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

Maryam Ekhlasi

Dec. 05th, 2024

Polytechnique Montreal

DORSAL Laboratory

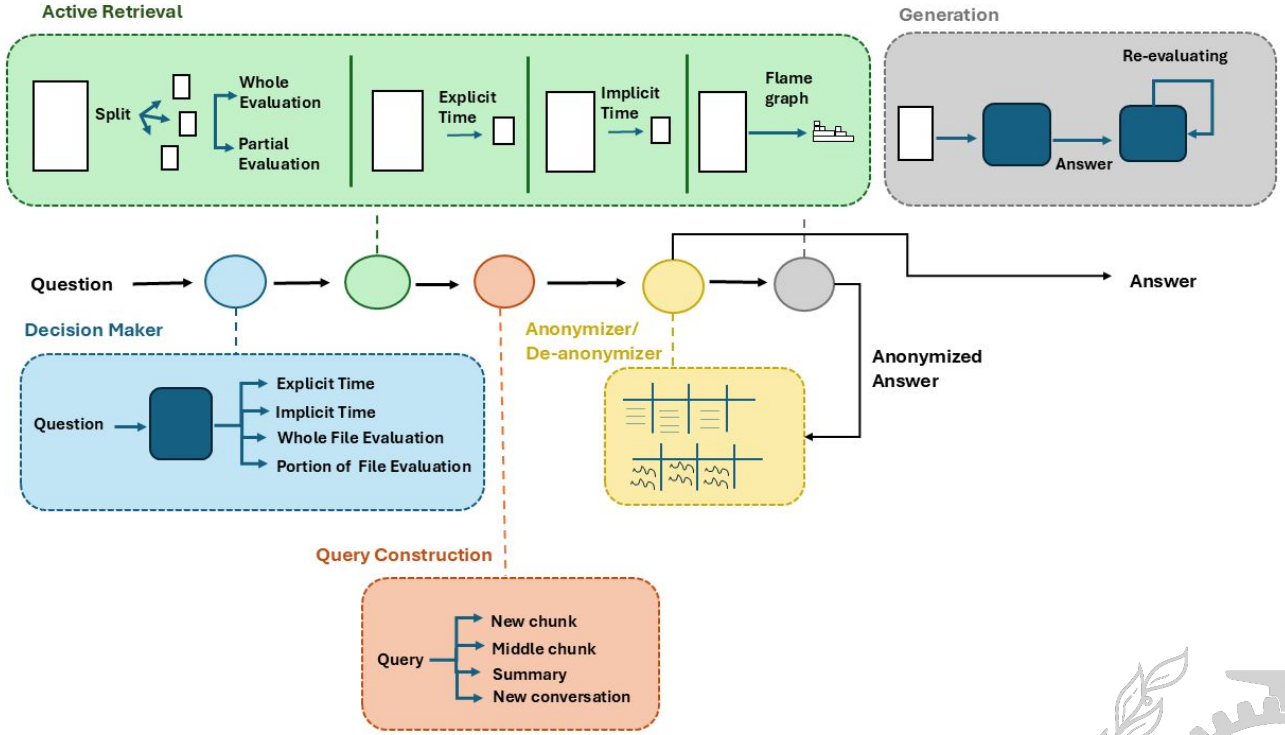
Motivations

- 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

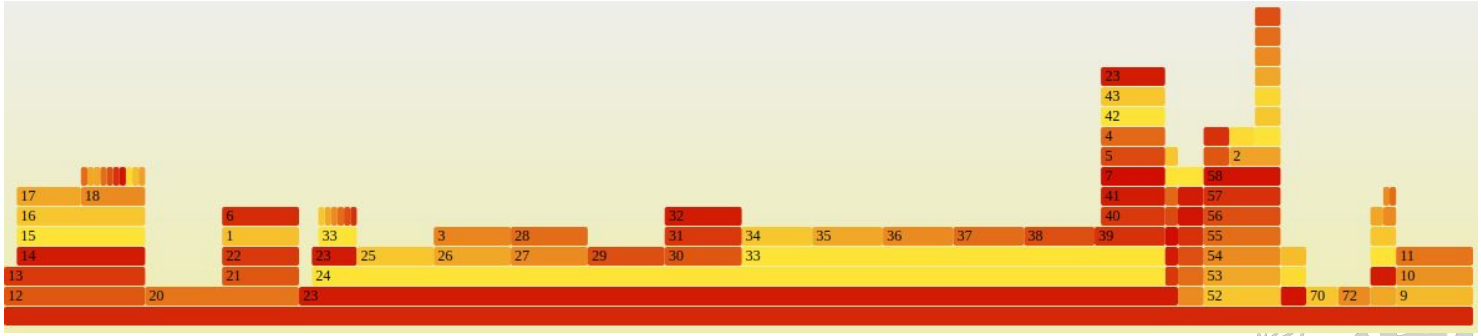
- An adaptive approach to efficiently analyze relevant logs based on user queries, optimizing token usage and reducing costs.
