AURA: MEASURING EVIDENCE FROM REPEATED CALLS TO HOSTED FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Repeated sampling is a standard way to improve foundation-model answers, but on a hosted service each sample is a billed API call. We introduce AURA, an output-only framework for asking how many independent categorical votes would match the plurality utility of a repeated-call trajectory. The resulting VOTE-ESS is an operational matched-budget index, not an estimator of latent correlation. We validate it on 180 Dirichlet–multinomial configurations and an exact block-replication control: it decreases with dependence in every fixed configuration and recovers the block budget exactly in 96.3% of cases. In a broader 288-cell stress test spanning four dependence mechanisms, an interpolated isotonic categorical VOTE-ESS reduces non-i.i.d. measurement error by 86.4% relative to the strongest output-only proxy. Its exact-block compatible intervals, however, cover only 83.5%, below a prespecified 90% coverage threshold, and its stopping controller incurs 0.0324 more regret than a raw-prefix controller. On frozen Qwen trajectories, the categorical raw/isotonic terminal ratios are 0.944/0.981 over 2,052 disjoint test trajectories, while the direct reference-minus-observed plurality gap is -0.0063 (95% question-bootstrap CI $[-0.0101, -0.0024]$). An exact-condition cross-fit over 752 test trajectories instead gives $+0.0011$ $[-0.0075, +0.0103]$. Breadth checks over 69 aliases and full matched runs on two official DeepSeek models show that the protocol transfers, but not that all providers are independent. Finally, prefix-only controllers save about 0.22–0.25 calls without detectable gains in accuracy or matched objective. These results support a disciplined way to measure repeated-call evidence while showing that measurable shrinkage need not justify a different test-time compute policy.

1 INTRODUCTION

Chain-of-thought prompting elicits intermediate reasoning (Wei et al., 2022; Kojima et al., 2022); self-consistency samples multiple paths and aggregates their answers (Wang et al., 2023). Adaptive stopping, difficulty-aware allocation, and confidence- or reasoning-aware sampling seek to reduce sampling cost (Aggarwal et al., 2023; Li et al., 2024; Wang et al., 2025b; Taubenfeld et al., 2025; Wan et al., 2025). Broader test-time scaling studies examine adaptive compute allocation (Snell et al., 2025) and solution coverage from increased sampling (Brown et al., 2024). For a hosted foundation model, however, a sample is not merely an experimental draw: it is an API call with cost, latency, a provider-controlled execution path, and possible drift over time (Atıl et al., 2025; Gao et al., 2025; Cai et al., 2025). This makes a simple question operationally important: *when does another call add evidence rather than mostly reproduce the evidence already seen?*

Raw answer agreement does not resolve this question. High agreement can mean either reliable concentration on a correct answer or redundant concentration on a mistake (Tan et al., 2025). Conversely, diverse wrong strings can obscure a stable categorical decision. We therefore measure the user-visible operation that self-consistency actually performs: plurality voting over canonical answer classes. For a trajectory of n calls, AURA asks how many independent draws from a matched categorical answer distribution would attain the same plurality utility. We call this reference-relative budget VOTE-ESS. It is designed for black-box model access: outputs, public request parameters, and provider usage counters are used, without model internals. Offline plurality utility additionally uses the benchmark's accepted gold answers; "output-only" does not mean label-free correctness evaluation. Online controllers receive only prefix-visible features, not the gold class or unseen suffix.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

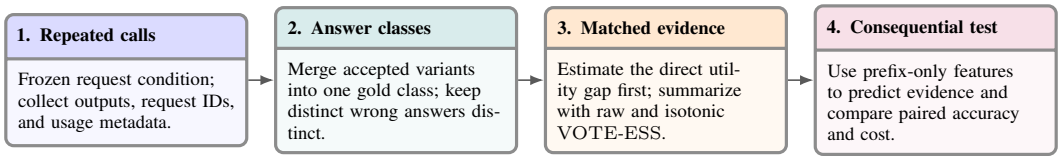

Figure 1: AURA separates data collection, categorical measurement, and the downstream decision test. Later stages may use only information available from earlier stages. A measured evidence difference is useful only if it can be estimated from prefix-visible information and improves a matched decision rule.

The main methodological risk is inversion: a flat or nonmonotone finite plurality-utility curve can amplify small utility perturbations in its inverse. First, we validate the index on known-dependence simulations and an exact block construction. Second, we make the non-inverted reference-minus-observed utility gap the primary empirical estimand, with raw and isotonic VOTE-ESS as diagnostics. Third, we test whether prefix-visible features predict apparent budget shrinkage and whether that prediction improves a matched stopping controller.

Our evaluation follows an evidence hierarchy. The deepest Qwen branch uses frozen repeated trajectories, disjoint test data, and unique-question bootstrap intervals. A stricter sensitivity analysis estimates each categorical reference only from other trajectories collected under an exactly matched condition. A 69-alias breadth panel tests scope but has only one trajectory per exact condition and is reported as plug-in evidence. Finally, two official DeepSeek models repeat seven frozen studies each, allowing a provider-portability check without converting it into a universal claim.

The study makes four contributions:

- We define categorical VOTE-ESS as an output-only, plurality-matched budget and pair it with a direct utility-gap estimand.
- We establish its behavior on a 180-condition Dirichlet–multinomial grid, an exact block control, finite-trajectory recovery tests, and a 288-cell stress test spanning four dependence mechanisms.
- We evaluate measurement, exact-condition cross-fitting, breadth, and cross-provider portability without conflating their evidence levels.
- We show a useful negative result: learned effective-budget controllers reduce calls modestly but do not detectably improve aggregate accuracy or a matched accuracy–cost objective over raw prefix length.

Figure 1 summarizes the path from repeated calls to a decision-relevant test. The separation matters: a measured budget difference is only operationally useful if it can be estimated from information available at stopping time and improves a matched decision rule.

This is deliberately not a claim that hosted outputs are statistically independent. It is a measurement and decision result for plurality voting under the observed endpoint, prompt, decoding, and collection contracts.

2 Related Work

Repeated sampling and test-time allocation. Self-consistency aggregates sampled reasoning paths (Wang et al., 2023); universal and ranked variants broaden the aggregation rule (Chen et al., 2024b; Wang et al., 2025a). Work on compound AI systems also shows that voting performance need not increase monotonically with call count (Chen et al., 2024a). Early stopping and difficulty-adaptive allocation reduce cost by estimating when further samples are unlikely to help (Aggarwal et al., 2023; Li et al., 2024; Wang et al., 2025b; Feng et al., 2025). Confidence- and reasoning-aware methods exploit richer signals when available (Taubenfeld et al., 2025; Wan et al., 2025; Huang et al., 2026). RPC combines internal probabilities with self-consistency in an initial study (Zhou et al., 2025b) and a later theoretical treatment (Zhou et al., 2025a). AURA is complementary: it tests whether raw call count is an adequate evidence budget before choosing an allocation rule.

Effective sample size. Classical effective sample size summarizes variance inflation for linear statistics or Markov-chain averages (Kish, 1965; Geyer, 1992). Repeat-prompting work shows that ignoring prompt-level clustering can understate uncertainty (Gallo et al., 2025); this does not by itself identify serving-induced dependence under a fixed question. Plurality vote is nonlinear and multicategorical, so the familiar $n/[1 + (n - 1)\rho]$ expression is not its target. Our quantity instead matches the utility of a concrete decision rule against an empirical categorical reference. Order-restricted isotonic regression is used only to stabilize a finite reference curve (Barlow et al., 1972); the raw curve and the direct, non-inverted utility gap remain visible.

Hosted-model evaluation. Broad benchmarks span diverse capabilities (Srivastava et al., 2023), while holistic evaluation emphasizes explicit scenarios, metrics, and reporting contracts (Liang et al., 2023). Black-box services may drift or conceal model substitutions, motivating output-level identity and nondeterminism tests (Atıl et al., 2025; Gao et al., 2025; Cai et al., 2025; Zhu et al., 2026). We instead ask how much plurality evidence repeated sampling delivers under a recorded API contract, storing trajectories and analyzing question-clustered uncertainty rather than treating calls within a question as independent experimental units.

3 Output-only measurement

3.1 Observed plurality utility

For question x, let $Y_{x1}, \ldots, Y_{xM}$ be canonical answer classes from M repeated calls and let g_x denote the accepted gold class. Multiple accepted surface forms are collapsed into this single class before voting; parse failures remain an explicit category and distinct wrong classes remain separate (Appendix B). For a subset S of n calls, the plurality score is one when g_x is the unique winner, $1/k$ when it is tied among k winners, and zero otherwise. The observed curve $U_x^{\mathrm{obs}}(n)$ averages this score over enumerated or seeded sampled subsets. Our terminal analysis uses $n = M$.

3.2 Categorical reference and matched budget

Let $\widehat{\boldsymbol{\pi}}_x$ be the empirical distribution over the gold class and every observed wrong-answer class. Independent draws from this full categorical law define a reference plurality curve

$$U_x^{\mathrm{ref}}(m) = \mathbb{E}_{Z_{1:m} \overset{\text{i.i.d.}}{\sim} \widehat{\boldsymbol{\pi}}_x} \left[s_{\mathrm{plurality}}(Z_{1:m}, g_x) \right]. \tag{1}$$

We estimate Eq. (1) by seeded Monte Carlo and retain both its raw curve and a nondecreasing isotonic fit. Given an observed utility u, the matched budget is

$$\text{VOTE-ESS}_x(n; u) = \max\left(\{1\} \cup \{m \in \{1, \ldots, n\} : U_x^{\mathrm{ref}}(m) \le u\} \right). \tag{2}$$

We set $u = U_x^{\mathrm{obs}}(n)$ when the argument is omitted. Thus $\text{VOTE-ESS}(8) = 5$ selects five as the largest tested budget whose reference utility does not exceed the observed utility; it need not be an exact equality. The floor returns one when no budget qualifies, and the cap $m \le n$ restricts the index to shrinkage. Plateaus need not identify a unique equivalent budget, and a nonmonotone raw curve need not admit an ordinary inverse. Our primary empirical quantity is therefore

$$\Delta_x(M) = U_x^{\mathrm{ref}}(M) - U_x^{\mathrm{obs}}(M), \tag{3}$$

where positive values indicate an observed utility deficit. We average first within repeated appearances of a question and bootstrap unique questions.

The plug-in reference reuses the terminal trajectory to estimate $\widehat{\boldsymbol{\pi}}_x$. It is useful for broad diagnostics but can mask estimation error. Our exact-condition cross-fit removes the target trajectory and estimates the reference only from other runs matching question, provider, model, prompt, temperature, top-p, completion cap, thinking setting, and a compatible collection period. Different run artifacts must have disjoint response IDs and recorded seed values. The permitted collection span is 168 hours (35.47 hours at most among eligible conditions); this excludes reuse of the target record, not every possible shared serving effect.

Table 1: Evidence hierarchy. Counts at different levels are not pooled into a single provider-wide estimate.

Evidence level	Collected scale	Inferential unit	Role and limitation
Synthetic grid	180 configurations	configuration	Known dependence and exact controls
Multi-mechanism stress test	5,760 cell–seed records	cell (288)	Point recovery, intervals, and control
Qwen disjoint test	2,052 trajectories	question (710)	Frozen plug-in measurement; clustered CI
Qwen exact cross-fit	752 trajectories	question (160)	Reference excludes target; eligible subset
69-alias breadth	one trajectory/condition	alias (69)	Scope only; no exact-condition cross-fit
DeepSeek full mirror	158,400 calls	question within study/model	Two models × seven studies; provider check

3.3 Curve processing and inverse uncertainty

Raw, isotonic, and cumulative-max curves have different roles. The raw reference retains nonmonotonicity; a least-squares nondecreasing isotonic fit stabilizes the inverse (Barlow et al., 1972). This is a reporting convention, not evidence that the population curve is nondecreasing. The cumulative maximum is only a controller envelope that prevents a later budget from lowering its running utility estimate. Neither transformation changes the direct gap in Eq. (3), which is computed from the unsmoothed reference.

The stress test uses a distinct, continuous inverse: each isotonic plateau is assigned its mean budget, adjacent distinct utilities are linearly interpolated, and the result is clipped to $[1, M]$. Its *utility-error-compatible interval* maps $\widehat{U} \pm 1.96\,\mathrm{SE}(\widehat{U})$ through this inverse. This propagates held-out utility error, but not reference-curve Monte Carlo error; it is not a calibrated 95% effective-budget confidence interval. The 90% coverage criterion is a prespecified minimum in the block-control evaluation, not the interval's nominal confidence level. Appendix J gives the complete implementation.

3.4 Prediction and control as a consequential test

We fit effective-budget predictors using only statistics visible in the answer prefix: prefix length, top-answer shares, number of answer classes, entropy, and plurality margins. Full-trajectory correctness is a label, never an online feature. The matched controllers share the same risk proxy, candidate stopping points, next-answer model, and per-call penalty; one substitutes the predicted effective budget for raw prefix length. We report paired accuracy, calls, tokens, and the pre-existing slice-specific objective accuracy $-\,\lambda$ calls. A smaller call count alone is not a controller improvement if accuracy or the matched objective deteriorates.

4 Evaluation design

Table 1 separates the six evidence levels. A trajectory is one question under one frozen API condition; repeated calls within it are technical observations, not independent inferential units. All random subsampling, Monte Carlo references, and bootstraps use recorded seeds. Unless stated otherwise, inferential intervals are 95% percentile bootstrap intervals, and we make no claim from an interval that includes zero. The utility-error-compatible inverse interval in Section 3.3 is a separate diagnostic, not an inferential interval of the same type.

