TURF: OUTCOME-GROUNDED UNCERTAINTY QUANTIFICATION FOR LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM agents execute multi-step workflows with consequences beyond their final responses, making uncertainty quantification important for identifying tasks that warrant verification or human oversight. We introduce TURF, a task-training-free method that ranks task-level failure risk by fusing observable outcome dispersion with frozen-NLI spectral diversity of terminal responses. TURF combines smoothed dispersion of observable outcomes with frozen-NLI spectral diversity of terminal responses through Gaussianized-rank fusion, **without** fitting correctness-supervised predictors or fusion weights. TURF-Fixed uses a prescribed trajectory budget, while TURF-Adaptive selects a prefix using Bayesian evidence of outcome and semantic recurrence relative to the first execution. We evaluate both variants against 15 baselines on six benchmarks, three backbones, and three seeds, across supported trajectory caps from 2 to 16. Equally averaging across datasets, backbones, seeds, and $B \in \{4, 6, 8, 12, 16\}$, TURF-Fixed improves AUROC/AUPR by $5.19/2.99$ percentage points over SGD, the strongest baseline with complete results on this grid. TURF-Adaptive retains gains of $4.68/2.60$ points while using 20.90% fewer trajectories for scoring than TURF-Fixed, with corresponding quality losses of $0.52/0.39$ points. The results support outcome-grounded, two-view uncertainty estimation across multiple sampling budgets. The code and datasets are available at: `https://anonymous.4open.science/r/TURF`.

1 INTRODUCTION

LLM agents can automate complex workflows in software development, database administration, operating-system management, and business operations (Liu et al., 2024b; Yang et al., 2024; Yao et al., 2024), making their reliability critical to the people and systems that depend on them. These workflows involve repeated planning, tool use, and interaction with environments (Yao et al., 2023), producing consequences beyond the final response. An agent may report successful completion while leaving an operation unfinished or a system in an unintended state (Yao et al., 2024). Uncertainty quantification (UQ) can help identify tasks that warrant verification, human review, or additional computation (Zhao et al., 2025; Duan et al., 2025). This motivates uncertainty signals grounded in what executions accomplish, alongside what their terminal responses express.

Existing methods provide complementary foundations. Semantic entropy and graph- or kernel-based measures characterize response disagreement beyond surface wording (Farquhar et al., 2024; Lin et al., 2024; Nikitin et al., 2024), while confidence–consistency fusion combines response confidence with semantic agreement (Vashurin et al., 2026). Agent-oriented methods aggregate or propagate step uncertainty (Zhao et al., 2025; Duan et al., 2025), or compare action-set consistency and outcome equivalence across resampled executions (Bouchard & Chauhan, 2026). Adaptive sampling addresses the cost of repeated generations through answer-consistency criteria (Aggarwal et al., 2023; Li et al., 2024) or Bayesian semantic-entropy estimation (Ciosek et al., 2025).

An important limitation of step-based summaries is that *local ambiguity need not imply uncertainty about the completed task's consequences*. High step-level uncertainty can be benign when alternative actions lead to essentially the same outcome, or when later observations resolve earlier ambiguity. For example, different database operations may produce the same task-relevant state. Weighting or propagating local uncertainty does not, by itself, establish whether alternative execution paths remain distinguishable at completion. Conversely, terminal-response agreement can conceal different environment changes. An explicit outcome-level view can therefore complement both process uncertainty and response semantics.

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.

We study whether *observable outcome dispersion and terminal semantic structure provide complementary evidence for ranking task-level failure risk*. Operational outcome categories retain consequences that similar responses may obscure; graded semantic relationships retain information that discrete categories discard. Neither view certifies correctness: an agent can consistently produce the same wrong result. Our objective is an informative ranking of tasks requiring scrutiny, not a calibrated probability of execution success.

TURF, *Trajectory Uncertainty via Rank Fusion*, combines these views using unlabeled reference ranks. Label-free adapters expose observable answers, environment states, and termination outcomes. A count-based statistic measures their dispersion with small-sample smoothing and an uncertainty penalty, while a frozen natural-language-inference (NLI) model supplies a spectral measure of terminal-response diversity. Gaussianized-rank Stouffer fusion aligns and combines the two scales. TURF thus summarizes execution consequences and terminal semantics, rather than explicitly modeling internal action–observation dynamics. It is *task-training-free*: its pretrained models remain frozen, and neither a risk predictor nor fusion weights are fitted to task correctness labels.

TURF-Fixed uses a prescribed trajectory budget. TURF-Adaptive retains the scorer and selects a prefix using Bayesian two-view stopping: it accumulates evidence that subsequent executions match the fixed first execution in observable outcome and terminal-response meaning. Under an explicit recurrence model, its certificate concerns majority recurrence in both views, not correctness or preservation of failure-ranking performance. Our theoretical analysis characterizes this model-relative guarantee and sampling efficiency, alongside structural and sensitivity properties of the scoring components (Section 4).

We compare TURF with 15 baselines on six benchmarks, three backbones, and three seeds, across budgets of two to sixteen trajectories. Equally averaging across the 54 dataset–backbone–seed cells and $B \in \{4, 6, 8, 12, 16\}$, TURF-Fixed improves AUROC/AUPR over SGD, the strongest baseline with complete results on this grid, by $5.19/2.99$ percentage points. TURF-Adaptive retains gains of $4.68/2.60$ points while reducing mean scoring-prefix length by 20.90%. These native-support comparisons concern retrospective failure ranking, not end-to-end execution savings.

Our contributions are threefold: (i) an outcome-grounded, task-training-free score combining observable execution outcomes with terminal semantics; (ii) Bayesian two-view prefix selection, with model-relative recurrence and sampling-efficiency analysis and structural results for the scorer; and (iii) a six-benchmark evaluation using information-access and scoring ablations, budget-policy controls, and sensitivity analyses to examine predictive gains and the sampling–quality tradeoff.

2 Related Work

Language-model uncertainty and semantic disagreement. Uncertainty measures use likelihoods (PPL, PE) (Malinin & Gales, 2021), relevance-weighted likelihoods (SentSAR) (Duan et al., 2024), lexical agreement (LS) (Fomicheva et al., 2020), or semantic-class entropy (SE) (Farquhar et al., 2024). Semantic structure is captured through response neighborhoods (SNNE) (Nguyen et al., 2025), kernel spectra (KLE) (Nikitin et al., 2024), graph degree and eigenvalues (Degree, EigV) (Lin et al., 2024), probability-weighted semantic density (SD) (Qiu & Miikkulainen, 2024), and semantic graph density (SGD) (Li et al., 2025). CoCoA-MaxProb and CoCoA-PPL combine semantic consistency with maximum-probability and perplexity confidence, respectively (Vashurin et al., 2026). Bayesian semantic entropy addresses sampling efficiency through posterior entropy estimation and adaptive allocation (Ciosek et al., 2025). TURF builds on these ideas by combining terminal semantic diversity with observable execution-outcome dispersion through unlabeled rank fusion.

Agent uncertainty and adaptive sampling. SAUP propagates step uncertainty with situation-aware weights (Zhao et al., 2025), while UProp distinguishes intrinsic uncertainty from uncertainty inherited across decisions (Duan et al., 2025). TRACER aggregates surprisal, repetition, and action–observation coherence through a tail-focused trajectory score (Tayebati et al., 2026); ProbeCal learns reward-based calibration from language-model representations (Liu et al., 2024a). Trajectory-adapted methods also compare action-set consistency and judged outcome equivalence across executions (Bouchard & Chauhan, 2026). For answer selection, self-consistency votes over sampled reasoning paths (Wang et al., 2023); Adaptive-Consistency uses probabilistic answer-count evidence (Aggarwal et al., 2023), and Early-Stopping Self-Consistency uses agreement within recent samples (Li et al., 2024). TURF instead ranks fixed archived task outcomes using execution and terminal-response dispersion. Its

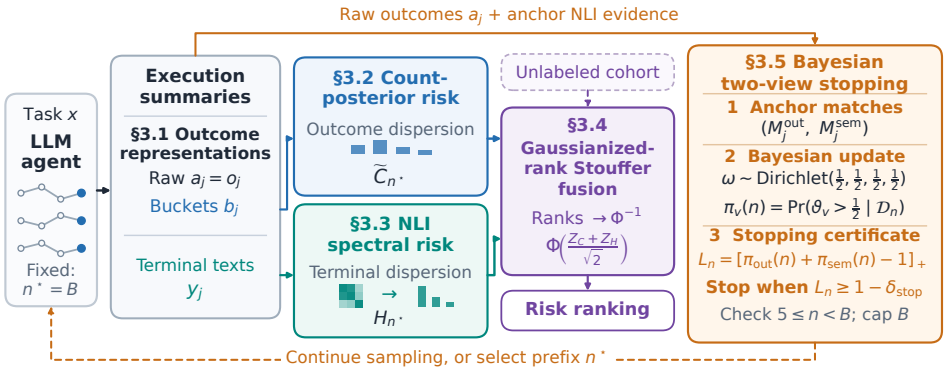

Figure 1: **TURF workflow.** Outcome and terminal-semantic risks are combined through rank fusion. Bayesian two-view stopping selects the scoring prefix using recurrence evidence. Orange arrows show recurrence evidence and sampling control; section numbers point to methodological details.

adaptive policy selects the scoring prefix through anchor recurrence, rather than changing the answer selected or certifying its correctness.

3 TURF: Trajectory Uncertainty via Rank Fusion

TURF combines observable-outcome dispersion and terminal semantic dispersion into a risk score, with a separate Bayesian policy selecting how many rollouts to inspect (Figure 1). From each rollout, TURF extracts an observable outcome and terminal text; at the selected prefix, it computes their dispersion scores and fuses their reference ranks. A separate recurrence rule selects that prefix for TURF-Adaptive.

Setting and scope. For task x, let $\tau_{1:B}$ be an ordered pool of complete stochastic agent rollouts with an externally specified cap $B \geq 2$. The index t enumerates *rollouts*, not steps within a rollout. A policy selects $n^\star \leq B$ and returns $S_{n^\star}$, with larger scores indicating greater estimated failure risk. TURF-Fixed uses $n^\star = B$; TURF-Adaptive uses the rule below. Both share the same scorer, and stopping parameters remain fixed across budgets. The appropriate cap depends on application-specific cost and quality requirements; we therefore evaluate a budget range rather than prescribe one operating point.

The prediction target is an archived task-level failure label Y_x, held fixed across budgets and policies; some targets aggregate the full rollout pool. Changing the scoring prefix n changes the evidence inspected, not Y_x. Evaluation therefore measures uncertainty about this fixed target, not the correctness of an answer generated using only the selected prefix (Appendix B.6).

3.1 Observable outcome representations

Agreement between executions depends on their observable consequences, not necessarily on their paths. For an environment-specific adapter indexed by d, define

$$o_t = g_d(x, \tau_t), \qquad b_t = T_d(o_{1:t}), \tag{1}$$

Here o_t is the extracted observable outcome, b_t is its count-scoring bucket, and y_t denotes the terminal text supplied to the semantic view. The adapter uses the task request and execution record, not reference answers, official rewards, or correctness labels. Extraction is deterministic conditional on the enriched record, including cached label-free preprocessing. The encoding T_d is prefix-causal: assigning b_t requires no later rollout. Its history dependence supports missing- and censored-output handling; ordinary completed outcomes can be encoded from o_t alone. These rules avoid uniformly treating missing outputs as agreement on a completed answer (Table 1).

Granularity is a modeling choice: coarse buckets can conceal consequential differences, while fine buckets can fragment equivalent outcomes. Buckets express agreement under the chosen representation, not correctness; the semantic view complements these categories with graded relationships.

Table 1: Observable outcome representations used by TURF. Appendix B.1 specifies normalization, preprocessing, and fallback rules.

Agent environment	Observable outcome representation	Evaluated examples
Answer-returning agents	Normalized answer or Boolean value; explicit missing-output handling.	HotpotQA, StrategyQA, GAIA (tool-assisted)
Database interaction	Answer for reads; database-state signature for writes when available; operation/status fallbacks.	AgentBench-DB
Operating-system interaction	Answer normalized to the task's output schema; invalid and missing-output outcomes.	AgentBench-OS
Interactive shopping	Purchase/no purchase; request-coverage, selected-option, and price-status signatures.	WebShop

3.2 Count posterior risk: observable-outcome dispersion

At prefix n, let K_n^{out} be the number of distinct scoring buckets and N_{nk} the count of bucket k. For ordinary outcomes, use

$$q_n \sim \text{Dirichlet}\left(N_{n1} + \tfrac{1}{2}, \ldots, N_{n,K_n^{\text{out}}} + \tfrac{1}{2}, \tfrac{1}{2}\right). \tag{2}$$

The final half-count reserves mass for an unseen outcome. Because the represented alphabet grows with observations, this is an observed-support *pseudo-posterior*, not a fixed-alphabet generative model. Let h_d denote Shannon entropy by default, with the evaluated environment-specific extension in Appendix B.1. With $\mu_n = \mathbb{E}[h_d(q_n)]$ and $v_n = \text{Var}[h_d(q_n)]$, define

$$\widetilde{C}_n = \mu_n - \kappa\sqrt{\max(v_n, 10^{-12})} + \epsilon_{d,B,n}\left(1 - e^{-\max(\bar{r}_n, 0)}\right), \qquad \kappa = \Phi^{-1}(0.975). \tag{3}$$

The mean μ_n measures expected outcome dispersion, the standard-deviation penalty discounts dispersion with greater uncertainty under the smoother, and the small final term distinguishes count ties. Here Φ is the standard normal CDF; chosen-output surprisal $\bar{r}_n$ averages valid steps within each rollout, then valid rollouts (Appendix B.2). Despite its conventional Gaussian multiplier, $\widetilde{C}_n$ is a ranking statistic, not a guaranteed 95% entropy bound or a failure probability.

3.3 NLI spectral risk: terminal-response dispersion

Following semantic uncertainty and kernel language entropy (Kuhn et al., 2023; Nikitin et al., 2024), a frozen NLI model supplies directed entailment probabilities between terminal texts $y_1, \ldots, y_n$, conditioned on the task prompt. Define the symmetric affinity

$$A_{jk} = \min\left\{p_{\text{NLI}}(y_j \Rightarrow y_k \mid x), p_{\text{NLI}}(y_k \Rightarrow y_j \mid x)\right\}, \quad j \neq k, \qquad A_{jj} = 1. \tag{4}$$

The minimum requires support in both directions. Let $\mathbf{K}_n$ be the unit-diagonal positive-semidefinite repair of $\mathbf{A}_n = (A_{jk})_{j,k=1}^n$ (Appendix B.3). Under uniform rollout weights, set $\rho_n = \mathbf{K}_n/n$ and compute normalized spectral entropy

$$H_n = -\frac{1}{\log n}\sum\nolimits_{j=1}^{n} \lambda_j \log \lambda_j, \qquad \lambda_j = \lambda_j(\rho_n), \qquad 0 \log 0 = 0. \tag{5}$$

Dividing the kernel by n makes its trace one; dividing entropy by $\log n$ scales its maximum to one. In exact arithmetic, a rank-one kernel gives $H_n = 0$ and the identity gives $H_n = 1$. This normalization does not imply prefix-length invariance, nor do policy-specific reference ranks automatically remove such dependence. The score measures graded semantic dispersion without hard clustering. Terminal decisions can serve as texts when final answers are absent; internal execution order is not encoded.

3.4 Gaussianized-rank Stouffer fusion

We align the views using unlabeled empirical midrank maps R_C, R_H from a reference cohort under the same scoring and sampling configuration. Each reference task contributes its statistics at its own selected prefix. Define

$$\begin{aligned} Z_C &= \Phi^{-1}\left(R_C(\widetilde{C}_n)\right), \quad Z_H = \Phi^{-1}(R_H(H_n)), \\ S_n^{\text{core}} &= \Phi\left(\frac{Z_C + Z_H}{\sqrt{2}}\right). \end{aligned} \tag{6}$$

The count reference includes all cohort tasks; the semantic reference includes only those with semantic evidence. Without semantic evidence, we use $S_n^{\text{core}} = R_C(\widetilde{C}_n)$. Using the unlabeled evaluation cohort makes the reported scoring transductive. Rank conventions and numerical details appear in Appendix B.4.

Equal weighting is a symmetric default requiring no correctness-fitted weights, not evidence that the views are equally predictive or optimally balanced. Its conditional rationale and stability properties are analyzed in Section 4. The outer Φ is a monotone rescaling, not probability calibration. We set $S_n = S_n^{\text{core}}$ except for the documented environment-specific adjustment (Appendix B), shared by both policies.

3.5 Bayesian two-view stopping

1. Anchor matches. The adaptive policy holds the first rollout fixed as an anchor. Let $a_j := o_j$ denote the extracted outcome of rollout j, before the count-scoring encoding $b_j = T_d(o_{1:j})$. For $j \geq 2$, define

$$M_j^{\text{out}} = \mathbf{1}\{a_j = a_1\}, \qquad M_j^{\text{sem}} = \mathbf{1}\{A_{1j} \geq \eta\}, \qquad \eta = 0.75. \tag{7}$$

Semantic matching uses the raw affinity before PSD repair. Stopping tests recurrence of observed outcomes, whereas count scoring treats missing outputs conservatively to avoid equating them with agreement on a completed answer. Thus repeated missing-output outcomes may match for stopping even when scoring separates them (Appendix B.1). These indicators are distinct from $\widetilde{C}_n, H_n$: persistence, including repeated missing outputs, does not imply correctness.

2. Posterior recurrence evidence. Let $m_v(n) = \sum_{j=2}^{n} M_j^v$ for $v \in \{\text{out}, \text{sem}\}$ and $\mathcal{D}_n = ((M_j^{\text{out}}, M_j^{\text{sem}}))_{j=2}^{n}$. Conditional on the anchor, the stopping model treats these pairs as iid categorical draws with probabilities $\boldsymbol{\omega} = (\omega_{00}, \omega_{01}, \omega_{10}, \omega_{11})$ and prior $\text{Dirichlet}(1/2, 1/2, 1/2, 1/2)$. The marginal match probabilities $\vartheta_{\text{out}} = \omega_{10} + \omega_{11}$ and $\vartheta_{\text{sem}} = \omega_{01} + \omega_{11}$ have $\text{Beta}(1, 1)$ priors, giving

$$\begin{aligned} \vartheta_v \mid \mathcal{D}_n &\sim \text{Beta}(1 + m_v(n),\, n - m_v(n))\,, \\ \pi_v(n) &:= \Pr\big(\vartheta_v > \tfrac{1}{2} \mid \mathcal{D}_n\big) = 2^{-n} \sum\nolimits_{k=0}^{m_v(n)} \binom{n}{k}. \end{aligned} \tag{8}$$

Thus $\pi_v(n)$ quantifies posterior evidence that view v matches the anchor more often than not. This model allows within-rollout dependence between views and is distinct from the scoring pseudo-posterior in Eq. (2).

3. Stopping certificate. For $\mathcal{R} = \{\vartheta_{\text{out}} > 1/2,\ \vartheta_{\text{sem}} > 1/2\}$, a posterior union bound gives

$$\Pr(\mathcal{R} \mid \mathcal{D}_n) \geq L_n := \big[\pi_{\text{out}}(n) + \pi_{\text{sem}}(n) - 1\big]_+, \qquad [z]_+ = \max(z, 0). \tag{9}$$

The bound certifies posterior support for majority recurrence in both views, not a joint match rate $\omega_{11} > 1/2$. For tolerance $\delta_{\text{stop}} \in (0, 1)$, select the first eligible prefix whose bound reaches $1 - \delta_{\text{stop}}$, or use the cap:

$$n^\star = \min\big(\big\{n \in \mathbb{N} : 5 \leq n < B,\ L_n \geq 1 - \delta_{\text{stop}}\big\} \cup \{B\}\big)\,. \tag{10}$$

We use $\delta_{\text{stop}} = 0.2$, equivalent to requiring $\pi_{\text{out}}(n) + \pi_{\text{sem}}(n) \geq 1.8$ for early stopping. All B rollouts are used when $B \leq 5$ or no eligible prefix qualifies. Without semantic evidence, the fallback uses $\pi_{\text{out}}(n) \geq 0.8$ with the same minimum and cap, supporting outcome recurrence alone. The semantic cutoff and minimum prefix are design choices, not consequences of the guarantee. Given the indicators, stopping requires two running counts and two binomial-CDF evaluations per prefix; the shared scorer then evaluates the selected prefix using its operating-point reference cohort.

4 Theoretical Analysis of TURF

We characterize the dispersion scores, sensitivity of rank fusion, and scope of the stopping certificate. Formal statements, proofs, and implementation qualifications appear in Appendix C.

Count posterior risk. Let P_n be the scoring smoother and $X_n = h_d(q_n)$ have exact mean μ_n and finite variance v_n. For $\kappa \geq 0$, Proposition C.1 gives

$$\inf_{\substack{\mathsf{Q} \ll \mathsf{P}_n \\ \chi^2(\mathsf{Q} \| \mathsf{P}_n) \leq \kappa^2}} \mathbb{E}_{\mathsf{Q}} X_n \geq \mu_n - \kappa\sqrt{v_n}. \tag{11}$$

Thus $\mu_n - \kappa\sqrt{v_n}$ lower-bounds expected dispersion under bounded reweightings of the smoother, relating the score to variance-based robustness (Duchi & Namkoong, 2019). This is not a confidence bound or robustness to task-distribution shift; equality requires an admissible reweighting specified in the appendix.

NLI spectral semantic risk. For a perfect semantic-partition kernel with group sizes c_g, the nonzero eigenvalues of ρ_n are c_g/n: spectral entropy recovers empirical group entropy normalized by $\log n$ (Nikitin et al., 2024). Proposition C.2 also bounds entropy changes by the trace distance between repaired, trace-normalized kernels at the same prefix length (Audenaert, 2007). This interprets the score as a continuous extension of semantic grouping, whose classes need not coincide with operational outcome buckets.

Gaussianized-rank Stouffer fusion. With fixed reference membership, exact-rank fusion is invariant to strictly increasing component transformations and nondecreasing in each rank. Proposition C.3 bounds fused-score changes caused by rank errors. Equal weighting induces the likelihood-ratio ranking under a Gaussian working model with equal positive class separation and shared exchangeable covariance (Appendix C.4). Gaussianization does not imply these assumptions; this is a conditional rationale, not general optimality.

Trajectory-budget policy. Given the task, fixed anchor, and joint category probabilities, assume iid recurrence pairs and the stated Dirichlet prior on those probabilities. The views may be dependent within a rollout. Proposition C.4 establishes

$$n^\star < B \quad \Longrightarrow \quad \Pr(\mathcal{R}^c \mid \mathcal{D}_{n^\star}) \le \delta_{\text{stop}}. \tag{12}$$

This is Bayesian posterior reliability, not a uniform frequentist false-stop guarantee. A companion finite-horizon savings bound increases with both marginal recurrence rates. Neither result guarantees correctness or preservation of failure-ranking performance. The statements concern the core scorer and two-view branch; the appendix qualifies numerical approximations, fallback, and outer adjustments. The full-pool semantic-availability mask is not shown to preserve the iid law.

Empirical implications. The propositions motivate complementary checks: penalty controls test the usefulness of variance discounting; component ablations test the information supplied by semantic dispersion; reference-bank and mean-rank controls probe rank sensitivity and the fusion choice; and same-cap, matched-cost, and shuffled-length comparisons test the stopping rule's sampling–quality tradeoff (Section 5; Appendices F, H.1, H.2, and F.3.2). These assess practical utility, not whether the Gaussian working model or iid recurrence law holds.

5 Experiments

Evaluation. We evaluate failure-positive AUROC and AUPR on AgentBench-DB and AgentBench-OS (Liu et al., 2024b), HotpotQA (Yang et al., 2018), WebShop (Yao et al., 2022), StrategyQA (Geva et al., 2021), and GAIA Levels 1–2 (Mialon et al., 2024), using Qwen 3.5 9B, Gemma 3 12B, GPT-OSS 20B, and three seeds. Metrics are computed within each of the 54 dataset–backbone–seed cells and averaged equally; repeated tasks are not independent replications. Archived pools contain $Z = 16$ trajectories per task, with evaluated caps $\mathcal{B} = \{2, 3, 4, 6, 8, 12, 16\}$. The 15 baselines retain their registered implementations and native task support. Degree and UProp cover all cells through $B = 12$; the other 13 also cover $B = 16$. Budget averages equally weight within-cell metrics over their stated grid, rather than pooling task scores or integrating curves. Appendix A specifies seeds, prediction targets, data provenance, baseline implementations, and inference protocols.

Failure ranking across budgets and environments. Figure 2 shows TURF's macro advantage across the budget range. Averaged over $B \in \{4, 6, 8, 12, 16\}$ and all 54 cells, TURF-Fixed reaches $0.7615/0.7221$ AUROC/AUPR, exceeding SGD, the strongest baseline under this aggregation, by $5.19/2.99$ percentage points (pp). Adaptive reaches $0.7563/0.7182$, retaining gains of $4.68/2.60$ pp while reducing mean trajectory usage by 20.90%. Both variants have higher macro means than all 13 fully covered baselines at each of these five budgets. The advantage is not universal: averaging the same budgets and seeds within each dataset–backbone panel, Fixed exceeds the strongest baseline in 15/18 AUROC and 14/18 AUPR panels; Adaptive does so in 14/18 and 13/18, respectively (Appendix D). For each panel and metric, the comparison uses the maximum baseline mean among the reported baselines. These descriptive native-support results do not establish coverage-matched statistical superiority.

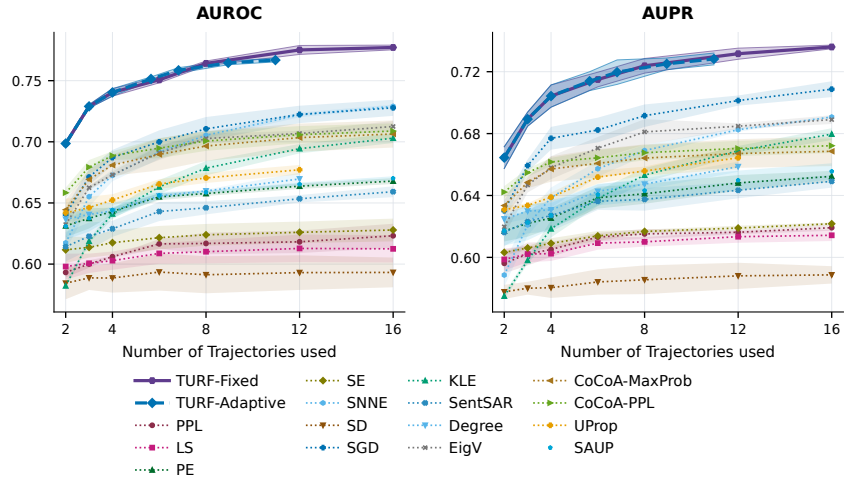

Figure 2: **Macro cost–performance comparison.** AUROC/AUPR versus source-reported mean trajectory usage, equally averaged over six datasets, three backbones, and three seeds. Markers correspond to caps $B \in \{2, 3, 4, 6, 8, 12, 16\}$ where supported; Adaptive markers use realized mean prefix lengths. Ribbons are ± 1 sample standard deviation across three seed macros, not confidence intervals. Baselines retain native task support and nominal-cost fallbacks where realized counts are unavailable; these curves are not matched-cost tests.

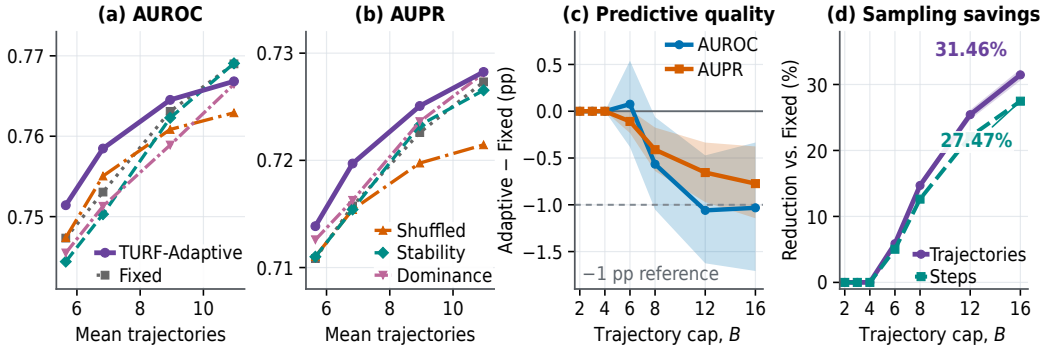

Figure 3: **Budget-policy comparison and sampling–quality tradeoff.** Results equally average 54 cells. **(a,b)** AUROC/AUPR at matched trajectory cost, with the TURF scorer fixed. Fixed, Stability, and Dominance use randomized mixtures with small matching errors; Shuffled preserves exact stopping-length histograms. Points correspond to caps $6, 8, 12, 16$. **(c)** Adaptive minus Fixed at the same cap, in pp. **(d)** Corresponding trajectory and recorded decision-step savings. Shading in (c,d) gives paired, pointwise 95% task-bootstrap intervals from 50,000 resamples; step savings have no interval. The dashed -1 pp line is a visual reference, not a validated noninferiority margin. Savings concern cached scoring prefixes, not measured end-to-end speedups.

How does the available budget change performance? The TURF variants coincide at $B = 2, 3, 4$, where the minimum stopping prefix prevents early termination. Increasing Fixed's budget from two to four raises AUROC/AUPR by $4.18/3.98$ pp; increasing it from twelve to sixteen adds $0.20/0.44$ pp (Figure 2). These macro trends suggest diminishing returns on this grid, not a universally appropriate cap: individual curves can flatten, continue improving, or become non-monotone, with substantial seed variation in some WebShop panels (Appendix D).

Sampling savings and their quality tradeoff. Across caps $B \in \{6, 8, 12, 16\}$, Adaptive saves 5.87–31.46% of trajectories relative to Fixed at the same cap (Figure 3(c,d)). Larger caps permit greater savings, with AUROC losses exceeding one pp at $B = 12, 16$. At $B = 16$, the 31.46% trajectory reduction accompanies AUROC/AUPR changes of $-1.03/-0.77$ pp and a 27.47% reduction in recorded decision steps. Savings vary substantially by dataset (Appendix G). These cached-prefix estimates exclude input-token, NLI, tool, and other unrecorded costs; they neither measure end-to-end speedups nor establish better quality at equal cost.

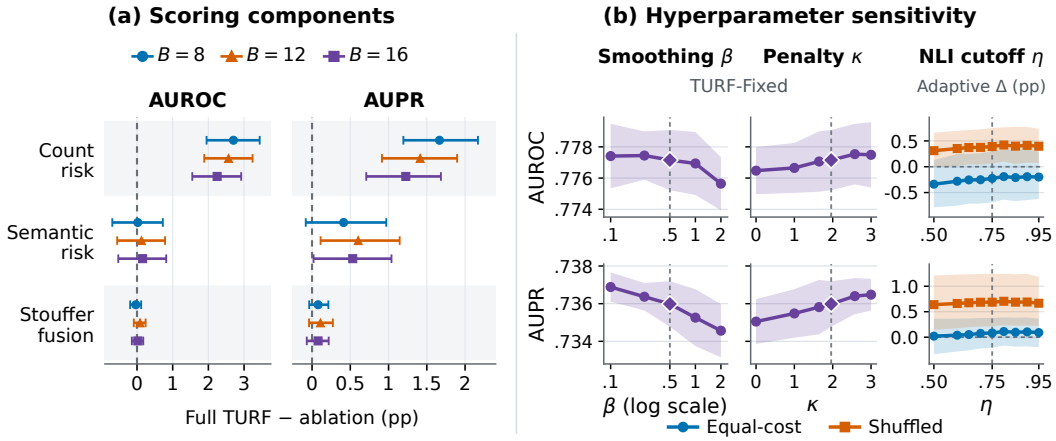

Figure 4: (a) **Scoring components across budgets.** Full TURF minus ablation in AUROC/AUPR (percentage points) at $B \in \{8, 12, 16\}$. Rows compare with the semantic arm (count fallback retained), count-only arm, and mean-rank fusion, respectively. Whiskers are paired, pointwise 95% task-bootstrap intervals; positive values favor Full TURF. (b) **Hyperparameter sensitivity at $B = 16$.** Left and middle: TURF-Fixed scores with ± 1 SD across three seed macro means. Right: TURF-Adaptive minus matched-cost controls (percentage points), with diagnostic 95% dataset-cluster intervals. Vertical dashed lines in (b) mark defaults.

