SANER 2026

Empirical Characterization of Logging Smells in Machine Learning Code.

GIGL, Polytechnique Montréal

By Patrick Loic Foalem, Leuson Da Silva, Foutse Khomh, Heng Li, Ettore Merlo

OUTLINE

- MOTIVATION
- PRIOR STUDY
- RESEARCH GAP
- RR PROTOCOL
- EXPECTED CONTRIBUTIONS
- TAKEAWAYS

MOTIVATION

What is logging in software engineering?

★ Common practice in software engineering is to record runtime system information.

Logging code

System in production

Log messages

```
logging.info("Training accuracy {:.2f}" .format(accuracy))
```

Verbosity Level

Static Message

Dynamic Variable

```
2022-02-01 16:10:08 INFO Training accuracy 0.79
```

Why Logging Matters Even More in ML Systems?

Engineering AI Systems: A Research Agenda

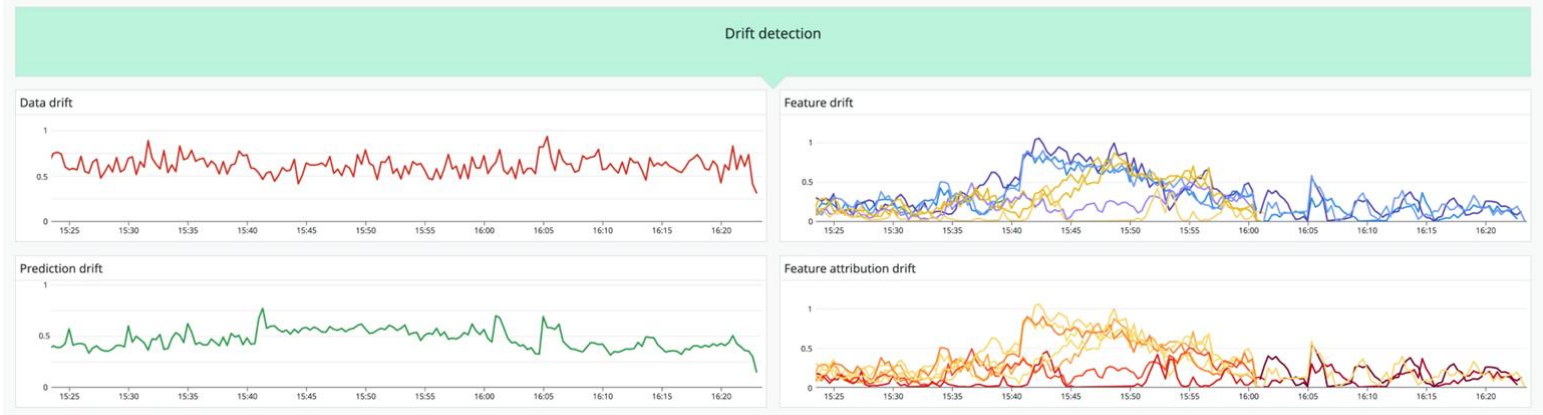
Jan Bosch
Computer Science and Engineering
Chalmers University of Technology
Gothenburg, Sweden
jan.bosch@chalmers.se

Helena Holmstrm Olsson
Computer Science and Media Technology
Malm University
Malm, Sweden
helena.holmstrom.olsson@mau.se

Ivica Crnkovic
Computer Science and Engineering
Chalmers University of Technology
Gothenburg, Sweden
ivica.crnkovic@chalmers.se

- Data drift
- Reproducibility of results
- Overfitting
- Monitoring and Logging

identified problems		strategic focus
Lack of labelled data Lack of metadata Shortage of diverse samples Heterogeneity in data Data granularity Imbalanced data sets	Data drift Data dependencies Managing categorical data Managing sequences in data Deduplication complexity Data streams for training	data quality management
Experiment management Dependency management Unintended feedback loops Effort estimation Cultural differences Specifying desired outcome	Lack of modularity Sharing and tracking techn. Reproducibility of results Data extraction methods Tooling	design methods and processes
Overfitting Scalable ML pipeline Quality attributes Statistical Understanding	Limited transparency Training/serving skew Sliced analysis of final model	model performance
Monitoring and Logging Testing Troubleshooting Data sources and distribution Glue code and support	Privacy and data safety Data silos Data storage Resource limitations	deployment & compliance