Known-dependence validation. We simulate exchangeable categorical draws from a Dirichlet–multinomial model (Mosimann, 1962) over correctness probability $p \in \{0.55, 0.70, 0.85\}$, wrong-answer categories $K \in \{2, 5, 20\}$, budget $M \in \{4, 8, 16, 32\}$, and intraclass correlation $\rho \in \{0, .05, .10, .25, .50\}$. The 180 configurations use 20,000 observed replicates and 65,536 reference replicates. A second construction repeats m independent categorical votes M/m times, so its final plurality is exactly that of m independent votes.

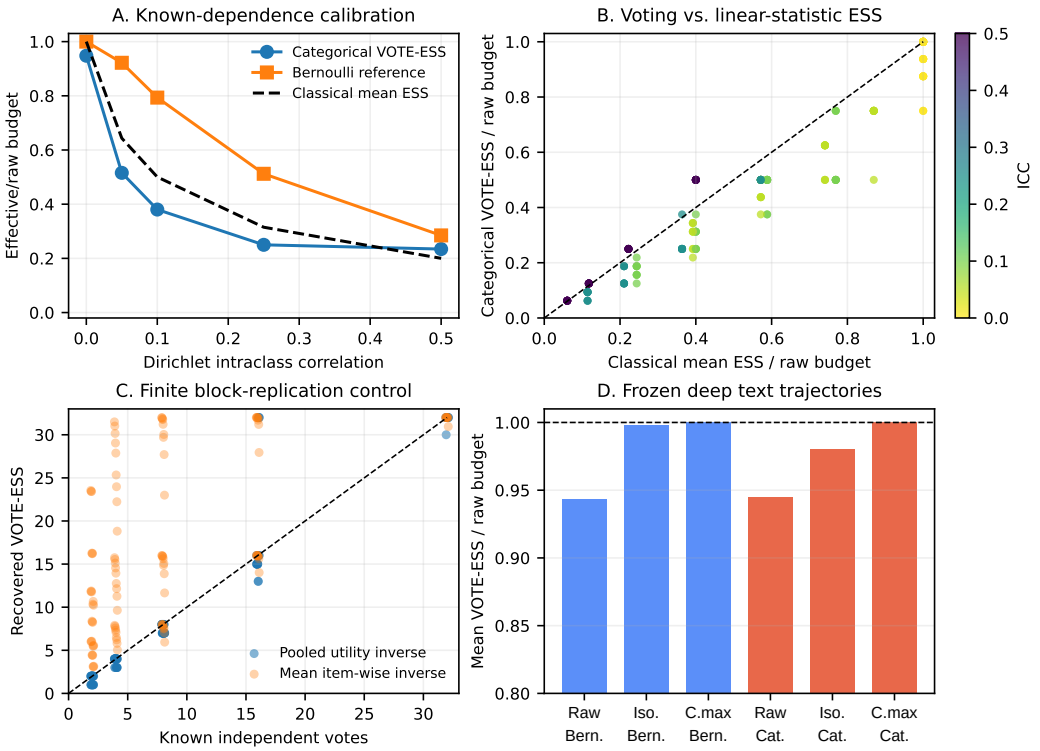

Figure 2: Categorical VOTE-ESS validation: (A) dependence calibration; (B) voting versus linear-statistic ESS; (C) finite block recovery; (D) frozen deep trajectories. Iso. denotes isotonic; C.max denotes the cumulative maximum, a controller envelope only. Bern. and Cat. distinguish Bernoulli and categorical references. Details and uncertainty analyses are in Appendix C.

Hosted trajectories. The Qwen branch covers exact-verifier questions from GSM8K (Cobbe et al., 2021), the MATH-500 subset of MATH (Hendrycks et al., 2021b; Lightman et al., 2024), MMLU-STEM (Hendrycks et al., 2021a), and ARC-Challenge (Clark et al., 2018), multiple prompt and decoding settings, and horizons through $M = 64$. The breadth panel spans 69 text, coding, and multimodal aliases under a shared answer-only contract, plus a nine-alias answer-only/short-CoT overlap. The independent-provider check uses the official DeepSeek API with two model variants and seven frozen study keys per model. Provider/model IDs, response IDs, usage counters, and contract fields are retained. Appendix K gives the full provenance and integrity checks.

5 Results

5.1 The categorical index responds to dependence

Figure 2 and Table 2 show three validation results. First, categorical VOTE-ESS/raw budget decreases monotonically in ρ for 100% of fixed (p, K, M) groups; its Spearman association with ρ is -0.761 ($p = 2.95 \times 10^{-35}$). Second, at $\rho = 0$, all 95% compatible-budget intervals contain the nominal budget. Third, the exact block control has 0.593-vote mean absolute error and 96.3% exact recovery, compared with 4.898 votes and 38.9% for a one-vs-rest Bernoulli reference. With independently simulated finite trajectories, pooled utility matching recovers the block budget with 0.907-vote MAE; averaging itemwise inversions instead gives 8.653 votes. The direct pooled estimand is therefore essential.

Table 2: Mean effective/raw-budget ratios on the synthetic grid. The final column is a comparison scale for a linear mean, not a plurality target.

ρ	Categorical ratio	Bernoulli ratio	Linear ESS ratio
0.00	0.948	1.000	1.000
0.05	0.516	0.922	0.643
0.10	0.380	0.793	0.500
0.25	0.250	0.512	0.315
0.50	0.234	0.285	0.200

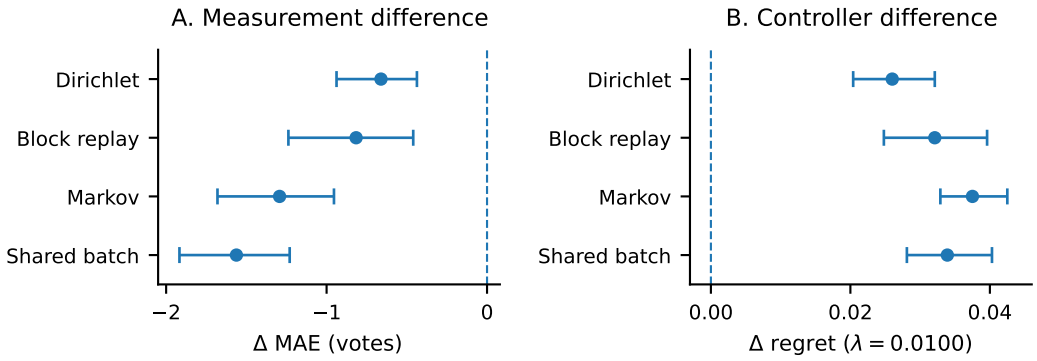

Figure 3: Measurement gains do not imply control gains. Points and 95% cell-bootstrap intervals are the paired contrasts in Table 4, including independence. (A) Categorical minus collision-ESS absolute error; (B) categorical minus raw-prefix controller regret. Negative favors categorical in both panels; the axes have different units. Full comparator and strength-level plots are retained in Appendix J.

5.2 Measurement improves across mechanisms, but stopping control does not

We next crossed three gold probabilities, two wrong-class counts, three budgets, four dependence mechanisms, and four strength levels, with 20 macro-seeds in each of 288 design cells. The mechanisms are a trajectory-specific Dirichlet law, balanced replication of independent vote prototypes, Markov copy/redraw, and a shared-batch answer mixture. Strength levels are ordered within a mechanism, not numerically interchangeable across mechanisms (Appendix J). Oracle references, categorical references, train, validation, and held-out test banks use disjoint seed namespaces. All 5,760 cell–seed records pass the parameter, Monte Carlo, provenance, and completion checks.

Point recovery uses the interpolated isotonic diagnostic in Section 3.3, not the discrete max-inverse in Eq. (2). Its reference bank is separate from the held-out trajectories used to estimate observed utility.

Across non-i.i.d. cells, collision ESS is the strongest prespecified output-only proxy with an equal-cell MAE of 1.206 votes. Interpolated isotonic categorical VOTE-ESS has MAE 0.163, an 86.4% reduction; the paired cell-bootstrap interval for its absolute-error difference is $[-1.225, -0.878]$ votes. Its budget ratio also tracks the known target pair-dependence parameter (Spearman -0.913 across cells). Across this mechanism set, the categorical point estimate is closer to the known target than the output-only proxies.

The uncertainty and control results are negative. Exact block replay has low point-estimate MAE (0.227 vote) but only 83.5% utility-error-compatible interval coverage, below the prespecified minimum of 90%; coverage falls to 76.3% at budget 32 (Table 24). At $\lambda = 0.010$, the categorical controller has mean regret 0.1472, compared with 0.1148 for the strongest raw-prefix controller. The paired difference is $+0.0324$ with 95% cell-bootstrap interval $[+0.0293, +0.0355]$. Dependence-rich regret reductions are negative for all four mechanisms (-15.5%, -27.4%, -22.5%, and -27.2%). Thus improved measurement does not yield a better stopping policy in this experiment.

Table 3: Prespecified evaluation criteria and outcomes. Each criterion is evaluated separately.

Check	Result	Criterion met?
Exact-block MAE	0.227 vote	yes (≤ 1.0)
Exact-block compatible coverage	83.5%	no ($\geq 90\%$)
Interpolated categorical MAE reduction	86.4%	yes ($\geq 15\%$)
Controller regret disadvantage	$+0.0324$	no ($\leq +0.002$)
Mechanisms with $\geq 10\%$ regret reduction	0 of 4	no (≥ 3)

Table 4: Mechanism-level measurement and control over all four strength levels, including independence. MAEs are in votes. Differences are categorical minus collision ESS for measurement, and categorical minus the globally selected raw-prefix controller at $\lambda = 0.010$ for regret; negative favors categorical. Intervals are 95% cell-bootstrap intervals.

Mechanism	MAE Cat. / collision	Δ MAE (95% CI)	Δ regret (95% CI)
Dirichlet–multinomial	0.143 / 0.804	-0.661 $[-0.938, -0.437]$	$+0.0260$ $[+0.0204, +0.0321]$
Exact block replay	0.227 / 1.043	-0.816 $[-1.238, -0.460]$	$+0.0321$ $[+0.0248, +0.0396]$
Markov copy/redraw	0.245 / 1.538	-1.293 $[-1.680, -0.954]$	$+0.0375$ $[+0.0329, +0.0425]$
Shared batch mixture	0.203 / 1.764	-1.562 $[-1.917, -1.230]$	$+0.0339$ $[+0.0281, +0.0403]$

What the controller comparison tests. In this simulator, all controllers use the same prefix risk proxy,

$$R_n(e_n) = \Phi\left(-\frac{\sqrt{e_n}(p_1 - p_2)}{\sqrt{p_1(1-p_1) + p_2(1-p_2)}}\right), \tag{4}$$

where Φ is the standard normal cumulative distribution function, p_1, p_2 are the two largest smoothed class shares, and e_n is the raw or effective prefix length. Smoothing adds α to each observed class and one unseen-class branch. The rule stops when the predictive one-step risk reduction is at most λ; $\alpha \in \{0.5, 1, 2\}$ is selected by validation objective, with smaller α breaking ties. Held-out regret uses a per-trajectory hindsight oracle only as a nondeployable lower bound (Appendix J). This tests effective-budget substitution in a matched rule, not every possible dependence-aware policy.

Figure 3 and Table 4 show lower point-recovery error but higher regret in every mechanism. These rows include independence; the 86.4% headline measurement reduction excludes independent cells.

5.3 Deep Qwen trajectories do not establish an aggregate utility deficit

Across 2,052 disjoint test trajectories and 710 questions, categorical raw, isotonic, and controller-cumulative-max terminal ratios are 0.944/0.981/1.000. Before inversion or smoothing, however, the aggregate reference-minus-observed plurality gap is -0.0063 with 95% CI $[-0.0101, -0.0024]$: the observed utility is slightly higher, not lower, than the plug-in independent reference. None of 13 slice intervals establishes a positive deficit. The inverse ratio and direct utility therefore tell a compatible practical story—no detected aggregate deficit under the stated reference—while also illustrating why the ratio alone should not be read as lost accuracy. This is not an equivalence test or a proof of independence.

The stricter cross-fit retains 752 test trajectories over 160 questions whose conditions have other separately collected runs. Its categorical-reference minus observed terminal gap is $+0.0011$ $[-0.0075, +0.0103]$, again not detectably positive. Its raw/isotonic ratios are 0.8818/0.8850, lower than the plug-in ratios despite the near-zero direct gap (Appendix D.3). The eligible subset also differs, so this is not an isolated effect of changing the reference estimator. One of seven run-level intervals is positive: the MATH-500 $M = 64$ short-CoT slice, $+0.0920$ $[+0.0195, +0.1825]$. Because this

Table 5: Reference-minus-observed plurality-utility gaps at the terminal budget. Positive values indicate an observed utility deficit. The positive MATH-500 slice is retained as a localized exception to the aggregate result.

Evidence set	Unit	Gap	95% CI
Qwen disjoint plug-in	710 questions	-0.0063	$[-0.0101, -0.0024]$
Qwen exact-condition cross-fit	160 questions	$+0.0011$	$[-0.0075, +0.0103]$
MATH-500, $M = 64$ cross-fit	40 questions	$+0.0920$	$[+0.0195, +0.1825]$
Answer-only breadth	69 aliases	-0.0082	$[-0.0128, -0.0041]$
DeepSeek matched panel	100 questions	-0.0268	$[-0.0489, -0.0078]$

Table 6: Unique-question clustered controller comparisons. Negative call and token differences mean that the learned controller stops earlier.