- Anonymizing log data to protect sensitive information while keeping the accuracy of our method.
- Having a chatbot for having an interactive platform between the model and developers.

Architecture



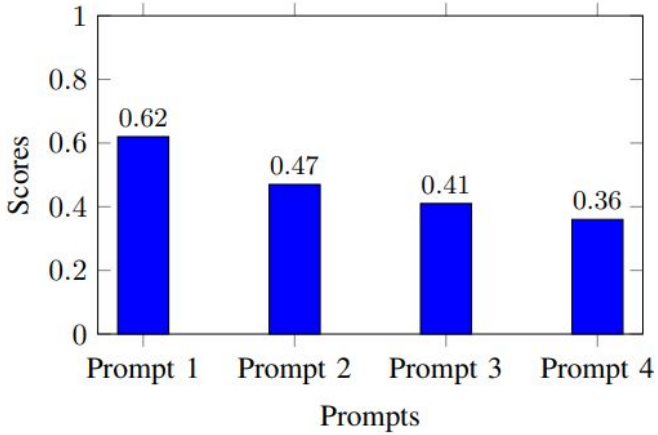
Active Retrieval

- 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

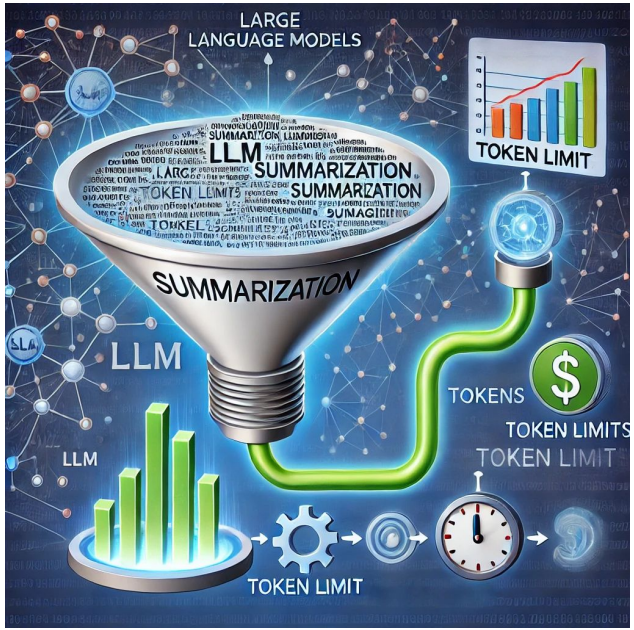
- Ip address
- Function names
- Specific names
- Module names
- Directories

