount to B_1^*, so the observed values need not be ordered as their limits are, and a better specification-only estimator than a twenty-seed majority could exceed B_2^*. A majority that also conditions on everything the STATE-ONLY prefix reveals, the cell, the glance duration, the onset bin, and the initial lane band and side, coincides with the specification-level majority on every cohort pair, because every prefix begins in the negligible band and the glance schedule is part of the cell. On peripheral the empirical modal prediction coincides with the cell-only rule on all 233 pairs; the modal class is the same, which does not imply that the conditional distributions are. The campaign-seed leave-one-out majority had a median of one comparison seed per specification group and had placed the 3B student above it on peripheral ($+0.040$ $[+0.005, +0.081]$); against B_2^* that difference is -0.005 $[-0.023, +0.011]$. B_Δ takes, over the same K_p re-executions, the modal factual band and the modal counterfactual band separately and predicts the sign of their difference, invariant when either mode is tied; a variant breaks a tied arm toward the milder band. It mirrors the arm-wise structure of a model that generates each arm separately and differences the results, as the students do, without reproducing the token-level greedy decoder, whose band is $h(\mathrm{Greedy}\, P_\theta(\tau \mid I))$ rather than $\arg\max_b \Pr_\theta[h(\tau) = b \mid I]$, and whose procedure contains no tie rule. It is weaker than B_2^* on both mechanisms: on all cohort pairs it reads 0.657 $[0.577, 0.729]$ against 0.773 on peripheral

and 0.535 $[0.468, 0.593]$ against 0.586 on hands (paired differences -0.116 $[-0.176, -0.062]$ and -0.051 $[-0.123, +0.018]$), disagreeing with B_2^* on 45 of 233 and 51 of 198 pairs, almost always by predicting invariant where the paired majority moves; where they disagree B_2^* is right on 35 and 27 pairs and B_Δ on 8 and 17. Ties drive most of the disagreement (40 and 32 pairs), and the mild tie rule brings B_Δ to 0.738 and 0.581, within 0.035 and 0.005 of B_2^*; the gap between the two rules is therefore mostly tie handling on a finite ensemble, not a structural cost of arm-wise prediction, and we report both readings. On each student's own set B_Δ reads 0.688, 0.658, and 0.658 on peripheral and 0.488, 0.484, and 0.488 on hands; the student differences are in Table 4, and under the mild tie rule the hands students read -0.104 $[-0.204, -0.016]$, -0.175 $[-0.271, -0.077]$, and -0.272 $[-0.401, -0.143]$, while the 3B peripheral difference falls to $+0.025$ $[-0.011, +0.063]$. Model-against-reference differences in Table 4 use each student's own incorporated-and-aligned pairs with a 10,000-draw pair bootstrap and a cell-cluster bootstrap. Table 7(a) gives, per mechanism and split by moving and invariant references, B_1^*, B_2^*, B_Δ under both tie rules, the campaign-seed leave-one-out majority, the cell-only majority, and a leave-one-out five-nearest-neighbour rule on the specification's features, all on the full cohorts; these all-cohort values do not replace the student-specific references of the main comparison.

A later, seed-informed boundary. A positive demonstration that a student uses seed-dependent information needs a boundary at which such information precedes the outcome. Among the narrated events, the time at which braking begins carries the driver's reaction-time draw: across the 20 re-executions of one specification its median within-specification standard deviation is 0.33 s and its median range 1.25 s, whereas the hazard-registration and decision times have median within-specification standard deviations no greater than 0.05 s. Eligibility for a cut immediately after the braking clause was fixed before any accuracy was computed: the clause must occur in both true traces after the intervention line and before the run ends, and at the cut, using only verbalized timestamps at their printed precision, the cumulative maximum rung must be 0 in both arms, no adverse event may have occurred, and no band written before the cut may exceed *slight*. 199 of 431 pairs qualify (107 peripheral, 92 hands, none in the curve context, which has no hazard), against a pre-fixed go criterion of 100; their canonical rung transition is invariant in 192 and moving in 7. Three rules were built from the 20 re-executions and evaluated on the canonical pair, which is held out by construction: the unconditioned majority over the rung transition, an arm-wise majority over the re-executions whose braking time falls in the same 0.25 s bin as the canonical arm's (falling back to the unconditioned majority below three seeds), and, as a pair-level oracle using more information than a student sees, a majority conditioned on both arms' braking times jointly.

On the eligible canonical pairs the arm-wise conditioned majority gained $+0.050$ $[+0.012, +0.102]$ in accuracy over the unconditioned one with no change in Brier score; between the re-executions themselves, leaving one seed out with both arms removed together, the gain was -0.007 $[-0.019, +0.003]$ and the Brier score worsened by 0.016 $[0.010, 0.025]$. The discrepancy is consistent with eligibility-induced distribution shift: the eligibility rule had been applied to the canonical pairs, whose outcomes are 96% invariant, but not to the re-executions, whose outcomes are 74% invariant. Applying the same rule to every re-execution, arm-wise, retains 2,597 of 3,527 (73.6%), matches the class mix (3.7% against 3.5% moving; 93% against 91% at rung 0 to 0), and leaves 185 pairs with at least three matched seeds. On that set the unconditioned majority, the conditioned majority, and the constant *invariant* rule score 0.951, 0.951, and 0.962 in accuracy, with Brier scores (three-class sum of squared errors) of 0.053, 0.047, and 0.076; the conditioned gain over the unconditioned majority is 0.000 $[-0.016, +0.017]$, the seed-based rules have lower Brier point estimates than the constant rule, with an interval that includes zero on the full matched set (-0.029 $[-0.069, +0.003]$ for the conditioned majority) and excludes it on hands (-0.066 $[-0.157, -0.008]$), and no rule predicts any of the seven moving references. Leave-one-seed-out on the matched set gives the same picture. The unmatched analysis was not repeated on the same 185 pairs, so selection is the leading explanation for the earlier gain rather than its confirmed sole cause. We did not run a student on the rung endpoint: the count criterion was met and is reported as met, and the endpoint that the leakage-free boundary leaves is nearly degenerate in the outcome the claim concerns. Table 7(b) gives the four rung-endpoint rows.

Table 7: Specification-level baselines: spec-matched and leave-one-out baselines for the lane-transition class, and feasibility baselines for the rung endpoint.