Provider	Learned minus raw-prefix	Estimate	95% CI
Qwen	Accuracy	-0.0014	$[-0.0043, +0.0000]$
	Calls	-0.2507	$[-0.2877, -0.2129]$
	Tokens	-52.63	$[-62.22, -42.73]$
	Matched objective	$+0.0011$	$[-0.0019, +0.0027]$
DeepSeek	Accuracy	-0.0018	$[-0.0043, +0.0001]$
	Calls	-0.216	$[-0.242, -0.190]$
	Matched objective	$+0.0002$	$[-0.0023, +0.0020]$

is one localized subgroup and the aggregate interval includes zero, we report it as a boundary case rather than a global effect.

Table 5 keeps the direct, non-inverted estimand visible across the main evidence levels. The estimates are not pooled because their reference construction and inferential units differ.

5.4 Breadth and provider portability bound the scope

The 69-alias answer-only panel has a hierarchical direct deficit of -0.0082 $[-0.0128, -0.0041]$. Text (-0.0073), multimodal (-0.0064), and coding (-0.0170) group estimates are all nonpositive. Three aliases have unadjusted positive intervals, but none survives Holm correction (Holm, 1979) across 69 comparisons. Because each exact condition has only one trajectory, these are plug-in scope checks rather than cross-fitted estimates.

Prompt wording is not invariant. On nine overlapping aliases, short-CoT minus answer-only changes the direct deficit by $+0.0056$ $[-0.0083, +0.0217]$, parse rate by $+3.63$ points $[+0.97, +7.19]$, and one-call accuracy by $+15.0$ points $[+7.12, +23.07]$. The evidence-budget conclusion is stable on this shard, but quality and parsability are not.

The matched 100-question Qwen/DeepSeek panel (Appendix F, Figure 4) tests portability. Raw categorical ratios are 0.956 and 0.974; their paired DeepSeek-minus-Qwen difference is $+0.0175$ $[-0.0262, +0.0612]$. The larger full mirror contains 14 model–study rows. Raw ratios range from 0.893 to 0.994 and isotonic ratios from 0.957 to 1.000; none of 14 paired isotonic intervals separates DeepSeek from Qwen. These data show protocol portability to a second provider, not a provider-independent law.

5.5 Measured shrinkage does not improve matched control

Table 6 reports the consequential test. On Qwen, averaging repeated appearances within 710 unique questions, the learned controller saves 0.251 calls and 52.6 tokens but changes accuracy by -0.0014 and the matched objective by $+0.0011$ with an interval spanning zero. The DeepSeek result is similar: 0.216 fewer calls, -0.0018 accuracy, and no detectable objective change. Thus prefix-visible effective-budget prediction moves the stopping point but does not establish better aggregate decisions than raw prefix length.

6 Discussion and limitations

The experiments yield three findings. The categorical matched-budget index responds to known dependence and is much more accurate than output-only proxies across four simulated mechanisms. Accurate point measurement does not guarantee calibrated inverse intervals: exact-block coverage fails. Neither the simulated stress test nor the hosted trajectories establish a better stopping controller. The positive measurement result thus coexists with negative uncertainty and decision results.

Replacing raw call count with an effective budget requires a positive direct gap that is stable under held-out reference estimation and a resulting prefix-only controller that improves a matched accuracy–cost objective. The aggregate tests do not satisfy these conditions, so they do not justify the tested replacement in these closed-answer settings; this does not establish that raw count is optimal. The MATH-500 $M = 64$ exception is a boundary to retest, not a reason to change the aggregate policy.

Prediction and decision are separate tests. On Qwen, prefix-only categorical prediction beats raw prefix length by held-out R^2 in 4/11 runs but by MAE in only 1/11; on DeepSeek the corresponding counts are 10/14 and 2/14 (Appendix G, Table 20). These metric-dependent, slice-level comparisons are not additional independent replications and do not establish the accuracy–cost improvement required by the controller test. Both prediction and decision tests are therefore needed to assess a candidate compute unit; terminal measurement alone does not establish policy value.

Answer-equivalence sensitivity. In the seven-alias translation follow-up, 589 of 1,408 responses were accepted variants that differed from the first gold string. Replacing single-string correctness with the task's accepted-answer classes changes the alias-mean raw matched budget from 4.7375 to 7.9375 at $M = 8$ (Table 21). Residual raw-inverse deviations remain, so the cumulative-max value of 8.0000 is not evidence of independence. This result identifies an answer-contract dependence of the apparent redundancy signal, not a change in the underlying hosted responses.

Inverse-boundary diagnostics. Across terminal stress-test records, 11.61% of categorical estimates are clipped at 1 or M, 22.22% of compatible intervals touch a boundary, and 15.69% of adjacent isotonic-curve steps are flat on average (Appendix J). The nominally independent $M = 32$ block condition covers only 45.00% of macro-seed records (Table 24). These diagnostics accompany, but do not by themselves identify the cause of, the failed exact-block coverage criterion. The intervals propagate held-out utility error but omit reference-curve Monte Carlo uncertainty. They should therefore remain diagnostic rather than serve as calibrated guarantees for deciding how many calls to purchase.

The provider study covers Qwen/DashScope and DeepSeek endpoints, not all hosted services. Exact-condition cross-fitting requires repeated frozen runs; breadth results reuse target trajectories to estimate their categorical laws. Canonicalization can merge or split semantically related answers, so answer contracts and accepted gold classes must be declared before analysis. Plurality utility suits scorable closed-form answers, not open-ended generation, where semantic diversity, factuality, and preference quality require different observables. In the stress test, compatible intervals under-cover exact block replay, especially at budget 32, and inversion boundaries remain common near nominal independence. Hosted endpoints can change after collection: recorded provider/model identifiers, collection windows, response IDs, and rerun contracts cannot guarantee that future calls follow the same hidden implementation.

7 Conclusion

AURA expresses repeated hosted-model calls as an output-only, plurality-matched evidence budget. Categorical VOTE-ESS responds to controlled dependence and reduces point-recovery error across four dependence mechanisms, but its interval undercoverage keeps the inverse secondary to direct utility. Direct gaps, cross-fitting, and clustered inference prevent that inverse from carrying more meaning than the data support. Across the evaluated Qwen and DeepSeek trajectories, the main result is not a new stopping win: measurable effective-budget shrinkage does not translate into a detectable aggregate improvement over raw prefix length; in the broader simulator, the categorical controller is decisively worse. This negative result is useful because it clarifies when a more faithful compute unit changes a description without yet changing a decision.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on the research methodology and experimental protocols, assist with implementing simulation and analysis methods, and support interpretation of the results. We also used these tools to identify and organize relevant literature, format references, prepare supporting code, tables, and scientific figures, and draft and revise the manuscript.

Separately, hosted foundation models generated the repeated-call response datasets evaluated in this study, including the translation outputs. The models, request conditions, and recorded responses are described in Section 4 and Appendices F, H, and K. These responses are experimental observations. The synthetic validation trajectories were instead produced by the seeded statistical simulators in Appendices C and J, with AI assistance in implementing the simulators.

We have reviewed all AI-assisted work. We checked the simulation and analysis code, verified the numerical summaries against the frozen datasets and API records, and checked the cited sources and the support for the reported claims, including the controller and interval-calibration results. We take responsibility for the final content of this work, including text, claims, code, analyses, and research artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

An anonymous artifact repository accompanies this submission: https://anonymous.4open.science/r/aura-artifact-AD85/.

Sections 3 and 4 define the measurement object, reference construction, evaluation units, and evidence hierarchy. Appendix B specifies answer canonicalization, reference-curve processing, exact-condition cross-fitting, and uncertainty calculations. Appendices C and J document the simulation designs, random seeds, reference banks, and evaluation criteria. Appendices D, E, and F report the hosted-model panels and their collection conditions, while Appendices G and H specify prefix-only features, stopping-rule evaluation, and answer-equivalence and information-timing checks. Appendix K, including Tables 25 and 26, records frozen analysis metadata, result-table hashes, and collection-integrity checks. As discussed in Appendix I, reproducing analyses of frozen outputs is distinct from repeating live API calls: a future collection constitutes a new time-stamped replication unless the provider identity and full request contract can be shown equivalent.

ETHICS STATEMENT

A potential misuse of AURA is to interpret repeated answer agreement or a near-nominal VOTE-ESS ratio as independently corroborated truth. The index matches plurality utility relative to a declared categorical reference; it does not certify factual correctness, safety, or fairness. Answer-equivalence rules also affect which valid responses are counted, as illustrated by the translation sensitivity analysis in Appendix H. We therefore distinguish agreement from correctness and retain the direct utility gaps, uncertainty limitations, and negative stopping results rather than presenting call savings as established improvements in reliability. Provider comparisons are limited to the recorded endpoints, request conditions, and collection periods; they do not identify hidden serving mechanisms or establish universal provider behavior. The findings should not be used as a stand-alone basis for high-stakes deployment or extrapolated to open-ended generation without task-specific validation. Relevant scope limitations are discussed in Section 6 and Appendix I.

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, 2023. doi: 10.18653/v1/2023.emnlp-main.761. URL https://aclanthology.org/2023.emnlp-main.761/.

Alibaba Cloud. Synchronous API for text embedding. Alibaba Cloud Model Studio documentation, 2026a. URL https://www.alibabacloud.com/help/en/model-studio/text-embedding-synchronous-api. Living documentation; accessed 2026-09-25.

Alibaba Cloud. Non-real-time speech recognition (Qwen-ASR) API reference. Alibaba Cloud Model Studio documentation, 2026b. URL https://www.alibabacloud.com/help/en/model-studio/qwen-asr-api-reference. Living documentation; accessed 2026-09-25.

Alibaba Cloud. Real-time speech recognition: User guide. Alibaba Cloud Model Studio documentation, 2026c. URL https://www.alibabacloud.com/help/en/model-studio/real-time-speech-recognition-user-guide. Living documentation; accessed 2026-09-25.

Alibaba Cloud. Qwen-Image API reference. Alibaba Cloud Model Studio documentation, 2026d. URL https://www.alibabacloud.com/help/en/model-studio/qwen-image-api. Living documentation; accessed 2026-09-25.

Alibaba Cloud. Qwen-Image Edit API reference. Alibaba Cloud Model Studio documentation, 2026e. URL https://www.alibabacloud.com/help/en/model-studio/qwen-image-edit-api. Living documentation; accessed 2026-09-25.

Alibaba Cloud. Audio and video translation: Qwen3-LiveTranslate-Flash API reference. Alibaba Cloud Model Studio documentation, 2026f. URL https://www.alibabacloud.com/help/en/model-studio/qwen3-livetranslate-flash-api. Living documentation; accessed 2026-09-25.

Alibaba Cloud. Real-time audio and video translation: Qwen3.5-LiveTranslate-Flash-Realtime. Alibaba Cloud Model Studio documentation, 2026g. URL https://www.alibabacloud.com/help/en/model-studio/qwen3-5-livetranslate-flash-realtime. Living documentation; accessed 2026-09-25.

Alibaba Cloud. Machine translation (Qwen-MT). Alibaba Cloud Model Studio documentation, 2026h. URL https://www.alibabacloud.com/help/en/model-studio/machine-translation. Living documentation; accessed 2026-09-25.

Alibaba Cloud. Text extraction (Qwen-OCR). Alibaba Cloud Model Studio documentation, 2026i. URL https://www.alibabacloud.com/help/en/model-studio/qwen-vl-ocr. Living documentation; accessed 2026-09-25.

Alibaba Cloud. Qwen-Omni. Alibaba Cloud Model Studio documentation, 2026j. URL https://www.alibabacloud.com/help/en/model-studio/qwen-omni. Living documentation; accessed 2026-09-25.

Alibaba Cloud. Qwen-Omni-Realtime. Alibaba Cloud Model Studio documentation, 2026k. URL https://www.alibabacloud.com/help/en/model-studio/realtime. Living documentation; accessed 2026-09-25.

Alibaba Cloud. OpenAI compatible – chat. Alibaba Cloud Model Studio documentation, 2026l. URL https://www.alibabacloud.com/help/en/model-studio/compatibility-of-openai-with-dashscope. Living documentation; accessed 2026-09-25.

Alibaba Cloud. Qwen-TTS speech synthesis API reference. Alibaba Cloud Model Studio documentation, 2026m. URL https://www.alibabacloud.com/help/en/model-studio/qwen-tts-api. Living documentation; accessed 2026-09-25.

Alibaba Cloud. Real-time speech synthesis: User guide. Alibaba Cloud Model Studio documentation, 2026n. URL https://www.alibabacloud.com/help/en/model-studio/realtime-tts-user-guide. Living documentation; accessed 2026-09-25.

Alibaba Cloud. Voice cloning: User guide. Alibaba Cloud Model Studio documentation, 2026o. URL https://www.alibabacloud.com/help/en/model-studio/voice-cloning-user-guide. Living documentation; accessed 2026-09-25.

Alibaba Cloud. Voice design: User guide. Alibaba Cloud Model Studio documentation, 2026p. URL https://www.alibabacloud.com/help/en/model-studio/voice-design-user-guide. Living documentation; accessed 2026-09-25.

Alibaba Cloud. Tingwu: Getting started. Alibaba Cloud documentation, 2026q. URL https://help.aliyun.com/zh/tingwu/getting-started-1. Chinese-language documentation; descriptive English title; accessed 2026-09-25.