Does task-dependent allocation help at matched cost? Holding the scorer fixed, Figure 3(a,b) compares Bayesian stopping with cost-matched Fixed, outcome stability, outcome dominance, and shuffled stopping lengths. The first three mix bracketing settings using costs, not labels; shuffling preserves each cell's exact length histogram. Adaptive has higher mean AUROC and AUPR than all four at caps $6, 8, 12$, including gains of $0.54/0.40$ pp over matched Fixed at $B = 8$. At $B = 16$, the corresponding differences are $-0.22/+0.09$ pp. Adaptive exceeds shuffled lengths in both metrics at all four caps, supporting task-dependent allocation but not uniform superiority over simpler policies. Appendix F.3.2 shows AUPR gains over matched Fixed in all six datasets at $B = 6, 8$. At $B = 8$, all four AUPR diagnostic intervals exclude zero, whereas all four AUROC intervals include it. These six-dataset-cluster intervals are descriptive rather than confirmatory.

Which scoring components matter across budgets? With sampling held fixed, Figure 4(a) shows positive count-view contributions to both metrics at $B = 8, 12, 16$, with all displayed intervals above zero. The full curves retain positive mean count contributions at every evaluated budget (Appendix F.2). Semantic risk supplies smaller, budget-dependent gains: its mean AUPR contribution is negative at $B = 2, 3$ and positive from $B = 4$ onward. Stouffer's differences from mean-rank fusion are small, with all displayed intervals crossing zero; the Gaussian construction is a theoretically motivated default, not an established empirical improvement. The arms retain count fallback and environment-specific wrappers; none changes the stopping rule's inputs. Appendix F reports full paired analyses. The displayed intervals are pointwise, without a multiplicity correction; the appendix's eight-test Holm correction applies only to its fixed-$B = 16$ family.

Hyperparameter sensitivity. We used the same fixed parameter values across datasets and backbones. Here, we summarize sensitivity to three key hyperparameters: half-count smoothing $\beta = 1/2$, the Gaussian-motivated penalty $\kappa = \Phi^{-1}(0.975)$, and semantic cutoff $\eta = 0.75$. At $B = 16$ across all 54 cells, varying $\beta \in \{0.1, 0.25, 0.5, 1, 2\}$ gives Fixed AUROC/AUPR ranges of $0.18/0.23$ pp; varying κ from 0 to 3 gives $0.11/0.14$ pp (Figure 4(b)). Across $\eta \in [0.50, 0.95]$, Adaptive's mean gains over shuffled allocation remain positive, while differences from equal-cost allocation remain inconclusive. Appendix H provides the full sensitivity analyses.

Does execution information add value beyond final responses? To test the outcome-grounded motivation directly, we compare execution-derived count buckets with normalized final-answer strings, using identical tasks, prefixes, censoring rules, a flat count estimator, and frozen semantic scores. Both arms omit the surprisal tie-breaker, hierarchy, and severity adjustment (Appendix F.5). Figure 5 shows positive macro effects at every displayed budget. Averaged over $B \in \{4, 6, 8, 12, 16\}$, execution information improves AUROC/AUPR by $1.59/1.24$ percentage points (pp). Gains concentrate in

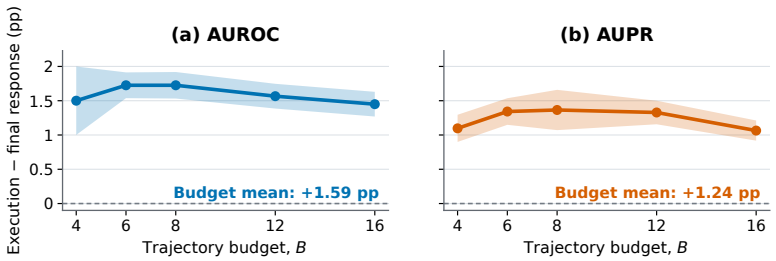

Figure 5: **Value of execution-outcome information.** AUROC/AUPR differences between execution-derived and final-response-derived count buckets, with the remaining scorer and sampled prefixes held fixed. Points equally average 54 cells; bands show ± 1 sample standard deviation of three seed-macro differences, not confidence intervals. Positive values favor execution-derived buckets.

Method	AUROC	AUPR	Trajectories
TURF-Fixed	**.7439**	**.7067**	5.833
TURF-Adaptive	.7412	.7049	**5.059**
SGD-Final	.6881	.6725	5.833
CoCoA-PPL-StepAgg	.6808	.6609	5.833
SGD-StepAgg	.6788	.6655	5.833

Table 2: **Matched-support Final/StepAgg comparison.** Equal-cell and equal-budget means over 54 cells and $B \in \{2, 3, 4, 6, 8, 12\}$ on the secondary candidate-enriched panel. All methods share 21,986 task–cell records. Displayed baselines comprise those with the highest mean AUROC or AUPR within each adaptation, determined from unrounded means. Trajectories are mean scoring-prefix lengths, excluding local-candidate acquisition; this is not a matched-cost test.

AgentBench-DB ($5.64/5.78$ pp) and WebShop ($3.99/1.48$ pp), with small or zero effects elsewhere. These results support outcome information where terminal responses can omit consequential state differences. The contrast isolates count-side information access; it is neither a completely response-only system comparison nor evidence for modeling within-trajectory dynamics.

Does the advantage persist under alternative baseline adaptations? We evaluate 13 baseline families in both final-response (Final) and step-aggregated (StepAgg) forms (Appendix E). TURF-Fixed and TURF-Adaptive exceed all 26 adaptations in macro AUROC and AUPR at every evaluated cap. In the budget averages of Table 2, their gains over SGD-Final, the strongest comparator on both metrics, are $5.58/3.42$ and $5.31/3.24$ pp, respectively. This checks sensitivity to baseline adaptation and task coverage, not statistical superiority or independent generalization: the panel uses different candidate-enriched pools and is not pooled with the primary comparison.

Additional robustness checks. We divide the tasks in each evaluation cell into five groups and compute reference ranks for each group using only tasks in the other four groups. Thus, no task contributes to its own reference distribution. This exclusion has little effect on either TURF variant, whereas small reference banks affect AUPR more strongly (Appendix H.2). Broader stopping sweeps show that the recurrence threshold, posterior tolerance, and minimum prefix produce larger changes in the sampling–quality tradeoff than the semantic cutoff (Appendix H.4). Together, these checks support limited sensitivity to including tasks in their own reference distributions and to the semantic cutoff. They also clarify how reference-bank size affects ranking quality and how other stopping parameters control the sampling–quality tradeoff within the tested settings.

6 Conclusion

TURF combines observable outcome dispersion and terminal semantics to rank task-level failure risk without correctness-supervised training. Its fixed-budget score has strong macro performance across the evaluated panel; Bayesian two-view stopping reduces scoring-prefix usage with a measurable quality tradeoff. Ablations most strongly support the count view, with smaller semantic contributions and competitive mean-rank fusion. These retrospective findings motivate prospective evaluation with decision-time semantic availability and independent rank references.

REFERENCES

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12375–12396. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.761. URL https://aclanthology.org/2023.emnlp-main.761/.

Koenraad MR Audenaert. A sharp continuity estimate for the von neumann entropy. *Journal of Physics A: Mathematical and Theoretical*, 40(28):8127–8136, 2007.

Dylan Bouchard and Mohit Singh Chauhan. Beyond single-turn confidence: Trajectory-adapted uncertainty quantification for LLM agents. *arXiv preprint arXiv:2608.11552*, 2026.

Kamil Ciosek, Nicolò Felicioni, and Sina Ghiassian. Hallucination detection on a budget: Efficient Bayesian estimation of semantic entropy. *TMLR*, 2025.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *ACL*, pp. 5050–5063, 2024.

Jinhao Duan, James Diffenderfer, Sandeep Madireddy, Tianlong Chen, Bhavya Kailkhura, and Kaidi Xu. UProp: Investigating the uncertainty propagation of LLMs in multi-step agentic decision-making. *arXiv preprint arXiv:2506.17419*, 2025.

John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *TACL*, 8:539–555, 2020.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did Aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *TACL*, 9: 346–361, 2021.

Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In *ICLR*, 2024.

Zhaoye Li, Siyuan Shen, Wenjing Yang, Ruochun Jin, Huan Chen, Ligong Cao, and Jing Ren. Enhancing uncertainty quantification in large language models through semantic graph density. In Silvia Chiappa and Sara Magliacane (eds.), *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 2537–2551. PMLR, 21–25 Jul 2025. URL https://proceedings.mlr.press/v286/li25b.html.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *TMLR*, 2024.

Hao Liu, Zi-Yi Dou, Yixin Wang, Nanyun Peng, and Yisong Yue. Uncertainty calibration for tool-using language agents. In *EMNLP*, pp. 16781–16805, 2024a.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. AgentBench: Evaluating LLMs as agents. In *ICLR*, 2024b.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *ICLR*, 2021.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *International Conference on Learning Representations*, volume 2024, pp. 9025–9049, 2024.

Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. Beyond semantic entropy: Boosting LLM uncertainty quantification with pairwise semantic similarity. In *Findings of ACL*, pp. 4530–4540, 2025.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. *NeurIPS*, 37:8901–8929, 2024.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. *NeurIPS*, 37:134507–134533, 2024.

Sina Tayebati, Divake Kumar, Nastaran Darabi, Davide Ettori, Ranganath Krishnan, and Amit Ranjan Trivedi. TRACER: Trajectory risk aggregation for critical episodes in agentic reasoning. *arXiv preprint arXiv:2602.11409*, 2026.

Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. CoCoA: A minimum Bayes risk framework bridging confidence and consistency for uncertainty quantification in LLMs. *NeurIPS*, 38:106236–106281, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

John Yang, Carlos Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37:50528–50652, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, pp. 2369–2380, 2018.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. *NeurIPS*, 35:20744–20757, 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Qiwei Zhao, Dong Li, Yanchi Liu, Wei Cheng, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Huaxiu Yao, Chen Zhao, et al. Uncertainty propagation on LLM agent. In *ACL*, pp. 6064–6073, 2025.

A Empirical Settings and Guide to the Appendix

This section defines the common protocol and separates three questions: failure ranking, quality versus inspected trajectories, and adaptive allocation at comparable sampling cost. Throughout, **TURF-Fixed** uses a prescribed number of trajectories, and **TURF-Adaptive** uses Bayesian two-view stopping with the same scorer. Endpoint rankings, same-cap comparisons, and matched-cost controls answer different questions.

Notation and reading conventions. We retain the main paper's notation throughout. A trajectory τ_t is one complete agent execution; t indexes executions, not decisions within one execution. The cap is B, a candidate prefix has length n, and the retained prefix has length $n^\star$. Observable outcomes o_t and scoring buckets b_t are distinct from the stopping outcomes $a_t := o_t$, which precede the censoring encoding. The scoring statistics $\widetilde{C}_n$ and H_n measure dispersion; the binary indicators $M_j^{\mathrm{out}}, M_j^{\mathrm{sem}}$ and their posterior probabilities $\pi_v(n)$ instead measure anchor recurrence. Neither is a correctness label. We reserve δ_{stop} for the Bayesian stopping tolerance and Y_x for the archived failure target. AUROC/AUPR differences are in percentage points (pp), whereas sampling savings are relative percentages.

A.1 Evaluation population and prediction target

Datasets, backbones, and seeds. We evaluate six datasets with three backbones—Qwen 3.5 9B, Gemma 3 12B, and GPT-OSS 20B—and generation seeds $101, 202, 303$. A *cell* is one dataset–backbone–seed combination, giving 54 cells. Table 3 specifies the evaluation population. Each cell has an ordered pool of 16 complete trajectories per task. The primary pools therefore contain 22,230 task–cell instances and 355,680 trajectories. These are repeated evaluations of 2,470 benchmark task identifiers, not 22,230 independent tasks. The same benchmark tasks recur across backbones and seeds. AgentBench-DB and AgentBench-OS are also abbreviated DBBench and OS, respectively, in the archived reports and figure labels.

Benchmark diversity. The six benchmarks cover complementary agent settings: database interaction (AgentBench-DB), operating-system operations (AgentBench-OS), constraint-driven shopping (WebShop), multi-hop evidence integration (HotpotQA), implicit multi-step reasoning (StrategyQA), and tool-assisted information gathering (GAIA). They differ in interaction environments, reasoning requirements, and observable outcomes, including answers, database states, purchases, and termination statuses. Although some share answer-based outputs, their distinct task demands allow us to examine whether outcome-grounded uncertainty remains useful across heterogeneous agent settings.

Table 3: **Primary evaluation population.** Task counts are per backbone–seed cell. Failure prevalence is the mean, followed by the minimum–maximum range, across the nine cells of each dataset, using the fixed archived target. These are not confidence intervals.

Dataset	Tasks	Evaluated scope	Failure prevalence (%)
AgentBench-DB	300	Frozen DBBench task set	46.85 [36.33, 64.00]
HotpotQA	1,000	Validation, distractor subset	42.49 [38.80, 46.20]
WebShop	200	Frozen shopping task set	84.50 [79.00, 89.50]
StrategyQA	687	Archived test split	7.73 [5.53, 11.06]
AgentBench-OS	144	Frozen OS task set	62.42 [49.31, 70.83]
GAIA	139	Validation, Levels 1 and 2	76.10 [61.87, 86.33]

The HotpotQA pool uses a fixed task-ID list from the distractor validation split, rather than the full validation population. StrategyQA uses the archived `ChilleD/StrategyQA` test split. GAIA excludes Level 3 tasks and the separate suitability pilot. We preserve each source campaign's finalized records and task identities; retry and repair shards are not additional evaluation examples.

What is predicted. For task x in a cell, $Y_x = 1$ denotes failure of the archived task-level result and $Y_x = 0$ denotes success. The target is held fixed across budgets, scoring ablations, and stopping policies within that panel. It is *not* a per-step error label, a first-rollout label, or a newly generated answer's correctness at the selected prefix. In particular, some targets aggregate the full 16-rollout

pool: HotpotQA, StrategyQA, and GAIA use the modal valid/nonempty answer under their respective normalization and checking rules; WebShop and AgentBench-OS use the modal available Boolean execution-correctness flag. AgentBench-DB selects the modal nonempty committed raw answer and aggregates the official flags of matching executions. Its target is distinct from its postprocessed-answer correctness field. The exact rules, including ties and missing flags, appear in Appendix B.6. In the five-dataset checkpoint materializer, a value other than Boolean true is mapped to failure; the GAIA loader negates its saved Boolean correctness.

Thus the experiment evaluates *prefix-based uncertainty about a fixed archived task result*. Early stopping changes the evidence used to score that target; it does not retroactively change the target or demonstrate that the same result could have been produced with fewer executions. Correctness labels are used for evaluation, not by the outcome adapters, rank fusion, or stopping decision.

A.2 Baseline methods and their adaptation to agent trajectories

We compare TURF-Fixed and TURF-Adaptive with 15 uncertainty baselines covering likelihood, output agreement, semantic structure, and within-trajectory uncertainty propagation. The selected implementations compute scores without fitting a task-specific correctness or reward predictor. Frozen pretrained semantic models are permitted: *task-training-free* does not mean that their encoders were never trained. All scores are oriented so that larger values indicate greater uncertainty, and performance is measured by ranking task failures rather than by interpreting the scores as calibrated failure probabilities. The method families below describe their underlying signals; the subsequent implementation specifications identify the versions actually evaluated.

Likelihood and relevance-based methods. **Perplexity (PPL)** summarizes token-level predictive surprise, while **predictive entropy (PE)** uses sampled sequence surprisals to estimate uncertainty in the predictive distribution (Malinin & Gales, 2021). They test whether uncertainty already present in the agent's generation probabilities is sufficient for failure ranking. **SentSAR** is the sentence-level relevance variant of SAR (Duan et al., 2024); it combines response likelihoods with relevance between sampled responses, reducing sensitivity to surface variation that does not change meaning. SentSAR is therefore a likelihood–similarity hybrid, rather than a purely likelihood-based statistic.

Output agreement and semantic structure. **Lexical similarity (LS)** measures variation among sampled outputs using lexical agreement (Fomicheva et al., 2020). **Semantic entropy (SE)** aggregates probability mass within groups of equivalent responses before measuring uncertainty (Kuhn et al., 2023; Farquhar et al., 2024). Several baselines instead retain graded pairwise relationships: **semantic nearest-neighbor entropy (SNNE)** summarizes similarity-based response neighborhoods (Nguyen et al., 2025); **kernel language entropy (KLE)** measures spectral dispersion through the von Neumann entropy of a semantic kernel (Nikitin et al., 2024); and **Degree** and **EigV** use degree-based and eigenvalue-based summaries of a response-similarity graph, respectively (Lin et al., 2024). **Semantic density (SD)** measures probability-weighted support around a response in semantic space (Qiu & Miikkulainen, 2024). **Semantic graph density (SGD)** combines fine-grained semantic relationships with response probabilities when weighting graph edges (Li et al., 2025). Finally, **CoCoA-MaxProb** and **CoCoA-PPL** combine semantic consistency with confidence derived from maximum probability and perplexity, respectively (Vashurin et al., 2026). Together, these ten baselines test alternatives to discrete outcome dispersion, including hard grouping, local density, graph structure, and confidence–consistency combinations.

Trajectory-oriented methods. **SAUP** aggregates step uncertainty using situation-aware weights, allowing the execution context to affect the contribution of an uncertain step (Zhao et al., 2025). Our comparison uses its registered distance-based implementation, not the learned hidden-Markov-model surrogate. **UProp** distinguishes uncertainty intrinsic to a decision from uncertainty propagated from earlier decisions (Duan et al., 2025). These methods provide process-oriented comparators to TURF's dispersion across observable execution outcomes and terminal responses. Their trajectory-level scores are aggregated over the evaluated rollout prefix; they are not replaced by generic response-set scores.

Which version appears in the primary comparison? A method name alone does not specify how a response-level estimator is applied to an agent. *Final* scores the terminal-response set from the sampled trajectories, with terminal likelihoods when required. *StepAgg* computes uncertainty at

Primary version	Baselines
Final	SNNE, KLE, EigV, SGD, CoCoA-MaxProb, CoCoA-PPL
StepAgg	PPL, PE, SentSAR, LS, SE, SD, Degree
Trajectory-specific	SAUP, UProp

Table 4: **Baseline versions in the primary comparison.** The 15 short names denote the registered implementations above. Degree and UProp retain an $N = 4$ local-candidate contract; N is distinct from the trajectory budget B. The separate matched adaptation study evaluates both Final and StepAgg for all 13 response-level families.

eligible agent decision steps, averages within each trajectory, and then averages across the sampled trajectories, retaining the registered eligibility and missing-value rules. It does not compare unrelated step positions across trajectories. SAUP and UProp retain their method-specific within-trajectory aggregation. Table 4 records the primary implementations. Plot legends omit the suffixes only for readability; each curve uses its registered adaptation at every budget.

Implementation fidelity and frozen similarity models. The results concern the documented adaptations, not necessarily exact reproductions of every configuration in the source papers. For example, the registered local LS, SE, SD, SentSAR, and Degree implementations use fuzzy-string decision similarity; their semantic labels should not be read as implying a shared NLI backend. SNNE uses the authors' `only_denom` variant with ROUGE-L F1, self-similarity included, and temperature one. The NLI-based Final implementations use cached frozen semantic evidence. SGD is an independent reimplementation and uses the shared DeBERTa-v2-xlarge-MNLI backend because its separately specified DeBERTa-large checkpoint was unavailable in the offline environment. SAUP uses its separate frozen `deepset/roberta-base-squad2` encoder with attention-mask mean pooling and maximum length 512; its prefix scores are recomputed from trajectory text rather than substituted from TURF's terminal NLI kernel. The archived SAUP variant follows the pinned Agent-Tracer code's generic-distance weighting, with unit coefficients and an additive-one term, and is not presented as an exact reproduction of every original SAUP surrogate. Historical likelihood spans, fallback rules, and score orientations are retained and recorded with the evaluation artifacts.

Budget, support, and adaptation controls. The primary comparison spans six datasets, three backbones, and three seeds, with trajectory budgets $B \in \{2, 3, 4, 6, 8, 12, 16\}$ where supported. Degree and UProp have complete 54-cell results only through $B = 12$; the other 13 baselines extend to $B = 16$. Method-specific task eligibility and source pools are retained, including the candidate-enriched pools used for Degree and UProp. Consequently, primary macro curves are not all-method matched-support comparisons, and equal trajectory counts do not imply equal total compute: local-candidate generation, likelihood access, and semantic-model inference can incur different costs. In particular, sharing the same 54 dataset–backbone–seed cells does not mean that every method scores identical task records or consumes identical inputs. Primary performance differences can reflect both the uncertainty method and these support or data-setup differences; they are descriptive comparisons, not isolated effects of the scoring algorithm on a common input set.

To examine adaptation separately, Appendix E evaluates both Final and StepAgg for the 13 response-level families, giving 26 baseline variants on a common, label-blind support of 21,986 task–cell records. This secondary panel uses candidate-enriched $Z = 12, N = 4$ pools and the shared budgets $B \in \{2, 3, 4, 6, 8, 12\}$. Its StepAgg variants use four local candidates per eligible step and require the stipulated step scores to be available rather than silently dropping missing required scores. These controlled adaptations are distinct from the historical primary implementations; their results are not pooled with the original $Z = 16$ comparison.

Scope of the selected baseline panel. We do not include ProbeCal's reward-trained calibration model (Liu et al., 2024a), which learns a mapping from language-model representations to execution reward. TRACER (Tayebati et al., 2026) is also outside the evaluated panel: its published parameter-identification procedure optimizes a failure–success separation objective using labeled trajectories. This distinction concerns that procedure, not an assertion that TRACER's underlying surprisal, repetition, and coherence features inherently require labels; a frozen-parameter adaptation could be

Table 5: **Maximum-pool launcher defaults.** Token limits are per generated decision, not total tokens per task. WebShop's three token limits correspond to Qwen, Gemma, and GPT-OSS, respectively. Finalized records and their provenance determine which cached executions are actually evaluated, including any per-run overrides.

Dataset	Maximum decision steps	Maximum output tokens/decision
AgentBench-DB	64	2,048
HotpotQA	6	512
WebShop	20	2,048 / 1,024 / 2,048
StrategyQA	6	512
AgentBench-OS	8	2,048
GAIA	15	768

evaluated separately. The comparison therefore covers these 15 selected implementations, not all agent-uncertainty methods.

A.3 Rollout generation, budgets, and data provenance

Agent backbones and generation settings. The generation configurations identify the checkpoints as `Qwen/Qwen3.5-9B`, `google/gemma-3-12b-it`, and `openai/gpt-oss-20b`. The maximum-pool launchers specify vLLM-backed sampling at temperature 1.0 and the source-code defaults in Table 5. Dataset-specific runners retain their task prompts, tools, response parsers, termination rules, and environment interfaces. A complete trajectory can terminate before its step limit. These are launcher defaults, not a census of effective settings for every cached execution. Archived per-run decoding overrides and accepted repair provenance take precedence when reproducing a record; reused executions retain their recorded settings. The separate label-free answer-normalization call for AgentBench-DB uses temperature zero.

Nested prefixes, not repeated generation. The primary nominal budget grid is $\mathcal{B} = \{2, 3, 4, 6, 8, 12, 16\}$. The budget B counts complete trajectories, not steps or candidate decisions. A smaller-budget evaluation reuses the first B trajectories in the archived order; we do not regenerate an independent pool at each budget. Fixed uses $n^\star = B$, while Adaptive selects $n^\star \leq B$ within that prefix. Task order and trajectory order are preserved. The three seeds index separate recorded generation cells; they are not training folds. TURF does not fit a risk predictor within or across those cells.

Executed trajectories versus local candidates. Four alternative responses at one intermediate decision do not constitute four complete trajectories. Candidate-dependent baselines can consume additional, separately cached responses at an executed context without replaying the environment along every alternative. The original pools, candidate-enriched pools, terminal NLI matrices, and local-step NLI matrices are distinct artifacts. We preserve their source selection rather than assume that a file with one materialized candidate has no cached alternatives, or that an $N = 4$ directory certifies every original context. Candidate-dependent source differences are separated below. In particular, Degree-anchored enriched pools need not equal the first trajectories of the original $Z = 16$ pool. Reusing candidates requires matching the executed path, decision context, parser, ordering, and sampling cursor, not merely task ID, seed, or position.

Missing parsed decisions are also distinct from missing generation. The finalized GAIA audit found four nonempty raw responses with valid likelihood metadata at all 153,714 recorded steps, but 294 GPT-OSS steps contained an empty parsed candidate. Those are representation failures, not missing responses. The Final/StepAgg panel declares its observable-text fallback; the primary baselines retain their own input contracts.

A.4 Frozen TURF configuration and label-free information

The primary results hold the following choices fixed across backbones, seeds, and budgets, except for the declared WebShop specialization. Implementation formulas and numerical details are given in Appendix B.

Scoring. Count posterior risk uses half-count smoothing for each observed outcome and a total unseen half-count, with dispersion penalty $\kappa = \Phi^{-1}(0.975)$ and the small chosen-output-surprisal tie term. UnknownCap2 handles censored outcomes on the five non-WebShop datasets. WebShop uses SeverityAware-v1, hierarchy weight $\gamma = 1/8$, and equal weighting of the core-score rank and severity rank; hierarchical moments use 2,048 deterministically seeded Dirichlet draws. These are label-free outcome constructions, not benchmark-correctness judgments. Their exact adapters and the WebShop specialization are specified in Appendix B.1.

The semantic view uses frozen `microsoft/deberta-v2-xlarge-mnli`, task-conditioned terminal texts, both entailment directions, and maximum tokenized pair length 512. It takes the minimum directional entailment affinity, applies the stated PSD repair, and computes normalized spectral entropy with uniform rollout weights. It does not encode the internal temporal sequence. The same cached semantic information is reused by scoring ablations where applicable; local-step baseline inference is a separate input contract.

Gaussianized-rank Stouffer fusion assigns equal weights to the two views. Component scores are rounded to 12 decimal places before midranking; normal-quantile inputs are clipped to $[10^{-6}, 1 - 10^{-6}]$. Each dataset–backbone–seed–policy–budget operating point has its own unlabeled reference cohort. Count ranks use all tasks in the cell, and semantic ranks use its covered subset; each reference task contributes the statistic at its own selected prefix. This is *transductive, task-training-free scoring*, not supervised fitting or an independently calibrated failure probability. The default includes each query in its own reference cohort. The 54-cell Fixed and Bayesian reference experiments separately test query exclusion and finite banks under their declared folds and support (Appendix B.4); neither changes the primary protocol.

Semantic availability and fallback. The primary semantic cache covers 22,158 of 22,230 task–cell instances. The remaining 72 use count-rank fallback: 51 AgentBench-OS, 14 HotpotQA, three GAIA, two AgentBench-DB, and two StrategyQA. TURF consequently retains full task coverage. Availability is inherited from the archived full-pool extraction and likelihood-metadata checks, not recomputed separately for every prefix. Thus a failure of a cache check on a later rollout can route a shorter prefix to count-only scoring and outcome-only stopping. This is future-dependent availability information, even though no correctness label is used by the rule. The replay is prefix-based only *conditional on this archived mask*; it is not a strictly online evaluation of the complete adaptive procedure. The reported analyses do not include a controlled re-evaluation with prefix-local availability, so its effects on stopping, coverage, and AUROC/AUPR remain unmeasured. Appendix B.3 distinguishes this cache restriction from the information mathematically needed by the scorer.

Bayesian two-view stopping. The default fixes rollout one as the anchor, thresholds raw symmetric anchor entailment at $\eta = 0.75$, and uses a four-category $\mathrm{Dirichlet}(1/2, 1/2, 1/2, 1/2)$ prior. Each marginal recurrence probability is tested against $1/2$. Starting at prefix five, the policy stops strictly before B when the two marginal posterior persistence probabilities sum to at least 1.8, corresponding to $\delta_{\mathrm{stop}} = 0.2$; otherwise it uses B. The semantic-missing branch requires only outcome persistence probability at least 0.8. Hence Fixed and Adaptive coincide at $B = 2, 3, 4$. Stopping compares pre-UnknownCap2 outcomes outside WebShop and mapped structured outcomes on WebShop; scoring retains its original encoded buckets. The same stopping parameters are used at every budget. The certificate concerns anchor recurrence under the stated model, not a uniform frequentist false-stop rate, correctness, or AUROC/AUPR preservation (Appendix C.5).

A.5 Baseline implementations and distinct comparison panels

Primary seven-budget comparison. The selected comparison contains the two TURF variants and 15 baselines. Its short legend names preserve specific implementations: SNNE, SGD, KLE, EigV, CoCoA-MaxProb, and CoCoA-PPL use terminal-response sets; PPL, LS, PE, SE, SD, SentSAR, and Degree use the archived step-aggregation implementations. UProp and SAUP retain their agent-oriented definitions. Degree and UProp retain their four-candidate contract and have complete six-dataset curves only through $B = 12$; the other selected methods extend to $B = 16$. Thus a native-endpoint table is not automatically a common-budget comparison. Dataset-specific panels

can retain additional supported points: Degree and UProp have $B = 16$ values for GAIA, but not a complete 54-cell point at that budget.

The primary comparison preserves method-specific task eligibility and source pools, including the candidate-enriched sources for Degree and UProp; it is not an all-method matched-support experiment. Semantic and likelihood-dependent baselines can have less than full task coverage. Score orientation is fixed by the registered implementation so that larger values indicate greater failure risk. Historical implementations must also be distinguished from a fresh reimplementation of the original papers: for example, SGD uses the shared DeBERTa-v2-xlarge backend because its separately specified DeBERTa-large checkpoint was unavailable offline. SAUP uses its separate frozen `deepset/roberta-base-squad2` semantic-distance encoder with attention-mask mean pooling and maximum length 512, and recomputes prefix scores rather than substituting the terminal NLI kernel.

Matched Final/StepAgg comparison. Appendix E answers a different question using 13 baseline families, each with both a Final and a StepAgg adaptation. It uses the candidate-enriched 12-trajectory pools on five datasets and the first 12 finalized GAIA trajectories, with $B \in \{2, 3, 4, 6, 8, 12\}$. Final scores terminal-response sets; StepAgg scores four local candidates per eligible step, averages within each trajectory, then averages over the trajectory prefix. The label-blind all-method $B = 12$ support contains 21,986 task–cell instances (98.90% of the population) and is held fixed at every budget. Missing required step scores are not silently dropped. The declared observable-text fallback and original likelihood-span semantics are retained. This panel uses its own outcome and semantic artifacts, including a full-12-trajectory semantic mask; Adaptive is recomputed on these pools, not imported from the original $Z = 16$ results. Its values must not be pooled with or substituted for the primary experiment. This is a complementary matched-support check on different pools; it does not make the original primary comparison coverage- or input-matched.

Component and allocation controls. Scoring ablations change a designated component while retaining the task support, outcome construction, other scoring components, and selected trajectory counts. In particular, a semantic-only core still has the declared fallback, and WebShop retains its outer wrapper. Holding lengths fixed imposes identical trajectory cost; it does not test whether a component could help another stopping policy. Budget-policy comparisons instead hold the scorer fixed. Fixed sampling tests a common cap, shuffled lengths preserve a policy's within-cell length histogram while removing task-specific allocation, and outcome-stability/dominance controls test simpler stopping signals. Matched-cost controls are constructed within cells using costs, not failure labels. Randomized mixtures of adjacent budgets or bracketing heuristics are distinct from a single fixed deterministic rule. The four-baseline campaign uses 100 schedules for its nonconstant matched Fixed/stability/dominance mixtures and reports residual cost mismatches; its approximate matches are not interchangeable with an exact-total-cost control from another campaign.