(a) Lane-transition class (decrease / invariant / increase), all cohort pairs

Baseline	Moving references	Invariant references	All pairs
Peripheral gain off (233 cohort pairs, 45 cells)			
B1* re-execution agreement	0.753 [0.688, 0.807]	0.584 [0.428, 0.684]	0.686 [0.644, 0.726]
B2* re-execution majority	0.864 [0.768, 0.938]	0.634 [0.400, 0.785]	0.773 [0.724, 0.820]
B_Δ, strict tie rule	0.621 [0.504, 0.728]	0.710 [0.513, 0.837]	0.657 [0.577, 0.729]
B_Δ, mild tie rule	0.786 [0.682, 0.872]	0.667 [0.450, 0.804]	0.738 [0.676, 0.797]
B2 LOO, cell × onset bin	0.850 [0.750, 0.927]	0.559 [0.333, 0.713]	0.734 [0.668, 0.795]
B3 LOO, cell only	0.864 [0.768, 0.938]	0.634 [0.400, 0.785]	0.773 [0.724, 0.820]
B5 LOO k-NN ($k = 5$)	0.836 [0.734, 0.917]	0.613 [0.378, 0.764]	0.747 [0.680, 0.809]
B1*, return-matched seeds	0.756 [0.690, 0.809]	0.587 [0.431, 0.687]	0.688 [0.648, 0.728]
B2*, return-matched seeds	0.864 [0.768, 0.938]	0.634 [0.400, 0.785]	0.773 [0.724, 0.820]
B_Δ strict, return-matched seeds	0.664 [0.550, 0.768]	0.731 [0.542, 0.849]	0.691 [0.614, 0.759]
Hands off wheel (198 cohort pairs, 41 cells)			
B1* re-execution agreement	0.517 [0.466, 0.564]	0.488 [0.401, 0.551]	0.505 [0.475, 0.530]
B2* re-execution majority	0.615 [0.482, 0.726]	0.543 [0.338, 0.701]	0.586 [0.517, 0.647]
B_Δ, strict tie rule	0.393 [0.270, 0.504]	0.741 [0.600, 0.853]	0.535 [0.468, 0.593]
B_Δ, mild tie rule	0.521 [0.384, 0.638]	0.667 [0.500, 0.793]	0.581 [0.516, 0.637]
B2 LOO, cell × onset bin	0.590 [0.451, 0.702]	0.457 [0.211, 0.642]	0.535 [0.440, 0.621]
B3 LOO, cell only	0.607 [0.463, 0.728]	0.506 [0.246, 0.701]	0.566 [0.464, 0.650]
B5 LOO k-NN ($k = 5$)	0.556 [0.422, 0.669]	0.568 [0.343, 0.721]	0.561 [0.476, 0.634]
B1*, return-matched seeds	0.519 [0.466, 0.567]	0.490 [0.404, 0.554]	0.507 [0.477, 0.532]
B2*, return-matched seeds	0.615 [0.482, 0.726]	0.543 [0.338, 0.701]	0.586 [0.517, 0.647]
B_Δ strict, return-matched seeds	0.427 [0.305, 0.538]	0.741 [0.600, 0.853]	0.556 [0.491, 0.612]

Accuracy with 95% cell-cluster intervals. B1*: agreement of the specification-matched re-executions with the reference class; B2*: their majority class (mode of the per-seed differences); B_Δ: sign of the difference between the two arms' modal bands over the same re-executions, a tied arm mode read as invariant (strict) or broken toward the milder band (mild). B1* and B2* use only 20 new driver seeds executed at each cohort pair's exact specification; the canonical pair is held out by construction. LOO: leave-one-out over the canonical pool with back-off to the next coarser group. $B_{\Delta\mathrm{mode}}$ uses the same re-executions as B2* and disagrees with it on 45 of 233 peripheral and 51 of 198 hands pairs (strict rule; 17 and 27 under the mild rule). Return-matched seeds: the cohort's eligibility rule up to the shared first gaze return (± 0.1 s) applied to every re-execution, which keeps 3,876 of 3,965 peripheral and 3,294 of 3,357 hands re-executions and every cohort pair. A prefix-informed LOO baseline coincides with B2 pair for pair on these cohorts and is not listed. Per-student comparisons against every baseline are in the supplementary material.

Table 7 (b) Rung endpoint (E6b): rung-transition class on the 199 eligible pairs

Evaluation (pairs)	(i) unconditioned	(ii) braking-time cond.	constant invariant
held out, all seeds (199)	0.744 [0.627, 0.830] Brier 0.316	0.794 [0.690, 0.872] Brier 0.317	0.965 [0.934, 0.987] Brier 0.070
leave one seed out, all seeds (199)	0.819 [0.755, 0.871] Brier 0.244	0.812 [0.746, 0.864] Brier 0.260	0.746 [0.652, 0.815] Brier 0.509
held out, matched seeds (185)	0.951 [0.908, 0.982] Brier 0.053	0.951 [0.907, 0.982] Brier 0.047	0.962 [0.928, 0.986] Brier 0.076
leave one seed out, matched seeds (185)	0.962 [0.936, 0.981] Brier 0.053	0.961 [0.935, 0.980] Brier 0.057	0.963 [0.938, 0.981] Brier 0.073

Held out: the canonical pair is the test item; matched: the eligibility rule is also applied to every re-execution. Each cell: accuracy [95% cell-cluster interval]; Brier score (three-class sum of squared errors, range 0–2). Eligible pairs: braking under way after the intervention in both arms with no rung reached before the cut; 7 of the 199 canonical references move. Selection-matched: the same eligibility rule applied to every re-executed seed (73.6% survive; pairs with at least three survivors). (ii) conditions each arm on the 0.25 s bin of its braking-start time and falls back to (i) below three seeds.

C EXECUTABLE REFERENCE PROCESS

Driver and scenarios. The driver runs on a 50 ms production cycle synchronized to each simulator tick. The two-point steering law combines a near lane-position cue and a far curvature or lead cue (Salvucci & Gray, 2004); longitudinal control regulates headway. When gaze is away, commands are held and calibrated zero-mean steering noise accumulates as drift. Eyes-only glances preserve near-channel information through a peripheral gain; head turns and body leans do not, and a body lean removes steering authority while the hand is off the wheel. Looming (Hoffmann & Mortimer, 1994) or brake lights initiate encoding, decision, motor, and residual reaction-time stages (Green, 2000); a re-acquisition penalty on the first percept after gaze returns and a motor penalty for a hazard met with a hand off the wheel are charged conditionally. Transitions consume a type-specific switch time on the way out and back. The parameter set is frozen under a recorded hash that the harness asserts at the start of every run (the value is in the released code); seed variation enters through reaction-time offsets, drift magnitude, control gains, the initial lateral position (± 0.05 m), and the speed (± 0.5 km/h). Table 8 lists every value.

