



# InsightAI: Root Cause Analysis in Large Hierarchical Log Files with Private Data Using Large Language Model

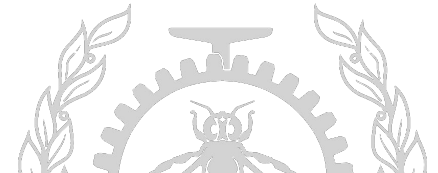
*Maryam Ekhlas*  
Dec. 05<sup>th</sup>, 2024

Polytechnique Montreal

**DORSAL** Laboratory

# Motivations

- Root cause analysis can be a time-intensive process.
- Modern software systems generate massive volumes of logs.
- Effective log analysis requires a deep understanding of the software architecture.



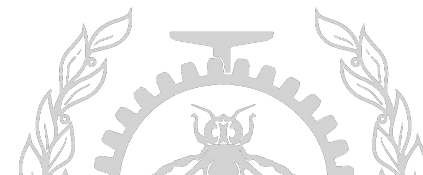
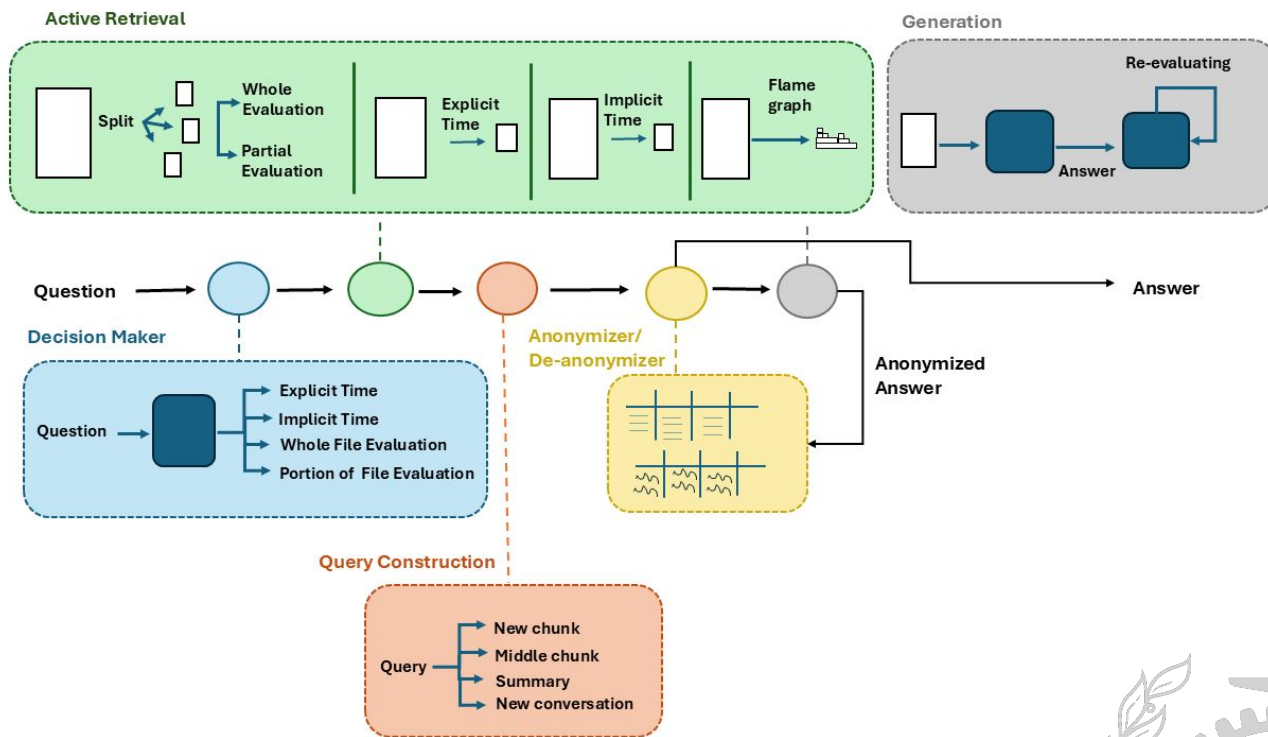
# Our Goal

---

- An adaptive approach to efficiently analyze relevant logs based on user queries, optimizing token usage and reducing costs.
- Anonymizing log data to protect sensitive information while keeping the accuracy of our method.
- Having a chatbot for having an interactive platform between the model and developers.

