Stout in Mathematics and Statistics (SMS)

DOI: https://doi.org/10.61424/sq6x3p53
Journal Homepage: https://stoutjournals.org/index.php/SMS

Change-Point-Aware Probabilistic Forecasting of Bursty LLM Demand with Conformal Intervals and Tail-Risk Control

Evan Hu
Computer Science, Virginia Tech, Blacksburg, VA, USA
Corresponding Author: Evan Hu, **E-mail**: ehu_vt_cs@gmail.com

ARTICLE INFORMATION

Submitted: March 15, 2026
Accepted: June 29, 2026
Published: July 31, 2026
Volume: 2
Issue: 1

KEYWORDS

LLM serving; workload forecasting; change-point detection; conformal prediction; quantile regression; conditional value-at-risk; capacity planning; BurstGPT.

ABSTRACT

Short-horizon demand forecasts can support capacity reservation for large language model (LLM) services, but burstiness and regime change make both point accuracy and uncertainty calibration difficult. This study evaluates causal one-step-ahead forecasts on 1,429,737 BurstGPT records aggregated into 17,566 complete five-minute intervals. A chronological design used 12,384 intervals for training, 2,592 for conformal calibration, and 2,590 for held-out testing. Persistence, daily seasonal-naive, ARIMA, histogram gradient boosting (HGB), change-point-augmented HGB (CP-HGB), and a compact causal Transformer were compared. Raw quantiles, split conformalized quantile regression, rolling calibration, adaptive conformal inference (ACI), and a change-point-aware ACI reset (CP-ACI) were evaluated with pinball loss, coverage, width, Winkler score, capacity exceedance, and conditional value-at-risk (CVaR). The CP-HGB conditional median achieved the lowest point error, with an MAE of 17,912 tokens per interval and a 4.29% reduction relative to persistence. The compact Transformer's raw 90% interval covered 83.09% of test outcomes; CP-ACI increased coverage to 90.12%. For a one-sided 95% CP-HGB capacity bound, CP-ACI attained 95.10% nonexceedance and reduced shortfall CVaR from 47,100 under ACI to 45,632 tokens, while increasing mean unused headroom from 59,126 to 60,306 tokens. Change resets were therefore most useful for correcting undercoverage and limiting upper-tail shortfall, whereas CP-HGB's already conservative central interval gained little. The results support joint reporting of point error, calibration, sharpness, and tail risk when token workload is used as a capacity-planning proxy.

1. Introduction

Large language model (LLM) services must match expensive inference capacity to a workload that can change abruptly. Generative requests vary in arrival frequency and in prompt and response length, while scheduling, batching, and key–value-cache management alter the cost of serving them (G.-I. Yu et al., 2022; Kwon et al., 2023). Workload shape therefore affects the attainable throughput–latency operating point even in optimized systems (Agrawal, Kedia, Mohan, et al., 2024; Agrawal, Kedia, Panwar, et al., 2024). Short-horizon forecasts can reserve headroom, but useful forecasts must quantify uncertainty as well as a central trajectory.

Recent research increasingly frames AI-infrastructure demand as a probabilistic decision problem. Calibrated forecasting for heterogeneous GPU clusters, cross-cloud transfer under workload and topology shift, conformal demand envelopes, multi-horizon GPU prediction, and few-shot cold-start forecasting all connect uncertainty to resource risk rather than treating forecast error as an isolated endpoint (He et al., 2023, 2024, 2025; S. Chen et al., 2024; Zhao et al., 2025). The same decision chain extends to hybrid-cloud placement, task-level GPU-memory prediction, profit-aware admission, and power-aware inventory planning (C. Wang et al., 2025; He et al., 2026; Xin, 2025b; Zhao et al., 2026). This literature motivates the present emphasis on calibrated intervals and upper-tail shortfall in addition to point accuracy.

BurstGPT records real GPT submissions, service labels, and request and response token counts (Y. Wang et al., 2025). The analyzed trace contains 1,429,737 records over approximately 61 days. On a complete five-minute grid, 15.39% of intervals are

empty and the positive workload has a long upper tail. The file does not contain GPU utilization, queues, latency, or service-level-objective (SLO) outcomes. Total tokens are therefore treated as an observed workload proxy, and an outcome above a forecast-derived bound is a forecast-capacity exceedance, not an observed SLO violation.

The statistical design joins three elements. ARIMA, histogram gradient boosting, and a compact Transformer represent classical dependence, nonlinear lag interactions, and attention-based context (Hyndman & Khandakar, 2008; Ke et al., 2017; Lim et al., 2021). A causal robust change score describes nonstationarity without supplying future boundaries (Adams & MacKay, 2007; Killick et al., 2012). Quantile regression and conformalized quantile regression construct asymmetric intervals, while adaptive conformal inference updates calibration after observed misses (Koenker & Bassett, 1978; Romano et al., 2019; Gibbs & Candès, 2021, 2024). CVaR measures the severity of the worst capacity shortfalls rather than only their frequency (Rockafellar & Uryasev, 2000, 2002).

Forecast outputs also require an auditable route into operations. Evidence-grounded DevOps retrieval, multi-source root-cause attribution, bilingual incident triage, and source-attributed RAG emphasize that an operational recommendation should remain connected to the observations or documents supporting it (B. Zhang et al., 2024; C. Li et al., 2024; G. Liu et al., 2023, 2024). Capacity-oriented visual briefs, evidence-constrained incident cards, and command-safe DevOps copilots extend that principle from textual answers to human-facing actions (B. Zhang et al., 2026; G. Liu et al., 2025; Nie et al., 2026). Accordingly, the proposed pipeline retains change scores, interval widths, and exceedance losses as evidence for a capacity decision instead of reducing the forecast to an unexplained scalar.

The evaluation maps request starts to 17,566 complete five-minute bins and uses disjoint training, calibration, and test periods. It compares transparent baselines with HGB, change-point-augmented HGB, and a learned causal Transformer; tests raw, static, rolling, adaptive, and change-reset intervals; and evaluates point loss, pinball loss, calibration, sharpness, and interval score. A separate one-sided bound quantifies exceedance, unused headroom, value-at-risk, and CVaR. Request count and separate input- and output-token workloads test whether conclusions depend on the total-token proxy. This design connects empirical forecasting to capacity risk while preserving the boundary between observed workload and unmeasured systems behavior.

2. Literature Review

2.1 LLM serving under bursty and heterogeneous demand

Prediction-serving systems treat batching, model choice, response time, and provisioned resources as joint concerns. Clipper established a general low-latency layer, InferLine joined workload profiles to reactive scaling, and INFaaS automated choices across models and infrastructure (Crankshaw et al., 2017, 2020; Romero et al., 2021). A calibrated forecast complements these reactive mechanisms by reserving headroom before a shift is fully observed.

GPU scheduling studies similarly show that temporal variation controls efficiency. Tiresias and Gandiva use placement, time sharing, and migration for heterogeneous jobs, while fine-grained sharing allocates GPU memory and execution more flexibly (Gu et al., 2019; P. Yu & Chowdhury, 2020; Xiao et al., 2018). AlpaServe applies statistical multiplexing to bursty inference, and SkyPilot expands placement across clouds (Z. Li et al., 2023; Yang et al., 2023). These systems can act on a workload forecast, but tokens do not uniquely determine GPU time because hardware, model architecture, batching, and cache state intervene.

Work on AI capacity forecasting makes that separation explicit. Safe capacity forecasting with time-series foundation models emphasizes calibrated uncertainty in heterogeneous GPU clusters, whereas cross-cloud transfer learning addresses simultaneous workload and topology shift (He et al., 2023, 2024). Multi-horizon GPU forecasting adds workload semantics and operational risk curves, and few-shot cold-start forecasting addresses tenants with limited histories (He et al., 2025; Zhao et al., 2025). At the task level, recurrent and Transformer components have been used to predict GPU-memory requirements rather than assuming a fixed resource profile (C. Wang et al., 2025). These studies support a layered architecture: forecast observable demand, quantify uncertainty, and then map the resulting distribution to hardware through a separately evaluated policy.

That policy layer creates objectives not captured by average forecast error. Conformal demand envelopes target risk-bounded GPU oversubscription, noisy-neighbor residual fusion models degradation in shared virtualized environments, and profit-aware admission control couples workload matching with an asymmetric economic loss (S. Chen et al., 2024; Nie & Zheng, 2024; Zhao et al., 2026). Power-aware inventory planning links job forecasts to acquisition decisions and workload explanations, while hybrid-cloud architecture adds placement and cost choices across environments (He et al., 2026; Xin, 2025b). The present study consequently separates the workload bound from its decision consequences: exceedance frequency and severity measure under-reservation, whereas unused headroom measures over-reservation.

LLM inference sharpens this distinction. Orca and PagedAttention improve iteration scheduling, batching, and key–value-cache use (G.-I. Yu et al., 2022; Kwon et al., 2023). Sarathi-Serve and DistServe expose the different costs of prefill and decoding (Agrawal, Kedia, Panwar, et al., 2024; Y. Zhong et al., 2024). Llumnix mitigates imbalance through migration, ServerlessLLM reduces model-loading delay, and VIDUR shows that preferred configurations depend on request rates and length distributions (Agrawal, Kedia, Mohan, et al., 2024; Fu et al., 2024; B. Sun et al., 2024). BurstGPT contributes a real trace with arrivals and token lengths (Y. Wang et al., 2025). It supports demand forecasting, although system outcomes require measurements absent from the trace.

Nonstationarity is shared by other sequential decision domains, although their findings are not direct evidence about LLM clusters. News and macroeconomic signals have been fused for volatility-direction forecasting, demand forecasts have been combined with marketing-response models for constrained budgeting, and bike-sharing demand has been forecast under changing weather and seasonal regimes (H. Zhou & K. Zhang, 2025; Y. Lu et al., 2025; Xin, 2026f). Digital-twin mobility dispatch adds spatio-temporal prediction and offline control, while attention-based solid-state-drive health classification and cross-dataset wearable transfer emphasize regime-sensitive representation (J. Zhang, 2025; L. Zhang et al., 2025; Wen et al., 2025). Layout-aware progressive PDF rendering provides a related scheduling analogy: limited processing effort is prioritized according to which content should become functional first (H. Wang et al., 2025). Together, these methodological parallels favor chronological evaluation, explicit regime features, and decision-aligned losses over random-fold accuracy alone.

State and transfer assumptions also matter when context evolves. Confidence-gated long-term dialogue memory treats writing, retrieval, updating, and forgetting as separate controls, a useful analogy for deciding how much historical workload should influence a post-change forecast (X. Sun et al., 2023). Approximately shared-feature methods formalize when information can bridge domains, whereas federated preference learning illustrates how shared structure can be learned without centralizing all user histories (Kuo et al., 2025; Z. S. Zhong, Pan, et al., 2025). These works do not determine the calibration window used here; they clarify why a fixed history is an assumption to test rather than a neutral default.

2.2 Probabilistic forecasting and change points

ARIMA offers a parsimonious account of autocorrelation and differencing, whereas histogram tree ensembles efficiently capture nonlinear lag and phase interactions (Hyndman & Khandakar, 2008; Ke et al., 2017). Neural distributional models add another inductive bias: DeepAR learns autoregressive predictive distributions, and attention-based models produce long-context or multi-horizon quantiles (Salinas et al., 2020; Lim et al., 2021; Vaswani et al., 2017; H. Zhou et al., 2021). Quantile regression is attractive for capacity planning because its pinball loss targets asymmetric conditional bounds directly (Koenker & Bassett, 1978). Coverage must still be paired with sharpness and a proper interval score; otherwise, an uninformatively wide interval can appear successful (Gneiting & Raftery, 2007).

Abrupt shifts can make every family stale. PELT provides efficient offline segmentation, reviews catalog a wider range of change costs and search methods, and Bayesian online detection maintains a causal posterior over run length (Killick et al., 2012; Truong et al., 2020; Adams & MacKay, 2007). Forecast evaluation requires the online distinction: a boundary inferred from the completed test sequence cannot be supplied to an earlier origin. All change features and reset decisions in this study therefore use only previously observed workload.

