Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Feedback Control for Adaptive Language Model Routing Under Non-Stationary Workloads

Harshil Lodhiya

Chief Software Architect, SlicedHealth Inc., Woodstock, Georgia, United States

Corresponding Author: Harshil Lodhiya, SlicedHealth Inc., E-mail: hlodhiya@slicedhealth.com; lodhiyaharshil@gmail.com

| ABSTRACT

Serving a stream of requests across large language models (LLMs) of differing cost and quality is an online allocation problem, usually framed as multi-armed bandits. We frame it as feedback control: a direct-acting Proportional–Integral–Derivative (PID) controller whose setpoint is the running fleet-average performance and whose bounded output adjusts each model's allocation share, with requests routed by weighted sampling over the allocation vector. This work contributes (i) a stability analysis of the closed loop — bounded-input bounded-output behaviour by anti-windup, exponential convergence of the performance estimates via a Lyapunov function, and a persistent-excitation condition guaranteeing recoverability after a regime change; (ii) a closed-form, analysis-grounded automatic tuning rule requiring no per-dataset search; and (iii) an honest head-to-head against static, round-robin, random, epsilon-greedy, UCB1, Thompson sampling, and the non-stationary bandits Sliding-Window UCB and Discounted UCB, on GSM8K with a checkable exact-match reward, reporting inference cost and request latency alongside quality. Under transient drift the auto-tuned controller is statistically tied with the best non-stationary bandit at lower cost; under a persistent regime shift it significantly outperforms both ($p < 0.03$). We further show the integral term helps only under a persistent shift — a proportional controller suffices for transient drift — and evaluate robustness to noisy rewards. The result is an interpretable, stability-analyzed, auto-tuned routing controller competitive with or superior to non-stationary bandits at lower cost.

| KEYWORDS

Feedback control; PID control; Lyapunov stability; non-stationary bandits; language-model routing; cost-aware inference

| ARTICLE INFORMATION

ACCEPTED: 01 June 2026 **PUBLISHED:** 29 July 2026 **DOI:** 10.32996/jcsts.2026.8.8.17

1. Introduction

Production language-model deployments increasingly expose several models — a strong expensive model, a cheaper mid-tier model, perhaps a small local one — and must decide, per request, which to use. The trade-off among answer quality, monetary cost, and latency shifts over time as providers change, models are updated, or endpoints degrade.

The dominant framing is the multi-armed bandit: each model is an arm, reward is answer quality, and algorithms such as UCB1 or Thompson sampling balance exploration and exploitation. Bandits are strong baselines, but their standard forms assume stationary arms; the non-stationary variants — Sliding-Window and Discounted UCB (Garivier & Moulines, 2011) — are the appropriate comparison under drift. We study an alternative, control-theoretic framing in which each model's observed performance is a measured signal, the fleet-average performance is a setpoint, and each model's traffic share is an actuator driven by a PID controller.

Relative to a bandit baseline, our contributions are: (i) a stability analysis of the routing loop, including a persistent-excitation condition (a minimum-allocation floor) that guarantees recoverability after a regime change; (ii) a closed-form auto-tuning rule grounded in that analysis, eliminating per-dataset hyperparameter search; and (iii) an honest, cost- and latency-aware

benchmark against seven baselines including non-stationary bandits, in which the auto-tuned controller ties the best bandit under transient drift and significantly beats it under a persistent shift, at lower cost. We also report a negative result: the integral and derivative terms are unnecessary under transient drift (the tuned controller is effectively proportional), and the integral earns its place only under a persistent shift.

2. Literature Review

Language-model routing and cascades. FrugalGPT (Chen et al., 2023) cascades cheap-to-expensive models with a learned stopping rule; RouteLLM (Ong et al., 2024) learns a binary router between a strong and weak model from preference data; Hybrid-LLM (Ding et al., 2024) routes queries by predicted difficulty. These learn a routing policy offline from data; we instead adapt allocation online with a controller and no training.