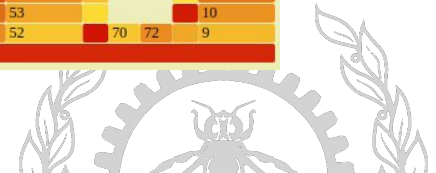
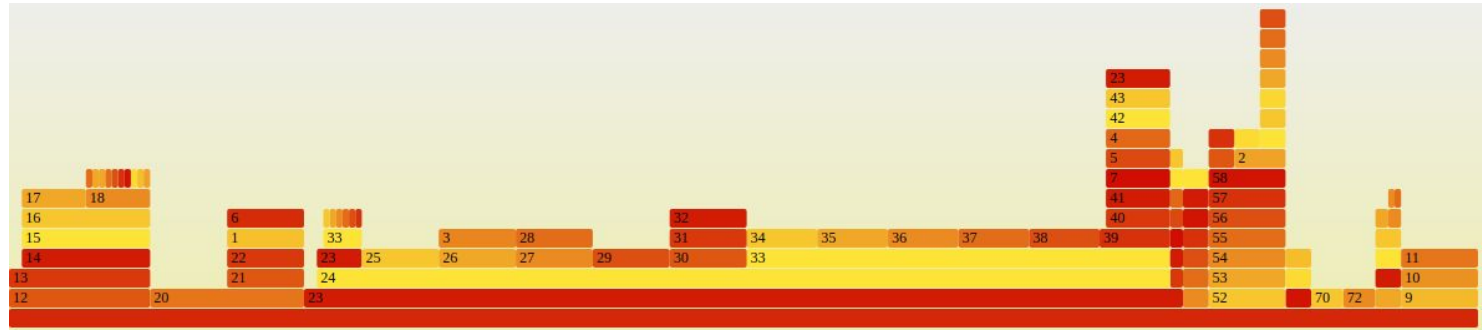


# Architecture



# Active Retrieval

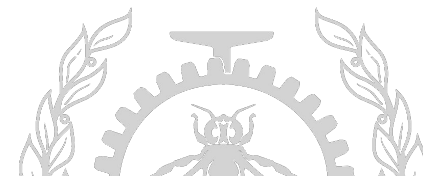
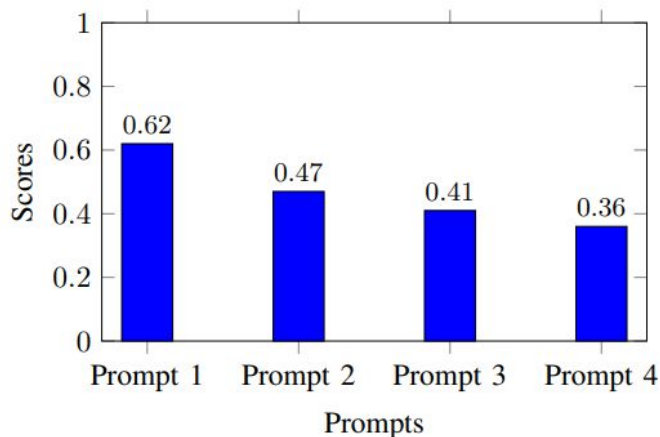
- Timestamp strategy
- Portion of content evaluation
- Full content evaluation
  - Token Count Tracking and Summarization Strategy.
  - Flame-graph-like Strategy.



# Query Construction

---

- Time-Specific Prompts.
- Initial Chunk Evaluation Prompt.
- Extended Evaluation Prompt.
- Self-Assessment Hallucination Mitigation.
- System Prompt (Instructor Prompt).
- Token Limit Management with Summarization.



