LLM meets ML: Data-efficient Anomaly Detection on Unstable Logs

Fatemeh Hadadi
University of Ottawa

Qinghua
Lero centre and University of Limerick

Domenico Bianculli
University of Luxembourg

Lionel Briand
University of Ottawa
Lero centre and University of Limerick

September 2025

uOttawa

uOttawa

Introduction: Reliability and Availability Impact

- **Software Dependability is Critical.**
  - Rising system complexity demands higher reliability.
  - Systems are growing smarter and riskier. AI now writes 25%+ of Google's new code.

- **Key Post-development Reliability Strategies**

✓ Proactive: Failure Prediction

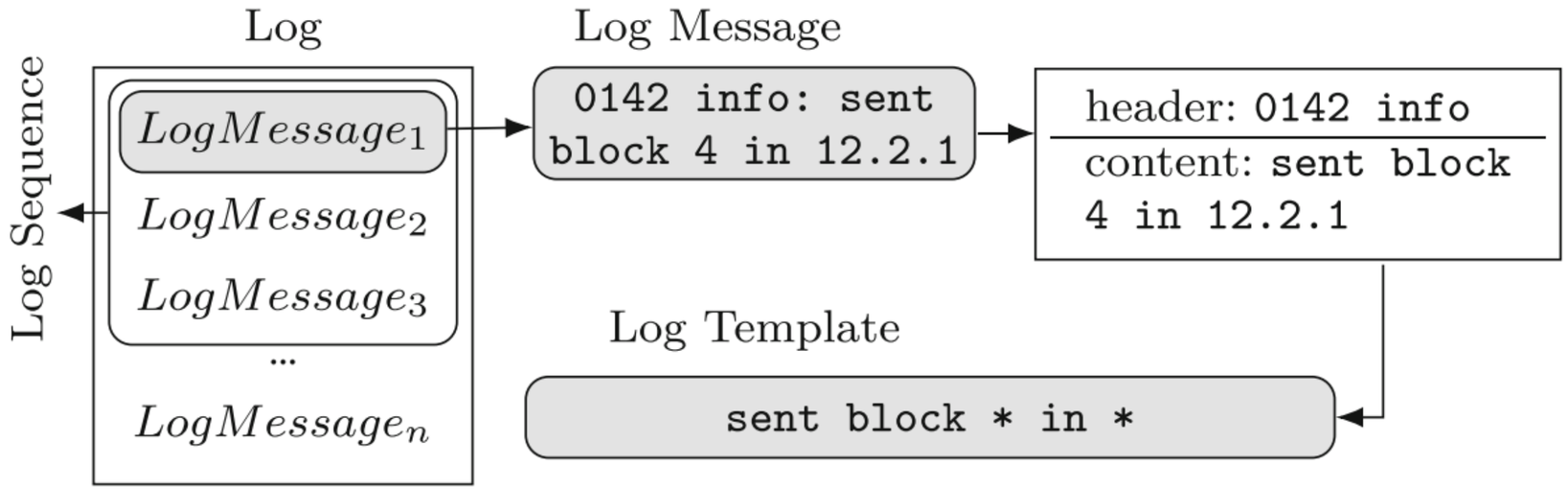
✓ Reactive: Anomaly Detection

uOttawa

Log Artifacts, a Valuable Resource for Insight

Log files capture detailed information during system execution.

- **Challenges & Opportunities**
  - Logs are **semi-structured.** → **Log Preprocessing**



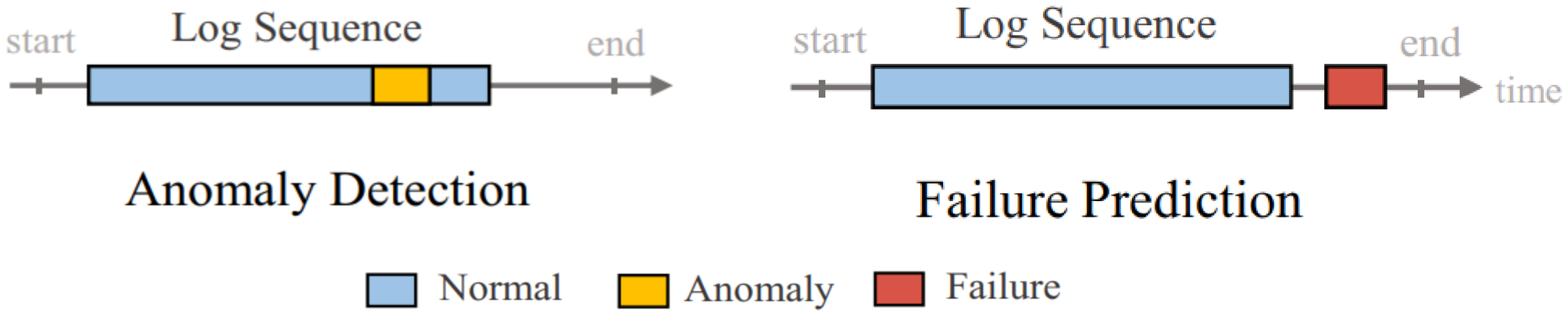
  - Modern systems generate **large volumes** of log data. → **Automatic Method**

uOttawa

Background: Log Analysis Tasks

- **Anomaly detection** is the task of identifying anomalous patterns in log data that do not conform to expected system behaviors.

- **Failure prediction** attempts to proactively generate alerts before the occurrence of failures, which often lead to unrecoverable outage.



Main differences: Mode of operation (reactive vs proactive) Input Data

uOttawa

Main challenges in Log Analysis

Log-based Failure Prediction

- **No Systematic Study** to evaluate the impact of different DL network and embedding strategies.
- **Lack Labeled Datasets** with different characteristics.

Log-based Anomaly Detection

- **Unstable Logs:** Most methods assume stable logs, but logs change with software and environment updates.
- **Substantial Reliability on Labeled Data:** ML, especially DL models, need large labeled datasets, which are costly.
- **Data Leakage:** Many datasets contain leaks, leading to inflated DL performance.

