VERDICTROUTE: QUERY-ONLY ROUTING FROM DEBATE VERDICT DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective computation succeeds when a router distinguishes inputs that are merely difficult from inputs that will improve under an additional inference step. We introduce *deliberation potential*, a supervisory abstraction built from the trajectory of an adjudicator's belief during teacher debate. VerdictRoute extracts verdict instability together with debate risk and pivotal-branch sensitivity, then distills this structured signature into a query-only Llama-3.1-8B router. Our controlled evaluation fixes the deployed self-critique action, a 18.3% validation firing-rate target, a bounded 256-token action, and $1.14\times$ average online cost; Repeated-Gain Direct-V uses 2032 stochastic action executions per item to estimate critique value. VerdictRoute reaches 64.1 average accuracy across GSM8K, MATH, TruthfulQA, and ToolVerify versus 63.3, a 0.8-point gain. It improves risk ECE from 0.071 to 0.058 and comes within 0.6 accuracy points of the matched-action oracle. Across five independent debates, verdict instability has 0.81 pairwise agreement and five-rollout consensus correlates at 0.79 with measured critique value. These results show that verdict-change trajectories can provide a useful query-only supervisory signal for bounded self-critique routing.

1 INTRODUCTION

Language models can spend extra inference compute by sampling another candidate, critiquing an initial response, invoking a verifier, or switching to a deeper reasoning mode (Snell et al., 2024; Ding et al., 2025; He et al., 2025; Tan et al., 2025). The allocation decision is consequential: another pass repairs some answers, leaves others unchanged, and can damage answers that were already correct. Initial-answer risk alone cannot separate these outcomes. A selective-compute policy therefore needs to predict *responsiveness to further reasoning*: the expected benefit of a configured intervention for the current input.

Learning responsiveness from action outcomes is statistically demanding and ties supervision to the sampled intervention. Repeated executions reduce this noise, but their labels still summarize only the designated action. Rich teacher traces offer another source of evidence; however, comparisons that simultaneously change the teacher process, student, action, or online budget cannot attribute a gain to supervision structure. We isolate this attribution problem by holding the deployed decision fixed and varying the information used to teach it.

Debate supplies competing arguments and rebuttals to inform a decision (Irving et al., 2018; Michael et al., 2023; Khan et al., 2024; Du et al., 2024). Our protocol tracks whether the judge's conclusion is stable, when its confidence erodes, and which branches redirect the decision. We name the predictable signal in these trajectories *deliberation potential*. VerdictRoute represents it as a structured signature: verdict instability is the primary routing target, while debate risk and pivotal-branch sensitivity provide auxiliary structure that makes this target predictable from the original query. The debate, graph construction, and branch scoring occur during training; deployment uses a query-only Llama-3.1-8B score to trigger one bounded self-critique action.

The primary experiment is a controlled supervision-source test. The policy rows use the Llama-3.1-8B model family, a 256-token self-critique action, the same validation firing-rate target of 18.3%, and $1.14\times$ average online cost. Repeated-Gain Direct-V provides the strongest displayed action-value control: it cross-fits expected critique gain from 2032 stochastic executions per training item. The study reports each policy's end-to-end outcome at this standardized deployment interface.

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.

The resulting evidence connects target quality and policy quality. VerdictRoute achieves 64.1 four-task average accuracy versus 63.3 for Repeated-Gain Direct-V and closes to 0.6 points of the matched-action oracle. Five independent debates produce 0.81 pairwise agreement on verdict instability; consensus increases correlation with measured critique value from 0.71 to 0.79. These target and policy measurements connect the trajectory-derived label to bounded self-critique allocation.

The paper delivers three connected advances. It defines deliberation potential from temporal changes in adjudicated belief and compiles the resulting signature into a query-only compute router. It introduces a fixed-action comparison that holds the deployed action, firing-rate procedure, and online budget constant while testing debate trajectories against direct action-value labels. Finally, teacher-resampling stability, per-task results, pooled evaluation counts, and an oracle decomposition connect verdict dynamics to improved compute allocation. Together, the method and evidence make teacher belief trajectories a useful supervisory interface for bounded student inference.

2 Related Work

Selective prediction and calibration estimate answer risk for abstention or deferral (Geifman & El-Yaniv, 2017; Jiang et al., 2021; Kadavath et al., 2022; Angelopoulos & Bates, 2021), while adaptive test-time-compute systems route among sampling budgets, reasoning modes, or models (Snell et al., 2024; Ding et al., 2025; He et al., 2025; Tan et al., 2025). These methods establish the importance of allocation, but risk and responsiveness are distinct targets. VerdictRoute predicts the latter from judge-belief dynamics and evaluates it with the student, action, firing rate, and budget held constant.

Self-refinement and reflection define interventions that can repair an answer (Madaan et al., 2023; Shinn et al., 2023). Their observed outcome gains naturally support action-conditioned value targets. Our Direct-V and Repeated-Gain Direct-V controls instantiate this approach for exactly the deployed action, including a repeated stochastic estimator of critique gain. VerdictRoute instead learns from the evolution of a teacher verdict and tests whether trajectory supervision improves routing for that fixed action.

Debate uses competing arguments and adjudication to improve supervision or final decisions (Irving et al., 2018; Michael et al., 2023; Khan et al., 2024; Du et al., 2024). Broader scalable-oversight work studies reward modeling and model assistance to human evaluators (Leike et al., 2018; Bowman et al., 2022). Process supervision extracts intermediate evidence from reasoning traces (Uesato et al., 2022; Lightman et al., 2023), and debate distillation trains students from multi-agent feedback (Zhou et al., 2025). VerdictRoute makes the temporal evolution of the adjudicated belief the supervisory object. Its matched-deployment protocol separates the information in that object from changes to the student policy or inference budget within the specified self-critique action.

3 Deliberation-Potential Supervision from Verdict Dynamics

For an input x, a policy selects between returning a base answer and executing one configured action a with incremental cost c_a. The controlled comparison fixes the student S, action a, firing rate q, gate family H, and online budget B; only the offline supervision source $z \in \{\text{debate}, \text{non-debate}\}$ changes. We evaluate $\pi_z(x) = \mathbb{1}[h_z(x) > \tau_z]$ with each threshold calibrated to the same validation firing rate. This definition makes the supervision effect identifiable under the deployed decision.

