



UZH, Dept. of Informatics, Binzmühlestr. 14, CH-8050 Zürich

Dario Alberto Monopoli
Matr. Nr: 21-717-632

Prof. Dr. Sven Helmer
Professor
Phone +41 44 635 67 28
Fax +41 44 635 68 09
helmer@ifi.uzh.ch

Zürich, September 10, 2026

MSc Thesis (30 points)

Topic: Time Series Foundation Models for Observability

Background and Motivation

Outages are expensive: the Uptime Institute's 2026 analysis reports that roughly 20% of major outages cost more than \$1M; ITIC (2024) finds 41% of enterprises put a single hour of downtime between \$1M and \$5M+. IBM (2024) reports that teams using AI and automation detect and contain incidents about 100 days faster.

At the same time the data keeps growing and the existing tools are noisy. In Chronosphere's 2024 survey log data grew about 250% year over year, and alert noise is a real problem: in Splunk's 2025 survey, 73% of teams reported outages caused by alerts that were ignored or suppressed. In Grafana's 2024 survey, more than three quarters of teams named anomaly detection as the AI capability they most want, yet most still detect problems with static thresholds and manual rules. Time series foundation models (TSFMs) seem to fit this task well.

Problem Statement

Almost all existing work points these models at metrics, not at signals built from logs. Log-derived series such as template counts and error rates are spiky, bursty, and unlike the clean metrics on which TSFMs are trained and evaluated, so it is not established that the models produce a meaningful anomaly signal on them at all. A second difficulty is heterogeneity: different companies and services log very differently, so one model for all of them is a poor fit.

A third is supervision, since ground truth comes from incident records and alert history, which are event-level and do not line up with the count series produced from logs. Recent work also warns that fine-tuned foundation models do not always beat small dedicated models [11] and that foundation models do not automatically win at anomaly detection [12]. Accordingly, the thesis addresses two questions:

1. Does clustering then fine-tuning one Chronos model per cluster beat, on the same data and metrics, (a) one global fine-tuned model, (b) zero-shot Chronos, (c) a single mixture-of-experts model [7], and (d) a cheap classical baseline?
2. Do time series foundation models work on log-derived signals like template counts and error rates at all, not just on clean metrics?

Proposed Approach

The overarching goal of this thesis is to parse logs into time series, cluster the log sources, and fine-tune one TSFM per cluster for anomaly detection. We propose exploring the following main ideas:

1. **Logs to Time Series:** Raw logs are mostly fixed templates with variable parameters. Using a streaming parser such as Drain3 [1] we map each log line to a stable template ID, and counting how often each template fires per time window turns a log stream into a set of time series: event-count series, error-rate series, and per-service counts [2].
2. **Clustering the Sources:** We group the series into clusters using shape-based features such as catch22 [3] or k-Shape distances [4], and we pick the number of clusters with a standard criterion such as the silhouette score. A model trained on one cluster should fit it better than a global model that averages over all of them. We test whether clustering helps by adding a baseline that fine-tunes one model globally with no clustering.
3. **Fine-Tuning per Cluster:** For each cluster we fine-tune a Chronos [5] model on that cluster's series, producing probabilistic forecasts. The Chronos-Bolt variant is small (roughly 9 to 200M parameters) and runs on CPU, so hosting tens of per-cluster models stays affordable. Chronos-2 [6] and Moirai-MoE [7] are the alternatives we benchmark against.
4. **Anomaly Detection:** With a fine-tuned model per cluster we flag anomalies where real values fall far outside the model's predicted range. Because the forecasts are probabilistic, the size of the deviation gives a natural anomaly score.

Data

We work on anonymised internal Dynatrace observability data. No NDA or IP agreement is required from UZH: the thesis itself is public and the data is never shared, with only aggregate, anonymised results appearing in the thesis.