Multi-armed bandits. UCB1 (Auer et al., 2002) and Thompson sampling (Thompson, 1933; Chapelle & Li, 2011) assume stationarity; the Sliding-Window and Discounted UCB policies (Garivier & Moulines, 2011) target non-stationarity and are our key baselines under drift.

Feedback control for computing systems. PID control (Åström & Hägglund, 2006) is standard for autoscaling and resource allocation. We apply it to language-model traffic allocation with a deliberately non-standard, direct-acting sign so that better performers gain share, and we analyze the resulting closed loop.

3. Methodology

3.1 Setup

With N models, at decision step t the router holds an allocation vector $a(t) \in \Delta^{N-1}$ and routes the request by sampling a model $m \sim \text{Categorical}(a(t))$. After observing reward $r \in [0,1]$ (answer quality) for the chosen model, it updates a.

3.2 Direct-acting PID controller

Each model has a controller. Its measured performance $p_i(t)$ is a rolling mean of the last w observed rewards. The setpoint is the fleet-average performance $s(t) = \frac{1}{N}\sum_j p_j(t)$, and the error is direct-acting, $e_i(t) = p_i(t) - s(t)$, so $e_i > 0$ means model i beats the fleet and should gain share (this inverts the classical setpoint − measured sign). Note $\sum_i e_i(t) = 0$. The controller output is

$$u_i(t) = \text{clip}\left(k_p e_i(t) + k_i \sum_{\tau \le t} e_i(\tau)\,\Delta t + k_d \frac{e_i(t) - e_i(t-1)}{\Delta t},\ -L,\ L\right),$$

with fixed logical timestep $\Delta t = 1$, the integral clipped to $[-I_{\max}, I_{\max}]$ (anti-windup), and output bound L. The bounded output is the allocation delta directly. Allocations are floored and renormalized: $\tilde{a}_i = \max(a_{\min}, a_i^{\text{base}} + u_i)$ and $a_i = \tilde{a}_i / \sum_j \tilde{a}_j$.

3.3 Routing, ablation, and the persistent-excitation floor

Routing is weighted sampling over $a(t)$; the controller is the primary driver of the decision. Freezing $a(t) = a^{\text{base}}$ recovers a static policy (the PID-disabled ablation). The floor $a_{\min} > 0$ guarantees every model is sampled with probability at least $a_{\min}$ each step — a persistent-excitation condition used in the analysis.

3.4 Closed-form auto-tuning

Rather than searching, we set gains from a one-shot system identification of the environment: the window w from the reward variance (a bias/variance/tracking trade-off), the proportional gain k_p so a typical performance gap maps near the bound L, the floor $a_{\min}$ from the persistent-excitation requirement, and $k_i, k_d = 0$ except when targeting a persistent shift.

4. Theoretical Analysis

We analyze the loop under three assumptions. **(A1) Exogenous performance:** a model's expected reward $\mu_i(t)$ is determined by the environment, not by the routing decision — calling a model does not change its quality; routing changes only sampling rates. **(A2) Estimator dynamics:** in expectation the estimate is a first-order filter, $\bar{p}_i(t+1) = \bar{p}_i(t) + \alpha(\mu_i - \bar{p}_i(t))$ with $\alpha = 1/w \in (0,1)$ when model i is sampled. **(A3) Persistent excitation:** $a_{\min} > 0$, so each estimate is refreshed at rate at least $a_{\min}$.

Proposition 1 (Boundedness / BIBO). By output clipping and integral anti-windup, $u_i(t) \in [-L, L]$ for all t and all inputs; after flooring and renormalization $a_i(t) \ge a_{\min}/(1 + NL) > 0$, so $a(t)$ lies in a compact subset of the simplex. The closed loop is thus bounded-input bounded-output stable by construction, with no assumptions.

