ONE SCHEDULE DOES NOT FIT ALL: ADAPTIVE INFERENCE SCHEDULES FOR DIFFUSION LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion language models commonly rely on a single, fixed denoising schedule across all workloads, even though the cheapest quality-preserving allocation of model evaluations varies substantially across tasks and checkpoints. We introduce **Task-Calibrated Denoising Allocation (TCDA)**, a confidence-controlled calibration framework that selects the lowest-cost schedule from a structured candidate library using a paired calibration sample and a user-specified quality tolerance. TCDA compares candidate topologies, including early prefix decoding, state-reusing online continuation, reduced full-range resampling, and the canonical baseline, deploying the cheapest schedule certified by one-sided simultaneous confidence bounds and task-specific degeneration guardrails.

Across three frozen ELF-B tasks, TCDA identifies three distinct operating regimes: early prefix stopping for XSum, state-reusing online continuation for WMT, and guarded full-range resampling for OpenWebText. At representative tolerances ($\varepsilon = 1.0$ for XSum and WMT; 20% relative perplexity for OpenWebText), selected schedules reduce denoiser evaluations by $1.33\times$ to $3.37\times$, yielding measured wall-clock speedups between $1.30\times$ and $2.91\times$ with held-out violation rates below 0.4%. Under strict margins, the method safely defaults to the full schedule. These findings establish that optimal inference reduction is workload-dependent, providing a disciplined methodology for budget-constrained diffusion inference.

1 INTRODUCTION

Generating text with continuous diffusion and flow models requires evaluating a neural denoiser repeatedly along a trajectory from noise to clean data. In models like Embedded Language Flows (ELF) (Hu et al., 2026), the standard recipe evaluates the network across all 64 steps of a logit-normal time grid. This is computationally expensive, and in practice, much of this compute is wasted after the sequence has already stabilized.

A straightforward fix would be to apply a single, smaller schedule across all workloads. However, our findings show that no universal schedule exists. The best way to spend a reduced compute budget depends entirely on the task: on summarization (XSum), core content stabilizes early, so stopping after step 19 and decoding immediately matches the 64-step baseline; on translation (WMT De–En), early stopping collapses and trajectories must reach the terminal step, but reusing internal states along a coarser continuation grid halves the cost; on unconditional generation (OpenWebText), neither early stopping nor state reuse succeeds, yet the model tolerates a coarser resampled grid across the full trajectory. Because the optimal topology varies by workload, schedule compression is an empirical model-selection problem: given a task, a deployed checkpoint, and an acceptable quality tolerance, which schedule maximizes speedup with certified statistical support?

We propose TCDA (Task-Calibrated Denoising Allocation) to resolve this trade-off automatically. TCDA is not a learned controller or an adaptive sampler. It is a simple, rigorous selection rule evaluated over a fixed library of schedule topologies: early prefix stops, state-reusing online continuations, resampled full-range grids, and uniform downsampling. Using a paired calibration set of modest size ($n = 128$), TCDA compares candidate schedules against the full 64-step baseline. It constructs one-sided simultaneous confidence bounds, verifies task-specific degeneration guardrails, and selects

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

the cheapest candidate certified to remain within a user-specified quality tolerance ε. If no candidate meets this bar, the method safely falls back to the full schedule.

Our contributions are as follows:

- **A confidence-controlled selection framework.** We frame schedule reduction as a tolerance-constrained decision with family-wise error rate control. On held-out splits at fixed $n = 128$, TCDA matches full-dataset reference policies in over 92% of runs, with empirical violation rates below 0.4% on WMT and 0% on XSum.

- **Identification of three distinct denoising regimes.** Across frozen ELF-B tasks, TCDA discovers fundamentally different schedules: prefix stopping for XSum ($3.37\times$ fewer evaluations, $2.91\times$ faster wall-clock), state-reusing continuation for WMT ($2.00\times$ fewer evaluations, $1.94\times$ faster wall-clock), and full-range resampling for OpenWebText ($1.33\times$ fewer evaluations, $1.30\times$ faster wall-clock).

- **A corrected sequential racing procedure.** We identify four fatal failure modes in standard sequential racing with optional stopping, and introduce a corrected multi-stage rule using nested calibration cohorts and stagewise α-spending. Because this corrected sequential test is conservative (median stopping size 64 to 512), we recommend fixed $n = 128$ calibration for practical use.

- **Practical calibration economics.** We track denoiser evaluation counts and actual wall-clock runtimes separately, deriving break-even deployment thresholds for each setting. At our primary operating points, calibration overhead amortizes after approximately 1,100 to 2,600 inference requests.

2 Embedded Language Flows and Schedule Topologies

Embedded Language Flows (ELF) (Hu et al., 2026) combine a frozen T5 encoder (Raffel et al., 2020) with a continuous latent flow-matching model and a token decoder. Generation works by integrating a learned velocity field over the normalized time interval $t \in [0, 1]$, moving from standard Gaussian noise toward an encoder-conditioned latent representation. Once integration is complete, a decoder head maps the final continuous representation to discrete tokens.

As with other flow and diffusion models (Ho et al., 2020; Song et al., 2021; Karras et al., 2022; Lipman et al., 2023; Liu et al., 2022), inference cost scales directly with the sampling schedule, which defines both the time grid and the final decoding point. In practice, evaluating the denoiser across the full canonical grid (64 steps for ELF-B and ELF-M on XSum and WMT; 32 steps on OpenWebText) wastes substantial compute on regions of the trajectory that no longer change the decoded text.

Rather than assuming all tasks can be accelerated identically, we define a structured candidate library spanning four distinct trajectory topologies (Figure 1, Table 5): (1) **Prefix Truncation (PREFIX):** Run the first K evaluations of the canonical schedule, halt early, and decode the current predicted clean latent x_{pred} at cost K. (2) **Online Continuation (HYBRID):** Run a short prefix on the canonical schedule and continue along a coarser monotone grid to $t = 1$, reusing intermediate latent and self-conditioning states without restarting (total cost H). (3) **Resampled Full-Range (RESAMPLED):** Traverse the full interval $[t_\varepsilon, 1]$ using a reduced budget of K points drawn from the logit-normal distribution (cost K). (4) **Uniform Downsampling (UNIFORM):** Traverse $[t_\varepsilon, 1]$ using K evenly spaced time steps as a negative control (cost K).

The canonical schedule is always included in the library as a guaranteed fallback. Candidate schedules share identical initial noise and random seeds with the canonical run; consequently, the first 19 evaluations of candidate `prefix19` are numerically identical to the first 19 steps of the full 64-step trajectory. This coupling isolates the effect of early truncation from sampling noise, allowing us to evaluate whether the remaining 45 steps contribute meaningful signal to the final decoded text.

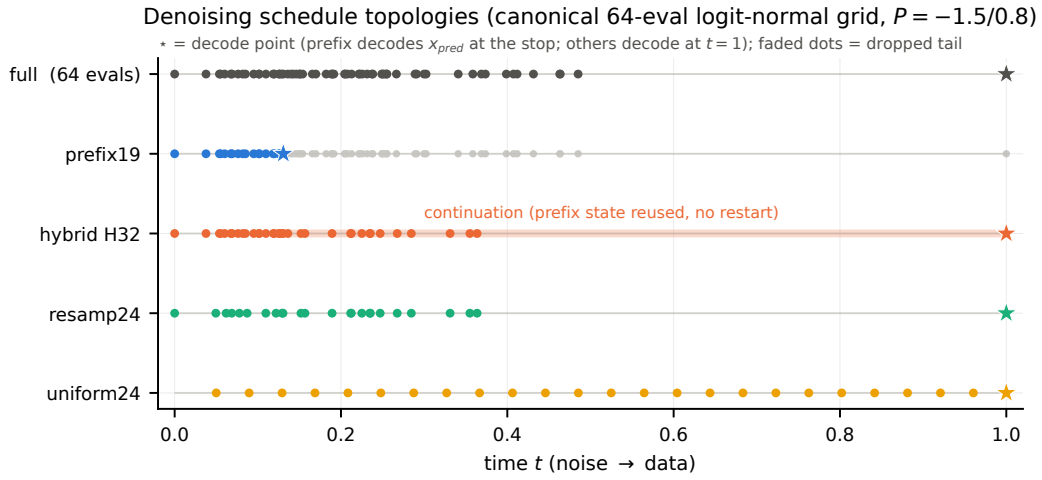

Figure 1: **Schedule action space.** The canonical schedule reaches the terminal endpoint; prefix schedules decode the predicted clean latent early; hybrid schedules reuse intermediate states before a coarse continuation; resampled and uniform baselines span the full horizon with fewer steps. Stars mark decode points.