https://www.datadoghq.com/blog/ml-model-monitoring-in-production-best-practices/

```python
import numpy as np
import bentoml
from bentoml.io import Text
from bentoml.io import NumpyNdarray

CLASS_NAMES = ["setosa", "versicolor", "virginica"]

iris_clf_runner = bentoml.sklearn.get("iris_clf:latest").to_runner()

svc = bentoml.Service("iris_classifier", runners=[iris_clf_runner])

@svc.api(
    input=NumpyNdarray.from_sample(np.array([4.9, 3.0, 1.4, 0.2], dtype=np.double)),
    output=Text(),
)
async def classify(features: np.ndarray) -> str:
    with bentoml.monitor("iris_classifier_prediction") as mon:
        mon.log(features[0], name="sepal_length", role="feature", data_type="numerical")
        mon.log(features[1], name="sepal_width", role="feature", data_type="numerical")
        mon.log(features[2], name="petal_length", role="feature", data_type="numerical")
        mon.log(features[3], name="petal_width", role="feature", data_type="numerical")

        results = await iris_clf_runner.predict.async_run([features])
        result = results[0]
        category = CLASS_NAMES[result]

        mon.log(category, name="prediction", role="prediction", data_type="categorical")
    return category
```

https://www.bentoml.com/blog/a-guide-to-ml-monitoring-and-drift-detection

Is logging within applications challenging?

Too **much** : runtime overhead, storage overhead, too many trivial logs.

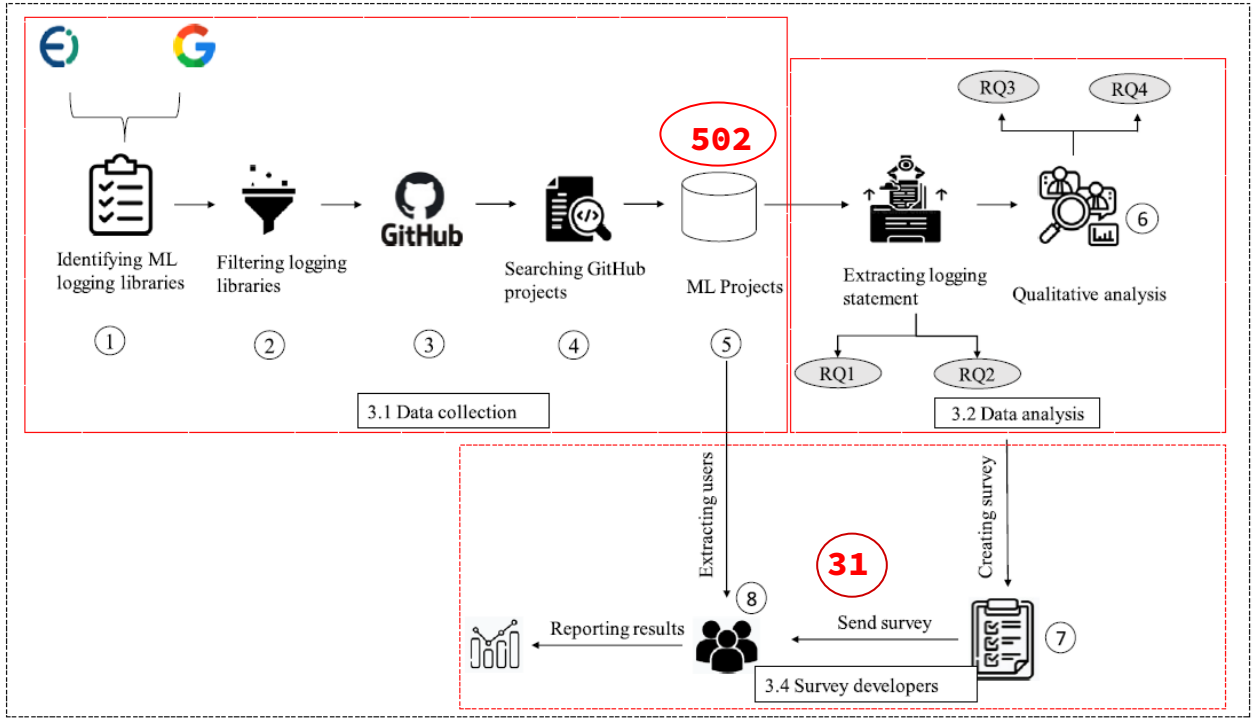
Too **little**: missing important information, increasing the difficulty for problem diagnose

PRIOR STUDY

Our Prior Study: Logging Practices in ML Systems.