uOttawa

Motivation

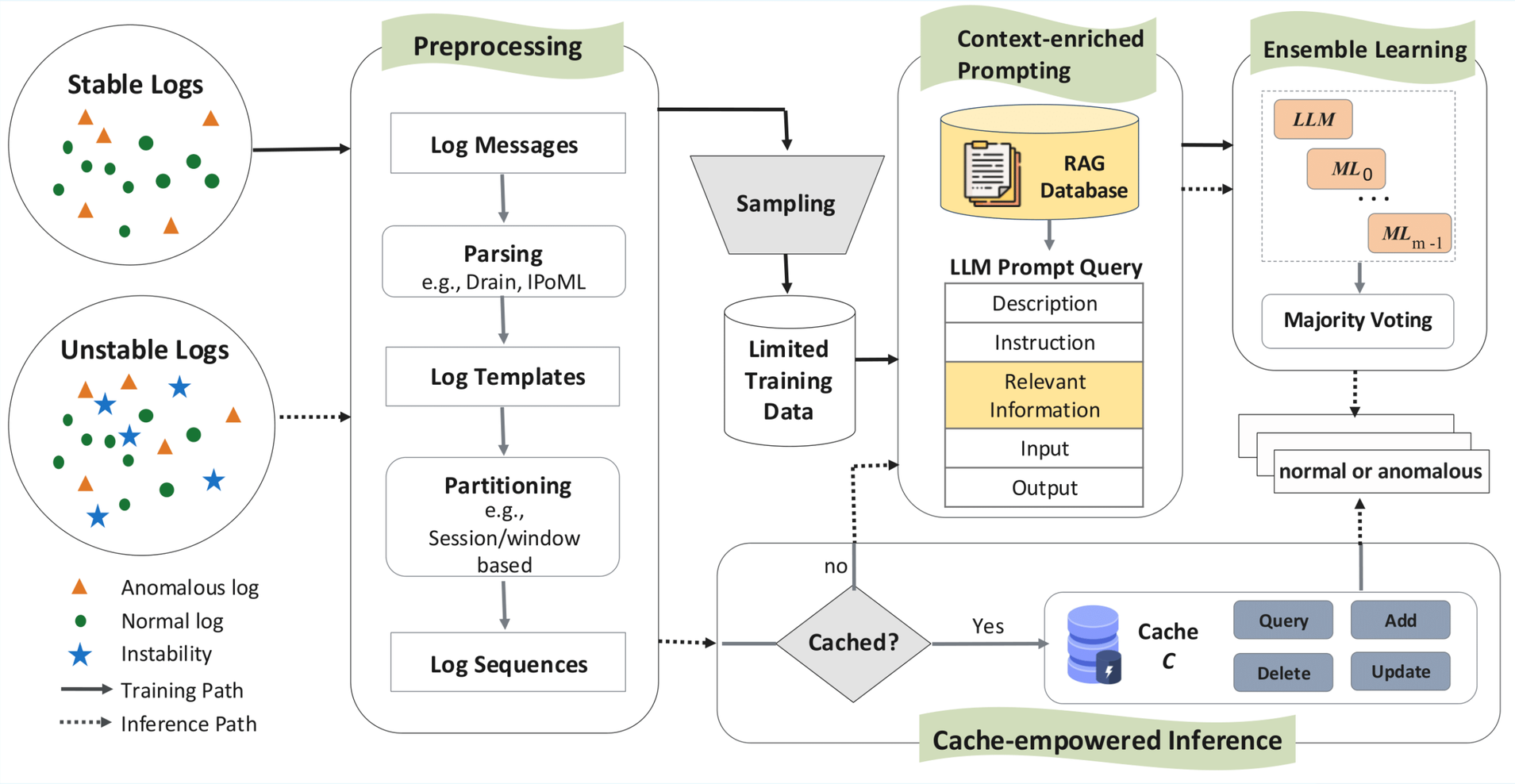
Exiting LAD Solutions mostly

- assume a **stable** data distribution,
- rely on **substantial labeled** training data,
- overlook the **contextual information** of logs,
- prone to **inflated effectiveness** due to data leakage issues.

uOttawa

Proposed Approach

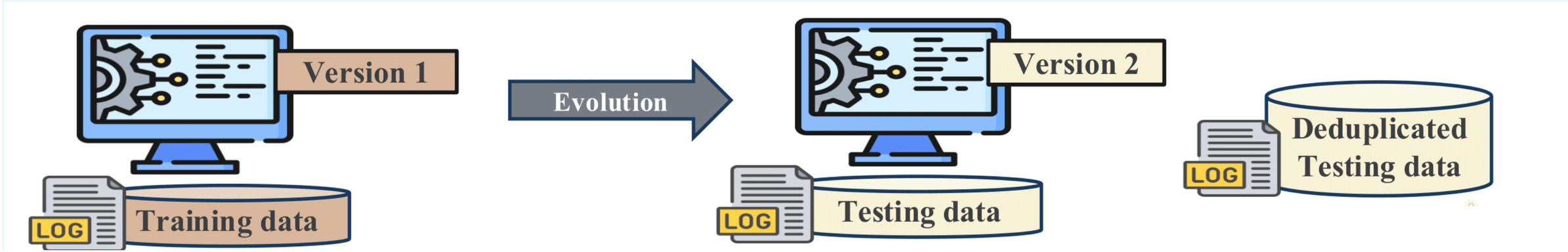
We propose **FlexLog**, a **data-efficient** anomaly detector on **unstable** datasets.



uOttawa

Unstable Data Configuration

Models are trained on stable data and tested on unstable data.



✓ **Expanded datasets to 4** (real-world + synthetic):
Real-world: ADFA-U, LOGEVOL-U, **Synthetic:** SynHDFS-U, SYNEVOL-U

✓ Removed overlapping test logs to **prevent information leakage**.

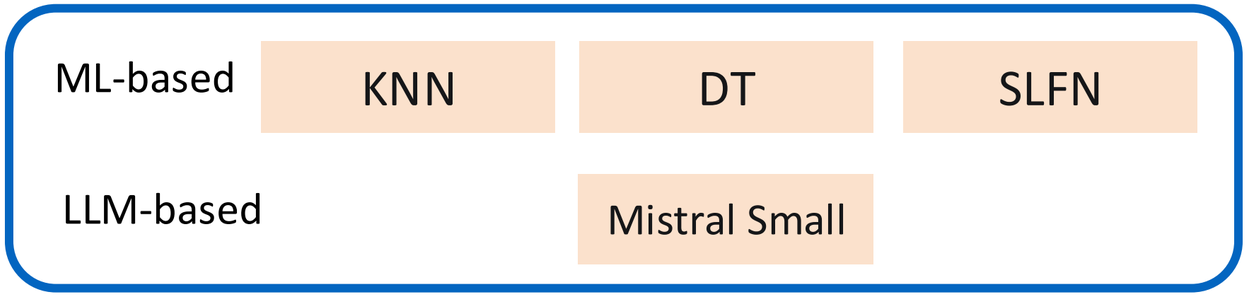
uOttawa

Experiment Setting

✓ We evaluated **nine ML baselines**: four unsupervised, one semi-supervised, and four supervised methods.

Supervised	Semi-supervised	Unsupervised
LightAD	PLELog	DeepLog
NeuralLog		LogAnomaly
LogRobust		LogCluster
CNN		PCA

✓ **Empirical Configuration of FlexLog:**



Evaluation Metrics and Statistical Testing:

Effectiveness: P (Precision), R (Recall), F1 (F1-score),

Efficiency: $\Delta u_{\%}$(labeling cost reduction), T (Training Time), I (Inference Time)

Statistical Testing: Mann-Whitney U test (also called Wilcoxon rank-sum test)

uOttawa

RQ1: How effective is FlexLog for ULAD compared to the baselines?