3 Task-Calibrated Denoising Allocation

3.1 Problem Formulation

Let $u_i(\pi)$ denote the generation utility of schedule π on example i (e.g., ROUGE-L $\times 100$ for summarization, chrF for translation, or negative log-likelihood for language modeling), with cost $c(\pi)$ defined as the number of denoiser evaluations. To isolate schedule compression from sampling noise, candidates share random seeds and initial latents with the canonical full schedule π_{full}, yielding paired utility differences $\Delta_i(\pi) = u_i(\pi) - u_i(\pi_{\text{full}})$.

Given a user-specified quality tolerance $\varepsilon \geq 0$, a candidate schedule is acceptable if its expected paired utility satisfies:

$$\mathbb{E}[\Delta(\pi)] \geq -\varepsilon. \tag{1}$$

Our objective is to identify the lowest-cost schedule satisfying this condition with high probability, defaulting to π_{full} when evidence for acceleration is insufficient.

3.2 Confidence-Controlled Selection Rule

Evaluating multiple candidate schedules on a shared calibration sample risks false discoveries under naive empirical selection. To ensure disciplined error control, TCDA computes simultaneous one-sided lower confidence bounds on $\mathbb{E}[\Delta(\pi)]$ across the candidate set via a studentized Westfall–Young max-T bootstrap over paired resamples (Westfall and Young, 1993), controlling the family-wise error rate at $\alpha = 0.05$. A Holm-corrected one-sided paired test (Holm, 1979) serves as a secondary check.

A candidate is certified as **SUPPORTED_FEASIBLE** if and only if: (i) its simultaneous lower confidence bound clears the non-inferiority margin, $\text{LCB}(\pi) \geq -\varepsilon$; and (ii) it passes all predeclared task guardrails against text degeneration (such as repetition or collapse in length). All other candidates remain **NOT_CERTIFIED**. Among certified schedules, TCDA deploys the policy with the lowest evaluation cost $c(\pi)$; if none qualifies, it defaults to the canonical full schedule π_{full}.

Testing non-inferiority against $-\varepsilon$ establishes critical thresholds from the distribution of $\max_\pi t_b^*(\pi)$ rather than absolute magnitudes. Additionally, if any bootstrap resample exhibits zero sample variance for a candidate, the procedure retains the replicate rather than dropping it, preserving conservative coverage across calibration splits.

Task	Model	Primary Metric	Canonical Budget	Tolerance Grid (ε)
XSum	ELF-B	ROUGE-L	64 evaluations	$\{0, 0.25, 0.5, 1.0\}$
WMT14 De–En	ELF-B	sentence chrF	64 evaluations	$\{0, 0.25, 0.5, 1.0\}$
OpenWebText	ELF-B	$-$NLL (PPL)	32 evaluations	$\{0, 5, 10, 20\}\%$ rel. PPL
XSum Transfer	ELF-M	ROUGE-L	64 evaluations	0.5 (prospective)

Table 1: Summary of experimental workloads, models, evaluation metrics, canonical evaluation budgets, and quality tolerances. Checkpoint paths, random seeds, and candidate library configurations are detailed in the appendix.

3.3 Candidate Libraries and Guardrails

Rather than searching an unconstrained combinatorial space of time steps, TCDA evaluates a structured library spanning the four schedule archetypes defined in Section 2: prefix truncation, online state-reusing continuations, resampled full-range grids, and uniform controls. Archetype parameterizations account for task-specific structural factors (such as target sequence lengths and baseline budgets) and are fixed prior to calibration. While the candidate set offers diverse trade-offs between evaluation cost and trajectory depth, TCDA itself remains agnostic to library composition, certifying viable accelerated schedules when statistical support exists and defaulting to the canonical baseline otherwise.

3.4 Sequential Calibration with Optional Stopping

While fixed-sample calibration ($n = 128$) serves as our primary protocol, deployments can also calibrate adaptively via an optional sequential racing variant. The sequential protocol evaluates nested sample sizes $n_1 < n_2 < \cdots < n_J$ under a predetermined α-spending schedule ($\alpha_j = \alpha_{\text{total}}/J$ via a Bonferroni union bound). At each intermediate checkpoint, it computes simultaneous two-sided confidence bounds to prune uncompetitive candidates and halt sampling once a candidate satisfies the non-inferiority criterion. Although this sequential test is conservative (yielding median stopping sizes between 64 and 512 examples), it provides mathematically valid error control under optional stopping for continuous calibration pipelines.

4 Experimental Setup

Table 1 summarizes the benchmark tasks, checkpoints, primary utilities, and evaluation tolerances.

ELF-B evaluation benchmarks: We evaluate three text generation workloads using public ELF-B checkpoints: (1) **Abstractive Summarization (XSum):** evaluated via ROUGE-L $\times 100$ (Lin, 2004; Narayan et al., 2018); (2) **Machine Translation (WMT14 De–En):** evaluated using sentence-level chrF (Popović, 2015; Bojar et al., 2014), with BLEU and SacreBLEU diagnostics (Papineni et al., 2002; Post, 2018); and (3) **Unconditional Language Modeling (OpenWebText):** evaluated by negative log-likelihood ($-$NLL) under a frozen GPT-2 Large model (Radford et al., 2019; Gokaslan and Cohen, 2019) and reported in perplexity (PPL), using classifier-free guidance during sampling (Ho and Salimans, 2022). For XSum and WMT, tolerances span $\varepsilon \in \{0, 0.25, 0.5, 1.0\}$ metric points; for OpenWebText, relative perplexity tolerances span $\{0, 5, 10, 20\}\%$.

Fixed-sample calibration evaluates cohorts of size $n = 128$ across 100 random splits per setting, while sequential racing evaluates 500 nested permutations. Out-of-sample performance is measured on the disjoint complement of each calibration split. All routines run offline and modularly alongside the ELF architecture without altering model weights or velocity fields.

Independently trained ELF-M model: To assess whether schedule behaviors transfer across model scales, we trained an ELF-M model from scratch on XSum (24 layers, hidden dimension 1,056, totaling 342,540,036 parameters). Training followed the standard synchronized-time flow-matching objective using the Muon optimizer with a global batch size of 512 distributed across eight NVIDIA B200 GPUs in bfloat16 precision for 39,800 steps (exponential moving average decay of 0.9999, random seed 42).

Task	Tolerance (ε)	Selected Schedule	Full → Selected Quality	Eval. Ratio	Wall-Clock Speedup
XSum	1.0 ROUGE-L	`prefix19`	28.10 → 28.50 ROUGE-L	3.37×	2.91×
WMT	1.0 chrF	`H32`	52.65 → 52.71 chrF	2.00×	1.94×
OWT	20% PPL	`resamp_K24`	24.94 → 27.91 PPL	1.33×	1.30×

Table 2: ELF-B descriptive operating points across full-dataset frontiers. Output quality reflects full-dataset evaluations. On OpenWebText, `resamp_K24` represents the guardrail-screened policy, which preserves generation diversity within the declared 20% relative-perplexity margin.

