Supplementary Information for Policy-TD: Teacher-Distilled Runtime Control for Frozen Small Language Models

Surya Teja Avvaru
avvarusuryateja@gmail.com

Krishna Sai Pokala
ok50250@umbc.edu

Varun Kumar Reddy Dodda
vdodda1@bowiestate.edu

Surya Teja Avvaru is an Independent Researcher and the corresponding author.
Supplementary Information - August 2026

S1. Evaluation construction and paired-row accounting

The internal heldout suite contains 100 prompts across ten families and is evaluated over three guide seeds and three frozen students, producing 900 paired runtime rows per condition. The external-style stress suite contains 300 deterministic prompts and is likewise evaluated over three seeds and three students, producing 2700 rows. The public external suite contains 400 prompts spanning SQuAD2-style answerable QA, SQuAD2-style unanswerable QA, ARC-Challenge, and GSM8K, producing 3600 rows per condition.

The evaluation unit used for helped/harmed accounting is a runtime row: the same prompt, student, and guide setting has both a baseline and a guided outcome. This pairing makes every accuracy change decomposable into recovered baseline failures and corrupted baseline successes. Because each prompt appears across multiple student/seed combinations, runtime rows are not independent prompt draws. The row-level intervals shown in the main paper are therefore descriptive paired intervals, not prompt-cluster generalization intervals.

S2. Paired metrics and intervention-quality diagnostics

Let baseline and guided correctness be $b_i, q_i \in \{0, 1\}$. Helped and harmed counts are

$$H = \sum_i \mathbf{1}[b_i = 0 \wedge q_i = 1], \qquad D = \sum_i \mathbf{1}[b_i = 1 \wedge q_i = 0],$$

and the paired commitment-control gain in percentage points is

$$\mathrm{CCG}_{\mathrm{pp}} = 100 \frac{H - D}{N}.$$

For behavior-changing interventions, let I be the number of interventions and T the number applied to baseline-incorrect rows. We distinguish targeting from outcome:

$$\mathrm{TargetingPrecision} = \frac{T}{I}, \qquad \mathrm{HelpYield} = \frac{H}{I}, \qquad \mathrm{HarmYield} = \frac{D}{I}.$$

Targeting precision is a post-hoc diagnostic because baseline correctness is not known at deployment. Help and harm yields quantify whether the chosen bounded action actually changes correctness.

Table 1: Intervention-quality decomposition from the frozen pooled result tables.

Suite	Setting	I	Targeting precision	Help yield	Harm yield
Internal heldout	Conservative	196	78.57%	73.98%	0.00%
Internal heldout	Higher-recall	198	78.79%	74.24%	0.00%
External-style stress	Conservative	190	84.21%	57.89%	0.00%
External-style stress	Higher-recall	590	94.92%	18.64%	0.00%
Public external	Conservative	543	88.95%	4.79%	1.10%
Public external	Higher-recall	572	87.94%	6.12%	2.62%

The contrast between targeting precision and help yield is a key transfer diagnostic. On the internal heldout, most conservative interventions that target a baseline failure also recover it. On public external evaluation, conservative targeting precision remains 88.95%, but help yield falls to 4.79%. Thus public transfer is limited not only by gate selectivity; many correctly targeted failures are not recoverable by the bounded action executor and frozen generator.

S3. Runtime action semantics

Table 2: Runtime action vocabulary and principal failure mode.

Action	Runtime operation	Principal risk or limitation
`finalize`	Accept the current candidate answer.	Preserves a wrong baseline answer if a failure is missed.
`say_unknown`	Abstain because the prompt lacks sufficient support.	Can harm answerable examples if over-applied.
`repair_final`	Apply a bounded correction when trace evidence supports a different final answer.	Can over-correct when the trace is unreliable.
`rollback`	Discard an unsafe or unsupported trajectory and request a constrained replacement.	Can discard useful partial reasoning.
`force_grounded_step`	Request a constrained continuation grounded in stated evidence.	Cannot create missing generator competence.
`continue`	Permit ordinary continuation without immediate override.	Can delay intervention if the state already warrants action.

S4. Seed-level runtime stability

Table 3: Internal heldout guide-seed results. Each seed is evaluated over three frozen students and 100 heldout prompts.

Seed/setting	Base acc.	Guided acc.	Delta pp	Helped	Harmed	Int. rate
seed13 conservative	31.00	47.00	16.00	48	0	21.67
seed17 conservative	31.00	47.33	16.33	49	0	22.00
seed23 conservative	31.00	47.00	16.00	48	0	21.67
seed13 higher-recall	31.00	47.33	16.33	49	0	22.00
seed17 higher-recall	31.00	47.33	16.33	49	0	22.00
seed23 higher-recall	31.00	47.33	16.33	49	0	22.00

S5. Student-level transfer

Model aliases are Qwen3 for Qwen3-1.7B, SmolLM2 for SmolLM2-1.7B-Instruct, and TinyLlama for TinyLlama-1.1B-Chat.

Table 4: Student-level behavior on internal heldout and public external evaluation.