Alibaba Cloud. Wan-Image generation and editing 2.7 API reference. Alibaba Cloud Model Studio documentation, 2026r. URL https://www.alibabacloud.com/help/en/model-studio/wan-image-generation-and-editing-api-reference. Living documentation; accessed 2026-09-25.

Alibaba Cloud. Z-Image API reference. Alibaba Cloud Model Studio documentation, 2026s. URL https://www.alibabacloud.com/help/en/model-studio/z-image-api-reference. Living documentation; accessed 2026-09-25.

Berk Atıl, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. Non-determinism of "deterministic" LLM system settings in hosted environments. In Mousumi Akter, Tahiya Chowdhury, Steffen Eger, Christoph Leiter, Juri Opitz, and Erion Çano (eds.), *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems*, pp. 135–148, Mumbai, India, December 2025. Association for Computational Linguistics. ISBN 979-8-89176-305-0. doi: 10.18653/v1/2025.eval4nlp-1.12. URL https://aclanthology.org/2025.eval4nlp-1.12/.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. arXiv preprint arXiv:2511.21631, 2025. URL https://arxiv.org/abs/2511.21631.

R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, 1972. ISBN 9780471049708.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. arXiv preprint arXiv:2407.21787, 2024. URL https://arxiv.org/abs/2407.21787.

Will Cai, Tianneng Shi, Xuandong Zhao, and Dawn Song. Are you getting what you pay for? auditing model substitution in LLM APIs. In *NeurIPS 2025 Workshop on Regulatable ML*, 2025. URL https://openreview.net/forum?id=thhrtv9P0s.

Ruisheng Cao, Mouxiang Chen, Jiawei Chen, Zeyu Cui, Yunlong Feng, Binyuan Hui, Yuheng Jing, Kaixin Li, Mingze Li, Junyang Lin, Zeyao Ma, Kashun Shum, Xuwu Wang, Jinxi Wei, Jiaxi Yang, Jiajun Zhang, Lei Zhang, Zongmeng Zhang, Wenting Zhao, and Fan Zhou. Qwen3-Coder-Next technical report. arXiv preprint arXiv:2603.00729, 2026. URL https://arxiv.org/abs/2603.00729.

Lingjiao Chen, Jared Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more LLM calls all you need? towards the scaling properties of compound AI systems. In *Advances in Neural Information Processing Systems*, volume 37, pp. 45767–45790, 2024a. doi: 10.52202/079017-1455. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/51173cf34c5faac9796a47dc2fdd3a71-Abstract-Conference.html.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language models. In *ICML 2024 Workshop on In-Context Learning*, 2024b. URL https://openreview.net/forum?id=LjsjHF7nAN.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388. PMLR, 2024. URL https://proceedings.mlr.press/v235/chiang24b.html.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. arXiv preprint arXiv:1803.05457, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

DeepSeek. DeepSeek-V3.2 release. Official DeepSeek release documentation, 2025. URL https://api-docs.deepseek.com/news/news251201. Released December 1, 2025.

DeepSeek. Your first API call. DeepSeek API documentation, 2026. URL https://api-docs.deepseek.com/. Living documentation; accessed 2026-09-25.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pp. 569–593. Springer, 1992.

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

Austin Feng, Marius Alonso, Ambroise Odonnat, Vasilii Feofanov, and Ievgen Redko. Optimal self-consistency for efficient reasoning with large language models. In *NeurIPS 2025 Workshop on Efficient Reasoning*, 2025. URL https://openreview.net/forum?id=vVvcoF8Rz3.

Robert J Gallo, Michael Baiocchi, Thomas R Savage, and Jonathan H Chen. Establishing best practices in large language model research: an application to repeat prompting. *Journal of the American Medical Informatics Association*, 32(2):386–390, 2025. doi: 10.1093/jamia/ocae294. URL https://doi.org/10.1093/jamia/ocae294.

Irena Gao, Percy Liang, and Carlos Guestrin. Model equality testing: Which model is this API serving? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QCDdI7X3f9.

Charles J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992. doi: 10.1214/ss/1177011137. URL https://projecteuclid.org/journals/statistical-science/volume-7/issue-4/Practical-Markov-Chain-Monte-Carlo/10.1214/ss/1177011137.full.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021b. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. CaTS: Calibrated test-time scaling for efficient LLM reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jrSc4RJXy1.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022. doi: 10.52202/068431-1613. URL https://proceedings.neurips.cc/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ndR8Ytrzhh.

Percy Liang et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW. Featured Certification, Expert Certification, Outstanding Certification.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

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

James E. Mosimann. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49(1–2):65–82, 1962. doi: 10.1093/biomet/49.1-2.65. URL https://academic.oup.com/biomet/article-abstract/49/1-2/65/279479.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2025. URL https://arxiv.org/abs/2412.15115v2. Version 2, January 2025; first submitted in December 2024.

Qwen Team. QVQ-Max: Think with evidence. Official Qwen release blog, 2025a. URL https://qwenlm.github.io/blog/qvq-max-preview/. Accessed 2026-09-25.

Qwen Team. Qwen3-Coder: Agentic coding in the world. Official Qwen release blog, 2025b. URL https://qwenlm.github.io/blog/qwen3-coder/. Accessed 2026-09-25.

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

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Aarohi Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj. Featured Certification.

Hexiang Tan, Fei Sun, Sha Liu, Du Su, Qi Cao, Xin Chen, Jingang Wang, Xunliang Cai, Yuanzhuo Wang, Huawei Shen, and Xueqi Cheng. Too consistent to detect: A study of self-consistent errors in LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4755–4765, Suzhou, China, November 2025. Association for Computational

Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.238. URL https://aclanthology.org/2025.emnlp-main.238/.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20090–20111, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1030. URL https://aclanthology.org/2025.findings-acl.1030/.

Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. Reasoning aware self-consistency: Leveraging reasoning paths for efficient LLM sampling. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3613–3635, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.184. URL https://aclanthology.org/2025.naacl-long.184/.

Weiqin Wang, Yile Wang, and Hui Huang. Ranked voting based self-consistency of large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14410–14426, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.744. URL https://aclanthology.org/2025.findings-acl.744/.

Xinglin Wang, Shaoxiong Feng, Yiwei Li, Peiwen Yuan, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. Make every penny count: Difficulty-adaptive self-consistency for cost-efficient reasoning. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 6919–6932, Albuquerque, New Mexico, April 2025b. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.383. URL https://aclanthology.org/2025.findings-naacl.383/.

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 *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-Omni technical report. arXiv preprint arXiv:2509.17765, 2025. URL https://arxiv.org/abs/2509.17765.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822, 2023a. doi: 10.52202/075280-0517. URL https://proceedings.neurips.cc/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract.html.

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*, 2023b. URL https://openreview.net/forum?id=WE_vluYUL-X.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv preprint arXiv:2506.05176, 2025. URL https://arxiv.org/abs/2506.05176.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

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

Zhi Zhou, Yuhao Tan, Zenan Li, Yuan Yao, Lan-Zhe Guo, Yu-Feng Li, and Xiaoxing Ma. A theoretical study on bridging internal probability and self-consistency for LLM reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-2922. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/7e9afa9a02857bce4515247842471444-Abstract-Conference.html.

Zhi Zhou, Yuhao Tan, Zenan Li, Yuan Yao, Lan-Zhe Guo, Xiaoxing Ma, and Yu-Feng Li. Bridging internal probability and self-consistency for effective and efficient LLM reasoning. arXiv preprint arXiv:2502.00511, 2025b. URL https://arxiv.org/abs/2502.00511.

Xiaoyuan Zhu, Yaowen Ye, Tianyi Qiu, Hanlin Zhu, Sijun Tan, Ajraf Mannan, Jonathan Michala, Raluca Ada Popa, and Willie Neiswanger. Auditing black-box LLM APIs with a rank-based uniformity test. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PSIe9mmF7a.

APPENDIX

A SCOPE OF THE SUPPORTING EVIDENCE

This appendix gives the full support for the empirical scope summarized in Table 7. Sections B–C define and validate the measurement object. Sections D– F give the Qwen, breadth, and DeepSeek results. Sections G and H examine prefix-only prediction, stopping, answer equivalence, and information timing. Section J reports the four-mechanism stress test, and Section K records provenance and reproduction checks. The evidence supports an operational plurality-vote result; it does not establish literal stochastic independence, universal provider behavior, or an open-ended semantic-diversity result.

Table 7: Empirical scope and its closest boundary.

Empirical result	Direct support	Boundary
Categorical VOTE-ESS responds to dependence	Dirichlet grid and exact block control	It is not a latent-correlation estimator
Interpolated isotonic categorical VOTE-ESS improves point recovery across four mechanisms	288-cell held-out stress test	Exact-block compatible intervals under-cover
No aggregate positive utility deficit in deep Qwen test	Direct clustered gap and exact-condition cross-fit	One MATH-500 subgroup is positive
Breadth pattern is mostly nonpositive	69-alias hierarchical summary	No cross-fit for one-trajectory conditions
Protocol transfers to DeepSeek	Matched panel and 14 full-matrix rows	Not a universal cross-provider law
Learned controller stops earlier	Paired unique-question comparisons	No detectable accuracy/objective improvement
Categorical controller has higher simulated regret	Four held-out mechanism families	A measurement gain is not a decision gain

B ESTIMAND AND IMPLEMENTATION DETAILS

B.1 CANONICAL ANSWER CLASSES

Each response is parsed into a canonical answer string under a frozen task-family contract. Parse failures remain an explicit category. All accepted gold variants are collapsed to one gold class before counts are formed. Wrong-answer classes remain separate: replacing them by a single "incorrect" category discards the competition among wrong pluralities and can substantially change the matched budget.

For terminal counts $(c_g, c_1, \ldots, c_K)$ with $\sum_k c_k = M$, the plug-in categorical law is

$$\widehat{\boldsymbol{\pi}} = M^{-1}(c_g, c_1, \ldots, c_K).$$

The reference simulation draws 32,768 trajectories for each cached empirical count pattern. For the synthetic experiments we use 65,536 reference draws. The seed is derived deterministically from the global seed and the count pattern, so identical patterns share an identical reference curve.

B.2 RAW, ISOTONIC, AND CUMULATIVE-MAX CURVES

Section 3.3 gives the main-text distinction; for implementation reference, we report three curve treatments with different roles:

1. **Raw**: no monotonicity assumption; this is the first diagnostic.
2. **Isotonic**: a least-squares nondecreasing fit to the finite curve; this is the primary stabilized diagnostic.

3. **Cumulative maximum**: an online envelope used only inside the controller to prevent a later budget from lowering the running utility estimate. It is not the primary empirical result.

The direct terminal gap in Eq. (3) is computed before either isotonic fitting or cumulative maximization.

B.3 Why the categorical reference is primary

A one-vs-rest Bernoulli majority curve depends only on the gold frequency. This is insufficient for plurality voting. For the same gold probability, mass distributed across 20 wrong classes yields a different decision curve from mass concentrated in two wrong classes. On the synthetic grid, the Bernoulli-minus-categorical effective ratio increases from $+0.187$ at $K = 2$ to $+0.276$ at $K = 20$. We retain the Bernoulli quantity only as a sensitivity comparison.

Table 8: Mean effective/raw-budget ratio by wrong-answer diversity, averaged over p, M, and ρ.

Wrong classes K	Categorical	Bernoulli	Difference
2	0.468	0.655	$+0.187$
5	0.467	0.714	$+0.247$
20	0.463	0.739	$+0.276$

B.4 Exact-condition cross-fitting

For target trajectory r, the reference pool contains only trajectories that match all fields below and come from a different run artifact with disjoint provider response IDs and recorded seed values:

question ID	provider endpoint	model identifier
prompt ID	temperature	top-p
completion cap	thinking setting	collection period.

The maximum allowed collection span is 168 hours; the maximum span among eligible conditions is 35.47 hours. Of 3,985 exact conditions, 540 have more than one run. They provide 1,274 eligible target trajectories over 280 questions and 752 test trajectories over 160 questions. The other 3,445 conditions are excluded because they appear in only one run. The sensitivity analysis therefore trades breadth for a cleaner reference.

B.5 Uncertainty and multiplicity

For aggregate effects we first average repeated run appearances within each unique question, then draw 20,000 bootstrap samples of questions. This uses nonparametric bootstrap resampling (Efron, 1992) at the declared question-level unit rather than resampling calls as independent observations. Synthetic summaries use 5,000 bootstrap replicates. Slice- and alias-level intervals are descriptive; where all 69 breadth aliases are inspected, Holm's step-down correction (Holm, 1979) is applied to the family of positive-deficit tests. We do not promote an uncorrected localized interval to the global conclusion.

C Known-dependence validation

C.1 Dirichlet–multinomial construction

Let the base categorical probabilities be $\boldsymbol{\pi} = (p, (1-p)/K, \ldots, (1-p)/K)$. For each trajectory we sample a probability vector from a Dirichlet distribution centered on $\boldsymbol{\pi}$ and then draw M categorical outputs. The concentration is chosen so that the exchangeable indicator correlation equals the designated ρ. The grid crosses three values of p, three values of K, four budgets, and five dependence levels, for 180 configurations.

Categorical VOTE-ESS is nonincreasing in ρ in all 36 fixed (p, K, M) groups. The mean ratio at $\rho = 0$ is 0.948 rather than exactly one because the inverse curve is discrete and capped at

Table 9: Frozen synthetic validation grid.