C1 is a lead vehicle braking at 80 km/h over headways 1.5, 2.0, 2.5, 3.0 s and decelerations 3.0, 4.5, 6.0 m/s^2. C2 is an adjacent-lane cut-in closing to gaps 0.8, 1.0, or 1.5 s. C3 is a trafficless curve of radius 300, 400, 600, or 1200 m; 200 m was excluded because the fixed 1,500 m arc wraps past a circle. Hazard onset is uniform over the glance window extended slightly on either side. The maximum rung is 0 (none), 1 (wheel past the lane edge), 2 (centre past it), 3 (vehicle wholly outside its lane), 4 (road departure), or 5 (collision); collision, departure, or clock expiry ends a run, which is what made the run length an outcome-dependent field.

Table 8: Frozen driver parameters; none was refit for this paper.

Parameter	Value	Origin
Cycle	0.050 s	ACT-R
Near / far / integral steering gains	8.0 / 2.0 / 8.0	Companion
Near / far look-ahead	1.0 / 3.0 s	Design; two-point
Steering ratio / max wheel angle	12.0 / 70°	Vehicle
Headway / closing-rate gain	1.0 / 0.8	Companion
Target headway / speed gain	1.5 s / 0.6	Design
Max acceleration / braking	3.0 / 8.0 m/s^2	Vehicle
Visual encoding / decision / motor	0.085 / 0.050 / 0.150 s	ACT-R
PBR medians: expected / unexpected / surprise	0.75 / 1.25 / 1.50 s	Green
PBR lognormal sigma: expected / unexpected	0.30 / 0.35	Calibrated
Re-acquisition penalty	0.32 s	Calibrated
Body-lean motor / concurrent-task penalty	0.30 / 0.10 s	Calibrated; threaded
Warning response	0.50 s	Calibrated
Switch time: owl / lizard / body lean	0.48 / 0.16 / 1.0 s	Protocol
Peripheral gain: owl / lizard / body lean	0.0 / 0.3 / 0.0	Design; calibrated
Steering noise SD per cycle	0.0045 rad	Calibrated
Looming threshold	0.003 s^{-1}	Calibrated
Gain CV / drift CV	0.15 / 0.25	Design

ACT-R: Anderson et al. (2004); two-point law: Salvucci & Gray (2004); longitudinal companion values: Salvucci (2006); Green: Green (2000); threaded cognition: Salvucci & Taatgen (2008); drift calibration: Godthelp et al. (1984); looming: Hoffmann & Mortimer (1994). PBR denotes perception-brake response. "Design" marks values fixed by the companion study's specification; "calibrated" marks values fitted once by that study to published lateral-control data and frozen.

The freeze gate measured lane deviations of 0.027 m on the straight road and 0.149 and 0.075 m on the 300 and 600 m curves, a settled headway of 1.501 s with ripple below 0.001, and a median time-to-line-crossing of 3.26 s. The companion study varied every literature-sourced value by $\pm 30\%$; this paper's analysis constants are swept separately in Appendix G.

Mechanisms and pair checks. The factual arm of every pair runs the untouched frozen parameters. Two mechanisms override parameters on a copy whose own hash is written on the pair; the

rest act on the world or the schedule. Every pair is checked before it is kept: the pre-intervention rows must agree on every field, and the first cycle at which any discrete cognitive state differs must lie after the intervention takes effect. Divergence is measured first on the discrete cognitive state and only second on control, because two CARLA runs of one seed differ at the physics noise floor even with no intervention. The generator wrote 1,800 pairs in the first expansion and 1,600 for the targeted offset mechanism with zero worker errors; after the warm-up filter the ladder uses 2,348 pairs across the seven mechanisms, of which 529 have a moved maximum rung (481 under the five cognitive mechanisms, 43 under the early offset, and 5 under the inert offset; Table 14); the channel decomposition of Appendix F uses the separately expanded set of 1,582 admissible early-offset pairs, of which 243 move.

Table 9: Counterfactual mechanisms. "Applies to" lists the glance types and contexts in which the mechanism can act at all.

Mechanism	Change	Applies to
Peripheral gain off	Gain set to zero for an eyes-only glance	Lizard; C1–C3
Hand kept on wheel	Steering authority kept during a body lean	Body lean; C1–C3
Re-acquisition zero	First-return perceptual penalty removed	All types; C1–C2
Brake-light cue off	Lead still brakes; lamps stop being a cue; looming remains	All types; C1
Switch-time swap	Owl and lizard exchange transition times, not gains	Owl, lizard; C1–C3
Initial offset (inert)	Start 0.15 m further out, glance at 3.0 s; corrected away before the eyes leave	All types; all contexts
Initial offset (early)	Start 0.45 m further out, glance at 1.0 s, while the correction is still running	All types; all contexts

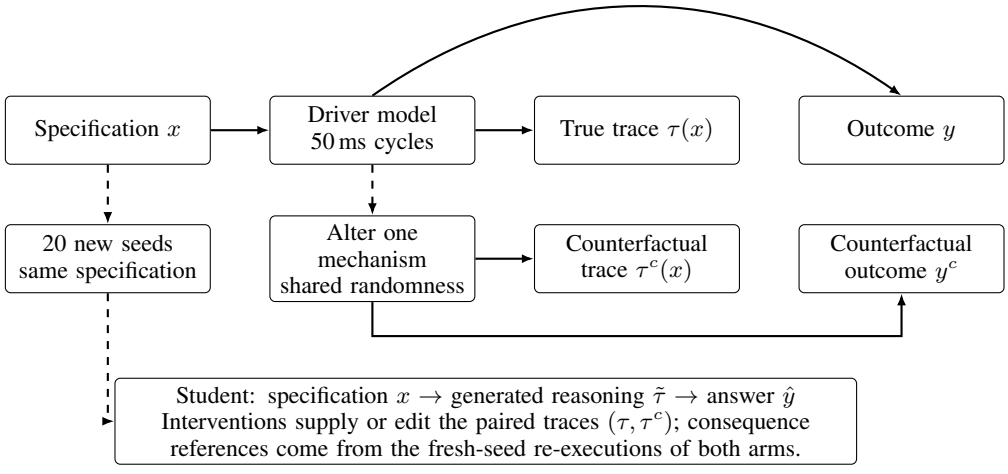


Figure 5: The setting (referenced from Section 3). The driver produces both trace and outcome; re-executing the same seed with one mechanism altered yields the counterfactual pair, and re-executing both arms from fresh seeds at the same specification yields the consequence references. The student is evaluated on the supplied or edited traces.