Suite	Student	Setting	Base	Guided	Delta pp	Helped	Harmed
Internal	Qwen3	Conservative	49.00	61.00	12.00	36	0
Internal	SmolLM2	Conservative	19.00	35.33	16.33	49	0
Internal	TinyLlama	Conservative	25.00	45.00	20.00	60	0
Internal	Qwen3	Higher-recall	49.00	61.00	12.00	36	0
Internal	SmolLM2	Higher-recall	19.00	36.00	17.00	51	0
Internal	TinyLlama	Higher-recall	25.00	45.00	20.00	60	0
Public	Qwen3	Conservative	30.75	32.67	1.92	23	0
Public	SmolLM2	Conservative	8.50	8.75	0.25	3	0
Public	TinyLlama	Conservative	16.75	16.25	-0.50	0	6
Public	Qwen3	Higher-recall	30.75	32.67	1.92	26	3
Public	SmolLM2	Higher-recall	8.50	9.00	0.50	6	0
Public	TinyLlama	Higher-recall	16.75	16.00	-0.75	3	12

The student decomposition shows that public transfer depends on the competence and behavior of the frozen generator. Qwen3 transfers positively, SmolLM2 weakly positively, and TinyLlama negatively. The result is consistent with the recoverability interpretation in the main paper.

S6. Internal family-level mechanism

Table 5: Internal heldout family-level effects, pooled over seeds and students. Values are percentages.

Family	Base	Conservative	Higher-recall	Higher-recall delta pp
answer-trace consistency	36.67	36.67	36.67	0.00
conservation	50.00	50.00	50.00	0.00
entity binding	66.67	66.67	66.67	0.00
missing information	13.33	97.78	100.00	86.67
quantifier/negation	13.33	13.33	13.33	0.00
state transition	3.33	3.33	3.33	0.00
style pressure	16.67	16.67	16.67	0.00
target lock	36.67	36.67	36.67	0.00
unknownability	23.33	100.00	100.00	76.67
unsupported assumption	50.00	50.00	50.00	0.00

S7. Public external domain decomposition

Table 6: Public external results by domain, pooled over seeds and students.

Domain	Setting	Base	Guided	Delta pp	Helped	Harmed
SQuAD2 answerable	Conservative	28.00	29.11	1.11	10	0
SQuAD2 unanswerable	Conservative	9.33	10.00	0.67	6	0
ARC-Challenge	Conservative	24.33	25.11	0.78	10	3
GSM8K math	Conservative	13.00	12.67	-0.33	0	3
SQuAD2 answerable	Higher-recall	28.00	28.11	0.11	10	9
SQuAD2 unanswerable	Higher-recall	9.33	11.00	1.67	15	0
ARC-Challenge	Higher-recall	24.33	25.11	0.78	10	3
GSM8K math	Higher-recall	13.00	12.67	-0.33	0	3

S8. Label-transfer diagnostics

Label-transfer accuracy is diagnostic evidence about guide-head learning, not the primary endpoint. Exact teacher-action imitation can be imperfect even when two actions are behaviorally near-equivalent, and exact action agreement does not guarantee a successful runtime repair if the frozen student cannot complete the requested action.

Table 7: Released label-transfer diagnostics pooled by student.

Student	support	answerable	entity	logic	trace	best action	intervene
SmolLM2	48.96	42.71	78.13	73.96	32.29	36.46	50.00
TinyLlama	26.04	56.25	64.58	62.50	66.67	20.83	56.25

The frozen public label-transfer table contains SmolLM2 and TinyLlama aggregates but no Qwen3 aggregate. This supplement therefore does not infer or fabricate a Qwen3 label-transfer value. Runtime evaluation, which is the primary endpoint, includes all three students.

S9. Claim-to-artifact map

Table 8: Traceability map for the principal manuscript claims.

Manuscript claim	Public supporting artifact class
Internal heldout gain and zero observed harm	frozen pooled runtime tables; seed/student/family decompositions; manifests
Router-off returns to baseline	router-off rows in frozen condition tables
Mechanistic concentration in missing-information and unknownability families	frozen family-level runtime table
External-style positive transfer with zero observed harmed cases	external-style benchmark report, manifest, and frozen tables
Public external boundary	public benchmark report and frozen student/domain tables
Teacher-label and gate structure	public JSON schema, typed label/action contracts, validation utilities, and diagnostic label-transfer tables
Public release integrity	result manifests, checksums, tests, and release-check scripts

S10. Practical deployment criterion

For a student-domain pair (S, D), a conservative empirical policy enables guided mode only when paired validation satisfies

$$\mathrm{CCG}_{\mathrm{pp}}(S, D) \geq 0 \qquad \text{and} \qquad D_{\mathrm{harm}}(S, D) \leq \delta,$$

where δ is the tolerated harmed-case threshold. A strict policy sets $\delta = 0$. This is an empirical admissibility rule, not a correctness or safety guarantee. If the condition is not met, router-off or ordinary baseline behavior is preferred.

S11. Public release boundary

The GitHub repository is intentionally curated. It exposes the stable method contract, label schema, runtime action vocabulary, paired metrics, frozen summary evidence, benchmark reports, provenance manifests, tests, and verification utilities. It does not expose raw provider transcripts or credentials, large trace directories, local exploratory scripts, all internal orchestration, or unreleased checkpoints. The public release therefore supports inspection and traceability of aggregate evidence; full from-scratch re-execution of the original research runs requires additional frozen runtime/checkpoint assets beyond the Git repository.

S12. Author contributions

Approximate contribution shares are Surya Teja Avvaru 75%, Krishna Sai Pokala 15%, and Varun Kumar Reddy Dodda 10%. Surya Teja Avvaru is the principal contributor, and author order reflects relative contribution. Formal declarations concerning competing interests, funding, data/code availability, and ethics appear in the main manuscript.