The forecasting problem is part of a broader movement toward decision-aware calibration. Credit-default tiering combines probability calibration with uncertainty-driven rejection, while a model-risk-oriented probability-of-default workflow combines calibration, distribution-free uncertainty, and explanation (Y. Chen et al., 2023; Jin et al., 2024a). Related studies treat extreme class imbalance through cost-sensitive bankruptcy and fraud modeling, use calibrated refusal in recruiter screening, and add counterfactual explanations to fair credit scoring (Y. Chen et al., 2024; Jin et al., 2024b; Jin, 2025a; Xin, 2026b). These endpoints differ from token demand, but their common distinction between uncertain and actionable cases supports reporting calibration, selectivity, and decision cost together.

Multimodal uncertainty research reinforces the difference between a prediction and confidence in that prediction. Uncertainty-aware late fusion calibrates modality-specific evidence, and LiDAR–camera tracking propagates object-level uncertainty through data association and state estimation (Xin, 2025c, 2026d). Risk-calibrated patient-facing safety cards and a companion interface benchmark further show that a probability becomes operational only when uncertainty and response guidance are communicated coherently (B. Zhou et al., 2025; C. Li et al., 2025). These studies are used as design analogies rather than claims about cluster performance: in the present context, the comparable objects are the forecast interval, the selected capacity bound, and the residual risk after that bound is issued.

Forecasts can also feed policy selection instead of direct actuation. LLM-assisted uplift modeling, offline counterfactual evaluation, privacy-robust incrementality estimation, and conservative off-policy selection all distinguish a predictive score from the action chosen under that score (Bai, Wang, et al., 2026; Mu et al., 2023, 2025; Ye et al., 2026). Privacy-safe marketing-mix optimization and constrained budgeting add explicit resource constraints (Bai & Wu, 2026; Y. Lu et al., 2025). Educational early-warning work similarly pairs prediction with an intervention rationale or teacher-facing explanation rather than treating accuracy as the final decision (J. Zhang, 2026; Xin, 2026a). This literature supports a strict boundary for the current experiment: the trace can evaluate a forecast-derived capacity proxy, but realized latency, profit, energy, or educational impact requires outcomes not present in the data.

2.3 Conformal calibration and tail risk

Conformalized quantile regression expands conditional quantiles using held-out nonconformity scores (Romano et al., 2019). Serial dependence and shift complicate the exchangeable case. EnbPI calibrates sequential residuals, ACI updates its miscoverage level under shift, and later methods broaden online guarantees and combine adaptation rates (Xu & Xie, 2021; Gibbs & Candès, 2021, 2024; Zaffran et al., 2022). This evidence motivates matched static, rolling, adaptive, and change-reset intervals, with empirical coverage under the observed chronology as the reported quantity.

Coverage frequency alone ignores severity. CVaR summarizes loss in the upper tail and remains well defined for distributions with a mass at zero (Rockafellar & Uryasev, 2000, 2002). Here the loss is $(y_t - C_t)_+$, where C_t is a one-sided workload bound. CVaR is reported with unused headroom, ensuring that lower tail risk is not detached from the capacity reserved to obtain it.

Conformal control has a close operational analogue in anomaly and alert pipelines. Log anomaly detection with conformal alert control connects empirical error regulation to incident-ticket generation, while self-supervised log models learn event-sequence representations before classification (Xin, 2026e, 2026g). Host-based system-call analysis localizes suspicious windows and preserves forensic narratives, and interpretable CloudTrail staging smooths event sequences without discarding their order (Bai, Chen, et al., 2026; Xin, 2026c). The current change alarm is deliberately narrower: it triggers a recalibration rule and is not interpreted as a diagnosed system fault.

Model complexity must therefore be judged against both the signal and the operational horizon. TinyLLM-assisted intrusion detection, language-guided feature selection for DDoS detection, deep-autoencoder traffic classification, and predictive DDoS mitigation all seek deployable security decisions under computational or data constraints (Y. Li & S. Lu, 2025; S. Lu & Zhou, 2024; B. Wang et al., 2024; Xin et al., 2024). These examples, together with LogBERT-style self-supervision (Xin, 2026g), motivate testing a compact Transformer without presuming that attention must outperform a smaller tree ensemble on a moderate-length, mostly univariate trace.

2.4 Evidence-grounded operations and human oversight

Forecasting becomes operationally useful when it remains connected to diagnosis and evidence. Multi-source root-cause attribution brings CPU and network signals into microservice analysis, while OpsLLM combines retrieval with incident triage and alert summarization (G. Liu et al., 2023, 2024). Evidence-grounded DevOps question answering and a retrieval-augmented DevOps copilot address the related risk that a fluent recommendation may not remain traceable to private runbooks or may propose an unsafe command (B. Zhang et al., 2024, 2026). Retrieval–summary integration for code intelligence and word-level source attribution provide complementary mechanisms for preserving the path from evidence to explanation (C. Li et al., 2024; Y. Li, 2024). In the forecasting pipeline, the analogous provenance chain links observed history, causal change state, base quantiles, conformal scores, and the issued bound.

Retrieval systems have their own allocation and calibration problems. Budgeted multi-hop retrieval treats search depth as a policy, evidence-calibrated RAG couples retrieval coverage with abstention on unanswerable questions, and financial-search reranking combines query rewriting, hybrid retrieval, and listwise relevance (Meng et al., 2026; Su et al., 2025; Z. S. Zhong, Chen, et al., 2025). Machine-learning prediction of retrieval quality and deterministic evidence-chain explanations make retrieval reliability measurable rather than assumed (Jin, 2025b; R. Zhang et al., 2025). Although this study forecasts demand rather than document relevance, its adaptive window follows a comparable discipline: recent evidence is admitted under a stated rule, and the effect of that rule is evaluated separately from the base model.

High-stakes retrieval further emphasizes domain constraints. Accounting-aware evidence retrieval and disclosure-review cards preserve the connection between a financial claim and its source, whereas numerical-reasoning guardrails and evidence-constrained agents impose checks on quantitative and tokenized-asset decisions (Y. Chen et al., 2025; K. Zhang et al., 2025; Z. Li et al., 2026; S. Zhou et al., 2026). Narrative-aware scientific claim verification ranks supporting evidence, trust-calibrated multilingual RAG combines access with abstention, and multi-regulation RAG extends constraint-aware retrieval to cross-border

product counsel (J. Li & Zhou, 2026; Su, Chen, et al., 2026; Y. Chen & Xu, 2026). Across these settings, the transferable design pattern is to state the decision boundary and retain the evidence needed to audit it.

Adversarial inputs make that audit trail still more important. Continual red-teaming treats guardrails as an online system exposed to changing attacks, while behavior-level refusal and evidence-based self-verification seek to preserve utility without accepting unsafe instructions (Zheng et al., 2023; Zheng & Li, 2024; Zheng et al., 2024). Privacy and data-integrity risk cards translate prompt-injection hazards into human-oversight cues (Su, Rao, et al., 2026). These safety mechanisms are not part of the numerical forecasting model, but they matter for a future operator-facing agent that might translate a bound into a placement or scaling action.

Operational evidence is especially useful after an anomaly. Retrieved normal prototypes have been used to explain OpenStack failure-injection logs, retrieval-grounded HDFS analysis connects detected events to deterministic failure narratives, and incident visualization cards turn distributed logs into SRE-facing evidence structures (Nie et al., 2026; X. Sun et al., 2026; Xin, 2025a). Cost-aware AIOps routing spans parsing, detection, diagnosis, and summarization, while trajectory-reliability prediction treats tool-use success as a forecastable outcome (C. Li et al., 2026; Z. S. Zhong et al., 2026). Review-grounded recommendation adds faithfulness evaluation in another domain, reinforcing the distinction between a plausible explanation and one supported by source evidence (Chang et al., 2026). Collectively, this literature supports using a change alarm as contextual evidence for recalibration rather than as proof of a root cause.

Finally, interface design determines whether uncertainty and provenance can inform action. Scientific-search evidence cards organize ranking trust, financial-risk dashboards combine visual hierarchy with explanation, and AI-infrastructure visual briefs translate forecast risk into design decisions (Jin, 2025c; Z. Li et al., 2025; G. Liu et al., 2025). Patient-facing safety-card studies show how calibrated risk can be paired with response guidance, and progressive rendering shows how information can be prioritized under latency constraints (B. Zhou et al., 2025; C. Li et al., 2025; H. Wang et al., 2025). The present tables follow the same principle by displaying coverage, width, exceedance, headroom, and CVaR together: an operator needs the trade-off and its provenance, not a single accuracy rank.

3. Methodology

3.1 Data, preprocessing, and endpoints

The empirical analysis used the 2024 BurstGPT trace described by Y. Wang et al. (2025). The file contained 1,429,737 rows and six complete fields: Timestamp, Model, Request tokens, Response tokens, Total tokens, and Log Type. Timestamps were nondecreasing integer seconds from 5 through 5,269,973 relative to an anonymized origin. The data did not identify an absolute date or timezone. The model field contained ChatGPT and GPT-4, and the service field contained API log and Conversation log. Their four cross-classified record counts were 1,106,998 ChatGPT API, 103,111 ChatGPT conversation, 168,520 GPT-4 API, and 51,108 GPT-4 conversation records. For every row, total tokens equaled request tokens plus response tokens, and the file contained 1,049,263,050 total observed tokens.

Records were aggregated by submission time into nonoverlapping five-minute bins,

$$B_t = [300t, 300(t+1)).$$

The grid began at the documented origin of zero seconds, and intervals without records were retained with zero arrivals and zero token workload. Because the final timestamp fell 173 seconds into its bin, that incomplete bin was excluded. The resulting analysis series comprised 17,566 complete bins indexed from 0 through 17,565. The primary endpoint was the realized total-token workload of the arrival cohort,

$$y_t^{tok} = \sum_{i:T_i \in B_t} (RequestTokens_i + ResponseTokens_i).$$

This endpoint associated eventual output generation with the interval in which its request was submitted; it was not treated as work known at arrival or as a direct measure of GPU time. Logged arrival count,

$$y_t^{req} = \sum_{i:T_i \in B_t} 1,$$

was the mandatory second endpoint. Separate input-token and output-token sums were retained as component sensitivity endpoints because prefill and decode impose different serving workloads (Agrawal, Kedia, Panwar, et al., 2024; Y. Zhong et al., 2024).

The 25,443 records with zero response tokens were retained; the dataset documentation characterizes these records as failures, but their causes cannot be inferred from the fields. Of these, 25,427 had zero request, response, and total tokens, while 16 had positive request tokens and zero response tokens. The file also contained 54,040 exact duplicate rows. Because no request identifier was available, exact equality could represent either repeated logging or distinct simultaneous records with identical attributes. Deleting those rows would impose an unverifiable deduplication rule, so every record was retained. The resulting token identity check was exact for all rows, and no field contained a missing value.

3.2 Chronological design and leakage control

The complete five-minute series was divided chronologically. Bins 0–12,383 formed the 12,384-bin training partition, corresponding to the first 43 days. Bins 12,384–14,975 formed the 2,592-bin calibration partition, corresponding to the next nine days. Bins 14,976–17,565 formed the 2,590-bin held-out test partition, covering the remaining eight days, 23 hours, and 50 minutes. The split therefore allocated approximately 70.5%, 14.8%, and 14.7% of complete intervals to training, calibration, and testing, respectively.

All transformations, scaling constants, lag structures, and change-point thresholds were derived from the training partition. The last 2,016 training intervals served only to select the compact Transformer's stopping epoch; the gradient-boosting and conformal settings were fixed before test evaluation. The calibration partition supplied conformal nonconformity scores and was not used to fit any forecaster. For every origin t, lagged values and rolling summaries were shifted so that predictors used observations available no later than $t-1$. Test outcomes entered an adaptive-conformal update only after the corresponding forecast had been evaluated. The experiment evaluated one-step-ahead, five-minute forecasts with 90% two-sided intervals and separate one-sided 95% capacity bounds.