D VERBALIZER, PARSER, AND LEAKAGE AUDIT

Segmentation and timestamp semantics. A line is constant in gaze state, hazard stage, hand position, peripheral availability, and the re-acquisition flag, and a segment closes wherever any of the five changes. A one-tick peripheral-only change folds forward; zero-length spans are dropped; 30 s chunks are snapped so that no glance is split. Each maximal run of on-road segments is compressed into one line whose clauses keep their own event times, so the bracket of a compressed line marks

the start of the span and the clauses after it carry later times within that span (in the example below, the line beginning at 6 s carries a decision at 6.45 s and braking at 7.9 s). The STATE-ONLY cut of Appendix A falls on an off-road line, which is never compressed, and the suite asserts that no later-event material precedes the cut. Bands are qualitative on purpose: the text narrates a band and not a number.

Table 10: Verbalizer vocabulary and band definitions.

Family	Template / definition
Gaze	*Eyes on the road / swinging away from the road toward* {*target*} */ fully away from the road / swinging back to the road*; {target} distinguishes a head turn (owl), an eyes-only glance (lizard), and a low target reached by leaning (body lean), written once per chunk.
Near channel	*lane edges in peripheral view* when the peripheral gain is active; otherwise *lane current* or *lane unsampled* {*age*} *s*.
Far channel	*road ahead current* or *road ahead unsampled* {*age*} *s*; *both channels current* replaces the pair when both ages are zero.
Hands	*hands on wheel* / *hands off wheel* (medium and full layouts).
Re-acquisition	*still re-acquiring the scene*, or ... *until* {*t*} *s* inside a merged on-road stretch.
Hazard	*a closing cue is present ahead, not yet taken in* / *the hazard registers on brake lights* or *on looming* / *encoding the hazard* / *deciding on a response* / *braking is under way*; inside a merged stretch each stage carries *at* {*t*} *s* or *from* {*t*} *s*.
Lateral band	*Lane offset* negligible < 0.25 m; slight < 0.50; noticeable < 0.90; large < 1.75; severe otherwise, with left/right.
Headway band	*Gap ahead* very short < 0.9 s; short < 1.5; normal < 2.5; long otherwise, or *No vehicle ahead*.

Band edges are the severity ladder's own geometry: 0.917 m is a wheel over the lane edge and 1.75 m the vehicle centre past it; 1.5 s is the model's target headway.

The two layouts, one run. The original layout wrote each span as `[t1–t2]`, a `#SEG start–end` line per chunk, and a header field `DUR`; all three ended at the run's termination. The revised layout writes `[t]`, `#SEG start`, no `DUR`, and a closing sentence with no time. The same run (C1, head-turning glance of 2 s at 3 s, headway 2.5 s, deceleration 3 m/s^2, run length 17.95 s) reads in the revised layout:

```
#P COG|CTX=C1|G=owl|PAT=single|CELL=h2.5_d3|SPD=80|LOD=medium
#SEG 0
[0] Eyes on the road: both channels current; hands on wheel. Lane offset
negligible. Gap ahead long.
[3] Eyes swinging away from the road toward an off-road target, head
turned with the eyes: both channels current; hands on wheel.
[3.5] Eyes fully away from the road: lane unsampled 2.00 s; road ahead
unsampled 2.00 s; hands on wheel.
[5.5] Eyes fully away from the road: lane unsampled 2.05 s; road ahead
unsampled 2.05 s; hands on wheel; a closing cue is present ahead, not
yet taken in.
[5.55] Eyes swinging back to the road: both channels current; hands on
wheel; a closing cue is present ahead, not yet taken in.
[6] Eyes on the road: both channels current; hands on wheel; still
re-acquiring the scene until 6.1 s; the hazard registers at 6 s on brake
lights; encoding the hazard from 6.05 s; deciding on a response from
6.45 s; braking is under way from 7.9 s. Lane offset slight left. Gap
ahead normal.
The run ends.
#PREDICT
```

In the original layout the header carried `DUR=17.95`, the segment line `#SEG 0–17.95`, the last span `[6–17.95]`, and a final bracket `[17.95]`; the clauses are identical. Both complete listings are in the archive.

Round-trip and length checks. At a 0.1 s tolerance the reverse parser recovers the five span families and the hazard transitions from 1,875 traces sampled evenly across the campaign tree with

start-F1, full-span F1, and state accuracy of 1.000 in both layouts, and no record contains an outcome word. In the revised layout a span's end is implied by the next line's start, so it is recovered exactly for every span but the last of each family; that one end is the field the layout removed and is counted rather than scored (0.85 of gaze, 0.76 of hazard, and 0.51 of hands spans state an end). Under the Qwen2.5 tokenizer the 95th-percentile record length falls from 1,767 tokens in the original layout to 1,536 in the revised one; prompt plus completion on the 8,000 training examples has a 99th percentile of 2,455 tokens in the revised layout, and no example exceeds the 3,072-token budget in either.

Leak measurements. Over the 8,000 original-layout training records the header `DUR`, the final `#SEG` end, and the final line's end agree to four decimals in every record. A validation-selected k-nearest-neighbour majority on any of them ($k = 9$) scores 0.857 and 0.832 in rung accuracy (in and out of distribution) and 0.964 and 0.959 in L3-time accuracy, which is the check that they are one channel and not three; the accuracy is 0.845/0.817 at $k = 3$ and 0.855/0.828 at $k = 25$. Three structural features survive into the revised layout: the last stamp written ($k = 15$; 0.769/0.690), the chunk count ($k = 1$; 0.763/0.693), and the line count ($k = 25$; 0.785/0.726), against a majority class of 0.763/0.693 and the specification-only baseline of 0.809/0.750.

Table 11: Whole-trace replacement on the original-layout arm decomposed, by whether the factual and counterfactual runs had the same length.

Subset	n moving	Number only	Words only	Both (published)
Different lengths	189	.735	.413	.915
Same length	292	.000	.812	.812

Direction agreement on the original-layout arm. Where the lengths agree the number-only texts are identical and must read exactly zero, and the words-only condition must reproduce the published value; both do. Over all 481 moving pairs the number alone reads .289 and the words alone .655 (Section 5.1).