3.1 Debate Trees and Argument Graphs

For question x, a teacher produces a debate tree $\mathcal{T}(x)$ of K argumentative turns from two debaters D_{PRO}, D_{CON} and a judge J. The debaters defend answer candidates a_{PRO} and a_{CON}, respectively. The tree's representation is the directed attributed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{r}, \boldsymbol{\delta})$: its argument nodes number $|\mathcal{V}| = K$, each carrying an embedding $\mathbf{h}_i \in \mathbb{R}^d$ from a frozen encoder; its directed edges have types SUPPORT, REBUTTAL, or JUDGEUPDATE; roles are assigned by $\mathbf{r} : \mathcal{V} \to \{\text{PRO}, \text{CON}, \text{JUDGE}\}$; and $\boldsymbol{\delta} : \mathcal{V}_{\text{JUDGE}} \to [-1, 1]$ records judge confidence states ($+1$ indicates full confidence in pro, -1 in con). At a judge-update node, $P_J(a_{\text{PRO}} \mid \mathcal{T})$ is the judge's normalized binary probability for the pro candidate, with $P_J(a_{\text{CON}} \mid \mathcal{T}) = 1 - P_J(a_{\text{PRO}} \mid \mathcal{T})$. The *verdict margin* is $m(\mathcal{T}) = P_J(a_{\text{PRO}} \mid \mathcal{T}) - 0.5$.

3.2 A DELIBERATION-POTENTIAL SIGNATURE

The ideal routing quantity for a configured action is

$$V_a^*(x) = \mathbb{E}[r(y_a, y^*) - r(y_0, y^*) \mid x]\,. \tag{1}$$

VerdictRoute distills $z^*(x) = (e^*(x), r^*(x), \Phi^*(x))$, the *deliberation-potential signature*, from teacher debates. Its components describe three aspects of responsiveness to additional reasoning: e^* records whether new arguments materially redirect adjudication, r^* measures competitive uncertainty, and Φ^* measures concentration of influence in pivotal branches. VerdictRoute uses e^* as the routing target and the remaining components as structured auxiliary supervision that makes this target predictable from the query. On a stratified 1,200-item held-out slice, e^* correlates with measured self-critique value at Spearman $\rho = 0.71$, compared with 0.48 for r^* and 0.41 for Φ^*. This alignment motivates using verdict instability as the action-allocation target.

A further target captures risk through a soft label derived from debate. For input x, inter-debater disagreement is $\sigma(x) = 1 - \left|P_{D_{\text{PRO}}}(a_{\text{PRO}}) - P_{D_{\text{CON}}}(a_{\text{CON}})\right| \in [0, 1]$: each probability gives the softmax mass that the respective debater assigns to its own defended answer at temperature 1.0; the *judge-margin deficit* is $\mu(x) = 1 - |m(\mathcal{T}(x))| \in [0.5, 1]$. The risk target is the soft debate-derived label

$$r^*(x) \;=\; \alpha\,\sigma(x) \;+\; (1-\alpha)\,\mu(x), \qquad \alpha = 0.6, \tag{2}$$

selected on the 10% validation split. This target lets the student learn when the adjudicator's margin is unreliable even though the teacher tree is absent online.

Branch sensitivity supplies another auxiliary target. For a subtree $b \subseteq \mathcal{V}$ with root node v_b, we ask the judge to score the tree after pruning, $\mathcal{T} \setminus b$, and define *branch-drop influence* as

$$\phi(b, \mathcal{T}) = |\, m(\mathcal{T}) - m(\mathcal{T} \setminus b)\,|, \tag{3}$$

When ϕ is high, removing that branch would have materially changed the verdict. Aggregating across branches gives verdict sensitivity $\Phi(x) = \max_{b \in \mathcal{T}(x)} \phi(b, \mathcal{T}(x))$. Exact evaluation requires another judge run for each dropped branch; we therefore compute ϕ for 25% of the debate corpus chosen at random (3,000 trees), using the offline labeler in Section 4.1 to impute the remaining values. It serves as an auxiliary denoising signal; Appendix D quantifies its fidelity to exact and continuation-aware counterfactuals.

Finally, we define the verdict-instability target. Let R be the number of judge rounds and $m_k(\mathcal{T})$ the judge margin after round k. The target aggregates material changes across the judge trajectory:

$$e^*(x) \;=\; \mathbb{1}\big[\exists\, k \in \{2, \ldots, R\} :\; \operatorname{sign}\big(m_k\big) \neq \operatorname{sign}\big(m_{k-1}\big) \;\text{ or }\; |m_k| - |m_{k-1}| \leq -0.15\big]\,. \tag{4}$$

Thus, $e^*(x) = 1$ iff a subsequent round flips the judge's argmax verdict or shrinks its margin by at least 0.15. The threshold 0.15 is selected on validation so that the $e^* = 1$ base rate matches the configured 18.3% deployment firing rate. Unlike a label from one sampled action outcome, e^* aggregates the adjudication trajectory and captures whether later reasons redirect or materially weaken the current conclusion.

3.3 STABILITY AND VALUE ALIGNMENT

We sample five independent GPT-4o debate trees for each item in the held-out value-validation slice, producing 6000 debate realizations over $1{,}200$ inputs. Table 1 reports the displayed repeatability statistics and their alignment with measured action value. The verdict-instability positive rate is 18.6% (across-rollout SD 1.2 percentage points) and pairwise agreement is 0.81. The corresponding nominal chance-corrected agreement is $\kappa = (0.81 - [0.186^2 + 0.814^2])/(1 - [0.186^2 + 0.814^2]) = 0.37$. Consensus increases correlation with measured critique value from 0.71 for one rollout to 0.79 for five, while independent single-rollout label sets yield 64.0 ± 0.3 average accuracy and a 1.1-point gain over critique-gain Direct-V. The trajectory-derived target is repeatable across teacher samples and consistently supports the student-side routing advantage.

4 COMPILING DELIBERATION POTENTIAL INTO A QUERY-ONLY ROUTER

VerdictRoute compiles teacher-side deliberation into a student-side routing score. Debate trees first produce a deliberation-potential signature; the student then predicts that signature from the original

Table 1: Stability of deliberation-potential supervision across five independent teacher debates on 1,200 held-out items. Agreement measures repeatability of verdict instability; rank statistics measure stability of auxiliary signals; value correlation connects the trajectory-derived target to measured critique utility. The positive-rate $\pm$ term is the across-rollout standard deviation in percentage points.

