Data Infrastructure for Robotaxi Services: Opportunities and Challenges

Yuxin Wang, Everett Richards, and Weisong Shi
Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716, USA
{yuxw, evrich, weisong}@udel.edu

Abstract—Autonomous driving research has traditionally focused on what an individual vehicle can perceive, predict, and decide in real time. A robotaxi extends this capability into a passenger service operated across a fleet. At fleet scale, an equally important question emerges: what should the fleet remember? A robotaxi fleet continuously encounters changing road conditions, unusual events, and system problems. These experiences become valuable when the fleet can learn from past events and when one vehicle can benefit from what other vehicles have encountered. We argue that robotaxi data infrastructure should therefore be viewed not simply as a logging or storage system, but as the mechanism that gives the fleet a shared, queryable memory. We call this capability fleet-level world memory. It links what happened in the physical world with the system conditions under which it was observed. This connection allows fleet experience to support both current operations and future system improvement. We propose a fourteen-stage reference architecture and identify eight open research challenges. Our central thesis is that scalable robotaxi services depend not on retaining the most data, but on preserving the right evidence with enough context to support trustworthy decisions and learning.

Index Terms—robotaxi, autonomous driving, data infrastructure, data provenance.

1 From Autonomous Driving to Robotaxi Service

Conventional driver-assistance systems support a human who remains responsible for supervision. In contrast, Level 4 automation performs the driving task without human fallback within a defined operational design domain [1]. A robotaxi applies this capability to passenger transport. Without a driver, exceptional situations and rider support must instead be handled by fleet operations and remote-assistance teams.

Consider a robotaxi entering a construction zone where a worker's instruction conflicts with the map. As traffic slows, the vehicle hesitates. The hesitation may reflect a weakness in perception or planning, or it may result from a problem with the vehicle, such as a sensor fault or calibration drift. The first case may call for system improvement, while the second may require immediate inspection.

To make either decision, the fleet needs more than an isolated sensor frame or diagnostic signal. It needs a record of what happened, how the situation developed, and the state of the vehicle at the time. Evidence becomes useful when observations of the external world remain linked to the system that produced them [2].

At fleet scale, these records must also connect experiences across vehicles and over time. We call this capability *fleet-level world memory*. It links a *world history* of what happened in the physical environment with a *system history* of how those observations were produced and used.

Robotaxi data infrastructure builds and maintains this memory. It collects observations, preserves their context, supports learning and evaluation, and records the results of fleet actions [3], [4]. This allows the fleet to learn from past experience and lets one vehicle benefit from what other vehicles have encountered.

Building this capability across a large fleet is a systems problem. Retaining more data increases storage, transfer, and processing costs, while filtering too aggressively may discard information that later becomes important. The goal is therefore not to retain the most data, but to preserve enough trustworthy evidence at practical cost and latency.

This paper formalizes fleet-level world memory, presents a fourteen-stage reference architecture organized into five technical directions, and identifies research challenges for building scalable, trustworthy and economical robotaxi data infrastructure. Robotaxi capability depends not only on what each vehicle can decide in the moment, but also on what the fleet can reliably remember and learn from over time.

2 Robotaxi Data Infrastructure: Definition and Fleet-Level World Memory

Robotaxi data infrastructure is the end-to-end system that turns observations from passenger service into evidence for fleet operation, system improvement, and assurance. It begins when an event is observed onboard, preserves enough context to interpret that event as data move between the vehicle and the cloud, and organizes the resulting information so that engineering results can be reproduced. It also returns validated results to fleet operation.

This makes robotaxi data infrastructure different from a conventional data platform. Its purpose is not simply to store logs or provide data for model training. It must support both current fleet operation and long-term system improvement. A field event may lead to an immediate action, while its outcome may later help improve the driving system. The important question is therefore not how much data are retained, but whether the evidence can support a decision and help evaluate what happened afterward.