Two further conditions were run as demonstrations rather than measurements: truncating both traces to a common horizon with equalized headers pins the prediction to rung 5 on 0.916 of generations because the median horizon of 7.15 s reads as a collision, and keeping each arm's own length above a truncated body fails to produce an answer in 43% of generations (agreements 0.048 and 0.040), so the collapse is the truncation and not the equalization. A masked true-trace probe on the original-layout arm, with the run-length fields replaced by the training median, reads 0.799/0.771 against an unaided 0.813/0.731, but a model trained on the original layout and handed a masked original-layout trace is off its own distribution, which is why the arm was retrained rather than masked. Retraining on the revised layout cost two accuracy points, 0.790/0.716 against 0.813/0.731, four and three standard deviations over three seeds a side, with every span family's fidelity within a point; at 24,000 examples the revised-layout arm reaches 0.802/0.743.

E Campaign, training, and reconstruction

Campaign and validity. The campaign ran 58,080 closed-loop runs across 1,452 cells with 40 seeds per cell, sized for coverage rather than depth. A cell is a context, a severity cell, a glance type, and a glance duration or pattern; single glances cover 0.5 to 5.0 s on a 0.25 s grid at a 3.0 s onset, cumulative patterns cover the ten-in-thirty rule and two stress variants (12 in 30, 10 in 20), and a protocol block runs six named protocols. Reference cells were re-measured ten times against an abort tolerance of 0.12 on their rate; the worst drift observed was 0.033, and the campaign closed with zero worker errors. The testbed's own geometry and brake-actuation filters dropped no rows; a run whose hazard was detected during the harness warm-up, before the run's clock starts, has no cognition to narrate and was removed (3,813 rows, 11.2% of C1 and none of C2 or C3), leaving 54,267 admissible runs with a minimum of 29 and a median of 38 seeds per cell.

Holdouts and splits. Three factor levels were withheld by construction: the interior C1 cell with headway 2.0 s and deceleration 4.5 m/s^2, the single-glance duration 3.25 s, and a low-speed protocol at 30 km/h in all three contexts; their absence from the written rows was verified after the campaign. Training, validation, and in-distribution test sets of 8,000, 1,000, and 1,000 examples are drawn by

a deterministic round robin over the non-holdout cells, split by seed, so that the 2,000- and 24,000-example training sets are nested prefixes of the same order. The 999 out-of-distribution examples comprise 620 from the withheld severity cell, 343 from the withheld duration, and 36 from the low-speed protocol. Class balance is skewed toward rung 0 (Table 12), which is why Table 2 reports macro-F1 beside accuracy and why direction agreement counts only pairs whose outcome moved.

Table 12: Class balance per split, counts and fractions.

Split	n	Rung 0	1	2	3	4	5
Train	8000	6058 (.757)	462 (.058)	234 (.029)	265 (.033)	194 (.024)	787 (.098)
Validation	1000	754 (.754)	69 (.069)	31 (.031)	30 (.030)	16 (.016)	100 (.100)
ID test	1000	763 (.763)	60 (.060)	26 (.026)	34 (.034)	24 (.024)	93 (.093)
OOD test	999	692 (.693)	73 (.073)	36 (.036)	41 (.041)	18 (.018)	139 (.139)

Recipe and reproducibility. Every arm uses LoRA with rank 16, $\alpha = 32$, no dropout, on the query, key, value, output, gate, up, and down projections; completion-only loss; the base model's own chat template; 8-bit AdamW with a cosine schedule, warmup ratio 0.03, learning rate 2×10^{-4}, batch size 4 with gradient accumulation 4, three epochs, bfloat16, and seed 3407 unless stated (101 and 202 for the three-seed rows). The sequence budget is 3,072 tokens for every arm; the Qwen and Llama 99th-percentile lengths are 2,455 and 2,294, so the budget truncates nothing. One NVIDIA RTX 6000 Ada trains the 1.5B 8k and 24k, the 3B 8k, and the 7B 8k trace arms in 1.25, 3.4, 2.3, and 3.6 GPU-hours. Greedy decoding is not bit-reproducible across batch compositions: three batched inference files with identical factual text disagree on 24 of 1,582 pairs when padding changes, a floor of about 0.015 under every ladder number and smaller than every difference the paper rests on; the matched assay and the re-executions use batch size one.

Each arm receives the same five-line specification block (context, speed, gaze, hazard onset, warning) and answers with `#ANSWER rung=`k `| L3_time=`t, parsed by a strict regular expression whose last match wins; unparseable outputs count as errors (0.000 to 0.004 across arms and splits). The TRACE instruction is word for word identical under both layouts. The FREE teacher sees the specification and the label and never the trace; the FREE-LONG teacher targets the trace's token length, reaching a median ratio of 0.83 with 98% of examples within a factor of two, while its rate of invented numerical quantities rises from 24% to 41%. One example: for an eyes-only glance in C1 the teacher wrote that the driver's gaze was fixed on a lizard, reading the glance-type token as an animal; the account passes every automatic check and is wrong about what the glance is, which is the fluency the faithfulness literature warns about. Teacher rewrites under a counterfactual premise cannot be parsed for what they changed and support no claim in the main text.

Specification and reconstruction references. The specification-only baseline of Equation 1 conditions on the cell crossed with the onset floored to 0.25 s. Over the 10,999 written examples there are 8,790 distinct keys, none carrying two different specification blocks; the leave-one-out estimator leaves 113 in-distribution and 154 out-of-distribution examples in singleton groups, which are excluded rather than scored as correct, so the baseline is evaluated on 887 and 845 examples while the arms in Table 2 are scored on the full 1,000 and 999; the two are therefore compared as a reference on the non-singleton subset, not as a bound. The in-sample majority, which scores a group against itself, would read 0.894 and 0.858 and is reported only to show the size of that optimism. The reproducibility reference $\bar{F}_f$ is computed on 400 examples per split by scoring each example's true trace against another seed's true trace in its group. On the hazard chain the 1.5B, 3B, and 7B trace arms read 0.773, 0.795, and 0.796 in distribution against 0.747, and 0.701, 0.740, and 0.751 out of distribution against 0.686; the families the specification determines sit on their reference. These are text-reconstruction comparisons, not recovery of an unobserved seed.