Quantity	Estimate	Support
e^* positive rate	$18.6 \pm 1.2\%$	Across-rollout SD in percentage points
e^* pairwise agreement	0.81	9720/12000 pairwise decisions
Nominal chance-corrected agreement	0.37	Derived from positive rate and pairwise agreement
Rank stability: r^* / Φ^*	0.74 / 0.68	Five-rollout summary
Consensus e^* vs. measured value	0.79	Five-rollout summary
Single-rollout policy average	64.0 ± 0.3	vs. critique-gain Direct-V: $+1.1$

query and converts its verdict-instability component into an escalation score for the configured self-critique action. Flat-Transcript, Direct-V, and Repeated-Gain Direct-V provide alternative supervision signals for the fixed-action comparison.

4.1 Offline target materialization

Frozen text embeddings, role tags, and judge-confidence updates parameterize a two-layer edge-typed graph encoder (Kipf & Welling, 2017; Veličković et al., 2018). On the 25% exact branch-drop subset, an MLP regresses branch influence from pooled tree and branch-root states, reaching held-out $R^2 = 0.71$; it then imputes Φ^* for the remaining teacher examples. Once (e^*, r^*, Φ^*) are materialized, this labeler is discarded. Deployment consumes only the query.

4.2 Query-only tuple prediction

At deployment, the pretrained language model Llama-3.1-8B serves as student S_θ, with only the original query x as input. Answer generation uses its standard language-modeling head; three additional lightweight auxiliary heads share $\mathbf{h}_q$, the query representation given by the final hidden state of the prompt's terminal token.

We fine-tune the standard language modeling head to generate the final answer supplied by the teacher:

$$\mathcal{L}_{\text{ANS}} = -\sum_{t=1}^{T} \log P_{S_\theta}(y_t^* \mid y_{<t}^*, x). \tag{5}$$

The risk head applies a linear logit $\tilde{h}_{\text{RISK}}$ to $\mathbf{h}_q$ and returns $h_{\text{RISK}} = \text{sigmoid}(\tilde{h}_{\text{RISK}}) \in (0, 1)$. Training minimizes mean-squared error relative to $r^*(x)$, the soft target derived from debate in Eq. 2, together with a small auxiliary BCE term for teacher–student answer disagreement:

$$\mathcal{L}_{\text{RISK}} \;=\; \big(h_{\text{RISK}}(\mathbf{h}_q) - r^*(x)\big)^2 \;+\; 0.2\,\text{BCE}\big(h_{\text{RISK}}(\mathbf{h}_q), \mathbb{1}[y_{\text{TEACHER}} \neq y^{\text{greedy}}_{\text{STUDENT}}]\big). \tag{6}$$

The first term transfers debate-internal uncertainty; the second grounds it in teacher–student answer disagreement before the estimate enters the escalation head.

The sensitivity head h_{SENS} is a two-layer MLP on $\mathbf{h}_q$ that regresses the aggregate teacher sensitivity target $\Phi^*(x)$:

$$\mathcal{L}_{\text{SENS}} \;=\; \big(h_{\text{SENS}}(\mathbf{h}_q) - \Phi^*(x)\big)^2. \tag{7}$$

This auxiliary head predicts from the query whether the teacher tree would contain a pivotal branch.

The escalation head applies a two-layer MLP logit $\tilde{h}_{\text{ESC}}$ to the concatenation of the query representation, predicted risk, and predicted sensitivity, $[\mathbf{h}_q;\, h_{\text{RISK}}(\mathbf{h}_q);\, h_{\text{SENS}}(\mathbf{h}_q)]$, and returns $h_{\text{ESC}} =$

$\mathrm{sigmoid}(\tilde{h}_{\text{ESC}})$. It is trained with BCE against the verdict-instability label from Eq. 4:

$$\mathcal{L}_{\text{ESC}} = \text{BCE}\Big(h_{\text{ESC}}\big([\mathbf{h}_q;\, h_{\text{RISK}}(\mathbf{h}_q);\, h_{\text{SENS}}(\mathbf{h}_q)]\big),\; e^*(x)\Big). \tag{8}$$

This is the deployment head. Every input is student-side, preserving the $1.14\times$ runtime boundary.

4.3 Training procedure

Training proceeds in two stages. First, the branch-drop subset is used to fit the offline graph labeler, after which we compute $\Phi^*(x)$ for every teacher example. The second stage trains the deployed student by combining the objectives as

$$\mathcal{L}_{\text{STUDENT}} = \mathcal{L}_{\text{ANS}} + \lambda_1 \mathcal{L}_{\text{RISK}} + \lambda_2 \mathcal{L}_{\text{SENS}} + \lambda_3 \mathcal{L}_{\text{ESC}}, \tag{9}$$

with $\lambda_1 = 0.3$, $\lambda_2 = 0.5$, $\lambda_3 = 0.2$ selected via validation on a held-out 10% split. Three phases stabilize training: Phase 1 (answer distillation, 3 epochs) uses $\mathcal{L}_{\text{ANS}}$ to train the language-modeling head only; Phase 2 (risk and sensitivity, 5 epochs) introduces $\mathcal{L}_{\text{RISK}}$ and $\mathcal{L}_{\text{SENS}}$, with answer-model fine-tuning continuing at a learning rate of $0.1\times$; Phase 3 (escalation policy, 3 epochs) introduces $\mathcal{L}_{\text{ESC}}$ and uses adversarial validation to disrupt correlations linking escalation decisions to surface features.

4.4 Inference policy

Using only the query at inference, the student initially produces a base answer along with scores for risk, sensitivity, and escalation. If $h_{\text{ESC}}(x) > \tau_{\text{ESC}} = 0.55$, the student takes *one* lightweight action: it regenerates the answer in a self-critique pass conditioned on a fixed critique prompt; reuse of the base generation's KV cache limits compute to the critique prefix and 256 newly generated output tokens. For Llama-3.1-8B, measurements use a single A100 80 GB in bf16 with vLLM $0.5.3$, CUDA 12.2 and driver $545.23.08$; batch size is 1, the prompt has 128 tokens, and output length is 256 tokens: the A100 wall time per self-critique pass is 634 ± 11 ms (0.62 A100-sec), which is $0.78\times$ the base student's full prefill+decode time of 812 ± 14 ms (0.80 A100-sec). A100-sec measures active accelerator time reported by the profiler for one A100 80 GB; runtime and invocation overhead are also counted in wall-clock latency. Escalation fires on 18.3% of inputs, giving an average inference cost of $1.0 + 0.183 \times 0.78 = 1.143\times$ (reported as $1.14\times$ throughout) and an average tokens/query of $256 + 0.183 \times 256 = 302.8$ (reported as 303).

The full pseudo-code (Algorithm 1) is in Appendix B.