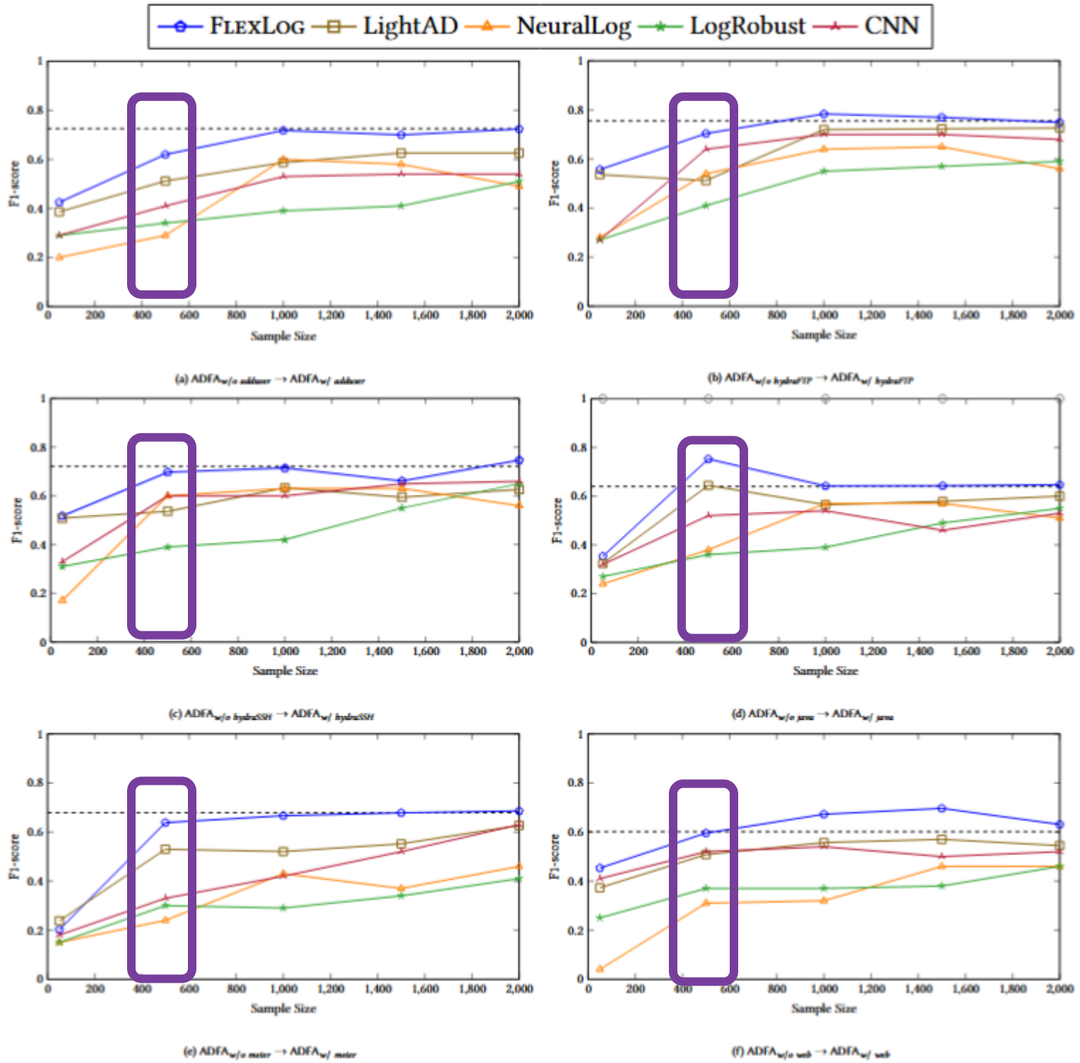
FlexLog achieves **state-of-the-art effectiveness** on both **real-world and synthesized datasets** while exhibiting **high data efficiency**.

On synthesized datasets, FlexLog **remains effective** under up to 30 % sequence- and template-level instability.

Data	Unstable	M	limited data	full training set			
				Supervised			
			FlexLog	LightAD	NeuralLog	LogRobust	CNN
ADFA	No	P	0.708	0.820	0.538	0.718	0.666
		R	0.894	0.814	0.602	0.708	0.842
		F1	0.791	**0.817**	0.568	0.713	0.744
average	Yes	P	0.633	0.747	0.357	0.667	0.748
		R	0.799	0.660	0.652	0.402	0.657
		F1	**0.704**	0.677	0.433*	0.494*	0.685
$Hadoop_{2\to2}$	No	P	0.999	0.999	0.997	0.984	0.997
		R	0.986	0.994	0.986	0.976	0.997
		F1	0.993	**0.997**	0.992	0.980	**0.997**
$Spark_{2\to2}$	No	P	0.999	0.999	0.999	0.941	0.999
		R	0.969	0.939	0.636	0.969	0.878
		F1	**0.984**	0.968	0.777	0.952	0.935
$Hadoop_{2\to3}$	Yes	P	0.998	0.998	0.914	0.905	0.992
		R	0.965	0.963	0.984	0.950	0.968
		F1	**0.982**	0.980	0.948	0.927	0.980
$Spark_{2\to3}$	Yes	P	0.999	0.981	0.916	0.696	0.992
		R	0.805	0.708	0.766	0.832	0.736
		F1	**0.892**	0.829	0.834	0.757	0.840
Average ($Spark_{2\to3}$, $Hadoop_{2\to3}$)	Yes	P	0.998	0.99	0.915	0.786	0.992
		R	0.871	0.83	0.875	0.863	0.852
		F1	**0.928**	0.898*	0.891*	0.833*	0.910*

* FlexLog yields a significant higher F1-score than baseline.

RQ2: How does the amount of labeled data impact FlexLog's data effectiveness?



FlexLog outperforms baselines in F1 scores **under varying data scarcity levels** of ADFA-U except at dataset size = 50, where all methods perform poorly due to insufficient data.

uOttawa

RQ3: How time-efficient is FlexLog, and what is the cache's memory overhead?

Data	Config	M	supervised				
			FlexLog	LightAD	NeuralLog	LogRobust	CNN
ADFA-U	adduser	T	16 161	5	922	221	271
		I	0.842	0.012	0.664	0.234	0.164
	hydraFTP	T	15 906	4	920	220	269
		I	0.818	0.014	0.635	0.229	0.165
	hydraSSH	T	14 147	5	923	220	269
		I	0.832	0.014	0.655	0.232	0.164
	java	T	13 933	6	921	220	270
		I	0.875	0.014	0.673	0.236	0.164
	web	T	14 079	5	918	220	270
		I	0.864	0.013	0.679	0.231	0.165
	meter	T	15 324	3	921	220	269
		I	0.888	0.014	0.682	0.238	0.165
LOGEVOL-U	Hadoop	T	4 031	30	1 550	178	316
		I	0.435	0.067	0.275	0.078	0.077
	Spark	T	23 706	0.2	712	232	136
		I	0.844	0.038	0.564	0.082	0.072
SynHDFS-U	average	T	13 587	22	1 260	293	355
		I	0.771	0.005	0.259	0.008	0.016
SYNEVOL-U	average	T	23 706	0.2	712	232	136
		I	0.988	0.038	0.703	0.116	0.076

While FlexLog is not the most time-efficient in ULAD inference, it processes each log sequence within **1 s on average**.