A.6 Metrics, aggregation, and resource accounting

We compute failure-positive AUROC and AUPR within each cell. AUPR is `average_precision_score`, not trapezoidal PR integration. It depends on failure prevalence; absolute AUPR should primarily be compared between methods on the same dataset and support. Let $M_{d,b,s}(B)$ be either metric for dataset d, backbone b, and seed s. The macro estimate is

$$\overline{M}(B) = \frac{1}{3} \sum_{s} \left[\frac{1}{18} \sum_{d} \sum_{b} M_{d,b,s}(B) \right]. \tag{13}$$

Thus all 54 cells have equal weight; the 1,000-task HotpotQA cells do not dominate the 139-task GAIA cells. We do not pool separately ranked task scores across cells. Dataset panels average three backbones and three seeds; backbone panels average six datasets and three seeds; dataset–backbone panels average the three seeds. A budget-averaged summary, when explicitly reported, averages the stated budget grid equally; it is not an area under the actual-cost curve.

For task i in cell c, let $n^{\star}_{ci}(B)$ denote its inspected trajectory count and N_c the number of tasks in that comparison. The realized trajectory-count coordinate and same-cap trajectory savings are

$$\overline{n}(B) = \frac{1}{54} \sum_{c} \frac{1}{N_c} \sum_{i} n^{\star}_{ci}(B), \qquad \text{Savings}(B) = 100 \left(1 - \frac{\overline{n}(B)}{B} \right). \tag{14}$$

Fixed lies at B; an Adaptive marker can lie to its left. Connecting markers does not create additional evaluated budgets. A same-cap quality comparison is not matched in realized cost, and a favorable curve alone does not identify a benefit from task-specific allocation. Paired differences are recomputed on the support declared for that comparison. Metric differences in percentage points are $100(M_{\text{method}} - M_{\text{control}})$; positive values favor the named method unless a caption explicitly states another contrast.

Trajectory counts measure inspected cached rollouts, not end-to-end wall-clock or GPU savings. The resource appendix additionally measures recorded decision steps and output tokens, using equal-cell averages of within-cell retained/full-cap resource ratios. Its denominators and complete-record restrictions are specified there. These counts do not include all input tokens, candidate-generation work, NLI inference, reference construction, tool latency, or unlogged retries. The number of trajectories required to produce an archived aggregate answer is also separate from the number inspected for uncertainty. Throughout the budget-policy analyses, "matched cost" means matched complete-trajectory counts under the stated control construction, not matched total compute across methods.

A.7 What the uncertainty bands and intervals mean

Different appendix analyses quantify different sources of variation. They should not all be read as the same "95% interval."

- **Cross-seed curve ribbons.** These show ± 1 sample standard deviation (`ddof=1`) across three seed-level estimates, after the specified dataset/backbone averaging. They are not confidence intervals, standard errors, or uncertainty over 54 independent domains.
- **Paired task-bootstrap intervals.** Principal component effects and same-cap adaptive retention use 50,000 paired task resamples. Within each dataset, canonical tasks are sampled jointly across all nine backbone–seed cells; paired methods receive the same resample. Metrics are recomputed within each cell before equal-cell aggregation. Primary ranks and stopping decisions remain fixed. The resulting pointwise 95% basic-bootstrap intervals are not obtained by averaging individual-cell confidence limits. They condition on the observed datasets and frozen score construction. Rank-rebuilding resamples, where used, are identified separately.
- **Dataset-block diagnostic intervals.** Several baseline and matched-cost plots instead use 50,000 resamples of the six dataset-level paired effects, retaining their backbone–seed averages. These descriptive percentile intervals characterize sensitivity to the mix of the six datasets, conditional on the observed task scores and reported configuration; they omit task-sampling uncertainty.
- **Final/StepAgg intervals and reference-bank variability.** The matched adaptation study resamples backbones and seeds within the six fixed datasets. Where shown, reference-bank bars describe variation across bank draws under that experiment's convention; the rotating-fold Fixed/Bayesian bank tables report draw means without new confidence intervals. Neither is the task-bootstrap analysis above.

Intervals are pointwise and unadjusted unless explicitly stated. The original Fixed-$B = 16$ ablation additionally uses Holm correction for its eight designated scoring/execution-information tests; this does not extend automatically to other budgets, Bayesian-policy contrasts, or all appendix comparisons. An interval crossing zero does not establish equivalence or absence of a useful effect. Conversely, exclusion of zero in a pointwise interval does not establish simultaneous significance across comparisons.

A.8 Reading the remaining appendix and reproducing the study

The appendix proceeds from implementation and interpretation to comparisons and controlled variations: Appendix B specifies outcome extraction, numerical scoring, references, and archived labels; Appendix C states and proves the four-component analysis, including the model-relative stopping properties; Appendix D presents macro and dataset–backbone curves; Appendix E isolates baseline adaptations; Appendix F examines component and policy effects; Appendix G separates

quality retention, allocation, and recorded resource reductions; and Appendix H varies scoring, reference, and Bayesian-policy settings. Their captions identify any departure from the common protocol, especially task support and interval construction.

Reproducibility and scope. The source inventory distinguishes original trajectories, local candidate caches, outcome checkpoints, terminal semantic caches, evaluation labels, and final metric tables. Reproduction from frozen records requires preserving task IDs, trajectory order, model-input representations, source hashes, availability masks, and the registered metric implementation; regeneration additionally requires the original prompts, environment and model versions, sampling configuration, and repair provenance. CPU-based replay and cached scoring should not be confused with the GPU inference that produced those inputs. The selected-baseline report retains 6,336 metric rows for two TURF variants and 15 baselines. Its source-parity receipt preserves the verified Bayesian replacement values: all 378 Fixed cell–budget pairs and all 378 Adaptive operating points, alongside unchanged selected baseline metrics.

Here, "task-training-free" means that TURF keeps its pretrained models frozen and fits neither a risk predictor nor fusion weights to task correctness labels.

B Implementation Details

This appendix specifies the frozen six-dataset TURF configuration: UnknownCap2 on five datasets and WebShop-SeverityAware-v1 on WebShop. All logarithms are natural. The empirical protocol is specified in Appendix A.

B.1 Outcome construction

The adapter consumes the task specification and generated execution, not the reference answer, benchmark reward, or official correctness. Outcome buckets are aligned with the persisted rollout order. Table 6 summarizes the source representations.

Table 6: Observable outcomes. The WebShop row describes the reported hierarchical mapping, not the original product-identity-only encoding.

Dataset	Label-free outcome
StrategyQA	Valid Boolean or stripped, case-folded yes/no answer; otherwise `no-answer:<status>`, defaulting to unknown.
HotpotQA	Stripped, lowercased, whitespace-collapsed answer; otherwise a no-answer status. No gold aliases, article removal, or punctuation matching are used for this outcome mapping.
AgentBench-DB	Cached normalized answer for read tasks; replayed database state for writes, falling back to write SQL/status signatures.
AgentBench-OS	Answer normalized using the task-provided answer schema (integer, binary-unit size, or text), with explicit invalid/no-answer outcomes.
GAIA	Stripped, case-folded, whitespace-collapsed final answer, or `<no-answer>`.
WebShop	Purchase with request-coverage, selected-option, and price signature; otherwise the detailed no-purchase outcome.

DBBench preprocessing. For answer-returning tasks, the runner can call the backbone at temperature zero to reformat the candidate answer into the requested answer type. This call sees the question and candidate, not the reference. The cached `semantic_answer` is preferred when forming the outcome. Write tasks instead use a replayed resulting-state signature when available; fallbacks distinguish executed write operations, errors, and no-write status. The stored UProp outcome sequence used by this report applies the bucket function independently to each rollout. It does not use the separate pooled semantic canonicalizer from the OE/SBOE evaluation paths. Thus outcome extraction is deterministic conditional on the enriched record, but not necessarily pure string parsing of the raw rollout. These preprocessing calls must be counted when reproducing the pipeline from uncached executions.

UnknownCap2. Normalize strings by case folding and whitespace collapse for censor recognition. A bucket is censored if it is `<no-answer>` or begins with `no-answer:` or `no-purchase:`. Its reason key consists of the family and first colon-delimited reason; `<no-answer>` has reason `no-answer:generic`. Let u_t be the number of preceding censored events with reason `no-answer:unknown`. Such an event receives a bucket indexed by $u_t \bmod 2$. Every other censored event receives a bucket containing its rollout index and normalized original bucket. Noncensored buckets are unchanged. In particular, generic missing answers do not receive the unknown-event cap. This encoding is applied on the five non-WebShop datasets for count scoring; no future bucket alters an earlier encoding. Bayesian two-view stopping instead compares the pre-UnknownCap2 buckets a_j. Its outcome recurrence signal can therefore treat repeated raw no-answer events as matches even when the scoring buckets b_j distinguish them. It is a design choice for censored outcomes, not an estimate of their latent true answers.

The choice of two is an operational compromise: one bucket treats all repeated unknown events as identical, whereas a fresh bucket per event makes their artificial support grow with the number of rollouts. Two is the smallest nontrivial cap that separates these repetitions while bounding the number of buckets assigned to this unknown-event type. It does not bound the total number of scoring buckets, which can also include observed answers and other censored events. This rationale does not establish two as an optimal cap or guarantee better failure ranking. The resulting dispersion is induced by the encoding, not evidence that two distinct latent answers occurred.

WebShop outcome mapping. The base adapter tracks executed shopping actions to identify the purchased product and selected options, or the most informative no-purchase state. The evaluated mapping replaces purchased-product identity with a request-constraint signature. Request tokens are matched against the observed product title and selected options using the frozen tokenizer and stopword list. Coverage is the number of distinct matched request tokens divided by $\max(1, |\text{request tokens}|)$. We discretize coverage into zero, low $(0, 0.25)$, mid $[0.25, 0.5)$, and high $[0.5, 1]$. The signature also records whether selected-option tokens overlap the request and whether the observed price satisfies an extracted upper-price limit. An unextractable limit or missing price is assigned *unknown*, not *pass*. No-purchase details are retained. These features use observable execution records and request constraints, not official reward or correctness judgments.

Hierarchical count risk. For WebShop, the count functional distinguishes purchase/no-purchase variation from variation within each branch:

$$\begin{aligned} h_{\mathrm{WS}}(q) = H_q(\chi) + \gamma q(\chi = p) H_q(\phi \mid \chi = p) \\ + q(\chi = f) H_q(F \mid \chi = f), \qquad \gamma = \tfrac{1}{8}. \end{aligned} \tag{15}$$

Here χ denotes purchase (p) or no purchase (f), ϕ is the purchased item's request-constraint signature, and F records no-purchase detail. The weight γ discounts differences among purchased configurations relative to purchase/no-purchase variation. The smoother adds a half-count to each observed leaf and reserves an unseen quarter-count in each branch. The functional's moments are estimated by Monte Carlo, as specified in Appendix B.2.

Severity adjustment. The frozen SeverityAware-v1 configuration additionally summarizes observable adverse execution outcomes. This is a third, rule-based scoring signal specific to WebShop, *in addition to* the shared count and terminal-semantic views; it is not merely a different outcome-bucket encoding. For rollout t, let χ_t denote its purchase status, covbin_t its coverage bin, and $\mathrm{priceflag}_t$ its price-check status. Define

$$V_n = \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}\Big\{ \chi_t = f \ \vee \ \mathrm{covbin}_t \neq \mathrm{high} \ \vee \ \mathrm{priceflag}_t \neq \mathrm{pass} \Big\}. \tag{16}$$

The logical connector is OR: a rollout contributes one if it makes no purchase, has zero/low/mid coverage, or has a nonpassing or unknown price status. Each rollout contributes at most one, regardless of how many conditions hold. The final uncertainty score is

$$S_n = \begin{cases} \frac{1}{2} R_S(S_n^{\mathrm{core}}) + \frac{1}{2} R_V(V_n), & d = \mathrm{WebShop}, \\ S_n^{\mathrm{core}}, & \text{otherwise}, \end{cases} \tag{17}$$

where R_S and R_V are empirical midrank maps constructed from the same dataset–backbone–seed–policy–budget reference cohort. Both TURF-Fixed and TURF-Adaptive use this frozen adjustment. It requires no additional model call and does not change the selected prefix. The Gaussian fusion rationale applies to S_n^{core}, not to this additional dataset-specific adjustment. Consequently, WebShop's reported results describe the two-view core *with* this extra signal, rather than the core alone. Appendix H.3 reports the core-only, severity-only, and mixed scores; the default gives equal weight to their two outer ranks.

B.2 Count posterior and likelihood stabilization

Ordinary entropy moments. At a fixed prefix, write $K = K_n^{\mathrm{out}} + 1$, including the unseen coordinate. In Eq. (2), $\alpha_k = N_{nk} + 1/2$ for observed buckets and $\alpha_K = 1/2$. Let $\alpha_\Sigma = \sum_k \alpha_k$ and let ψ, ψ_1 denote digamma and trigamma. For $H(q) = -\sum_k q_k \log q_k$, the analytic moments are

$$\begin{aligned}
\mu &= \psi(\alpha_\Sigma + 1) - \sum_k \frac{\alpha_k}{\alpha_\Sigma} \psi(\alpha_k + 1), \\
\mathbb{E}[H(q)^2] &= \sum_k \frac{\alpha_k(\alpha_k+1)}{\alpha_\Sigma(\alpha_\Sigma+1)} \left\{ [\psi(\alpha_k+2) - \psi(\alpha_\Sigma+2)]^2 + \psi_1(\alpha_k+2) - \psi_1(\alpha_\Sigma+2) \right\} \\
&\quad + \sum_{k \neq j} \frac{\alpha_k \alpha_j}{\alpha_\Sigma(\alpha_\Sigma+1)} \left\{ [\psi(\alpha_k+1) - \psi(\alpha_\Sigma+2)][\psi(\alpha_j+1) - \psi(\alpha_\Sigma+2)] - \psi_1(\alpha_\Sigma+2) \right\}.
\end{aligned} \tag{18}$$

The cross sum is over ordered pairs. We use $v = \max(0, \mathbb{E}[H^2] - \mu^2)$ and the variance floor in Eq. (3); the stored penalty constant is $\kappa = \Phi^{-1}(0.975) \simeq 1.959964$. These are moments of the fixed-prefix smoother, not a correctness posterior.

WebShop moments. Sort the observed leaf labels and assign each count plus 0.5. Append two unseen leaves, purchase and no-purchase, with parameters 0.25 each. Evaluate Eq. (15) on 2,048 Dirichlet draws; use their sample mean and population variance (`ddof=0`). For exact reproducibility, the NumPy default generator seed is the first eight SHA-256 digest bytes, interpreted little-endian and reduced modulo 2^{32}, of

```
webshop-hierarchical|<sample_id>|<n>|
    hierarchical-gamma=0.25|0.5.
```

The historical seed string retains `gamma=0.25` even though the scored functional uses $\gamma = 0.125$. Consequently, equal count patterns on different task identifiers can have slightly different Monte Carlo moment estimates.

Two-level surprisal. For rollout t, let $\mathcal{J}_t$ contain steps with finite chosen-output log-probability sum ℓ_{tj} and positive chosen-output token count m_{tj}. Set

$$r_t = \frac{1}{|\mathcal{J}_t|} \sum_{j \in \mathcal{J}_t} \frac{-\ell_{tj}}{m_{tj}}, \qquad \bar{r}_n = \frac{1}{|\mathcal{T}_n|} \sum_{t \in \mathcal{T}_n} r_t, \tag{19}$$

$\mathcal{T}_n$ contains prefix rollouts with finite r_t; an empty $\mathcal{J}_t$ makes r_t missing. If $\mathcal{T}_n$ is empty, use $\bar{r}_n = 0$. This averages steps and rollouts equally, not all tokens equally. It uses chosen outputs throughout the rollout, not only the final response; it does not encode their temporal order.

Numerical scale. For each integer partition of each total $m \in \{2, \ldots, M\}$, compute the ordinary entropy statistic $\mu - \kappa\sqrt{\max(v, 10^{-12})}$. Round these values to 14 decimals, keep distinct values, and let Δ_M be the smallest adjacent gap strictly greater than 10^{-12}. Set $\epsilon(M) = \Delta_M/4$, or 10^{-12} if no such gap exists. The five ordinary datasets use $\epsilon_{d,B,n} = \epsilon(B)$; WebShop uses $\epsilon_{\mathrm{WS},B,n} = \epsilon(n)$. The added term lies in $[0, \epsilon_{d,B,n})$. It is intended to resolve discrete count ties at a small scale, but the rounded grid excludes tiny gaps and does not characterize WebShop's hierarchical Monte Carlo scores. We therefore make no universal strict-order-preservation claim.

B.3 SEMANTIC INPUTS AND NUMERICAL KERNEL CONSTRUCTION

Frozen encoder and terminal text. We use the frozen DeBERTa-v2-xlarge MNLI model, identified by `microsoft/deberta-v2-xlarge-mnli`. The extractor normally takes the committed final answer. If absent or marked as hard-finalized, it uses the last realized decision; a terminal action can therefore serve as the text. The existing semantic cache also requires finite terminal decision log-probability metadata, a positive integer decision-token count, and average log probability at least -100. A task enters that cache only if all sixteen terminal responses pass extraction. This historical availability mask is held fixed when replaying smaller prefixes, even though Eq. (5) uses uniform weights and does not require likelihood weighting.

Why the archived check uses the full pool. The sixteen-rollout requirement is inherited from the task-level cache-admission procedure, not from NLI, spectral entropy, or the Bayesian recurrence model. At a prefix n, spectral scoring requires terminal texts and affinities only for the selected n rollouts; two-view stopping requires only their outcome and anchor-match indicators. A likelihood-metadata requirement is likewise not intrinsic to the uniformly weighted spectral score. Nevertheless, a later extraction or metadata failure can change the branch used to score and stop at an earlier prefix under the archived protocol. Retaining this mask reproduces the reported comparison; it does not establish that an online implementation would make the same decisions.

A prospective protocol must define semantic availability using only the inspected prefix, specify how missing terminal evidence is handled, and then recompute stopping decisions, reference ranks, and scores. No results for that controlled replacement are reported here. The small number of full-pool fallback cases does not by itself show that the replacement has negligible effects, particularly because reference ranks depend on which tasks supply semantic statistics.

Each NLI input concatenates the original task prompt and one terminal text. Both pair directions are evaluated with joint tokenization, truncation, and maximum length 512. Only entailment probabilities are used; the neutral-class contribution is zero. The cached class order is contradiction, neutral, entailment. The NLI weights are not adapted on the experimental tasks.

PSD repair. Starting from the symmetric affinity in Eq. (4), clip entries to $[0, 1]$ and set the diagonal to one. If $\mathbf{A}_n = U\Lambda U^\top$, reconstruct $\mathbf{K}_n^{(0)} = U \max(\Lambda, 0)U^\top$. Let $d_j = \sqrt{\max((\mathbf{K}_n^{(0)})_{jj}, 10^{-12})}$ and set $(\mathbf{K}_n)_{jk} = (\mathbf{K}_n^{(0)})_{jk}/(d_j d_k)$. The implementation symmetrizes, clips numerical entries to $[-1, 1]$, and restores the unit diagonal. This is an eigenvalue-based repair, not a nearest-correlation optimization. For entropy evaluation, density eigenvalues are clipped below at zero and renormalized to sum to one; eigenvalues no greater than 10^{-12} are omitted from the log sum. In exact arithmetic the repaired correlation matrix is PSD with trace n.

B.4 RANK REFERENCES AND MISSING SEMANTIC COVERAGE

For each dataset–backbone–seed–policy–budget cell, compute each task's statistics at its selected prefix. After rounding a component to 12 decimal places, its reference rank is

$$R_v(\xi_i) = \frac{\mathrm{midrank}(\mathrm{round}_{12}(\xi_i)) - 1/2}{M_v}, \qquad v \in \{C, H\}. \tag{20}$$

Here ξ_i denotes task i's statistic for the indicated view. Midranks are one-indexed and ties receive their average rank. The count reference contains all scored tasks (M_C); the semantic reference contains only covered tasks (M_H). For a covered task, its count rank still comes from the full count cohort, not a re-ranking within the covered subset. Ranks supplied to Φ^{-1} are clipped to $[10^{-6}, 1 - 10^{-6}]$. WebShop's outer core/severity ranks follow the same midrank convention. To reproduce the archived WebShop ties, its core Stouffer score is cast to `float32` and back before the outer core rank is computed.

The primary results include each task in its own unlabeled reference cohort. This is a transductive transformation, not supervised fitting. The final score therefore requires the reference cohort; the individual stopping decision does not, conditional on its fixed semantic-availability mask. The archived mask itself is not a prefix-local deployment rule. Exact midranks are invariant to strictly increasing transformations; the implemented version has this invariance only when the transformation preserves rounded ordering and tie groups.

Missing semantic evidence. For a task without semantic evidence, the core score falls back to the count reference rank:

$$S_n^{\text{core}} = R_C(\widetilde{C}_n).$$

No semantic statistic is imputed, and the count reference remains the full scored cohort. Any documented environment-specific scoring adjustment is applied afterward. TURF-Fixed and TURF-Adaptive use the same scoring and fallback rules; their reference values can nevertheless differ because each task contributes statistics at its policy-selected prefix.

Score interpretation. Rank normalization aligns component scales without using correctness labels, and symmetric Stouffer fusion requires no label-fitted weights. The outer normal CDF is a monotone rescaling, not calibration to a failure probability. Scores constructed from different reference cohorts need not be numerically comparable. The conditional Gaussian rationale and rank-sensitivity analysis concern the two-view core fusion, rather than the count-only fallback or an environment-specific outer adjustment (Appendix C).

Separate reference experiments. The reported 54-cell Fixed and Bayesian checks remap cached components using the tie-aware external CDF $[\#\{v_j < v\} + \frac{1}{2}\#\{v_j = v\}]/M$ after 12-decimal rounding, with the same clipping before Φ^{-1}. Five-fold out-of-fold maps use the other four folds for every query. Rotating-fold finite-bank analyses score all tasks using banks of 25, 50, or 100 reference tasks and ten deterministic draws per fold; Fixed and Adaptive use separately seeded draws. The distinct held-out-query sensitivity study reserves one fold and uses nested bank sizes, with 200-task banks restricted to 27 cells. These protocols and their numerical results are distinguished in Appendices F and H.2. They test the specified cached-score remappings, not invariance to every reference construction, shifted-domain banks, or prospective acquisition and availability policies. A deployment with a frozen unlabeled bank is therefore distinct from the primary full-cell protocol.

B.5 Bayesian stopping and execution contract

Bayesian two-view fixes the first rollout as the anchor. On the five non-WebShop datasets, its outcome signal compares raw adapter buckets before UnknownCap2; on WebShop it compares the mapped structured outcomes. Its semantic signal uses the minimum of the two raw anchor entailment probabilities, thresholded at $\eta = 0.75$, before any PSD repair. For each view v, the prefix has $n-1$ post-anchor observations and $m_v(n)$ matches. The evaluated four-category prior has half-counts per category, giving marginal posterior $\mathrm{Beta}(1 + m_v(n), n - m_v(n))$.

Beginning at prefix five, evaluate the two marginal tail probabilities in Eq. (8). The first prefix strictly below B at which their sum is at least 1.8 is selected. Otherwise use B. Thus $B \le 5$ implies fixed-budget sampling. If semantic evidence is unavailable under the archived mask, use $\pi_{\text{out}}(n) \ge 0.8$ with the same minimum prefix and cap. This fallback provides only outcome-recurrence evidence. The posterior certificate and its assumptions are addressed in Appendix C.5.

Replay procedure. For each task, retain its ordered outcomes, rollout surprisals, and semantic cache identity. For each budget and policy: (i) construct stopping outcomes a_j and raw anchor affinities, and scan prefixes to select $n^\star$; (ii) use encoded scoring buckets b_j to compute the count input, and compute the available spectral input on that same prefix; (iii) form the operating-point reference ranks; (iv) fuse or use count fallback, then apply the WebShop wrapper if relevant; (v) evaluate the fixed archived label against the resulting score. Correctness is not an argument to steps (i)–(iv). Some source files contain labels and scores together; this is a functional separation, not a claim that every original artifact was physically label-isolated.

Cost interpretation. Counting outcomes requires one pass after outcome construction; analytic moment evaluation adds a quadratic sum in the number of represented categories. WebShop uses a fixed 2,048-draw calculation. Evaluating both directions of anchor affinity for the stopping rule requires $O(n)$ directed NLI pairs through prefix n. Given those indicators and a precomputed binomial-CDF table, the stopping calculation maintains two counts and performs two lookups per prefix. Computing the spectral score at the selected prefix requires $O(n^2)$ directed NLI pairs in total and an $O(n^3)$ eigendecomposition; anchor pairs can be reused. Reference sorting costs $O(M \log M)$ per component. The experiments replay already generated pools and cached semantic matrices.

Their $n^\star$ measures how many complete rollouts are retained in the uncertainty-scoring prefix; it does not measure actual saved GPU time, tokens, tool calls, or preprocessing calls. The full pools were generated before this replay. Shorter scoring prefixes therefore demonstrate reduced evidence use, not measured avoidance of the original generation work. A prospective implementation must also specify a prefix-local semantic availability policy and a reference bank/cohort. It must account separately for producing the task-level answer when that answer requires rollouts beyond $n^\star$.

B.6 Evaluation target and aggregation

The primary report uses one archived task-level `is_correct` field per sixteen-rollout record, converted to a failure label before scoring budgets. For the five-dataset checkpoint materializer, any value other than Boolean true becomes failure; the GAIA loader uses the negation of its saved Boolean correctness. These labels are fixed across TURF's budgets and policies. Table 7 specifies the source-generator aggregation; in particular, no common first-rollout correctness target was used.

Table 7: Archived task-level target produced by the source generators. These evaluation aggregations are not inputs to TURF's score.

Dataset	Source of archived correctness
AgentBench-DB	Select the modal nonempty committed raw answer. Correct if any matching-answer execution has an official true flag; false if matching flags exist but none is true; otherwise missing. This differs from the separate postprocessed-answer correctness field.
HotpotQA	Compare the modal nonempty normalized answer with the normalized reference; first-occurrence tie breaking.
StrategyQA	Compare the modal valid yes/no normalized answer with the reference; first-occurrence tie breaking.
WebShop	Modal Boolean official correctness among available execution results; first-occurrence tie breaking.
AgentBench-OS	Modal Boolean official correctness among available execution results; first-occurrence tie breaking.
GAIA	Evaluate the modal nonempty normalized final answer with the GAIA answer checker; first-occurrence tie breaking.

Population, the 72 count-fallback instances, and equal-cell metric aggregation are specified in Appendix A. Fallback retains coverage but need not preserve rankings relative to semantically covered scores. Both the target and the full-pool availability mask remain fixed as the inspected prefix changes; producing that target can require more executions than the uncertainty score inspects. The separate Final/StepAgg panel inherits its own native labels and full-12-trajectory cache mask (Appendix E).

C Supporting Theory: Statements, Proofs, and Scope

This appendix supplies the four supporting propositions summarized in Section 4. It distinguishes exact properties of the declared scoring construction from the Gaussian working-model rationale and the Bayesian recurrence certificate. The stopping subsection also derives the posterior calculation, gives a fixed-parameter counterexample to frequentist false-stop control, and states the scope of the savings bound. None of these results alone guarantees improved AUROC or AUPR on the evaluated tasks.

C.1 Count risk: reweighting robustness and numerical effects

At a fixed prefix, let P_n denote the declared distribution over outcome probabilities q_n, and let $X_n = h_d(q_n)$. Write $\mu_n = \mathbb{E}_{\mathsf{P}_n} X_n$ and $v_n = \mathrm{Var}_{\mathsf{P}_n}(X_n)$. This is the scoring smoother, not a posterior over task correctness. The idealized count statistic is $C_n^\circ = \mu_n - \kappa\sqrt{v_n}$, with $\kappa \geq 0$ and finite v_n.

Proposition C.1 (A lower bound under reweightings of the smoother) *For every probability law* $\mathsf{Q} \ll \mathsf{P}_n$ *satisfying* $\chi^2(\mathsf{Q}\|\mathsf{P}_n) \le \kappa^2$,

$$\mathbb{E}_{\mathsf{Q}} X_n \ge C_n^\circ, \qquad \chi^2(\mathsf{Q}\|\mathsf{P}_n) = \mathbb{E}_{\mathsf{P}_n}\left[\left(\frac{d\mathsf{Q}}{d\mathsf{P}_n} - 1\right)^2\right]. \tag{21}$$

If $v_n > 0$ *and* $1 - \kappa(X_n - \mu_n)/\sqrt{v_n} \ge 0$ P_n*-almost surely, the bound is attained by this density.*

Proof of Proposition C.1. Let $w = d\mathsf{Q}/d\mathsf{P}_n$, so $\mathbb{E}_{\mathsf{P}_n} w = 1$. By Cauchy–Schwarz,

$$\begin{aligned}\mathbb{E}_{\mathsf{Q}} X_n - \mu_n &= \mathbb{E}_{\mathsf{P}_n}[(w-1)(X_n - \mu_n)] \\ &\ge -\sqrt{\mathbb{E}_{\mathsf{P}_n}(w-1)^2}\sqrt{\mathbb{E}_{\mathsf{P}_n}(X_n-\mu_n)^2} \\ &\ge -\kappa\sqrt{v_n}.\end{aligned}$$

When $v_n > 0$ and $w_\star = 1 - \kappa(X_n - \mu_n)/\sqrt{v_n}$ is nonnegative, it is an admissible probability density with $\chi^2(\mathsf{Q}_\star\|\mathsf{P}_n) = \kappa^2$ and attains the lower bound. If $v_n = 0$, X_n is constant almost surely and every absolutely continuous reweighting has that same expectation. Otherwise, nonnegativity can prevent attainment: the proposition supplies a lower bound, not an unconditional equality with a distributionally robust optimum. □

The argument is an elementary moment inequality, consistent with the broader connection between variance penalties and distributional robustness (Duchi & Namkoong, 2019). It does not require P_n to be a correctly specified generative posterior and does not assert robustness of the true task-outcome distribution. Here P_n is the declared fixed-prefix scoring law, which can be a pseudo-posterior.

Variance floor, tie-breaking, and estimated moments. Let $v_0 = 10^{-12}$ and $\zeta_n = \epsilon_{d,B,n}(1 - e^{-\max(\bar{r}_n, 0)})$. For exact moments, the implemented statistic satisfies

$$\widetilde{C}_n - \zeta_n = \mu_n - \kappa\sqrt{\max(v_n, v_0)} \le C_n^\circ.$$

Thus the robust lower-bound interpretation applies after removing the known stabilizer; it need not apply to $\widetilde{C}_n$ itself. If approximate moments obey $|\widehat{\mu}_n - \mu_n| \le d_\mu$ and $|\widehat{v}_n - v_n| \le d_v$, then

$$\left|\widehat{\mu}_n - \kappa\sqrt{\max(\widehat{v}_n, v_0)} - \mu_n + \kappa\sqrt{\max(v_n, v_0)}\right| \le d_\mu + \kappa\sqrt{d_v}.$$

This follows because $v \mapsto \max(v, v_0)$ is 1-Lipschitz and $|\sqrt{x} - \sqrt{y}| \le \sqrt{|x-y|}$ for nonnegative x, y. Without certified moment errors, a Monte Carlo implementation is not an exact numerical certificate.

For ordinary Shannon-entropy buckets, identical count multisets at the same prefix give the same exact smoother entropy moments, regardless of bucket names or execution paths. The tie-breaker can still distinguish such prefixes through their surprisal summaries. This consequence-level invariance does not imply temporal sensitivity, nor universal invariance of prefix-causal censoring encodings under reordering.

C.2 Spectral semantic risk

The repaired semantic kernel defines a positive-semidefinite, trace-one matrix $\rho_n = \mathbf{K}_n/n$. Its entropy represents graded semantic agreement, with ordinary cluster entropy as a special case.