F INTERVENTION CHANNELS AND CONTINUATION DIAGNOSTICS

Consequence channels. For each of the 1,582 admissible targeted-offset pairs, `band_only` replaces the factual trace's lateral-band and headway-band sentences, line by line, with the counterfactual's, and `all_but_band` restores the factual band sentences in an otherwise counterfactual

trace; nothing is invented. Nearest line starts within 0.1 s aligned 508 pairs, and 53 pairs kept their own band where no partner existed. Both texts are scored exactly as REPLACE. Over the 243 moving pairs, direction agreement is 0.412, 0.136, and 0.481 for band-only, all-but-band, and the whole counterfactual trace, with invariance 0.926, 0.953, and 0.899; the same-cell donor restricted to this mechanism reads 0.620 $[0.533, 0.701]$ over 137 moving pairs with invariance 0.845. On 911 pairs the two true traces differ in nothing but their bands; restricted to the 671 pairs where both channels change text, the three read 0.368, 0.200, and 0.477, and the donor 0.657 over 67 moving pairs. The two channels' signed predicted deltas sum to the whole trace's exactly on 0.922 of pairs; the departure from additivity is 0.011 $[-0.018, 0.038]$ over all pairs and -0.066 $[-0.169, 0.029]$ over the moving ones. The channels' own predictions differ on 0.175 of pairs and 0.605 of moving pairs, above the 0.015 batch-order floor; the per-pair results are in the released result files.

Clause propagation versus span-start lead. For a continuation $\tilde{\tau}'$ the span-start lead on family f is

$$\text{lead}_f = \text{F1}_f(\tilde{\tau}', \tau^c) - \text{F1}_f(\tilde{\tau}', \tau), \tag{4}$$

positive when the continuation moved toward the counterfactual run. A state carried forward unchanged creates no span start, so the lead is dominated by the edited line, which the model did not write: after the cut it rests on 2 of 120 hands continuations. For brake-light removal, 84 of 92 pairs write nothing after the cut and the -0.281 lead at the edited line is shared by the edited, deleted, and unedited conditions; it is the factual hazard chain narrated into the prefix. The clause-state view instead scores generated clauses only at times when the two true traces disagree (Table 13).

Table 13: Fraction of lines after the cut, on which the two true traces disagree, whose clause for the touched family matches the counterfactual trace's.

Condition	Line 1	Line 2	Line 3	Line 4+
Rewritten hands	1.000 (120)	1.000 (102)	1.000 (57)	1.000 (21)
Rewritten peripheral	1.000 (120)	1.000 (118)	1.000 (91)	.989 (88)
Deleted hands, where stated	1.000 (23)	1.000 (35)	1.000 (22)	–
Deleted peripheral	.667 (120)	.667 (120)	.621 (95)	.920 (163)
Unedited hands	.000 (120)	.247 (93)	.552 (58)	.947 (19)
Unedited peripheral	.000 (120)	.000 (120)	.609 (115)	1.000 (45)

Parentheses are evaluated line counts. Deleted continuations omit the hand clause on 1.000 of first lines and .742 of second lines; whenever they name the hands they name the counterfactual's state.

Unedited matches rise only where the factual glance ends too. The unaligned consequence counts do not follow the edit: on lines that state a lane band, a rewritten peripheral continuation names the counterfactual's band on 0.074 and 0.143 against 0.130 and 0.054 unedited, intervals overlapping, and the hazard stage reads 0.111 and 0.136 against 0.111 and 0.038, with *neither* the commonest outcome throughout. These unaligned counts are the metric that the aligned, reference-compared assay of Section 5.3 replaces.

G MODEL-FAMILY RESULTS AND ROBUSTNESS

Training-arm accuracy and layout. Once the shortcut is gone, reasoning does not buy accuracy: the DIRECT arm leads by 3.6 and 6.4 points at 8,000 examples on the 1.5B base and by 4.3 and 7.7 on the 3B base, and all trace students sit within two points of one another and below the specification-only baseline, which is computed on the non-singleton subsets (Table 2, Appendix E); the 7B arm reads 0.800 and 0.723. Out of distribution the errors sit in the lateral-band family, the family the specification cannot supply and the assay scores. A terse layout drops the hand clause and both band sentences and changes nothing else: trained on it, the 1.5B arm is more accurate, 0.799 and 0.724 against 0.790 and 0.716, reconstructs the families it still carries to within a point, and its REPLACE falls from 0.735 to 0.439 $[0.399, 0.478]$. The fall is not interventions becoming invisible: textually identical pairs rise from 346 of 2,348 to 529 of 2,346, almost all offset pairs (175), and the hand and peripheral mechanisms that carry the collapse gain none. The sentences that make the reasoning intervenable are the ones that cost accuracy.

Model family. Repeating the arms on Llama-3.2-3B-Instruct and the TRACE arm on Qwen2.5-7B-Instruct with identical example ids, recipe, and seed reproduces the order of every comparison: REPLACE reads 0.726 and 0.784 against 0.735, CONSEQUENCE 0.465 against 0.419 on the same 43-pair prefix at 3B, and on the hazard chain the 3B arm clears the reproducibility reference by $+0.048$ and $+0.054$ at 8,000 examples, and the 7B arm by $+0.049$ and $+0.065$. What changes with size is not the order of the ladder but the size of the shortfall below B_2^* on hands and the appearance of the state-to-answer offset in the two larger students under different keys (Section 5.3). Table 14 gives REPLACE per mechanism for the three students; the donor controls exist for the 1.5B and 7B students only and are in Table 3.

Analysis constants. At parser tolerances of 0.05, 0.10, and 0.20 s the start-F1 of the deterministic families moves by at most 0.001. On the first 400 runs of each split, where the reproducibility reference is computed, the hazard chain reads 0.746, 0.776, and 0.823 in distribution and 0.670, 0.694, and 0.748 out of it, against references of 0.687, 0.747, and 0.842 and 0.615, 0.686, and 0.787; the margin over the reference holds at 0.05 and 0.10 s and reverses at 0.20 s. The EDIT suffix-inconsistency rate reads 0.883, 0.877, and 0.865. The onset bin is swept in Table 15: halving it to 0.125 s moves the two splits' baselines in opposite directions, which is the estimator running out of data, and doubling it moves them by at most half a point. Merging the five lane bands into three widens the error-localization gap in distribution from 0.44 to 0.59 and leaves it at 0.53 out of distribution; a seven-band scheme cannot be scored from five-band generations. Defining a moved outcome by any rung flag rather than the maximum leaves direction agreement at 0.735 and raises invariance from 0.938 to 0.959 by removing 94 pairs on which a lower rung moved; predictions change on 0.447 of those against 0.041 where nothing moved, so the published maximum-rung invariance is the stricter number.

Table 15: The specification-only baseline against the onset bin.