5 Task-Dependent Denoising Regimes

Evaluating candidate libraries across the full dataset shows that optimal budget allocation varies fundamentally across tasks (Figure 2, Table 2). At target operating points ($\varepsilon = 1.0$ for XSum and WMT; 20% relative perplexity for OpenWebText), selected policies deliver substantial accelerations within declared tolerances: on XSum, `prefix19` cuts denoiser evaluations from 64 to 19 ($3.37\times$) with a $2.91\times$ wall-clock speedup; on WMT De–En, online continuation `H32` halves evaluations from 64 to 32 ($2.00\times$) with a $1.94\times$ speedup; and on OpenWebText, `resamp_K24` reduces evaluations from 32 to 24 ($1.33\times$) with a $1.30\times$ speedup. Across all settings, wall-clock speedups are slightly lower than theoretical evaluation reductions due to fixed tokenization, embedding, and vocabulary decoding overheads, showing why both measures must be reported separately.

Relying solely on perplexity or likelihood can be deceptive in unconditional generation. A primary-metric-only criterion initially selected `resamp_K16`, but text entropy analysis revealed repetitive loops that artificially depressed cross-entropy loss. We therefore introduced explicit guardrails monitoring n-gram repetition, generation length, and decoding entropy (Table 9). These guardrails filter out `resamp_K16` and select `resamp_K24`, which preserves genuine output diversity within the 20% perplexity margin.

To formalize how workloads spend their denoising budgets, we report the Endpoint Dependence Index (EDI) in task-native units with 95% confidence intervals. Because metrics operate on different scales, EDI is evaluated within each task rather than compared across them: (1) **Early-Decodable Regime (XSum):** XSum yields a negative EDI of -0.32 (CI: $[-0.48, -0.16]$), confirming that early truncation preserves quality. Core content stabilizes early, making prefix stopping optimal. (2) **Online-Continuation Regime (WMT):** WMT yields a positive EDI of $+35.54$ chrF points (CI: $[+34.94, +36.13]$). Direct prefix stopping destroys translation quality, proving the trajectory must reach the terminal step; however, successful state reuse allows a coarse continuation grid to halve inference cost. (3) **Reduced Full-Range Regime (OpenWebText):** OpenWebText also requires endpoint coverage (EDI $+0.80$ NLL, CI: $[+0.78, +0.83]$), but does not support state-reusing hybrids, tolerating coarser discretization across the entire integration interval instead.

6 Tolerance-Dependent Calibration Reliability

Reliability and fallback behavior. We evaluate TCDA across all 12 task-by-tolerance configurations using 100 independent splits at fixed calibration size $n = 128$ (Figures 3 and 4), measuring selected policies out of sample on the disjoint held-out cohort. Empirical held-out violation rates remain below 0.4% across all cells, well within the target family-wise error budget ($\alpha = 0.05$). At primary operating points ($\varepsilon = 1.0$ for XSum and WMT; 20% relative perplexity for OpenWebText), TCDA consistently recovers the full-dataset reference policy: selections match in 95.4% of trials on XSum, 92.8% on WMT, and 90.0% on OpenWebText (Table 3). When tolerances tighten ($\varepsilon \leq 0.25$), candidates cannot guarantee non-inferiority under sampling noise; TCDA conservatively defaults to the canonical full schedule to prevent quality degradation.

Sequential racing and unconstrained selection. Benchmarking TCDA against unconstrained empirical selection (choosing the candidate with the highest sample mean at $n = 128$) yields held-out violation rates up to 32.6% (Table 8), confirming that safety stems from simultaneous confidence bounds rather than empirical search alone. Our corrected sequential racing procedure matches this reliability (violations $\leq 0.4\%$). However, because sequential racing enforces simultaneous coverage across candidates and evaluation checkpoints, it requires larger cohorts to stop early under small

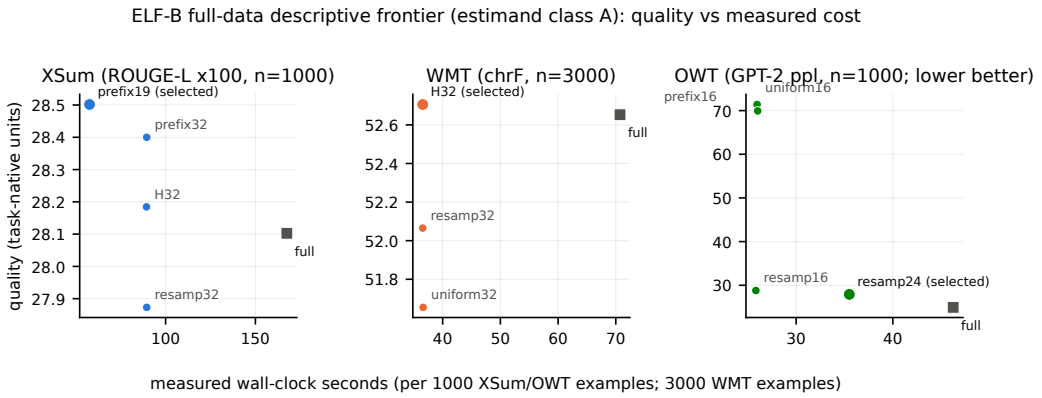

Figure 2: Quality versus computational cost trade-offs across ELF-B benchmarks. Budget reductions require different trajectory topologies depending on the task: XSum favors early prefix truncation, WMT requires endpoint continuation with state reuse, and OpenWebText relies on coarse full-range resampling.

Task	Tolerance (ε)	Modal Policy	Reference Match	Violation	Mean Evals
XSum	1.0 ROUGE-L	`prefix19`	95.4%	0.0%	19.7
WMT	1.0 chrF	`H32`	92.8%	$\leq 0.4\%$	32.7
OWT	20% PPL	`resamp_K24`	90.0%	0.0%	23.8

Table 3: Fixed-n=128 calibration across primary tolerances. TCDA reliably recovers the reference schedule at permissive tolerances and safely defaults to the full schedule under tight margins (see Appendix L).

effect sizes (median sample size 64 to 512). Consequently, fixed $n = 128$ calibration provides the more practical, predictable protocol for standard deployments.

7 Prospective Transfer to ELF-M

Experimental design and preregistration: To test whether the XSum prefix-stop regime transfers across model scale, we evaluate an independently trained ELF-M model on XSum. Rather than conducting an unconstrained search, we preregistered a single-candidate prospective confirmation protocol prior to inference (Appendix E; SHA-256 `0dd07054...`).

We compare candidate `prefix19` (which runs the first 19 canonical steps and decodes the predicted clean latent x_{pred}) against the 64-step baseline, `original64`. Evaluations were performed on an untouched cohort of 1,000 XSum examples (IDs 1000–1999) using common initial noise across three fresh seeds (126, 127, and 128). Non-inferiority is evaluated at tolerance $\varepsilon = 0.5$ ROUGE-L points using 10,000 paired bootstrap resamples per seed, alongside a hierarchical bootstrap resampling seeds and examples within seed. Because only one candidate is evaluated, no family-wise multiplicity correction is needed.

Empirical confirmation: As detailed in Table 5 and Figure 5, `prefix19` satisfies all preregistered non-inferiority criteria. Across all three seeds, the mean paired ROUGE-L differences are strictly positive ($+0.1424$, $+0.1913$, and $+0.2195$), and the corresponding one-sided 95% lower confidence bounds ($+0.0028$, $+0.0670$, and $+0.1006$) comfortably clear the -0.5 non-inferiority threshold. The hierarchical bootstrap yields an overall mean difference of $+0.1844$ with a one-sided lower bound of $+0.1037$. In addition, all task guardrails tracking sequence length, empty outputs, copy rates, and n-gram diversity pass across every seed.

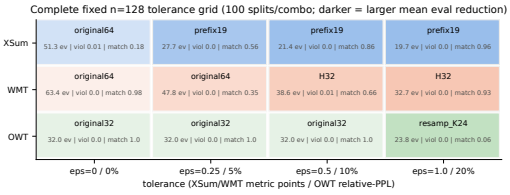

Figure 3: Complete fixed-n=128 tolerance grid showing modal schedule, mean evaluations, held-out violation rate, and reference match.