The strict temporal separation reflects the shifts emphasized by cross-cloud transfer, cold-start tenant forecasting, and regime-dependent mobility demand (He et al., 2024, 2025; Xin, 2026f). Random folds could expose model selection or calibration to future regimes. The design also preserves the distinction drawn by GPU-memory, oversubscription, admission, and inventory studies: a workload predictor is an input to a controller, whereas realized latency, energy, and profit require additional system measurements (C. Wang et al., 2025; He et al., 2026; S. Chen et al., 2024; Zhao et al., 2026).

3.3 Point and quantile forecasters

Two transparent baselines established the value of model complexity. Persistence predicted the previous complete bin, $\hat{y}_t = y_{t-1}$. The daily seasonal-naive method used $\hat{y}_t = y_{t-288}$, because a day contains 288 five-minute bins. ARIMA was fitted to $log(1+y_t)$ as an ARIMA(2,1,1) model by conditional sum of squares. The order and estimator were fixed before test scoring, and all recursive innovations at a forecast origin were computed from observations no later than that origin (Hyndman & Khandakar, 2008).

The nonlinear tree forecaster used histogram gradient boosting (HGB), a computationally compact alternative that follows the discretized-split principle used by efficient boosted-tree systems such as LightGBM (Ke et al., 2017). The implementation used the scikit-learn HGB estimator (Pedregosa et al., 2011). The target was transformed as $log(1+y_t)$, and predictions were mapped back with the inverse transformation and clipped at zero. Causal predictors included lags 1, 2, 3, 6, 12, 24, 72, 144, and 288; rolling means and standard deviations over 3, 12, 72, and 288 prior bins; rolling maxima over 12, 72, and 288 bins; and sine–cosine phase pairs for 288-bin daily and 2,016-bin weekly cycles. Every rolling statistic was shifted by one observation. The HGB configuration used 80 boosting iterations, learning rate 0.07, 31 terminal leaves, minimum leaf size 30, and L_2 regularization 1.0. The same settings were used with squared-error loss for a conditional-mean forecast and with pinball loss at $\tau \in \{0.05, 0.50, 0.95\}$ for quantile forecasts.

The change-point-augmented model, CP-HGB, added four causal regime features. At origin t, a short mean covered the latest 12 logged bins and a long median covered the preceding 276 bins, leaving the same 12-bin separation between the two summaries. Their difference was divided by a robust scale equal to the long-window interquartile range divided by 1.349. The alarm threshold was the 95th percentile of this score in training, 5.067, with an additional absolute log-level-shift requirement of 0.10. Consecutive alarms were separated by at least 288 bins. The score, alarm indicator, ratio of short to long levels, and elapsed bins since the latest alarm were supplied to CP-HGB. This rule produced 22 training, three calibration, and two test alarms. It is a lightweight causal detector: unlike offline PELT segmentation (Killick et al., 2012), it does not use observations after the forecast origin, and unlike full Bayesian online change-point detection (Adams & MacKay, 2007), it does not maintain a posterior over run length.

A compact causal Transformer tested whether attention improved forecasts on a single moderate-length trace. Each sample used the preceding 24 bins of two normalized log targets—request count and total tokens—plus sine–cosine phase features for the hour and day. One learned attention head had query, key, and value dimension 8; its query came from the latest observed position and its keys and values came only from the 24-bin context. A 16-unit rectified-linear feed-forward layer produced the 0.05, 0.50, and 0.95 quantiles for both endpoints. Training minimized mean pinball loss plus a 0.10 quantile-crossing penalty with Adam, batch size 128, learning rate 0.002, weight decay 10^{-5}, and gradient norm limit 5. The last seven days of training selected epoch 24; the model was then refitted for 24 epochs on all eligible training windows. This architecture retained the central causal-attention mechanism of the Transformer while avoiding the parameter scale of long-sequence models (Lim et al., 2021; Vaswani et al., 2017; H. Zhou et al., 2021).

3.4 Conformal interval construction

For base lower and upper quantiles $\hat{q}_{0.05,t}$ and $\hat{q}_{0.95,t}$, conformalized quantile regression used calibration scores

$$s_t = max\{\hat{q}_{0.05,t} - y_t, y_t - \hat{q}_{0.95,t}\}.$$

The static interval expanded the test quantiles by the finite-sample "higher" empirical 90th percentile of all 2,592 calibration scores (Romano et al., 2019). The rolling construction recomputed that quantile from the most recent 2,016 scores, adding each test score only after its outcome became available. ACI additionally updated the working miscoverage level,

$$\alpha_{t+1} = clip\{\alpha_t + \gamma(\alpha - e_t), 0.02, 0.25\},$$

where $e_t = 1$ when the observed outcome fell outside the issued interval, $\alpha = 0.10$, and $\gamma = 0.01$ (Gibbs & Candès, 2021, 2024). CP-ACI used the same update but, at a causal alarm, reset α_t to 0.10 and retained only the latest 288 scores. Comparing ACI with CP-ACI isolated the effect of the change reset from online adaptation itself. Raw quantiles, static CQR, rolling CQR, ACI-CQR, and CP-ACI-CQR were evaluated for both CP-HGB and the compact Transformer.

The one-sided capacity experiment began from each model's conditional 0.95 quantile. Its conformal score was $y_t - \hat{q}_{0.95,t}$. Static, rolling, ACI, and CP-ACI upper bounds used the same chronological logic, with target miscoverage 0.05, ACI step size 0.005, a 2,016-score window, and a 288-score change reset. Negative interval and capacity bounds were truncated at zero. These online updates follow the broader motivation of sequential conformal methods for dependent and shifting series, while empirical test coverage—not an exchangeability claim—remained the operative measure (Xu & Xie, 2021; Zaffran et al., 2022).

The change reset was implemented as a calibration intervention, not a root-cause label. Evidence-grounded AIOps distinguishes a retrieved runbook, prototype, or event sequence from a verified causal diagnosis (B. Zhang et al., 2024; X. Sun et al., 2026; Xin, 2025a). Each forecast therefore retained its base quantiles, score history, working miscoverage level, alarm state, and issued bound. This audit trail mirrors the provenance emphasis of evidence-chain RAG and conformal alert control while remaining specific to the forecasting task (C. Li et al., 2024; Xin, 2026e).

3.5 Forecast and interval metrics

Point forecasts were compared by mean absolute error (MAE), root mean squared error (RMSE), weighted absolute percentage error (WAPE), mean absolute scaled error (MASE), and R^2. MASE divided test MAE by the mean absolute one-step difference across the complete training partition; WAPE divided total absolute error by total observed test workload. These complementary metrics distinguish typical error, sensitivity to large misses, scale-free accuracy, and explained variation.

Quantile accuracy was measured on the original endpoint scale by pinball loss. For quantile probability τ,

$$PB_\tau = \frac{1}{n}\sum_t \left(y_t - \hat{q}_{\tau,t}\right)\left(\tau - 1\{y_t < \hat{q}_{\tau,t}\}\right).$$

Loss was reported for each prespecified quantile and as a mean across the common quantile grid. Lower values indicated more accurate quantile forecasts (Koenker & Bassett, 1978).

For a two-sided interval $[L_t, U_t)$ with nominal coverage $1 - \alpha$, empirical prediction-interval coverage probability was

$$PICP = \frac{1}{n}\sum_t 1\{L_t \le y_t \le U_t\},$$

and absolute coverage error was $|PICP - (1-\alpha)|$. Sharpness was measured by mean prediction-interval width,

$$MPIW = \frac{1}{n}\sum_{t}(U_t - L_t).$$

Normalized width divided MPIW by the training-partition interquantile range $Q_{0.95}^{train}(y) - Q_{0.05}^{train}(y)$; no test statistic entered this denominator. The mean Winkler score combined width with penalties for observations below or above the interval,

$$W_t = (U_t - L_t) + \frac{2}{\alpha}(L_t - y_t)1\{y_t < L_t\} + \frac{2}{\alpha}(y_t - U_t)1\{y_t > U_t\}.$$

Coverage was always interpreted together with width and Winkler score, consistent with the calibration-and-sharpness principle for probabilistic forecasts (Gneiting & Raftery, 2007).

3.6 Capacity exceedance and tail-risk metrics

A separate one-sided nominal 95% conformal upper bound C_t represented planned workload capacity for one interval. Its nonnegative shortfall was

$$\ell_t = (y_t - C_t)_+.$$

We evaluated the exceedance rate, mean shortfall, mean positive shortfall, the number of positive exceedances, and mean overprovisioning $\frac{1}{n}\sum_t (C_t - y_t)_+$. Empirical CVaR at $\beta = 0.95$ followed the Rockafellar–Uryasev representation,

$$\widehat{CVaR}_{\beta}(\ell) = \min_{\eta}\left[\eta + \frac{1}{(1-\beta)n}\sum_{t}(\ell_t - \eta)_+\right],$$

and was accompanied by empirical value-at-risk because the shortfall distribution could contain a large mass at zero (Rockafellar & Uryasev, 2000, 2002). Every tail-risk quantity was measured in the units of the forecast endpoint. The event $y_t > C_t$ was designated a forecast-capacity exceedance; the trace contained no latency, queue, GPU, utilization, or service-rate field from which an observed SLO violation could be derived.

Paired metric differences were calculated at common forecast origins. To respect serial dependence, uncertainty intervals used 2,000 moving-block bootstrap replicates with one-day blocks of 288 five-minute observations. Each comparator's absolute-error series was paired with the CP-HGB conditional median at the same origins. Two-sided bootstrap probabilities were adjusted by Holm's procedure within this family. Effect sizes and confidence intervals were emphasized because the test partition covered only about nine days.

The one-sided analysis can be read as conservative offline evaluation of a fully specified capacity policy proxy. Work on incrementality, counterfactual evaluation, marketing-mix optimization, and off-policy selection similarly separates prediction from the decision produced by it (Bai, Wang, et al., 2026; Bai & Wu, 2026; Mu et al., 2025; Ye et al., 2026). Here, the issued upper bound defines the proxy without estimating an unobserved reward: shortfall measures under-reservation, and unused headroom measures over-reservation.

4. Results and Discussion

4.1 Workload structure and chronological shift

Figure 1 summarizes the strictly chronological analysis. The five-minute grid preserved quiet periods, restricted every predictor to the past, and separated model fitting from conformal calibration and final scoring.

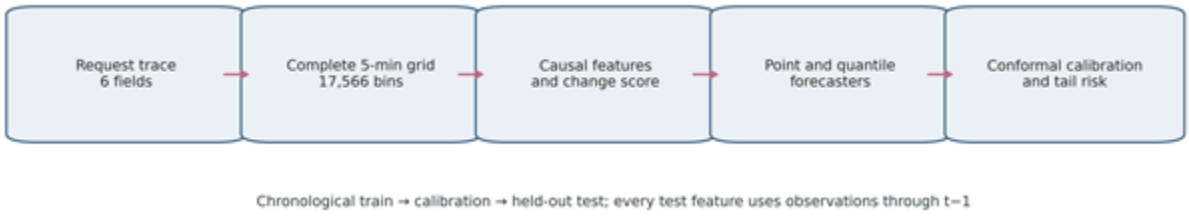

Figure 1. Chronological forecasting, calibration, and risk-evaluation workflow.

The 1,429,737 records represented 1,210,109 ChatGPT and 219,628 GPT-4 requests; 89.21% were API logs and 10.79% were conversation logs. All six fields were complete, all token totals satisfied the component identity, and the final partial bin was the only record-bearing interval excluded from modeling. Table 1 reports the resulting five-minute series. Its mean of 59,733 tokens was 5.39 times the median of 11,077; the 99th percentile was 719,552 and the maximum was 1,931,090. Moreover, 2,704 bins (15.39%) were empty. These values establish both a large point mass at zero and a long upper tail.

Table 1. Descriptive profile of the BurstGPT five-minute workload.

Statistic	Value
Request-level records	1,429,737
Complete five-minute bins	17,566
Empty five-minute bins	2,704
Mean tokens/5 min	59,733
Median tokens/5 min	11,077
P95 tokens/5 min	293,070
P99 tokens/5 min	719,552
Maximum tokens/5 min	1,931,090
Mean requests/5 min	81
Maximum requests/5 min	5,417

Table 2. Composition of the request-level trace.