- 502 open-source ML repositories from GitHub
- Python ML ecosystem
- Analysis of logging statements and frameworks
- Practitioner survey

Study Design



Replication package
Try it out!

Finding 1: Logging is Less Prevalent in ML Systems

Result (1): Prevalence of Logging in ML vs. Traditional Applications

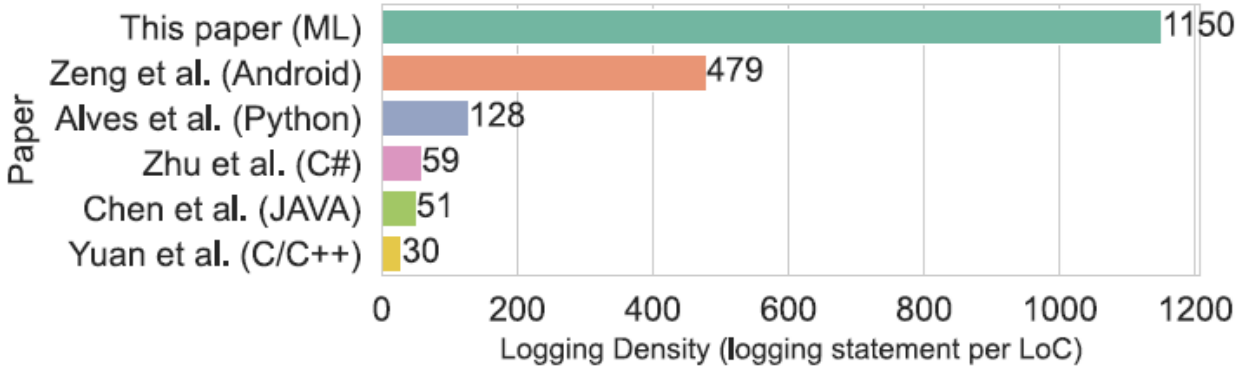


Fig. 4. Logging density across different papers.

Metric:
Log density = SLOC/NL

Finding 1: Our study reveals that logging practices in ML-based applications are less prevalent compared to traditional software systems.

Finding 2: Hybrid Logging Ecosystem.

Result (2): Logging libraries types.

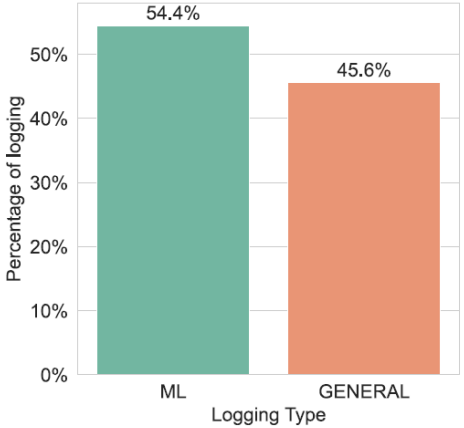


Fig. 6. Logging type distribution.

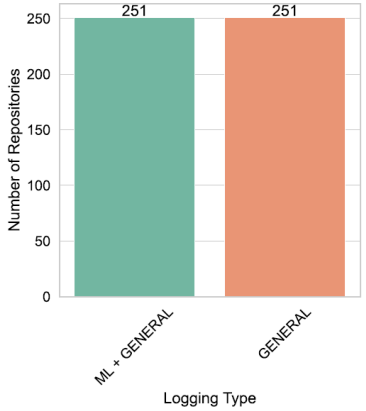


Fig. 7. Distribution of number of repository by logging type.

Logging libraries type	Libraries
ML-specific logging	Comet_ml Whylogs Wandb Tensorboard MLflow Tensorflow Neptune Dowel Sacred Ml-logger
General logging libraries	Logging Warnings

60% of 31 practitioners

Finding 2:

- Traditional applications primarily rely on general logging libraries.
- ML-based applications use a combination of ML-specific and general logging libraries.

Finding 3: Logging Concentrated in Training.