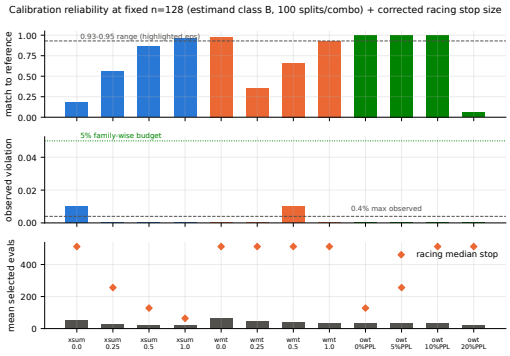

Figure 4: Calibration reliability profiles across target tolerances, including median stopping sample sizes for sequential racing.

Seed	Mean Δ ROUGE-L	One-sided 95% LCB	Two-sided 95% CI
126	+0.1424	+0.0028	[−0.0239, +0.3068]
127	+0.1913	+0.0670	[+0.0465, +0.3354]
128	+0.2195	+0.1006	[+0.0771, +0.3609]
Hierarchical	+0.1844	+0.1037	[+0.0879, +0.2762]

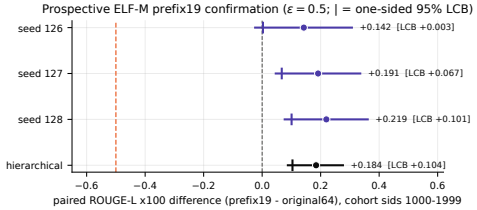

Figure 5: **Left (Table 5):** Prospective validation of candidate `prefix19` on an untouched 1,000-example XSum cohort using the independently trained ELF-M model. Differences are computed as `prefix19` minus `original64`; the hierarchical bootstrap explicitly resamples generation seeds and examples within seed. **Right (Figure 5):** Prospective ELF-M validation results. All seed-level one-sided 95% lower confidence bounds exceed zero, comfortably satisfying the predeclared non-inferiority margin of -0.5 ROUGE-L.

These results confirm that the early-stopping regime identified on XSum transfers cleanly from ELF-B to the larger, independently trained ELF-M checkpoint. The candidate schedule reduces denoiser evaluations from 64 to 19, saving 70.3% of the forward passes (3.37× reduction).

To establish practical latency gains, we benchmarked both policies under a rigorous timing protocol on the same frozen checkpoint and 1,000-example validation slice. Model weights and tokenizers were preloaded prior to timing, an unmeasured warm-up run was executed for each policy, and execution times were recorded across seven interleaved repetitions in a randomized sequence. Explicit GPU synchronizations (`torch.cuda.synchronize()`) were placed immediately before and after the timed generation loops, and metric computation was excluded entirely from the measured interval.

Under this protocol, total runtime per 1,000 examples dropped from 503.51 ± 0.03 seconds for `original64` to 154.49 ± 0.01 seconds for `prefix19`. This corresponds to an empirical wall-clock speedup of 3.26× (bootstrap 95% CI: $[3.259, 3.259]$). The ratio between the empirical speedup and the theoretical evaluation reduction ($3.259/3.368 = 0.968$) indicates that denoiser forward passes account for nearly the entire inference runtime, with fixed tokenization and decoding overheads consuming only 3.2% of total execution time.

8 Multi-Metric Guardrails

Table 9 details all guardrail thresholds and their provenance. For the prospective ELF-M evaluation on XSum, four guardrails were frozen prior to candidate generation: (i) generated token length within $\pm 10\%$ of the canonical schedule; (ii) an empty-output rate increase ≤ 1 percentage point; (iii) a source-copy coverage increase ≤ 0.05 absolute; and (iv) a novel-trigram rate reduction ≤ 0.05 absolute.

Task / Model	Policy	Eval Break-Even	Time Break-Even
XSum / ELF-B	`prefix19`	1,084	1,215
WMT / ELF-B	`H32`	1,772	1,865
OWT / ELF-B	`resamp_K24`	2,592	2,976
XSum / ELF-M	`prefix19`	1,374	1,433

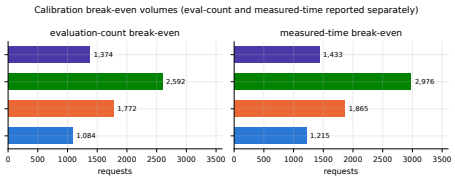

Figure 6: **Left (Table 6):** Deployment requests required to amortize calibration overhead, reported separately for evaluation counts and measured wall-clock runtimes. **Right (Figure 6):** Cumulative net compute savings as a function of served inference requests. Conservative fallbacks yield infinite break-even.

For ELF-B, guardrail thresholds were fixed prior to candidate selection. On OpenWebText, n-gram repetition and entropy thresholds were established after observing degenerate text loops on seed 125 that were artificially masked by low cross-entropy loss.

Candidate `prefix19` passed all four guardrails across all three seeds in the prospective ELF-M trial. On seed 126, for instance, relative length shifted by $+0.05\%$, empty outputs remained at 0%, source copying increased by $+0.0047$, and the novel-trigram rate changed by $+0.0017$. These guardrails evaluate degradation relative to the canonical schedule rather than absolute generation quality. Because they track cohort-level means, they do not capture isolated, example-level anomalies if aggregate statistics match the canonical baseline.

9 Calibration Economics

Table 6 and Figure 6 report break-even deployment volumes computed separately from denoiser evaluation counts and measured wall-clock runtimes. A one-time fixed calibration ($n = 128$) across the candidate library requires 48,768 denoiser evaluations (133.5 GPU-s) on XSum (ELF-B), 56,704 evaluations (21.3 GPU-s) on WMT (ELF-B), 20,736 evaluations (32.0 GPU-s) on OpenWebText (ELF-B), and 61,824 evaluations on ELF-M.

When TCDA certifies an accelerated schedule, this upfront compute amortizes rapidly over deployed queries: break-even is achieved after 1,084 evaluation requests (1,215 timed requests) on XSum with `prefix19`, 1,772 evaluations (1,865 timed) on WMT with `H32`, 2,592 evaluations (2,976 timed) on OpenWebText with `resamp_K24`, and 1,374 evaluations (1,433 timed) on prospective ELF-M with `prefix19`. Wall-clock break-evens are 7% to 15% higher than evaluation-based estimates due to fixed tokenization and decoding overheads. In strict-tolerance regimes where TCDA defaults to the full schedule, inference costs remain unchanged and calibration compute is not recouped (infinite break-even). Model pre-training compute is accounted for separately and excluded from inference accounting.

10 Related Work

Efficient sampling in diffusion and flow models. Sampling from continuous diffusion and flow-matching models requires integrating an ODE or SDE across discrete time steps (Ho et al., 2020; Song et al., 2021; Karras et al., 2022; Lipman et al., 2023; Liu et al., 2022; Sahoo et al., 2024). Prior work compresses this trajectory primarily through step distillation (Salimans and Ho, 2022), optimal discretization grids (Sabour et al., 2024), or key-value caching and parallel decoding (Wu et al., 2025). These approaches typically search for a single improved schedule or retrain the network. In contrast, TCDA leaves model weights untouched and treats schedule reduction as an empirical selection problem, identifying the lowest-cost schedule topology supported for a specific task and checkpoint.

Adaptive compute and early exit. Adaptive computation frameworks dynamically allocate capacity across input tokens or network layers (Moon et al., 2024). While these methods decide when to halt computation on an example-by-example basis during inference, TCDA operates at deployment

time. It evaluates whether a checkpoint consistently stabilizes early on a target task distribution, selecting a fixed, reproducible schedule that bypasses unnecessary downstream integration steps.

Statistical model selection and risk control. Risk-controlling prediction methods provide formal quality guarantees under exchangeability assumptions (Angelopoulos et al., 2021). In sequential testing and multi-armed bandits, racing algorithms eliminate suboptimal configurations as evidence accumulates (Maron and Moore, 1994). TCDA adapts these principles to denoising trajectories by constructing paired, simultaneous one-sided confidence bounds (Westfall and Young, 1993; Holm, 1979) against the canonical baseline. This guarantees family-wise error control across candidate topologies and ensures a reliable fallback to the standard schedule whenever accelerated candidates fail certification.