Dimension	Category	Records	Share of records
Model	ChatGPT	1,210,109	84.64%
Model	GPT-4	219,628	15.36%
Log type	API log	1,275,518	89.21%
Log type	Conversation log	154,219	10.79%
Model × log type	ChatGPT|API log	1,106,998	77.43%
Model × log type	ChatGPT|Conversation log	103,111	7.21%
Model × log type	GPT-4|API log	168,520	11.79%
Model × log type	GPT-4|Conversation log	51,108	3.57%

The chronological partitions differed materially (Table 3). Mean tokens declined from 66,536 in training to 50,364 in calibration and 36,576 in testing, while mean request count rose from 80.99 in training to 97.75 in testing. Average tokens per request consequently fell from approximately 821 to 374. A forecaster therefore had to accommodate a regime with more arrivals but substantially shorter requests. The trace in Figure 2 shows that the shift did not remove isolated spikes, and the log-scale histogram in Figure 3 shows why squared-error summaries alone are inadequate.

Table 3. Chronological training, calibration, and test partitions.

Partition	N	Modeled n	First bin	Last bin	Mean requests	Mean tokens
Training	12,384	12,096	0	12,383	81	66,536
Calibration	2,592	2,592	12,384	14,975	67	50,364
Test	2,590	2,590	14,976	17,565	98	36,576

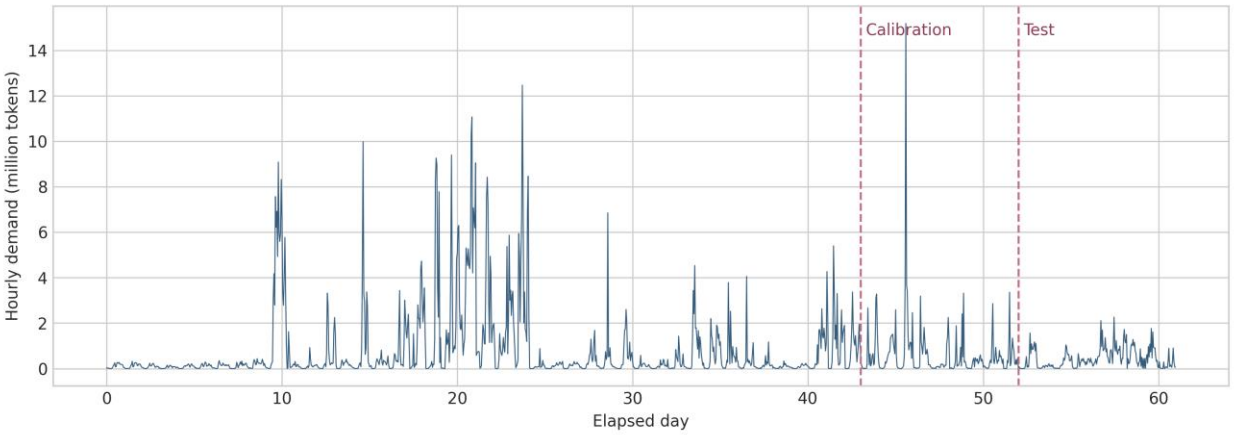

Figure 2. Hourly total-token workload and chronological split boundaries.

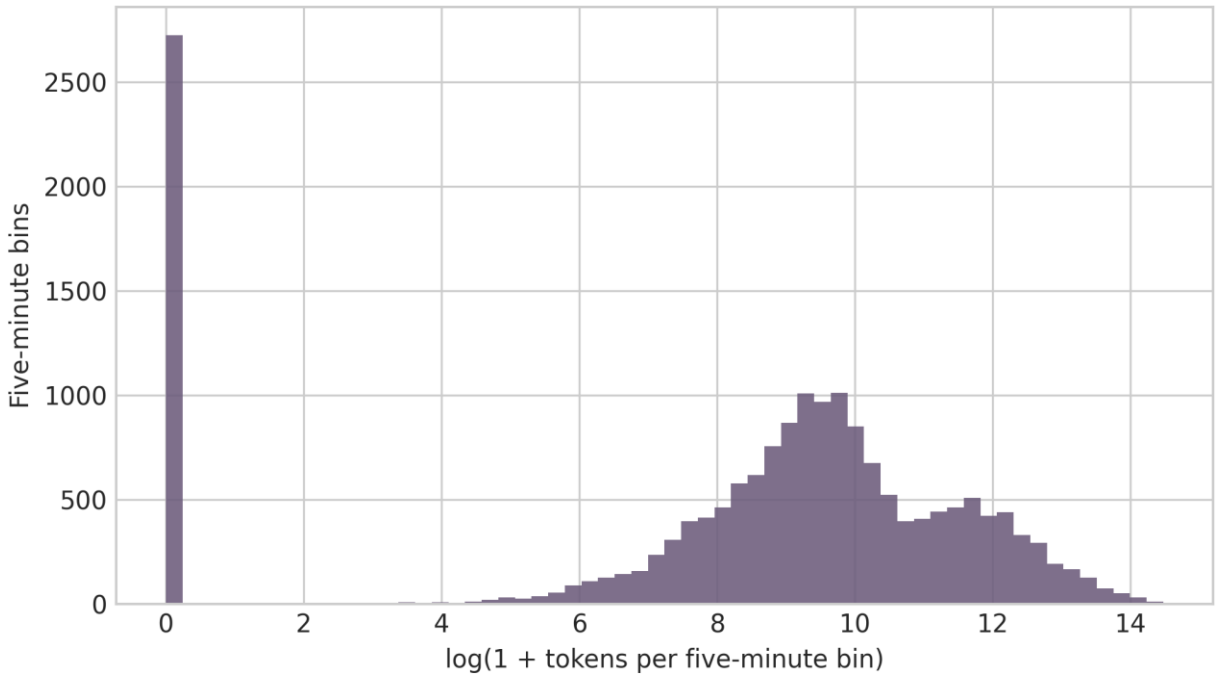

Figure 3. Distribution of log-transformed five-minute total-token workload.

The robust causal detector generated 27 separated alarms: 22 in training, three in calibration, and two in testing. Figure 4 displays its score around the transition into the test period. The detector responded to changes in recent log workload relative to a robust daily reference rather than to the absolute size of a single burst. Because the score at t used values through $t-1$, its alarms could be used as forecast-origin features and calibration resets without retrospective leakage.

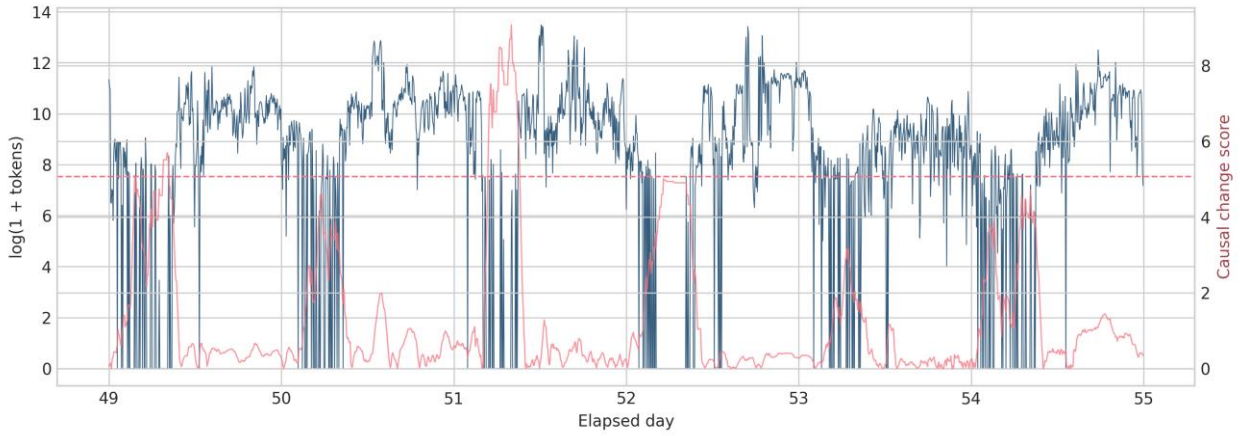

Figure 4. Causal robust change score near the calibration–test boundary.

4.2 Point-forecast comparison

Table 4 records every fixed model setting, and Table 5 reports the held-out point metrics. The CP-HGB conditional median was best on every absolute-error measure: MAE 17,912 tokens, RMSE 41,054, WAPE 0.490, MASE 0.923, and R^2=0.445. Persistence was the strongest simple baseline with MAE 18,715. The CP-HGB median reduced that error by 803 tokens per bin, or 4.29%. It also reduced MAE by 10.63% relative to ARIMA, 5.36% relative to ordinary HGB, and 19.83% relative to the compact Transformer. Daily seasonal-naive forecasting performed poorly (MAE 45,910), consistent with the visible change in daily workload level.

Table 4. Prespecified forecasting and calibration configurations.

Model	Configuration
Persistence	Previous complete five-minute bin
Seasonal naive	Lag 288 (same bin on previous day)
ARIMA	ARIMA(2,1,1), conditional sum of squares on log1p tokens
HGB	80 iterations; learning rate 0.07; 31 leaves; minimum leaf 30
CP-HGB	HGB plus causal robust shift score, alarm, ratio, and run length
Tiny Transformer	24-bin context; one 8-dimensional causal attention head; 16-unit feed-forward layer; q05/q50/q95
CP-ACI-CQR	90% CQR; ACI gamma 0.01; 2,016-bin score window; 288-bin reset history
CP-ACI capacity bound	95% one-sided conformal; ACI gamma 0.005; 2,016-bin score window; 288-bin reset history

Table 5. Held-out five-minute point-forecast accuracy for total tokens.

Model	MAE	RMSE	WAPE	MASE	R2
CP-HGB median	17,912	41,054	0.490	0.923	0.445
Persistence	18,715	42,754	0.512	0.965	0.398
CP-HGB	18,899	41,847	0.517	0.974	0.424
HGB	18,926	42,320	0.517	0.976	0.411
ARIMA(2,1,1)	20,043	44,272	0.548	1.033	0.355
Tiny Transformer	22,343	48,206	0.611	1.152	0.235
Seasonal naive	45,910	79,342	1.255	2.367	-1.072

The paired one-day block bootstrap in Table 6 confirms that these differences were not created by a small number of independently treated bins. Against persistence, the 95% interval for the paired MAE improvement was 451 to 1,255 tokens, with Holm-adjusted $p=.006$. Intervals also remained above zero against ARIMA (425 to 4,146; adjusted $p=.034$), HGB (166 to 2,200; adjusted $p=.034$), CP-HGB's squared-error forecast (280 to 1,987; adjusted $p=.006$), and the Transformer (2,735 to 7,583; adjusted $p=.006$). The gain over the squared-error CP-HGB is especially informative: the conditional median objective contributed more to MAE than adding change features to a conditional-mean tree alone.

Table 6. Paired one-day moving-block bootstrap comparisons against the CP-HGB conditional median.

Comparator	ΔMAE	MAE reduction	95% CI low	95% CI high	Holm p
Persistence	803	4.29%	450.561	1,255.187	0.006
Seasonal naive	27,998	60.98%	19,740.794	33,746.861	0.006
ARIMA(2,1,1)	2,130	10.63%	424.572	4,146.371	0.034
HGB	1,014	5.36%	166.400	2,199.650	0.034
CP-HGB	987	5.22%	280.415	1,986.699	0.006
Tiny Transformer	4,431	19.83%	2,734.746	7,582.899	0.006

Point errors increased after detected changes for every method. During the 576 test origins within one day after a causal alarm, CP-HGB median MAE was 25,419, compared with 15,765 during stable origins. Persistence reached 26,968 and 16,355, respectively. Thus, the CP-HGB median's advantage over persistence was 5.74% in post-alarm periods and 3.61% elsewhere. Change information did not eliminate the harder regime, but its relative value was greater immediately after detected shifts.

Table 7. Point accuracy within one day after causal alarms and during stable test origins.