Proposition C.2 (Hard-partition limit and entropy continuity) *Suppose* $n \ge 2$. *If* $\mathbf{K}_n$ *has entries one within semantic groups and zero between groups, with group sizes* $c_1, \dots, c_J$, *then*

$$H_n = -\frac{1}{\log n}\sum_{g=1}^{J}\frac{c_g}{n}\log\frac{c_g}{n}. \tag{22}$$

More generally, for two repaired trace-one matrices ρ_n, ρ_n', *let* $d = \frac{1}{2}\|\rho_n - \rho_n'\|_1$, *where* $\|\cdot\|_1$ *is the trace norm. Their normalized entropies satisfy*

$$|H_n - H_n'| \le \frac{d\log(n-1) + h_2(d)}{\log n}, \qquad h_2(d) = -d\log d - (1-d)\log(1-d). \tag{23}$$

Here $0\log 0 = 0$, *including at* $d = 0, 1$.

Proof of Proposition C.2. After a permutation, the hard-partition kernel is block diagonal with blocks $\mathbf{1}_{c_g}\mathbf{1}_{c_g}^\top$. Each block has one nonzero eigenvalue c_g. The nonzero eigenvalues of $\rho_n = \mathbf{K}_n/n$ are consequently c_g/n, which gives Eq. (22). This recovers entropy of empirical semantic-group frequencies, normalized by $\log n$, not by $\log J$. It does not equate this quantity to the smoothed count statistic.

For the second statement, apply the Fannes–Audenaert continuity inequality to the two n-dimensional density matrices (Audenaert, 2007) and divide by $\log n$. The inequality uses the actual trace distance $d \in [0, 1]$ and natural logarithms; changing the logarithm base rescales numerator and denominator equally. □

In particular, $0 \le H_n \le 1$, with rank-one and identity correlation kernels attaining the extremes. Permuting the sampled terminal responses conjugates the kernel by a permutation matrix and leaves its eigenvalues unchanged. These are structural facts, not evidence that inferred semantic groups correspond to correctness or execution outcomes.

The continuity result compares the *repaired* density matrices at the same prefix length. A bound on raw NLI probabilities must also control the repair map before it can be substituted into this result. Comparing two different prefix lengths additionally changes the normalization $\log n$; the displayed same-dimension inequality does not by itself handle that change. With eigenvalues below a numerical threshold omitted from the entropy sum, the computed entropy also has a truncation error. The ideal matrix statement should not be described as exact floating-point invariance without accounting for rounding and eigensolver error.

C.3 Rank fusion and its reference cohort

For a task with both views, write the clipped reference ranks as $u_{Ci}, u_{Hi} \in [a, 1-a]$, where $0 < a < 1/2$, and define $T_i = (\Phi^{-1}(u_{Ci}) + \Phi^{-1}(u_{Hi}))/\sqrt{2}$. The core score $\Phi(T_i)$ has the same ordering as T_i.

Proposition C.3 (Rank invariance and reference sensitivity) *With fixed reference membership and exact midranks, applying a strictly increasing transformation separately to each component, including its reference values, leaves the fused score unchanged. The fusion is nondecreasing in each rank coordinate. For two sets of clipped ranks satisfying $|u_{vi} - u'_{vi}| \le e_{vi}$ for $v \in \{C, H\}$,*

$$|T_i - T'_i| \le \frac{e_{Ci} + e_{Hi}}{\sqrt{2}\,\varphi(\Phi^{-1}(a))}, \tag{24}$$

where φ is the standard normal density.

Proof of Proposition C.3. Strictly increasing component transformations preserve all strict comparisons and ties within a fixed reference cohort, hence preserve midranks. Applying the same clipping, inverse normal CDF, sum, and outer CDF therefore leaves scores unchanged. Both CDF transformations are increasing, so the fused score is nondecreasing in each rank coordinate.

For $u \in [a, 1-a]$,

$$\frac{d}{du}\Phi^{-1}(u) = \frac{1}{\varphi(\Phi^{-1}(u))} \le \frac{1}{\varphi(\Phi^{-1}(a))}.$$

The mean-value theorem applied to each coordinate, followed by the triangle inequality, gives Eq. (24). The outer CDF additionally implies $|S_i^{\mathrm{core}} - S_i^{\mathrm{core}\prime}| \le |T_i - T'_i|/\sqrt{2\pi}$. □

The result applies to fixed semantic-coverage status and fixed reference membership. Different count and semantic cohort sizes are allowed. It does not assert score comparability between separately normalized cohorts. For a count-only fallback, the direct bound is instead $|R_C - R'_C| \le e_C$. The separate environment-specific outer rank mixture requires its own analysis and is not covered by the core formula.

A deterministic bound for changing component values. For one component, let $x_1, \ldots, x_M$ and $x'_1, \ldots, x'_M$ be aligned reference values with $|x_j - x'_j| \le d_j$. Let u_i, u'_i be their inclusive midrank

plotting positions. Then

$$|u_i - u_i'| \le \frac{1}{M} \sum_{j \neq i} \mathbf{1}\{|x_i - x_j| \le d_i + d_j\}. \tag{25}$$

Indeed, the comparison with reference task j cannot change unless the original gap is at most $d_i + d_j$; each comparison contributes at most $1/M$ to the rank change, and self-comparison is unchanged. Clipping cannot enlarge this difference. The bound applies directly to rounded values if their perturbation bounds are defined after rounding. It reveals why component perturbations can be amplified when many reference tasks have nearly tied values.

C.4 Conditional Gaussian rationale for equal weighting

Consider a Gaussian *working-model proxy* $G = (G_C, G_H)^\top$ for the transformed coordinates, with failure indicator $Y \in \{0, 1\}$ and

$$G \mid Y = y \sim \mathcal{N}(\boldsymbol{\nu}_y, \Sigma), \qquad \boldsymbol{\nu}_1 - \boldsymbol{\nu}_0 = \Delta \mathbf{1}, \quad \Delta > 0, \qquad \Sigma = s^2 \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}, \quad s > 0, \quad |r| < 1.$$

The log likelihood ratio is, up to a constant,

$$(\boldsymbol{\nu}_1 - \boldsymbol{\nu}_0)^\top \Sigma^{-1} G = \frac{\Delta}{s^2(1+r)} (G_C + G_H).$$

Thus the equal-weight sum has the likelihood-ratio ranking under this model, by the standard common-covariance Gaussian discrimination argument (Hastie et al., 2009). For independent draws from the failure and success classes, the difference of their summed scores is Gaussian with mean 2Δ and variance $4s^2(1+r)$, giving

$$\mathrm{AUROC}(G_C + G_H) = \Phi\left(\frac{\Delta}{s\sqrt{1+r}}\right), \qquad \mathrm{AUROC}(G_v) = \Phi\left(\frac{\Delta}{\sqrt{2}\, s}\right).$$

The symmetric model therefore predicts a benefit from complementary, equally informative views, with diminishing benefit as $r \to 1$.

This is not an exact distributional assertion about finite, clipped empirical ranks. Marginal Gaussianization does not imply Gaussian class-conditional distributions, equal signal strengths, or shared covariance; exact population Gaussianization of a mixed cohort must not be conflated with an exact Gaussian mixture model for those same coordinates. Unequal signals or covariances generally change the likelihood-ratio weights. The calculation supplies a conditional design rationale, not a claim that Stouffer empirically dominates arithmetic rank fusion.

C.5 Bayesian two-view stopping

Conditional recurrence model. We use Section 3.5's notation and condition on the task and fixed first-rollout anchor. For $j \ge 2$, the recurrence pair $(M_j^{\mathrm{out}}, M_j^{\mathrm{sem}})$ takes values in $\{0,1\}^2$. The Bayesian model assigns $\boldsymbol{\omega} = (\omega_{00}, \omega_{01}, \omega_{10}, \omega_{11})$ the prior $\mathrm{Dirichlet}(1/2, 1/2, 1/2, 1/2)$ and treats the pairs as iid conditional on $\boldsymbol{\omega}$. Dependence between views within a rollout is unrestricted. Write $\mathcal{D}_n$ for the $n-1$ post-anchor pairs and $m_v(n) = \sum_{j=2}^{n} M_j^v$ for $v \in \{\mathrm{out}, \mathrm{sem}\}$. The marginal recurrence probabilities are $\vartheta_{\mathrm{out}} = \omega_{10} + \omega_{11}$ and $\vartheta_{\mathrm{sem}} = \omega_{01} + \omega_{11}$. The event $\mathcal{R} = \{\vartheta_{\mathrm{out}} > 1/2, \vartheta_{\mathrm{sem}} > 1/2\}$ concerns both marginal rates, not necessarily $\omega_{11} > 1/2$. Bayesian probabilities below refer to this model; they do not assert that benchmark executions or the archived semantic-availability strata satisfy its assumptions.

Marginal posterior calculation. Let $J_{ab}(n)$ count pairs equal to (a, b) among the $n-1$ post-anchor observations. Dirichlet conjugacy gives

$$\boldsymbol{\omega} \mid \mathcal{D}_n \sim \mathrm{Dirichlet}\big((\tfrac{1}{2} + J_{ab}(n))_{a,b \in \{0,1\}}\big).$$

Aggregating the two matching and two nonmatching categories for either view yields a uniform beta prior and posterior

$$\vartheta_v \mid \mathcal{D}_n \sim \mathrm{Beta}(1 + m_v(n),\, n - m_v(n)).$$

For positive integers α, β and $z \in [0,1]$, the beta–binomial identity is

$$\Pr\{\mathrm{Beta}(\alpha,\beta) > z\} = \sum_{k=0}^{\alpha-1} \binom{\alpha+\beta-1}{k} z^k (1-z)^{\alpha+\beta-1-k}.$$

Substituting $\alpha = 1 + m_v(n)$, $\beta = n - m_v(n)$, and $z = 1/2$ recovers Eq. (8):

$$\pi_v(n) := \Pr(\vartheta_v > 1/2 \mid \mathcal{D}_n) = 2^{-n} \sum_{k=0}^{m_v(n)} \binom{n}{k}.$$

The beta marginals need not be independent; the joint Dirichlet model supplies coherence without cross-view independence.

Policy and savings-bound notation. For $\delta_{\mathrm{stop}} \in (0,1)$, the two-view rule selects the first prefix $5 \le n < B$ satisfying $\pi_{\mathrm{out}}(n) + \pi_{\mathrm{sem}}(n) \ge 2 - \delta_{\mathrm{stop}}$, and selects B if none qualifies, as in Eq. (10). Let $F_{N,p}(r) = \Pr\{\mathrm{Binomial}(N,p) \le r\}$. For $5 \le n < B$, define the sufficient marginal threshold

$$c_n = \min\left(\left\{s \in \{0,\dots,n-1\} : F_{n,1/2}(s) \ge 1 - \frac{\delta_{\mathrm{stop}}}{2}\right\} \cup \{n\}\right). \tag{26}$$

The value $c_n = n$ marks an unattainable threshold. Define

$$G_n = \left[1 - F_{n-1,\vartheta_{\mathrm{out}}}(c_n - 1) - F_{n-1,\vartheta_{\mathrm{sem}}}(c_n - 1)\right]_+, \qquad [z]_+ = \max(z, 0). \tag{27}$$

Proposition C.4 (Posterior reliability and finite-horizon savings) Posterior reliability. *Under the stated Bayesian model, every early-stopped history satisfies*

$$n^\star < B \quad \Longrightarrow \quad \Pr(\mathcal{R}^c \mid \mathcal{D}_{n^\star}) \le \delta_{\mathrm{stop}}. \tag{28}$$

Under the prior-predictive law, if $\Pr(n^\star < B) > 0$*, then also* $\Pr(\mathcal{R}^c \mid n^\star < B) \le \delta_{\mathrm{stop}}$.

Sampling efficiency. *For any fixed iid paired-indicator law* $\boldsymbol{\omega}$*, retain the same policy and let* G_n *be as in Eq.* (27)*. Then*

$$\Pr_{\boldsymbol{\omega}}(n^\star \le n) \ge G_n \qquad (5 \le n < B), \tag{29}$$

$$\mathbb{E}_{\boldsymbol{\omega}}[B - n^\star] \ge \sum_{n=5}^{B-1} G_n. \tag{30}$$

For $B \le 5$*, both the savings and the empty sum are zero. Dependence within a pair is permitted in both statements.*

Proof of Proposition C.4: reliability. Since $\mathcal{R}^c = \{\vartheta_{\mathrm{out}} \le 1/2\} \cup \{\vartheta_{\mathrm{sem}} \le 1/2\}$, the posterior union bound gives

$$\begin{aligned}\Pr(\mathcal{R}^c \mid \mathcal{D}_n) &\le \Pr(\vartheta_{\mathrm{out}} \le 1/2 \mid \mathcal{D}_n) + \Pr(\vartheta_{\mathrm{sem}} \le 1/2 \mid \mathcal{D}_n) \\ &= 2 - \pi_{\mathrm{out}}(n) - \pi_{\mathrm{sem}}(n).\end{aligned}$$

At an early stop, the rule makes this upper bound at most δ_{stop}, proving Eq. (28).

The stopped history records both the indicator pairs and their length. Since the rule is a deterministic function of that history, conditioning additionally on its selection does not change the posterior. Writing $E = \{n^\star < B\}$ and averaging under the joint Bayesian model, the tower property yields

$$\Pr(\mathcal{R}^c, E) = \mathbb{E}[\mathbf{1}_E \Pr(\mathcal{R}^c \mid \mathcal{D}_{n^\star})] \le \delta_{\mathrm{stop}} \Pr(E).$$

Division by $\Pr(E) > 0$ proves the prior-predictive assertion. No frequentist optional-stopping claim is used. □

Proof of Proposition C.4: efficiency. Consider a complete potential sequence of recurrence pairs through the cap, including uninspected pairs after an early stop. If both counts at prefix n are at least c_n, monotonicity of the binomial CDF gives

$$\pi_{\text{out}}(n) \geq 1 - \delta_{\text{stop}}/2, \qquad \pi_{\text{sem}}(n) \geq 1 - \delta_{\text{stop}}/2.$$

The stopping criterion holds at n, so the first qualifying prefix is no later than n. Hence

$$\{n^\star > n\} \subseteq \{m_{\text{out}}(n) < c_n\} \cup \{m_{\text{sem}}(n) < c_n\}.$$

Under any fixed iid paired-indicator law, each marginal count is $\text{Binomial}(n-1, \vartheta_v)$. The union bound and nonnegativity give $\Pr_{\boldsymbol{\omega}}(n^\star \leq n) \geq G_n$ without independence between views. Finally,

$$B - n^\star = \sum_{n=5}^{B-1} \mathbf{1}\{n^\star \leq n\},$$

because the selected prefix is either at least five or the cap $B \leq 5$. Taking expectations proves the savings bound, including the empty-sum case. □

Why the tolerance is not a frequentist false-stop rate. The setting $\delta_{\text{stop}} = 0.2$ bounds the posterior probability of $\mathcal{R}^c$ under the specified Bayesian model at a certified stop; it does *not* guarantee that at most 20% of nonpersistent tasks will stop early under every fixed recurrence law. Take $\delta_{\text{stop}} = 0.2$, $B = 16$, and the fixed iid pair law $\omega_{01} = \omega_{11} = 1/2$, $\omega_{00} = \omega_{10} = 0$. Then $\vartheta_{\text{out}} = 1/2$ and $\vartheta_{\text{sem}} = 1$, so $\mathcal{R}$ is false. Nevertheless, exact enumeration gives

$$\Pr_{\boldsymbol{\omega}}(n^\star < B) = \frac{6383}{16384} = 0.38958740234375 > 0.2. \tag{31}$$

Semantic matches are certain, so $\pi_{\text{sem}}(n) = 1 - 2^{-n}$. The minimum outcome-match counts permitting stopping at $n = 5, \ldots, 15$ are, respectively,

$$(4, 4, 5, 5, 6, 6, 7, 7, 8, 9, 9).$$

These thresholds use the actual summed-tail rule, not the sufficient equal-split thresholds c_n. Among the 2^{14} equally likely outcome sequences through prefix 15, the first-stop counts at prefixes $5, 6, 8, 10, 12, 15$ are $1024, 2048, 1280, 960, 760, 311$; the other eligible prefixes have zero first stops. Their sum is 6383. Thus the reliability part of Proposition C.4 is model-relative, not a uniform bound under fixed pair laws or a demonstration of model adequacy for the evaluated tasks.

Interpretation of the savings bound. The equal split of posterior tolerance in Eq. (26) is a *sufficient* condition; the implemented rule constrains only the sum of the two tails. The proof also lower-bounds stopping by n using a single-prefix event. Both relaxations can be conservative, and G_n may be zero for weak recurrence or short horizons. For a fixed policy, G_n is nondecreasing in each marginal recurrence rate, because the corresponding lower-tail binomial CDF is nonincreasing in that rate. This explains the direction of the bound, not an ordering of actual savings for all pair laws with different within-pair dependence. Unlike the posterior-reliability statement, the savings bound applies under any fixed iid pair law; the prior need not generate that law. Neither part guarantees correctness, score stability, or preservation of AUROC/AUPR. The bound counts inspected trajectories, not total compute or end-to-end execution time.

Recurrence signals, fallback, and selection. Outcome matching uses pre-UnknownCap2 buckets on the five non-WebShop datasets, so repeated raw no-answer events can count as recurrence while scoring separates them. WebShop uses its mapped structured outcomes. Semantic matching thresholds the anchor's unrepaired symmetric NLI affinities, not spectral entropy. The NLI-missing fallback has a one-view posterior interpretation under its Beta–Bernoulli model, rather than the two-view conclusion of Proposition C.4; its marginal threshold is $1 - \delta_{\text{stop}}$.

The replay fixes semantic availability from the complete sixteen-rollout pool. Prefix causality conditional on that mask does not imply that the mask is available online or that iid sampling survives conditioning on it. The propositions do not justify selection on an arbitrary future-dependent availability event. Prospective use must specify decision-time availability and assess the recurrence assumptions under that protocol. The empirical-settings appendix separately specifies the cached-replay, reference-cohort, cost-accounting, and evaluation protocols; the theoretical statements do not validate those assumptions from observed predictive performance.

C.6 Connecting score perturbations to ranking changes

Consider a fixed evaluation cell containing $N_+ > 0$ failure tasks and $N_- > 0$ success tasks. Suppose two latent core-score vectors satisfy $|T_i - T_i'| \le b_i$. Then, with half credit for ties,

$$|\,\mathrm{AUROC}(T') - \mathrm{AUROC}(T)| \le \frac{1}{N_+ N_-} \sum_{i:Y_i=1} \sum_{j:Y_j=0} \mathbf{1}\{|T_i - T_j| \le b_i + b_j\}. \tag{32}$$

Outside the indicated set, perturbations cannot change the sign of any failure–success score difference. Each remaining pair can change its AUROC contribution by at most one. Summing proves the inequality. Equal-cell macro effects are bounded by equally averaging the corresponding cell bounds.

Equation (24) can supply b_i when valid rank-error bounds are available; Eq. (25) can in turn translate component-error bounds into rank-error bounds. The count and spectral results describe possible ingredients in such an analysis, but they do not automatically bound the change between different selected prefix lengths or different availability regimes. In particular, anchor recurrence alone does not bound b_i. The right-hand side uses evaluation labels and, for a full-versus-stopped comparison, information about the full-budget scores. It is an offline deterministic diagnostic, not a label-free online stopping rule, a generalization guarantee, or a bound on AUPR.

D Additional Cost–Performance and Dataset–Backbone Comparisons

We examine the tradeoff between predictive performance and trajectory usage, and whether the macro-level patterns hold across individual dataset–backbone combinations. Figure 2 summarizes the six-dataset comparison; Figures 6 and 7 show its 18 constituent combinations. The comparison includes TURF-Fixed, TURF-Adaptive, and 15 baselines: PPL, LS, PE, SE, SNNE, SD, SGD, KLE, SentSAR, Degree, EigV, CoCoA-MaxProb, CoCoA-PPL, UProp, and SAUP.

Protocol and figure interpretation. The evaluation spans AgentBench-DB, HotpotQA, WebShop, StrategyQA, AgentBench-OS, and GAIA Level 1/2; Qwen 3.5 9B, Gemma 3 12B, and GPT-OSS 20B; and seeds 101, 202, and 303. Curves connect evaluated nominal budgets $B \in \{2, 3, 4, 6, 8, 12, 16\}$ where supported, with markers positioned at the source-reported mean trajectory count. TURF-Adaptive uses measured selected-prefix lengths; baseline coordinates retain historical nominal-count fallbacks when realized costs were absent. UProp and Degree have an audited common six-dataset endpoint at $B = 12$; the other displayed methods extend to $B = 16$. For the macro figure, we average the 18 dataset–backbone cells within each seed and then average across seeds. In the individual panels, we average only the three seeds. Shading denotes ± 1 sample standard deviation across these seed-level quantities, rather than a confidence interval. In particular, narrow macro ribbons do not establish uniformly small variation across tasks or datasets.

We use short baseline names throughout. SNNE, SGD, KLE, EigV, CoCoA-MaxProb, and CoCoA-PPL retain their final-response implementations; PPL, LS, PE, SE, SD, SentSAR, and Degree retain their step-aggregation implementations. UProp and Degree retain the existing four-candidate contract. These figures compare the registered implementations and do not constitute a paired comparison of the two adaptation strategies.

Macro performance and sampling efficiency. TURF-Fixed attains macro AUROC/AUPR of $0.7772/0.7360$ at $B = 16$. The strongest selected baseline endpoint is SNNE for AUROC (0.7290) and SGD for AUPR (0.7087), corresponding to descriptive gaps of 4.82 and 2.73 percentage points, respectively. TURF-Adaptive attains $0.7668/0.7282$ while using 10.966 trajectories on average at the same nominal cap. It therefore saves 31.46% of trajectories relative to TURF-Fixed, with decreases of 1.03 AUROC and 0.77 AUPR percentage points. Its endpoint remains above the strongest selected baseline macro endpoint on each metric, despite lower mean trajectory usage. These comparisons retain the source report's method-specific task coverage and should be interpreted descriptively.

Comparison across trajectory budgets. Table 8 separates the nominal cap from realized sampling cost. At small budgets ($B = 2, 3, 4$), the two TURF variants coincide because early stopping is

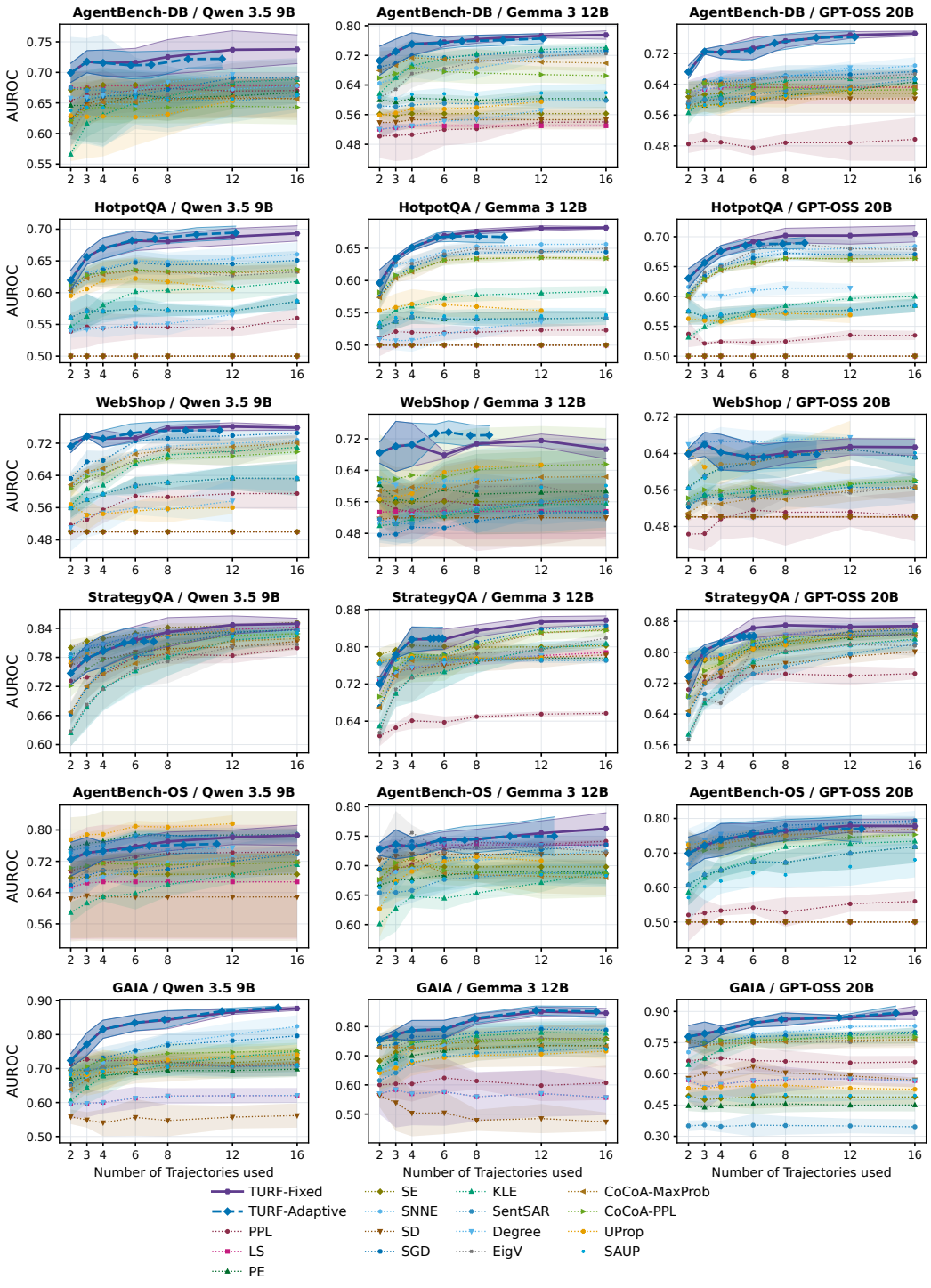

Figure 6: **AUROC by dataset and backbone.** Rows identify datasets and columns identify backbones. Curves show three-seed means; ribbons show ± 1 sample standard deviation across seeds. Markers use actual mean trajectory counts within each combination, retaining baseline nominal fallbacks where applicable. Each panel preserves its original report's X- and Y-axis limits; vertical scales can differ across both datasets and backbones. All panels use the same method colors and markers as Figure 2.

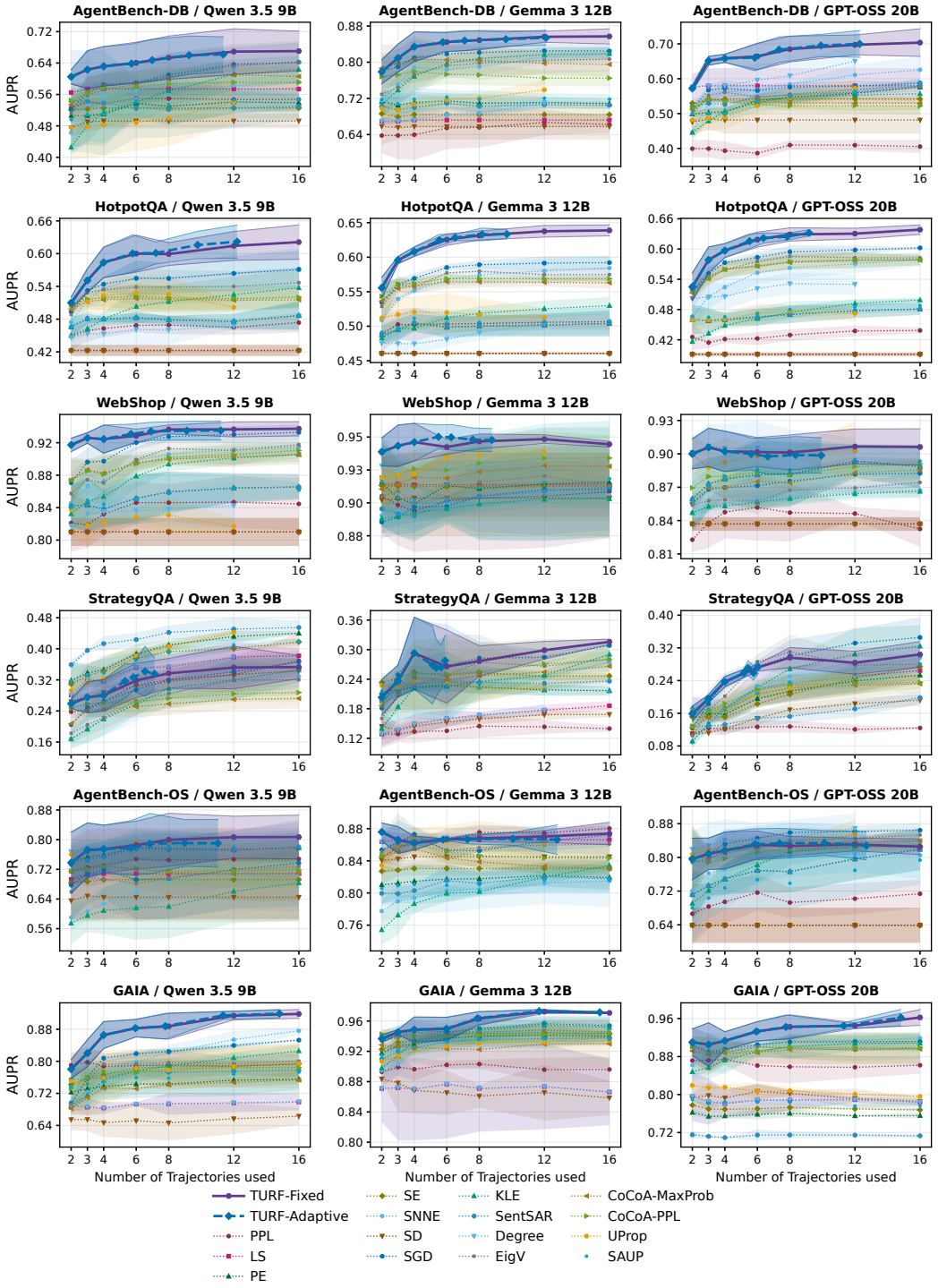

Figure 7: **AUPR by dataset and backbone.** The layout and aggregation match Figure 6. Each dataset–backbone panel preserves its original report's axis limits. Because AUPR depends on failure prevalence, its absolute level should primarily be compared between methods within the same panel, rather than across datasets. Thin TURF ribbon boundaries and distinct line styles improve readability without altering estimates.

Table 8: **Budget-wise macro comparison.** Fixed uses exactly B trajectories. Adaptive uses the same nominal cap but can stop earlier. Savings are $100(1 - \bar{n}_{\text{Adaptive}}/B)$. Scores are means over the same dataset–backbone–seed aggregation as the macro plot; the figure shows seed variability.

	TURF-Fixed		TURF-Adaptive			
B	AUROC	AUPR	Mean cost	Savings (%)	AUROC	AUPR
2	0.6986	0.6645	2.000	0.00	0.6986	0.6645
3	0.7290	0.6893	3.000	0.00	0.7290	0.6893
4	0.7404	0.7043	4.000	0.00	0.7404	0.7043
6	0.7507	0.7150	5.648	5.87	0.7514	0.7139
8	0.7641	0.7238	6.824	14.70	0.7585	0.7197
12	0.7751	0.7316	8.949	25.43	0.7645	0.7251
16	0.7772	0.7360	10.966	31.46	0.7668	0.7282

permitted only from prefix five. Even in this low-budget regime, their macro AUROC exceeds the strongest available baseline at each corresponding budget by 4.03–5.17 percentage points, and AUPR by 2.23–3.00 points. The strongest baseline can differ by metric and budget: CoCoA-PPL leads baseline AUROC at these three budgets, whereas SGD leads baseline AUPR at $B = 3, 4$.