Factor	Frozen values
Gold probability p	0.55, 0.70, 0.85
Wrong-answer classes K	2, 5, 20
Terminal budget M	4, 8, 16, 32
Intraclass correlation ρ	0, 0.05, 0.10, 0.25, 0.50
Observed replicates	20,000 per configuration
Reference replicates	65,536 per reference curve
Global seed	20260725

M; nevertheless every 95% compatible-budget interval covers the nominal budget. The classical $M/[1+(M-1)\rho]$ value is plotted only to show how a linear-statistic scale differs from the plurality estimand.

C.2 Exact and finite block controls

For a divisor m of M, draw m independent categorical votes and repeat each vote M/m times. The repeated trajectory has exactly the same plurality winners and tie score as the m original votes. Matching against the full categorical law therefore has a known operational target. Across the frozen configuration set, categorical matching recovers it with 0.593-vote MAE and 96.3% exact recovery. Bernoulli matching has 4.898-vote MAE and 38.9% exact recovery.

An additional finite control independently simulates both observed and reference trajectories. Pooled plurality-utility matching gives 0.907-vote MAE. If each noisy trajectory is first inverted and the inverses are then averaged, MAE rises to 8.653 votes. This noncommutativity motivates the paper's emphasis on the pooled direct gap.

D Deep Qwen results

D.1 Disjoint test summary

The frozen deep test contains 2,052 trajectories from 710 unique questions and 639 cached categorical count patterns. Table 10 separates the diagnostic ratio from the direct outcome.

Table 10: Terminal measurement on the disjoint Qwen test set.

Quantity	Estimate	95% question-bootstrap CI
Categorical raw ratio	0.944	descriptive
Categorical isotonic ratio	0.981	descriptive
Controller cumulative-max ratio	1.000	descriptive
Reference minus observed utility	-0.0063	$[-0.0101, -0.0024]$
Largest positive slice mean	$+0.0025$	includes zero
Positive slice intervals	0/13	–

The negative direct gap means the observed plurality is slightly better than the plug-in independent categorical curve at the terminal budget. We do not interpret this as "negative dependence": the categorical law is estimated from finite terminal counts, the metric is capped, and conditions differ. The supported conclusion is only that the aggregate data do not reveal a positive plurality-utility deficit.

D.2 Slice-complete same-batch record

Table 11 retains all 13 prespecified Qwen slices, including small procedural probes and the broad public-benchmark extensions. The categorical reference is estimated from the target trajectory in this table; the cross-fitted analysis in Appendix D.3 is the stricter sensitivity. Several intervals are

Table 11: Complete same-batch terminal record for the 13 Qwen slices. Direct gap is categorical reference minus observed plurality utility. These rows are plug-in diagnostics; they are not described as held-out references.

ID	Slice	Traj./ questions	Raw ratio (95% CI)	Isotonic ratio (95% CI)	Direct gap (95% CI)
D1	Core pilot	40/40	0.9297 [0.8359, 1.0000]	0.9766 [0.9297, 1.0000]	+0.0025 [−0.0007, +0.0082]
D2	Four-prompt sweep	32/8	0.8242 [0.7070, 0.9121]	0.9414 [0.8535, 1.0000]	−0.0135 [−0.0682, +0.0311]
D3	Trap probe, $T = 0.2$	16/8	1.0000 [1.0000, 1.0000]	1.0000 [1.0000, 1.0000]	−0.0000 [−0.0000, +0.0000]
D4	Trap probe, $T = 0.7$	16/8	1.0000 [1.0000, 1.0000]	1.0000 [1.0000, 1.0000]	−0.0763 [−0.1713, −0.0100]
D5	Compact trap probe	16/8	0.4141 [0.1797, 0.7070]	0.8242 [0.6484, 0.9414]	−0.0296 [−0.1370, +0.0577]
D6	Redundancy stress	192/8	0.8975 [0.8389, 0.9561]	0.9512 [0.9121, 0.9854]	−0.0226 [−0.0493, +0.0017]
D7	Benchmark pilot, $M = 4$	32/32	1.0000 [1.0000, 1.0000]	1.0000 [1.0000, 1.0000]	−0.0051 [−0.0151, +0.0000]
D8	Benchmark core, $M = 16$	660/660	0.9688 [0.9545, 0.9815]	0.9915 [0.9844, 0.9972]	−0.0035 [−0.0071, −0.0004]
D9	Benchmark prompt probe	480/240	0.9512 [0.9277, 0.9727]	0.9883 [0.9785, 0.9961]	−0.0024 [−0.0082, +0.0028]
D10	MATH-500 decoding stress	240/40	0.9180 [0.8555, 0.9688]	0.9583 [0.9232, 0.9870]	−0.0054 [−0.0209, +0.0077]
D11	Broad benchmark, $M = 32$	160/160	0.9939 [0.9818, 1.0000]	0.9939 [0.9818, 1.0000]	−0.0131 [−0.0222, −0.0055]
D12	Hard-task extension, $M = 32$	160/80	0.9092 [0.8547, 0.9576]	0.9709 [0.9430, 0.9939]	+0.0003 [−0.0122, +0.0115]
D13	MATH-500 short-CoT, $M = 64$	40/40	0.9262 [0.8277, 1.0000]	1.0000 [1.0000, 1.0000]	−0.0007 [−0.0051, +0.0039]

Table 12: Exact-condition cross-fitted test results for all seven run families. Intervals resample unique questions within each run. The MATH-500 short-CoT row is the only positive run-level interval; it remains a localized result within an aggregate interval that includes zero.

Run family	Traj./ questions	Raw ratio (95% CI)	Isotonic ratio (95% CI)	Direct gap (95% CI)
Core benchmark	160/160	0.8594 [0.8066, 0.9063]	0.8594 [0.8066, 0.9121]	+0.0119 [−0.0615, +0.0851]
Hard-prompt, $M = 32$	160/80	0.8182 [0.7453, 0.8850]	0.8363 [0.7637, 0.9029]	+0.0261 [−0.0578, +0.1082]
High-budget	160/160	0.8789 [0.8305, 0.9273]	0.8789 [0.8244, 0.9273]	−0.0177 [−0.0928, +0.0583]
MATH-500 decoding stress	40/40	0.6719 [0.5313, 0.8125]	0.6719 [0.5313, 0.8125]	+0.0553 [−0.0255, +0.1399]
MATH-500 short-CoT, $M = 64$	40/40	0.7293 [0.5816, 0.8523]	0.7293 [0.5816, 0.8523]	+0.0920 [+0.0195, +0.1825]
Prompt probe	160/80	0.8652 [0.7891, 0.9297]	0.8711 [0.8008, 0.9355]	−0.0668 [−0.1738, +0.0398]
Smoke family	32/32	0.8125 [0.6953, 0.9297]	0.8125 [0.6953, 0.9063]	+0.0589 [−0.1003, +0.2200]

negative, most include zero, and none is positive. The heterogeneous pattern is more informative than the aggregate ratio alone.

D.3 Cross-fitted run-level result

Aggregating the seven eligible test families yields 752 trajectories over 160 unique questions. The aggregate raw/isotonic/cumulative-max ratios are 0.8818/0.8850/0.8874, and the primary direct gap is $+0.0011$ $[-0.0075, +0.0103]$. Cumulative maximum is reported here only to complete the frozen sensitivity record; it is not used as primary evidence.

The cross-fitted inverse ratios are lower than the plug-in ratios, while the direct gap remains near zero. Finite held-out categorical curves can be flat or weakly nonmonotone, so this discrepancy is expected when inversion is ill-conditioned. It is also why the main text does not convert a ratio difference into an implied accuracy loss.

E Breadth and prompt sensitivity

E.1 Group-level answer-only panel

Table 13: Answer-only breadth results from the frozen 69-alias panel. Ratios are equal-alias means; direct-gap intervals use the hierarchical panel estimator. Each exact condition has one trajectory, so these are plug-in, not cross-fitted, references.

Group	Aliases	Positive CIs	Raw ratio	Isotonic ratio	Direct gap (hierarchical 95% CI)
Text-chat	44	3	0.8082	1.0000	-0.0073 $[-0.0135, -0.0017]$
Multimodal-chat	17	0	0.9449	1.0000	-0.0064 $[-0.0124, -0.0015]$
Coding	8	0	1.0000	1.0000	-0.0170 $[-0.0326, -0.0049]$
Overall	69	3	0.8641	1.0000	-0.0082 $[-0.0128, -0.0041]$

Three alias-level intervals are positive before multiplicity correction; none survives Holm correction across 69 aliases. These local rows are preserved in the released result table rather than discarded, but they do not overturn the hierarchical group estimate.

Table 14: The three unadjusted positive alias-level intervals. The printed lower bound for `qwen3-max` rounds to zero but is positive before rounding. The adjustment is exploratory and does not establish equivalence.

Alias	Direct gap (95% CI)	69-way Holm-adjusted p
`qwen-plus-2025-12-01`	$+0.0147$ $[+0.0004, +0.0358]$	1.0
`qwen-plus-latest`	$+0.0078$ $[+0.0001, +0.0210]$	1.0
`qwen3-max`	$+0.0055$ $[+0.0000, +0.0128]$	1.0

E.2 Prompt-overlap panel

Nine aliases were evaluated with both an answer-only instruction and a short-chain-of-thought instruction while holding the benchmark manifest and terminal budget fixed. The prompt changes parse rate and one-call accuracy materially, even though the difference in terminal direct deficit is uncertain.

Table 15: Paired prompt-overlap summary.

Short-CoT minus answer-only	Estimate	95% CI
Direct utility deficit	$+0.0056$	$[-0.0083, +0.0217]$
Parse rate (percentage points)	$+3.63$	$[+0.97, +7.19]$
One-call accuracy (points)	$+15.0$	$[+7.12, +23.07]$

This shard rules out neither strong prompt effects nor other prompts that create a different evidence regime. It shows only that the answer-only breadth result is not reversed on the tested short-CoT overlap.

E.3 Complete alias-level breadth results

Table 16 reports every alias rather than selecting rows by the sign of the estimate. All rows use the same 40 held-out questions and terminal budget $M = 16$. Raw and isotonic columns are inverse diagnostics; the direct gap and its question-bootstrap interval are the outcome. The family codes are T (text-chat), V (multimodal-chat), and C (coding). The aliases are provider-exposed identifiers recorded at collection time, not claims that every alias is a distinct weight snapshot. For lineage context only, public descriptions cover Qwen2.5 (Qwen et al., 2025), Qwen3 (Yang et al., 2025), Qwen3-Coder (Qwen Team, 2025b), Qwen3-Coder-Next (Cao et al., 2026), Qwen3-Omni (Xu et al., 2025), Qwen3-VL (Bai et al., 2025), and DeepSeek-V3.2 (DeepSeek, 2025). These sources describe model families, not the identity of every hosted alias at collection time.

Table 16: Complete 69-alias answer-only panel. A dagger marks the three unadjusted positive intervals; none survives the exploratory 69-way Holm adjustment.