FlexLog's cache memory remains **below 4 MB** for most datasets (up to 19.6 MB for ADFA-U), confirming its memory efficiency.

RQ4: How does the performance of FlexLog vary under different configurations?

Config	ADFA-U	LOGEVOL-U	SYNEVOL-U	SynHDFS-U
FlexLog	**0.853**	**0.628**	**0.941**	**0.771**
FlexLog w/o cache	0.904	0.940	0.988	0.794

Config	adduser	hydraFTP	hydraSSH	java	meter	web	Average
FlexLog	0.718	0.784	0.723	0.642	0.682	0.672	**0.704**
FlexLog w/o RAG	0.688	0.705	0.648	0.622	0.659	0.641	0.660

Config	ADFA-U							LOGEVOL-U			SynHDFS-U	SYNEVOL-U
	adduser	hydraFTP	hydraSSH	java	meter	web	average	Hadoop	Spark	average	average	average
FlexLog	**0.718**	**0.784**	**0.723**	**0.642**	**0.682**	**0.672**	**0.704**	0.982	0.892	0.937	**0.972**	0.971
w/o Mistral	0.645	0.769	0.692	0.628	0.639	0.616	0.664*	0.975	0.841	0.908*	0.939*	0.936*
w/o KNN	0.683	0.768	0.654	0.621	0.651	0.579	0.659*	0.980	0.892	0.936†	0.945*	0.971†
w/o DT	0.667	0.749	0.690	0.640	0.654	0.623	0.670*	0.973	0.875	0.924†	0.948*	0.979*
w/o SLFN	0.677	0.691	0.613	0.641	0.647	0.556	0.637*	0.978	0.871	0.924†	0.934*	0.948*
w/o Simples	0.579	0.591	0.630	0.628	0.674	0.571	0.612*	**0.998**	**0.962**	**0.980‡**	0.928*	**0.979†**

Config	Source	ADFA-U	LOGEVOL-U	SynHDFS-U	SYNEVOL-U
FlexLog	open	0.704	**0.928**	0.972	**0.971**
FlexLog (*Mistral* → *Llama*)	open	0.664*	0.895*	0.949*	0.970†
FlexLog (*Mistral* → *GPT*)	closed	**0.710†**	0.919†	**0.976†**	**0.971†**

The **full FlexLog configuration**—combining cache, RAG, and an ensemble of KNN, DT, SLFN, and Mistral—performs best, with each component improving efficiency or effectiveness.

Among LLMs, **Mistral is a cost-effective choice**, matching the performance of GPT-4o while outperforming Llama.

uOttawa

Publication

- This work is accepted in *ACM Transactions on Software Engineering and Methodology (TOSEM)*.

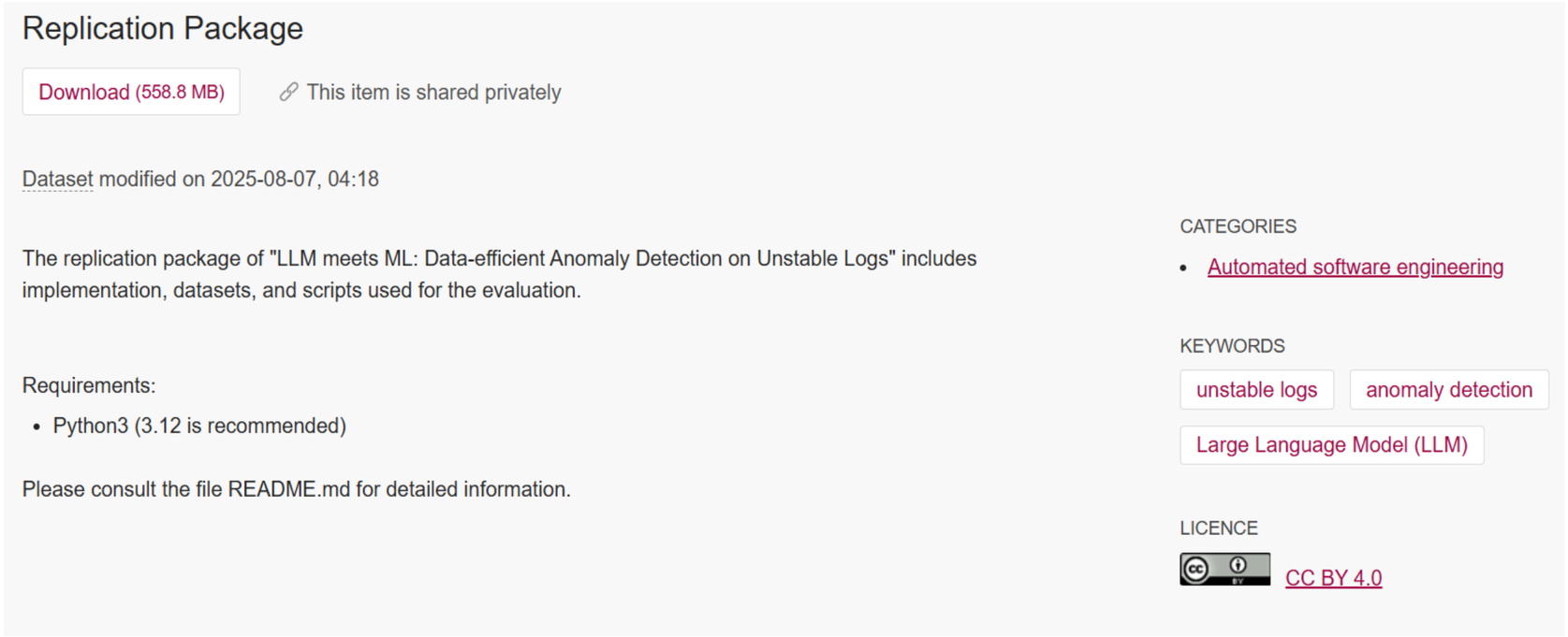
LLM meets ML: Data-efficient Anomaly Detection on Unstable Logs

FATEMEH HADADI, University of Ottawa, Canada
QINGHUA XU, Research Ireland Lero Centre for Software and University of Limerick, Ireland
DOMENICO BIANCULLI, University of Luxembourg, Luxembourg
LIONEL BRIAND, University of Ottawa, Canada and Research Ireland Lero Centre for Software and University of Limerick, Ireland

uOttawa

Date and Code

- We have included a replication package including datasets, FlexLog and baselines, and statistical evaluation.