5 Evaluating Fixed-Action Compute Routing

5.1 Setup

GPT-4o (OpenAI, 2024) supplies both debaters and the judge for the reported three-to-five-round debates, with a $2{,}048$-token cap per turn. The primary student architecture is Llama-3.1-8B (Grattafiori et al., 2024). We construct $12{,}000$ training trees from GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), a category-balanced truthfulness pool based on the 38 TruthfulQA categories (Lin et al., 2022), and ToolVerify: $4{,}000/3{,}000/2{,}500/2{,}500$ trees, respectively. The truthfulness pool is disjoint from the 817 evaluation questions. ToolVerify contains $2{,}500$ executable problems split by problem hash into $2{,}000/250/250$ train/validation/test items; its training allocation comprises one tree for every training item plus a second independent tree for 500 stratified training items. Evaluation uses the official GSM8K and MATH tests, TruthfulQA's 817-item MC1 set, and the held-out 250-item ToolVerify test split.

The five-seed policy comparison uses the Llama-3.1-8B model family, the configured self-critique action, and validation thresholds targeting a common 18.3% firing rate. The reported metric is each resulting policy's end-to-end outcome. GSM8K and MATH use normalized exact match, TruthfulQA uses MC1, and ToolVerify requires the predicted invocation to pass its deterministic executable verifier. Risk-head ECE uses 15 equal-mass bins and measures risk-score calibration, separately from end-to-end policy calibration and escalation-rate stability. For each seed, the four-task point estimate is the unweighted arithmetic mean of the four per-task accuracies; the table reports the mean of these seed-level estimates where stated.

The direct-value controls estimate the value of the identical self-critique action. Direct-V teacher labels average six independent $k = 8$ GPT-4o self-consistency groups, using one response in each group as its CoT reference. Direct-V student labels record whether one sampled critique repairs the base answer. Repeated-Gain Direct-V estimates expected critique gain from repeated stochastic executions and cross-fits the resulting estimator. Each policy uses the same action, threshold-calibration procedure, and online operating target as VerdictRoute. Appendix E provides the target construction and deployment map.

Training uses AdamW (Loshchilov & Hutter, 2019) (learning rate 2×10^{-5}, cosine decay, weight decay 0.01, batch size 32, gradient clip 1.0) with DeepSpeed ZeRO-2 (Rasley et al., 2020) on $8\times$A100 80GB GPUs (NVLink, CUDA 12.2, driver $545.23.08$, PyTorch $2.3.1$, bf16, FlashAttention-2). The three phases take $6.1/8.0/4.1$ hours. Evaluation uses vLLM $0.5.3$, batch size 1, a 128-token prompt, and up to 256 generated tokens.

5.2 Self-critique routing against action-conditioned value

Table 2 tests deliberation potential within one designated action. Every row uses the same self-critique action and validation firing-rate target; all four task columns are shown, so each average is directly reconstructible. Repeated-Gain Direct-V is the strongest displayed action-conditioned control and spends 2032 critiques per item estimating expected gain.

Table 2: Fixed-action selective computation. Rows use the Llama-3.1-8B model family, the same self-critique action, and a validation firing-rate target of 18.3%; supervision signals vary by row. Accuracy and average are percentages over five seeds; parenthesized values are seed standard deviations where reported. Wrong-Answer Hit Rate at the Fixed Gate is $P(\text{base wrong} \mid h_{\text{ESC}}(x) > \tau_{\text{ESC}})$ over deployed escalations.

Gate supervision	GSM	MATH	TQA	TV	Avg (SD)	ECE	Wrong-answer hit
Predictive entropy	77.2	45.1	57.8	66.7	61.7	0.091	0.66
Logit margin	77.0	44.8	57.5	66.3	61.4	0.098	0.64
Conformal defer	77.6	45.4	58.2	66.8	62.0	0.079	0.67
PRM-triggered	77.8	45.8	58.5	67.5	62.4 (0.31)	0.085	0.68
Flat transcript	77.4	45.3	58.0	66.9	61.9	0.094	0.65
Direct-V teacher	78.0	46.2	58.8	67.8	62.7	0.081	0.69
Direct-V critique gain	78.2	46.4	58.9	68.1	62.9 (0.28)	0.076	0.70
Repeated-Gain Direct-V	78.6	46.9	59.3	68.4	63.3 (0.25)	0.071	0.72
VerdictRoute	79.4	47.7	60.0	69.3	64.1 (0.26)	0.058	0.74

At $1.14\times$ average cost, VerdictRoute improves the four-task average from 63.3 to 64.1 over Repeated-Gain Direct-V, a 0.8-point gain. It also leads critique-gain Direct-V by $+1.2$ points and PRM-triggered escalation by $+1.7$. The ordering holds on every task, the fixed gate targets more initially wrong answers, and risk-head ECE improves from 0.071 to 0.058. The ECE improvement concerns the risk head; it does not by itself establish a fixed escalation rate under a shifted task mixture. Table 4 reports pooled correct/total counts for the exact five runs.

Figure 1 localizes the improvement to allocation. The verdict-instability precision–recall curve separates most strongly where missed escalations are costly. The figure's e^* precision and Table 2's wrong-answer hit rate condition on different events; the threshold sweep shows that VerdictRoute's selectivity advantage persists beyond the validation-fixed deployment point.