Alias	Family	Raw ratio	Isotonic ratio	Direct gap (95% CI)
ccai-pro	C	1.0000	1.0000	−0.0156 [−0.0394, −0.0003]
deepseek-v3.2	T	1.0000	1.0000	−0.0133 [−0.0397, −0.0000]
qwen-coder-plus	C	1.0000	1.0000	−0.0285 [−0.0742, 0.0000]
qwen-flash	T	1.0000	1.0000	−0.0147 [−0.0379, 0.0009]
qwen-flash-character	T	0.1250	1.0000	0.0025 [−0.0163, 0.0246]
qwen-max	T	1.0000	1.0000	−0.0252 [−0.0655, 0.0022]
qwen-max-2025-01-25	T	1.0000	1.0000	−0.0092 [−0.0275, 0.0000]
qwen-max-latest	T	1.0000	1.0000	−0.0123 [−0.0331, 0.0009]
qwen-plus	T	0.0625	1.0000	0.0094 [−0.0001, 0.0265]
qwen-plus-2025-01-25	T	1.0000	1.0000	−0.0016 [−0.0226, 0.0158]
qwen-plus-2025-04-28	T	1.0000	1.0000	−0.0100 [−0.0299, −0.0000]
qwen-plus-2025-07-14	T	1.0000	1.0000	−0.0040 [−0.0125, 0.0017]
qwen-plus-2025-09-11	T	0.7500	1.0000	0.0005 [−0.0055, 0.0084]
qwen-plus-2025-12-01†	T	0.0625	1.0000	0.0147 [0.0004, 0.0358]
qwen-plus-character	T	1.0000	1.0000	−0.0124 [−0.0436, 0.0086]
qwen-plus-latest†	T	0.0625	1.0000	0.0078 [0.0001, 0.0210]
qwen-turbo	T	1.0000	1.0000	0.0000 [0.0000, 0.0000]
qwen-turbo-2024-11-01	T	1.0000	1.0000	−0.0457 [−0.1025, −0.0040]
qwen-turbo-2025-04-28	T	1.0000	1.0000	−0.0000 [−0.0000, −0.0000]
qwen-turbo-latest	T	0.0625	1.0000	0.0000 [0.0000, 0.0000]
qwen-vl-max	V	1.0000	1.0000	−0.0035 [−0.0102, −0.0000]
qwen-vl-max-2025-04-08	V	0.0625	1.0000	0.0007 [−0.0009, 0.0027]
qwen-vl-max-2025-08-13	V	1.0000	1.0000	−0.0073 [−0.0218, 0.0000]
qwen-vl-max-latest	V	1.0000	1.0000	−0.0010 [−0.0025, −0.0000]
qwen-vl-plus	V	1.0000	1.0000	−0.0107 [−0.0236, −0.0001]
qwen-vl-plus-latest	V	1.0000	1.0000	−0.0161 [−0.0341, −0.0018]
qwen2.5-14b-instruct	T	1.0000	1.0000	−0.0099 [−0.0280, 0.0004]
qwen2.5-14b-instruct-1m	T	1.0000	1.0000	−0.0263 [−0.0650, −0.0000]
qwen2.5-32b-instruct	T	1.0000	1.0000	−0.0289 [−0.0743, −0.0023]
qwen2.5-72b-instruct	T	1.0000	1.0000	−0.0044 [−0.0101, −0.0001]
qwen2.5-7b-instruct	T	1.0000	1.0000	−0.0052 [−0.0168, 0.0013]
qwen2.5-7b-instruct-1m	T	0.0625	1.0000	0.0185 [−0.0040, 0.0487]
qwen3-235b-a22b-instruct-2507	T	1.0000	1.0000	−0.0077 [−0.0341, 0.0113]
qwen3-235b-a22b-thinking-2507	T	1.0000	1.0000	−0.0113 [−0.0341, 0.0079]
qwen3-30b-a3b-instruct-2507	T	1.0000	1.0000	−0.0078 [−0.0195, 0.0001]
qwen3-30b-a3b-thinking-2507	T	0.2500	1.0000	0.0062 [−0.0007, 0.0171]
qwen3-coder-480b-a35b-instruct	C	1.0000	1.0000	−0.0204 [−0.0484, 0.0005]
qwen3-coder-flash	C	1.0000	1.0000	−0.0046 [−0.0210, 0.0067]
qwen3-coder-next	C	1.0000	1.0000	−0.0072 [−0.0403, 0.0176]
qwen3-coder-plus	C	1.0000	1.0000	−0.0282 [−0.0633, −0.0005]
qwen3-coder-plus-2025-07-22	C	1.0000	1.0000	−0.0165 [−0.0439, 0.0019]
qwen3-coder-plus-2025-09-23	C	1.0000	1.0000	−0.0153 [−0.0358, −0.0007]
qwen3-max†	T	0.0625	1.0000	0.0055 [0.0000, 0.0128]
qwen3-max-2025-09-23	T	0.0625	1.0000	0.0024 [−0.0084, 0.0136]
qwen3-max-2026-01-23	T	1.0000	1.0000	−0.0097 [−0.0342, 0.0079]
qwen3-max-preview	T	1.0000	1.0000	−0.0027 [−0.0080, 0.0000]
qwen3-next-80b-a3b-instruct	T	1.0000	1.0000	−0.0031 [−0.0086, −0.0000]
qwen3-next-80b-a3b-thinking	T	1.0000	1.0000	−0.0110 [−0.0318, −0.0000]
qwen3-omni-flash	V	1.0000	1.0000	−0.0121 [−0.0369, 0.0035]
qwen3-omni-flash-2025-09-15	V	1.0000	1.0000	−0.0022 [−0.0051, −0.0001]
qwen3-omni-flash-2025-12-01	V	1.0000	1.0000	−0.0022 [−0.0105, 0.0040]
qwen3-vl-235b-a22b-instruct	V	1.0000	1.0000	−0.0039 [−0.0304, 0.0232]
qwen3-vl-235b-a22b-thinking	V	1.0000	1.0000	−0.0057 [−0.0134, 0.0004]
qwen3-vl-flash	V	1.0000	1.0000	−0.0130 [−0.0296, −0.0001]
qwen3-vl-flash-2025-10-15	V	1.0000	1.0000	−0.0048 [−0.0182, 0.0079]
qwen3-vl-flash-2026-01-22	V	1.0000	1.0000	−0.0014 [−0.0044, 0.0005]
qwen3-vl-plus	V	1.0000	1.0000	−0.0102 [−0.0420, 0.0153]

Continued on next page

Alias	Family	Raw ratio	Isotonic ratio	Direct gap (95% CI)
qwen3-vl-plus-2025-09-23	V	1.0000	1.0000	-0.0086 $[-0.0211, -0.0001]$
qwen3-vl-plus-2025-12-19	V	1.0000	1.0000	-0.0076 $[-0.0299, 0.0062]$
qwen3.5-122b-a10b	T	1.0000	1.0000	-0.0145 $[-0.0391, 0.0016]$
qwen3.5-27b	T	1.0000	1.0000	-0.0162 $[-0.0449, 0.0000]$
qwen3.5-35b-a3b	T	1.0000	1.0000	-0.0124 $[-0.0390, 0.0089]$
qwen3.5-397b-a17b	T	1.0000	1.0000	-0.0180 $[-0.0416, -0.0001]$
qwen3.5-flash	T	1.0000	1.0000	-0.0018 $[-0.0471, 0.0383]$
qwen3.5-flash-2026-02-23	T	1.0000	1.0000	-0.0173 $[-0.0545, 0.0126]$
qwen3.5-plus	T	1.0000	1.0000	-0.0149 $[-0.0467, 0.0062]$
qwen3.5-plus-2026-02-15	T	1.0000	1.0000	-0.0089 $[-0.0425, 0.0188]$
qwen3.6-plus	T	1.0000	1.0000	-0.0077 $[-0.0458, 0.0239]$
qwen3.6-plus-2026-04-02	T	1.0000	1.0000	-0.0013 $[-0.0269, 0.0228]$

F Independent-provider results

F.1 Matched 100-question panel

The provider-matched panel contains 25 questions each from GSM8K, MATH-500, MMLU-STEM, and ARC-Challenge, with eight calls per question, answer-only prompts, temperature 0.7, and frozen Qwen outputs. The DeepSeek run records 93,672 prompt tokens and 72,152 completion tokens.

Table 17: Matched 100-question provider panel. The paired raw-ratio difference is $+0.0175$ $[-0.0262, +0.0612]$.

Provider	Raw ratio	Isotonic ratio	Cumulative max
Qwen	0.956	0.991	1.000
DeepSeek	0.974	0.983	1.000

DeepSeek's direct reference-minus-observed gap is -0.0268 $[-0.0489, -0.0078]$, with parse rate 0.963 and one-call accuracy 0.782. These values are useful for checking that the analysis pipeline transfers, but a single model panel cannot characterize provider diversity.

F.2 Full two-model matrix

The full mirror covers `deepseek-v4-flash` and `deepseek-v4-pro`, each under seven study keys spanning public benchmarks, prompt contracts, decoding stress, procedural tasks, and horizons $M = 16, 32, 64$. Each model has 79,200 successful calls, 79,200 unique call keys, and 79,200 unique response IDs. There are no empty responses, recorded model mismatches, provider-sent seeds, or thinking-mode mismatches. Length caps occur in 4,793 flash and 4,998 pro responses and remain part of the frozen outputs.

Across the 14 model–study rows, raw categorical ratios range from 0.893 to 0.994, isotonic ratios from 0.957 to 1.000, and cumulative-max ratios equal 1.000 throughout. None of the 14 paired DeepSeek-minus-Qwen isotonic intervals excludes zero. This convergence supports portability of the operational measurement pattern under the matched studies, not equality of the providers' response distributions.

F.3 Provider request and collection integrity

The DeepSeek mirror uses the official API with thinking disabled and no provider seed. Requests retain the Qwen study's prompts, temperatures, top-p values, completion limits, and answer parsers where the APIs expose matching controls. The absence of a provider seed is a recorded contract difference, not an unreported attempt to emulate Qwen's seed schedule. Every logical call stores a call key, provider response ID, returned model string, timestamps, usage counters, finish reason, and length-cap indicator.

These two model names are products of one provider and therefore do not count as two independent providers. Likewise, successful response logging does not show that hidden serving implementations were identical across providers. It shows only that the declared request and output record is complete enough for the bounded portability comparison reported here.

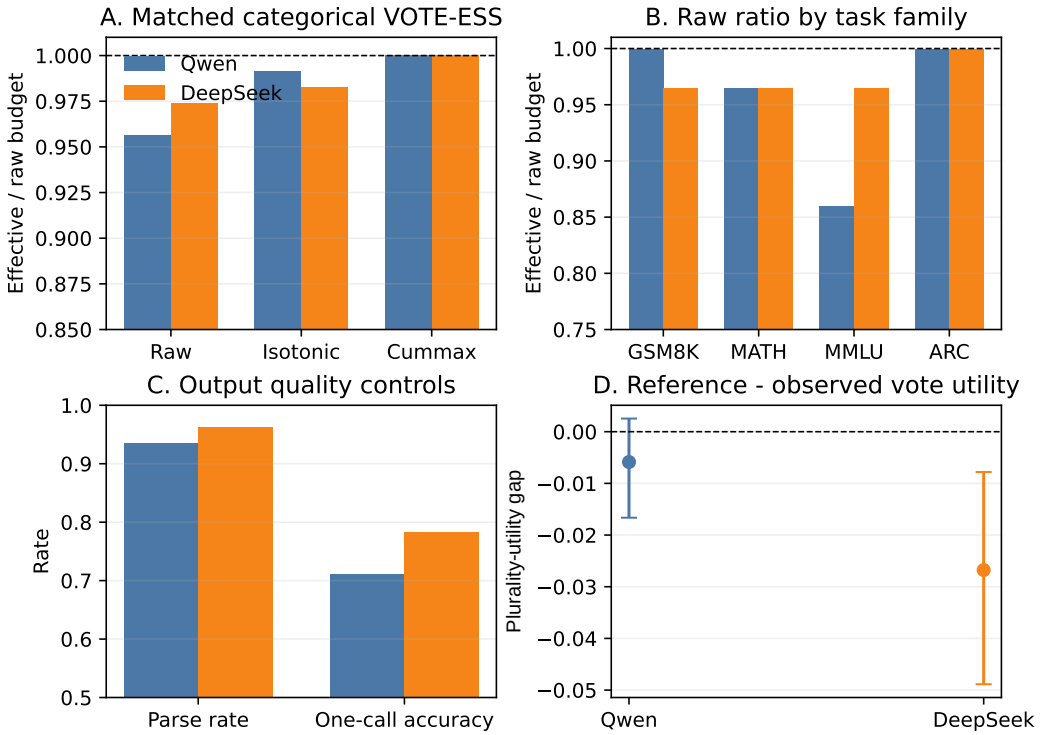

Figure 4: Matched 100-question provider check. (A) Curve-treatment ratios; (B) raw ratios by task; (C) parse rate and one-call accuracy; (D) reference-minus-observed plurality utility with question-bootstrap intervals. Bars in A–C are point summaries, not confidence intervals; C.max/Cummax is the controller envelope. The paired raw-ratio difference is $+0.0175$ $[-0.0262, +0.0612]$ (Table 17), so the data neither establish a provider difference nor prove equivalence.

Table 18: Complete DeepSeek full-matrix test results. Flash denotes `deepseek-v4-flash` and Pro denotes `deepseek-v4-pro`. Direct gap is reference minus observed plurality utility. No row is interpreted as evidence of provider equality.

Study	DeepSeek model	Raw ratio (95% CI)	Isotonic ratio (95% CI)	Direct gap (95% CI)
Core benchmark	Flash	0.9689 $[0.9557, 0.9803]$	0.9922 $[0.9867, 0.9969]$	-0.0078 $[-0.0120, -0.0040]$
	Pro	0.9615 $[0.9472, 0.9749]$	0.9877 $[0.9791, 0.9950]$	-0.0026 $[-0.0073, +0.0018]$
Hard-prompt, $M = 32$	Flash	0.9576 $[0.9152, 0.9879]$	0.9879 $[0.9697, 1.0000]$	$+0.0012$ $[-0.0042, +0.0071]$
	Pro	0.9697 $[0.9273, 1.0000]$	0.9770 $[0.9479, 1.0000]$	$+0.0011$ $[-0.0112, +0.0140]$
High-budget	Flash	0.9939 $[0.9818, 1.0000]$	0.9939 $[0.9818, 1.0000]$	-0.0037 $[-0.0131, +0.0018]$
	Pro	0.9879 $[0.9697, 1.0000]$	0.9939 $[0.9818, 1.0000]$	$+0.0026$ $[-0.0009, +0.0088]$
MATH-500 decoding stress	Flash	0.9331 $[0.8630, 0.9880]$	0.9701 $[0.9388, 0.9935]$	$+0.0039$ $[-0.0080, +0.0177]$
	Pro	0.9531 $[0.9102, 0.9883]$	0.9781 $[0.9539, 0.9966]$	-0.0072 $[-0.0195, +0.0034]$
MATH-500 short-CoT, $M = 64$	Flash	0.9016 $[0.8031, 0.9754]$	0.9754 $[0.9262, 1.0000]$	-0.0019 $[-0.0222, +0.0136]$
	Pro	0.9754 $[0.9262, 1.0000]$	1.0000 $[1.0000, 1.0000]$	-0.0117 $[-0.0342, +0.0001]$
Prompt probe	Flash	0.9805 $[0.9648, 0.9922]$	0.9941 $[0.9863, 1.0000]$	-0.0019 $[-0.0078, +0.0026]$
	Pro	0.9805 $[0.9629, 0.9941]$	0.9885 $[0.9768, 0.9980]$	-0.0018 $[-0.0063, +0.0025]$
Redundancy probe	Flash	0.8926 $[0.8242, 0.9562]$	0.9756 $[0.9561, 0.9951]$	-0.0003 $[-0.0049, +0.0041]$
	Pro	0.9121 $[0.8486, 0.9707]$	0.9567 $[0.9176, 0.9860]$	-0.0100 $[-0.0263, +0.0045]$

Table 19: Collection-integrity counts for the full DeepSeek mirror. The last column reports empty responses and recorded model/thinking mismatches; both are zero for both models. Length-capped responses remain in the frozen data.