Replication Package

uOttawa

Summary

- ✓ We introduce FlexLog, a hybrid ML + LLM ensemble for unstable logs.
- ✓ FlexLog outperforms baselines by ≥1.2 pp F1 using 63% less labeled data.
- ✓ Inference remains efficient (<1 s per log sequence).
- ✓ Caching and RAG modules improve robustness and efficiency.

uOttawa

Future Work

- ✓ Integrate stronger open-source LLMs (e.g., DeepSeek R1).
- ✓ Develop dynamic ensemble strategies beyond majority voting.
- ✓ Extend caching to handle approximate log similarities.
- ✓ Use data augmentation (GANs, synthetic logs) to balance datasets.
- ✓ Explore agentic detection via repair–execute–feedback loops with LLMs.

uOttawa

Acknowledgment

LLM meets ML: Data-efficient Anomaly Detection on Unstable Logs

Fatemeh Hadadi
University of Ottawa

Qinghua
Lero centre and University of Limerick

Domenico Bianculli
University of Luxembourg

Lionel Briand
University of Ottawa
Lero centre and University of Limerick

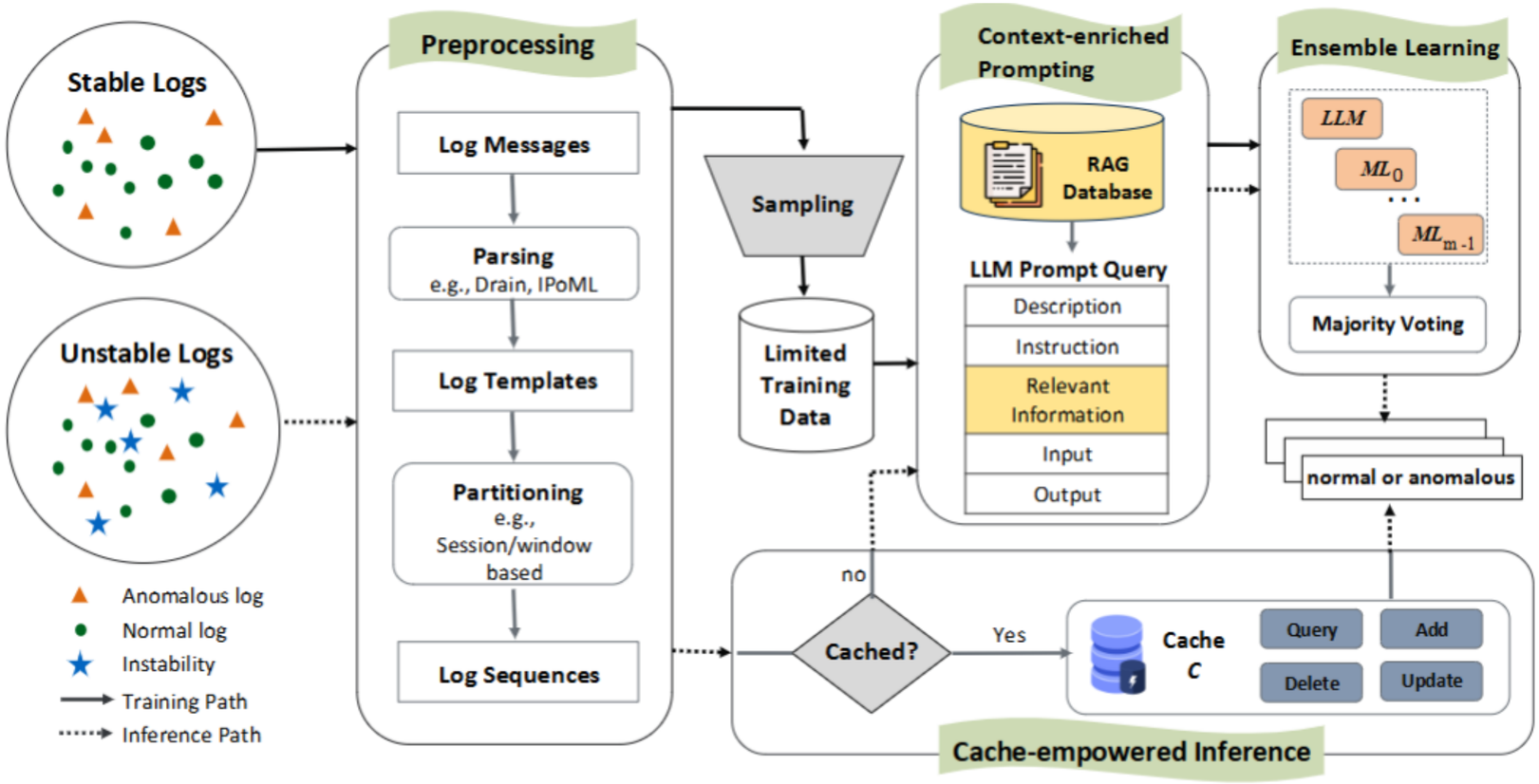
September 2025

uOttawa

uOttawa

Proposed Approach

FlexLog combines ML-based methods with fine-tuned LLMs by **ensemble averaging**.
Contextual information are integrated by LLMs using **RAG-based prompting**.
The speed of inference is improved by **cache-empowered inference**.



uOttawa

Impact of De-duplication

Table A.2: Effectiveness of FLEXLOG and baselines for ULAD on ADFA-U, LOGEVOL-U, SynHDFS-U, and SYNEVOL-U with and without de-duplication.

