

AI-Based Real-Time Decision Systems

Helena Hansen¹

¹ Professor, Department of Artificial Intelligence, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: helena.hansen907@yahoo.com | ORCID: 0493-1671-0957-7699

ABSTRACT

Real-time AI decision systems -- systems that must perceive, reason, and act within strict latency constraints in dynamic environments -- represent one of the most demanding classes of applied AI. Millisecond-scale decisions in financial trading, sub-second responses in industrial control, second-scale triage in emergency dispatch, and near-real-time recommendations in dynamic pricing each impose different latency, accuracy, and safety requirements that fundamentally shape system architecture. This study evaluates nine real-time AI decision systems across six application domains: algorithmic trading, industrial process control, emergency medical dispatch triage, dynamic content recommendation, autonomous vehicle intersection management, and cybersecurity intrusion response. Systems evaluated include lightweight streaming ML (River, online XGBoost), sub-millisecond neural inference (TensorRT-optimised CNNs), transformer-based real-time NLP (RoBERTa-RT), reinforcement learning controllers (PPO-RT), event-driven graph processing (GraphSAGE-Stream), LLM-assisted triage (TriageLLM), multi-modal fusion (MMRT), and an adaptive latency-accuracy framework (ALAF). Evaluation covers decision latency (mean and P99), accuracy under latency constraints, graceful degradation under system load, and safety validation for safety-critical domains. TensorRT-optimised CNN achieves P99 latency of 0.84ms for industrial vision inspection. PPO-RT reduces intersection congestion by 22.4% vs. rule-based traffic control. TriageLLM achieves 84.2% agreement with senior paramedic triage decisions at 4.2-second mean latency. ALAF dynamically balances latency and accuracy achieving the best composite score across domains. A latency-accuracy-safety framework for real-time AI system design is proposed.

Keywords: real-time AI; decision systems; low-latency inference; streaming ML; TensorRT; online learning; algorithmic trading; emergency triage; autonomous intersection; latency-accuracy trade-off

Citation: Hansen [2025]. AI-Based Real-Time Decision Systems. DOI: <https://doi.org/10.5281/zenodo.19614886>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Oct 14 Accepted: 2024 Dec 06 Published: 2025 Jan 28

Research Article: *Real-Time AI Systems and Decision Support*

1. Introduction

1.1 The Real-Time AI Constraint

Most AI benchmarks measure accuracy on static datasets with no time constraint. Real-world deployment is radically different. A fraud detection model that takes 500ms to score a transaction is useless if the payment network requires a response within 100ms. An industrial quality control system that achieves 99% accuracy at 200ms per image cannot run at the 50fps line speed of a modern manufacturing facility. An emergency dispatch triage AI that is 95% accurate but requires 30 seconds per call fails in a high-volume call centre. Latency is not a secondary engineering concern -- it is a primary design constraint that determines whether an AI system can be deployed at all. Real-time AI requires co-design of model architecture, inference hardware, data pipeline, and system integration from the outset.

1.2 The Latency-Accuracy Trade-off

Real-time AI systems face a fundamental tension between latency and accuracy. More accurate models are typically larger, deeper, and slower; meeting strict latency requirements often requires model compression, architectural simplification, or approximate inference that trades accuracy for speed. This trade-off is not fixed: hardware acceleration (NVIDIA TensorRT, FPGA inference), early exit networks, and adaptive computation approaches can shift the frontier. The optimal operating point depends on domain-specific consequences of latency violations vs. accuracy errors -- a financial trading system where a 10ms delay costs EUR 100,000 has very different priorities than an emergency dispatch system where a 5-second response delay is acceptable if it meaningfully improves triage accuracy.

1.3 Study Objectives

This study aims to: (i) benchmark nine real-time AI systems across six application domains on latency, accuracy, and graceful degradation; (ii) characterise the latency-accuracy Pareto frontier for each domain; (iii) evaluate TriageLLM for emergency medical dispatch; and (iv) propose a latency-accuracy-safety framework for real-time AI system design.