Model	Trajectories	Calls	Unique keys/IDs	Length capped	Empty/mismatch
`deepseek-v4-flash`	4,215	79,200	79,200/79,200	4,793	0/0
`deepseek-v4-pro`	4,215	79,200	79,200/79,200	4,998	0/0

G Prefix prediction and stopping

G.1 Feature contract

Full-trajectory single-sample accuracy can be used in an offline diagnostic, but it is unavailable to a stopping rule at an intermediate prefix. The deployable pipelines therefore use only prefix-visible answer statistics. Eleven fitted categorical predictors were checked and contain zero features that depend on unseen suffixes. Label and terminal columns remain in training files for supervision and verification, but are not selected by the fitted pipeline. Online answer canonicalization does not consult the held-out question's gold answers. Gold-based merging for offline utility evaluation is not used to construct prefix features.

For the 11 Qwen run artifacts, unweighted terminal categorical ratios are 0.903 raw, 0.968 isotonic, and 0.999 under the controller cumulative maximum. For DeepSeek, the categorical predictor beats raw prefix length by held-out R^2 in 10/14 slices but by MAE in only 2/14. Prediction results therefore depend on metric and slice; the controller comparison is the more consequential test.

Table 20: Held-out predictor comparison against raw prefix length. A win count is not evidence of repeatable aggregate superiority.

Target/features	R^2 wins	MAE wins	Mean ΔR^2
Offline full-trajectory feature set	1/11	1/11	0.0000
Bernoulli, prefix only	2/11	0/11	0.0000
Categorical, prefix only	4/11	1/11	0.0261

Table 21: Sensitivity of all seven frozen translation branches to the answer-equivalence contract. "Nonprimary/affected" gives accepted non-primary rows and affected trajectories. Values are alias-mean effective votes, not ratios.

Branch	Aliases	Traj.	Nonprimary/affected	M	Single-string raw/cummax	Accepted-set raw/cummax
Translation text	6	50	184/23	8	5.1611/5.1611	8.0000/8.0000
Translation deep	1	20	95/12	8	3.8000/3.8000	8.0000/8.0000
Seven-alias follow-up	7	176	589/76	8	4.7375/4.7688	7.9375/8.0000
Follow-up rerun	1	20	120/15	8	2.7500/2.7500	8.0000/8.0000
Prompt probe	7	672	2,131/277	8	5.1667/5.1771	7.9375/8.0000
Second branch	7	224	1,494/97	16	9.5045/9.6384	15.7321/16.0000
Second branch	7	224	2,997/99	32	18.1607/18.7143	30.8929/32.0000

G.2 Controller units and interpretation

Qwen controller outputs contain 1,316 aligned run-level question rows, which are first averaged to 710 unique questions. The 20,000-replicate bootstrap is then applied to questions. For DeepSeek, repeated slices are first averaged within model/question. This prevents repeated appearances of one question from acting as independent evidence.

The Qwen learned controller's accuracy interval includes zero and its matched objective interval includes zero. The DeepSeek result has the same pattern. Although call-count intervals exclude zero in 8/11 Qwen and 10/14 DeepSeek slices, fewer calls are not sufficient to establish a better controller. The aggregate conclusion is therefore a modest operating-point shift without a detectable accuracy or objective gain.

H Answer equivalence and information timing

H.1 Translation accepted-answer classes

Plurality evidence depends on the declared answer-equivalence relation. A single-string reference treats only the first listed gold string as correct, even when the task scorer accepts several equivalent variants. In the seven-alias panel, 589 of 1,408 rows are accepted variants that differ lexically from the first string; across the completed branches the count is 7,610. Under the task-consistent contract, all accepted variants collapse to one gold voting class, substantially reducing the apparent sub-nominal signal; residual raw-inverse deviations remain in some branches.

The task-consistent analysis shows that the large single-string reduction is not robust to the accepted-answer contract. Some raw inverses remain below nominal, so the cumulative-max values alone do not establish absence of a utility deficit or independence. No primary result uses the single-string values; the discussion in Section 6 reports them only as an answer-contract sensitivity. Semantic evaluation remains a separate possible analysis and is not mixed with exact-match voting.

H.2 Information available at stopping time

An offline predictor may use full-trajectory correctness, whereas an online controller cannot observe it before deciding to stop. All deployable results therefore use the prefix-only feature contract in Appendix G. The translation and controller analyses share the same principle: an operational conclusion may depend only on equivalences and information available at the decision point.

I ENDPOINT AND TASK BOUNDARIES

Open-ended generation. Canonical plurality is meaningful when an answer contract maps outputs to a finite set of decision classes. It does not capture novelty, discourse quality, factual coverage, or preference diversity in open-ended responses. Applying VOTE-ESS to such outputs would require a declared semantic equivalence rule and validation of that rule. Semantic-entropy methods aggregate uncertainty over meanings rather than surface strings (Farquhar et al., 2024); this is not the plurality-matched budget estimated here. Likewise, model-judge evaluation (Zheng et al., 2023), human-preference comparison (Chiang et al., 2024), and verifiable instruction-following constraints (Zhou et al., 2023) define different evaluation targets. No open-ended semantic-equivalence result is reported here.

Adaptive calls and external feedback. Tree of Thoughts branches over intermediate reasoning states (Yao et al., 2023a), ReAct interleaves reasoning with actions (Yao et al., 2023b), and Self-Refine and Reflexion use feedback to revise outputs or subsequent attempts (Madaan et al., 2023; Shinn et al., 2023). Process-supervised verification scores intermediate reasoning steps (Lightman et al., 2024). These approaches change the information or selection rule available across calls. They are distinct from repeated sampling under one fixed request contract; the present plurality-budget results do not establish their effective sample size or stopping performance.

Non-text endpoints. Embeddings, image generation/edit, realtime sessions, and speech synthesis have different user-visible operations. Endpoint-specific analyses may use retrieval stability, judge or hash diversity, transcript agreement, or waveform variation. Those quantities are not pooled into the categorical plurality result here. Table 22 groups the relevant provider documentation by operation. It is a map of interfaces outside the primary estimand, not an additional experimental panel.

Table 22: Endpoint-operation boundaries. Documentation identifies interfaces; it does not supply empirical evidence for AURA. Candidate observables in the right column would require separate validation. Living documentation is cited as accessed on September 25, 2026, not as a snapshot of the collection period.

Operation	Relevant model/interface sources	Distinct observable or boundary
Text-call protocol	OpenAI-compatible Qwen chat (Alibaba Cloud, 2026l)	Request and response schema; compatibility does not establish identical serving behavior.
Embeddings and reranking	Qwen3 Embedding (Zhang et al., 2025); text-embedding API (Alibaba Cloud, 2026a)	Vector or retrieval-rank stability, not a categorical answer vote.
Multimodal and realtime interaction	Qwen-Omni and realtime interfaces (Alibaba Cloud, 2026j;k)	Declared modality and session unit; repeated turns are not interchangeable fixed calls.
Translation	Qwen-MT (Alibaba Cloud, 2026h); audio/video and realtime translation (Alibaba Cloud, 2026f;g)	An explicitly declared translation-equivalence rule; documentation does not define the accepted-answer classes in our frozen translation analysis.
Visual reasoning and extraction	QVQ-Max visual reasoning (Qwen Team, 2025a); Qwen-OCR extraction (Alibaba Cloud, 2026i)	Separate reasoning or extraction contracts; these references do not add evaluated aliases to the reported panel.
Image generation and editing	Qwen-Image, Wan 2.7, and Z-Image (Alibaba Cloud, 2026d;r;s); Qwen-Image Edit (Alibaba Cloud, 2026e)	Perceptual or judge-defined similarity; identical hashes and semantic equivalence are not interchangeable.
Speech recognition and media workflows	File and realtime recognition (Alibaba Cloud, 2026b;c); Tingwu (Alibaba Cloud, 2026q)	Transcript agreement and a declared segmentation/session boundary.
Speech synthesis and voices	TTS and realtime TTS (Alibaba Cloud, 2026m;n); cloning and design (Alibaba Cloud, 2026o;p)	Waveform, intelligibility, or speaker variation; none is directly a plurality-accuracy budget.

API drift. Hosted services can change routing, weights, decoding, safety layers, and response formatting without exposing every change (Gao et al., 2025; Cai et al., 2025). Model identifiers are not necessarily immutable weight identifiers: current DeepSeek documentation permits legacy names to route to a replacement model (DeepSeek, 2026). This is a live-rerun caveat, not a finding about which implementation produced the frozen records. Frozen outputs support offline reproduction of this paper. A live rerun supports only a new time-stamped replication unless provider identity and the full request contract can be shown equivalent.

J Additional multi-mechanism evidence

J.1 Calibration and decision regret

This experiment tests whether categorical VOTE-ESS improves operational-budget recovery over simpler output-only proxies beyond the Dirichlet–multinomial construction. The 288-cell grid crosses three gold probabilities, two wrong-class counts, three budgets, four dependence mechanisms, and four strengths. Each cell has 20 macro-seeds. Oracle-reference, training, validation, and test banks use disjoint seed namespaces, and every target's categorical reference comes from an independent bank. The primary measurement outcome is absolute error in oracle vote-equivalent calls; the primary decision outcome is held-out regret for plurality accuracy minus λ times calls.

The four mechanisms create dependence in different ways. A Dirichlet–multinomial trajectory first draws a trajectory-specific categorical law. Exact block replay draws q independent prototypes and repeats them in a balanced cyclic schedule. Markov copy/redraw copies the preceding answer with probability s and otherwise draws afresh. Shared-batch mixture uses one trajectory-level answer with probability s at each position and otherwise draws afresh. For the first, third, and fourth mechanisms, $s \in \{0, 0.10, 0.25, 0.50\}$. For block replay, $q = \max(2, M/d)$ with $d \in \{1, 2, 4, 8\}$. The four strength levels are therefore ordered within a mechanism, not numerically comparable across mechanisms. For the cross-mechanism rank diagnostic we instead use the known target pair dependence: s for Dirichlet–multinomial, the fraction of call pairs sharing a prototype for block replay, $\sum_{\ell=1}^{M-1}(M-\ell)s^{\ell}/[M(M-1)/2]$ for Markov copy/redraw, and s^2 for shared-batch mixture.

Table 23: Design of the multi-mechanism stress test. The oracle is a regret lower bound, not a candidate method. All baselines were specified before held-out evaluation.

Design component	Frozen specification
Factorial grid	$p_g \in \{0.55, 0.70, 0.85\}$; $K \in \{2, 20\}$; $M \in \{8, 16, 32\}$; four mechanisms; four strengths
Independent repetitions	20 macro-seeds within each of 288 design cells
Banks per cell	65,536 oracle-reference; 65,536 categorical-reference; 65,536 categorical-curve; 1,024 train; 1,024 validation; 4,096 held-out
Fixed comparators	Raw count, distinct classes, collision/design-effect, Bernoulli VOTE-ESS, classical indicator-ICC ESS, categorical VOTE-ESS
Controller costs	$\lambda \in \{0.0025, 0.005, 0.010, 0.020\}$
Inferential unit	Equal-cell contrast after averaging macro-seeds within cell
Execution record	16 low-priority CPU workers; 862.77 s; 152,600 KiB peak RSS; no API/GPU

The six seed namespaces are `oracle_reference`, `categorical_reference`, `categorical_curve`, `train`, `validation`, and `test`. The categorical-reference bank estimates the marginal answer law; a separate i.i.d. categorical-curve bank maps budget to plurality utility. We fit that curve by nondecreasing isotonic regression. For each distinct fitted utility, its inverse location is the mean of all budgets on the corresponding plateau; linear interpolation between adjacent distinct utilities then defines a fractional budget, clipped to $[1, M]$. We call this stress-test diagnostic the *interpolated isotonic categorical* VOTE-ESS; it is distinct from the discrete max-inverse in Eq. (2). At each budget, we form a *utility-error-compatible interval* by inverting $\widehat{U} \pm 1.96\,\mathrm{SE}(\widehat{U})$ from the held-out trajectories through that isotonic categorical curve, with endpoints clipped to the supported range $[1, M]$. This interval propagates held-out utility sampling error through the inverse but not Monte Carlo uncertainty in the estimated reference curve. It is therefore a diagnostic interval, not a nominal 95% confidence interval for the effective budget.

The three output-only comparators are raw prefix length, the mean number of distinct answer classes, and collision/design-effect ESS. For prefix length n, the last is $n/[1+(n-1)\widehat{\rho}_{\mathrm{coll}}]$, where

$$\widehat{\rho}_{\mathrm{coll}} = \max\left\{0, \frac{\widehat{A} - \sum_k \widehat{\pi}_k^2}{1 - \sum_k \widehat{\pi}_k^2}\right\},$$

and $\widehat{A}$ is mean within-trajectory pair agreement. The Bernoulli and gold-indicator ICC quantities use the gold label and are therefore sensitivity comparators rather than output-only baselines.

All adaptive controllers in this stress test apply the prefix risk rule in Eq. (4). After adding α to each observed class and to one unseen-class branch, let p_1, p_2 be the two largest smoothed shares and let e_n be the controller's effective prefix length. The current error proxy is $\Phi\{-\sqrt{e_n}(p_1 - p_2)/[p_1(1-p_1) + p_2(1-p_2)]^{1/2}\}$; the next-step risk averages this expression over the corresponding smoothed predictive branches. The controller stops when the expected one-step risk reduction is at most λ. For each controller and λ, $\alpha \in \{0.5, 1, 2\}$ is chosen solely by validation objective, with smaller α breaking ties, and the selected rule is evaluated once on the held-out test bank. Regret is test objective loss relative to a per-trajectory hindsight oracle, which is reported only as a lower bound and is not deployable.