11 Discussion and Limitations

Prospective confirmation on ELF-M proved critical: evaluating a single preregistered candidate on untouched data avoids the multi-testing inflations inherent to retrospective candidate search. Deployments seeking certification for more aggressive schedules require larger calibration cohorts; our corrected sequential test provides formal error control for adaptive sample sizes, though its stopping thresholds are conservative ($n = 64$–512).

12 Conclusion

In this work, we formulate inference schedule acceleration for diffusion language models as a confidence-controlled calibration problem. Rather than imposing a single, uniform schedule reduction across diverse workloads, TCDA evaluates candidate schedule topologies against user-declared quality tolerances using a modest calibration budget. It deploys the cheapest statistically certified schedule and reliably falls back to the canonical full schedule when empirical evidence is insufficient.

Our evaluation establishes that the optimal denoising topology is fundamentally task-dependent: early prefix stopping for XSum, state-reusing online continuation for WMT translation, and coarse full-range resampling for OpenWebText. These policies reduce denoiser evaluations by $1.33\times$ to $3.37\times$, yielding measured wall-clock accelerations between $1.30\times$ and $2.91\times$ with held-out violation rates below 0.4%. Furthermore, a prospective confirmation demonstrates that the XSum prefix-stopping regime transfers to an independently trained ELF-M model, maintaining generation quality within a strict 0.5 ROUGE-L margin while delivering a $3.26\times$ empirical speedup. By combining task-specific candidate topologies with paired confidence control, TCDA provides a practical, statistically grounded framework for budget-constrained diffusion inference.

AI Use Statement: AI-assisted coding tools were used under direct author supervision for script implementation, debugging. All research questions, experimental designs were conceived, executed, and verified solely by the authors. No AI system trained models, generated synthetic benchmark data, or made autonomous research decisions.

Reproducibility Statement: The supplementary archives contain: (1) the preregistered ELF-M protocol with its SHA-256 hash; (2) per-example evaluation arrays across three fresh seeds; (3) execution logs, synchronization traces, and economic break-even calculations; (4) evaluation records for the 12-cell grid and selector baselines; and (5) regression tests verifying one-sided max-T directionality, zero-variance bootstrap handling, reproducibility of ELF-M `prefix19`, OpenWebText `resamp_K24` enforcement, and distinct tracking of evaluations versus wall-clock speedups.

Ethics and Data Statement: Experiments rely entirely on public benchmarks (XSum, WMT14 De–En, OpenWebText) and publicly released model weights (T5, GPT-2). No private, proprietary, or human-subject data were used. Pre-training the ELF-M checkpoint required 17.9 wall-clock hours on an $8\times$NVIDIA B200 node (approximately 143.2 GPU hours); this cost is accounted for independently and excluded from inference calibration break-even. Degeneration guardrails monitor relative output stability on evaluated cohorts without claiming universal safety guarantees.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia, editors, *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3302. URL https://aclanthology.org/W14-3302/.

Aaron Gokaslan and Vanya Cohen. OpenWebText corpus. https://Skylion007.github.io/OpenWebTextCorpus/, 2019.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Keya Hu, Linlu Qiu, Yiyang Lu, Hanhong Zhao, Tianhong Li, Yoon Kim, Jacob Andreas, and Kaiming He. ELF: Embedded language flows. *arXiv preprint arXiv:2605.10938*, 2026.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Oded Maron and Andrew W. Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. In *Advances in Neural Information Processing Systems 6*, pages 59–66. Morgan Kaufmann, 1994.

Taehong Moon, Moonseok Choi, EungGu Yun, Jongmin Yoon, Gayoung Lee, Jaewoong Cho, and Juho Lee. A simple early exiting framework for accelerated sampling in diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Maja Popović. chrF: Character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, 2015.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your steps: optimizing sampling schedules in diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander M. Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Peter H. Westfall and S. Stanley Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley, New York, 1993.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Fast-dLLM: Training-free acceleration of diffusion LLM by enabling KV cache and parallel decoding. *arXiv preprint arXiv:2505.22618*, 2025.

A Complete Tolerance Grid

Table 7 below lists all 12 task×tolerance cells of the fixed-n=128 grid (source: `latest_tcda_reviewer_stat_audit/a5/tolerance_grid_complete.csv`, regenerated exactly against the frozen `tcda_recomputed_fixed_calibration.csv` by the audit; the regeneration requires consuming the frozen RNG stream over all six calibration sizes, a reproducibility trap documented for future re-runs). Fixed-split held-out paired differences with 95% bootstrap intervals are in `a5/fixed_split_heldout_quality.csv` (class B rows). The estimand classes are kept separate throughout: (A) full-data descriptive frontier; (B) repeated calibration/held-out reliability; (C) racing; (D) prospective ELF-M held-out confirmation.

B Corrected Sequential Racing

The earlier racing variant was invalid under optional stopping for four documented reasons: one-round stopping at n=16 (via `len(survivors)<=1` rather than the documented two-consecutive-stability rule); elimination of the UNDECIDED state (treating not-proven-safe as unsafe); a fixed $\alpha = 0.05$ at every stage with no spending; and no upper bound (it could never prove infeasibility). Its "16–32 calibration examples" conclusion is retired. The corrected procedure uses nested calibration sizes (each stage a prefix of the next), equal-spend $\alpha_j = \alpha_{\text{total}}/J = 0.05/6$ (Bonferroni union bound; valid under optional stopping, conservative), both one-sided lower and upper simultaneous bounds, and a three-state decision. Corrected results: median stopping size 64–512 across task/tolerance cells; violation at most 0.4% across 500 nested permutations per cell; match-to-reference up to 1.0 for large-effect cells and deliberately conservative for OWT 10–20%-PPL (conservative-unresolved rate 0.82–0.84, which correctly avoids the degenerate resamp_K16 policy). Source: `tcda_racing_corrected_summary.csv`, `tcda_racing_alpha_spending.json`, `RACING_CORRECTNESS_AUDIT.md`.

C Statistical Implementation Details

Estimand. For candidate π, $\Delta_i(\pi) = u_i(\pi) - u_i(\text{full})$; feasibility means $\mathbb{E}[\Delta] \geq -\varepsilon$. **Test statistic.** One-sided studentized Westfall–Young max-T: per bootstrap replicate b, $t^*_{b,\pi} = d^*_{b,\pi}/\text{se}^*_{b,\pi}$, and the critical value is the $(1-\alpha)$-quantile of $\max_\pi + t^*_{b,\pi}$ — the maximum of *positive* studentized deviations only. The max-$|t|$ (two-sided) variant is the implementation error documented in Appendix D; the two are not interchangeable. **Degenerate denominators.** If a bootstrap resample has zero variance for some candidate, its t^* is undefined; the predeclared policy treats that replicate's statistic as $+\infty$ (it votes for a larger critical value). It is never silently dropped. With the ELF-M calibration data's degeneracy this makes the one-sided critical value unbounded, so certification abstains rather than resolving by artifact. **Intervals.** Lower (feasibility) and upper bounds are each one-sided simultaneous bounds at level $1 - \alpha$; two-sided joint intervals are $1 - 2\alpha = 0.90$. **Cross-check.** Holm-corrected one-sided paired tests; on the discrete ELF-M data the bootstrap p-values under step-down are anti-conservative (measured FWER 0.0775), so Holm is reported as a cross-check with a discreteness caveat, never as a retroactive fallback. **Categories.** Only SUPPORTED_FEASIBLE and NOT_CERTIFIED are operational; the retired "PROVEN_FEASIBLE / UNRESOLVED / proven-infeasible" vocabulary relied on an upper side whose empirical coverage was invalid (Appendix D).