# Anonymizer/De-anonymizer

- Ip address
- Function names
- Specific names
- Module names
- Directories

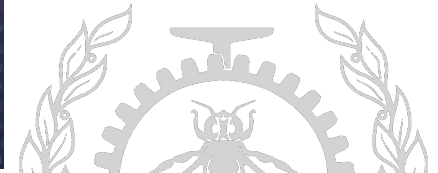
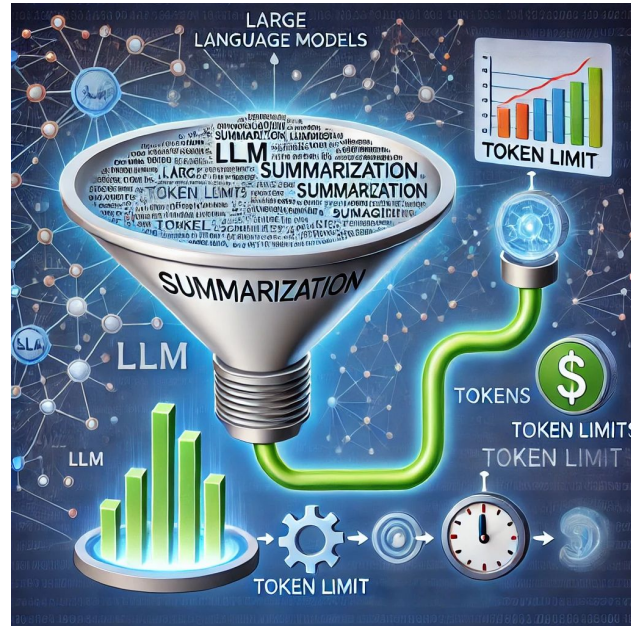
Message content and timestamp

Anonymization	Precision (%)	Recall (%)	F1 Score (%)
RadomValue (Baseline)	86.67	23.64	37.26
FunctionNameRandomValue	<b>100.00</b>	<b>80.00</b>	<b>88.89</b>
FunctionName_RandomValue	100.00	76.36	86.57
FunctionName_RandomValue()	0.00	0.00	0.00
RandomValue()	100.00	27.27	42.86
functionNameRandomValue	100.00	27.27	42.86
RandomValue_FunctionName	86.67	20.00	32.47

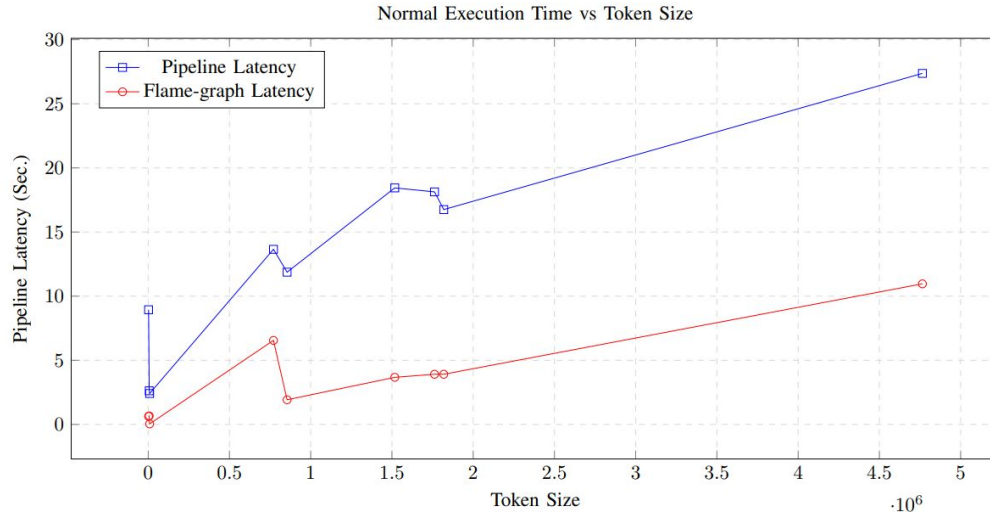


# Generation

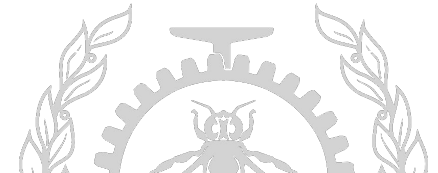
- Counting tokens
- Summarize the current conversation
- Initiate a new conversation
- Self assessment strategy