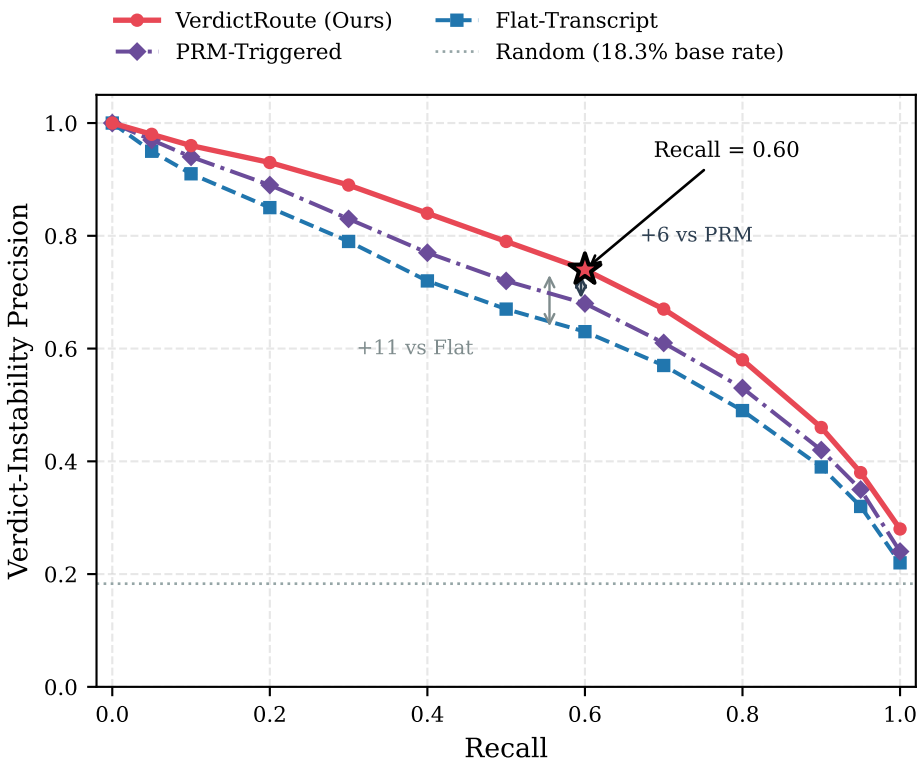


Figure 1: Verdict-instability selectivity while the score threshold varies. Precision is $P(e^* = 1 \mid h_{\text{ESC}}(x) > t)$; at recall 0.60, VerdictRoute reaches 0.74 versus 0.68 for PRM and 0.63 for Flat-Transcript.

5.3 Oracle headroom and deployment cost

The matched-action oracle reaches 64.7, leaving 0.6 points of gate-side headroom. Relative to each baseline's own oracle gap, VerdictRoute captures 57.1% of the lift beyond Repeated-Gain Direct-V, 66.7% beyond critique-gain Direct-V, and 74% beyond PRM. The policies tie at 1.00×, and the ordering persists from 1.07× to 1.50×. Appendix G provides the complete sweep.

The deployed path remains bounded and directly measured. On one A100 80GB GPU, VerdictRoute averages 928 ms and 303 tokens per query with 1,446 ms p95 latency; arbitrary query content can trigger at most one configured action. Appendix C reports the measured primitives.

6 Conclusion

VerdictRoute demonstrates that the evolution of a teacher verdict can supervise where student inference compute has value. With a bounded self-critique action, a validation-fixed firing-rate target, and 1.14× average cost, it reaches 64.1 average accuracy versus 63.3 for Repeated-Gain Direct-V and approaches within 0.6 points of the matched-action oracle. Teacher-rollout agreement and direct correlation with measured action value ground these policy gains in stable belief-change trajectories. The evidence supports a general design principle for selective computation: retain how informed judgments move as reasons accumulate, then distill that trajectory into a query-only score. VerdictRoute turns this principle into a practical interface between rich teacher deliberation and bounded student inference.

AI USE STATEMENT

Generative AI was used as the teacher model for the debate construction described in Section 5 and for manuscript writing, editing, and presentation. The authors reviewed all AI-assisted material and remain responsible for the final content.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv preprint arXiv:2107.07511*, 2021. URL https://arxiv.org/abs/2107.07511.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. Measuring Progress on Scalable Oversight for Large Language Models. *arXiv preprint arXiv:2211.03540*, 2022. URL https://arxiv.org/abs/2211.03540.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Dujian Ding, Ankur Mallick, Shaokun Zhang, Chi Wang, Daniel Madrigal, Mirian Del Carmen Hipolito Garcia, Menglin Xia, Laks V. S. Lakshmanan, Qingyun Wu, and Victor Rühle. BEST-Route: Adaptive LLM Routing with Test-Time Optimal Compute. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 13870–13884, 2025. URL https://proceedings.mlr.press/v267/ding25d.html.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving Factuality and Reasoning in Language Models through Multiagent Debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 11733–11763, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Yonatan Geifman and Ran El-Yaniv. Selective Classification for Deep Neural Networks. *Advances in Neural Information Processing Systems*, 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Yang He, Xiao Ding, Bibo Cai, Yufei Zhang, Kai Xiong, Zhouhao Sun, Bing Qin, and Ting Liu. Self-Route: Automatic Mode Switching via Capability Estimation for Efficient Reasoning. *arXiv preprint arXiv:2505.20664*, 2025. URL https://arxiv.org/abs/2505.20664.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018. URL https://arxiv.org/abs/1805.00899.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with More Persuasive LLMs Leads to More Truthful Answers. *arXiv preprint arXiv:2402.06782*, 2024. URL https://arxiv.org/abs/2402.06782.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1609.02907.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018. URL https://arxiv.org/abs/1811.07871.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *Proceedings of the 7th International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R. Bowman. Debate Helps Supervise Unreliable Experts. *arXiv preprint arXiv:2311.08702*, 2023. URL https://arxiv.org/abs/2311.08702.

OpenAI. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 3505–3506, 2020. doi: 10.1145/3394486.3406703. URL https://doi.org/10.1145/3394486.3406703.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. *arXiv preprint arXiv:2408.03314*, 2024. URL https://arxiv.org/abs/2408.03314.

Zhendong Tan, Xingjun Zhang, Chaoyi Hu, Yancheng Pan, and Shaoxun Wang. Adaptive Rectification Sampling for Test-Time Compute Scaling. *arXiv preprint arXiv:2504.01317*, 2025. URL https://arxiv.org/abs/2504.01317.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. URL https://arxiv.org/abs/2211.14275.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1710.10903.

Xiaofeng Zhou, Heyan Huang, and Lizi Liao. Debate, Reflect, and Distill: Multi-Agent Feedback with Tree-Structured Preference Optimization for Efficient Language Model Enhancement. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9122–9137, 2025. doi: 10.18653/v1/2025.findings-acl.475. URL https://aclanthology.org/2025.findings-acl.475/.

A Additional Results and Protocols

The appendix provides target-stability, fixed-action control, deployment, branch-fidelity, shortcut-control, and budget-sweep detail.

Table 3: Oracle decomposition at the $1.14\times$ operating budget. Oracle escalation changes only gate selection and chooses the same number of escalations with hindsight.

Policy	GSM	MATH	TQA	TV	Avg	ECE	Wrong-answer hit
PRM-Triggered	77.8	45.8	58.5	67.5	62.4	0.085	0.68
Direct-V critique gain	78.2	46.4	58.9	68.1	62.9	0.076	0.70
Repeated-Gain Direct-V	78.6	46.9	59.3	68.4	63.3	0.071	0.72
VerdictRoute	79.4	47.7	60.0	69.3	64.1	0.058	0.74
Oracle	80.1	48.5	60.8	69.5	64.7	0.044	0.82
Lift share over Repeated-Gain / Direct-V / PRM					$57.1/66.7/74.0\%$		

A.1 Pooled counts for the fixed-action comparison

Table 4: Pooled correct/total counts over the five runs used in Table 2. Fixed denominators are 6,595 / 25,000 / 4,085 / 1,250 for GSM8K / MATH / TruthfulQA / ToolVerify.