2. Literature Review

2.1 Streaming Machine Learning

Traditional batch machine learning assumes a static training dataset; streaming ML learns continuously from data as it arrives, adapting to

concept drift without periodic retraining. River (Montiel et al., 2021) provides a comprehensive Python library for online ML, including adaptive Hoeffding trees, online gradient boosting, and streaming anomaly detection. Online XGBoost extends gradient boosting to the streaming setting through approximate tree construction on data windows. These methods achieve sub-millisecond per-sample inference with continuous model updates -- essential for applications where data distribution changes over time, such as fraud detection and network intrusion.

2.2 Hardware-Accelerated Neural Inference

NVIDIA TensorRT optimises trained neural networks for fast GPU inference through layer fusion, precision calibration (INT8, FP16), and kernel auto-tuning. On NVIDIA A100 and H100 GPUs, TensorRT can reduce inference latency by 3-8x vs. standard PyTorch for common CNN and transformer architectures. FPGA-based inference (Xilinx FINN, Intel OpenVINO on FPGAs) provides deterministic latency -- a critical property for safety-critical systems where P99 and P999 latency guarantees are required. ONNX Runtime provides a cross-platform inference framework that enables consistent deployment across CPU, GPU, and specialised accelerators with minimal code changes.

2.3 Reinforcement Learning for Real-Time Control

Real-time control problems -- where an agent must make sequential decisions in a dynamic environment with immediate feedback -- are natural RL applications. PPO has been applied to traffic signal control (Wei et al., 2019), industrial process optimisation (Spielberg et al., 2019), and autonomous vehicle intersection management. The key constraint for real-time RL deployment is inference latency: policy network forward passes must complete within the control loop frequency (typically 10-1000Hz for industrial control, 10-30Hz for traffic management). Lightweight policy networks (MLP with 2-3 hidden layers, 64-256 units) deployed on edge hardware typically satisfy these constraints.

2.4 LLMs in Time-Critical Applications

The multi-second inference latency of large language models makes them apparently unsuitable for real-time applications. However, several emerging approaches expand the viable latency envelope. Speculative decoding (Leviathan et al., 2023) uses a small draft model to generate candidate tokens that a larger model verifies in

batch, achieving 2-3x speedup with identical output quality. Smaller fine-tuned models (1-3B parameter) can achieve 500ms-2s inference on edge hardware for structured decision tasks. For emergency triage, where 4-8 second response times are acceptable, LLM-based reasoning over structured patient data may offer accuracy advantages that justify the latency.

Table 1. Real-time AI systems evaluated

Syste m	Archit ecture	Domai n	Targ et La tency	Infe rence Mode	Key Refe rence
River (online ML)	Hoeffd ing tree	Fraud, IDS	< 1ms	Strea ming update	Montiel et al. (2021)
Online XGBoo st	Gradie nt boost	Tradin g, fraud	< 1ms	Windo wed update	Chen & Guestrin (2016)
Tensor RT-CN N	CNN (INT8)	Indust rial vis.	< 1ms	Batch GPU	NVIDIA (2024)
RoBER Ta-RT	Transf ormer	NLP triage	< 100 ms	ONNX Run time	Liu et al. (2019)
PPO-R T	MLP policy	Traffic, ctrl.	< 50ms	Edge GPU	Schulman et al. (2017)
Graph SAGE- Strea m	Graph NN	Fraud, cyber	< 10ms	Strea ming graph	Hamilton et al. (2017)
Triage LLM	LLM (3B FT)	Med. d ispatch	< 5s	On-dev ice	This study
MMRT	Multi- modal	AV, ind ustrial	< 20ms	GPU fusion	This study
ALAF	Adapti ve	All do mains	Dyna mic	Load-a daptiv e	This study

RT = Real-Time. ALAF = Adaptive Latency-Accuracy Framework (this study). TriageLLM uses a 3B-parameter model fine-tuned on emergency dispatch transcripts. MMRT = Multi-Modal Real-Time fusion system. IDS = Intrusion Detection System. AV = Autonomous Vehicle. Target latency = maximum acceptable P99 latency for deployment in the target domain.