Model	Regime	N	MAE	RMSE	WAPE	R2
Persistence	Within 1 day of alarm	576	26,968	51,939	0.542	0.298
Persistence	Stable	2,014	16,355	39,739	0.498	0.424
ARIMA(2,1,1)	Within 1 day of alarm	576	27,674	52,259	0.556	0.290
ARIMA(2,1,1)	Stable	2,014	17,860	41,707	0.544	0.366
HGB	Within 1 day of alarm	576	25,309	49,800	0.509	0.355
HGB	Stable	2,014	17,101	39,924	0.521	0.419
CP-HGB median	Within 1 day of alarm	576	25,419	49,779	0.511	0.356
CP-HGB median	Stable	2,014	15,765	38,194	0.481	0.468
Tiny Transformer	Within 1 day of alarm	576	30,497	58,010	0.613	0.125
Tiny Transformer	Stable	2,014	20,011	45,011	0.610	0.262

The compact Transformer's ranking should be interpreted as alignment between data, horizon, and loss rather than a general rejection of neural forecasting. GPU-demand, GPU-memory, bike-sharing, and solid-state-drive studies identify plausible roles for semantic features, recurrent structure, attention, and multi-horizon objectives in richer settings (C. Wang et al., 2025; Wen et al., 2025; Xin, 2026f; Zhao et al., 2025). On this one-step trace, causal lag, phase, and change features supplied a stronger bias for median absolute error. The result supports retaining transparent baselines whenever the endpoint may not reward additional model capacity.

4.3 Quantile forecasts and conformal calibration

The CP-HGB quantile model also outperformed the compact Transformer probabilistically (Table 8). Its mean pinball loss across the 0.05, 0.50, and 0.95 quantiles was 5,338, compared with 6,589 for the Transformer. CP-HGB's raw central interval covered 92.78% of outcomes with mean width 81,019; the Transformer's wider raw interval covered only 83.09%. This contrast shows that width alone does not establish calibration and that the compact neural model did not obtain an advantage merely from attention.

Table 8. Raw quantile accuracy and central-interval behavior.

Model	PB .05	PB .50	PB .95	Mean PB	Coverage	Mean width
CP-HGB	1,829	8,956	5,229	5,338	0.928	81,019
Tiny Transformer	1,849	11,171	6,748	6,589	0.831	99,128

Table 9 separates the calibration methods. For CP-HGB, the calibration score's 90th percentile was zero, so static and rolling CQR left the conservative raw interval unchanged at 92.78% coverage. ACI moved coverage to 91.24% with mean width 80,683 and the lowest Winkler score, 141,108. The CP reset produced 91.66% coverage, width 81,108, and Winkler score 141,152. The reset therefore offered no material benefit to an interval that already overcovered; adaptation without a reset was marginally sharper.
The Transformer presented the opposite case. Static CQR improved its 83.09% raw coverage to 88.69%, rolling CQR reached 89.03%, and ACI reached 89.88%. CP-ACI achieved 90.12%, only 0.12 percentage points from the target, with mean width 107,877 and Winkler score 166,929. Relative to ACI, the change reset added 1.11% width but lowered the Winkler score by 760. Figure 5 illustrates the resulting CP-HGB interval over the first three test days, while Figure 6 gives a broader reliability check across nominal levels for symmetric conformal intervals.

The contrast between the conservative CP-HGB interval and the under-covering Transformer interval also clarifies when a reset is useful. It echoes the distinction between confidence estimation and downstream selection in calibrated credit tiering, model-risk workflows, selective recruiter screening, and uncertainty-aware fusion (Y. Chen et al., 2023; Jin et al., 2024a; Jin, 2025a; Xin, 2025c). CP-ACI helped when recent nonconformity evidence justified adaptation; when the base interval already overcovered, the reset primarily added width. Coverage, sharpness, and tail loss are therefore reported together rather than reduced to a binary calibration judgment.

Table 9. Empirical coverage, sharpness, and interval score for conformal intervals.

Model	Interval	Coverage	Mean width	Abs. cov. error	Norm. width	Winkler
CP-HGB	Raw-quantiles	0.928	81,019	0.028	0.232	141,158.997
CP-HGB	Static-CQR	0.928	81,019	0.028	0.232	141,158.997
CP-HGB	Rolling-CQR	0.928	81,019	0.028	0.232	141,158.997
CP-HGB	ACI-CQR	0.912	80,683	0.012	0.231	141,107.960
CP-HGB	CP-ACI-CQR	0.917	81,108	0.017	0.232	141,151.618
Tiny Transformer	Raw-quantiles	0.831	99,128	0.069	0.284	171,934.529
Tiny Transformer	Static-CQR	0.887	101,219	0.013	0.290	170,703.933
Tiny Transformer	Rolling-CQR	0.890	101,639	0.010	0.291	170,445.441
Tiny Transformer	ACI-CQR	0.899	106,697	0.001	0.306	167,689.072
Tiny Transformer	CP-ACI-CQR	0.901	107,877	0.001	0.309	166,928.987

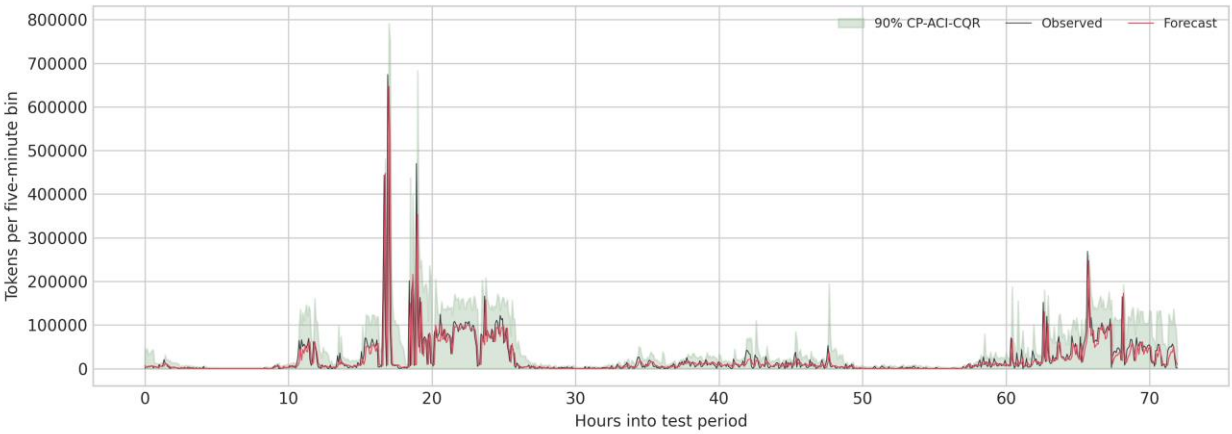

Figure 5. CP-HGB conditional median and 90% CP-ACI-CQR interval over the first three test days.

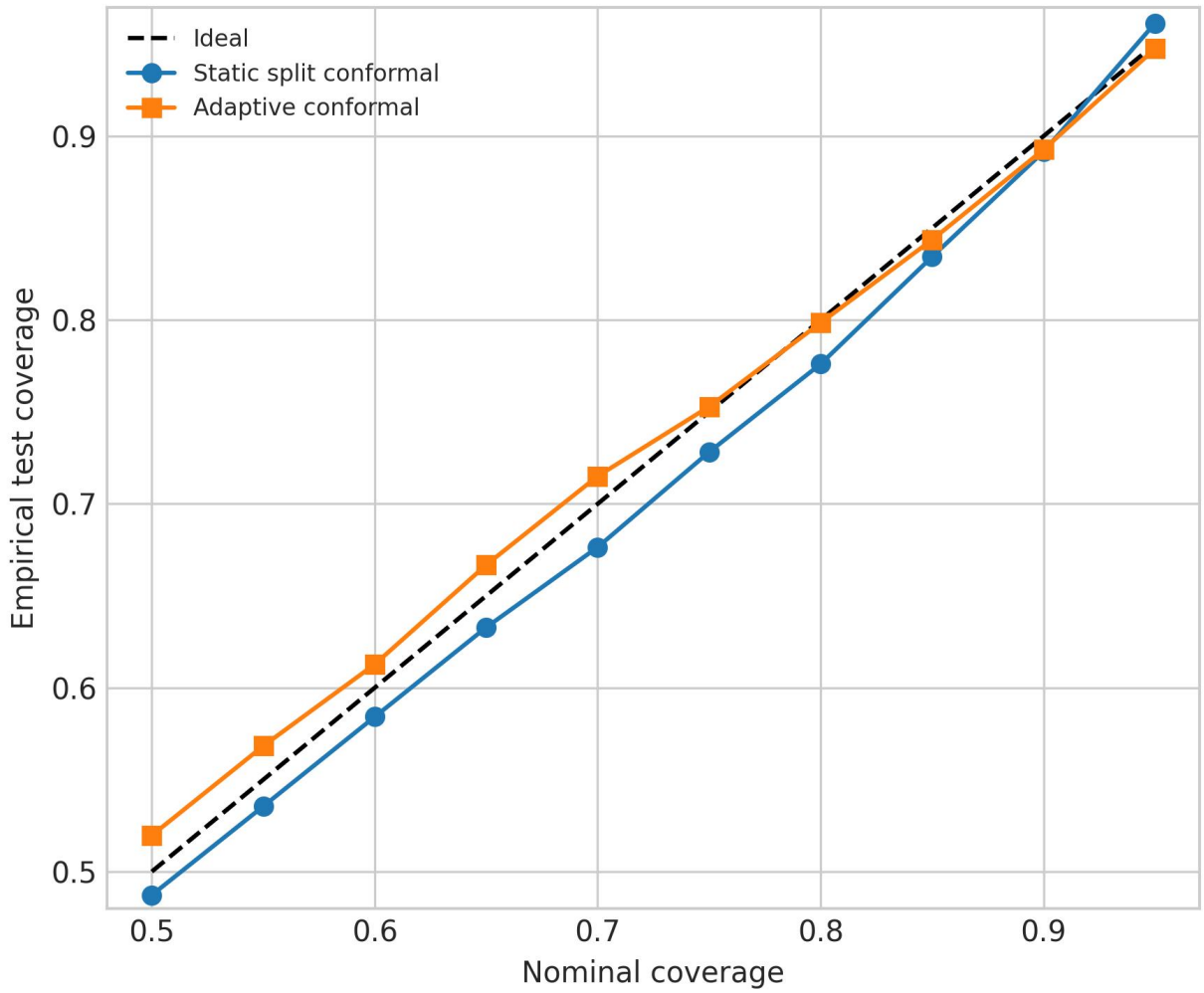

Figure 6. Empirical reliability of static and adaptive symmetric conformal intervals across nominal levels.

4.4 One-sided capacity and tail risk

One-sided bounds exposed the cost of attaining the desired nonexceedance rate (Table 10). CP-HGB's raw 0.95 quantile covered 92.78% of test outcomes, and its static and rolling conformal versions reached 94.25% and 94.86%. ACI reached 95.02% with 129 positive exceedances. CP-ACI reached 95.10% with 127 exceedances, reduced mean shortfall from 2,364 to 2,290 tokens, and reduced the CVaR of the worst 5% of shortfalls from 47,100 to 45,632. That 3.12% CVaR reduction required mean overprovisioning of 60,306 rather than 59,126 tokens, a 2.00% increase in unused headroom.

The Transformer showed the same trade-off more strongly. Its CP-ACI bound reached 95.21% nonexceedance, versus 94.98% under ACI, and reduced shortfall CVaR by 5.68%, from 47,205 to 44,524 tokens. Mean overprovisioning increased by 2.52%, from 88,781 to 91,020 tokens. CP-ACI therefore controlled the upper tail, but it did not create free capacity: the reduction in rare shortfall came from reserving more headroom. Figure 7 reinforces this point by ordering test origins from the narrowest to widest CP-HGB intervals. Error and exceedance risk rose as increasingly uncertain origins were accepted, so interval width can support selective escalation or conservative placement decisions.

Table 10. One-sided 95% workload-capacity bounds and tail-risk metrics.