Model	Dedup	ADFA-U							LOGEVOL-U			SynHDFS-U	SYNEVOL-U
		adduser	hydraFTP	hydraSSH	java	meter	web	average	Hadoop	Spark	average	average	average
FLEXLOG	Yes	0.718	0.784	**0.723**	0.642	0.682	0.672	0.704	0.982	0.892	**0.937**	0.972	**0.971**
	No	0.723	0.790	0.726	0.652	0.701	0.673	0.711	0.996	0.895	0.945	0.974	0.987
LightAD	Yes	0.725	0.754	0.666	0.639	0.679	0.602	0.677	0.980	0.829	0.898	0.959	0.956
	No	0.745	0.778	0.745	0.646	0.695	0.618	0.704*	0.995	0.876	0.935*	0.968	0.981*
NeuralLog	Yes	0.412	0.388	0.368	0.511	0.461	0.457	0.433	0.948	0.834	0.891	0.946	0.765
	No	0.606	0.681	0.601	0.570	0.645	0.492	0.599*	0.961	0.859	0.910*	0.951	0.905*
LogRobust	Yes	0.524	0.408	0.374	0.558	0.651	0.449	0.494	0.927	0.757	0.833	0.760	0.941
	No	0.636	0.597	0.504	0.662	0.660	0.496	0.592*	0.981	0.789	0.885*	0.929*	0.966*
CNN	Yes	0.641	0.750	0.750	0.635	0.711	0.621	0.685	0.980	0.840	0.910	0.942	0.918
	No	0.703	0.765	0.777	0.644	0.724	0.634	0.707*	0.989	0.863	0.925	0.961*	0.925
PLELog	Yes	0.361	0.388	0.473	0.430	0.253	0.233	0.356	0.709	0.165	0.437	0.499	0.164
	No	0.405	0.428	0.494	0.443	0.311	0.277	0.393*	0.761	0.223	0.492*	0.528*	0.236*
LogAnomaly	Yes	0.291	0.451	0.480	0.422	0.218	0.343	0.368	0.310	0.216	0.263	0.446	0.304
	No	0.305	0.464	0.486	0.399	0.237	0.351	0.373	0.619	0.212	0.415*	0.498*	0.336*
DeepLog	Yes	0.292	0.481	0.458	0.339	0.253	0.353	0.363	0.367	0.122	0.244	0.741	0.253
	No	0.340	0.499	0.450	0.341	0.249	0.369	0.374	0.685	0.141	0.413*	0.776*	0.291*
LogCluster	Yes	0.334	0.418	0.461	0.348	0.175	0.304	0.340	0.430	0.485	0.458	0.519	0.714
	No	0.326	0.431	0.523	0.317	0.211	0.336	0.357*	0.798	0.614	0.706*	0.759*	0.786*
PCA	Yes	0.165	0.144	0.140	0.158	0.241	0.197	0.174	0.360	0.108	0.234	0.404	0.103
	No	0.155	0.152	0.163	0.277	0.211	0.198	0.192*	0.454	0.097	0.275*	0.567*	0.189*

* The same model shows a significant F1 score difference between deduplicated and original test data.

uOttawa

Alternative Ensembling Strategies

Table A.1: F1 scores of using alternative Ensembling Strategies in FLEXLOG.

Ensembling Config	ADFA-U							LOGEVOL-U			SynHDFS-U	SYNEVOL-U
	adduser	hydraFTP	hydraSSH	java	meter	web	average	Hadoop	Spark	average	average	average
Majority Voting (FLEXLOG)	**0.718**	**0.784**	**0.723**	**0.642**	**0.682**	**0.672**	**0.704**	**0.982**	**0.892**	**0.937**	**0.972**	**0.971**
Majority Voting (alternative)	0.641	0.711	0.691	0.615	0.656	0.635	0.650*	0.964	0.840	0.902*	0.936*	0.929*
SNAIL	0.671	0.714	0.683	0.624	0.628	0.651	0.661*	0.965	0.854	0.909*	0.958*	0.949*
MetaFormer	0.666	0.706	0.691	0.615	0.607	0.663	0.658*	0.976	0.866	0.921	0.954*	0.951*

* FLEXLOG yields a significant higher F1 score compared to using the alternative ensembling strategy.

uOttawa

Research Questions

RQ1. (**effectiveness**) How effective is FlexLog for ULAD compared to the baselines?

RQ1.1 Can FlexLog trained on limited labeled data achieve comparable effectiveness to baselines trained on full datasets?

RQ1.2 What impact does the level of log instability have on FlexLog and the baselines?

RQ2. (**data efficiency**) How does the amount of labeled training data impact FlexLog's effectiveness, and can it maintain robust effectiveness under varying degrees of data scarcity?

RQ3. (**time and memory efficiency**) What is the performance of FlexLog in terms of time efficiency during training and inference, and how much memory overhead does the cache incur during inference?

RQ4. (**configuration impact**) How does the performance of FlexLog vary under different configurations, including ablations of base models, RAG, and the cache, as well as alternative LLM choices?

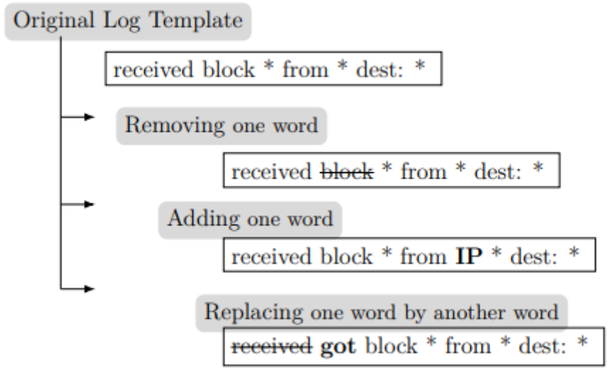
uOttawa

Another Contribution: Unstable Data Configuration

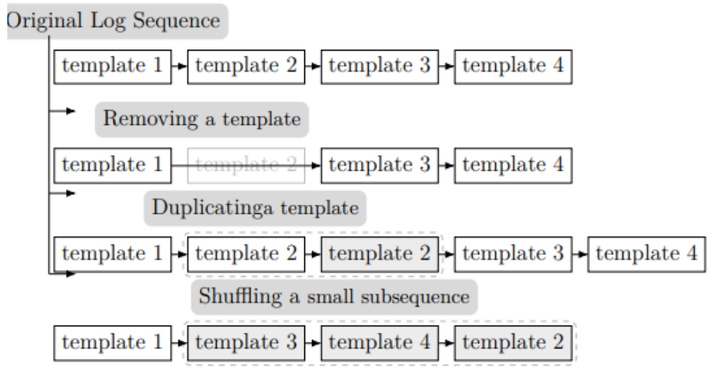
Real-world: ADFA-U, LOGEVOL-U,
Synthetic: SynHDFS-U, SYNEVOL-U

Name	Sys	#Log Messages	#Anomalous Messages	#Sessions	#Log Templates	Session Length avg	min	max
ADFA	Linux	2,747,550	317,388 (11.5%)	5,951	175	461.69	75	4,474
LOGEVOL	Hadoop 2	2,120,739	35,072 (1.6%)	333,699	319	6.35	1	1,963
	Hadoop 3	2,050,488	30,309 (1.4%)	343,013	313	5.97	1	1,818
	Spark 2	931,960	1,702 (0.1%)	13,892	130	67.08	1	1125
	Spark 3	1,600,273	2,430 (0.1%)	21,232	134	75.37	1	1977
HDFS	Hadoop	11,110,850	284,818 (2.9%)	575,061	48	19.32	2	30