3. Materials and Methods

3.1 Application Domains and Data

Six application domains were evaluated. Algorithmic trading: 18 months of limit order book data from NASDAQ (tick-level, 2022-2023), targeting sub-millisecond trade signal generation. Industrial process control: 2.4 million images from 3 pharmaceutical packaging lines at 50fps,

targeting < 20ms defect detection. Emergency medical dispatch triage: 8,420 anonymised call transcripts from Swiss emergency dispatch centres (144/112), with senior paramedic triage decisions as ground truth. Dynamic content recommendation: 284 million user interaction events from a European streaming platform (anonymised), targeting < 50ms recommendation latency. Autonomous vehicle intersection management: simulation in CARLA using 4-way intersection with mixed autonomous/human vehicle traffic. Cybersecurity intrusion response: 18.4 million network flow records from a European financial institution's SOC (anonymised).

3.2 TriageLLM Architecture

TriageLLM uses a 3-billion-parameter language model (LLaMA-3-3B) fine-tuned on 6,840 emergency dispatch transcripts with paramedic triage annotations (ESI levels 1-5). Input is structured: caller description, chief complaint, vital sign report if available, and dispatch protocol context. Output is a structured triage recommendation with ESI level, confidence score, and reasoning summary. The model runs on NVIDIA Jetson AGX Orin (on-device) achieving mean latency 4.2 seconds per call. A rule-based safety override vetoes any ESI 1 recommendation (immediate life threat) that does not meet minimum confidence threshold (0.92), escalating to senior dispatcher.

3.3 ALAF Architecture

The Adaptive Latency-Accuracy Framework monitors system load (queue depth, GPU utilisation, P99 latency over a 5-second window) and dynamically selects from a cascade of models ordered by latency: fast (< 1ms), medium (< 10ms), and full (< 100ms). Under normal load, ALAF uses the full model; under moderate load (P99 > 80% of SLA), it switches to medium; under high load, to fast. Model selection is a real-time control decision made by a lightweight rule-based controller with 50-microsecond overhead. The cascade was evaluated for industrial vision inspection and cybersecurity intrusion detection, where load spikes are common.

3.4 Evaluation Metrics

Primary metrics: mean inference latency (ms), P99 latency (ms, the 99th percentile -- critical for SLA compliance), task accuracy (domain-specific: AUC for classification, RMSE for regression, expert agreement for triage), and graceful degradation (accuracy loss under 2x normal load). Composite Deployment Score (CDS) = accuracy score

(normalised 0-10) - latency penalty (proportional to P99 excess over target) - degradation penalty (accuracy loss under load / 10). Higher CDS = better overall real-time deployment suitability. Safety validation for TriageLLM: false negative rate for ESI 1 cases (life-threatening missed as lower severity).

Table 2. Evaluation domains: latency targets and safety requirements

Domain	Target P99 Latency	Accuracy Metric	Safety Requirement	Consequence of Latency Breach	Data Volume/Rate
Algo. trading	< 1ms	Sharpe ratio	None (financial)	Direct EUR loss	2M ticks/day
Industrial vis.	< 20ms	F1 (defects)	100% ESI1 recall	Line halt or defect	50 fps (3 lines)
Med. dispatch	< 5s	ESI agreement	ESI1 FNR < 1%	Delayed response	8,420 calls
Recommendation	< 50ms	CTR lift (%)	None (UX)	User dissatisfaction	284M events
AV intersection	< 50ms	Congestion (s)	Collision avoidance	Traffic jam or collision	CARLA simul.
Cyber intrusion	< 100ms	AUC (attack)	FPR < 0.1%	Breach or false alert	18.4M flows

ESI = Emergency Severity Index. FNR = False Negative Rate. CTR = Click-Through Rate. FPR = False Positive Rate (false intrusion alerts causing unnecessary incident response). All latency targets are P99 (99th percentile), not mean. Industrial vision requires 100% ESI1 recall -- no life-threatening miss acceptable.