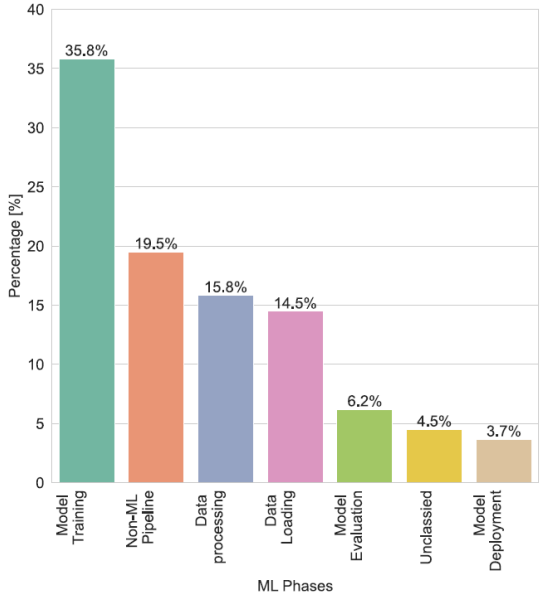


Fig. 12. Percentage of logging in different ML phases.

- Model training: need for extensive experimentation and configuration adjustments.
- Model Deployment: monitoring inference performance and detecting anomalies.

Finding 3: All five phases of the entire ML pipeline contain logging statements. However, the **model training** phase has the largest proportion of logging statements and the **model deployment** phase has the smallest proportion of logging statements.

RESEARCH GAP

Research Gap: Understanding Logging Quality in ML Systems.

What we know

- ✓ Logging used in ML.
- ✓ Hybrid logging tools.
- ✓ Logging across pipeline.

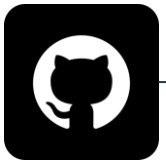
What we don’t know

- ✗ What logging smells exist?
- ✗ ML-specific bad practices?
- ✗ Practitioner perception?

Research Opportunity.

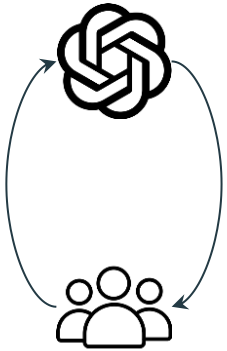
- Identify ML logging smells.
- Build an empirical taxonomy
- Understand practitioner perceptions

RR PROTOCOL



444 ML GitHub repositories

Extract function + logging statements

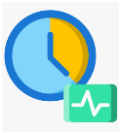


LLM + Human

Smell Taxonomy

Survey Practitioners

Agreement

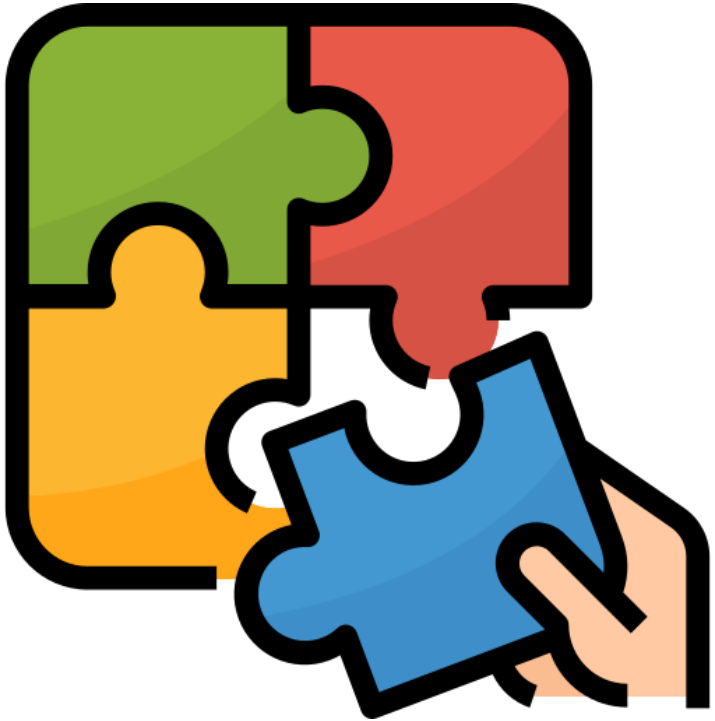


Frequency

Relevance

Severity

5-point Likert Scale



EXPECTED CONTRIBUTIONS

Expected contributions.

Empirical Catalog
Logging smells in ML systems

Large-scale Evidence
Mining ML repositories

Practitioner Insights
Survey of ML developers

Impact:
Improving observability in ML systems

TAKEAWAYS

Takeaways.

Logging plays a critical role in ML systems;

ML logging practices remain poorly understood;

Our study aims to systematically identify ML logging smells.