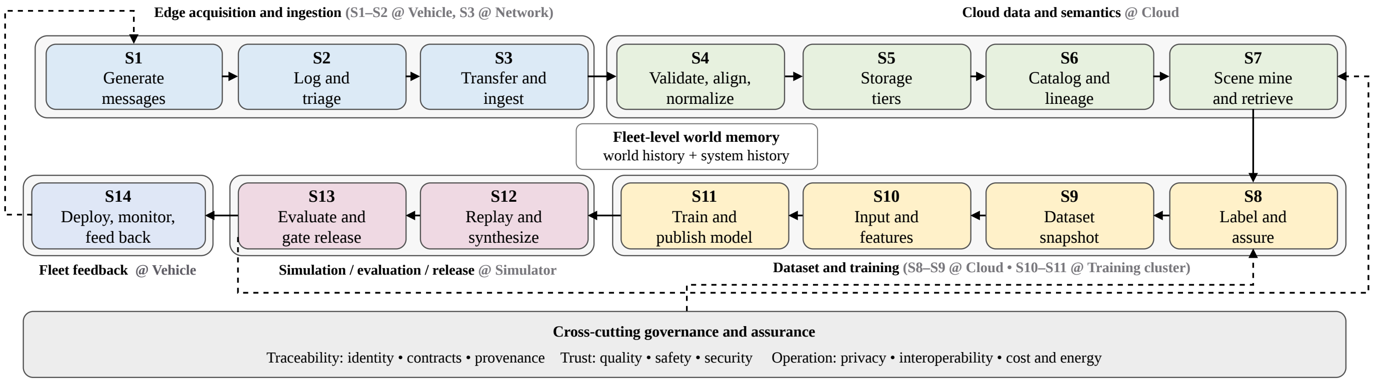

Fig. 1. Lifecycle of fleet-level world memory across fourteen stages and five technical directions. Solid arrows show the flow of observations and derived artifacts, while dashed arrows show feedback to earlier stages. World history records changes in the physical environment, while system history records the conditions under which observations were produced. Governance and assurance apply across the full lifecycle. Labels marked with @ show typical execution domains.

It is not a complete model of the road. Instead, it is a queryable record that allows events to be reconstructed and compared across vehicles and over time. Unlike a file archive, fleet memory preserves enough context to explain what the fleet encountered, how the observation was produced, and how it was later used.

This memory contains two connected histories. *World history* records what happened in the physical environment and how it changed over time. *System history* records the state of the vehicle and system that produced each observation. These histories must remain linked because the same situation may be observed differently under different sensor, calibration, or software conditions.

Within this model, we define *scene* as a bounded, time-ordered record of an encounter, including what the vehicle did and what happened afterward. A *scenario* groups related scenes so they can be retrieved and compared. *Evidence lineage* connects these scenes to the datasets, models, evaluations, and fleet actions derived from them.

Figure 1 shows how this memory moves through the system. Stages 1–3 capture observations and establish trust at the source; Stages 4–7 validate and organize them for fleet-wide search and analysis; Stages 8–11 support learning; Stages 12–13 support evaluation and release; and Stage 14 returns results to fleet operation. These stages represent major data and responsibility boundaries rather than fourteen required software services.

3 Five Technical Directions

We organize the robotaxi data infrastructure into five technical directions that form, organize, use, test, and update fleet-level world memory. Their latency and retention requirements differ, but data identity and temporal context must remain consistent across the full lifecycle.

3.1 Edge Acquisition and Ingestion (Stages 1–3)

Trustworthy fleet memory begins at the source. On the vehicle, Stage 1 publishes typed messages [5], Stage 2 retains selected evidence, and Stage 3 transfers it through the network. During acquisition, the system should preserve when and where an observation was made, the state of the vehicle, why the data were collected, and what information may be missing. Once this context is lost, it may be difficult or impossible to recover in the cloud.

IEEE 1588 provides a foundation for clock synchronization [6], but synchronized clocks alone do not establish the sensor's actual capture time, geometric consistency, or the relation between vehicle health and behavior [7]. The complete record must still fit within tight onboard compute, storage, and bandwidth budgets.

3.2 Cloud Data and Semantics (Stages 4–7)

In the cloud, Stage 4 receives the data, checks its integrity, and aligns incoming packages. Stages 5–7 organize observations from many vehicles into searchable fleet memory. The cloud preserves data identity and lineage while allowing related events to be found across vehicles, locations, and time. This makes it possible to identify patterns that may not be visible in a single vehicle log.