At intermediate caps, adaptation begins to exchange a small amount of quality for lower sampling cost. At $B = 6$, it saves 5.87% of trajectories, with changes of $+0.08$ AUROC and -0.11 AUPR points relative to Fixed. At $B = 8$, savings increase to 14.70%, while the corresponding quality decreases are 0.57 and 0.41 points. At larger caps, the tradeoff becomes more pronounced: $B = 12$ saves 25.43% with losses of $1.06/0.66$ AUROC/AUPR points, compared with 31.46% savings and $1.03/0.77$ points lost at $B = 16$. Both variants remain above every displayed baseline macro mean on both metrics at each corresponding reported budget. These are descriptive comparisons, not significance claims; baseline availability and task coverage remain those of the source evaluation.

Increasing the budget also reveals diminishing aggregate returns. For Fixed, moving from $B = 2$ to $B = 4$ improves AUROC/AUPR by $4.18/3.98$ points; moving from $B = 12$ to $B = 16$ improves them by only $0.20/0.44$ points. For Adaptive, raising the cap from 12 to 16 adds 2.018 trajectories on average for gains of $0.23/0.32$ points. Thus the largest cap yields the highest macro quality for each variant, but an intermediate cap can offer a more economical operating point. This pattern is not a universal monotonicity guarantee for individual datasets or backbones.

Nominal-budget versus realized-cost comparisons. The table compares variants under the same maximum allowance, not equal realized expenditure. The curves instead place each marker at its realized cost. For example, Adaptive at cap $B = 12$ achieves $0.7645/0.7251$ using 8.949 trajectories, versus Fixed at $B = 8$ with $0.7641/0.7238$ using eight trajectories. Their similar quality does not establish superior allocation by Adaptive, which uses more samples in this particular comparison. Connecting evaluated markers does not create additional observations or establish matched-cost superiority; that claim requires dedicated matched-cost controls.

Variation across datasets and backbones. The individual panels show that strong macro performance coexists with meaningful heterogeneity. Comparing each selected baseline at its native endpoint, TURF-Fixed exceeds the strongest baseline mean in 13 of 18 combinations for AUROC and 14 of 18 for AUPR. The corresponding counts for TURF-Adaptive are 12 and 11. These are descriptive counts of three-seed means, not independent statistical wins; native budgets and task coverage can differ. Thus the aggregate advantage should not be read as dominance in every environment and backbone.

The budget trajectories also differ across panels. In AgentBench-DB and HotpotQA, much of the improvement occurs at small budgets, with flatter curves thereafter. GAIA retains a visible benefit from larger budgets, and Adaptive generally follows Fixed closely in realized cost. WebShop illustrates why the macro trend should not be assumed locally: its Gemma 3 AUROC curve is non-monotone and has substantial seed variation. StrategyQA shows a clearer separation between the Fixed and Adaptive high-cap operating points. These patterns favor reporting the complete budget grid rather than selecting only the largest-budget endpoint or a single favorable operating point.

Dataset-level averages further clarify the adaptive tradeoff. At $B = 16$, StrategyQA uses only 6.359 trajectories on average, a 60.25% reduction, but AUROC/AUPR falls from $0.8583/0.3242$ to $0.8244/0.2956$. GAIA, in contrast, uses 15.014 trajectories, saving 6.16%, with similar quality: $0.8747/0.9508$ for Adaptive versus $0.8722/0.9504$ for Fixed. On WebShop, Adaptive uses 9.969 trajectories and changes AUROC/AUPR from $0.7020/0.9293$ to $0.7066/0.9271$. The same stopping rule therefore produces different amounts of sampling and different quality tradeoffs across domains. The observed differences alone do not identify why a task benefits from additional trajectories.

E Final-response versus step-aggregated uncertainty baselines

Question and matched evaluation. Can response-level uncertainty methods become more competitive when applied to intermediate agent decisions? We evaluate 13 external baseline families in two adaptations: *Final* scores terminal-response sets, whereas *StepAgg* scores four local candidate decisions at each method-eligible step, averages those step scores within a trajectory, and then averages the trajectory scores over the ordered budget prefix. Every required step must yield a finite score; missing or zero-opportunity trajectories are not silently omitted or imputed. We compare these 26 adaptations with TURF-Fixed and TURF-Adaptive on six datasets, three backbones, and three seeds (54 dataset–backbone–seed cells), at nominal caps $B \in \{2, 3, 4, 6, 8, 12\}$.

The five non-GAIA datasets use their candidate-enriched $Z = 12, N = 4$ trajectory pools; GAIA uses the first 12 trajectories of its finalized $Z = 16, N = 4$ pool. All methods share ordered trajectories and native task labels within a cell. Macro AUROC and AUPR weight the six datasets, three backbones, and three seeds equally. The label-blind, all-method $B = 12$ support retains 21,986 of 22,230 task–cell records (98.90%) and is held fixed at every cap. This is a *separate secondary panel*: its metrics must not be pooled with, or substituted for, the main paper's original seven-budget $Z = 16$ experiment.

For every local and terminal candidate, a predetermined extractor uses identifiable decision or answer text without requiring an executable action; otherwise it uses the full observable response. It applies the same rule without consulting correctness labels. Empty observable responses remain missing. Historical method-specific likelihood fields and token-span semantics are unchanged, so a full-text fallback can be paired with a decision-span likelihood. No trajectories or candidate responses were generated or substituted for this comparison.

Adaptive policy in this panel. TURF-Adaptive uses *Bayesian two-view stopping*; TURF-Fixed and all baseline scores remain unchanged. The first rollout anchors two prefix-observable recurrence indicators: a later rollout matches its pre-UnknownCap2 outcome (or frozen mapped WebShop outcome), and its raw symmetric NLI affinity to the anchor is at least 0.75. A four-category $\mathrm{Dirichlet}(1/2, 1/2, 1/2, 1/2)$ model supplies marginal posterior probabilities $\pi_{\mathrm{out}}(n)$ and $\pi_{\mathrm{sem}}(n)$ that their recurrence rates exceed $1/2$. Starting at prefix five, the policy stops before the cap when

$$\pi_{\mathrm{out}}(n) + \pi_{\mathrm{sem}}(n) \geq 1.8, \qquad n \in \{5, \ldots, B-1\};$$

otherwise it uses all B rollouts. When NLI is unavailable, its declared fallback requires only $\pi_{\mathrm{out}}(n) \geq 0.8$. Thus $B \leq 5$ uses the full cap. The posterior criterion concerns *recurrence*, not correctness or preservation of AUROC/AUPR. This panel characterizes the reported adaptive policy under the secondary matched-support protocol.

E.1 Does step aggregation improve the baselines?

Figure 8 reports paired StepAgg-minus-Final effects at $B = 12$ on identical matched task support. StepAgg is not a uniform improvement: PE gains 2.17 AUROC and 1.90 AUPR percentage points; SentSAR gains 4.17 and 3.97 points. SNNE, Degree, SD, and LS decline on both metrics. The remaining family-level intervals overlap zero on at least one metric. Descriptively, PE and SentSAR gain on both metrics at all six caps, whereas SD declines on both throughout. The unadjusted 95% hierarchical bootstrap intervals resample backbones and seeds *within six fixed datasets*; they neither account for multiple testing nor establish generalization to unseen datasets. PE's positive $B = 12$ AUROC interval only narrowly excludes zero.

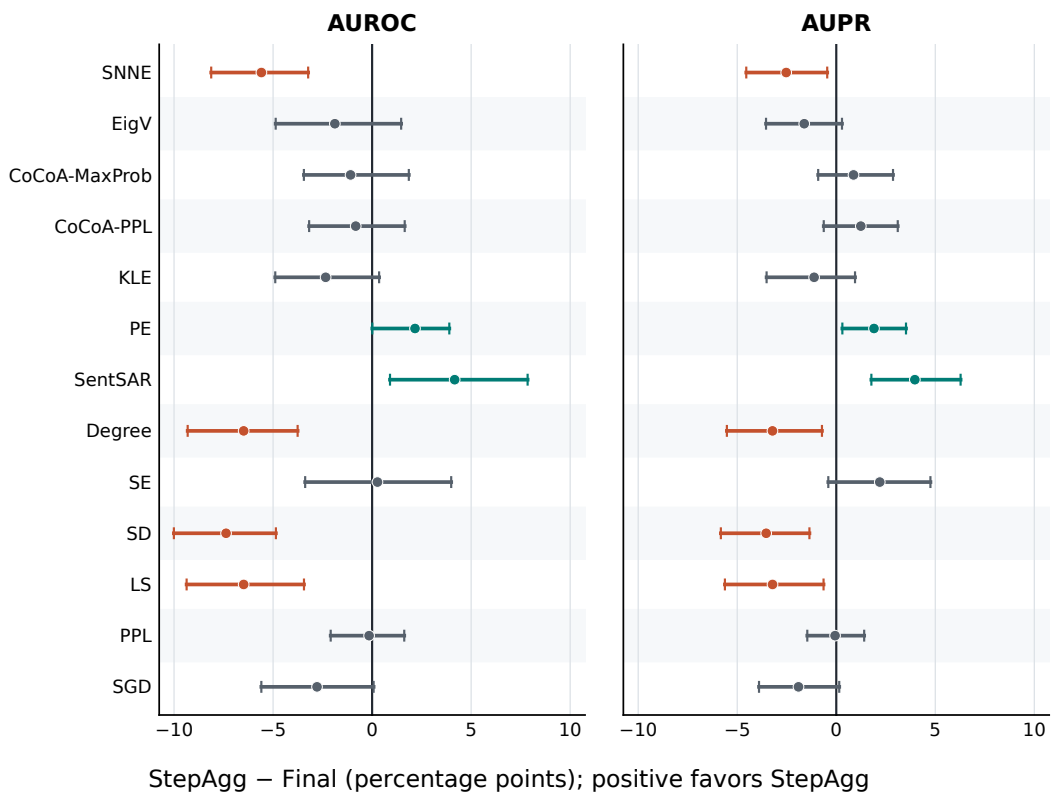

Figure 8: Paired StepAgg-minus-Final macro effects at $B = 12$, in percentage points. Bars are unadjusted 95% hierarchical bootstrap intervals over backbones and seeds within each fixed dataset. Teal and orange indicate intervals wholly above and below zero; gray intervals include zero. All 13 pairs use the same label-blind task support.

	AUROC		AUPR	
Family	Final	StepAgg	Final	StepAgg
SNNE	**0.7228**	0.6669	0.6836	0.6583
EigV	0.7051	0.6863	0.6839	0.6677
CoCoA-MaxProb	0.7031	0.6922	0.6667	0.6754
CoCoA-PPL	0.7054	**0.6971**	0.6699	**0.6823**
KLE	0.6933	0.6699	0.6681	0.6569
PE	0.6510	0.6727	0.6362	0.6552
SentSAR	0.6183	0.6600	0.6146	0.6542
Degree	0.7025	0.6376	0.6664	0.6342
SE	0.6772	0.6799	0.6382	0.6601
SD	0.6752	0.6014	0.6382	0.6028
LS	0.7025	0.6376	0.6664	0.6342
PPL	0.6192	0.6177	0.6133	0.6127
SGD	0.7212	0.6934	**0.7005**	0.6814

Table 9: Six-dataset macro AUROC/AUPR at nominal $B = 12$ on the common, label-blind matched support. Bold entries indicate the highest mean within each metric and adaptation.

Table 9 reports the corresponding absolute scores. Boldface identifies the highest mean within each adaptation and metric. These highlights summarize the panel; they do not define a deployable baseline-selection rule.

Cap B	Policy	AUROC	AUPR	Mean trajectories	Saved (%)
6	Fixed	0.7517	0.7167	6.000	0.00
	Adaptive	0.7520	0.7160	5.642	5.97
8	Fixed	0.7642	0.7253	8.000	0.00
	Adaptive	0.7583	0.7214	6.806	14.92
12	Fixed	0.7749	0.7327	12.000	0.00
	Adaptive	0.7642	0.7264	8.906	25.78

Table 10: Fixed and Bayesian two-view stopping on identical scorer and task support at caps permitting early stopping. At $B = 2, 3, 4$ the policies coincide. Savings count complete trajectories only.

E.2 TURF VERSUS BOTH ADAPTATIONS ACROSS BUDGETS

Budget-averaged matched-support comparison. Equally averaging all 54 cells and $B \in \{2, 3, 4, 6, 8, 12\}$ on this secondary panel gives AUROC/AUPR of $0.7439/0.7067$ for TURF-Fixed and $0.7412/0.7049$ for TURF-Adaptive. Their mean prefix lengths are 5.833 and 5.059, respectively. SGD-Final is the strongest of the 26 baseline adaptations on both budget-averaged metrics ($0.6881/0.6725$). Fixed and Adaptive therefore have descriptive advantages of $5.58/3.42$ and $5.31/3.24$ pp. Among StepAgg adaptations, CoCoA-PPL leads mean AUROC (0.6808) and SGD leads mean AUPR (0.6655); each metric therefore has its own StepAgg comparator. These six-budget means use the secondary matched-support pools and must not be substituted for the primary panel's five-budget means.

Same-cap tradeoffs and endpoint comparisons. The two TURF policies coincide for $B \in \{2, 3, 4\}$ because stopping cannot yet trigger. At $B = 6, 8, 12$, Adaptive uses 5.642, 6.806, and 8.906 complete trajectories on average—savings of 5.97%, 14.92%, and 25.78% relative to the corresponding cap (Table 10). Relative to Fixed at those same caps, its AUROC changes by $+0.03$, -0.59, and -1.07 percentage points, and its AUPR changes by -0.08, -0.39, and -0.63 points. These are common-*cap* comparisons, not controlled equal-trajectory-count experiments.

As one endpoint example, at $B = 12$ the best observed Final baseline reaches 0.7228 AUROC (SNNE) or 0.7005 AUPR (SGD), while CoCoA-PPL-StepAgg is the strongest StepAgg variant on both metrics at $0.6971/0.6823$. Relative to the highest baseline mean for each metric, TURF-Fixed has descriptive margins of 5.22 AUROC and 3.22 AUPR percentage points. These margins summarize observed performance and do not establish statistical superiority. Figures 9 and 10 show the full budget dependence.

Figures 9 and 10 show *all 13 baseline families* in separate Final and StepAgg quality–trajectory views. Each point corresponds to one of the six nominal caps. The horizontal coordinate is the actual equal-cell mean number of complete trajectories used: baseline and Fixed points lie at B, whereas Adaptive points can lie to the left. Both TURF policies exceed every Final and StepAgg adaptation in six-dataset macro AUROC and AUPR at each nominal cap. At $B \leq 4$, Fixed and Adaptive coincide because the stopping rule cannot yet trigger. The curves connect operating points across caps; they do not imply a matched-cost comparison between methods at arbitrary horizontal coordinates. In particular, the horizontal coordinate excludes acquisition of the four local candidates used by StepAgg at each eligible step, as well as semantic inference and other overheads. These plots compare failure ranking against complete-trajectory usage, not end-to-end compute efficiency.

E.3 VARIATION ACROSS DATASETS AND BACKBONES

Figures 11–14 resolve each global curve into 18 dataset–backbone panels, separately for Final and StepAgg and for AUROC and AUPR. Each curve is the mean over seeds 101, 202, and 303 on the same label-blind task support for all 28 methods and all six caps. Its translucent ribbon is ± 1 sample standard deviation across those three seeds, *not* a confidence interval. The horizontal coordinate is mean *actual* complete trajectories used in that dataset–backbone combination; Adaptive markers are placed at their realized mean trajectory counts. Vertical limits are shared by the three backbones within a dataset row, but can differ between dataset rows. Thus within-row backbone comparisons are visually comparable without compressing every dataset into one scale.

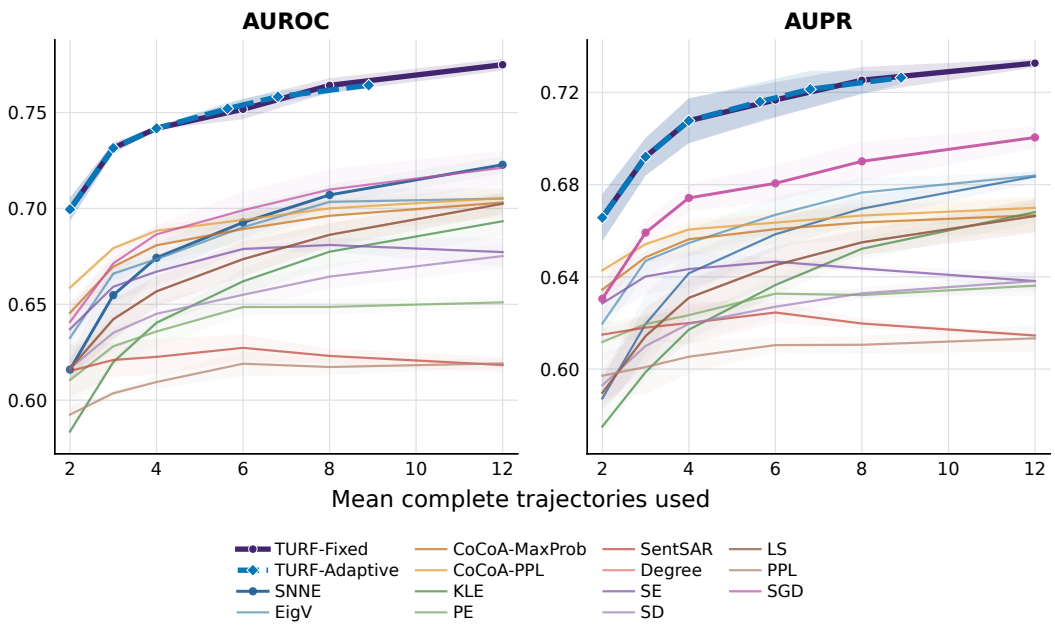

Figure 9: Final-response adaptations: AUROC and AUPR against mean complete trajectories used. All 13 named baseline families use their `-Final` variants; the legend omits that shared suffix for legibility. TURF-Adaptive is Bayesian two-view stopping. Translucent bands show ± 1 sample standard deviation across the three seed-level macro means, not confidence intervals or uncertainty in trajectory count. The axis measures scoring-prefix length, not total computation.

Dataset	Retained / total	Retained (%)	Failure (retained / excluded)
AgentBench-DB	2,642 / 2,700	97.85	47.2 / 62.1
GAIA	1,249 / 1,251	99.84	76.1 / 50.0
HotpotQA	8,976 / 9,000	99.73	42.3 / 33.3
AgentBench-OS	1,137 / 1,296	87.73	60.9 / 74.8
StrategyQA	6,182 / 6,183	99.98	7.7 / 0.0
WebShop	1,800 / 1,800	100.00	84.3 / —

Table 11: Label-blind all-method support at $B = 12$, held fixed at every cap. Failure percentages describe retained and excluded records and are not selection criteria; the excluded GAIA and StrategyQA rates are based on only two and one records, respectively.

The macro advantage is heterogeneous across both domains and budgets. For each dataset–backbone combination, metric, and cap, the comparison uses the maximum mean among the reported external adaptations. TURF-Fixed exceeds this maximum in 8 of 18 AUROC panels and 6 of 18 AUPR panels at $B = 2$. For each larger evaluated cap $B \in \{3, 4, 6, 8, 12\}$, it is higher in 11–12 of 18 panels on each metric. In particular, all three HotpotQA backbones favor Fixed on both metrics throughout the six caps; all three AgentBench-OS backbones favor an external adaptation; and WebShop's GPT-OSS backbone is an exception to Fixed's otherwise positive WebShop comparisons. StrategyQA remains mixed, especially for AUPR. Adaptive, at a lower mean trajectory cost once stopping can trigger, is higher than the same per-panel external envelope in 10 of 18 AUROC and 11 of 18 AUPR panels at each of $B = 6, 8, 12$. These counts compare observed per-panel baseline maxima at a common nominal cap. They describe ranking performance and do not constitute matched-compute or statistical-significance tests.

Support and scope. Table 11 shows the matched support behind both global and per-combination curves. In particular, only 87.73% of AgentBench-OS records are retained; excluded OS records have a 74.8% failure rate versus 60.9% among retained records. Although support selection is label-blind, these OS estimates are conditional on available inputs, not full-population performance.

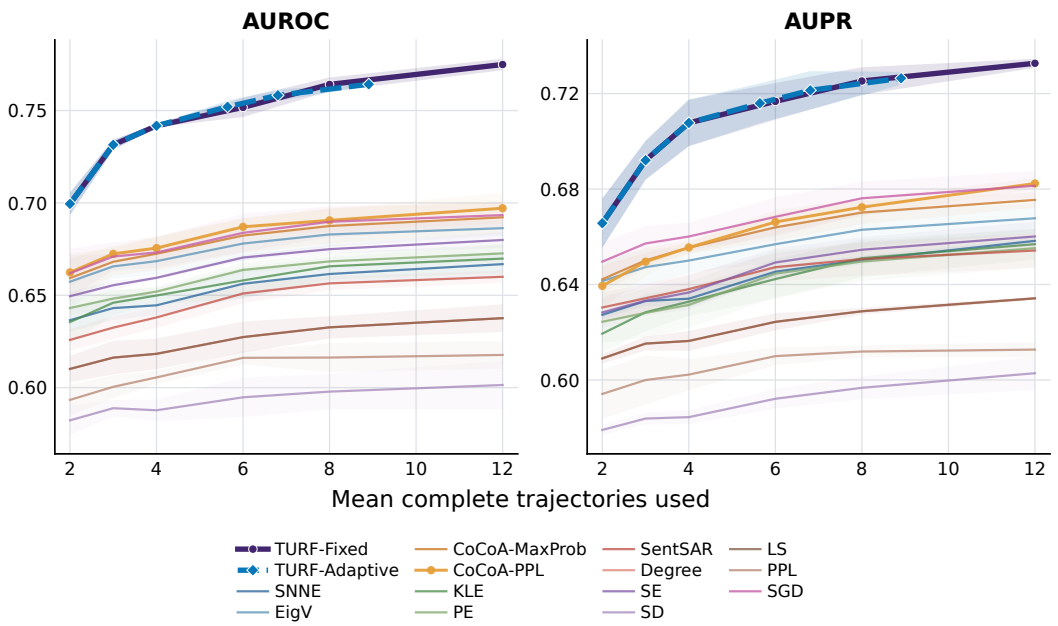

Figure 10: Step-aggregated adaptations on the same panel and complete-trajectory-count axis. All 13 named baseline families use their `-StepAgg` variants; the legend omits that shared suffix for legibility. Each StepAgg method additionally consumes four existing local candidates per eligible step, whose acquisition cost is *not* represented on the horizontal axis. Bands are descriptive ± 1 seed standard deviation, not confidence intervals.

The extraction audit recorded 602,390 local-candidate fallbacks to full observable responses; fallback text and historical likelihoods can refer to different token ranges. The cost axis counts complete trajectories, not local-candidate generation or all NLI inference. The ribbons summarize only three seeds, while the paired intervals treat the six datasets as fixed.

F Ablation Study

We organize the study around three questions: which information and scoring components improve failure prediction, whether task-dependent sampling improves allocation at matched cost, and how sensitive these findings are to reference-task construction and functional choices. Scoring ablations hold sampling fixed; allocation ablations hold the TURF scorer fixed. This separation distinguishes predictive contributions from changes in sampling cost and avoids attributing the latter to interventions that cannot change stopping decisions.

F.1 Experimental Protocol and Statistical Analysis

We evaluate AgentBench-DB, HotpotQA, WebShop, StrategyQA, AgentBench-OS, and GAIA with Qwen 3.5 9B, Gemma 3 12B, and GPT-OSS 20B, using seeds 101, 202, and 303. This gives 54 dataset–backbone–seed cells and 22,230 task–cell instances; repeated tasks across backbones and seeds are not independent observations. We compute failure-positive AUROC and AUPR within each cell and average cells equally. The evaluated nominal trajectory caps are $B \in \{2, 3, 4, 6, 8, 12, 16\}$.

The scoring ablations retain outcome construction, available frozen-NLI information, and evaluation support. UnknownCap2 is retained on the five non-WebShop datasets, and SeverityAware-v1 is retained on WebShop. The *Count posterior risk only* arm removes semantic risk from the core score; the *NLI Spectral Semantic Risk Only* arm removes the count view where semantic information is available but retains count fallback otherwise. *Mean Rank Fusion* replaces Gaussianized-rank Stouffer fusion with an equal-weight arithmetic mean of the two percentile scores. Thus "only" describes core scoring, not removal of fallback or the WebShop wrapper.

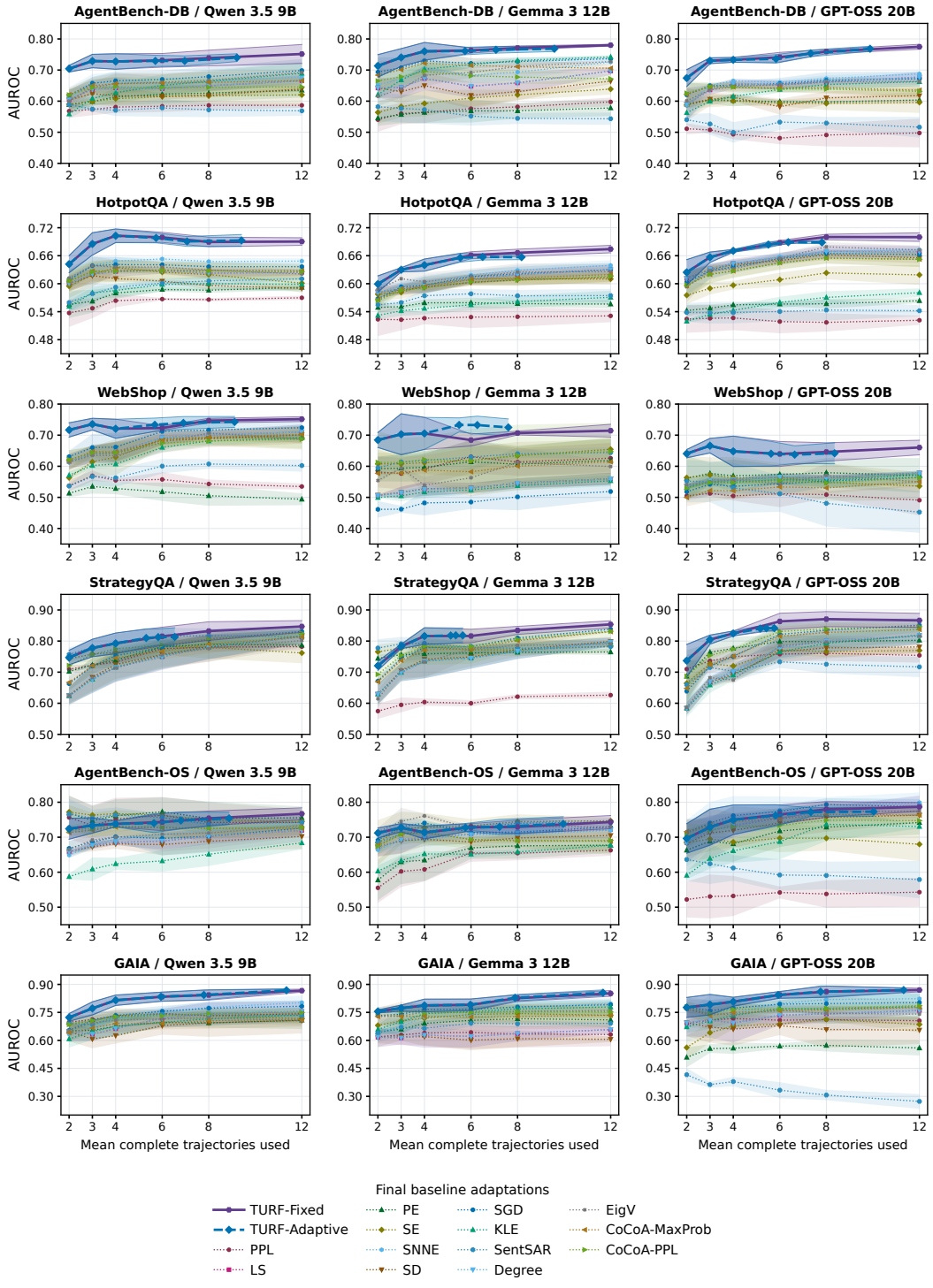

Figure 11: AUROC by dataset (rows) and backbone (columns), with all 13 Final adaptations and both TURF policies. Legend family names denote the corresponding `-Final` variant. Each panel shows six nominal caps at actual mean trajectory usage; faint ribbons are ± 1 standard deviation over three seeds, not confidence intervals. Vertical scales are shared within each dataset row. Trajectory counts measure scoring-prefix lengths, not total compute.

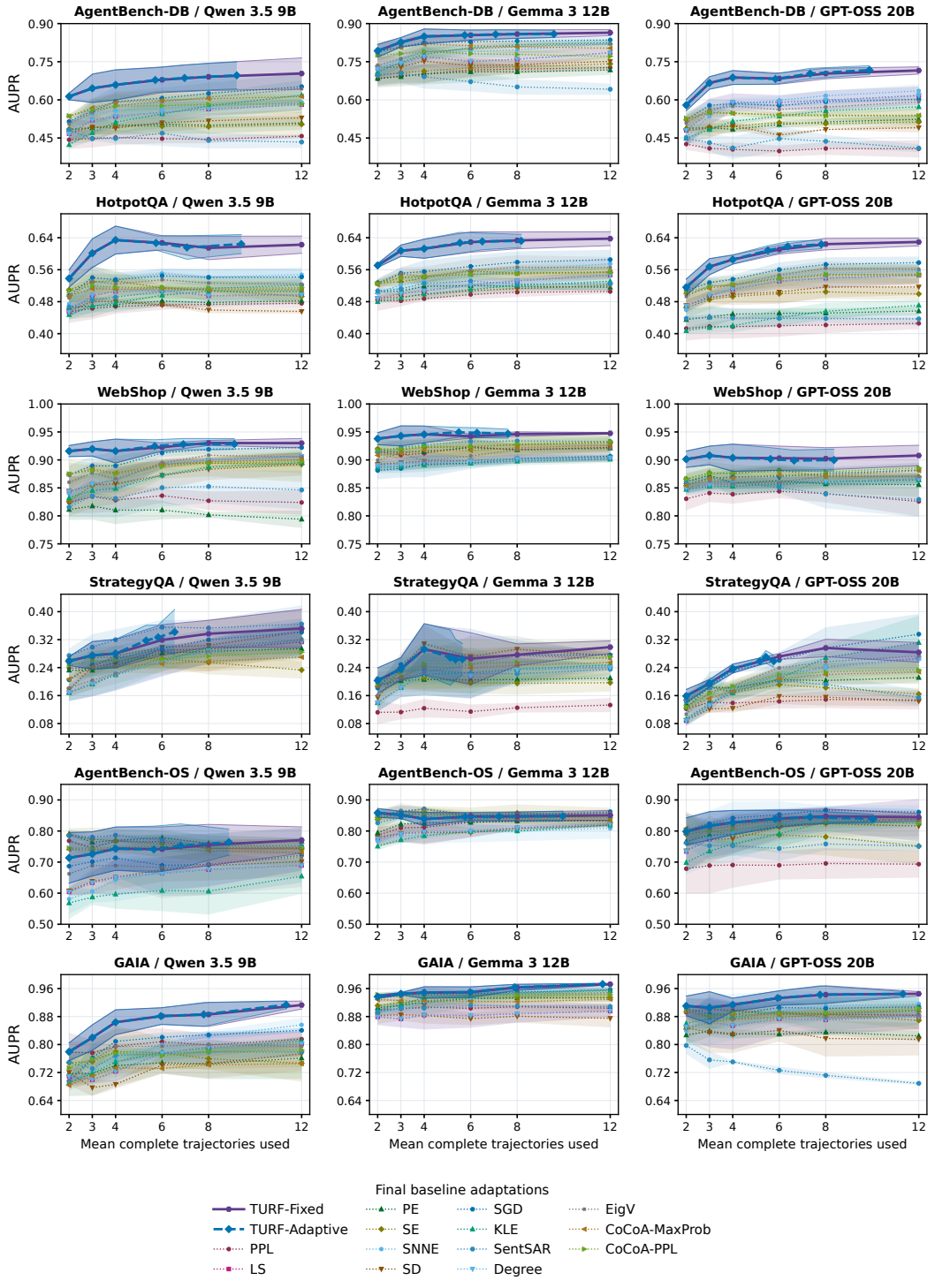

Figure 12: AUPR for the same 18 dataset–backbone combinations and Final adaptations as Figure 11. The vertical scale varies between dataset rows, reflecting their different failure prevalence and score distributions; ribbons are descriptive three-seed standard deviations.