D Max-T/Holm Audit and the Retired H40 Result

This appendix is the only place H40 appears. **H40 is retired as a selected or validated policy.** The audited ELF-M experiment (frozen checkpoint step 36,616, n=128 calibration split seed 20260923) is excluded from the evidence base because:

1. *Two-sided versus one-sided mismatch.* The implementation computed the max-$|t|$ quantile (7.0572) while the hashed preregistration declared one-sided max-T simultaneous lower bounds; the same replicates under the declared statistic give $c = 3.001$ and select prefix19. `EXTERNAL_PROTOCOL_DEVIATIONS.md` recorded "No deviations."
2. *Zero-denominator handling.* 550 of 10,000 bootstrap replicates contained exactly-zero bootstrap SEs (H40: 62; H48: 498) that were silently dropped from the max; counted at their $\pm\infty$ limit, the critical value is unbounded. H40's calibration difference vector had 123/128 exact zeros (5 informative examples), and H40 attained the max studentized deviation in 411/500 top-5% replicates — the criterion's tail was H40's own denominator pathology, and H40 survived it via its own tiny original-sample SE.
3. *Invalid simultaneous upper-side coverage.* The submitted criterion's intervals failed empirical simultaneous coverage (0.8045 vs 0.95 nominal over 2,000 boundary-null replicates), so "UNRESOLVED"/infeasibility labels had no empirical support; the preregistered one-sided method showed FWER 0.048 / coverage 0.937. Holm's FWER on these discrete data was 0.0775 (inflated), and the ELF-M preregistration names Holm a cross-check only — not a retroactive fallback.

Under the implemented two-sided criterion, H40 was the cheapest of two candidates with lower bounds clearing $-\varepsilon$ (H40 -0.2672; H48 -0.2214), with held-out paired difference $+0.0545$ on the seed-124 held-out subset and -0.0029 on the seed-125 confirmation pair. **Permitted reading:** H40 is a descriptive result produced by an implementation that deviated from the preregistered one-sided statistic and suffered degenerate bootstrap denominators. It must not be presented as a validated TCDA outcome, as evidence of a checkpoint-specific topology change, or as a break-even/timing source (its $1.70\times$ timing and 2,576-request break-even are retired). Selector-sensitivity, split-stability, and library-sensitivity evidence for prefix19 is in Appendices F and G and in `latest_tcda_reviewer_stat_audit/` (A1–A7).

E Prospective Confirmation Preregistration

The full preregistration text is reproduced verbatim in the reproducibility packages (`protocols/PREFIX19_CONFIRMATION_PREREGISTRATION.md`); its SHA-256 is `0dd07054...ad516e`, hashed before any generation command of the run executed, and recorded in `CONFIRMATION_PROTOCOL_MANIFEST.json`. It freezes: the single candidate (prefix19)

and baseline (original64); the untouched cohort (XSum validation IDs 1000–1999, n=1000; prior ELF-M artifacts cover only IDs 0–999); fresh seeds 126/127/128 with common random numbers; $\varepsilon = 0.5$ ROUGE-L points; 10,000-resample paired bootstrap with frozen RNG streams (PCG64 seeded from 20260926); the hierarchical seeds-then-examples bootstrap; the four success criteria; the frozen guardrails; and the exact claim semantics. The frozen checkpoint's provenance note: `frozen_checkpoint_sha256.txt` (`9f20d916...`) is a directory-tree hash per the freeze procedure; the plain file SHA-256 is `7ae326c9...` — both were re-verified and both match the artifacts that produced every prior ELF-M result.

F Library Sensitivity

Across five library variants (3, 6, 8, 9, and the full 11-candidate ELF-M library), the submitted two-sided criterion's selection moves (H40; H32/H48 at other floors; prefix19 in one variant), while every other procedure — preregistered one-sided max-T, both fixed-denominator conventions, both non-studentized variants — selects prefix19 in every library (source: `a3/elfm_selector_sensitivity.csv`, `a7/library_sensitivity.csv`). Editing the library therefore cannot repair a methodological defect, and library edits are not a permitted remedy.

G Selector Baselines

Table 8 compares nine selectors across 100 pool splits per task/tolerance at the highlighted tolerances (source: `a6/selector_baseline_comparison.csv`). Headlines: full (no selection) never violates but never accelerates; fixed N/2 and unconstrained calibration-argmax lose quality or violate (unconstrained dev-set selection violates at up to 32.6%); uncorrected mean-within-ε is unsafe; Holm is over-conservative at strict tolerances; TCDA (one-sided max-T) tracks the full-data reference selection at a fraction of the cost. On ELF-M, across 100 pool splits the submitted two-sided criterion's modal choice is prefix19 (31%) with H40 at 23% — the frozen split landed in the minority branch; Holm picks prefix19 at 90%.

H Exact Guardrails

Table 9 is the authoritative list (threshold, direction, source file, SHA-256, mtime, timing class). ELF-M XSum: truly pre-generation (hashed external preregistration). ELF-B XSum/WMT/OWT: fixed before selection but after generation (`tcda_metric_definition.md`, frozen 2026-09-16). OWT `rep_adj`/`decode_entropy`: decisive application post-hoc after the seed-125 anomaly (documented correction). Metric definitions are those of the frozen generation path (`ev_gen_candidates.py::metrics`): ROUGE-L/R-1/R-2 F×100 (`rouge_score`, stemming on); source-copy coverage = |gen∩src|/|gen|; novel-trigram rate = |gen-3g\src-3g|/|gen-3g|; empty flag on blank output.

I Anchor and Async-ELF Negative Results

These are appendix/negative evidence only; they motivate the TCDA design (fixed candidate library; no learned router; no anchor) and never support the statistical guarantees. **Learned anchors (CAT / richer heads / LoRA adapter):** every deployable learned anchor predictor failed the offline cosine gate (best 0.840 vs gate 0.85) and produced no CI-backed generation gain; the failure is a generalization/train–serve skew finding, not under-training. **Task-label-free router:** the adaptive router ablations show no reliable task-label-free routing signal. **Async-ELF H0/H1/H2:** a frozen-ELF + heterogeneous-timestep adapter is *negative* for free-running commitment (H1 negative; H2 inconclusive/negative at full n=128); the hypothesis remains open only as training from scratch with heterogeneous per-token timesteps. Sources: `evidence/appendix_negative/` in the reproducibility package (`ASYNC_ELF_H1_REPORT.md`, `d3_static_token_anchor_ablation.csv`, `learned_anchor_precision_audit.csv`, `adaptive_router_ablation.csv`, `APPENDIX_CLAIM_SCOPE.md`).

Table 4: T1: Tasks, checkpoints, metrics, and tolerances.

task	model	genre	dataset	n_eval_examples	seeds	primary_metric	notes
xsum	ELF-B	summarization	XSum	1000	124	ROUGE-L F x100	base 64-eval schedule; tolerances eps in 0, 0.25, 0.5, 1.0 metric points; highlighted eps=1.0
wmt	ELF-B	translation	WMT14 De-En	3000	124	sentence chrF (BLEU secondary)	base 64-eval schedule; eps in 0, 0.25, 0.5, 1.0; highlighted eps=1.0
owt	ELF-B	language modeling	OpenWebText	1000	124	GPT-2-large ppl on generated text (utility = -NLL)	base 32-eval schedule; sampler=SDE gamma=1.5 cfg=3.0; relative-PPL tolerances 0, 5, 10, 20%; highlighted 20%
elfm_xsum	ELF-M	summarization	XSum	1000 (cohort sids 1000-1999)	126/127/128	ROUGE-L F x100	342,540,036 params; 24L/1056h; frozen ckpt step 36616; prospective confirmation cohort

Table 5: T2: Schedule topology definitions (denoiser evaluations; state reuse for hybrids).

topology	label	definition	denoiser_evals	notes
full	original64 / original32	full step grid of the base sampler (64 evals ELF-B/ELF-M; 32 evals OWT)	64 / 32	decode at final point; always in the library as the fallback
prefix	prefix_K	run the first K evaluations of the canonical grid, then decode the predicted clean latent x_pred at the truncation point	K	state reused; no restart; K in 19, 32, 48 ELF-B XSum / 19,32,48 ELF-M
hybrid	H_K	shared canonical-grid prefix up to step p, then a monotone resampled continuation to t=1; latent and self-conditioning state reused, no restart; all prefix evaluations counted	H	H in 32, 40, 48
resampled	resamp_K	K evaluations drawn spanning the full [t_eps, 1] range from the same logit-normal family	K	K in 16, 24, 32, 40, 48
uniform	uniform_K	evenly spaced K-step full-range control	K	negative control; K in 16, 24, 32, 48

J Retired Claims Register

"ELF-M selects H40 (40 evals) as PROVEN_FEASIBLE"; "ELF-M selects a different schedule topology than ELF-B"; H40/H48 proven-feasible statuses; "simultaneous_crit = 7.0572" as a one-sided value; the old H40 $1.70\times$ timing and 2,576-request break-even; the old ELF-M candidate-status plot; TCDA_MODEL_SIZE_TRANSFER_SUPPORTED wording when it implies full-library selector transfer; the old OWT resamp_K16 selection, break-even 1,296/1,566, and the transcription error "$\sim$2,596"; "16–32 calibration examples"; "oracle" terminology (replaced by full-data reference selection); step ratios presented as wall-clock speedups; XSum-only/WMT-deferred headline framings; any task-agnostic, zero-shot, distribution-free, new-task, or universal-transfer claim; any formal no-loss guarantee.