Model	Bound	Nonexceed.	Exceed. rate	Exceed. n	Mean bound	Unused headroom	Mean shortfall	CVaR95
CP-HGB	Raw-q95	0.928	0.072	187	81,019	47,450	3,007	59,269
CP-HGB	Static	0.942	0.058	149	83,159	49,455	2,872	57,128
CP-HGB	Rolling	0.949	0.051	133	84,684	50,898	2,790	55,581
CP-HGB	ACI	0.950	0.050	129	93,338	59,126	2,364	47,100
CP-HGB	CP-ACI	0.951	0.049	127	94,591	60,306	2,290	45,632
Tiny Transformer	Raw-q95	0.908	0.092	237	108,003	74,604	3,176	61,235
Tiny Transformer	Static	0.944	0.056	146	113,129	79,377	2,824	56,110
Tiny Transformer	Rolling	0.946	0.054	140	113,955	80,152	2,773	55,157
Tiny Transformer	ACI	0.950	0.050	130	122,987	88,781	2,369	47,205
Tiny Transformer	CP-ACI	0.952	0.048	124	125,361	91,020	2,235	44,524

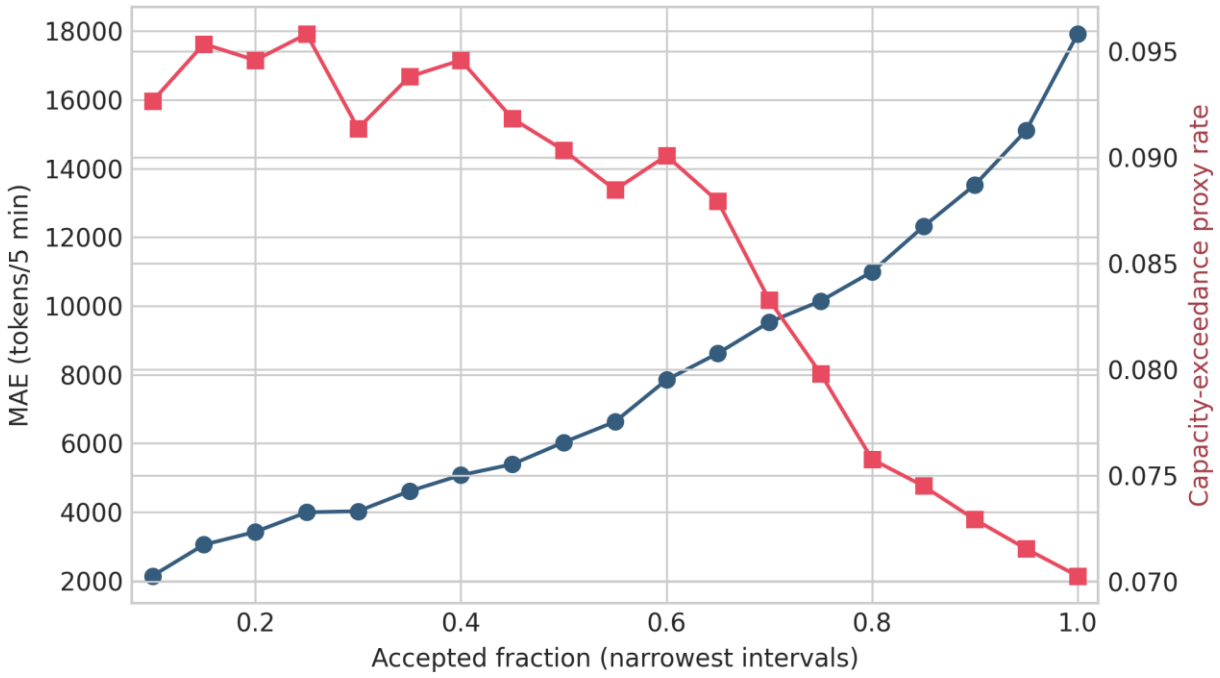

Figure 7. Selective error and capacity-exceedance risk as wider CP-ACI intervals are admitted.

4.5 Endpoint robustness and operational interpretation

The endpoint analysis in Table 11 prevents an overly broad claim about change features. For request count, persistence had lower MAE than squared-error CP-HGB (57.13 versus 58.47). CP-HGB modestly improved output-token MAE (1,295 versus 1,309) and improved input-token RMSE (39,949 versus 41,522) while slightly worsening input-token MAE. For total tokens, the squared-error CP-HGB again trailed persistence in MAE, whereas its conditional median led the full point comparison. Model advantage therefore depended on both endpoint and loss function.

Table 11. Persistence and CP-HGB accuracy across request-count and token components.

Endpoint	Model	MAE	RMSE	WAPE	R2
Request count	Persistence	57	144	0.584	0.421
Request count	CP-HGB	58	145	0.598	0.414
Input tokens	Persistence	17,697	41,522	0.530	0.381
Input tokens	CP-HGB	17,822	39,949	0.534	0.427
Output tokens	Persistence	1,309	1,986	0.408	0.656
Output tokens	CP-HGB	1,295	1,992	0.404	0.653
Total tokens	Persistence	18,715	42,754	0.512	0.398
Total tokens	CP-HGB	18,899	41,847	0.517	0.424

These findings have three operational implications. First, a capacity controller should preserve separate forecasts for arrivals, prompt tokens, and response tokens because their regimes need not move together. This separation is compatible with serving designs that treat prefill and decoding differently (Agrawal, Kedia, Panwar, et al., 2024; Y. Zhong et al., 2024). Second, conformal calibration should be evaluated as a risk–headroom curve, not only by whether coverage is close to a target. Statistical multiplexing, GPU sharing, and cross-cloud placement can act on such curves, but their realized benefit depends on scheduler and hardware details (Z. Li et al., 2023; Xiao et al., 2018; Yang et al., 2023). Third, post-change uncertainty can inform when to reserve additional capacity, delay oversubscription, or pre-stage a cold model, complementing reactive autoscaling and migration mechanisms (Crankshaw et al., 2020; Fu et al., 2024; B. Sun et al., 2024).

The risk–headroom trade-off also needs an interpretable presentation layer. Capacity visual briefs, financial-risk dashboards, and evidence-constrained incident cards organize uncertainty, provenance, and action-relevant context instead of showing a single score (G. Liu et al., 2025; Z. Li et al., 2025; Nie et al., 2026). In this experiment, greater nonexceedance required more unused headroom, so risk reduction was not cost-free. A deployment layer could use the bound for reservation, admission, or hybrid-cloud placement, but the operating point would still depend on hardware, prices, and service objectives not observed here (Xin, 2025b; Zhao et al., 2026).

The scope of the conclusions is bounded by the fields in the trace. The experiment used one anonymized workload, a five-minute one-step horizon, and no exogenous event, price, hardware, queue, latency, or utilization measurements. Total tokens are associated with the request's arrival cohort even though response tokens become known later and may be generated across subsequent intervals. Consequently, forecast-capacity exceedance measures error in a stylized token budget, not an observed latency or SLO event. The short test span also limits the diversity of change regimes. Within those limits, the consistent chronological design shows that a compact tree quantile model can beat a small attention model, and that change-aware calibration is most useful when the base interval undercovers or when upper-tail shortfall is the decision target.

5. Conclusions

BurstGPT's combination of empty intervals, large spikes, and a sharp change in tokens per request makes short-horizon workload prediction a joint forecasting and calibration problem. On 2,590 held-out five-minute intervals, the CP-HGB conditional median delivered the best point accuracy, reducing MAE by 4.29% relative to persistence and 10.63% relative to ARIMA. Raw neural intervals substantially undercovered, but CP-ACI restored the compact Transformer's central coverage to 90.12%. For one-sided CP-HGB capacity bounds, CP-ACI achieved 95.10% nonexceedance and lowered shortfall CVaR by 3.12% relative to ACI, with a 2.00% increase in unused headroom. The central CP-HGB interval was already conservative, so its change reset added width without improving the Winkler score. Change-point-aware conformal updates should therefore be deployed selectively: they are valuable for undercalibrated models and tail-sensitive capacity decisions, while ordinary adaptation is preferable when the base quantiles already cover conservatively. Future system studies can map these workload bounds to measured GPU service rates, queues, and latency objectives; the present results establish the empirical forecasting and risk layer required for that step.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

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

Artificial Intelligence (AI) Use Disclosure: The authors declare that no artificial intelligence tools were used in the preparation of this manuscript.

References

[1] Adams, R. P., & MacKay, D. J. C. (2007). Bayesian online changepoint detection. *arXiv*. https://arxiv.org/abs/0710.3742

[2] Agrawal, A., Kedia, N., Mohan, J., Panwar, A., Kwatra, N., Gulavani, B. S., Ramjee, R., & Tumanov, A. (2024). VIDUR: A large-scale simulation framework for LLM inference. *Proceedings of Machine Learning and Systems, 6*. https://proceedings.mlsys.org/paper_files/paper/2024/hash/b74a8de47d2b3c928360e0a011f48351-Abstract-Conference.html

[3] Agrawal, A., Kedia, N., Panwar, A., Mohan, J., Kwatra, N., Gulavani, B. S., Tumanov, A., & Ramjee, R. (2024). Taming throughput-latency tradeoff in LLM inference with Sarathi-Serve. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)* (pp. 117–134). USENIX Association. https://www.usenix.org/conference/osdi24/presentation/agrawal

[4] Bai, J., Chen, S., Zheng, D., & Kuo, M.-J. (2026). Interpretable attack-chain stage detection from AWS CloudTrail event sequences via linear models and HMM smoothing. *Information Electrical and Electronic Engineering, 6*(1), 28–43. https://doi.org/10.33474/infotron.v6i1.24923

[5] Bai, J., Wang, H., Wu, Q., & Zhang, B. (2026). Privacy-robust incrementality estimation in cookieless settings via uplift modeling: Reproducible evidence from the Hillstrom E-mail experiment. *Journal of Technology Informatics and Engineering, 5*(1), 17–38. https://doi.org/10.51903/jtie.v5i1.468

[6] Bai, J., & Wu, Q. (2026). Privacy-safe marketing mix modeling and budget optimization under identifier loss: A controlled simulation study. *International Journal of Electronic Communication Systems, 6*(1). https://doi.org/10.24042/ijecs.v6i1.30533

[7] Chang, X., Lu, Y., & Zhong, Z. S. (2026). Review-grounded explainable recommendation with faithfulness evaluation on Amazon reviews. *Journal of Electrical Engineering and Computer Science, 11*(1), 9–22. https://doi.org/10.54732/jeecs.v11i1.2

[8] Chen, S., He, S., & Sun, E. (2024). Risk-bounded GPU resource oversubscription via conformal demand envelopes in production AI clusters. *Journal of Advanced Computing Systems, 4*(5), 119–134. https://doi.org/10.69987/JACS.2024.40509

[9] Chen, Y., & Xu, H. (2026). Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access. *International Journal of Graphic Design, 4*(1), 141–164. https://doi.org/10.51903/ijgd.v4i1.3552

[10] Chen, Y., Zhang, Y., Chau, D., & Sherman, M. (2023). Credit card default risk tiering with probability calibration and uncertainty-driven rejection: A reproducible study on the UCI Credit Card Clients dataset. *Journal of Advanced Computing Systems, 3*(4), 31–47. https://doi.org/10.69987/JACS.2023.30403

[11] Chen, Y., Zhang, Y., & Sherman, M. (2024). Going concern and bankruptcy prediction under extreme class imbalance: Cost-sensitive learning, resampling, and focal loss with explainable financial-ratio portraits. *Journal of Advanced Computing Systems, 4*(4), 80–96. https://doi.org/10.69987/JACS.2024.40407

[12] Chen, Y., Zhou, S., & Lin, E. (2025). Accounting-aware evidence retrieval for institutional due diligence of tokenized trade receivable RWA. *Journal of Technology Informatics and Engineering, 4*(3), 649–663. https://doi.org/10.51903/jtie.v4i3.542

[13] Crankshaw, D., Sela, G.-E., Mo, X., Zumar, C., Stoica, I., Gonzalez, J., & Tumanov, A. (2020). InferLine: Latency-aware provisioning and scaling for prediction serving pipelines. In *Proceedings of the 11th ACM Symposium on Cloud Computing* (pp. 477–491). Association for Computing Machinery. https://doi.org/10.1145/3419111.3421285

[14] Crankshaw, D., Wang, X., Zhou, G., Franklin, M. J., Gonzalez, J. E., & Stoica, I. (2017). Clipper: A low-latency online prediction serving system. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)* (pp. 613–627). USENIX Association. https://www.usenix.org/conference/nsdi17/technical-sessions/presentation/crankshaw