Dataset	Configuration train/fine-tune	Configuration test	Duplication Ratio	#Log Sequences train	fine-tune	test
ADFA-U	ADFA $_{wo\ java}$	ADFA $_{w\ java}$	0.32	4786	1000	1165
	ADFA $_{wo\ hydraSSH}$	ADFA $_{w\ hydraSSH}$	0.31	4734	1000	1217
	ADFA $_{wo\ hydraFTP}$	ADFA $_{w\ hydraFTP}$	0.31	4748	1000	1203
	ADFA $_{wo\ meter}$	ADFA $_{w\ meter}$	0.34	4835	1000	1116
	ADFA $_{wo\ web}$	ADFA $_{w\ web}$	0.33	4792	1000	1159
	ADFA $_{wo\ adduser}$	ADFA $_{w\ adduser}$	0.33	4819	1000	1132
LOGEVOL-U	Hadoop 2	Hadoop 3	0.84	302312	8558	34495
	Spark 2	Spark 3	0.50	11114	1134	4246
SYNEVOL-U	Spark 2	Spark $2_{5\%\_sequence}$	0.6	11114	1134	2778
		Spark $2_{10\%\_sequence}$	0.55			
		Spark $2_{15\%\_sequence}$	0.50			
		Spark $2_{20\%\_sequence}$	0.44			
		Spark $2_{25\%\_sequence}$	0.37			
		Spark $2_{30\%\_sequence}$	0.32			
		Spark $2_{5\%\_template}$	0.54	11114	1134	2778
		Spark $2_{10\%\_template}$	0.45			
		Spark $2_{15\%\_template}$	0.36			
		Spark $2_{20\%\_template}$	0.28			
		Spark $2_{25\%\_template}$	0.22			
		Spark $2_{30\%\_template}$	0.18			
SynHDFS-U	HDFS	SynHDFS $_{5\%\_sequence}$	0.93	460048	5772	51000
		SynHDFS $_{10\%\_sequence}$	0.88			
		SynHDFS $_{20\%\_sequence}$	0.78			
		SynHDFS $_{30\%\_sequence}$	0.69			



Template-Level Changes



Sequence-Level Changes

uOttawa

Experiment Setting

Baselines:
Four supervised, one semi-supervised, and four unsupervised models

Empirical configuration of FlexLog:
KNN (K-Nearest Neighbor), **DT** (Decision Tree), and **SLFN** (Single-Layer Feed-forward Network) as ML models chosen due to their data and time efficiency seen in a recent work

Mistral (Mistral Small Instruct 22B) as LLM model chosen due to its open-source and comparable performance to the closed-source LLMs.

Evaluation Metrics and Statistical Testing:
Effectiveness: P (Precision), R (Recall), F1 (F1-score),
Efficiency: $\Delta u_{\%}$(labeling cost reduction), T (Training Time), I (Inference Time)
Statistical Testing: Mann-Whitney U test (also called Wilcoxon rank-sum test)

uOttawa

RQ1: Effectiveness on Real-world Data

- In ADFA-U, FlexLog **achieves state-of-the-art effectiveness**. FlexLog outperforms all baselines significantly, except LightAD and CNN where the difference is statistically insignificant.
- In LOGEVOL-U, on average, FlexLog **significantly outperforms all the baselines** in terms of F1 score, with a minimum margin of 1.8 pp (0.928 – 0.910).

Data	Unstable	M	limited data	full training set								
				Supervised				Semi-S	Unsupervised			
			FlexLog	LightAD	NeuralLog	LogRobust	CNN	PLELog	LogAnomaly	DeepLog	LogCluster	PCA
ADFA	No	P	0.708	0.820	0.538	0.718	0.666	0.736	0.357	0.334	0.255	0.196
		R	0.894	0.814	0.602	0.708	0.842	0.216	0.430	0.523	0.907	0.139
		F1	0.791	**0.817**	0.568	0.713	0.744	0.334	0.390	0.408	0.398	0.162
average	Yes	P	0.633	0.747	0.357	0.667	0.748	0.503	0.286	0.297	0.236	0.164
		R	0.799	0.660	0.652	0.402	0.657	0.404	0.531	0.494	0.901	0.157
		F1	**0.704**	0.677	0.433*	0.494*	0.685	0.356*	0.368*	0.363*	0.340*	0.174*
$Hadoop_{2\to2}$	No	P	0.999	0.999	0.997	0.984	0.997	0.648	0.263	0.384	0.952	0.267
		R	0.986	0.994	0.986	0.976	0.997	0.888	0.616	0.352	0.320	0.867
		F1	0.993	**0.997**	0.992	0.980	**0.997**	0.749	0.368	0.367	0.479	0.408
$Spark_{2\to2}$	No	P	0.999	0.999	0.999	0.941	0.999	0.172	0.501	0.512	0.771	0.072
		R	0.969	0.939	0.636	0.969	0.878	0.129	0.393	0.443	0.818	0.471
		F1	**0.984**	0.968	0.777	0.952	0.935	0.243	0.441	0.475	0.794	0.125
$Hadoop_{2\to3}$	Yes	P	0.998	0.998	0.914	0.905	0.992	0.626	0.221	0.384	0.510	0.225
		R	0.965	0.963	0.984	0.950	0.968	0.819	0.522	0.352	0.371	0.898
		F1	**0.982**	0.980	0.948	0.927	0.980	0.709	0.310	0.367	0.430	0.360
$Spark_{2\to3}$	Yes	P	0.999	0.981	0.916	0.696	0.992	0.105	0.141	0.08	0.347	0.061
		R	0.805	0.708	0.766	0.832	0.736	0.377	0.458	0.606	0.805	0.484
		F1	**0.892**	0.829	0.834	0.757	0.840	0.165	0.216	0.122	0.485	0.108
Average ($Spark_{2\to3}$, $Hadoop_{2\to3}$)	Yes	P	0.998	0.99	0.915	0.786	0.992	0.366	0.181	0.232	0.428	0.143
		R	0.871	0.83	0.875	0.863	0.852	0.887	0.49	0.479	0.588	0.691
		F1	**0.928**	0.898*	0.891*	0.833*	0.910*	0.437*	0.263*	0.244*	0.458*	0.234*

* FlexLog yields a significant higher F1-score than baseline.

uOttawa

RQ1: Effectiveness on Synthetic Data

- In **sequence-level**, FlexLog outperforms all baselines significantly, except LightAD where the difference is statistically insignificant.
- In **template-level**, FlexLog outperforms all baselines significantly.

FlexLog achieves **state-of-the-art effectiveness on both real-world and synthesized datasets** while exhibiting **high data efficiency** (with a reduction in labeled data usage of ranging from 62.87 pp to 78.43 pp.)