Policy	GSM8K	MATH	TruthfulQA	ToolVerify
PRM-triggered	5131/6595	11450/25000	2390/4085	844/1250
Direct-V critique gain	5157/6595	11600/25000	2406/4085	851/1250
Repeated-Gain Direct-V	5184/6595	11725/25000	2422/4085	855/1250
VerdictRoute	5236/6595	11925/25000	2451/4085	866/1250

A.2 Deliberation-potential stability under teacher resampling

Figure 2 expands Table 1. For $1/3/5$ rollouts per item, correlation with measured action value is $0.71/0.76/0.79$, downstream policy average is 64.0/64.2/64.3, and positive-rate dispersion is 1.2/0.8/0.5 points. Aggregation steadily reduces teacher-sampling variance while preserving the single-rollout policy result.

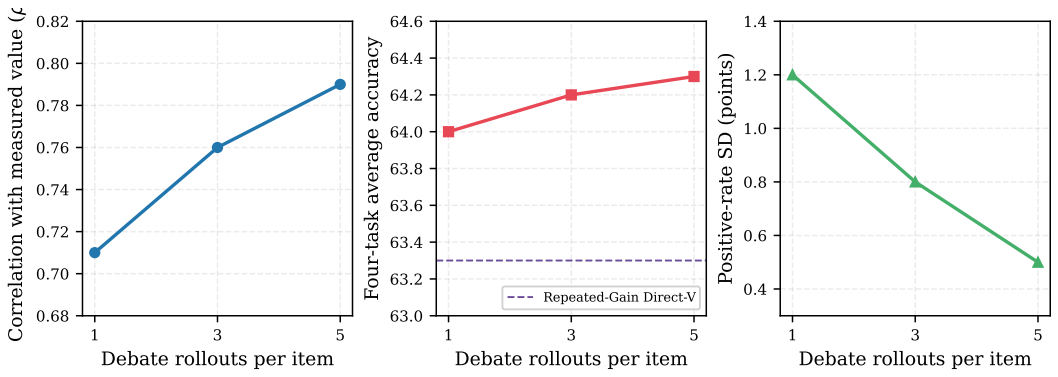


Figure 2: Deliberation-potential stability as independent teacher debates are aggregated. Left: rank correlation with measured action value. Center: four-task policy accuracy, with Repeated-Gain Direct-V as an action-conditioned reference. Right: across-rollout standard deviation of the e^* positive rate.

B VERDICTROUTE TRAINING AND DEPLOYMENT ALGORITHM

Algorithm 1 lists the full offline-label-extraction and three-phase student training procedure referenced in Section 4.

Algorithm 1 Offline target construction and three-phase student training. The graph labeler is used only before deployment; the deployed student sees only the query and may run one self-critique pass.

Require: Debate corpus $\mathcal{D} = \{(x_i, \mathcal{T}_i)\}_{i=1}^{N}$, student S_θ, loss weights $\lambda_1, \lambda_2, \lambda_3$
Ensure: Trained student S_θ with answer model and heads $h_{\text{RISK}}, h_{\text{SENS}}, h_{\text{ESC}}$

Offline preprocessing: extract the deliberation-potential signature
for each $(x_i, \mathcal{T}_i) \in \mathcal{D}$ **do**
  Construct argument graph $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i, \mathbf{r}_i, \boldsymbol{\delta}_i)$
  Compute node embeddings $\{\mathbf{h}_j\}_{j=1}^{K_i}$ via frozen encoder
  Extract disagreement $\sigma(x_i)$, verdict margin $m(\mathcal{T}_i)$
end for
Fit the graph labeler by computing branch-drop influence $\phi(b, \mathcal{T})$ on a 25% random sample and regressing the predictor on that subset
Compute risk target $r^*(x_i)$ (Eq. 2), aggregate sensitivity target $\Phi^*(x_i)$, and verdict-instability label $e^*(x_i)$ (Eq. 4)
Phase 1: Answer distillation (3 epochs)
Train the language-modeling head with $\mathcal{L}_{\text{ANS}}$ (Eq. 5)
Phase 2: Risk and sensitivity distillation (5 epochs)
Train $h_{\text{RISK}}, h_{\text{SENS}}$ with $\mathcal{L}_{\text{ANS}} + \lambda_1 \mathcal{L}_{\text{RISK}} + \lambda_2 \mathcal{L}_{\text{SENS}}$
Phase 3: Escalation training (3 epochs)
Train h_{ESC} with full $\mathcal{L}_{\text{STUDENT}}$ (Eq. 9), adversarial validation on surface features
Inference
for test input x **do**
  Generate answer $\hat{y} \leftarrow S_\theta(x)$, risk $\hat{c} \leftarrow h_{\text{RISK}}(x)$, sensitivity $\hat{s} \leftarrow h_{\text{SENS}}(x)$, escalation $\hat{e} \leftarrow h_{\text{ESC}}(x)$
  if $\hat{e} > \tau_{\text{ESC}}$ **then**
    $\hat{y} \leftarrow \text{SelfCritique}(S_\theta, x, \hat{y})$
  end if
  return $\hat{y}, \hat{c}, \hat{s}$
end for

Algorithm 1 makes the training–deployment boundary explicit: every debate-tree computation precedes online use, and inference consists only of the student plus its bounded action. Section 5 and Appendix C provide the reported student-side measurements.

C DEPLOYMENT TRADEOFFS (FULL TABLE)

Table 5: Student-side deployment tradeoffs measured on one A100 80GB.

Method	Avg acc.	Esc. rate	Avg ms	$p95$ ms	Tokens
Answer-Only	54.9	0.0	812	843	256
CoT Distill	59.2	0.0	1,218	1,261	384
VerdictRoute (Ours)	64.1	18.3	928	1,446	303
Self-Cons. ($k = 5$)	58.4	—	4,060	4,215	1,280

Table 5 reports the student-side operating context for the fixed-action result. VerdictRoute remains close to answer-only latency while retaining its accuracy gain. The latency tail is bounded by a single extra self-critique pass, and the paired average/$p95$ measurements make that bound directly inspectable.