Proposition 2 (Exponential estimator convergence; Lyapunov). Under (A1)–(A2) with stationary μ, let $V(t) = 1/2 \sum_i (\bar{p}_i(t) - \mu_i)^2$. Along the mean dynamics $V(t+1) = (1-\alpha)^2 V(t)$, so V is a Lyapunov function and $\bar{p}_i \to \mu_i$ globally exponentially at rate $(1-\alpha)^2$. Because μ is exogenous (A1), no choice of gains can destabilize the estimate — there is no algebraic loop through the plant.

Proposition 3 (Allocation convergence). The map from errors to allocation (clip, affine, floor, renormalize) is Lipschitz. Composed with the converging bounded errors of Proposition 2 ($e_i \to e_i^* = \mu_i - \bar{\mu}$), the allocation converges to a bounded fixed point: for the proportional controller it reaches the fixed point in one step; for the full PID a persistent nonzero error drives the integral into anti-windup saturation (Proposition 1), and the output converges to the saturated fixed point.

Proposition 4 (Recoverability under a regime shift). If μ changes at time T, (A3) ensures each model is resampled within expected $1/a_{\min}$ steps, and by (A2) its estimate re-converges with time constant $1/\alpha = w$. Hence the controller detects and re-tracks the shift within $O(w + 1/a_{\min})$ steps. The floor $a_{\min}$ is therefore essential: without persistent excitation a model driven to zero share would never be re-measured and the loss could not be undone.

Gains and scope. Because μ is exogenous, k_p cannot destabilize the loop; it only sharpens the static error-to-allocation map and hastens saturation. The binding design quantities are the estimator time constant w and the excitation floor $a_{\min}$, which motivates the closed-form tuner of Section 3.4. These results hold for the mean/EMA idealization with exogenous performance and bounded excitation; they establish boundedness, exponential estimator convergence, allocation convergence, and recoverability, but not almost-sure convergence of the full stochastic sampling process, regret bounds, or behaviour if (A1) fails. The derivative term is excluded from the guarantees as it amplifies measurement noise.

5. Results and Findings

5.1 Experimental setup

The reward is numeric exact-match on GSM8K (Cobbe et al., 2021) after answer extraction — an objective, checkable signal. We use 300 test items and two OpenAI models, gpt-4o (strong, higher cost) and gpt-4o-mini (cheaper); each model's response to each item is recorded once and replayed offline, so every run is deterministic and free. Baselines are static (equivalent to PID-off), round-robin, random, epsilon-greedy, UCB1, Thompson sampling, Sliding-Window UCB, Discounted UCB, best-single-model, and the PID controller. Scenarios are stationary, transient drift (the strong model degrades over steps 100–200 then recovers), and persistent drift (degradation begins mid-run and does not recover); perturbations are seeded quality flips plus a latency spike. We report means over five seeds with 95% confidence intervals and a two-sided paired-bootstrap significance test. Every tunable baseline (PID and both non-stationary bandits) is tuned by the same protocol — grid search on validation seeds — for fairness; the reported PID additionally uses the closed-form auto-tuner. All reported numbers use disjoint held-out evaluation seeds.

5.2 Stationary and transient drift

In the stationary regime both models are near ceiling, so mean quality ties across strategies (≈0.955–0.963, overlapping intervals; $p(\text{pid vs. ucb1}) = 0.61$); cost is the only differentiator, and adaptive or rotating policies keep ≈99% of the strong model's quality at ≈55% of its cost. Table 1 reports transient drift.

Strategy	Quality	Cost	Latency
PID (auto-tuned)	0.937 ± 0.008	$0.183	3.9 s
Discounted UCB	0.933 ± 0.003	$0.249	3.7 s
Sliding-Window UCB	0.933 ± 0.009	$0.244	3.6 s
UCB1	0.927 ± 0.006	$0.239	3.9 s
Thompson sampling	0.926 ± 0.028	$0.127	4.0 s
Round-robin	0.861 ± 0.006	$0.303	4.8 s
Static (PID off)	0.850 ± 0.026	$0.313	4.9 s
Best single model	0.775 ± 0.027	$0.566	5.8 s

Table 1. Transient drift (held-out seeds). Adaptive methods also reduce latency by routing away from the latency-spiked model.