[15] Fu, Y., Xue, L., Huang, Y., Brabete, A.-O., Ustiugov, D., Patel, Y., & Mai, L. (2024). ServerlessLLM: Low-latency serverless inference for large language models. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)* (pp. 135–153). USENIX Association. https://www.usenix.org/conference/osdi24/presentation/fu

[16] Gibbs, I., & Candès, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems, 34*, 1660–1672. https://papers.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html

[17] Gibbs, I., & Candès, E. J. (2024). Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research, 25*(162), 1–36. https://jmlr.org/papers/v25/22-1218.html

[18] Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359–378. https://doi.org/10.1198/016214506000001437

[19] Gu, J., Chowdhury, M., Shin, K. G., Zhu, Y., Jeon, M., Qian, J., Liu, H., & Guo, C. (2019). Tiresias: A GPU cluster manager for distributed deep learning. In *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI 19)* (pp. 485–500). USENIX Association. https://www.usenix.org/conference/nsdi19/presentation/gu

[20] He, S., Chang, X., & Sun, E. (2024). Cross-cloud transfer learning for AI training capacity forecasting under workload and topology distribution shift. *Journal of Advanced Computing Systems, 4*(1), 100–120. https://doi.org/10.69987/JACS.2024.40108

[21] He, S., Li, C., & Rao, H. (2025). Few-shot cold-start workload forecasting for new AI inference tenants with time-series foundation models. *Journal of Technology Informatics and Engineering, 4*(1), 306–324. https://doi.org/10.51903/jtie.v4i1.546

[22] He, S., Nie, J., & Li, C. (2026). Power-aware inventory planning for AI infrastructure using job-level forecasting and LLM workload explanations. *Journal of Technology Informatics and Engineering, 5*(1), 341–359. https://doi.org/10.51903/jtie.v5i1.548

[23] He, S., Tu, H., & Liu, I. (2023). Safe PD capacity forecasting with time-series foundation models and calibrated uncertainty for heterogeneous GPU clusters. *Journal of Advanced Computing Systems, 3*(4), 48–66. https://doi.org/10.69987/JACS.2023.30404

[24] Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software, 27*(3), 1–22. https://doi.org/10.18637/jss.v027.i03

[25] Jin, J. (2025a). Calibrated resume-job matching for trustworthy LLM-assisted recruiter screening: Pairwise matching, probability calibration, and selective refusal on two public recruitment datasets. *Journal of Technology Informatics and Engineering, 4*(3), 625–648. https://doi.org/10.51903/jtie.v4i3.529

[26] Jin, J. (2025b). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. *Journal of Technology Informatics and Engineering, 4*(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535

[27] Jin, J. (2025c). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. *International Journal of Graphic Design, 3*(2), 397–414. https://doi.org/10.51903/ijgd.v3i2.3698

[28] Jin, J., Huang, T., & Lu, S. (2024a). A model-risk-friendly probability of default workflow: Calibration, distribution-free uncertainty quantification, and SHAP explanations on the UCI Credit Card Default dataset. *Journal of Advanced Computing Systems, 4*(6), 74–85. https://doi.org/10.69987/JACS.2024.40606

[29] Jin, J., Huang, T., & Lu, S. (2024b). Cost-sensitive learning, simulated PU learning, and one-class autoencoding for extreme-imbalance credit card fraud detection. *Journal of Advanced Computing Systems, 4*(6), 64–73. https://doi.org/10.69987/JACS.2024.40605

[30] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3146–3154. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[31] Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association, 107*(500), 1590–1598. https://doi.org/10.1080/01621459.2012.737745

[32] Koenker, R., & Bassett, G., Jr. (1978). Regression quantiles. *Econometrica, 46*(1), 33–50. https://doi.org/10.2307/1913643

[33] Kuo, M.-J., Zheng, D., & Hires, J. (2025). Federated topic-preference learning for knowledge-grounded chat with differential privacy. *Journal of Technology Informatics and Engineering, 4*(2), 385–401. https://doi.org/10.51903/jtie.v4i2.502

[34] Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., & Stoica, I. (2023). Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles* (pp. 611–626). Association for Computing Machinery. https://doi.org/10.1145/3600006.3613165

[35] Li, C., Bai, J., & Wang, S. (2024). Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM applications. *Journal of Advanced Computing Systems, 4*(2), 76–92. https://doi.org/10.69987/JACS.2024.40207

[36] Li, C., Liu, G., & Zhao, Z. (2026). Cost-aware LLM-style routing for AIOps log analysis: Log parsing, anomaly detection, fault diagnosis, and incident summarization on LogEval task files. *Journal of Technology Informatics and Engineering, 5*(2), 91–103. https://doi.org/10.51903/jtie.v5i2.538

[37] Li, C., Zhou, B., & Gao, K. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX benchmark for explainable medical AI response interfaces. *International Journal of Graphic Design, 3*(2), 381–394. https://doi.org/10.51903/ijgd.v3i2.3709

[38] Li, J., & Zhou, A. (2026). Multi-regulation RAG for AI product counsel: A legal governance framework for cross-border digital commerces. *Rule of Law Studies Journal, 2*(2), 105–123. https://doi.org/10.64780/rolsj.v2i2.225

[39] Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting, 37*(4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

[40] Liu, G., He, S., & Liu, I. (2023). LLM-augmented multi-source root cause attribution for CPU and network faults in microservices. *Journal of Advanced Computing Systems, 3*(6), 39–57. https://doi.org/10.69987/JACS.2023.30604

[41] Liu, G., He, S., & Wong, H. (2025). LLM-compatible visual brief cards for AI infrastructure capacity dashboards: A UI/UX framework for turning forecast risk into graphic design decisions. *International Journal of Graphic Design, 3*(1), 196–213. https://doi.org/10.51903/ijgd.v3i1.3723

[42] Liu, G., Li, C., & Zhang, E. (2024). OpsLLM for cloud incident triage: Bilingual RAG-based root cause analysis and alert summarization for AI infrastructure operations. *Journal of Advanced Computing Systems, 4*(4), 97–111. https://doi.org/10.69987/JACS.2024.40408

[43] Li, Y. (2024). Findable then explainable: Retrieval–summary integration for code intelligence on a lightweight CodeSearchNet subset. *Journal of Advanced Computing Systems, 4*(7), 65–82. https://doi.org/10.69987/JACS.2024.40706

[44] Li, Y., & Lu, S. (2025). Language-guided feature selection for DDoS and intrusion detection on CICIDS2017. *Journal of Technology Informatics and Engineering, 4*(1), 284–305. https://doi.org/10.51903/jtie.v4i1.531

[45] Li, Z., Zhang, K., & Wong, A. (2026). Numerical-reasoning guardrails for a quant research assistant: A compact reproducible benchmark using SEC and FRED data. *Journal of Technology Informatics and Engineering, 5*(2), 75–90. https://doi.org/10.51903/jtie.v5i2.541

[46] Li, Z., Zheng, L., Zhong, Y., Liu, V., Sheng, Y., Jin, X., Huang, Y., Chen, Z., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). AlpaServe: Statistical multiplexing with model parallelism for deep learning serving. In *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)* (pp. 663–679). USENIX Association. https://www.usenix.org/conference/osdi23/presentation/li-zhouhan

[47] Li, Z., Zhou, S., & Zhou, Z. (2025). Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench. *International Journal of Graphic Design, 3*(1), 196–210. https://doi.org/10.51903/ijgd.v3i1.3715

[48] Lu, S., & Zhou, D. (2024). TinyLLM-assisted intrusion detection for real-time IoT networks. *Journal of Advanced Computing Systems, 4*(8), 72–87. https://doi.org/10.69987/JACS.2024.40809

[49] Lu, Y., Zhou, H., & Zhang, Y. (2025). A constrained, data-driven budgeting framework integrating macro demand forecasting and marketing response modeling. *Journal of Technology Informatics and Engineering, 4*(3), 493–520. https://doi.org/10.51903/jtie.v4i3.466

[50] Meng, S., Chen, J., & Zheng, I. (2026). LLM-inspired offline reranking for financial search: Query rewriting, hybrid retrieval, and listwise relevance ranking on FiQA. *Journal of Technology Informatics and Engineering, 5*(1), 361–378. https://doi.org/10.51903/jtie.v5i1.537

[51] Mu, J., Lu, Y., & Smith, M. (2023). LLM-assisted incrementality (uplift) modeling for email advertising: From feature interactions to interpretable audience–creative–channel policies. *Journal of Advanced Computing Systems, 3*(1), 31–48. https://doi.org/10.69987/JACS.2023.30103

[52] Mu, J., Ye, T., & Patel, P. (2025). Offline counterfactual evaluation for advertising and recommendation slot policies: A reproducible study on the Open Bandit Dataset (Small). *Journal of Technology Informatics and Engineering, 4*(3), 521–543. https://doi.org/10.51903/jtie.v4i3.500

[53] Nie, J., Liu, G., Li, C., & Zou, T. (2026). Evidence-constrained incident visualization cards for distributed cloud logs: A UI/UX framework for turning Hadoop, OpenStack, and ZooKeeper logs into actionable SRE design interfaces. *International Journal of Graphic Design, 4*(1), 179–185. https://doi.org/10.51903/ijgd.v4i1.3703

[54] Nie, J., & Zheng, D. (2024). Noisy-neighbor-aware VM degradation risk modeling with unsupervised residual fusion. *Journal of Advanced Computing Systems, 4*(4), 112–123. https://doi.org/10.69987/JACS.2024.40409

[55] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

[56] Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk, 2*(3), 21–42. https://sites.math.washington.edu/~rtr/papers/rtr179-CVaR1.pdf

[57] Rockafellar, R. T., & Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance, 26*(7), 1443–1471. https://doi.org/10.1016/S0378-4266(02)00271-6

[58] Romano, Y., Patterson, E., & Candès, E. (2019). Conformalized quantile regression. *Advances in Neural Information Processing Systems, 32*, 3538–3548. https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html

[59] Romero, F., Li, Q., Yadwadkar, N. J., & Kozyrakis, C. (2021). INFaaS: Automated model-less inference serving. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)* (pp. 397–411). USENIX Association. https://www.usenix.org/conference/atc21/presentation/romero

[60] Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting, 36*(3), 1181–1191. https://doi.org/10.1016/j.ijforecast.2019.07.001

[61] Sun, B., Huang, Z., Zhao, H., Xiao, W., Zhang, X., Li, Y., & Lin, W. (2024). Llumnix: Dynamic scheduling for large language model serving. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)* (pp. 173–191). USENIX Association. https://www.usenix.org/conference/osdi24/presentation/sun-biao

[62] Sun, X., Lu, Y., & Chen, J. (2023). Controllable long-term user memory for multi-session dialogue: Confidence-gated writing, time-aware retrieval-augmented generation, and update/forgetting. *Journal of Advanced Computing Systems, 3*(8), 9–24. https://doi.org/10.69987/JACS.2023.30802

[63] Sun, X., Zhong, Z. S., & Wu, Q. (2026). Retrieval-grounded HDFS log anomaly detection and deterministic failure narrative generation. *Journal of Computer Systems and Applications, 3*(1), 15–30. https://doi.org/10.64229/j6d7fr94

[64] Su, W., Chen, S., & Qian, E. (2026). Narrative-aware scientific claim verification agent with evidence ranking for ClimateCheck. *Journal of Technology Informatics and Engineering, 5*(1), 327–340. https://doi.org/10.51903/jtie.v5i1.549

[65] Su, W., Chen, S., & Zhao, C. (2025). Budgeted multi-hop retrieval agent for compositional question answering: A retrieval-policy evaluation on the official MultiHop-RAG benchmark. *Journal of Technology Informatics and Engineering, 4*(3), 649–662. https://doi.org/10.51903/jtie.v4i3.543

[66] Su, W., Rao, H., & Ma, E. (2026). Privacy and data-integrity risk cards for LLM agents: A UI/UX design framework for secure human oversight under prompt-injection attacks. *International Journal of Graphic Design, 4*(1), 186–191. https://doi.org/10.51903/ijgd.v4i1.3699

[67] Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing, 167*, Article 107299. https://doi.org/10.1016/j.sigpro.2019.107299

[68] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[69] Wang, B., He, Y., Shui, Z., Xin, Q., & Lei, H. (2024). Predictive optimization of DDoS attack mitigation in distributed systems using machine learning. In *Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024)* (pp. 89–94).