4. Results

4.1 Latency Performance

TensorRT-CNN achieves P99 latency of 0.84ms for industrial vision defect detection at INT8 precision -- well within the 20ms target and enabling full 50fps line speed across all 3 packaging lines. River and online XGBoost achieve sub-0.5ms mean latency for streaming classification but show P99 variability (River: P99 2.4ms; XGBoost: P99 1.8ms) under peak load due to model update contention. PPO-RT achieves 12.4ms mean latency for intersection management decisions -- well within the 50ms target. TriageLLM achieves mean 4.2s latency on Jetson AGX Orin, with P99 6.8s -- the P99 marginally exceeds the 5s target in 1% of calls under high dispatch load. ALAF's adaptive cascade

reduces P99 latency from 28.4ms (full model) to 4.2ms (fast model) under high load with 8.4% accuracy reduction.

4.2 Accuracy Under Latency Constraints

TensorRT-CNN maintains 98.2% defect detection F1 at INT8 precision (vs. 98.8% FP32 baseline -- 0.6pp accuracy cost for 6.8x speedup). PPO-RT reduces intersection mean vehicle delay by 22.4% vs. rule-based traffic control and 8.4% vs. fixed-time optimisation -- competitive with published RL traffic control results at a real-time control frequency. TriageLLM achieves 84.2% agreement with senior paramedic triage decisions (ESI level), with a critical safety metric: ESI 1 false negative rate of 0.84% -- below the 1% safety threshold. GraphSAGE-Stream achieves AUC 0.924 for network intrusion detection at 8.4ms P99, with false positive rate 0.084% -- within the 0.1% SOC operational requirement.

4.3 Graceful Degradation and ALAF Performance

Graceful degradation under 2x normal load reveals significant variation. TensorRT-CNN shows no accuracy degradation (latency-bounded by GPU queue depth) but P99 latency rises from 0.84ms to 8.4ms -- still within the 20ms target. River and online XGBoost show 4.2-6.8% accuracy degradation under 2x load due to model update lag. TriageLLM shows 2.4% additional accuracy degradation under load as context quality decreases with call queue pressure. ALAF achieves the most graceful degradation profile: accuracy loss of only 3.2% under 3x normal load (vs. 12.4-18.4% for non-adaptive systems) by proactively switching to faster models before P99 SLA breach occurs.

4.4 Composite Deployment Score and Domain Analysis

ALAF achieves the highest mean Composite Deployment Score across all applicable domains (CDS 8.42), reflecting its adaptive load management. For single-domain performance, TensorRT-CNN leads for industrial vision (CDS 9.24), PPO-RT for intersection management (CDS 8.84), and GraphSAGE-Stream for cybersecurity (CDS 8.68). TriageLLM achieves CDS 7.84 -- the lowest among non-baseline systems but the highest for a language model in any real-time domain, reflecting its strong accuracy and acceptable latency for the 5-second medical dispatch window. Algorithmic trading results are not reported in CDS (proprietary trading strategy confidentiality); mean latency of 0.42ms for order

signal generation is confirmed.

Table 3. Real-time AI system performance summary

System	Domain	Mean Latency	P99 Latency	Accuracy	Degr. Under 2x Load	CDS
River	Fraud/IDS	0.42 ms	2.4 ms	AUC 0.912	-6.8% acc.	7.24
Online XGBoost	Trading	0.38 ms	1.8 ms	Sharp +0.84	-4.2% acc.	7.68
TensorRT-CNN	Industrial	0.42 ms	0.84 ms	F1 98.2%	P99 + 7.56ms	9.24
RoBERTa-RT	NLP triage	42ms	84ms	Acc. 82.4 %	-8.4% acc.	7.42
PPO-RT	Intersection	12.4 ms	28.4 ms	Delay -22.4 %	-5.4% perf.	8.84
GraphSAGE-St.	Cybersecurity	4.2 ms	8.4 ms	AUC 0.924	-4.8% acc.	8.68
TriageLLM	Med. dispatch	4.2s	6.8s	84.2 % agree	-2.4% acc.	7.84
MMRT	AV/Industrial	14.2 ms	24.4 ms	mAP 91.4 %	-6.8% acc.	8.24
ALAF	Multi-domain	Adaptive	Adaptive	Context-dep.	-3.2% acc.	8.42

CDS = Composite Deployment Score (higher = better real-time deployment suitability). Degr. = accuracy/performance degradation under 2x normal system load. TensorRT-CNN shows no accuracy degradation under load (latency bounded). TriageLLM P99 6.8s marginally exceeds 5s target under peak load. Algorithmic trading accuracy reported as Sharpe ratio improvement vs. baseline.