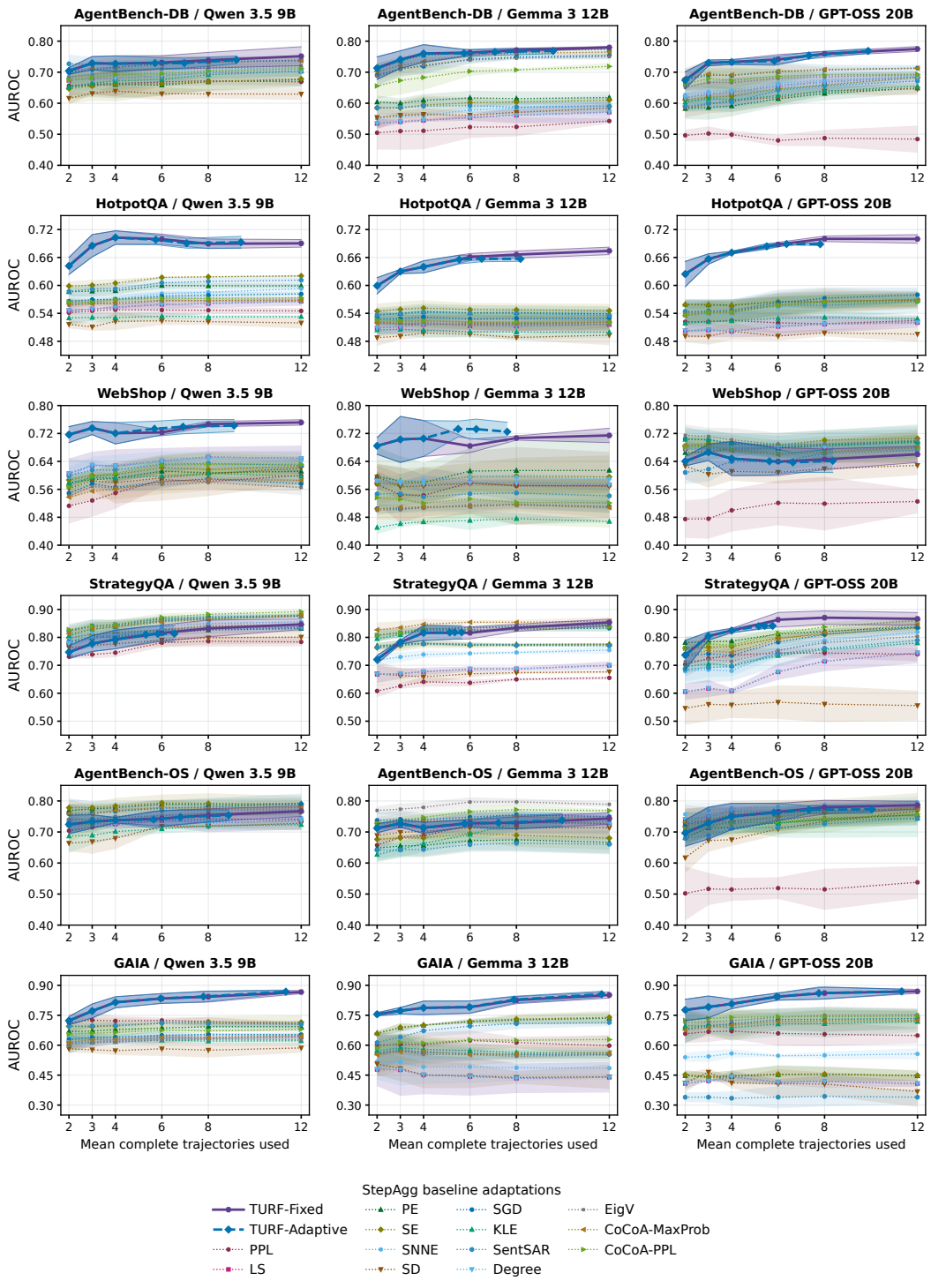

Figure 13: AUROC by dataset and backbone with all 13 StepAgg adaptations and both TURF policies. Legend family names denote the corresponding `-StepAgg` variant. Adaptive coordinates show realized mean trajectory usage; ribbons are descriptive three-seed standard deviations. The axis excludes acquisition of four local candidates per eligible step and other inference overheads.

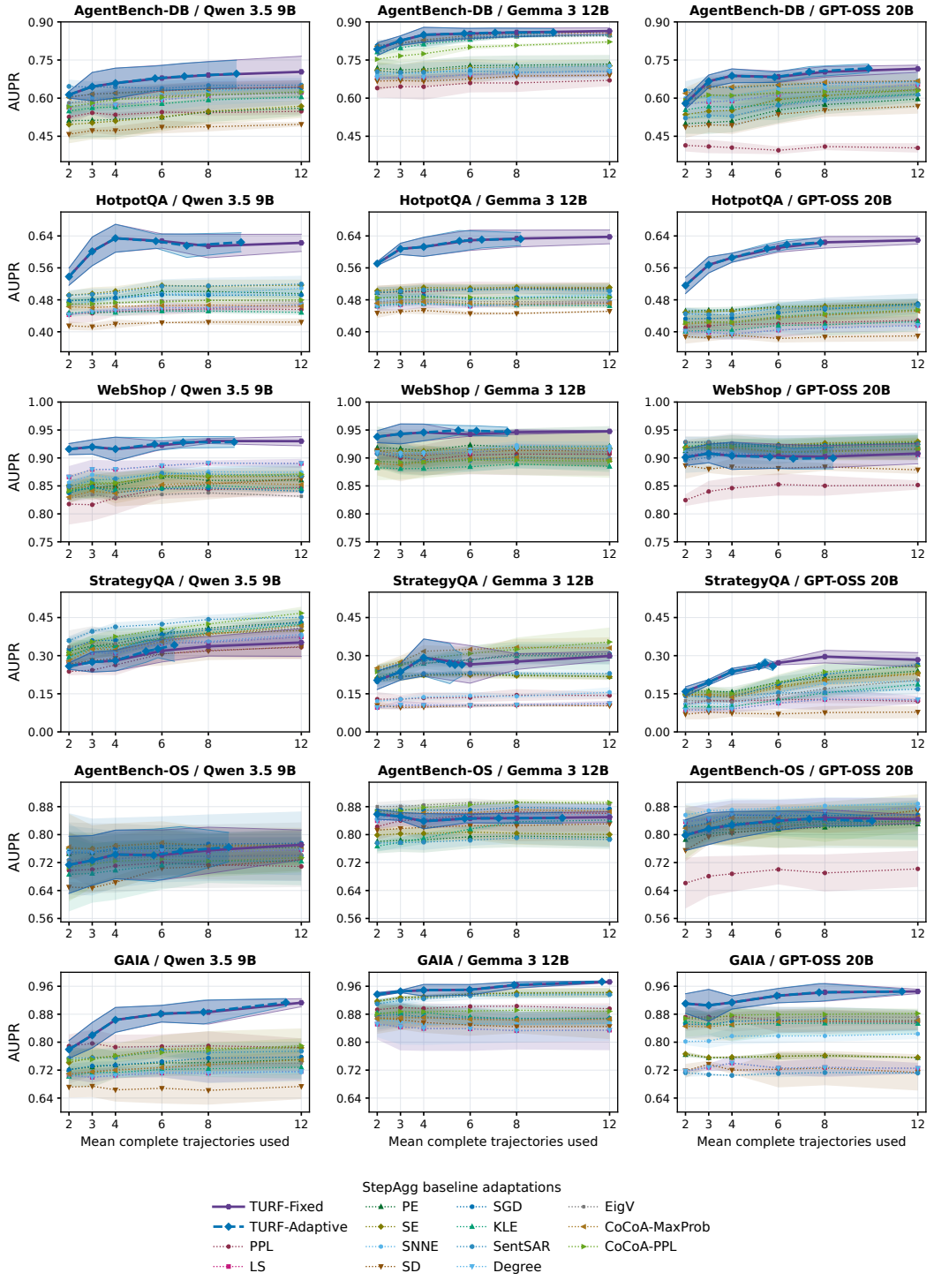

Figure 14: AUPR for the same 18 dataset–backbone combinations and StepAgg adaptations as Figure 13. StepAgg's four local candidates per eligible step are not included in the horizontal complete-trajectory count. Ribbons are descriptive three-seed standard deviations, not confidence intervals.

Our principal forests (Figures 15–17) use *Fixed sampling* for the first three rows: Full minus the semantic arm, Full minus the count arm, and Stouffer minus mean-rank fusion. Only the fourth row changes sampling, comparing Full TURF with Bayesian two-view stopping to Fixed Full TURF at the same nominal cap. Consequently, zero savings in the first three rows are imposed by the experiment, not evidence that scoring information could never influence a different stopping rule.

The Bayesian policy uses semantic affinity cutoff 0.75, recurrence threshold 0.5, joint Dirichlet concentration 0.5 per category, posterior tolerance 0.2, and minimum prefix five. It monitors outcome recurrence and semantic agreement with the first rollout, stopping at the first qualifying prefix or at B. Its posterior criterion concerns recurrence, not a guarantee on AUROC or AUPR. Separate scoring ablations at its selected prefixes retain exactly the same stopping lengths across all scoring arms.

Task-level inference for component effects. The forests show paired 95% basic-bootstrap intervals from 50,000 task resamples. Within each dataset, canonical tasks are resampled jointly across its nine backbone–seed cells; metrics are recomputed within cells before equal-cell aggregation. Primary rank maps and stopping decisions are held fixed. These are pointwise intervals, not simultaneous coverage statements. The original Fixed-$B = 16$ analysis additionally applies Holm correction to eight scoring and execution-information tests. That correction does not extend to the new Bayesian contrasts or the additional budgets. Supporting analyses rebuild rank maps in 5,000 resamples and resample the six dataset effects in 50,000 draws; the latter targets sensitivity to domain composition, not the same uncertainty as task resampling.

Dataset-level diagnostics for matched-cost policies. The four-baseline policy comparison uses a different uncertainty analysis: we compute Bayesian-minus-control effects within each cell, average its nine cells within each dataset, and form percentile 95% intervals from 50,000 resamples of the six dataset effects. Randomized-control metrics are first averaged over schedules within a cell. Task scores and the reported policy remain fixed. These pointwise intervals omit task-sampling uncertainty and are not multiplicity-adjusted. They are descriptive diagnostics, not substitutes for the component-level task bootstrap.

F.2 Scoring-Component Ablations

Count risk provides the strongest evidence. At Fixed $B = 16$, Full TURF reaches $0.7772/0.7360$ AUROC/AUPR, compared with $0.7547/0.7237$ for the semantic arm (Table 12). The paired count-view contributions are $+2.25$ AUROC and $+1.23$ AUPR percentage points (pp), with task-bootstrap intervals $[1.55, 2.92]$ and $[0.71, 1.68]$ pp. Both survive the original Holm correction; the corresponding dataset-resampling intervals are also positive. This supports the count-based view, rather than establishing that every detail of its posterior formulation is indispensable.

Semantic risk is complementary, especially for AUPR. Relative to count-only scoring, Full TURF improves Fixed-$B = 16$ AUROC/AUPR by $0.16/0.53$ pp. The AUROC interval includes zero. The AUPR pointwise interval is positive, but its test does not survive the eight-test Holm correction. Under Bayesian sampling, Full improves mean AUPR over count-only scoring from $B = 4$ onward, whereas AUROC differences are small (Figure 18). At nominal $B = 16$, the AUPR gain is 0.36 pp, with a task interval crossing zero. We therefore interpret the semantic view as a modest complementary signal rather than a uniformly significant improvement across settings.

Stouffer versus arithmetic rank fusion. At Fixed $B = 16$, Stouffer exceeds Mean Rank Fusion by only $0.023/0.079$ pp in AUROC/AUPR, with both intervals crossing zero. At Bayesian-selected prefixes, the differences are $-0.026/+0.054$ pp, again with intervals crossing zero. These results support mean-rank fusion as a competitive simple control. They do not establish either equivalence or an empirical necessity for the Gaussian transformation; Stouffer should be understood as the theoretically motivated default, not a demonstrated source of large gains.

Cross-budget scoring summaries. For the descriptive Fixed-policy summary over $\mathcal{B}_{\text{shared}} = \{2, 3, 4, 6, 8, 12\}$, equally averaging the 54 cells and these six budgets, Full TURF exceeds the semantic arm by $2.96/1.88$ pp in AUROC/AUPR. That arm retains count fallback, so this is not a universally count-free control. Adding semantic risk to the count-only arm gives smaller mean gains of $0.02/0.17$ pp. Its AUPR contribution is negative at $B = 2, 3$ and positive from $B = 4$

Table 12: Scoring ablations at nominal $B = 16$, averaged equally over 54 cells. Each policy block uses identical sampling lengths for all four arms. "Semantic only" retains count fallback; WebShop retains its wrapper. These are descriptive means, not significance rankings.

Sampling	Core score	AUROC	AUPR	Mean trajectories
Fixed	Full TURF (Stouffer)	.7772	.7360	16.00
	Count posterior only	.7756	.7307	16.00
	NLI spectral semantic risk only	.7547	.7237	16.00
	Mean Rank Fusion	.7769	.7352	16.00
Bayesian	Full TURF (Stouffer)	.7668	.7282	10.97
	Count posterior only	.7671	.7247	10.97
	NLI spectral semantic risk only	.7387	.7141	10.97
	Mean Rank Fusion	.7671	.7277	10.97

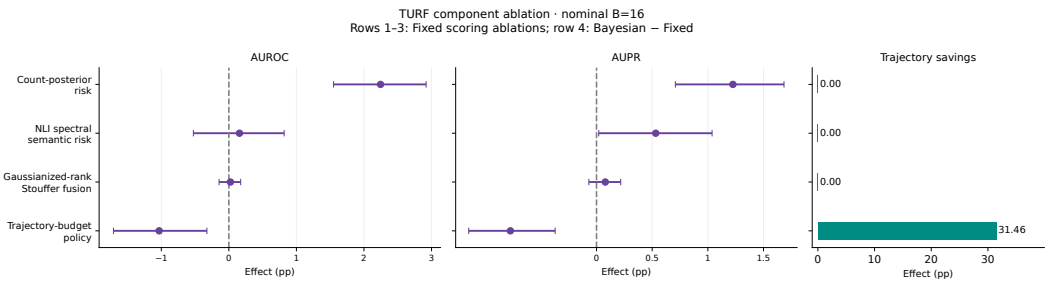

Figure 15: Component effects at $B = 16$. Rows 1–3 retain Fixed-budget scoring ablations; row 4 compares Bayesian two-view stopping with Fixed. Positive quality effects favor Full in rows 1–3 and Bayesian in row 4. Intervals are paired 95% task-bootstrap intervals. Savings are relative to the nominal cap and are zero by design in the scoring rows; they do not measure NLI runtime.

onward, reaching 0.41, 0.60, and 0.53 pp at $B = 8, 12, 16$. Stouffer minus arithmetic mean-rank fusion averages $-0.036/+0.029$ pp over the six-budget grid, with directions varying across budgets. These observations neither establish Stouffer's superiority nor demonstrate equivalence to mean-rank fusion.

All scoring arms use identical sampled prefixes within a policy and retain their native fallback and environment-specific adjustments. This six-budget summary differs from the main baseline comparison's $B \in \{4, 6, 8, 12, 16\}$ average and the component figure's $B \in \{8, 12, 16\}$ contrasts. The original eight-test Holm correction applies only to the fixed-$B = 16$ test family, not to these descriptive budget averages or the additional pointwise intervals.

F.3 Trajectory-Budget Policy Ablations

We now freeze outcome construction, the NLI model, fallback, percentile normalization, WebShop severity treatment, and the final TURF score, varying only trajectory allocation. We distinguish savings relative to the same nominal cap from predictive quality relative to alternatives at the same actual sampling cost.

F.3.1 Same-Cap Quality–Cost Tradeoffs

Bayesian two-view stopping reduces mean trajectory counts by 14.70%, 25.43%, and 31.46% at caps 8, 12, and 16, respectively, relative to Fixed at the *same cap* (Table 13). This is accompanied by lower mean predictive quality. At $B = 16$, the AUROC/AUPR losses are $1.03/0.77$ pp; the AUROC loss exceeds the one-point visual reference even at the point-estimate level. That reference is not a validated noninferiority margin. These are measured quality–cost tradeoffs, not noninferiority guarantees or savings against the cheapest qualifying fixed budget.

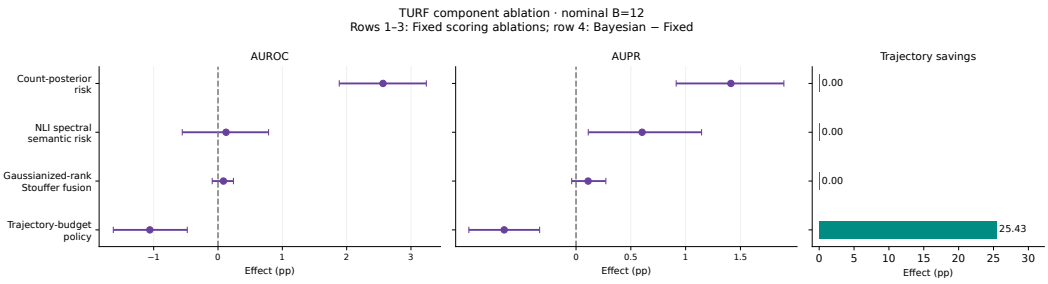

Figure 16: Component effects at $B = 12$, using the same contrast definitions as Figure 15 and budget-specific bootstrap intervals. Axis ranges accommodate all intervals; compare numerical effects rather than apparent horizontal distances across figures.

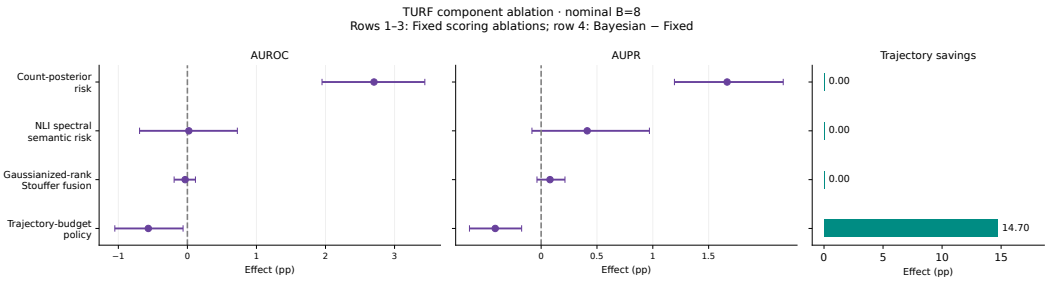

Figure 17: Component effects at $B = 8$. Fixed scoring ablations occupy rows 1–3; Bayesian versus Fixed occupies row 4. The intervals are pointwise and are not adjusted across budgets. Zero savings in scoring rows reflect fixed sampling.

F.3.2 Matched-Cost Comparisons

We use the four-baseline comparison campaign consistently for all matched-cost results in this subsection. Its randomized matched-Fixed control is distinct from the exact-total-cost EqualCost control in a separate sensitivity campaign; their numerical results are not pooled or treated as interchangeable.

Controls. *Fixed (matched)* randomly mixes adjacent integer sampling lengths to approach the Bayesian policy's mean cost within each cell; it need not assign one common length to every task in a schedule. *Shuffled lengths* permutes Bayesian stopping lengths among tasks within the same cell, preserving their histogram and total cost exactly. *Outcome stability* stops after a unique leading outcome persists for a given number of prefixes; ties reset the run. Its tested patience values are $2, 3, 4, 6, 8, 10, 12$. *Outcome dominance* stops when the leading outcome reaches a share in $\{0.55, 0.60, 0.70, 0.80, 0.90, 1.0\}$, with minimum prefix 2, 4, or 8. Native stability and dominance policies stop by 16 trajectories.

For matched-cost evaluation, Stability and Dominance use randomized mixtures of bracketing native settings selected using sampling costs alone, not correctness labels. They are thus mixtures of policies rather than a single deterministic threshold. Fixed, Stability, and Dominance use 100 prespecified schedules in nonconstant cells. The largest absolute per-cell cost mismatches are 0.0114, 0.0131, and 0.0392 trajectories, respectively. Matching occurs separately within every cell, rather than only in the macro average.

Benefits at intermediate budgets, not uniform dominance. At caps 6, 8, and 12, Bayesian two-view stopping has higher mean AUROC and AUPR than all four matched-cost controls (Figure 19). At $B = 8$, its improvements over matched Fixed, Stability, and Dominance are $0.542/0.395$, $0.817/0.429$, and $0.718/0.343$ pp, respectively. This is consistent with a modest allocation benefit at intermediate budgets. At $B = 16$, the AUROC differences versus matched Fixed and Stability are -0.217 and -0.224 pp, whereas AUPR differences are $+0.092$ and $+0.171$ pp. Differences

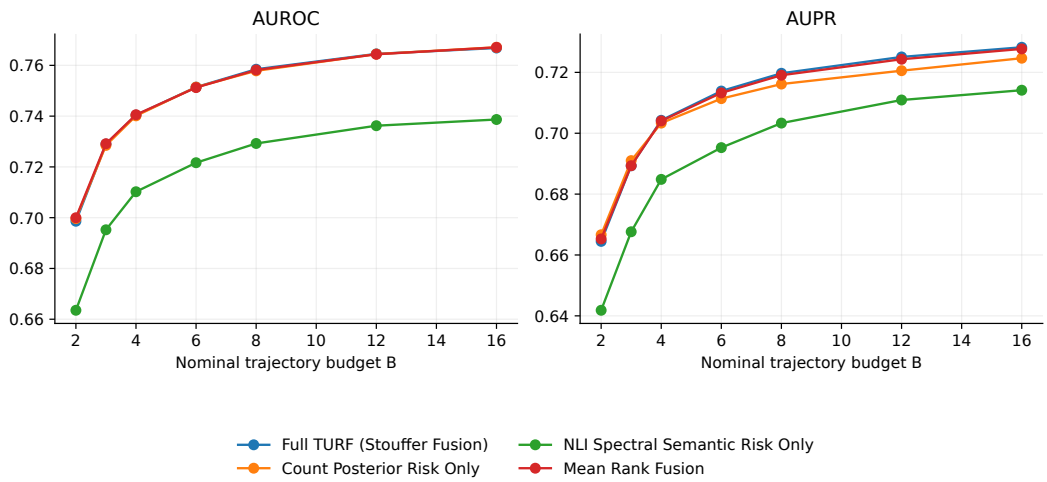

Figure 18: Scoring components at identical Bayesian-selected prefixes. The horizontal axis is the nominal cap, not mean sampling cost. All curves share stopping decisions, including both recurrence views. Curves are equal-cell means; no confidence bands are shown. Count-only, Full, and Mean Rank Fusion nearly overlap in AUROC, while their AUPR differences are more visible at moderate and larger caps.

Table 13: Full TURF with Bayesian stopping versus Fixed at the same cap. Quality differences are Bayesian minus Fixed, in percentage points. Sampling savings exclude semantic-inference overhead.

Cap B	Mean trajectories	Savings (%)	ΔAUROC (pp)	ΔAUPR (pp)
8	6.82	14.70	−0.566	−0.411
12	8.95	25.43	−1.060	−0.655
16	10.97	31.46	−1.032	−0.774

versus matched Dominance are only $+0.036/+0.012$ pp. The benefit is therefore budget- and metric-dependent, not uniform superiority over simple policies.

F.3.3 Task-Dependent Allocation and Dataset Heterogeneity

Reassigning stopping lengths degrades mean quality. Shuffled lengths removes the association between task identity and assigned budget without changing the amount or distribution of sampling. Bayesian Two-View exceeds this control in mean AUROC/AUPR by $0.403/0.284$, $0.342/0.429$, $0.368/0.533$, and $0.392/0.680$ pp at caps 6, 8, 12, and 16, respectively. This is the most direct evidence in the control set that *which tasks receive additional trajectories* matters.

Uncertainty remains substantial. Figure 20 shows the dataset-cluster diagnostic intervals defined in Section F.1. At $B = 8$, all four AUPR intervals lie above zero, whereas all four AUROC intervals include zero. At $B = 16$, Bayesian-minus-Shuffled intervals are $[0.075, 0.741]$ pp for AUROC and $[0.225, 1.206]$ pp for AUPR. These intervals are descriptive, conditional on the reported configuration, and based on only six dataset clusters; they are not multiplicity-adjusted.

Macro trends conceal domain-specific tradeoffs. At $B = 6$ and $B = 8$, AUPR exceeds matched Fixed on all six datasets. AUROC is less uniform: at $B = 8$, StrategyQA and WebShop improve by 1.907 and 1.379 pp, but AgentBench-DB and HotpotQA decrease by 0.547 and 0.214 pp. At $B = 16$, AUROC exceeds matched Fixed on only two datasets and AUPR on four. Against Shuffled lengths, five datasets improve in each metric at $B = 16$, with AgentBench-DB the AUROC exception and WebShop the AUPR exception. Figures 21 and 22 report the complete dataset breakdown rather than only favorable cells.

Table 14: Bayesian minus matched-Fixed quality in the four-baseline campaign. Unlike Table 13, the comparator here is matched to Bayesian sampling cost, not the nominal cap. Differences are in pp; values are descriptive equal-cell means.

Cap B	Bayesian AUROC	Bayesian AUPR	ΔAUROC	ΔAUPR
6	.7514	.7139	+0.409	+0.302
8	.7585	.7197	+0.542	+0.395
12	.7645	.7251	+0.143	+0.248
16	.7668	.7282	−0.217	+0.092

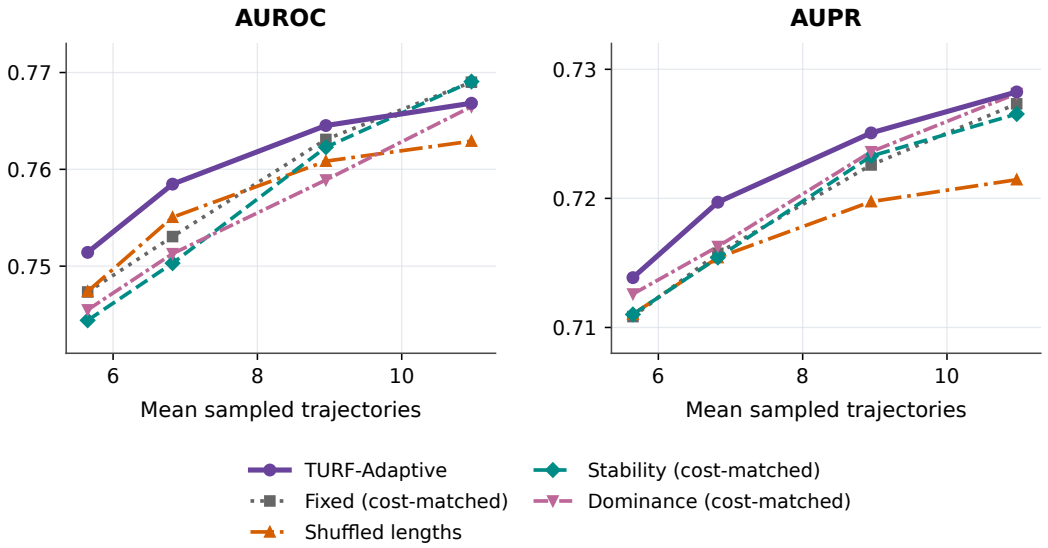

Figure 19: Predictive quality at approximately matched sampling cost for caps $6, 8, 12, 16$. Fixed, Stability, and Dominance are randomized mixtures matched within cells; Shuffled lengths preserves Bayesian cost exactly. Points are equal-cell means over 54 cells; connecting lines are visual guides, not interpolation guarantees. These curves do not include confidence bands.

Native operating points provide context, not matched-cost tests. Figure 23 additionally shows Bayesian and Shuffled at caps $\{2, 3, 4, 6, 8, 12, 16\}$, Fixed at integer budgets 2–16, Stability at its seven patience settings, and Dominance at its three minimum-prefix branches. These native settings generally have different costs; the matched-cost comparisons above are the direct allocation tests.

F.4 Sensitivity to Reference Tasks

Reference tasks supply the score distributions used to convert uncertainty scores into percentiles. Five-fold query-excluded maps are built from the other four folds, preventing each task from contributing to its own map. At $B = 16$, this changes Fixed AUROC/AUPR from $0.7772/0.7360$ to $0.7767/0.7356$, and Bayesian scores from $0.7668/0.7282$ to $0.7665/0.7284$. These small descriptive changes suggest that self-inclusion is not a major driver on the evaluated panel; they do not test domain shift.

Smaller reference subsets have a clearer effect (Figure 24). With 25 reference tasks, Fixed reaches $0.7703/0.7236$ and Bayesian reaches $0.7597/0.7172$. With 100, the respective scores recover to $0.7760/0.7324$ and $0.7656/0.7255$. The remaining AUPR losses are approximately 0.36 and 0.27 pp. These results average ten deterministic draws per fold and cell, with queries excluded and scores assembled over all five folds. Policies use separately seeded draws; these are not paired draw-by-draw effects, confidence intervals, or a universal reference-size requirement.

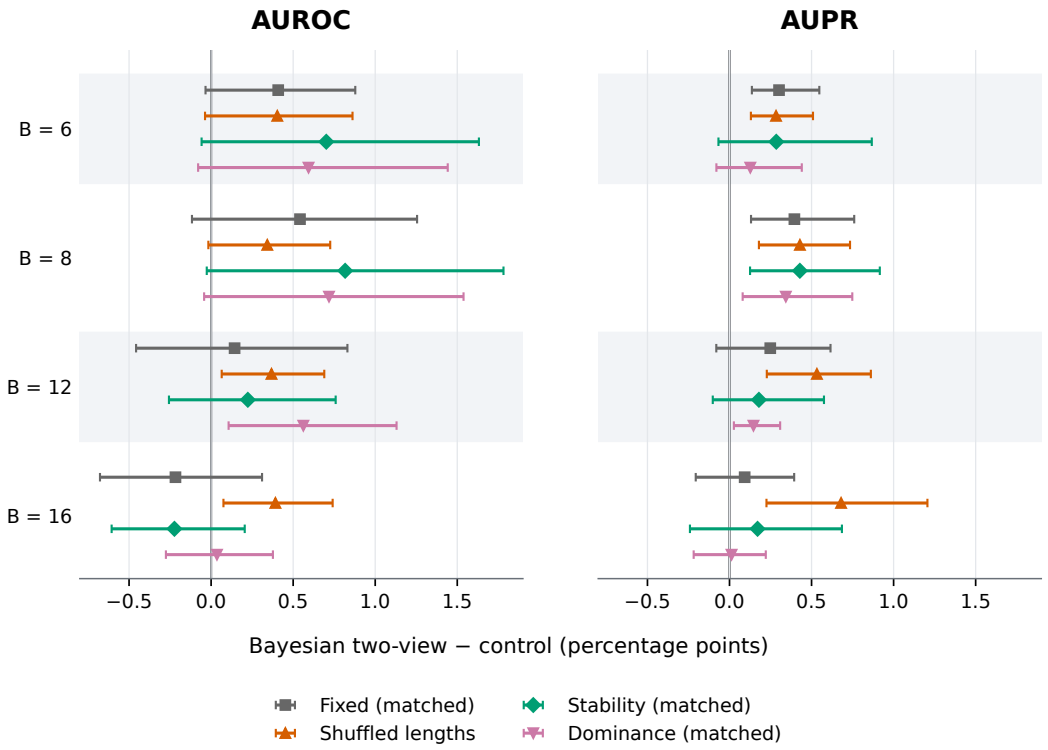

Figure 20: Bayesian-minus-control effects across budgets. Positive values favor Bayesian Two-View. Intervals are percentile 95% intervals from 50,000 resamples of six dataset effects, retaining the nine cells together within each dataset. These descriptive pointwise intervals condition on observed task scores and the reported policy, omit task-sampling uncertainty, and are not adjusted for multiple comparisons. They differ from the task-bootstrap component forests.