Message content and timestamp

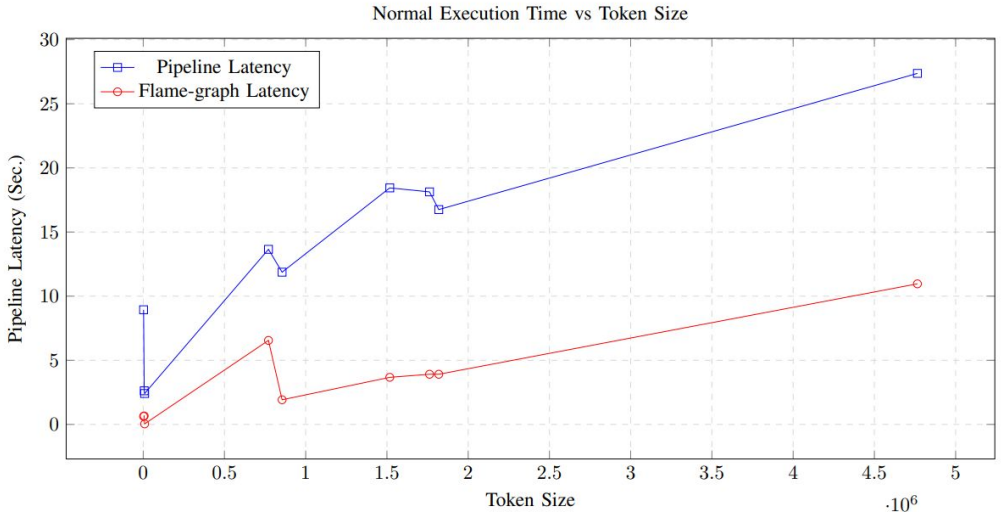
Anonymization	Precision (%)	Recall (%)	F1 Score (%)
RadomValue (Baseline)	86.67	23.64	37.26
FunctionNameRandomValue	**100.00**	**80.00**	**88.89**
FunctionName_RandomValue	100.00	76.36	86.57
FunctionName_RandomValue()	0.00	0.00	0.00
RandomValue()	100.00	27.27	42.86
functionNameRandomValue	100.00	27.27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

Generation

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



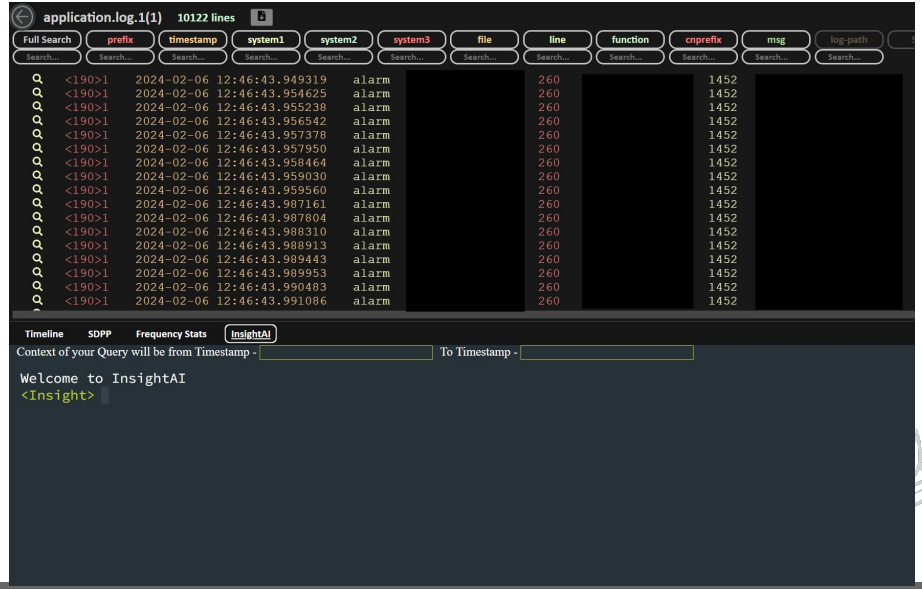
Experimental Evaluation



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.



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