D BRANCH-DROP FIDELITY DETAIL

Table 6 reports branch-drop validation: Spearman ρ and downstream four-task average accuracy when each reference or proxy replaces branch-drop as the sensitivity-head supervision. Table 7 reports Spearman rank correlation, mean absolute error, and systematic bias between the predicted influence $\hat{\phi}_\psi$ and exact branch-drop evaluation on small trees. Negative bias reflects underestimation of rebuttal-node influence.

Table 6: Branch-drop validation against alternative supervision signals. Spearman ρ is averaged over tree sizes 4–10 and Downstream Avg is the resulting four-task student accuracy.

Reference / Proxy	Spearman ρ	Downstream avg
Exact branch-drop (ours)	1.00	64.1
Continuation-aware resampled counterfactual	0.83	64.3
Judge-margin Δ at branch root (heuristic)	0.61	62.7
Student GAT attention weight	0.48	62.2
Node depth in rebuttal chain	0.34	61.8

Table 7: Branch-drop predictor fidelity on small trees where exact counterfactual evaluation is feasible. Negative bias indicates underestimation.

Tree Size	Spearman ρ	Mean Abs. Error	Systematic Bias
4-node	0.81	0.043	-0.018
6-node	0.78	0.056	-0.024
8-node	0.76	0.068	-0.031
10-node	0.73	0.082	-0.039

Together, Table 6, Table 7, and Figure 3 establish branch-drop as a useful auxiliary target. The continuation-aware reference improves downstream accuracy by 0.2 points, while heuristic proxies lose both fidelity and student accuracy. This evidence supports the target's denoising role without assigning it the primary decision signal carried by e^*.

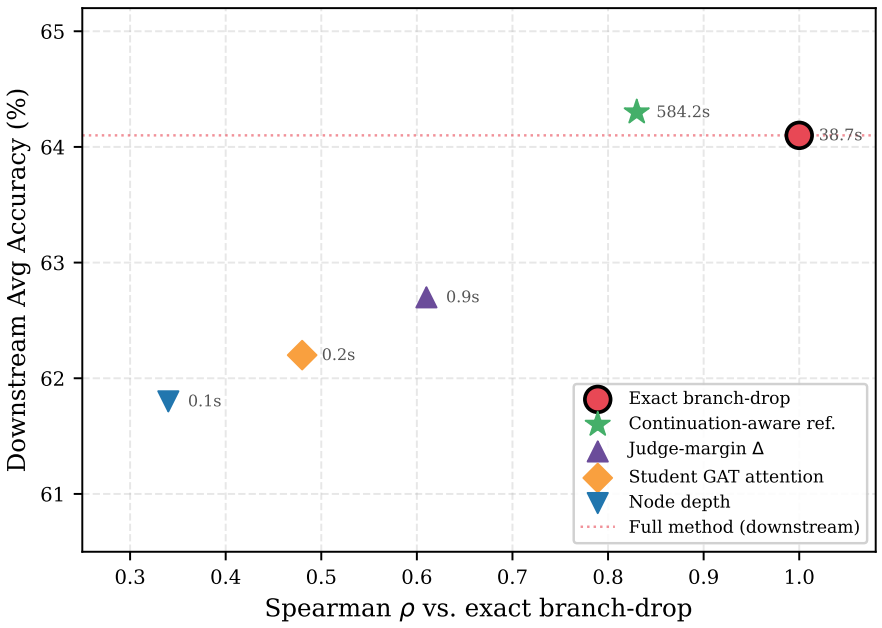


Figure 3: Rank agreement with exact branch-drop versus downstream four-task accuracy.

E Dataset Construction and Action-Conditioned Value Targets

ToolVerify is a fixed suite of 2,500 tool-use verification problems. Each natural-language task is paired with a canonical JSON invocation and a deterministic verifier executing in Python 3.11 with a fixed stub registry, disabled network access, and scratch-directory-only file access. Correctness requires exit status zero and the expected stdout hash. Problem-hash splitting assigns 2,000 items to training, 250 to validation, and 250 to test. The 2,500 ToolVerify training trees use all 2,000 training items once and a category-stratified subset of 500 items a second time with independent debate sampling; validation and test items never enter teacher construction. Table 2 reports every per-task cell, and Table 4 gives the corresponding pooled test counts.

TruthfulQA teacher-pool accounting. The 817 released questions are reserved for MC1 evaluation. For teacher construction, we generate 2,500 category-balanced questions independently from the benchmark's 38 truthfulness categories and filter candidates against the evaluation set by normalized-string identity and MinHash similarity above 0.8. These questions produce teacher debates and the same (r^*, Φ^*, e^*) targets used for the other training pools; no evaluation question enters target construction. All TruthfulQA numbers in the main tables are computed on the official 817-question set under MC1 scoring.

Table 8: Direct-value target map for the fixed-action comparison. Each row uses the configured self-critique action and the validation firing-rate target.

Baseline	Target construction	Deployment condition
Direct-V teacher	Mean gain of six independent $k = 8$ GPT-4o votes over a fixed CoT baseline	Llama-3.1-8B family; self-critique; 1.14× cost
Direct-V critique gain	Binary correctness improvement from one sampled student critique	Llama-3.1-8B family; self-critique; 1.14× cost
Repeated-Gain Direct-V	Cross-fitted mean gain from 2032 stochastic critiques	Llama-3.1-8B family; self-critique; 1.14× cost

For Repeated-Gain Direct-V, the base answer $y_{0,i}$ is fixed and each stochastic critique yields

$$g_{im} = r(y_{im}, y_i^*) - r(y_{0,i}, y_i^*), \qquad \bar{g}_i = \frac{1}{M} \sum_{m=1}^{M} g_{im}, \quad M = 2032.$$

Critiques use temperature 0.7, top-p 0.95, and the deployed 256-token cap with independent seeds. We partition inputs into 5 problem-hash folds; for each fold, a query-only value regressor is fit on the other folds' $\bar{g}_i$ values and produces out-of-fold targets for the held-out fold. The resulting policy uses the same configured action and validation threshold procedure as the fixed-action comparison.

We measure V_a^* on a stratified $1{,}200$-item held-out slice by executing the Llama-3.1-8B critique action and recording the correctness change. The verdict-instability target correlates with measured signed gain at Spearman $\rho = 0.71$ on this slice and 0.66 on a matched OOD CommonsenseQA check, exceeding risk (0.48) and branch sensitivity (0.41). Five-rollout consensus raises this correlation to 0.79; the agreement and downstream stability results in Table 1 establish that the signal persists across teacher sampling.