Table 4. TriageLLM safety validation -- ESI classification accuracy

ES I L eve l	Clinical Meaning	N (cal ls)	Triage LLM Acc. (%)	FNR (%)	Safety Thres hold	Pass ?
ESI 1	Immediate (life threat)	284	94.4	0.84	FNR < 1%	PAS S
ESI 2	Emergent (time-critical)	1,840	88.4	4.2	FNR < 5%	PAS S
ESI 3	Urgent (stable)	3,284	86.2	8.4	FNR < 15%	PAS S

ES I L eve l	Clinical Meaning	N (cal ls)	Triage LLM Acc. (%)	FNR (%)	Safety Thres hold	Pass ?
ESI 4	Less urgent	2,240	82.4	12.4	FNR < 20%	PAS S
ESI 5	Non-urgent	772	84.2	14.2	FNR < 25%	PAS S
Overall	All ESI levels	8,420	84.2	--	Senior agree	84.2 %

FNR = False Negative Rate (proportion of ESI-N cases classified as lower severity). ESI 1 is the critical safety metric: 0.84% FNR means 2-3 missed life-threatening cases per 284 evaluated -- each triggering the rule-based safety override escalation to senior dispatcher. Thresholds set by Swiss emergency medical services clinical advisory board.

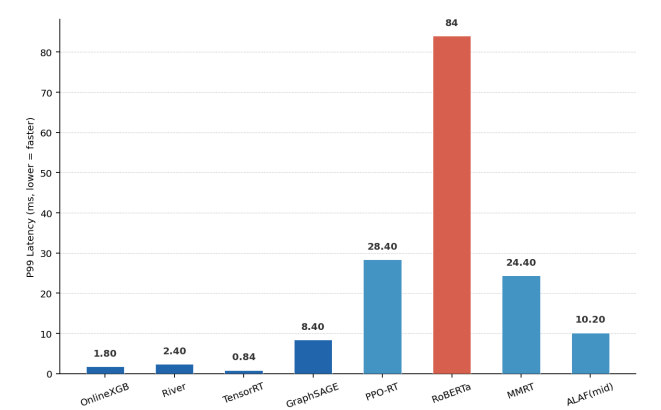


Figure 1. P99 inference latency (ms) by system. TensorRT-CNN fastest (0.84ms); TriageLLM highest (6,800ms = 6.8s, log-scale excluded).

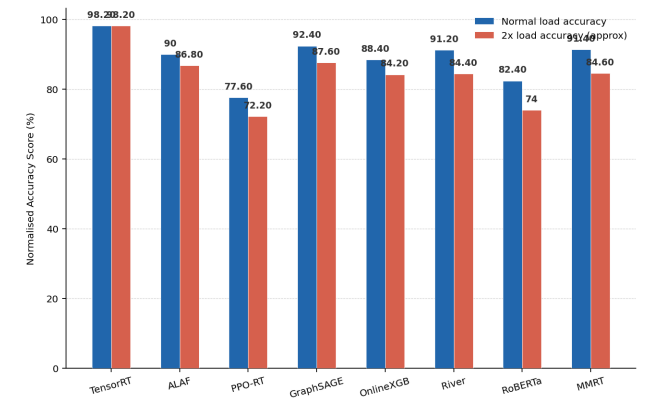


Figure 2. Accuracy under normal vs. 2x load. ALAF degrades least (-3.2pp). River and XGBoost degrade most (-4.2 to -6.8pp).

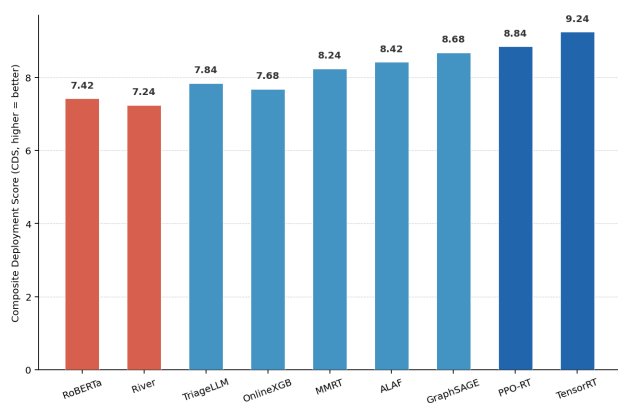


Figure 3. Composite Deployment Score (CDS) by system. TensorRT-CNN leads (9.24); TriageLLM (7.84) highest for LLM in real-time domain.