The auto-tuned controller is nominally highest and cheapest in the top tier; versus the non-stationary bandits the difference is not significant ($p = 0.18$ vs. Discounted UCB, $p = 0.26$ vs. Sliding-Window UCB) — a statistical tie at lower cost. Figure 1 is the cost–quality Pareto and Figure 2 the drift-response time series.

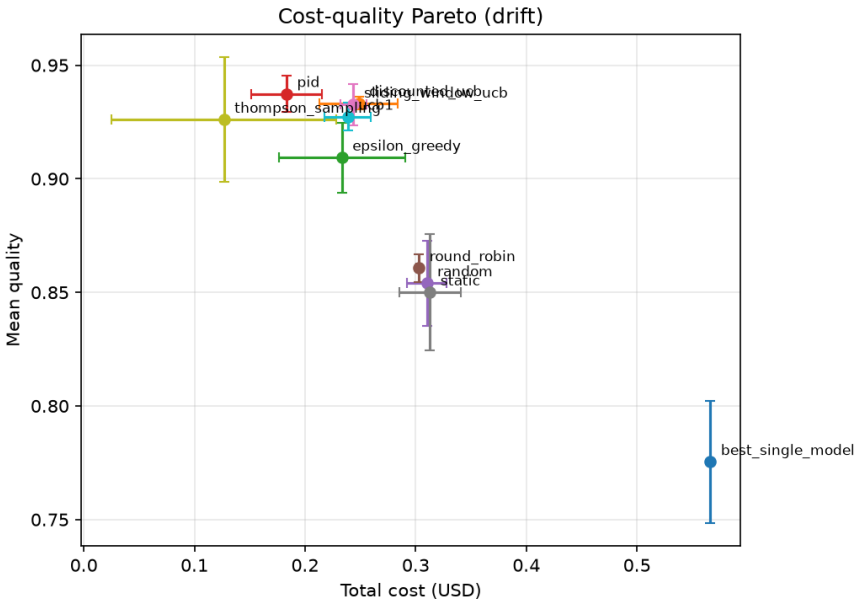

Figure 1. Cost–quality Pareto across strategies under transient drift (means with 95% confidence intervals). Up and to the left is better.

Figure 2. Drift response: quality versus step (moving average), with the drift window shaded. The auto-tuned controller reacts fastest at drift onset and tracks the non-stationary bandits, while non-adaptive baselines sag and the single-model policy collapses.

5.3 Persistent drift

Under a sustained shift (Table 2) the auto-tuned controller significantly outperforms both non-stationary bandits ($p = 0.011$ vs. Discounted UCB, $p = 0.026$ vs. Sliding-Window UCB), tying only the high-variance Thompson sampling, and at lower cost than the UCB variants.

Strategy	Quality	Cost
Thompson sampling	0.923 ± 0.028	$0.126
PID (auto-tuned)	0.923 ± 0.005	$0.151

Strategy	Quality	Cost
Discounted UCB	0.917 ± 0.008	$0.154
Sliding-Window UCB	0.916 ± 0.007	$0.161
UCB1	0.905 ± 0.006	$0.170

Table 2. Persistent drift (held-out seeds).