Onset bin	ID accuracy [CI]	OOD accuracy [CI]	Median group ID / OOD	Singletons ID / OOD
0.125 s	.833 [.806, .860]	.734 [.700, .767]	3 / 2	244 / 329
0.25 s (reported)	0.809 [.784, .833]	0.750 [.721, .780]	4 / 3	113 / 154
0.5 s	.814 [.790, .839]	.750 [.722, .779]	7 / 5	40 / 58

Monitorability. Six inconsistency flags are computed on the generated reasoning and the specification alone. On the revised-layout arm five never fire in distribution and fire on at most 1.5% of runs out of it. The sixth, a mismatch between the narrated and the specified glance duration, fires on 12.5% of in-distribution runs and on 41% of wrong answers against 5% of right ones (flag-count AUROC 0.676); it fires when a revised-layout trace stops inside a glance whose end no line wrote, which is a statement about when the run ended, and a variant that declines that one comparison fires on 0.0 to 0.4% of runs with AUROC 0.494 and 0.496. The parser was verified layout-aware on 200 example ids in both layouts. The answer's own log-probability reaches AUROC 0.948 and 0.894, an out-of-fold logistic on the flags 0.685 and 0.563, and their combination 0.948 and 0.874; selective accuracy under the log-probability reaches 0.98 at 70% coverage. These checks do not establish the reliability of arbitrary monitors.

Duration generalization. Two training splits differ only in which single-glance duration is withheld; a 3.25 s glance opens a window neither of whose endpoints occurs elsewhere in training, a 3.5 s glance opens one whose endpoints both do. Withholding a duration costs nothing: with 3.5 s withheld the DIRECT arm scores 0.557 on 3.5 s glances ($n = 1000$) and 0.571 on 3.25 s; with 3.25 s withheld it scores 0.528 on 3.5 s ($n = 36$) and 0.598 on the withheld 3.25 s ($n = 343$); the TRACE arm reads 0.458 and 0.438, and 0.417 and 0.499, respectively. Across the grid, DIRECT reads 1.00, 0.86, 0.62, 0.60, 0.53, and 0.43 at 0.5, 2, 3, 3.25, 3.5, and 4 s, while the per-cell majority reads 1.00, 0.85, 0.57, 0.54, 0.51, and 0.50 at the same points; the withheld cell sits on the curve its seen neighbours trace. The trace arm writes the withheld return instant to within one tick in 0.997 of traces. The out-of-distribution gap in the main tables is a mixture effect, since the holdout over-weights long glances whose outcomes the specification does not determine.

Table 14: Whole-trace replacement by mechanism and student family.

(a) Whole-trace replacement, direction agreement on outcome-moving pairs

Mechanism	n_{mov}	Qwen2.5-1.5B	Llama-3.2-3B	Qwen2.5-7B
brake-light cue off	22	0.727 [0.545, 0.909]	0.727 [0.545, 0.909]	0.818 [0.636, 1.000]
hands off wheel	187	0.797 [0.743, 0.850]	0.791 [0.738, 0.850]	0.882 [0.835, 0.928]
initial offset shift	5	0.200 [0.000, 0.600]	0.200 [0.000, 0.600]	0.200 [0.000, 0.667]
initial offset shift, early	43	0.419 [0.256, 0.558]	0.465 [0.326, 0.605]	0.628 [0.475, 0.766]
peripheral gain off	202	0.802 [0.743, 0.856]	0.777 [0.718, 0.837]	0.787 [0.733, 0.847]
re-acquisition cost zero	1	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]
swap: lizard takes owl time	3	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
swap: owl takes lizard time	66	0.636 [0.515, 0.743]	0.621 [0.500, 0.742]	0.667 [0.556, 0.774]

(b) Whole-trace replacement, invariance on outcome-invariant pairs

Mechanism	n_{inv}	Qwen2.5-1.5B	Llama-3.2-3B	Qwen2.5-7B
brake-light cue off	362	0.961 [0.939, 0.978]	0.992 [0.983, 1.000]	0.959 [0.938, 0.978]
hands off wheel	206	0.835 [0.786, 0.884]	0.864 [0.820, 0.913]	0.854 [0.802, 0.900]
initial offset shift	293	0.986 [0.973, 0.997]	0.990 [0.976, 1.000]	0.986 [0.973, 1.000]
initial offset shift, early	255	0.898 [0.859, 0.933]	0.910 [0.875, 0.945]	0.894 [0.856, 0.928]
peripheral gain off	192	0.885 [0.833, 0.932]	0.896 [0.849, 0.938]	0.885 [0.837, 0.926]
re-acquisition cost zero	187	0.995 [0.979, 1.000]	0.989 [0.973, 1.000]	0.984 [0.963, 1.000]
swap: lizard takes owl time	193	1.000 [1.000, 1.000]	0.990 [0.974, 1.000]	0.979 [0.958, 0.995]
swap: owl takes lizard time	131	0.916 [0.870, 0.954]	0.931 [0.893, 0.969]	0.885 [0.824, 0.937]

95% pair-bootstrap intervals. 1.5B and 3B cells are the frozen per-mechanism results of the ladder analysis; 7B cells come from the same run as Table 3 (1,000-resample pair bootstrap). The three students are scored on the same pairs, so n is shared. "swap": switch-time swap, split by which glance type receives the other's transition time.

H REPRESENTATIVE PAIRS AND PROMPTS

Three examples. These are three of the original six replacement examples and two continuations, chosen to show a followed pair, an unfollowed pair, and a continuation; the remaining examples are in the archive's trace listings. Brake-light cue removed, followed (seed 6, C1, headway 1.5 s, deceleration 4.5 m/s^2, eyes-only glance of 2.5 s): the true rung moves 0 to 5 and the 1.5B TRACE arm's prediction moves 0 to 5 under REPLACE; the traces differ in whether the hazard registers at 2.75 s on brake lights before the glance or is still a closing cue at 3.25 s. Hand kept on the wheel, not followed (seed 7, same cell, body lean of 1.5 s): the true rung moves 1 to 0 and the prediction stays at 0, already the counterfactual value. A peripheral EDIT+CONTINUE continuation (eyes-only glance of 5 s, same cell; true rung 0 to 5): the factual reasoning is cut after its third line and one clause changed from *lane edges in peripheral view* to *lane current*, and the model continues

```
[3.5] Eyes fully away from the road: lane unsampled 4.05 s; road ahead
unsampled 4.05 s; hands on wheel.
The run ends.
#ANSWER rung=5 | L3_time=6.95
```

against a reference that continues to 8 s and narrates a return, a re-acquisition, and a severe lane offset. The edited state propagates at once and the answer moves with the simulator; the consequence is not written. In the deleted-hands continuation of the same cell the model also writes *brake lights are on* on three lines, a phrase outside the verbalizer's vocabulary that the parser scores as a hallucinated span; the continuation is the model's own text, not the architecture's.