K Artifact Traceability

Every table cell traces to a source file and column/key (`tables/TABLE_SOURCE_TRACEABILITY.csv`); figure sources are documented in `figures/FIGURE_SOURCE_NOTES.md`. Key upstream hashes: ELF-B TCDA hardening run `tcda_hardening_20260916_113124`; ELF-B external preregistration SHA-256 `4c88514a...`; ELF-M training preregistration SHA-256 `e667e15d...`; ELF-M external protocol SHA-256 `1d7e2213...`; ELF-M calibration decision SHA-256 `db812fd3...`; prospective confirmation preregistration SHA-256 `0dd07054...`; frozen ELF-M checkpoint tree SHA `9f20d916...` / file SHA `7ae326c9...`. The full statistical audit (`tcda_reviewer_stat_audit_20260925_225345`, verdict MAXT_NUMERIC_OR_IMPLEMENTATION_FAILURE) is bundled with the packages.

L Full Reproducibility Tables

The following machine-generated tables preserve the complete source fields used by the claim and artifact audits. They are intentionally more detailed than the compact main-text tables.

Table 6: T3: ELF-B full-data descriptive frontier (estimand class A; n=1000/3000, seed 124).

task	full_policy	full_evals	full_quality	selected_policy_guardrail_corrected	selected_evals	selected_quality	delta_vs_full_native	one_sided_lcb_vs_full_native	speedup_step	wallclock_full_s_per_1000	wallclock_selected_s_per_1000	speedup_time
xsum	original64	64	28.1026	prefix19	19	28.5016	0.3991	0.1797	3.368421052631578 8	167.51	57.63	2.9066
wmt	original64	64	52.6527	H32	32	52.7056	0.0529	-0.0273	2.0	70.73	36.54	1.9357
owt	original32	32	24.9396	resamp_K24	24	27.9091	-0.1125	held-out fixed-split CI instead (see T4 note)	1.3333	46.24	35.5	1.3025

Table 7: T4: Complete fixed n=128 tolerance grid (12 combos, 100 splits each; estimand class B).

task	tolerance	full_data_reference_selection	ref_cost_evals	final_guardrail_corrected	modal_selected_n128	modal_freq	full_fallback_freq	mean_selected_evals	mean_eval_reduction	match_to_reference_freq	heldout_violation_rate	heldout_violation_count	n_splits	heldout_loss_q95	fixed_split_heldout_diff_ci95
xsum	0.0	prefix19	19	prefix19	original64	0.68	0.68	51.28	0.1987	0.18	0.01	1	100	0.0	0.0 [0.0, 0.0]
xsum	0.25	prefix19	19	prefix19	prefix19	0.56	0.08	27.74	0.5666	0.56	0.0	0	100	0.0	0.0 [0.0, 0.0]
xsum	0.5	prefix19	19	prefix19	prefix19	0.86	0.01	21.44	0.665	0.86	0.0	0	100	-0.0834	0.4385 [0.2478, 0.6263]
xsum	1.0	prefix19	19	prefix19	prefix19	0.96	0.0	19.7	0.6922	0.96	0.0	0	100	-0.3316	0.4385 [0.2478, 0.6263]
wmt	0.0	original64	64	original64	original64	0.98	0.98	63.36	0.01	0.98	0.0	0	100	0.0	0.0 [0.0, 0.0]
wmt	0.25	H32	32	H32	original64	0.41	0.41	47.84	0.2525	0.35	0.0	0	100	0.0076	-0.0123 [-0.0599, 0.0351]
wmt	0.5	H32	32	H32	H32	0.66	0.15	38.64	0.3962	0.66	0.01	1	100	0.0021	0.0468 [-0.0203, 0.1121]
wmt	1.0	H32	32	H32	H32	0.93	0.01	32.72	0.4888	0.93	0.0	0	100	-0.0027	0.0468 [-0.0203, 0.1121]
owt	0%PPL	original32	32	resamp_K24	original32	1.0	1.0	32.0	0.0	1.0	0.0	0	100	0.0	0.0 [0.0, 0.0]
owt	5%PPL	original32	32	resamp_K24	original32	1.0	1.0	32.0	0.0	1.0	0.0	0	100	0.0	0.0 [0.0, 0.0]
owt	10%PPL	original32	32	resamp_K24	original32	1.0	1.0	32.0	0.0	1.0	0.0	0	100	0.0	0.0 [0.0, 0.0]
owt	20%PPL	resamp_K16	16	resamp_K24	resamp_K24	0.9	0.04	23.84	0.255	0.06	0.0	0	100	0.1546	-0.1125 [-0.1264, -0.0982]

Table 8: T5: Selector-baseline comparison at the highlighted tolerances (100 pool splits; estimand class B).

task	tolerance	selector	modal_selected	modal_freq	mean_evals	mean_eval_reduction_vs_full	mean_heldout_util_diff	violation_rate	fallback_rate	final_guardrail_corrected_policy
xsum	1.0	full (no selection)	original64	1.0	64.0	0.0	0.0	0.0	1.0	prefix19
xsum	1.0	fixed N/2	prefix32	1.0	32.0	0.5	0.3005	0.0	0.0	prefix19
xsum	1.0	manual task-specific	prefix19	1.0	19.0	0.7031	0.4052	0.0	0.0	prefix19
xsum	1.0	unconstrained calibration argmax (max cal mean)	prefix19	0.5	25.62	0.5997	0.2835	0.0	0.0	prefix19
xsum	1.0	uncorrected mean-within-epsilon	prefix19	1.0	19.0	0.7031	0.4052	0.0	0.0	prefix19
xsum	1.0	Holm-corrected one-sided	prefix19	1.0	19.0	0.7031	0.4052	0.0	0.0	prefix19
xsum	1.0	TCDA (one-sided max-T)	prefix19	0.97	19.27	0.6989	0.4	0.0	0.0	prefix19
xsum	1.0	fixed-denominator max-T	prefix19	1.0	19.0	0.7031	0.4052	0.0	0.0	prefix19
xsum	1.0	full-data reference selection	prefix19	1.0	19.0	0.7031	0.4052	0.0	0.0	prefix19
wmt	1.0	full (no selection)	original64	1.0	64.0	0.0	0.0	0.0	1.0	H32
wmt	1.0	fixed N/2	prefix19	1.0	19.0	0.7031	-35.4833	1.0	0.0	H32
wmt	1.0	manual task-specific	H32	1.0	32.0	0.5	0.0528	0.0	0.0	H32
wmt	1.0	unconstrained calibration argmax (max cal mean)	H32	0.52	37.84	0.4087	-0.0337	0.0	0.0	H32
wmt	1.0	uncorrected mean-within-epsilon	H32	1.0	32.0	0.5	0.0528	0.0	0.0	H32
wmt	1.0	Holm-corrected one-sided	H32	1.0	32.0	0.5	0.0528	0.0	0.0	H32
wmt	1.0	TCDA (one-sided max-T)	H32	0.98	32.64	0.49	0.0515	0.0	0.02	H32
wmt	1.0	fixed-denominator max-T	H32	1.0	32.0	0.5	0.0528	0.0	0.0	H32
wmt	1.0	full-data reference selection	H32	1.0	32.0	0.5	0.0528	0.0	0.0	H32
owt	20%PPL	full (no selection)	original32	1.0	32.0	0.0	0.0	0.0	1.0	resamp_K24
owt	20%PPL	fixed N/2	prefix_K16	1.0	16.0	0.5	-1.0358	1.0	0.0	resamp_K24
owt	20%PPL	manual task-specific	resamp_K24	1.0	24.0	0.25	-0.1122	0.0	0.0	resamp_K24
owt	20%PPL	unconstrained calibration argmax (max cal mean)	resamp_K24	0.81	22.48	0.2975	-0.1198	0.0	0.0	resamp_K24
owt	20%PPL	uncorrected mean-within-epsilon	resamp_K16	0.77	17.84	0.4425	-0.137	0.0	0.0	resamp_K24
owt	20%PPL	Holm-corrected one-sided	resamp_K24	0.9	23.84	0.255	-0.1105	0.0	0.04	resamp_K24
owt	20%PPL	TCDA (one-sided max-T)	resamp_K24	0.9	24.0	0.25	-0.109	0.0	0.05	resamp_K24
owt	20%PPL	fixed-denominator max-T	resamp_K24	0.91	23.92	0.2525	-0.1101	0.0	0.04	resamp_K24
owt	20%PPL	full-data reference selection	resamp_K16	1.0	16.0	0.5	-0.1426	0.0	0.0	resamp_K24