Systems such as Delta Lake [8] demonstrate how versioned data storage can preserve access to earlier data states. Robotaxi infrastructure must extend this idea beyond tables to operational and learning evidence. For example, the system should help determine whether repeated hesitation in construction zones comes from similar road conditions or from a common vehicle or software condition.

3.3 Dataset and Training (Stages 8–11)

Stages 8–11 turn selected fleet experience into training data and models while preserving provenance. Scenes are labeled, organized into versioned datasets, converted into reproducible training inputs, and used to produce models. A dataset is therefore a task-specific view of fleet memory rather than an isolated collection of examples.

Provenance also matters for data quality. Pseudo-labeling can propagate model errors [9], while related scenes can leak across training and evaluation splits and produce overly optimistic results [10]. Information about sensor health and calibration should therefore remain connected to the scenes used for learning.

3.4 Evaluation and Release (Stages 12–13)

Stages 12–13 test proposed system changes before they are released to the fleet. Evaluation may include replay, simulation, physical testing, and controlled deployment. CARLA shows how simulation can support autonomous-driving development and validation [11], but evaluation methodology matters: open-loop evaluation can be misleading for long-term planning [12].

A release decision should therefore not depend on a single score. It should preserve the evidence needed to explain what improved, what regressed, and why the change was approved [13].

3.5 Fleet Feedback (Stage 14)

Stage 14 closes the loop by using fleet memory to guide action and recording what happened afterward. In the short term, evidence may determine whether a vehicle remains in service or requires intervention. Over time, the same evidence can improve data collection, maintenance, learning, and future releases.

Each action should remain linked to the evidence that supported it and to its eventual outcome [2], [4]. This feedback allows the fleet to learn whether a decision worked and to use that experience when similar situations occur again.

4 Research Challenges and Opportunities

Fleet memory supports two questions: *Is this vehicle fit to serve now?* and *What should the fleet learn for later?* These questions share evidence but differ in latency, uncertainty, and risk. The challenge is therefore not simply to collect more data, but to understand what the data mean, what may be missing, and whether the evidence is reliable enough for a particular decision.

4.1 Temporal and Physical Integrity

Even basic information such as time and sensor pose can be uncertain. A camera timestamp may refer to exposure, transmission, decoding, or delivery to the driving system. These times are not always the same. Physical calibration has a similar problem because sensor pose can change with vibration, temperature, maintenance, or mechanical movement.

The open challenge is to represent these uncertainties and understand how they affect downstream tasks. Research should determine when errors in time, pose, or calibration become large enough to change sensor fusion, scene reconstruction, incident analysis, or learning results.

A useful benchmark could introduce controlled timing and calibration errors and measure both reconstruction accuracy and their effect on downstream decisions.

4.2 Selection and Coverage

A fleet cannot retain every sensor observation at full quality. Selective collection reduces storage and bandwidth, but it may miss rare events or early signs of failure. Once these data are discarded, they may be impossible to recover.

The research challenge is to understand what a collection policy misses and whether the retained data still represent important fleet conditions. This becomes especially difficult when models trained on past data also decide what data should be collected next, since the fleet may repeatedly collect examples of problems it already knows.

A useful benchmark could compare collection policies against a high-retention reference stream and measure missed events, fleet coverage, downstream usefulness, and retained data volume.

4.3 Cloud Data Lifecycle and Retrieval

Cloud systems combine observations from many vehicles to find patterns across the fleet. However, similar scenes may have been recorded under different vehicle or system conditions. If these differences are hidden during aggregation, problems affecting only part of the fleet may be missed.

The open challenge is to support joint fleet analysis while preserving the system conditions of each observation. Retrieval should consider both what happened in the physical world and how the observation was produced.

A second challenge is maintaining evidence across the full lifecycle. A field observation may later become a scene, label, dataset, model input, evaluation result, and fleet action. If the original data are later corrected or found unreliable, the system should identify which downstream artifacts and decisions are affected. Research is needed to preserve this lineage and determine what must be rechecked when earlier evidence changes.