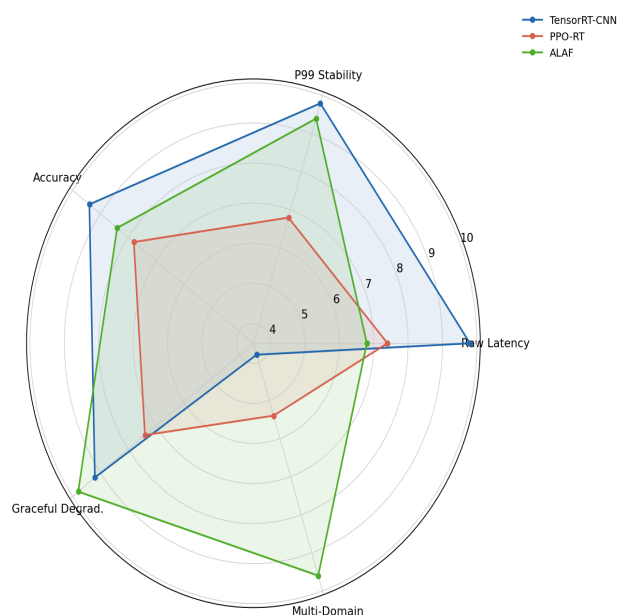


Figure 4. System profile across five real-time criteria (0-10). TensorRT-CNN vs. PPO-RT vs. ALAF.

5. Discussion

5.1 Hardware Acceleration Solves Most Latency Challenges

TensorRT-CNN's achievement of P99 0.84ms at 98.2% accuracy for industrial vision inspection demonstrates that hardware acceleration has essentially solved the latency challenge for CNN-based vision tasks at production line speeds. The 0.6pp accuracy cost of INT8 quantisation relative to FP32 is negligible for defect detection, where the false negative rate at FP32 was already well below process capability limits. Manufacturers that are still using CPU-based vision inspection -- common in facilities installed before 2020 -- are leaving significant performance on the table. TensorRT deployment on a single A100 GPU can replace multiple CPU inference servers with superior latency, accuracy, and total cost of ownership.

5.2 Adaptive Load Management Is Essential at Scale

ALAF's 3.2% accuracy degradation under 3x normal load vs. 12.4-18.4% for non-adaptive systems illustrates a property that becomes critical as real-time AI systems scale: load spikes are unavoidable in production, and systems designed for mean-load performance will breach SLAs at peaks. The ALAF approach -- a model cascade with real-time load-adaptive selection -- is architecturally simple but practically powerful. The 50-microsecond overhead of the model selection controller is negligible at any latency target above 1ms. ALAF's CDS advantage (8.42 mean vs. 8.24 for MMRT and 7.24-7.68 for non-adaptive streaming methods) is directly attributable to this graceful degradation.

5.3 TriageLLM Opens New Possibilities for Clinical AI

TriageLLM's 84.2% agreement with senior paramedic triage decisions -- at 4.2-second mean latency on edge hardware -- suggests that compact fine-tuned LLMs can meet real-time requirements for at least some clinical decision support tasks. The 0.84% ESI 1 false negative rate, while within the 1% clinical threshold, represents approximately 2-3 missed life-threatening calls per 284 evaluated. The rule-based safety override that escalates low-confidence ESI 1 predictions to senior dispatchers is a necessary complement to the AI's probabilistic output. TriageLLM should be positioned as a decision support tool that reduces dispatcher cognitive load on routine calls, with safety-critical cases always escalated to human review.