Data	Unstable	M	limited data	full training set								
				supervised				Semi-S	Unsupervised			
			FlexLog	LightAD	NeuralLog	LogRobust	CNN	PLELog	LogAnomaly	DeepLog	LogCluster	PCA
SynHDFS 0%	No	P	0.954	0.9763	0.949	0.957	0.933	0.630	0.243	0.725	0.999	0.924
		R	0.999	0.990	0.985	0.980	0.951	0.966	0.971	0.927	0.346	0.667
		F1	0.976	**0.983**	0.966	0.969	0.942	0.763	0.389	0.814	0.514	0.762
SynHDFS average	Yes	P	0.949	0.948	0.915	0.684	0.932	0.398	0.355	0.63	0.962	0.374
		R	0.992	0.972	0.979	0.965	0.95	0.778	0.762	0.903	0.355	0.688
		F1	**0.971**	0.959	0.946*	0.760*	0.942*	0.499*	0.446*	0.741*	0.519*	0.404*
SYNEVOL 0%	No	P	0.999	0.999	0.999	0.941	0.999	0.172	0.501	0.512	0.771	0.072
		R	0.969	0.939	0.636	0.969	0.878	0.129	0.393	0.443	0.818	0.471
		F1	**0.984**	0.968	0.777	0.952	0.935	0.243	0.441	0.475	0.794	0.125
SYNEVOL Sequence Average	Yes	P	0.999	0.999	0.999	0.972	0.994	0.175	0.247	0.229	0.764	0.062
		R	0.966	0.94	0.622	0.914	0.888	0.144	0.422	0.42	0.811	0.47
		F1	**0.982**	0.969	0.767*	0.942*	0.938*	0.158*	0.302*	0.288*	0.786*	0.11 *
SYNEVOL Template Average	Yes	P	0.995	0.999	0.974	0.93	0.937	0.121	0.237	0.156	0.547	0.053
		R	0.934	0.894	0.628	0.953	0.864	0.29	0.399	0.419	0.79	0.471
		F1	**0.963**	0.944*	0.763*	0.941*	0.899*	0.17*	0.305*	0.218*	0.643*	0.096*

* FlexLog yields a significant higher F1-score than compared baseline.

uOttawa

RQ3: Time Efficiency

Data	Config	M	supervised					Semi-S	Unsupervised			
			FlexLog	LightAD	NeuralLog	LogRobust	CNN	PLELog	LogAnomaly	DeepLog	LogCluster	PCA
ADFA-U	adduser	T	16 161	5	922	221	271	285	276	184	4	0.017
		I	0.842	0.012	0.664	0.234	0.164	0.211	0.025	0.012	<0.001	≪ 0.001
	hydraFTP	T	15 906	4	920	220	269	285	276	181	4	0.017
		I	0.818	0.014	0.635	0.229	0.165	0.211	0.024	0.011	<0.001	≪ 0.001
	hydraSSH	T	14 147	5	923	220	269	285	275	183	4	0.017
		I	0.832	0.014	0.655	0.232	0.164	0.211	0.025	0.012	<0.001	≪ 0.001
	java	T	13 933	6	921	220	270	285	275	180	4	0.017
		I	0.875	0.014	0.673	0.236	0.164	0.211	0.025	0.011	<0.001	≪ 0.001
	web	T	14 079	5	918	220	270	285	276	181	4	0.017
		I	0.864	0.013	0.679	0.231	0.165	0.211	0.025	0.012	<0.001	≪ 0.001
	meter	T	15 324	3	921	220	269	285	276	182	4	0.017
		I	0.888	0.014	0.682	0.238	0.165	0.211	0.025	0.012	<0.001	≪ 0.001
LOGEVOL-U	Hadoop	T	4 031	30	1 550	178	316	48	1 340	1 020	21	0.180
		I	0.435	0.067	0.275	0.078	0.077	0.072	0.004	0.001	<0.001	≪ 0.001
	Spark	T	23 706	0.2	712	232	136	220	944	514	9	0.015
		I	0.844	0.038	0.564	0.082	0.072	0.006	0.007	0.003	< 0.001	≪ 0.001
SynHDFS-U	average	T	13 587	22	1 260	293	355	42	976	1 110	20	0.067
		I	0.771	0.005	0.259	0.008	0.016	0.007	0.004	0.002	< 0.001	≪ 0.001
SYNEVOL-U	average	T	23 706	0.2	712	232	136	220	944	514	9	0.015
		I	0.988	0.038	0.703	0.116	0.076	0.008	0.013	0.004	< 0.001	≪ 0.001

- While FlexLog is not the fastest method in inference, **its practicality depends on system constraints**.
- When detection effectiveness is paramount, FlexLog remains a strong choice despite higher time cost. The latter **can easily be alleviated with more powerful, parallel computations.**

While FlexLog is not the most time-efficient in ULAD inference, it processes each log sequence within **1 s on average**. FlexLog's cache memory remains **below 4 MB** for most datasets (up to 19.6 MB for ADFA-U), confirming its memory efficiency.

RQ4: How does the performance of FlexLog vary under different configurations?

Config	ADFA-U	LOGEVOL-U	SYNEVOL-U	SynHDFS-U
FlexLog	**0.853**	**0.628**	**0.941**	**0.771**
FlexLog w/o cache	0.904	0.940	0.988	0.794
Difference	-0.051*	-0.312*	-0.047*	-0.023†

Config	adduser	hydraFTP	hydraSSH	java	meter	web	Average
FlexLog	0.718	0.784	0.723	0.642	0.682	0.672	**0.704**
FlexLog w/o RAG	0.688	0.705	0.648	0.622	0.659	0.641	0.660
Difference (pp)	3	7.9	7.5	2	2.3	3.1	4.4*

Config	ADFA-U							LOGEVOL-U			SynHDFS-U	SYNEVOL-U
	adduser	hydraFTP	hydraSSH	java	meter	web	average	Hadoop	Spark	average	average	average
FlexLog	**0.718**	**0.784**	**0.723**	**0.642**	**0.682**	**0.672**	**0.704**	0.982	0.892	0.937	**0.972**	0.971
w/o Mistral	0.645	0.769	0.692	0.628	0.639	0.616	0.664*	0.975	0.841	0.908*	0.939*	0.936*
w/o KNN	0.683	0.768	0.654	0.621	0.651	0.579	0.659*	0.980	0.892	0.936†	0.945*	0.971†
w/o DT	0.667	0.749	0.690	0.640	0.654	0.623	0.670*	0.973	0.875	0.924†	0.948*	0.979*
w/o SLFN	0.677	0.691	0.613	0.641	0.647	0.556	0.637*	0.978	0.871	0.924†	0.934*	0.948*
w/o Simples	0.579	0.591	0.630	0.628	0.674	0.571	0.612*	**0.998**	**0.962**	**0.980**‡	0.928*	**0.979**†

* FlexLog yields a significant higher F1-score than the ablation configuration.
† indicates no significant difference between FlexLog and the ablation configuration.
‡ FlexLog yields a significant lower F1-score than the ablation configuration.

Config	Source	ADFA-U	LOGEVOL-U	SynHDFS-U	SYNEVOL-U
FlexLog	open	0.704	**0.928**	0.972	**0.971**
FlexLog (*Mistral* → *Llama*)	open	0.664*	0.895*	0.949*	0.970†
FlexLog (*Mistral* → *GPT*)	closed	**0.710**†	0.919†	**0.976**†	**0.971**†

- **Without cache**, the inference time increased for all datasets.
- **Without RAG**, Flexlog performance significantly drops by 4.4 pp on average.
- **Removing any base model** generally decreases F1 scores, except for LOGEVOL-U Spark and SYNEVOL-U, where extreme dataset imbalance hinders ML base model performance.
- **Mistral** remains the best LLM choice as an open-source LLM with insignificant or small different to GPT-4o.