4.4 Learning and Operational Decisions

Improving one component of a robotaxi system does not necessarily improve the system as a whole. For example, a perception update may improve local accuracy while changing downstream prediction or planning behavior and introducing new regressions.

The challenge is even greater at fleet scale. A change that performs well on several vehicles may not generalize to the entire fleet because vehicles can differ in sensor condition, calibration, hardware, software configuration, or operating environment. Research should therefore determine when a local improvement translates into end-to-end improvement and when limited evidence is sufficient to justify fleet-wide deployment.

A useful benchmark should introduce a change to one subsystem, such as perception, and measure both its direct improvement and its downstream effects on other components and end-to-end driving behavior. It should then evaluate the same change across vehicles with different system conditions and configurations, measuring whether observed gains remain consistent and whether the available evidence is sufficient to support a fleet-wide deployment decision.

4.5 Reasoning, Diagnosis, and Maintenance

Fleet memory may show what happened without explaining why it happened. Similar behavior can arise from very different causes, such as an unusual scene, perception degradation, calibration error, software regression, or hardware failure.

The open challenge is to turn heterogeneous fleet evidence into grounded, uncertainty-aware diagnoses. Causal models, multimodal retrieval, or Vision-language Model (VLM) may help retrieve evidence and compare hypotheses, but their outputs cannot be treated as ground truth; even modest representation changes can alter safety-relevant VLM decisions [14]. Diagnoses should remain traceable to supporting observations and system state, and guide maintenance and return-to-service decisions.

A useful benchmark should present similar symptoms with different known causes and measure whether the system retrieves the relevant evidence, identifies the likely cause with appropriate uncertainty, and verifies that the corrective action resolves the problem.

4.6 Simulation and Replay

Simulation can extend fleet memory by replaying real events and testing conditions that are difficult or costly to reproduce in the physical world. However, simulated environments cannot perfectly reproduce real sensor behavior, vehicle dynamics, environmental complexity, or system interactions.

The open challenge is how to account for this simulation-to-reality gap when using simulated evidence for engineering or fleet decisions. Research should quantify where simulation diverges from real-world observations, identify which conclusions remain reliable despite that gap, and determine when real-world validation is still required.

A useful benchmark should reconstruct real fleet events in simulation and compare simulated outcomes with their real-world counterparts. It should measure the gap between them and evaluate whether simulation can reliably predict the direction and magnitude of downstream effects.

4.7 Privacy and Security

Fleet memory becomes more useful when observations can be linked across time, but these links can also reveal people, places, and patterns of activity.

The privacy challenge is to preserve the temporal and spatial relations needed for fleet decisions without keeping unnecessary identity or tracking information. These restrictions must also follow derived data such as labels and datasets.

Security has a related problem: cryptographic integrity can show that data were not modified, but cannot prove that the original physical observation was correct. Research should combine sensor consistency, vehicle health, provenance, and observations from other vehicles to better judge whether physical evidence can be trusted.

4.8 Scale, Cost, and Resource Use

Robotaxi data infrastructure creates significant costs in storage, network transfer, and computation. For N vehicles generating d GB per vehicle-day,

$$D = 30Nd, \qquad D_r = rD, \tag{1}$$

where D is the 30-day raw-data volume, r is the retained fraction, and D_r is the retained volume.