Table 9: T6: Guardrail thresholds and provenance.

task	metric	direction	threshold	timing_class	note
elfm_xsum	gen_len	within 10% of full	/len_p - len_full/ ¡= 0.10*len_full	truly fixed before candidate generation	hashed preregistration, frozen before reduced-schedule generation (A0)
elfm_xsum	empty	increase ¡= 1pp	inc ¡= 0.01	truly fixed before candidate generation	hashed preregistration, frozen before reduced-schedule generation (A0)
elfm_xsum	source_copy	increase ¡= 0.05	inc ¡= 0.05	truly fixed before candidate generation	hashed preregistration, frozen before reduced-schedule generation (A0)
elfm_xsum	novel_trigram	drop ¡= 0.05	drop ¡= 0.05	truly fixed before candidate generation	hashed preregistration, frozen before reduced-schedule generation (A0)
xsum	gen_len	relative deviation ¡= 10%	/d/ffull/ ¡= 0.10	fixed before selection but after generation	tcda_metric_definition.md frozen 2026-09-16; XSum/WMT/OWT generated 2026-09-15 (tcda_20260915_130737)
xsum	empty	increase ¡= 1pp	inc ¡= 0.01	fixed before selection but after generation	
xsum	source_copy	increase ¡= 0.05	inc ¡= 0.05	fixed before selection but after generation	
xsum	novel_trigram	drop ¡= 0.05	drop ¡= 0.05	fixed before selection but after generation	
wmt	len_ratio	absolute deviation ¡= 0.10	/d/ ¡= 0.10	fixed before selection but after generation	
wmt	empty	increase ¡= 1pp	inc ¡= 0.01	fixed before selection but after generation	
owt	gen_length	relative deviation ¡= 10%	/d/ffull ¡= 0.10	fixed before selection but after generation	
owt	distinct2	drop ¡= 0.05	drop ¡= 0.05	fixed before selection but after generation	
owt	rep_adj	increase ¡= 0.05	inc ¡= 0.05	fixed before selection but after generation; DECISIVE APPLICATION POST-HOC	screen existed at hardening time but the K16-¿K24 retirement was triggered by the observed seed-125 anomaly (documented
owt	decode_entropy	increase ¡= 0.5	inc ¡= 0.5	fixed before selection but after generation; DECISIVE APPLICATION POST-HOC	same as rep_adj
owt	empty	increase ¡= 1pp	inc ¡= 0.01	fixed before selection but after generation	

Table 10: T7: Prospective ELF-M prefix19 confirmation (estimand class D; cohort 1000–1999, seeds 126/127/128).

item	value	source_file	source_key
preregistration_sha256	0dd070541ed5dc2b1c85cc81333fc6027649b6cba1df06104e806e9d6cad516e	latest_elfm_prefix19_confirmation/prefix19_confirmation_preregistration.sha256	sha256
seed 126 paired delta (prefix19 - original64)	+0.1424	quality/quality_summary.json	seed_stats[126].mean
seed 126 one-sided 95% LCB	+0.0028	quality/quality_summary.json	seed_stats[126].lcb_one_sided_95
seed 126 two-sided 95% CI	[-0.0239, +0.3068]	quality/quality_summary.json	seed_stats[126].ci_two_sided_95
seed 127 paired delta (prefix19 - original64)	+0.1913	quality/quality_summary.json	seed_stats[127].mean
seed 127 one-sided 95% LCB	+0.0670	quality/quality_summary.json	seed_stats[127].lcb_one_sided_95
seed 127 two-sided 95% CI	[+0.0465, +0.3354]	quality/quality_summary.json	seed_stats[127].ci_two_sided_95
seed 128 paired delta (prefix19 - original64)	+0.2195	quality/quality_summary.json	seed_stats[128].mean
seed 128 one-sided 95% LCB	+0.1006	quality/quality_summary.json	seed_stats[128].lcb_one_sided_95
seed 128 two-sided 95% CI	[+0.0771, +0.3609]	quality/quality_summary.json	seed_stats[128].ci_two_sided_95
hierarchical mean delta	+0.1844	quality/quality_summary.json	hierarchical.mean
hierarchical one-sided 95% LCB	+0.1037	quality/quality_summary.json	hierarchical.lcb_one_sided_95
hierarchical two-sided 95% CI	[+0.0879, +0.2762]	quality/quality_summary.json	hierarchical.ci_two_sided_95
guardrails seed 126	PASS	quality/guardrail_results.csv	row seed=126:guardrail_pass
guardrails seed 127	PASS	quality/guardrail_results.csv	row seed=127:guardrail_pass
guardrails seed 128	PASS	quality/guardrail_results.csv	row seed=128:guardrail_pass
pairing smoke (n=8, seed 126)	PAIRING_PASS	pairing_smoke/pairing_smoke.json	verdict
criterion 1 (per-seed LCB ¿= -0.5)	True	quality/quality_summary.json	criteria.c1_per_seed
criterion 2 (hierarchical LCB ¿= -0.5)	True	quality/quality_summary.json	criteria.c2_hierarchical
decision	PREFIX19_MODEL_SIZE_TRANSFER_CONFIRMED	quality/quality_summary.json	decision

Table 11: T8: Evaluation-count and measured-time economics (reported separately).

task	tolerance	selected	cal_denoiser_evals	cal_gpu_seconds	break_even_requests_eval	break_even_requests_time
xsum	0.0	prefix19	48768	133.47	1083.7	1214.7
xsum	0.25	prefix19	48768	133.47	1083.7	1214.7
xsum	0.5	prefix19	48768	133.47	1083.7	1214.7
xsum	1.0	prefix19	48768	133.47	1083.7	1214.7
wmt	0.0	original64	56704	21.26	inf	inf
wmt	0.25	H32	56704	21.26	1772.0	1865.3
wmt	0.5	H32	56704	21.26	1772.0	1865.3
wmt	1.0	H32	56704	21.26	1772.0	1865.3
owt	0%PPL	original32	20736	31.97	inf (full selected)	inf (full selected)
owt	5%PPL	original32	20736	31.97	inf (full selected)	inf (full selected)
owt	10%PPL	original32	20736	31.97	inf (full selected)	inf (full selected)
owt	20%PPL	resamp_K24	20736	31.97	2592.0	2976.4
elfm_xsum	eps=0.5	prefix19 (prospective)	61824	not itemized (not wall-clock timed)	1373.9	1432.9