F.5 Functional Controls and Execution-Outcome Information

Functional choices have smaller aggregate effects. At Fixed $B = 16$, replacing the native count summary with posterior mean changes AUROC/AUPR by $-0.07/-0.09$ pp; plug-in entropy changes them by $+0.04/-0.06$ pp; mean disagreement gives $-0.07/+0.06$ pp. Removing step-surprisal produces the largest decline among these controls, $-0.35/-0.24$ pp. These are descriptive controls without paired significance tests in the primary family. They indicate limited sensitivity to several functional choices, not statistical equivalence or evidence that each native choice is essential.

Execution-outcome versus final-response information. A separate matched experiment compares execution-derived count buckets with normalized final-answer strings at every evaluated fixed budget. Both arms use identical tasks, prefixes, censoring rules, a flat count estimator, and frozen semantic scores; both omit the step-surprisal tie-breaker, WebShop hierarchy, and severity adjustment. Final-answer strings use Unicode NFKC normalization, whitespace normalization, and case folding. This isolates count-side information access rather than comparing a trajectory-aware system with a wholly response-only system. At $B = 16$, execution information improves macro AUROC/AUPR by $1.45/1.06$ pp, with task intervals $[0.67, 2.23]/[0.42, 1.70]$ pp; both survive the original Holm correction. Gains concentrate in AgentBench-DB ($4.92/4.45$ pp) and WebShop ($3.56/1.35$ pp), with small effects in AgentBench-OS and approximately zero effects in the remaining datasets. Dataset-resampling intervals cross zero, so this does not establish uniform cross-domain benefit. These matched arms are not interchangeable with the native scoring-ablation arms above.

The main paper's five-budget summary equally averages the 54 cells and $B \in \{4, 6, 8, 12, 16\}$. On this grid, execution information adds $1.59/1.24$ pp overall, $5.64/5.78$ pp on AgentBench-DB, and $3.99/1.48$ pp on WebShop. AgentBench-OS has mixed, small changes ($-0.08/+0.17$ pp); the other

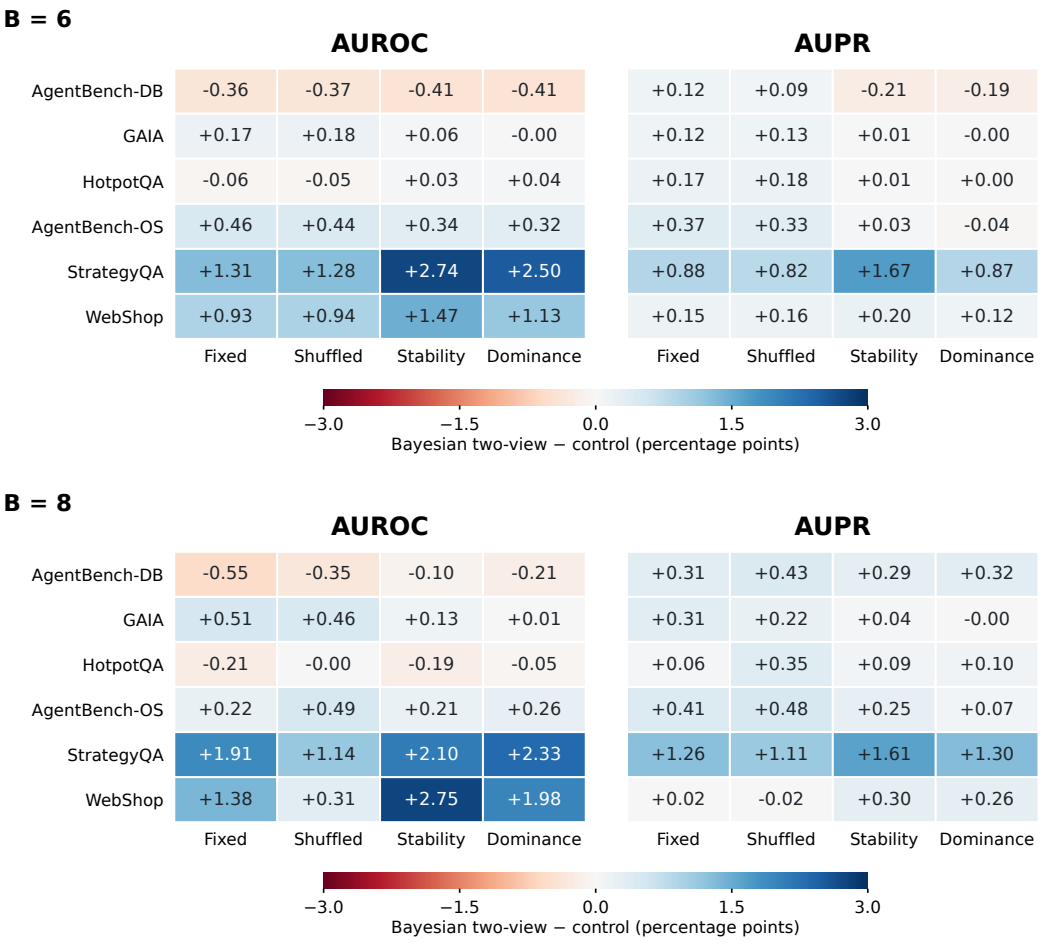

Figure 21: Dataset-level Bayesian-minus-matched-control effects at $B = 6$ (top) and $B = 8$ (bottom), in pp. Each entry averages three backbones and three seeds. Blue favors Bayesian and red favors the control. The color scale is shared across metrics and all four budget heatmaps. These are point estimates, not nine independent task-level replications per dataset.

datasets have essentially zero effects. The cross-budget figure shows descriptive three-seed standard deviations, not task-bootstrap intervals. The $B = 16$ intervals and Holm correction above do not transfer to these budget averages.

The concentration of gains is consistent with execution information being especially useful when terminal text does not fully represent the task outcome, as in database interactions and shopping actions. Conversely, small effects are compatible with substantial overlap between answer-derived buckets and the observable outcome representation. This is evidence for environment-dependent information value, not a uniform benefit across all six datasets or a direct test of internal trajectory-dynamics modeling.

Dataset-specific controls also require matched support. Censoring controls exclude WebShop and cover 45 cells; WebShop hierarchy/severity controls cover only nine. They should not be mixed into the 54-cell component forest.

F.6 Scope and Limitations

The present evidence supports count-based risk, complementary semantic information, and execution-outcome access where it changes the relevant representation. It provides weaker evidence for the necessity of the Stouffer transform or superiority of Bayesian allocation over simple matched-cost

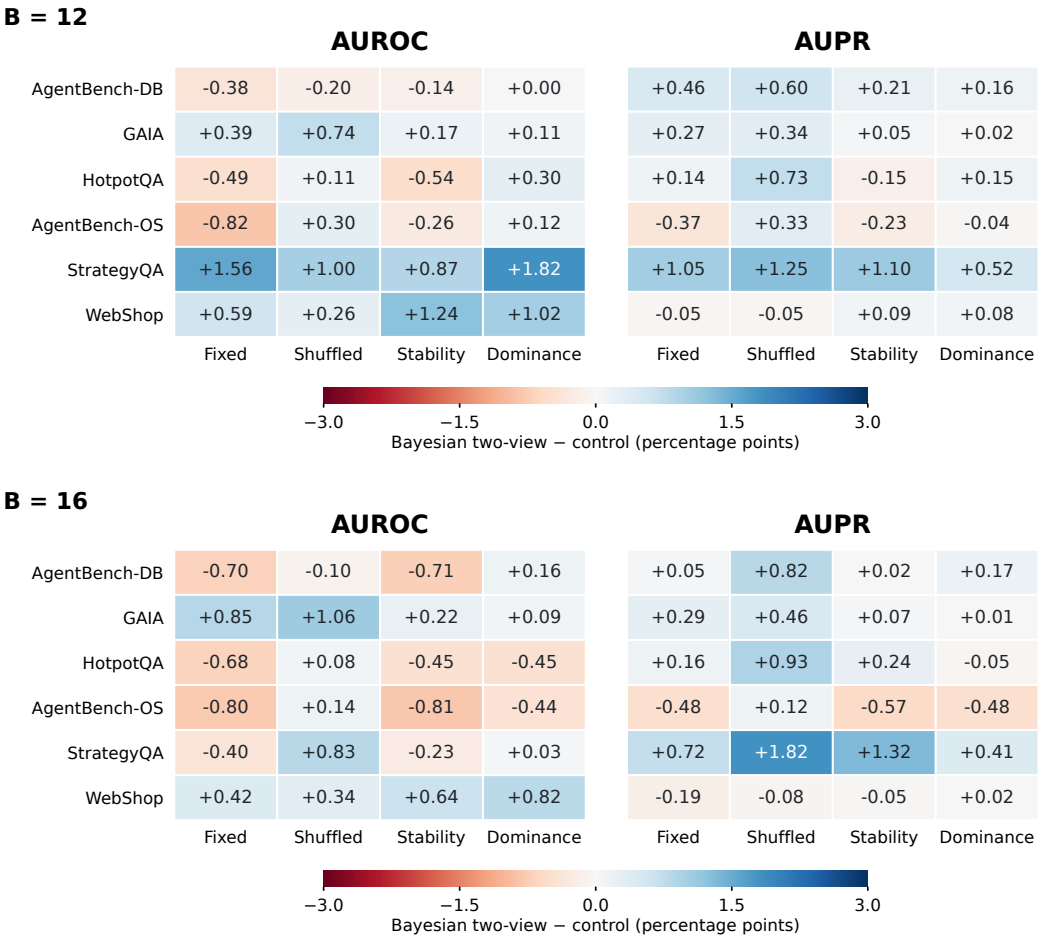

Figure 22: Dataset-level effects at $B = 12$ (top) and $B = 16$ (bottom), using the same definitions and color scale as Figure 21. Higher-budget macro comparisons conceal both positive and negative effects.

alternatives. Cached full-pool semantic availability is inherited by these replays, and trajectory savings should not be equated with measured end-to-end speedups. The policy comparisons do not separately remove the count and semantic recurrence views from the stopping mechanism, nor do they constitute a hyperparameter-sensitivity study. The Bayesian posterior statement concerns anchor recurrence under its model, not preservation of uncertainty rankings.

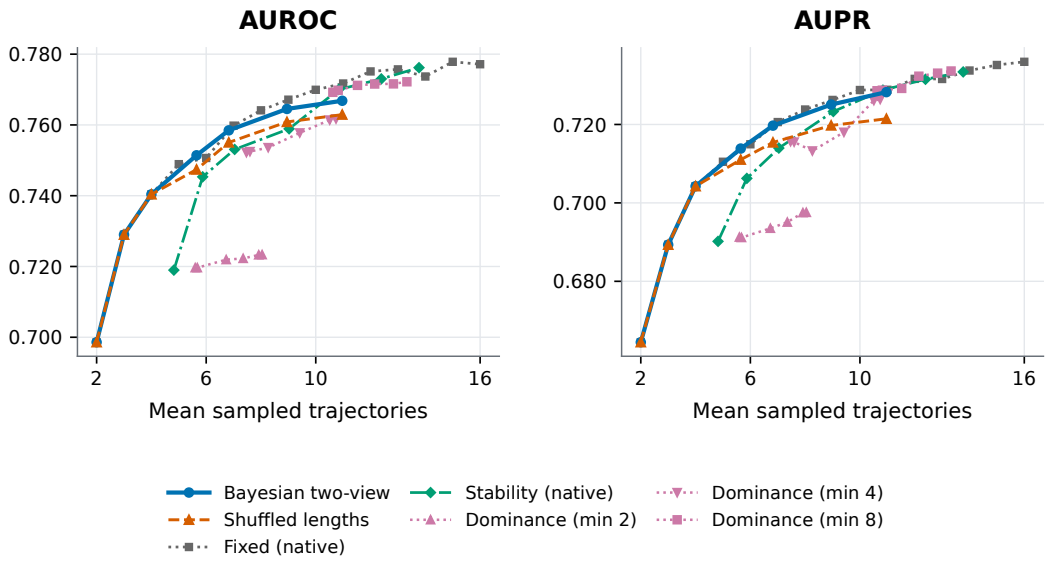

Figure 23: Native quality–cost operating points over 54 cells. The horizontal axis is actual mean trajectories. Heuristic points are not cost-matched to each Bayesian point; use Figure 19 for that comparison. Connected points are descriptive tradeoffs.

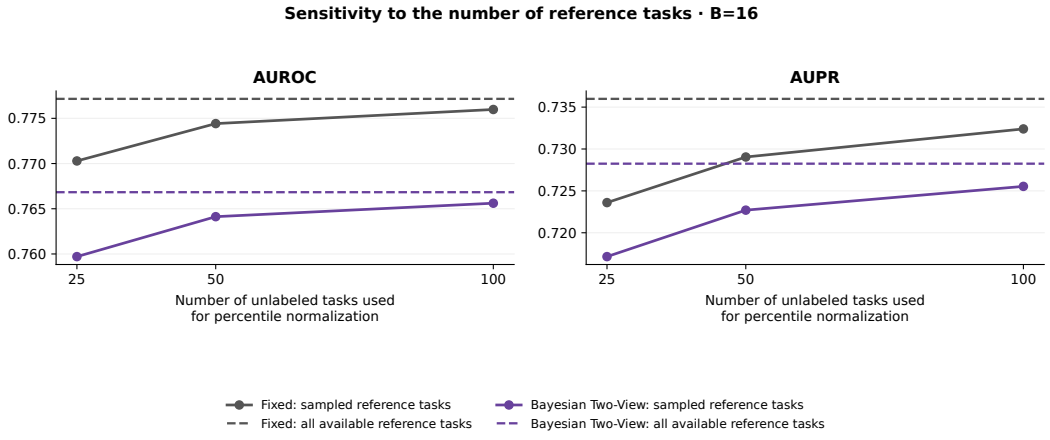

Figure 24: Sensitivity to the number of unlabeled tasks used for percentile normalization, at nominal $B = 16$. Solid curves use sampled reference tasks that exclude the evaluated query; dashed lines use all available tasks in the same cell, including the query. Reference-task counts are distinct from trajectories sampled per task. Points average ten deterministic draws per fold and cell; no uncertainty bands are shown.

G Adaptive Performance and Sampling Efficiency

TURF's strong failure-ranking performance need not require the full trajectory budget for every task. We examine whether its Bayesian two-view policy can retain this predictive utility while reducing sampling, and whether the allocation of trajectories matters beyond the reduction in average budget. Three complementary comparisons separate these questions: adaptive versus fixed sampling with the *same maximum budget*, TURF-Adaptive versus other uncertainty methods at *lower mean trajectory count*, and alternative allocation policies at *matched trajectory cost*. Across the 54-cell evaluation panel, TURF-Adaptive saves up to 31.46% of trajectories, with corresponding mean AUROC/AUPR losses of $1.03/0.77$ percentage points (pp) relative to TURF-Fixed. The results support a useful quality–sampling tradeoff, while the matched-cost controls identify both the benefits and limits of task-dependent allocation.

G.1 Evaluation protocol

We use the six datasets from the main evaluation: AgentBench-DB, HotpotQA, WebShop, StrategyQA, AgentBench-OS, and GAIA, with three backbones and seeds 101, 202, and 303. Each dataset–backbone–seed combination is one cell; the panel contains 54 cells and 22,230 task–cell instances. Metrics are computed within each cell and then averaged with equal cell weights, rather than pooling tasks across datasets.

What changes and what is held fixed. TURF-Adaptive uses the default Bayesian two-view stopping policy; TURF-Fixed uses all B trajectories. Count posterior risk, NLI spectral semantic risk, Gaussianized-rank Stouffer fusion, outcome construction, semantic fallback, percentile normalization, and the WebShop-specific treatment are held fixed. We replay the same ordered trajectory records at $B \in \{2, 3, 4, 6, 8, 12, 16\}$. All model components remain frozen, and the scoring and stopping configurations remain fixed throughout this analysis. The default rule first permits an early stop at prefix five; hence the two policies coincide when $B \leq 4$.

Cost and effect definitions. Let $n^{\star}_{ci}(B) \leq B$ denote the retained trajectory count for task i in cell c, and let $R_{ci}(n)$ denote cumulative recorded resource use through trajectory n. For trajectories, $R_{ci}(n) = n$; for decision steps and output tokens, it is the corresponding cumulative recorded count. We report

$$\text{Savings}_R(B) = \frac{100}{54} \sum_{c=1}^{54} \left(1 - \frac{\sum_i R_{ci}(n^{\star}_{ci}(B))}{\sum_i R_{ci}(B)} \right). \tag{33}$$

Quality effects are adaptive-minus-comparator differences in pp; positive values favor TURF-Adaptive. This definition distinguishes sampling savings from predictive improvements and does not equate trajectory counts with end-to-end runtime.

G.2 Substantial sampling reductions with modest quality losses

Table 15 and Figure 25 show a clear budget-dependent tradeoff. At $B = 8$, TURF-Adaptive uses 6.82 trajectories on average, a 14.70% reduction, while AUROC and AUPR decrease by 0.57 and 0.41 pp. At $B = 12$ and $B = 16$, savings increase to 25.43% and 31.46%, respectively, with mean AUROC losses of approximately one point and AUPR losses below one point. At $B = 6$, the quality changes are small in either direction and savings are 5.87%. Thus the adaptive variant provides a lower-sampling operating point when fewer inspected trajectories are worth the observed loss in ranking quality. This retrospective result does not establish a reduction in live trajectory-generation cost.

These observations should not be interpreted as lossless stopping or as guaranteed one-point quality retention. The observed AUROC loss exceeds one point at $B = 12, 16$, and the $B = 8$ AUROC interval extends slightly beyond a one-point loss. The evidence instead quantifies the cost of choosing a lower-sampling operating point.

Savings are also visible in recorded steps and output tokens. To check that the trajectory reduction does not merely discard unusually short rollouts, we audit costs in the original trajectory

Maximum B	Mean trajectories	Trajectory savings (%)	ΔAUROC (pp)	ΔAUPR (pp)	Decision-step savings (%)
6	5.65	5.87	+0.08	−0.11	5.03
8	6.82	14.70	−0.57	−0.41	12.62
12	8.95	25.43	−1.06	−0.66	22.11
16	10.97	31.46	−1.03	−0.77	27.47

Table 15: Operating points of TURF-Adaptive relative to TURF-Fixed at the same maximum budget. All quantities are equal-cell means over 54 cells; decision-step reductions use complete full-panel records. Quality effects are in percentage points, not relative percentages. These are cached-replay results; uncertainty for quality and trajectory savings is shown in Figure 25.

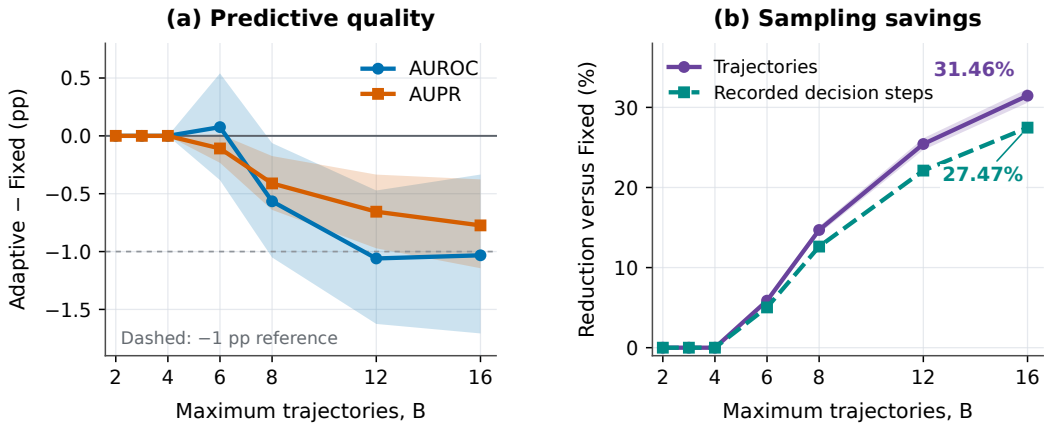

Figure 25: **Quality retention and sampling savings across budgets.** Left: adaptive-minus-Fixed AUROC and AUPR at the same cap. Right: trajectory and recorded decision-step reductions. Shaded regions are paired pointwise 95% task-bootstrap intervals from 50,000 resamples; the decision-step curve is a point estimate without an interval. The dashed −1 pp line is a visual reference, not a validated noninferiority margin. Both favorable and unfavorable quality changes and the full intervals are displayed.

files. All 54 source hashes match their inventory, and task IDs align with the stopping records. Across 1,009,787 recorded decision steps, the full-panel reduction at $B = 16$ is 27.47%. Its difference from the 31.46% trajectory reduction illustrates the importance of accounting for variable rollout lengths.

Token metadata is incomplete for 15 decision steps affecting 10 task–cell instances. For the token-inclusive comparison, we exclude those ten instances entirely, retaining the same 22,220 instances in all 54 cells for every curve and budget in Figure 26. At $B = 16$, retained-response output tokens decrease by 24.97%, and tokens from all candidates recorded in the original files decrease by 22.90%. Missing counts are not imputed. These complementary measures show that selected prefixes retain less recorded step and token work, not merely fewer trajectories. They do not measure work avoided during live generation, since the full pools were already collected.

G.3 Strong failure ranking at a lower trajectory count

We next compare TURF-Adaptive with six representative implementations from the main benchmark: KLE-Final, SNNE-Final, CoCoA-PPL-Final, PPL-StepAgg, SE-StepAgg, and SAUP. These cover semantic/kernel, confidence-combination, likelihood, stepwise semantic-entropy, and trajectory-scoring approaches. The `-Final` and `-StepAgg` suffixes identify the existing final-response and step-aggregation implementations; this comparison does not substitute one adaptation for another.

At nominal $B = 8$, TURF-Adaptive uses approximately 6.82 trajectories, yet its paired AUROC gains over these implementations range from 5.13 to 14.16 pp and its AUPR gains from 4.78 to 10.45 pp (Figure 27). All six mean comparisons remain positive for both metrics at $B = 12$ and $B = 16$

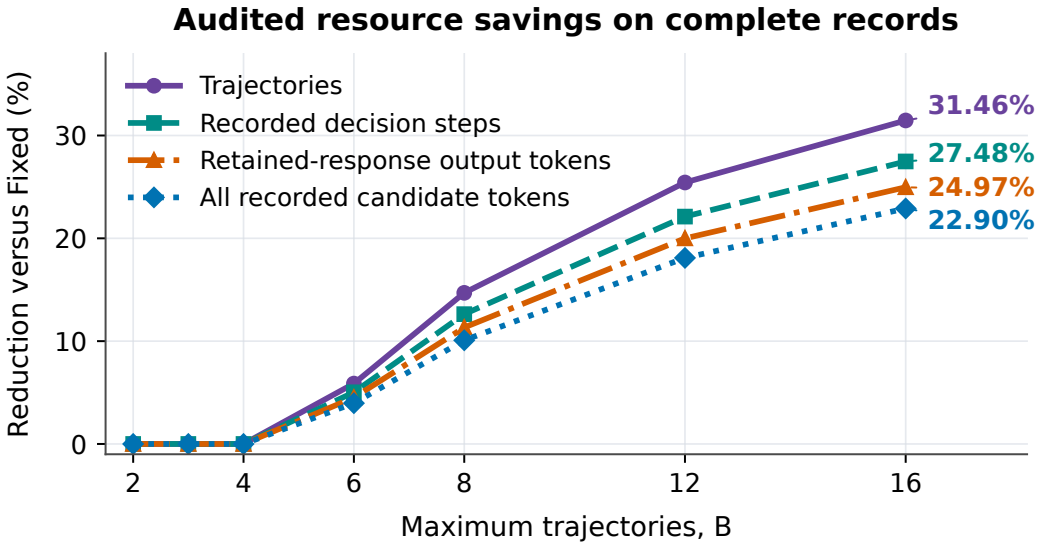

Figure 26: **Resource reductions on identical complete-record support.** All curves use 22,220 task–cell instances across 54 cells, excluding ten instances with incomplete token metadata. Ratios are computed within cells and averaged as in Eq. 33; curves are descriptive point estimates. Endpoint labels give reductions at $B = 16$. "All recorded candidate tokens" includes only candidates present in the original trajectory files, not separately cached alternatives or unlogged generation attempts. The trajectory/step values therefore differ slightly from the full-panel values in Table 15.

(Figure 28). Thus the sampling reduction does not erase TURF's observed predictive advantage over these representative uncertainty scores.

Task IDs and failure labels are paired separately for each baseline and cell. TURF metrics are recomputed on these intersections without changing its percentile maps; all 54 cells remain represented. Individual-cell coverage is 92.36%–100%. The native baseline scores reconstruct all 972 method–budget–cell source metric rows to numerical precision. These comparisons use a common nominal cap, not identical realized cost: they demonstrate predictive performance at reduced sampling, but do not isolate the contribution of the allocation policy or account for baseline-specific inference overhead.

G.4 Does task-dependent allocation help at matched cost?

A same-cap comparison alone cannot distinguish adaptive allocation from simply using a smaller fixed budget. We therefore hold the TURF scorer fixed and compare four allocation controls at the Bayesian policy's mean cost within each cell: randomized adjacent *fixed budgets*, *outcome stability*, *outcome dominance*, and *shuffled stopping lengths*. The last control preserves the exact per-cell distribution of stopping lengths while permuting their assignment to tasks, directly testing whether *which tasks receive more samples* matters.

Fixed, stability, and dominance controls use randomized mixtures of bracketing settings chosen by cost, not failure labels. Results average 100 prespecified schedules within nonconstant cells before averaging cells. The largest absolute per-cell cost mismatch is 0.0392 trajectories; shuffling preserves cost exactly. These controls should not be confused with deterministic native heuristic settings at unmatched costs.

Figure 19 shows higher mean AUROC and AUPR for TURF-Adaptive than for all four controls at $B = 6, 8, 12$. At $B = 8$, improvements over matched Fixed are $0.54/0.40$ pp. The shuffled-length comparison is particularly informative: at $B = 12$ and $B = 16$, TURF-Adaptive gains $0.37/0.53$ and $0.39/0.68$ pp, respectively, without changing the stopping-length distribution. These effects support useful task-dependent allocation beyond average cost alone.

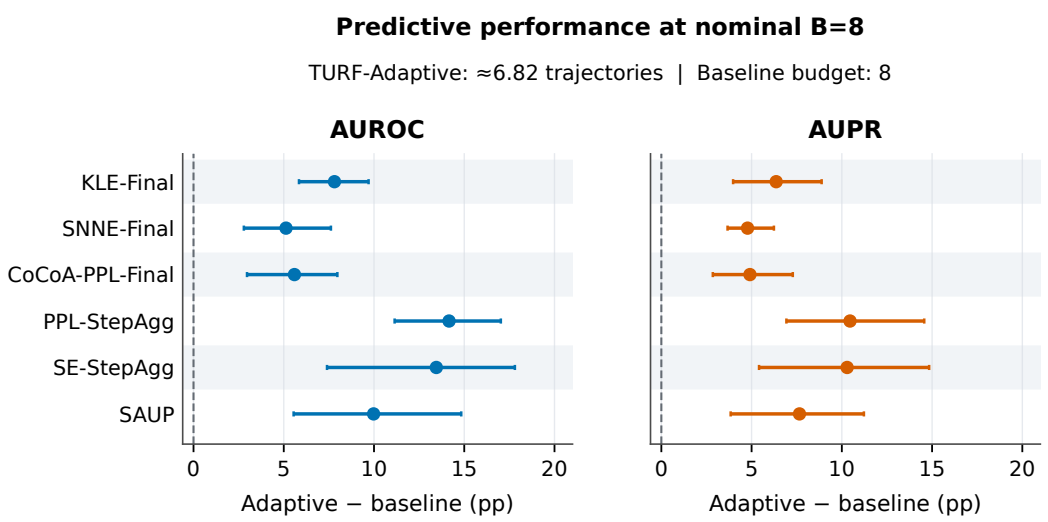

Figure 27: **Strong predictive performance with fewer sampled trajectories.** TURF-Adaptive minus six selected baseline implementations at nominal $B = 8$. Positive effects favor TURF-Adaptive. Each comparison uses exact paired task support within all 54 cells; the mean adaptive count varies slightly with the method-specific intersection. Whiskers are diagnostic 95% dataset-cluster percentile intervals from 50,000 resamples, as defined in Section G.6. These are not the task-bootstrap intervals used in Figure 25.

The advantage is nevertheless budget-dependent. At $B = 16$, differences against matched Fixed are $-0.22/+0.09$ pp; against matched Stability they are $-0.22/+0.17$ pp. We therefore do not claim uniform dominance over simple allocation rules. Figure 20 shows all comparisons and their diagnostic intervals, including the negative AUROC effects and intervals crossing zero.

G.5 Where are trajectories saved?

Figure 29 makes the allocation behavior visible across datasets. At $B = 16$, StrategyQA uses approximately 6.4 trajectories on average, whereas GAIA uses approximately 15.0. The policy therefore does not impose an almost uniform reduction across domains: the same rule permits early exits in some domains while using nearly the full cap in others. This is consistent with different persistence patterns in the two recurrence views, but does not by itself establish that early-exiting tasks are easier or more accurately ranked. The terminal column includes tasks that simply exhaust their budget; it does not imply that they met the posterior stopping criterion.

G.6 Uncertainty, accounting boundary, and interpretation

Two resampling schemes answer different questions. For same-cap quality retention and trajectory savings, we use 50,000 paired task-bootstrap replicates. Canonical tasks are sampled jointly across the nine backbone–seed cells within each dataset, cell metrics are recomputed, and cells are then averaged equally. The six datasets, percentile maps, and stopping decisions are fixed. Intervals are two-sided pointwise 95% basic intervals. For baseline and matched-cost effects, diagnostic percentile intervals instead resample the six dataset-level paired effects 50,000 times, retaining the nine cells within each dataset. These condition on the observed task scores and do not capture task-sampling uncertainty. Both schemes yield descriptive pointwise intervals conditional on the reported configuration, without adjustment for multiple comparisons.