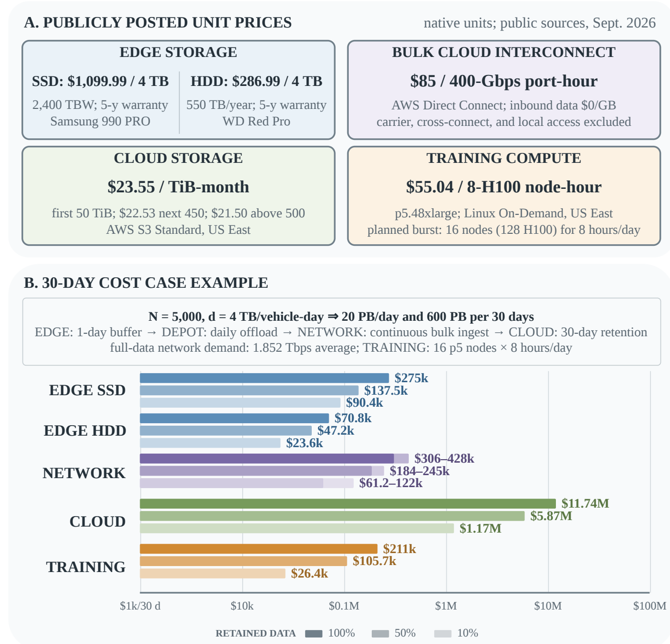

Fig. 2. Cost sensitivity to the fraction of generated data retained across edge storage, network transfer, cloud storage, and training. Part A shows the public unit prices used in the calculation; Part B compares the resulting 30-day cost exposures under 100%, 50%, and 10% retention.

The 30-day cost can be estimated as:

$$C(r) = \underbrace{\frac{30K_e(r)p_e}{L_e(r)}}_{\text{edge}} + \underbrace{720M_n(r)p_n}_{\text{network}} + \underbrace{D_r\bar{p}_s}_{\text{cloud}} + \underbrace{H_t(r)p_t}_{\text{training}}. \tag{2}$$

Here, K_e, p_e, and L_e are the number of storage devices, storage unit price, and replacement interval of storage devices; M_n and p_n are the number and hourly price of network ports; $\bar{p}_s$ is the effective cloud-storage price; and H_t and p_t are training node-hours and price per node-hour.

Figure 2(A) lists public prices used in our example [15]–[19]. Figure 2(B) applies them to 5,000 vehicles generating 4 TB/vehicle-day. If we only consider storage, network and computing cost, retaining all data produces roughly $12.3–12.7M in 30-day infrastructure exposure, while retaining 10% reduces it to about $1.28–1.41M, an approximately 89% reduction. The estimates are not intended to represent an operator's actual bill, which may differ because of private infrastructure, negotiated pricing, transfer strategies, and workload design. Instead, they show how data-retention decisions can translate into large downstream economic differences.

The open challenge is to connect technical data-management choices to their economic impact. Decisions such as what to retain at the edge, how to transfer data, how long to keep it in the cloud, and what data to use for training affect different cost terms in different ways. Reducing retained data can substantially lower downstream storage and transfer requirements, but aggressive reduction may discard rare or valuable evidence and increase later collection, replay, or retraining effort. The goal is therefore to optimize data selection, movement, storage, and computation jointly, minimizing total resource demand while preserving the evidence needed for fleet operation and learning.

5 Conclusion

A scalable robotaxi service cannot treat each trip as an isolated collection of logs. Data infrastructure gives the fleet a shared memory that connects what happened in the physical world with the state of the system that observed it. This allows events to be compared across vehicles and over time, and turns past experience into evidence for future decisions.

The fourteen-stage architecture explains how this memory is formed, organized, used for learning, tested, and returned to fleet operation. It must serve the people and vehicles operating now while supporting the engineering choices that shape future service. Although these challenges span multiple technical and operational disciplines, decision validity and end-to-end cost provide common objectives for evaluating the overall system.

Robotaxi service is a demanding case study for fleet robotics more broadly. What generalizes beyond robotaxi services is not any particular vendor's technology stack, but the underlying problem of turning distributed observations and system histories into traceable and trustworthy memory for immediate operation and long-term improvement. The same problem extends beyond robotaxis. Any large fleet of robots operating in a changing environment must learn from past experience and from one another, although each domain will define its own standards of risk and accountability.

The long-term intelligence of a robotaxi service will depend not only on what each vehicle perceives and decides in the moment, but also on what the fleet can remember about the world, how faithfully it can reconstruct change, and how responsibly it turns that memory into action.