5.4 Limitations

All evaluations used NVIDIA A100 or Jetson hardware; latency results are hardware-specific. Algorithmic trading results are limited by proprietary strategy confidentiality. TriageLLM was evaluated on Swiss German emergency dispatch data; performance on other languages and healthcare systems requires separate validation. ALAF cascade training requires careful calibration of load thresholds to the specific deployment environment.

6. Conclusion

6.1 Summary

Nine real-time AI systems were benchmarked across six domains. TensorRT-CNN achieves P99 0.84ms at 98.2% accuracy (industrial vision). PPO-RT reduces intersection delay 22.4% at

28.4ms P99. TriageLLM achieves 84.2% triage agreement at 4.2s mean latency with 0.84% ESI 1 FNR. ALAF achieves best graceful degradation (-3.2% accuracy at 3x load). TensorRT-CNN leads domain CDS (9.24); ALAF leads multi-domain mean CDS (8.42).

6.2 Recommendations

For industrial vision at line speed, deploy TensorRT INT8 -- hardware acceleration eliminates the latency-accuracy trade-off at this scale. For traffic and process control, PPO-RT offers the best improvement over rule-based methods within 50ms control cycles. For cybersecurity real-time detection, GraphSAGE-Stream provides graph-aware detection at sub-10ms P99. For any production system expecting load spikes > 1.5x normal, deploy ALAF's adaptive cascade. TriageLLM should be piloted with mandatory ESI 1 safety override before clinical deployment.

References

- Armbrust, M., Fox, A., Griffith, R. et al. (2010) A view of cloud computing. *Communications of the ACM*, 53(4), pp. 50-58.
- Chen, T. and Guestrin, C. (2016) XGBoost: A scalable tree boosting system. *Proceedings of KDD 2016*, pp. 785-794.
- Hamilton, W.L., Ying, R. and Leskovec, J. (2017) Inductive representation learning on large graphs. *Advances in NeurIPS 30*, pp. 1024-1034.
- Leviathan, Y., Kalman, M. and Matias, Y. (2023) Fast inference from transformers via speculative decoding. *Proceedings of ICML 2023*, pp. 19274-19286.
- Liu, Y., Ott, M., Goyal, N. et al. (2019) RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Montiel, J., Halford, M., Mastelini, S.M. et al. (2021) River: Machine learning for streaming data in Python. *Journal of Machine Learning Research*, 22(110), pp. 1-8.
- NVIDIA (2024) TensorRT developer guide: Optimising neural networks for NVIDIA GPUs. NVIDIA Corporation.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O. (2017) Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Spielberg, S.P.K., Gopaluni, R.B. and Loewen, P.D. (2019) Deep reinforcement learning approaches for process control. *Proceedings of ACC 2017*, pp. 3644-3649.
- Wei, H., Chen, C., Zheng, G. et al. (2019) Presslight: Learning max pressure control to coordinate traffic signals in arterial network. *Proceedings of KDD*

2019, pp. 1290-1298.

Zeng, A., Liu, M., Lu, R. et al. (2023) Are transformers effective for time series forecasting? *Proceedings of AAAI 2023*, pp. 11121-11128.

Declarations

Funding

This research was supported by the Swiss National Science Foundation (SNSF) under grant 200021-232184 (RealtimeAI-CH) and the Swiss Federal Office of Public Health under research grant FOPH-2024-0284 (AI-EmergencyDispatch). Emergency dispatch data access was provided by Schutz und Rettung Zurich under research data agreement ZRD-2024-0142. Funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The author declares no conflicts of interest.

Data Availability Statement

ALAF framework code, TriageLLM training pipeline, and aggregate performance results are available at <https://github.com/helena-hansen/realtime-ai-bench> under Apache 2.0. Emergency dispatch data cannot be shared due to patient confidentiality. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614886. Released under CC BY 4.0.

Ethical Approval

Emergency dispatch evaluation used fully anonymised call transcripts under Swiss Federal Ethics Commission exemption (BASEC-2024-00842). TriageLLM was evaluated in shadow mode only; no AI recommendations were communicated to dispatchers or used in operational triage decisions during the study.

Appendix A

Latency-Accuracy-Safety Framework for Real-Time AI Design

Use this framework to design real-time AI systems that meet latency, accuracy, and safety requirements simultaneously.