What the resource audit does and does not measure. Decision-step counts exclude forced terminal-finish records. Token counts come from recorded metadata, with retained-response counts cross-checked against the selected candidate where both records exist. Complete-record exclusion may be selective. Resource totals omit input tokens, unlogged retries, separately cached alternatives, NLI inference, percentile-reference construction, and tool/environment runtime. Consequently, the

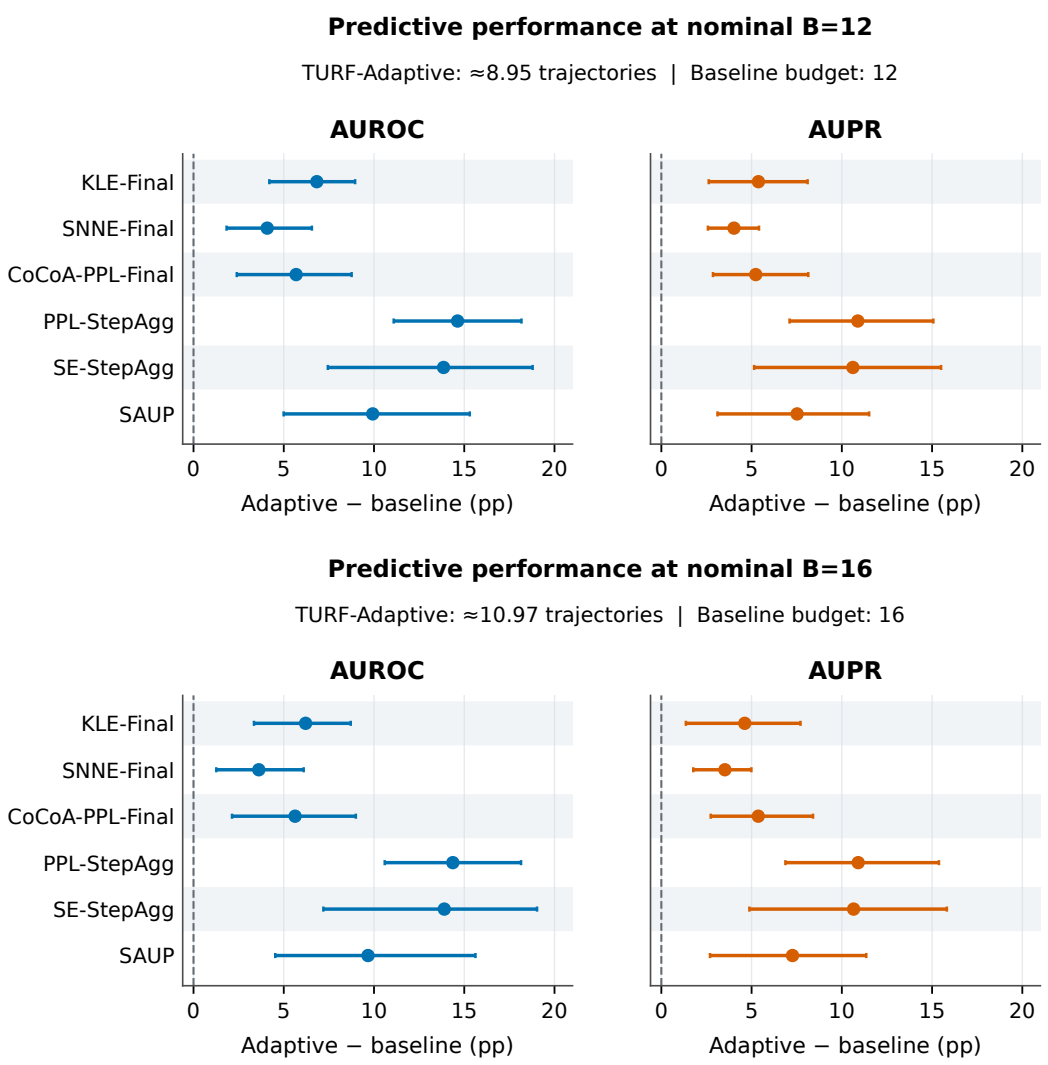

Figure 28: **Reduced-sampling comparisons at higher caps.** Top: $B = 12$; bottom: $B = 16$. Definitions, paired-support checks, and diagnostic intervals are identical to Figure 27. Both metrics and all six selected methods are retained at every cap. The same horizontal scales are used across budgets to avoid exaggerating differences through rescaling. These results concern the named implementations, not all possible uncertainty estimators or adaptations.

reductions measure recorded costs under cached-prefix replay, not live wall-clock or end-to-end compute speedups.

H Hyperparameter Sensitivity and Robustness

We examine whether TURF requires precise parameter choices, how its reference cohort affects ranking, and how its sampling policy trades predictive quality for trajectory cost. The Fixed scoring sweeps vary one parameter at a time around the reported default; Bayesian stopping sweeps additionally evaluate parameter interactions. All results use the existing six datasets, three backbones, and seeds $101, 202, 303$, with equal weight for each of the 54 cells. WebShop-specific experiments average its nine cells. The original task support, outcome adapters, frozen NLI inputs, and semantic availability mask are retained. Default configurations are reproduced before comparison.

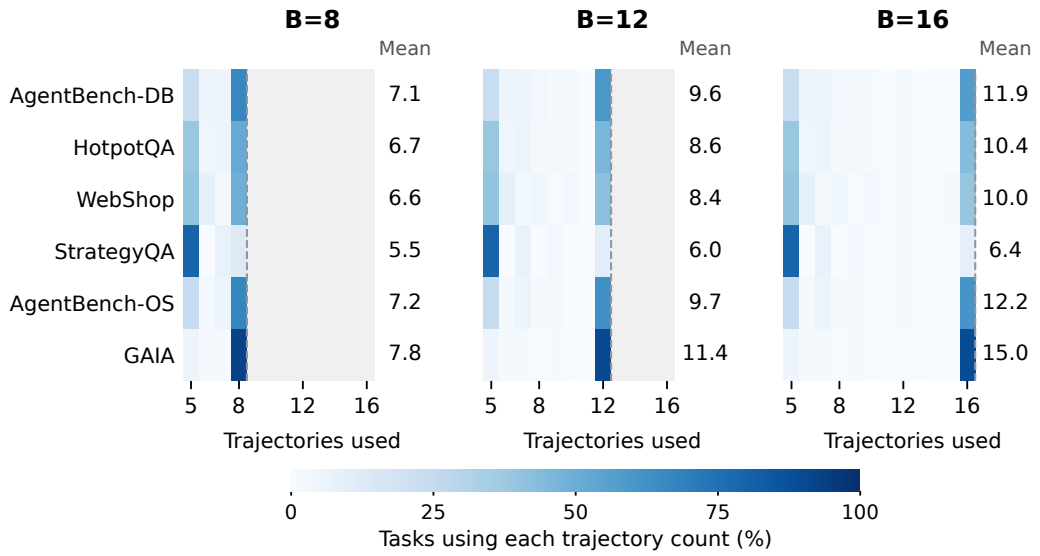

Figure 29: **Heterogeneous stopping behavior across datasets.** Rows average within-cell stopping fractions equally over three backbones and three seeds. All panels share a linear 0–100% color scale; numbers on the right give mean trajectories. Counts below five cannot be selected at these caps and are omitted; gray columns exceed the cap. The dashed vertical line marks each cap's right boundary. Stopping at the cap includes budget exhaustion, not only criterion-triggered stops.

Table 16: Scoring sensitivity at fixed $B = 16$. AUROC/AUPR are equal-cell macro means. Ranges cover every listed setting and are not uncertainty intervals. The default achieves $0.7772/0.7360$.

Parameter	Tested values	AUROC range	AUPR range
β	$0.1, 0.25, 0.5, 1, 2$	0.7756–0.7774	0.7346–0.7369
κ	$0, 1, 1.645, 1.960, 2.576, 3$	0.7765–0.7775	0.7350–0.7365
w_C	$0, 0.25, 0.4, 0.5, 0.6, 0.75, 1$	0.7547–0.7797	0.7237–0.7362

Unless stated otherwise, changes are in *percentage points* (pp). Scoring ribbons show one sample standard deviation across three seed-level macro means, each averaging the dataset–backbone cells. Reference-bank error bars instead summarize variation across ten bank draws. Neither is a confidence interval. Sampling cost counts inspected trajectories in cached replay; it excludes reference acquisition and model-inference overhead.

H.1 Count regularization and fusion weights

For TURF-Fixed at $B = 16$, we vary the Dirichlet pseudocount β, the count-risk penalty κ, and the count-view weight w_C. The default uses $\beta = 1/2$ and $\kappa = \Phi^{-1}(0.975)$ in Eq. (3), retaining its variance floor and bounded surprisal term. Numerical grid labels round the Gaussian multipliers for display; the default computation uses 1.959963984540054. Smoothing changes both the observed-outcome pseudocount and the total unseen mass; WebShop splits its unseen mass equally across two leaves. Weighted Stouffer fusion is

$$S_n^{\text{core}}(w_C) = \Phi\left(\frac{w_C Z_C + (1 - w_C) Z_H}{\sqrt{w_C^2 + (1 - w_C)^2}}\right), \qquad w_C = 0.5 \text{ by default.} \tag{34}$$

Tasks without semantic information retain the count fallback; WebShop's separate severity mixture is fixed in these sweeps. The native surprisal tie-break scale is held fixed. WebShop moments use 2,048 posterior draws, so its smoothing sweep includes finite Monte Carlo variation.

Count regularization has a small aggregate effect over the tested ranges (Table 16). The largest absolute changes from the default are $0.151/0.142$ pp in AUROC/AUPR for smoothing and $0.068/0.094$

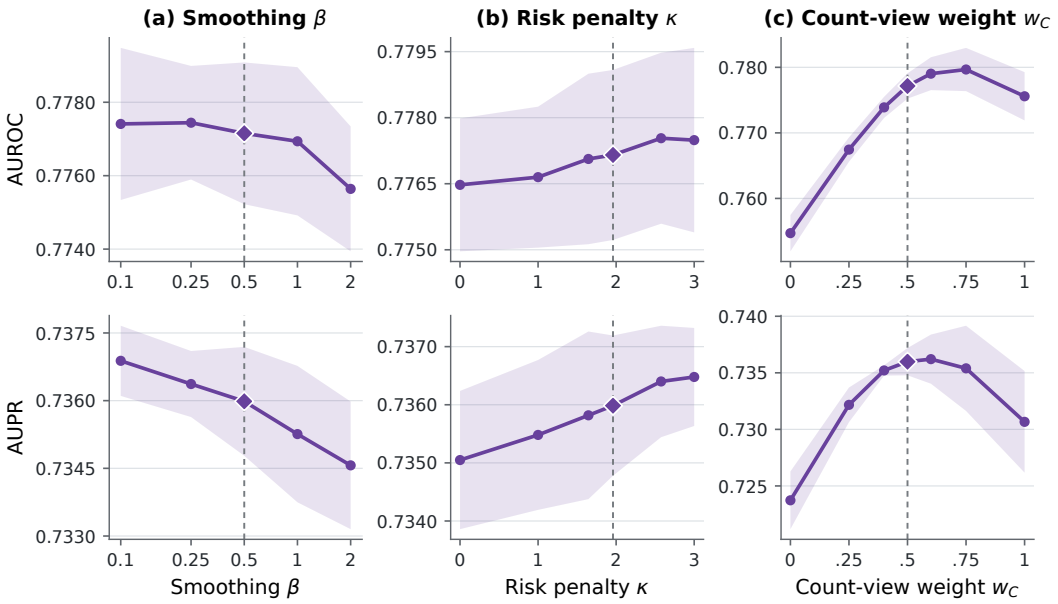

Figure 30: One-factor scoring sensitivity across 54 cells at fixed $B = 16$. Vertical dashed lines and diamond markers indicate defaults. The smoothing axis is logarithmic. Bands are ± 1 seed-level macro standard deviation. All settings use the same number of trajectories.

pp for the penalty. Fusion weights have a larger effect (Figure 30). The default equal mixture achieves $0.7772/0.7360$, compared with $0.7547/0.7237$ at the semantic endpoint and $0.7756/0.7307$ at the count endpoint. A moderate shift toward counts ($w_C = 0.6$) yields $0.7790/0.7362$; $w_C = 0.75$ further raises AUROC to 0.7797 but lowers AUPR to 0.7354. Equal weighting is therefore a *simple symmetric default*, not the empirically best balance on this grid. In particular, $w_C = 0.6$ has higher observed means on both metrics, while $w_C = 0.75$ trades higher AUROC for lower AUPR relative to the default. These descriptive differences do not establish a statistically superior or generally optimal weight. The equal-signal Gaussian working model provides a conditional symmetry rationale, not evidence that its assumptions hold in these data (Appendix C.4). The sweep characterizes sensitivity around the reported default and does not establish Stouffer's superiority over a different fusion function.

H.2 Reference exclusion and finite unlabeled banks

We evaluate both TURF-Fixed and TURF-Adaptive, where Adaptive denotes Bayesian two-view stopping, at nominal cap $B = 16$ in all 54 cells. Only the rank-reference construction changes: component statistics, semantic availability, and each task's policy-selected prefix remain fixed. Reference tasks contribute statistics at their own selected prefixes under the same policy. We retain the established tie-aware external empirical-CDF remapping of cached component ranks, count fallback, and WebShop outer ranking described in Appendix B.4.

First, five-fold out-of-fold references score each task using only the other four folds. A deterministic, label-independent hash assigns balanced folds. Relative to full-cell scoring, macro AUROC/AUPR change by $-0.045/-0.039$ percentage points (pp) for Fixed and $-0.036/+0.011$ pp for Adaptive (Table 17). These small observed changes indicate limited sensitivity to query inclusion under this protocol; they establish neither invariance nor robustness to a reference cohort from another domain.

Second, we reserve fold zero as a fixed evaluation cohort and sample disjoint reference banks from the other four folds. We use ten deterministic draws, with nested bank sizes within each draw and identical bank membership for Fixed and Adaptive. Every size and its corresponding full-cell baseline score identical held-out queries within a cell. Banks of 25, 50, and 100 tasks are feasible in all 54 cells. A 200-task bank is feasible only in AgentBench-DB, HotpotQA, and StrategyQA; all sizes are repeated on those 27 cells to isolate bank-size effects within a common evaluation support.

Table 17: Five-fold out-of-fold minus full-cell rank-reference performance at $B = 16$, equally averaged over the same 54 cells and evaluated on all tasks. Adaptive uses Bayesian two-view stopping. Entries are observed effects in percentage points, not confidence intervals.

Policy	ΔAUROC	ΔAUPR
TURF-Fixed	−0.045	−0.039
TURF-Adaptive	−0.036	+0.011

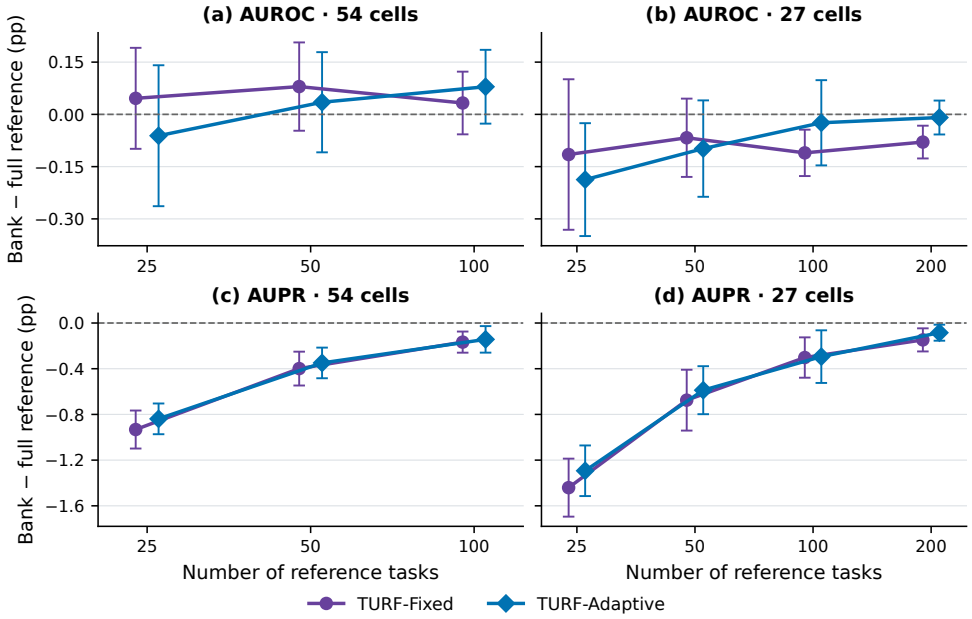

Figure 31: **Reference-bank sensitivity of Fixed and Bayesian Adaptive TURF.** At nominal $B = 16$, effects compare each disjoint bank with the same policy's full-cell reference on identical held-out queries. Left: all 54 cells. Right: all sizes on the 27 cells eligible for 200-task banks. Error bars show one sample standard deviation of the ten bank-draw macro effects, not task-sampling confidence intervals. Bank membership is shared across policies; scoring-prefix costs differ.

On the 54-cell support, increasing the bank from 25 to 100 tasks reduces the observed AUPR loss from 0.932 to 0.167 pp for Fixed and from 0.839 to 0.143 pp for Adaptive (Figure 31). The corresponding AUROC changes are +0.046 and +0.033 pp for Fixed, and −0.061 and +0.079 pp for Adaptive. Thus small banks affect AUPR more strongly than AUROC in this study; reference exclusion does not necessarily decrease every metric. On the separate 27-cell support, the 200-task AUPR losses are 0.148/0.085 pp for Fixed/Adaptive. These results do not identify a universal minimum bank size or establish robustness under domain shift.

The out-of-fold table evaluates all tasks, whereas the finite-bank figure evaluates only the held-out fold. The separate rotating-fold study in Appendix F scores all tasks and uses separately seeded banks across policies; its finite-bank values are not interchangeable with this paired-bank, held-out-query study. All analyses remap cached statistics and retain the full-pool semantic-availability mask. They do not evaluate prospective bank acquisition, decision-time semantic availability, or the cost of constructing reference statistics.

H.3 WebShop hierarchy and severity weights

We examine the two WebShop-specific weights for both TURF-Fixed and Bayesian TURF-Adaptive across nine backbone–seed cells and all seven budgets $B \in \{2, 3, 4, 6, 8, 12, 16\}$. The hierarchy weight $\gamma \in \{0, 0.125, 0.25, 0.5, 1\}$ multiplies the purchase-branch conditional entropy term; its default is 0.125. The separate severity weight $w_S \in \{0, 0.25, 0.5, 0.75, 1\}$ enters

$(1 - w_S)R_S(S_n^{\text{core}}) + w_S R_V(V_n)$, where V_n is the recorded severity fraction, and defaults to 0.5. Each sweep holds the other weight fixed. Neither weight changes the outcome buckets or raw semantic affinities used for stopping, so each task's selected prefix is held identical across weights within a policy and cap. The hierarchy sweep shares the same 2,048 posterior samples per task and prefix across all γ values. Default scores and metrics reproduce the archived Fixed and Bayesian results.

Figure 32 displays the $B = 16$ slice. With default weights, Fixed achieves AUROC/AUPR of $0.7020/0.9293$ using 16 trajectories, whereas Adaptive achieves $0.7066/0.9271$ using 9.969 trajectories on average across these nine WebShop cells. These policy values are common-cap, not equal-cost comparisons. Setting $\gamma = 0$ gives $0.6943/0.9281$ for Fixed and $0.7065/0.9268$ for Adaptive; $\gamma = 0.25$ gives $0.7034/0.9296$ and $0.7089/0.9278$, respectively. The observed effect of within-purchase dispersion is therefore policy-dependent, and the default does not maximize either metric over the tested hierarchy grid.

Severity weighting has a larger effect on the displayed means. Without the severity view ($w_S = 0$), Fixed and Adaptive achieve $0.6454/0.9185$ and $0.6405/0.9160$; severity alone ($w_S = 1$) achieves $0.6873/0.9094$ and $0.6924/0.9079$, respectively. The default mixture exceeds both single-view endpoints in both metrics for each policy. Increasing w_S to 0.75 raises AUROC to $0.7055/0.7132$ for Fixed/Adaptive but reduces AUPR to $0.9277/0.9264$. These observations support combining the core and severity views on this panel while exposing a metric-dependent tradeoff among interior weights. They do not establish statistical superiority, universal optimality, or a reason to replace the defaults.

Numerical correction to the severity control. This sweep computes severity ranks directly from recorded severity fractions. The earlier sensitivity implementation recovered them by subtracting the core rank from twice the fused score; floating-point roundoff could split exact ties, particularly at $w_S = 1$. The corrected Fixed severity-only endpoint is therefore $0.6873/0.9094$, replacing $0.6823/0.9107$. Default scores, stopping decisions, and default performance are unchanged.

H.4 Bayesian two-view stopping parameters

TURF-Adaptive uses Bayesian two-view stopping with the frozen scorer. It monitors recurrence of the first rollout's outcome and semantic agreement with that rollout. Its default parameters are semantic cutoff $\eta = 0.75$, recurrence threshold $\theta_0 = 0.5$, joint Dirichlet concentration $a_B = 0.5$ per category, posterior tolerance $\delta_{\text{stop}} = 0.2$, and minimum prefix $n_{\min} = 5$. The induced marginal prior is $\text{Beta}(2a_B, 2a_B)$. The Bayesian concentration a_B and scoring pseudocount β are distinct parameters. Here the subscript in a_B identifies the Bayesian prior; it does not denote dependence on the trajectory cap B. For this sensitivity experiment, the posterior becomes

$$\vartheta_v \mid \mathcal{D}_n \sim \text{Beta}\big(2a_B + m_v(n),\, 2a_B + n - 1 - m_v(n)\big), \qquad \pi_v(n) = \Pr(\vartheta_v > \theta_0 \mid \mathcal{D}_n).$$

The binomial-CDF identity in the main paper applies to the evaluated default $a_B = \theta_0 = 1/2$, not to every setting in this sweep. The policy uses raw outcomes before UnknownCap2 for non-WebShop recurrence, mapped WebShop outcomes, and an outcome-only branch when semantics are unavailable; the score retains its original encoded outcomes.

We evaluate 97 distinct settings at all seven budgets across 54 cells. One-factor grids are $\eta \in \{0.50, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95\}$, $\theta_0 \in \{0.40, 0.45, 0.50, 0.55, 0.60, 0.65\}$, $a_B \in \{0.125, 0.25, 0.5, 1, 2\}$, $\delta_{\text{stop}} \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.40\}$, and $n_{\min} \in \{2, 3, 4, 5, 6, 8\}$. We also evaluate the full $\eta \times \delta_{\text{stop}}$ and $a_B \times \theta_0$ grids with other parameters fixed. Varying θ_0 changes the recurrence event being certified; lowering $n_{\min}$ can also change small-budget behavior in the outcome-only branch, so those budgets are evaluated rather than assumed identical.

The semantic cutoff produces comparatively small changes over its full range, supporting limited dependence on the specific value 0.75. Recurrence threshold, posterior tolerance, and minimum prefix affect cost more substantially (Table 18). For example, raising $n_{\min}$ from 5 to 8 improves AUROC/AUPR by $0.528/0.385$ pp but raises mean trajectory cost from 10.966 to 12.217. The interaction grids (Figure 34) expose the joint effects hidden by one-factor sweeps.

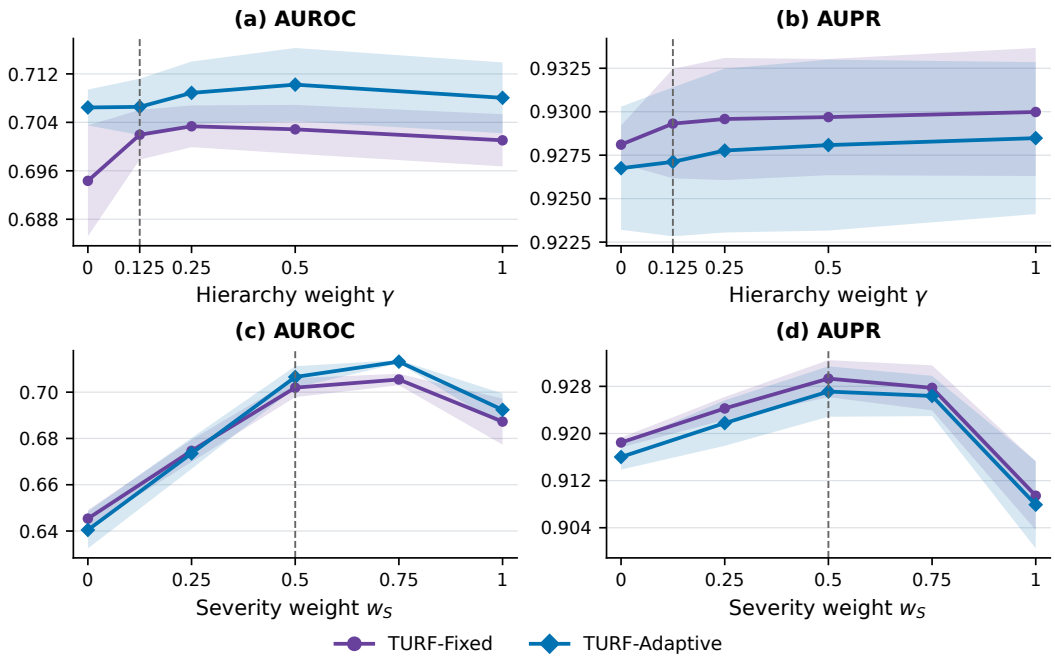

Figure 32: **WebShop scoring sensitivity under Fixed and Bayesian Adaptive sampling.** At $B = 16$, hierarchy weight γ (top) and severity weight w_S (bottom) vary separately, with the other weight fixed at its default. Each panel equally averages nine cells. Bands show one sample standard deviation across three seed macro means, not confidence intervals; vertical dashed lines mark defaults. Within each policy, trajectories and stopping decisions are unchanged across weights: Fixed uses 16 trajectories and Adaptive uses 9.969 on average.

Table 18: Bayesian policy sensitivity at $B = 16$. Each row covers its entire one-factor grid. Predictive ranges are changes from the Bayesian default in percentage points, not confidence intervals.

Parameter	Savings (%)	ΔAUROC (pp)	ΔAUPR (pp)
η	30.33–32.52	$[-0.149, +0.054]$	$[-0.101, +0.028]$
θ_0	23.32–37.75	$[-0.535, +0.125]$	$[-0.891, +0.270]$
a_B	26.66–33.75	$[-0.258, 0.000]$	$[-0.487, +0.215]$
δ_{stop}	24.59–37.25	$[-0.540, +0.125]$	$[-0.910, +0.269]$
$n_{\min}$	23.65–36.26	$[-0.541, +0.528]$	$[-0.276, +0.385]$

Allocation comparisons at equal cost. For every setting we evaluate two randomized controls with 100 schedules per cell: adjacent-integer allocations matching the exact total cost (*EqualCost*), and shuffled stopping lengths preserving both total cost and the stopping-length histogram (*Shuffled*). Ranks and scores are recomputed for every schedule; metrics are averaged over schedules within cells before macro averaging. At $B = 16$, the default saves 31.46% of trajectories, with AUROC/AUPR $0.7668/0.7282$. Against Shuffled, its effects are $+0.392/+0.687$ pp, with diagnostic intervals $[0.060, 0.750]/[0.214, 1.242]$ pp. Against EqualCost, the effects are $-0.226/+0.083$ pp, with intervals $[-0.688, 0.301]/[-0.207, 0.380]$ pp. The policy's task-specific allocation therefore improves observed ranking relative to random reassignment of its own lengths, while the evidence does not establish better AUROC and AUPR than adjacent-integer equal-cost allocation. Across the semantic-cutoff sweep, the same distinction persists in the point estimates: AUROC/AUPR effects are $+0.311$–$+0.420/+0.638$–$+0.705$ pp against Shuffled, versus -0.338–$-0.191/+0.021$–$+0.111$ pp against EqualCost (Figure 33).

The diagnostic intervals use 50,000 resamples of six dataset clusters, conditional on observed task scores, randomized-schedule averages, and the reported configuration. They are descriptive and pointwise, omit task-sampling uncertainty, and are not multiplicity-adjusted. Maximum macro Monte

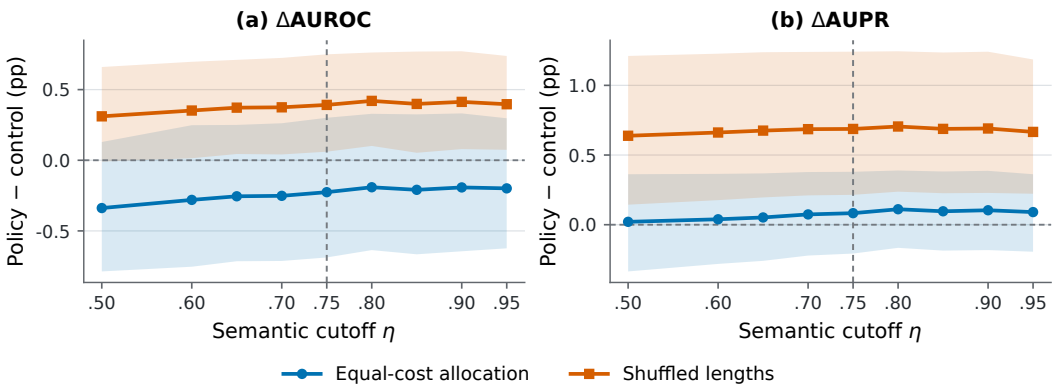

Figure 33: Bayesian semantic-cutoff sensitivity at $B = 16$, including comparisons with controls at exactly matched trajectory cost. Diagnostic intervals resample datasets and condition on observed task scores, schedule means, and the reported configuration. They are descriptive, pointwise, and not multiplicity-adjusted.

Carlo standard errors across controls are $0.0165/0.0154$ pp for AUROC/AUPR. These intervals and errors summarize different sources of variation. The posterior tolerance has a model-relative interpretation and is not a frequentist guarantee on predictive quality.

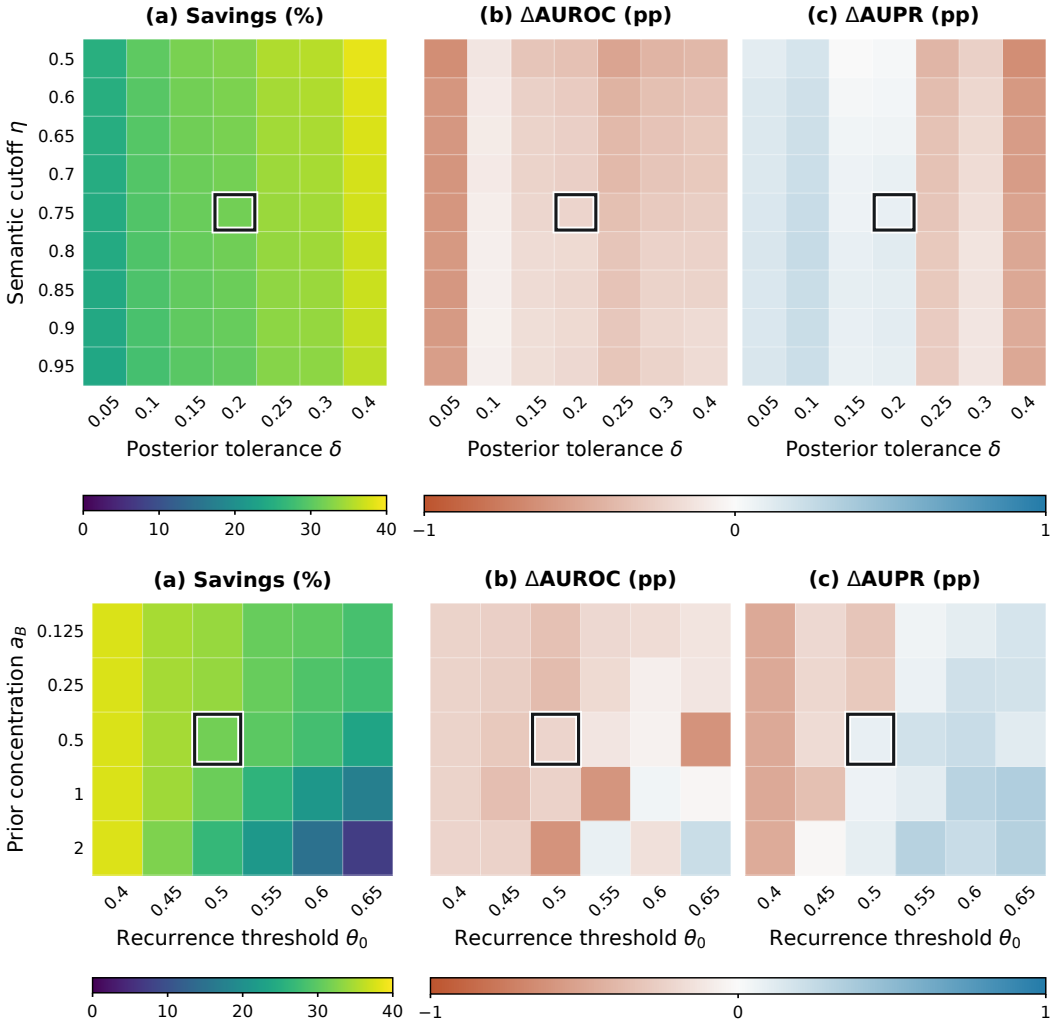

Figure 34: Bayesian stopping interactions at $B = 16$: semantic cutoff and posterior tolerance (top), prior concentration and recurrence threshold (bottom). Quality effects are relative to EqualCost. Both grids use common color scales, with predictive effects centered at zero. Outlined cells mark the default. The axis abbreviation δ denotes δ_{stop}; a_B is the per-category Bayesian prior concentration, not the cap. All grid settings are shown; exact values accompany the source tables.