# Experimental Evaluation

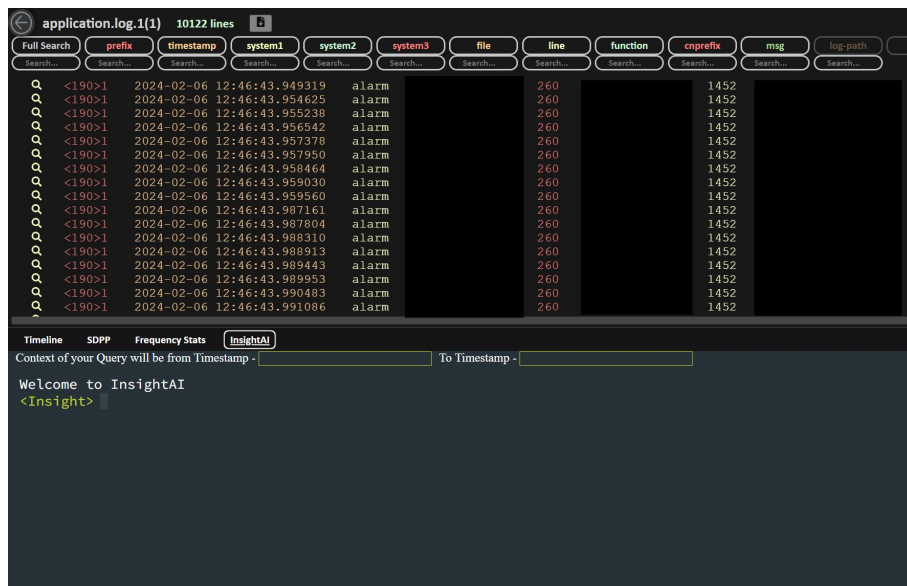


Name	Total Tokens	Flame-graph Tokens	Pipeline Latency (Sec.)	Flame-graph Latency (Sec.)
Log A	854 754	47 347	11.87	1.92
Log B	2179	278	8.94	0.63
Log C	1 517 318	32 772	18.44	3.67
Log D	770 823	141 728	13.65	6.55
Log E	8682	1994	2.41	0.04
Log F	1 762 090	61 854	18.13	3.91
Log G	4 764 844	177 248	27.37	10.96
Log H	1 819 004	16 074	16.75	3.91
Log I	6216	261	2.62	0.66



# Accomplished

- Implemented Chatbot for user interaction.
- Our flame-graph-like methodology reduces input tokens by 93.61% and processing latency by 77.45%.
- Our anonymization results show an improvement of 138.63% over the baseline.



The screenshot displays a log viewer application titled "application.log.1(1)" with 10122 lines. The interface includes a search bar and several filter tabs: "Full Search", "prefix", "timestamp", "system1", "system2", "system3", "file", "line", "function", "cnprefix", "msg", and "log-path". Below the search bar, a table of log entries is shown. Each entry consists of a source identifier, a timestamp, a message, and a line number. The chatbot interface at the bottom is titled "InsightAI" and shows a welcome message and a prompt for the user's query.

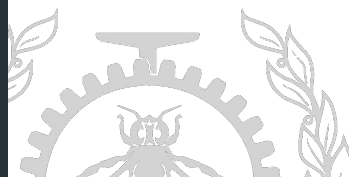
Source	Timestamp	Message	Line
<190>1	2024-02-06 12:46:43.949319	alarm	260
<190>1	2024-02-06 12:46:43.954625	alarm	260
<190>1	2024-02-06 12:46:43.955238	alarm	260
<190>1	2024-02-06 12:46:43.956542	alarm	260
<190>1	2024-02-06 12:46:43.957378	alarm	260
<190>1	2024-02-06 12:46:43.957950	alarm	260
<190>1	2024-02-06 12:46:43.958464	alarm	260
<190>1	2024-02-06 12:46:43.959030	alarm	260
<190>1	2024-02-06 12:46:43.959560	alarm	260
<190>1	2024-02-06 12:46:43.987161	alarm	260
<190>1	2024-02-06 12:46:43.987804	alarm	260
<190>1	2024-02-06 12:46:43.988310	alarm	260
<190>1	2024-02-06 12:46:43.988913	alarm	260
<190>1	2024-02-06 12:46:43.989443	alarm	260
<190>1	2024-02-06 12:46:43.989953	alarm	260
<190>1	2024-02-06 12:46:43.990483	alarm	260
<190>1	2024-02-06 12:46:43.991086	alarm	260

Timeline SDPP Frequency Stats InsightAI

Context of your Query will be from Timestamp - [ ] To Timestamp - [ ]

Welcome to InsightAI

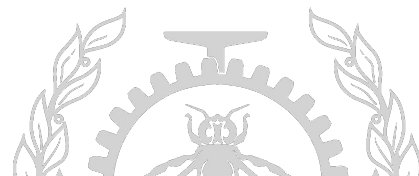
<Insight>



# Lesson Learned

---

- A larger token limit led to higher latency and cost, but it also made the selected chunks more relevant. This shows a trade-off between performance and cost.
- The flame-graph approach reduces token size and latency, optimizing processing speed and lowering costs.
- Using structured prefixes like 'FunctionName' for anonymized entities improves model accuracy by keeping key details in sensitive data.



---

# Thank you

**Email:** [maryam.ekhlasi@polymtl.ca](mailto:maryam.ekhlasi@polymtl.ca)