5.4 Integral term, noise robustness, cost-aware control, and auto-tuning

The auto-tuner selects an effectively proportional controller ($k_i = k_d = 0$) for transient drift; adding an integral term does not help there and, under noisy rewards, hurts (quality 0.87 versus 0.92 at 20% reward-flip noise). Under persistent drift the integral term does help (0.919 versus 0.908 for pure proportional), consistent with Proposition 3. Sliding-Window UCB is the most noise-robust baseline (0.914 at 30% noise), with the proportional controller close behind (0.903). With a cost-aware reward ($r - \lambda$ cost, $\lambda = 0.5$) the controller trades a 0.8-point quality drop for a 78% cost reduction ($0.303 → $0.068), moving deliberately along the cost–quality frontier; UCB1 behaves comparably. Finally, the closed-form auto-tuner, identifying the plant from the recorded responses, selects $k_p = 5.0$, $w = 5$, $a_{\min} = 0.02$ and matches or exceeds the grid-search optimum (transient drift 0.937 versus 0.928; persistent 0.923 versus 0.908) with no per-dataset search.

Limitations. Two models on one task leave both near ceiling in the stationary regime, so the informative differences are under drift. The theory holds under the exogenous-performance and bounded-excitation idealizations and does not give regret or almost-sure-convergence guarantees for the full stochastic system. Thompson sampling remains a strong, cheap baseline that ties the controller under persistent drift. The drift is synthetic (seeded), not an observed provider outage. We do not benchmark against learned language-model routers (e.g., RouteLLM), which require training and a different setup; a confidence-cascade comparison is left to future work.

6. Conclusion

Adaptive language-model routing admits a clean feedback-control formulation that is bounded-input bounded-output stable, has an exponentially convergent performance estimator, and is recoverable after regime changes under a persistent-excitation floor. A closed-form, analysis-grounded auto-tuner removes hyperparameter search and matches grid search. Empirically, the controller ties the best non-stationary bandit under transient drift and significantly beats it under a persistent shift, at lower cost, while remaining interpretable. Future work includes cost-aware setpoints, more and weaker models, learned–control hybrids, and evaluation on real provider latency and quality traces.

Statements and Declarations

Funding: This research received no external funding.
Conflicts of Interest: The author is employed by SlicedHealth Inc. The work uses only the public GSM8K benchmark and public OpenAI model APIs and involves no proprietary or customer data. The author declares no other conflict of interest.
Data Availability: All code, the recorded response cache, tuning, auto-tuning, and analysis that reproduce every number and figure are available in the project repository; a single command regenerates the results and figures deterministically.
ORCID iD: 0009-0005-5798-9176
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Åström, K. J., & Hägglund, T. (2006). *Advanced PID control*. ISA—The Instrumentation, Systems, and Automation Society.
[2] Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning, 47*(2–3), 235–256. https://doi.org/10.1023/A:1013689704352
[3] Chapelle, O., & Li, L. (2011). An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems* (Vol. 24, pp. 2249–2257).
[4] Chen, L., Zaharia, M., & Zou, J. (2023). *FrugalGPT: How to use large language models while reducing cost and improving performance*. arXiv. https://doi.org/10.48550/arXiv.2305.05176
[5] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). *Training verifiers to solve math word problems.* arXiv. https://doi.org/10.48550/arXiv.2110.14168
[6] Ding, D., Mallick, A., Wang, C., Sim, R., Mukherjee, S., Rühle, V., Lakshmanan, L. V. S., & Awadallah, A. H. (2024). Hybrid LLM: Cost-efficient and quality-aware query routing. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[7] Garivier, A., & Moulines, E. (2011). On upper-confidence bound policies for non-stationary bandit problems. In *Algorithmic Learning Theory (ALT)* (pp. 174–188). https://doi.org/10.1007/978-3-642-24412-4_16

[8] Ong, I., Almahairi, A., Wu, V., Zhang, H., Willmott, D., Kim, S., Fox, A., Gonzalez, J. E., & Stoica, I. (2024). *RouteLLM: Learning to route LLMs with preference data*. arXiv. https://doi.org/10.48550/arXiv.2406.18665

[9] Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika, 25*(3–4), 285–294. https://doi.org/10.1093/biomet/25.3-4.285