References

[1] SAE International, *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, SAE International Std. SAE J3016_202 104, 2021, accessed: 2026-07-21. [Online]. Available: https://www.sae.org/standards/content/j3016_202104/

[2] World Wide Web Consortium, *PROV-O: The PROV Ontology*, World Wide Web Consortium Std. W3C Recommendation, 2013, accessed: 2026-07-21. [Online]. Available: https://www.w3.org/TR/prov-o/

[3] D. Baylor, E. Breck, H.-T. Cheng, N. Fiedel, C. Y. Foo, Z. Haque, S. Haykal, M. Ispir, V. Jain, L. Koc, C. Y. Koo, L. Lew, C. Mewald, A. N. Modi, N. Polyzotis, S. Ramesh, S. Roy, S. E. Whang, M. Wicke, J. Wilkiewicz, X. Zhang, and M. Zinkevich, "TFX," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017. [Online]. Available: https://doi.org/10.1145/3097983.3098021

[4] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden Technical Debt in Machine Learning Systems," in *Advances in Neural Information Processing Systems*, vol. 28, 2015, pp. 2503–2511. [Online]. Available: https://proceedings.neurips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

[5] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot Operating System 2: Design, architecture, and uses in the wild," *Science Robotics*, 2022. [Online]. Available: https://doi.org/10.1126/scirobotics.abm6074

[6] IEEE, *IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems*, IEEE Std. IEEE 1588-2019, 2020, accessed: 2026-07-21. [Online]. Available: https://standards.ieee.org/ieee/1588/6825/

[7] P. Furgale, J. Rehder, and R. Siegwart, "Unified Temporal and Spatial Calibration for Multi-Sensor Systems," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 1280–1286.

[8] M. Armbrust, T. Das, L. Sun *et al.*, "Delta Lake: High-Performance ACID Table Storage over Cloud Object Stores," *Proceedings of the VLDB Endowment*, vol. 13, no. 12, pp. 3411–3424, 2020. [Online]. Available: https://www.vldb.org/pvldb/vol13/p3411-armbrust.pdf

[9] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[10] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, 2023. [Online]. Available: https://doi.org/10.1016/j.patter.2023.100804

[11] A. Dosovitskiy, G. Ros, F. Codevilla, A. M. Lopez, and V. Koltun, "CARLA: An Open Urban Driving Simulator," in *Proceedings of the 1st Conference on Robot Learning*, vol. 78, 2017, pp. 1–16, arXiv:1711.03938; accessed 2026-07-21. [Online]. Available: https://proceedings.mlr.press/v78/dosovitskiy17a.html

[12] H. Caesar, J. Kabzan, K. S. Tan, W. K. Fong, E. Wolff, A. Lang, L. Fletcher, O. Beijbom, and S. Omari, "NuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles," http://arxiv.org/abs/2106.11810, 2021, arXiv:2106.11810; accessed 2026-07-21.

[13] E. Breck, S. Cai, E. Nielsen, M. Salib, and D. Sculley, "The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction," in *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 1123–1132.

[14] E. Richards, "Task-aligned stability analysis of vision-language models for autonomous driving hazard detection," in *2026 International Conference on Machine Learning (ICML) Workshop on Combining Theory and Benchmarks (CTB)*, 2026.

[15] Samsung Electronics America, "990 PRO PCIe 4.0 NVMe SSD, 4 TB," https://www.samsung.com/us/business/memory-storage/nvme-ssd/990-pro-pcie-4-0-nvme-ssd-4tb-sku-mz-v9p4t0b-am/, 2026, official price and specification; accessed 2026-08-31.

[16] B&H Photo Video, "WD 4TB Red Pro 7200 rpm SATA III 3.5-Inch NAS HDD," https://www.bhphotovideo.com/c/product/1830093-REG/wd_wd4005ffbx_spcaun0_red_pro_nas_hard.html, 2026, authorized-retailer price and manufacturer specifications for model WD4005FFBX; accessed 2026-09-03.

[17] Amazon Web Services, "AWS Direct Connect Pricing," https://aws.amazon.com/directconnect/pricing/, 2026, official dedicated-port and inbound-data pricing; accessed 2026-09-03.

[18] ——, "Amazon S3 Pricing," https://aws.amazon.com/s3/pricing/, 2026, public US East S3 Standard list pricing and binary-GB definition; accessed 2026-09-03.

[19] ——, "Amazon EC2 On-Demand Pricing," https://aws.amazon.com/ec2/pricing/on-demand/, 2026, public Linux On-Demand list price for p5.48xlarge in US East (N. Virginia), $55.04 per node-hour; accessed 2026-09-03.