The evaluation required all 288 cells and 5,760 cell–seed records, dependence and marginal calibration within 0.02, terminal-utility Monte Carlo standard error at most 0.002, and no reference leakage. Support for the controller extension additionally required block-replay MAE at most 1.0 vote with at least 90% compatible-interval coverage; at least 15% lower equal-cell MAE than the best prespecified output-only baseline with a favorable paired interval; and, at $\lambda = 0.010$, overall regret no worse by more than 0.002 while improving regret by at least 10% in dependence-rich cells for three of four mechanisms. Because the interval-coverage and controller criteria were not met, we retain only the measurement result. We make no decision-aware controller claim, and the planned hosted-API extension was not run.

All 5,760 cell–seed records completed. The independent verification pass found exactly 288 cells with macro-seeds 0–19, 34,560 globally unique namespace seeds, 34,560 measurement rows, and 184,320 controller rows. The largest marginal/dependence miss is 0.00684 and the largest terminal-reference Monte Carlo standard error is 0.001468. Source, simulation, and analysis hashes match the recorded metadata.

Table 4 gives equal-cell summaries. Interpolated isotonic categorical VOTE-ESS improves point recovery for every mechanism, whereas its controller has higher regret for every mechanism. The intervals use 20,000 bootstrap resamples of design cells after the 20 macro-seeds are averaged within cell.

The complete mechanism-level table is in the main text (Table 4); the following diagnostics retain the coverage and boundary behavior needed to interpret it.

Low exact-block MAE does not imply calibrated compatible intervals. Table 24 shows that undercoverage is concentrated but not isolated to one setting. In particular, budget-32 coverage is only 76.25% overall, and the nominally independent budget-32 block condition covers only 45% of macro-seed records.

Table 24: Exact-block compatible-interval coverage (percent). Levels use prototype counts $M, M/2, M/4, \max(2, M/8)$; the two coincident $M = 8$ schedules remain distinct design levels.

Budget	Level 0	Level 1	Level 2	Level 3	Overall
8	100.0	100.0	70.8	77.5	87.1
16	80.0	97.5	100.0	71.7	87.3
32	45.0	65.0	95.0	100.0	76.3

The sign diagnostics do not indicate reference leakage or failed parameter calibration. They do show inversion boundaries: 11.61% of terminal categorical estimates are clipped at 1 or M, 22.22% of compatible intervals touch a boundary, and isotonic reference curves are flat across 15.69% of adjacent steps on average. Exact-block level 3 has signed error -0.0258 $[-0.0469, -0.0067]$ and coincident

boundary behavior; shared-batch level 2 has a smaller negative error, -0.0067 $[-0.0141, -0.0008]$. We therefore do not treat the inverse intervals as calibrated uncertainty and do not advance the controller result.

J.2 Full comparator and dependence-level plots

Figures 5 and 6 retain the full comparator and dependence-level views underlying the compact paired contrasts in Figure 3. All panels average the 20 macro-seeds within each design cell and include the independent level. Across all 288 cells, categorical and collision MAE are 0.204 and 1.288 votes (an 84.1% reduction), whereas the 86.4% headline excludes independent cells.

J.3 Temporal replication as a remaining boundary

A time-separated, three-provider replication remains future work. Using paired standard deviations from the existing 100-question panel, 160 complete questions would provide 80% power for differences of 0.05 in the raw ratio and 0.03 in the direct gap; 180 would allow for attrition. Because the present interval-calibration and controller results do not support that extension, we make no three-provider or temporal-stability claim.

K Reproducibility and integrity

Anonymous artifact access. The companion artifact repository is available at https://anonymous.4open.science/r/aura-artifact-AD85/. The metadata and result-table hashes below identify the frozen analyses reported in this paper.

K.1 Frozen analysis metadata

Table 25: Core frozen metadata used by the reported analyses.

Component	Recorded metadata
Synthetic/reference validation	seed 20260725; 20,000 synthetic draws; 65,536 synthetic-reference draws; 32,768 empirical-reference draws; 5,000 bootstrap replicates
Exact-condition cross-fit	seed 20260726; 40,000 reference draws; 20,000 unique-question bootstrap replicates; 168-hour maximum collection window
Controller comparison	seed 20260726; 20,000 unique-question bootstrap replicates; objective accuracy $-\ \lambda$ calls using each frozen run's λ
Multi-mechanism test	base seed 20260901; 288 cells; 20 macro-seeds per cell; 65,536 draws in each oracle-reference, categorical-reference, and categorical-curve bank; 1,024/1,024/4,096 train/validation/test trajectories; 20,000 cell-bootstrap replicates (gate seed 20260902; diagnostic seed 20260904; figure seed 20260905); 16 low-priority CPU workers; 862.77 s; peak RSS 152,600 KiB; no API/GPU
Provider mirror	provider/model IDs, call keys, response IDs, token counters, length-cap flags, prompt and decoding contracts

K.2 Result-table provenance

Each principal number in the appendix is tied to a compact canonical result table. Table 26 records the frozen SHA-256 values used for the reported analyses. The hashes identify the analyzed tables; they do not by themselves prove a scientific claim, which also requires the stated unit, reference construction, and uncertainty calculation.

K.3 Integrity checks

The answer-equivalence and information-timing checks pass: accepted-set translation classes reproduce 7.9375/7.9688/8.0000 raw/isotonic/cumulative-max effective votes, held-out test values are 8/8 for all three variants, all 11 fitted categorical predictors exclude forbidden online features, and stored label columns are used only for supervision or verification. DeepSeek collection checks find

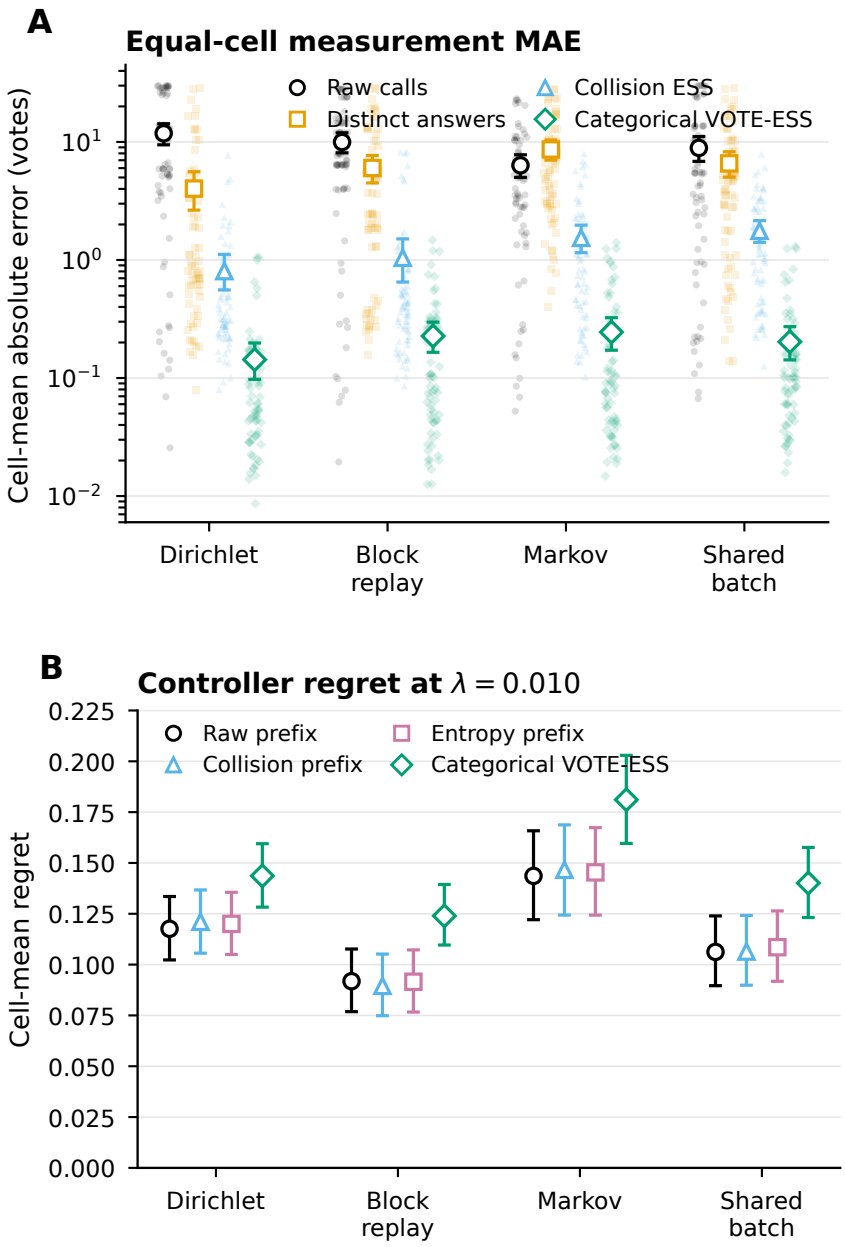

Figure 5: Full multi-mechanism comparator views. (A) Measurement MAE for interpolated isotonic categorical VOTE-ESS and three output-only proxies, on a logarithmic vertical axis; light points are design-cell summaries. (B) Held-out controller regret at $\lambda = 0.010$. Points are equal-cell means and bars are 95% cell-bootstrap intervals. Lower is better in both panels. Categorical measurement improves, whereas its controller is consistently worse than the raw-prefix baseline.

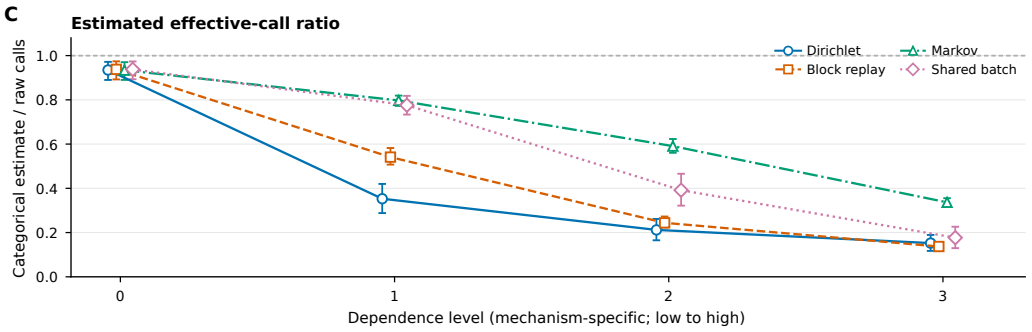

Figure 6: Full stress-test strength diagnostic (panel C of the original comparator view). The categorical effective/raw-budget ratio is shown for four mechanism-specific dependence levels; points are equal-cell means and bars are 95% cell-bootstrap intervals. Levels run from low to high within each mechanism and are not numerically interchangeable across mechanisms.

Table 26: Frozen hashes for the compact result tables used in this appendix.

Canonical result	SHA-256
Synthetic grid	`0bbc26c15b3ee88dcde8aaabb8ca49211848847ae4f39609971b47b9e6b42396`
Synthetic summary	`0920d87242279e4c3b4d0ea063be3648a1ce4c7c911651aa9b306da97b248339`
Qwen plug-in ratios	`cf620ab6495eb375eac5dba5f7e53d2682d7de5a71dc67a2299416a90234dc56`
Qwen direct gaps	`334207d44afa3b5ec038205a029ee86d58a2824dc59ec893a7cecc7b07506df0`
Cross-fit summary	`ae5b2c61ed49c929ca3b8393ff933f1cebc0b4fc5773755423569769f88b0a67`
Cross-fit run table	`1147aa4358bd67bcd946f3fb1d803d3c6d847c6a7d40a68bf3812011a99f06a8`
Qwen controller	`9bcd7c43f38cc279ef127cacc659a41d2533dafa01c3061fb75a7599db151f2b`
DeepSeek full matrix	`64cc3779e86a44e34b5906bc31437b2bc54bcc9e277e05c19f6ceaad62fad9b8`
DeepSeek controller	`b83710d1f0bed476e56e8f92287d3fff464a366d8fc10fd637cb90d117cca384`
69-alias groups	`236f2bd5e2f7696004dee02d3f99b12d44939e0f65f3ff2f6e5971f55fdd364d`
69-alias rows	`aa014274dd56ce8ab806beb656788f5f3348a83c6b23c26bb15f373b0d530e14`
Translation branches	`82377655c88da69975efc36125a500b34e8ee0e660c9d51fccf46ec40481ba4f`
Multi-mechanism simulation	`f8484e958e1c997d46fe198465e9b55a62d401933d2827b86c12c95765e75d65`
Multi-mechanism measurement	`626870bfc028115a17dae1ffa0b753a9e46a768924878b49cee7061b49b2688f`
Multi-mechanism controller	`76665cc1f39d972bdbb5688ebd1bb0bc9258773e46378d181664074df6e05eea`
Multi-mechanism gate	`a1826af4c542d56715a4f3baf29881b3f35d700d617f39e4f0c117137df5129b`
Manual diagnostic packet	`e77acf2cc1dda5a30f28309877b056e09f5541eca41d19b25d08ac851225f6fe`

no duplicate call keys or response IDs, empty outputs, model mismatches, provider-sent seeds, or thinking-setting mismatches among the 158,400 completed calls.