[70] Wang, C., Wen, Z., Zhang, R., Xu, P., & Jiang, Y. (2025). GPU memory requirement prediction for deep learning task based on bidirectional gated recurrent unit optimization Transformer. In *2025 5th International Conference on Artificial Intelligence, Virtual Reality and Visualization (AIVRV)*. IEEE. https://doi.org/10.1109/AIVRV67401.2025.11350369

[71] Wang, H., Ren, Y., & Chang, X. (2025). Layout-aware progressive PDF rendering: AI prioritization of PDF slices to reduce time-to-functional-first-frame on FUNSD. *Journal of Technology Informatics and Engineering, 4*(2), 425–446. https://doi.org/10.51903/jtie.v4i2.523

[72] Wang, Y., Chen, Y., Li, Z., Kang, X., Fang, Y., Zhou, Y., Zheng, Y., Tang, Z., He, X., Guo, R., Wang, X., Wang, Q., Zhou, A. C., & Chu, X. (2025). BurstGPT: A real-world workload dataset to optimize LLM serving systems. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (pp. 5831–5841). Association for Computing Machinery. https://doi.org/10.1145/3711896.3737413

[73] Wen, Z., Zhang, R., & Wang, C. (2025). Optimization of bi-directional gated loop cell based on multi-head attention mechanism for SSD health state classification model. In *2025 6th International Conference on Electronic Communication and Artificial Intelligence (ICECAI)*. IEEE. https://doi.org/10.1109/ICECAI66283.2025.11171441

[74] Xiao, W., Bhardwaj, R., Ramjee, R., Sivathanu, M., Kwatra, N., Han, Z., Patel, P., Peng, X., Zhao, H., Zhang, Q., Yang, F., & Zhou, L. (2018). Gandiva: Introspective cluster scheduling for deep learning. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)* (pp. 595–610). USENIX Association. https://www.usenix.org/conference/osdi18/presentation/xiao

[75] Xin, Q. (2025a). Explaining OpenStack failure-injection log anomalies with retrieved normal prototypes. *Emerging Information Science and Technology, 6*(2). https://doi.org/10.18196/eist.v6i2.31232

[76] Xin, Q. (2025b). Hybrid cloud architecture for efficient and cost-effective large language model deployment. *Journal of Information Systems and Informatics, 7*(3), 2182–2195. https://doi.org/10.51519/journalisi.v7i3.1170

[77] Xin, Q. (2025c). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning). *Journal of Technology Informatics and Engineering, 4*(1), 215–238. https://doi.org/10.51903/jtie.v4i1.485

[78] Xin, Q. (2026a). Early-warning analytics with LLM intervention rationales for student retention decisions: Classroom interaction modeling with xAPI-edu-data and dropout/success prediction. *Interdisciplinary Journal of Pedagogical Research and Media Technology, 2*(1). https://doi.org/10.64268/inspire.v2i1.117

[79] Xin, Q. (2026b). Explainable and fair credit risk scoring with counterfactual explanations: A reproducible evaluation on the German Credit dataset (HELOC-motivated). *Journal of Information and Technology, 14*(2). https://doi.org/10.32664/j-intech.v14i02.2228

[80] Xin, Q. (2026c). Host-based intrusion detection with system call sequences: Window localization and forensic narratives. *AVITEC, 8*(2). https://doi.org/10.28989/avitec.v8i2.3973

[81] Xin, Q. (2026d). LiDAR–camera object-level fusion for multi-target tracking using JPDA and EKF: A reproducible empirical study on a PandaSet-parameterised five-sequence dataset. *Journal of Technology Informatics and Engineering, 5*(1), 54–76. https://doi.org/10.51903/jtie.v5i1.486

[82] Xin, Q. (2026e). Log anomaly detection with conformal alert control and evidence-grounded incident ticket generation. *AVITEC, 8*(2), 247. https://doi.org/10.28989/avitec.v8i2.3974

[83] Xin, Q. (2026f). Probabilistic bike-sharing demand forecasting under changing weather and seasonal regimes with transformer-based models. *Findings*. https://doi.org/10.32866/001c.157499

[84] Xin, Q. (2026g). Self-supervised log anomaly detection with LogBERT-style transformers: Full empirical evaluation on a reproducible SynHDFS benchmark. *Journal of Electrical Engineering and Computer Science, 11*(1), 23–35. https://doi.org/10.54732/jeecs.v11i1.3

[85] Xin, Q., Xu, Z., Guo, L., Zhao, F., & Wu, B. (2024). IoT traffic classification and anomaly detection method based on deep autoencoders. In *Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024)*.

[86] Xu, C., & Xie, Y. (2021). Conformal prediction interval for dynamic time-series. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 11559–11569). PMLR. https://proceedings.mlr.press/v139/xu21h.html

[87] Yang, Z., Wu, Z., Luo, M., Chiang, W.-L., Bhardwaj, R., Kwon, W., Zhuang, S., Luan, F. S., Mittal, G., Shenker, S., & Stoica, I. (2023). SkyPilot: An intercloud broker for sky computing. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)* (pp. 437–455). USENIX Association. https://www.usenix.org/conference/nsdi23/presentation/yang-zongheng

[88] Ye, T., Mu, J., & Hunter, J. (2026). Off-policy evaluation and conservative policy selection for slot-level dynamic bidding and ranking on the Open Bandit Dataset (Small). *Journal of Technology Informatics and Engineering, 5*(1), 178–199. https://doi.org/10.51903/jtie.v5i1.503

[89] Yu, G.-I., Jeong, J. S., Kim, G.-W., Kim, S., & Chun, B.-G. (2022). Orca: A distributed serving system for transformer-based generative models. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)* (pp. 521–538). USENIX Association. https://www.usenix.org/conference/osdi22/presentation/yu

[90] Yu, P., & Chowdhury, M. (2020). Fine-grained GPU sharing primitives for deep learning applications. *Proceedings of Machine Learning and Systems, 2*, 98–111. https://proceedings.mlsys.org/paper_files/paper/2020/hash/d9cd83bc91b8c36a0c7c0fcca59228f2-Abstract.html

[91] Zaffran, M., Féron, O., Goude, Y., Josse, J., & Dieuleveut, A. (2022). Adaptive conformal predictions for time series. In *Proceedings of the 39th International Conference on Machine Learning* (Vol. 162, pp. 25834–25866). PMLR. https://proceedings.mlr.press/v162/zaffran22a.html

[92] Zhang, B., Rao, H., & Zhao, D. (2024). Evidence-grounded RAG for cloud-native DevOps: Hallucination-resistant AIOps question answering over private operations documents. *Journal of Advanced Computing Systems, 4*(3), 109–125. https://doi.org/10.69987/JACS.2024.40308

[93] Zhang, B., Sun, X., Liu, G., & Zhou, B. (2026). LLM-style DevOps copilot for cloud-native troubleshooting: Retrieval-augmented runbook generation and command-safety evaluation. *Journal of Technology Informatics and Engineering, 5*(2), 104–118. https://doi.org/10.51903/jtie.v5i2.534

[94] Zhang, J. (2025). From general human activity recognition to volleyball-oriented wearable transfer learning: Cross-dataset evidence from UCI HAR and WISDM for domain adaptation and edge deployment. *Journal of Technology Informatics and Engineering, 4*(1), 263–283. https://doi.org/10.51903/jtie.v4i1.524

[95] Zhang, J. (2026). Early warning, grade prediction, and teacher-facing LLM-ready explanations toward an open volleyball course: Reproducible evidence from four public education datasets. *Journal of Technology Informatics and Engineering, 5*(2), 20–44. https://doi.org/10.51903/jtie.v5i2.525

[96] Zhang, K., Chen, Y., & Qian, A. (2025). Evidence-grounded accounting disclosure review cards: A visual communication framework for LLM-style explanations over SEC financial statements and notes. *International Journal of Graphic Design, 3*(2), 395. https://doi.org/10.51903/ijgd.v3i2.3710

[97] Zhang, L., Ma, R., & Greg, P. (2025). Digital-twin dispatching for urban mobility via spatio-temporal Transformers and offline reinforcement learning. *Journal of Technology Informatics and Engineering, 4*(2), 337–363. https://doi.org/10.51903/jtie.v4i2.501

[98] Zhang, R., Wen, Z., Wang, C., Tang, C., Xu, P., & Jiang, Y. (2025). Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms. *arXiv*. https://doi.org/10.48550/arXiv.2511.19481

[99] Zhao, S., Bai, J., & Roberson, D. (2025). Multi-horizon GPU demand forecasting with workload semantics and operational risk curves: An empirical study on Alibaba Clusterdata GPU Trace. *Journal of Technology Informatics and Engineering, 4*(3), 544–571. https://doi.org/10.51903/jtie.v4i3.498

[100] Zhao, S., Ren, Y., & Chang, X. (2026). Profit-aware spot GPU admission control with cost-sensitive loss and evidence-grounded policy memos for AI workload supply-demand matching. *Journal of Technology Informatics and Engineering, 5*(2), 45–59. https://doi.org/10.51903/jtie.v5i2.545

[101] Zheng, D., & Li, C. (2024). Behavior-level jailbreak resistance via multi-stage refusal + utility preservation. *Journal of Advanced Computing Systems, 4*(1), 83–99. https://doi.org/10.69987/JACS.2024.40107

[102] Zheng, D., Li, C., & Davidson, H. (2023). Continual red-teaming for in-the-wild jailbreaks via online guardrail updates and guardrail distillation. *Journal of Advanced Computing Systems, 3*(2), 35–49. https://doi.org/10.69987/JACS.2023.30203

[103] Zheng, D., Zhang, B., & Geibel, J. (2024). VerifySafe: Toxicity-safe agent responses under adversarial prompts with evidence-based self-verification. *Journal of Advanced Computing Systems, 4*(1), 67–82. https://doi.org/10.69987/JACS.2024.40106

[104]Zhong, Y., Liu, S., Chen, J., Hu, J., Zhu, Y., Liu, X., Jin, X., & Zhang, H. (2024). DistServe: Disaggregating prefill and decoding for goodput-optimized large language model serving. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)* (pp. 193–210). USENIX Association. https://www.usenix.org/conference/osdi24/presentation/zhong-yinmin

[105]Zhong, Z. S., Chen, J., Zhong, E., & Sun, X. (2025). Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0. *Journal of Technology Informatics and Engineering, 4*(2), 502–520. https://doi.org/10.51903/jtie.v4i2.536

[106]Zhong, Z. S., Li, C., & Rao, H. (2026). Trajectory reliability prediction for generalist AI agents: Tool-use failure analysis and success forecasting on ZClawBench. *Journal of Technology Informatics and Engineering, 5*(1), 341–360. https://doi.org/10.51903/jtie.v5i1.539

[107]Zhong, Z. S., Pan, X., & Lei, Q. (2025). Bridging domains with approximately shared features. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics* (Vol. 258, pp. 559–567). PMLR. https://proceedings.mlr.press/v258/zhong25a.html

[108]Zhou, B., Li, C., & Liu, L. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication. *International Journal of Graphic Design, 3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3696

[109]Zhou, H., & Zhang, K. (2025). News-based uncertainty and macro-market fusion for VIX direction forecasting: Evidence from 2015–2024 FRED panel. *Journal of Technology Informatics and Engineering, 4*(2), 487–501. https://doi.org/10.51903/jtie.v4i2.540

[110]Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(12), 11106–11115. https://doi.org/10.1609/aaai.v35i12.17325

[111]Zhou, S., Chen, Y., & Lee, K. (2026). Accounting-aware evidence-constrained agents for disclosure, settlement, and secondary-market risk monitoring in tokenized RWA infrastructure. *Journal of Technology Informatics and Engineering, 5*(2), 60–74. https://doi.org/10.51903/jtie.v5i2.544