F Escalation Shortcut Controls

Table 9: Escalation precision at matched recall $= 0.60$ under shortcut scrambling. The recall operating point is fixed so that only precision changes across rows.

Condition	Precision @ Recall $= 0.60$
Full Features	0.74
Scrambled Length	0.72
Scrambled Position	0.71
Removed Lexical Markers	0.73
Removed Teacher Confidence	0.70
All Proxies Scrambled	0.68

Under simultaneous scrambling of length, position, and lexical proxies, VerdictRoute retains 0.68 precision compared with the 0.63 Flat-Transcript reference. Removing teacher confidence produces the largest single change, from 0.74 to 0.70, aligning the stress test with the method's use of judge-state dynamics. The retained advantage under every perturbation shows that verdict-state information contributes beyond surface cues and complements the matched-action accuracy comparison.

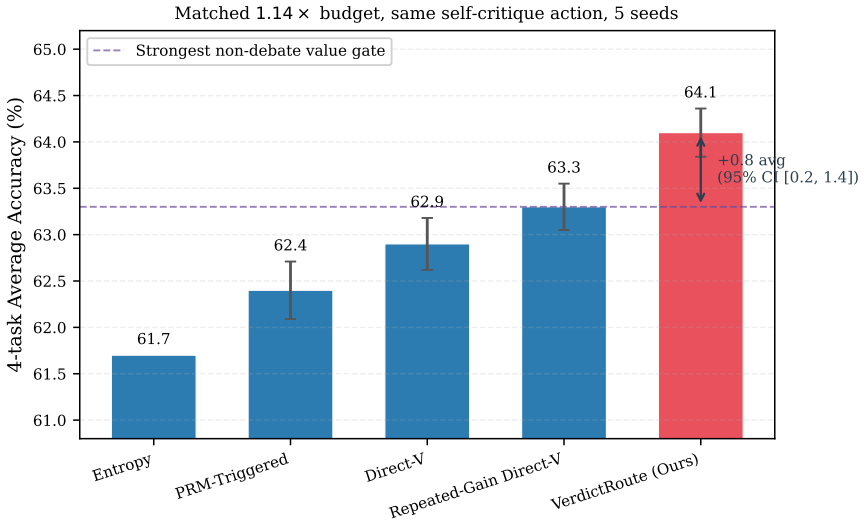


Figure 4: Fixed-action selective-compute comparison from Table 2. Error bars show one seed standard deviation where reported in the table; none is shown for entropy.

In Figure 4, the dashed line marks Repeated-Gain Direct-V, the strongest displayed non-debate comparator at 63.3.

G Oracle Gating Ceiling: Budget Sweep and Per-Task Detail

Table 10 expands the main-text gap decomposition (Table 3) by sweeping the average inference budget from 1.00× to 1.50× for three policies using the Llama-3.1-8B model family and the 256-token self-critique action. At each budget the escalation threshold is calibrated on the validation split so that escalation fires at the stated rate; the Oracle row instead chooses the same number of escalations with hindsight on the evaluation set. At 1.00× no escalation fires; the VerdictRoute ordering then persists from the first nonzero escalation budget at 1.07× through 1.50×. Figure 5 visualizes the three curves.

Table 10: Budget sweep for three gating policies using the Llama-3.1-8B model family and one 256-token self-critique action. The Oracle row is hindsight-optimal under the same budget.

Gating policy	1.00×	1.07×	1.14×	1.22×	1.30×	1.50×
PRM-Triggered (avg)	60.5	61.6	62.4	63.0	63.4	63.9
VerdictRoute (avg)	60.5	62.4	64.1	64.7	65.1	65.5
Oracle (avg, hindsight)	60.5	62.6	64.7	65.4	65.9	66.4
VerdictRoute lead over PRM (avg pts)	0.0	+0.8	+1.7	+1.7	+1.7	+1.6
Oracle minus VerdictRoute gap (avg pts)	0.0	0.2	0.6	0.7	0.8	0.9
VerdictRoute share of Oracle lift over PRM	—	80%	74%	71%	68%	64%
Per-task accuracy at 1.14× *(reference, identical to Table 3):*						
PRM-Triggered	GSM 77.8	MATH 45.8	TQA 58.5	TV 67.5	Avg 62.4	ECE 0.085
VerdictRoute	GSM 79.4	MATH 47.7	TQA 60.0	TV 69.3	Avg 64.1	ECE 0.058
Oracle	GSM 80.1	MATH 48.5	TQA 60.8	TV 69.5	Avg 64.7	ECE 0.044

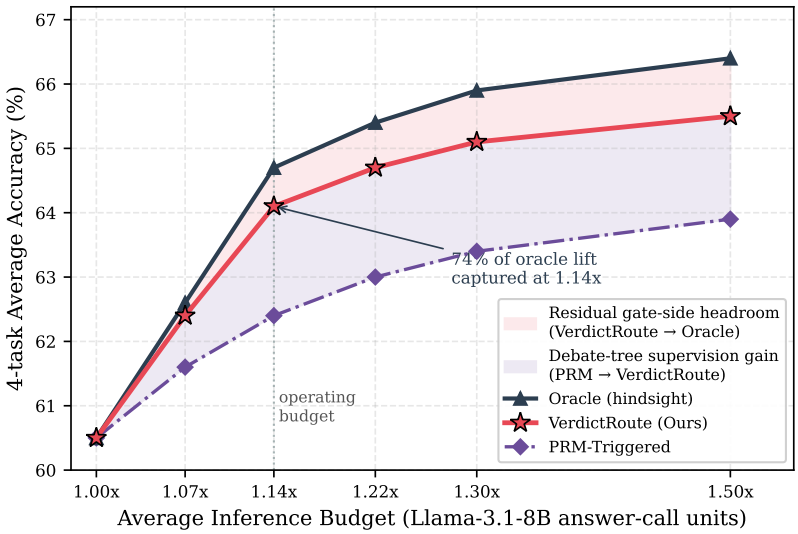


Figure 5: Budget sweep for the gap-decomposition package. Vertical spread isolates the gating decision because the answer model and action are fixed.

The sweep isolates the mechanism: at the operating budget only 0.6 points remain between VerdictRoute and the matched-action oracle. Above $1.30\times$, all three policies show sharp diminishing returns; the marginal lift from $1.30\times$ to $1.50\times$ is at most 0.5 points for every policy, supporting $1.14\times$ as an efficient operating point. The hindsight oracle serves as a ceiling that isolates remaining gate-side headroom.