Prompt components. The prompts are factored here into a common suffix and instruction-specific prefixes only to avoid printing repeated text; the complete original listings, with their line breaks, are in the archive. Each student receives its instruction, a blank line, and the five-line specification block. Teacher generation is Qwen2.5-7B-Instruct, local, greedy.

Common student suffix:

```
#ANSWER rung=<0-5> | L3_time=<seconds or none>
rung is the highest severity level reached: 0 none, 1 a wheel past the
lane edge, 2 the vehicle centre past it, 3 a completed lane exit, 4 off
the road, 5 a collision. L3_time is when the lane exit first happened,
or none.
```

DIRECT prefix:

```
You are given a driving scenario specification. Predict the outcome.
Reply with the answer line only, in exactly this form:
```

TRACE prefix (identical across layouts):

```
You are given a driving scenario specification. Write the driver's
cognitive trace: one bracketed line per interval, narrating gaze, held
control and hazard processing as the architecture would produce them.
Then give the answer line, in exactly this form:
```

FREE prefix:

```
You are given a driving scenario specification. Explain, in 4 to 8
sentences, what the driver's cognition does during this scenario and why
it leads to the outcome. Then give the answer line, in exactly this form:
```

FREE-LONG replaces "in 4 to 8 sentences," with "in as much detail as the scenario needs," and changes nothing else.

FREE teacher system prompt:

```
You are a human-factors researcher describing how a driver's cognition
unfolds in a simulated driving scenario. You will be given a scenario
specification and the outcome that actually occurred. Write a plausible
explanation of the driver's cognition that leads to that outcome.
Rules:
```

```
- Between 4 and 8 sentences. No lists, no headings, no timestamps in
brackets.
- Write about perception, attention, what the driver can and cannot see
while looking away, held control actions, hazard detection and response
latency.
- You may refer to times that appear in the specification. Do not invent
any other numeric quantity: no lateral offsets, no speeds you were not
given, no reaction times, no distances.
- Do not state the answer line and do not write the words rung or L3.
- Plain prose. No em-dashes.
```

The user text is "Scenario specification: {spec}", a blank line, "Outcome that occurred: {outcome}", a blank line, and "Write the explanation." {outcome} is one of *the driver stayed in the lane and did not collide; a wheel crossed the lane edge; the vehicle centre crossed the lane edge; the vehicle completely left its lane; the vehicle left the road entirely; the vehicle collided*, followed, when a lane exit occurred, by *, and the lane exit first happened at* $\{t\}$ *s*. Generation is capped at 400 new tokens.

Length-matched teacher. The first rule becomes *Write the number of sentences you are asked for. No lists, no headings, no timestamps in brackets.* The requested count is the TRACE arm's own reasoning length for the same example, in the student's tokens, divided by a measured 29 tokens per sentence and clipped to $[4, 90]$; generation is capped at 2,600 new tokens, above the trace arm's maximum of 2,559. The final user instruction becomes:

```
Write the explanation in about {n} sentences. Cover the whole scenario
in order, one moment at a time, at a level of detail that fills that
length: what the driver is attending to, what is and is not visible
while the eyes are away, what the hands and the steering are doing, when
and how the hazard is noticed, and what the response is. Do not pad with
repetition.
```

Counterfactual rewrite system prompt (FREE arms only):

```
You rewrite a short account of a driver's cognition so that it describes
the same scenario under one changed assumption. Change only what the
assumption changes and its consequences. Keep the same length, the same
order and the same style. Write plain prose. Do not state an outcome, do
not write the words rung or L3, and do not add a numbered answer.
```

The user turn is the account, "Changed assumption:" followed by the premise in Table 16, and "Rewrite the account." Generation is greedy at 512 new tokens; the rewritten prose is handed back to the arm as its reasoning and the answer is generated over it. No parser verifies what the rewrite changed, and these rows support no claim in the main text.

Table 16: Mechanism premises for the teacher rewrite.

Mechanism	Premise
Peripheral	the driver gets no lane information from peripheral vision during the glance, so looking away costs the lane as completely as it costs the road ahead
Hands	the driver keeps both hands on the wheel throughout, so steering authority is never given up
Brake light	the lead vehicle's brake lights are not usable as a cue, so the hazard can only be noticed from the vehicle ahead visibly growing larger
Re-acquisition	the driver pays no re-acquisition cost on looking back, so a hazard met just after the eyes return is processed without that extra delay
Switch time	the eyes take the other glance type's time to swing away and back, so the fully blind part of the glance is a different length
Initial offset	the vehicle starts further from the centre of its lane, so every drift during the glance begins from further out

I SCOPE AND REMAINING LIMITATIONS

"True" means relative to the frozen architecture, not to human cognition or driving safety. The evidence covers one testbed, templated traces, and 1.5B to 7B instruction-model fine-tunes; only the TRACE arm was trained at 7B, and the donor controls were run at 1.5B and 7B only, so statements about them are statements about two models. The FREE teachers saw the label, so that arm says what a rationalized chain of thought does under intervention and not what free reasoning would do. Template edits are not samples from the student's conditional distribution; only the continuation is the model's own text, for the model and not for the architecture. The lane assay covers two mechanisms and one consequence family, excludes trajectories without a return opportunity, and reports conditional performance separately from coverage; hands incorporation is weak evidence because its clause is the corpus default. The re-execution references use at most twenty seeds and are empirical rules, not bounds on all specification-only predictors; the arm-wise modal reference mirrors the students' arm-wise structure but not their greedy decoder, and depends on how tied modes are read, so both readings are reported; the campaign-seed baseline is sensitive to sparse onset groups; the lane references are reported both unfiltered and under the cohort's own shared-return rule, which changes one reading (Appendix B). The brake-light and re-acquisition mechanisms move the outcome in 22 and 1 pairs, and are kept because their tiered expectation is itself a check. The lead metric counts span starts and is blind to a carried-forward state; the clause fraction is the measure of propagation. None of clause propagation, supplied-trace effects, or consequence accuracy diagnoses hidden computation: a model could compute an unspoken consequence, and a monitor reading the same text could not see it either. The results motivate controls and matched references; they do not establish their failure rate outside this setting.