Primary Goals and Tasks

The project will focus on the following core objectives:

1. **Feasibility Probe:** Run zero-shot inference first, on one service's count series, to check that the models produce a meaningful anomaly signal on log-derived series at all. If they do not, that is itself an early finding, and we revise the plan before spending time on clustering and fine-tuning.
2. **Implementation:** Build the pipeline from raw logs to time series (Drain3 parsing, template-count and error-rate aggregation), the clustering stage, and the per-cluster fine-tuning and scoring code.
3. **Label Alignment:** Derive window-level ground truth from event-level incident records and alert history. Use fixed time windows (for example one minute), a rule that a window is anomalous if it overlaps a recorded incident, and injected anomalies with known timestamps as a controlled check where incident coverage is thin. Run a pilot on one service to confirm this gives sensible labels before scaling, and report how sensitive the results are to the window size.
4. **Evaluation:** Empirically compare the per-cluster approach against zero-shot Chronos, one global fine-tuned Chronos, a single mixture-of-experts model, classical detectors (Isolation Forest, Matrix Profile), and a production detector, all on the same data and metrics. Use chronological train and test splits to avoid leaking the future into training, and threshold-free metrics like VUS-PR [8] and affiliation-based precision and recall [9] instead of point-adjusted F1 [10]. Include an ablation analysis and a cost and scaling analysis. Analyse the false positives and missed anomalies on a validation split and improve the per-cluster approach through continuous refinement, keeping the held-out test split for the final numbers.
5. **Write-Up:** Describe the implementation, results, evaluation, and other findings in a thesis, and present and defend it in a group meeting.

Project Schedule (26 Weeks)

The thesis is planned to be completed over a 26-week period with the following tentative schedule:

- **Weeks 1–4 (October):** Data access, pipeline, and the zero-shot feasibility probe on count series.
- **Weeks 5–9 (November):** Drain3 parsing, clustering, and the label-alignment pilot on one service.
- **Weeks 10–13 (December):** Core baselines complete, all on the same data and metrics.
- **Weeks 14–18 (January):** Per-cluster fine-tuning and comparison with the mixture-of-experts model.

- **Weeks 19–22 (February):** Full evaluation, including ablation analysis and continuous refinement.
- **Weeks 23–26 (March):** Write-up, including cost and scaling analysis.

University of Zurich
Department of Informatics



Prof. Dr. Sven Helmer
Professor

Candidate: Dario Alberto Monopoli, Matr. Nr: 21-717-632

External supervisor: Dr. Martin Flechl, Head of Generative AI Research, Dynatrace Research, Vienna

Start date: October 1st, 2026

End date: March 31st, 2027

References

- [1] P. He, J. Zhu, Z. Zheng, M. R. Lyu, Drain: An Online Log Parsing Approach with Fixed Depth Tree, ICWS 2017, pp. 33–40.
- [2] S. He, J. Zhu, P. He, M. R. Lyu, Experience Report: System Log Analysis for Anomaly Detection, ISSRE 2016, pp. 207–218.
- [3] C. H. Lubba, S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, N. S. Jones, catch22: CAnonical Time-series CHaracteristics, Data Mining and Knowledge Discovery 33(6):1821–1852, 2019.
- [4] J. Paparrizos, L. Gravano, k-Shape: Efficient and Accurate Clustering of Time Series, SIGMOD 2015, pp. 1855–1870.
- [5] A. F. Ansari et al., Chronos: Learning the Language of Time Series, Transactions on Machine Learning Research, 2024. arXiv:2403.07815.
- [6] A. F. Ansari et al., Chronos-2: From Univariate to Universal Forecasting, arXiv:2510.15821, 2025.
- [7] X. Liu et al., Moirai-MoE: Empowering Time Series Foundation Models with Sparse Mixture of Experts, ICML 2025. arXiv:2410.10469.
- [8] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, M. J. Franklin, Volume Under the Surface: A New Accuracy Evaluation Measure for Time-Series Anomaly Detection, PVLDB 15(11):2774–2787, 2022.
- [9] A. Huet, J. M. Navarro, D. Rossi, Local Evaluation of Time Series Anomaly Detection Algorithms, KDD 2022, pp. 635–645.
- [10] S. Kim, K. Choi, H.-S. Choi, B. Lee, S. Yoon, Towards a Rigorous Evaluation of Time-Series Anomaly Detection, AAAI 2022, 36:7194–7201.
- [11] N. Karaouli, D. Coquenot, E. Fromont, M. Mermillod, M. Reyboz, How Foundational are Foundation Models for Time Series Forecasting?, NeurIPS 2025 Workshop on Recent Advances in Time Series Foundation Models. arXiv:2510.00742.
- [12] C. Shyalika, H. K. Bagga, A. Bhatt, R. Prasad, A. Al Ghazo, A. Sheth, Time Series Foundational Models: Their Role in Anomaly Detection and Prediction, AAAI 2025 Workshop on Anomaly Detection in Scientific Domains. arXiv:2412.19286.

