AI in SRE

Unlocking Prometheus Insights with Natural Language



OUTLINE

| INTRODUCTION

The Problem Our Solution

RESULT

Lessons Limitations Future Plans Contributing



APPROACH

Architecture Key Components

CLOSING Contact Thank You



PROMCHAT

With Node Exporter With Custom Exporter

01

Introduction

The ProblemOur Solution

Understanding the Problem & Motivation

- > The Need for Faster Incident Response
- Democratizing Monitoring Data Access
- > The PromQL Learning Curve
- Inspired by "Chat with Your Data"

The Solution: Natural Language to PromQL

- Ask Questions in Natural Language
- PromQL query generated from Natural Language
- > Result fetched from prometheus using generated PromQL Query
- Result presented back in Natural Language



Approach

Architecture OverviewKey Components Review

Architecture Overview



Prometheus

Notation

Typically written in the format:

<metric name>{<label name>=<label value>, ...}

For example:

api_http_requests_total{method="POST", handler="/messages"}

/api/v1/metadata

```
"status": "success",
"data": {
  "cadvisor_version_info": [
      "type": "gauge",
     "unit": "",
     "help": "A metric with a constant '1' value labeled by k
  "container_blkio_device_usage_total": [
     "type": "counter",
     "unit": "",
     "help": "Blkio Device bytes usage"
  "container_cpu_load_average_10s": [
     "type": "gauge",
     "unit": "",
     "help": "Value of container cpu load average over the la
  "container_cpu_system_seconds_total": [
      "type": "counter",
     "unit": "",
      "help": "Cumulative system cpu time consumed in seconds.
```

LLM TOOLs



LLM Tool def _query_prometheus_tool(self, promql_query, source): """Wrapper for query_prometheus to return string output for tool.""" try: data_handler = self.data_handlers.get(source) if not data_handler: return f"Error: unknown data source: {source}" result = data_handler.query_prometheus(promql_query) return json.dumps(result) # Convert result to JSON string for tool output except Exception as e: return f"Error querying Prometheus: {e}"

LLM Tool def _get_metric_metadata_tool(self, source): """Wrapper for get_metric_metadata to format output for LLM tool use.""" try: data_handler = self.data_handlers.get(source) if not data_handler: return f"Error: unknown data source: {source}" metadata_description = data_handler.get_metric_metadata_description_for_llm() return metadata_description except Exception as e: return f"Error retrieving Prometheus metric metadata: {e}"



PromChat

Node ExporterCustom Exporter

NODE EXPORTER

Sample metrics and Queries

rate(node_cpu_seconds_total{ mode="system"}[1m])	The average amount of CPU time spent in system mode, per second, over the last minute (in seconds)
node_filesystem_avail_bytes	The filesystem space available to non-root users (in bytes)
rate(node_network_receive_bytes_total[1m])	The average network traffic received, per second, over the last minute (in bytes)



CUSTOM EXPORTER

Q	PromLens					Resources	¥
~	https://demo.promlabs.com					🗘 < Show hotkey	is: ?
	inter query or edit tree view below					× 🖾 ×	
F	ilter metrics by:						
	demo					×	
	<pre>demo_api_http_requests_in_progress</pre>	Q Explore labels	Ŧ	œ	gauge	The current number of API HTTP requests in progress.	
	<pre>demo_api_request_duration_seconds_bucket</pre>	Q Explore labels	•	œ	histogram	A histogram of the API HTTP request durations in seconds.	
	<pre>demo_api_request_duration_seconds_count</pre>	Q Explore labels	1	•	histogram	A histogram of the API HTTP request durations in seconds.	
	<pre>demo_api_request_duration_seconds_sum</pre>	Q Explore labels	٦	œ	histogram	A histogram of the API HTTP request durations in seconds.	
	<pre>demo_batch_last_run_duration_seconds</pre>	Q Explore labels	•	œ	gauge	The duration in seconds of the last batch job run.	
	<pre>demo_batch_last_run_processed_bytes</pre>	Q Explore labels	•	Ċ	gauge	The number of bytes processed by the demo batch job in the last run.	
	<pre>demo_batch_last_run_timestamp_seconds</pre>	Q Explore labels	•	٢	gauge	The Unix timestamp in seconds since the last demo batch job run.	
	<pre>demo_batch_last_success_timestamp_seconds</pre>	Q Explore labels	•	œ	gauge	The Unix timestamp in seconds since the last successful demo batch job completion.	
	<pre>demo_cpu_usage_seconds_total</pre>	Q Explore labels	•	œ	counter	The CPU usage in seconds split by mode.	
	<pre>demo_disk_total_bytes</pre>	Q Explore labels	•	œ	gauge	Total disk space in bytes.	
	<pre>demo_disk_usage_bytes</pre>	Q Explore labels	+	œ	gauge	Disk usage in bytes.	
	demo_intermittent_metric	Q Explore labels	٦	œ	gauge	A metric that is only present intermittently to test staleness handling.	
	demo_is_holiday	Q Explore labels	•	٢	gauge	Set to 1 if it is currently a holiday, 0 otherwise.	
	<pre>demo_items_shipped_total</pre>	Q Explore labels	Ŧ	•	counter	The total number of shipped items. Affected heavily by whether it's currently a holiday.	
	demo_memory_usage_bytes	Q Explore labels	•	Ē	gauge	The memory usage by type.	
	demo_num_cpus	Q Explore labels	1	œ	gauge	The number of CPUs.	



INFO:main: User query: Is today an holiday?

INFO:agent.agent:Generated PromQL Query: demo_is_holiday

INFO:agent.agent:Prometheus Result: [{"metric": {"__name__": "demo_is_holiday",

"instance": "demo-service-2:10002", "job": "demo"}, "value": [1744152905.074, "1"]}]

INFO:main: Agent response: Yes, `demo-service-2` reports that today is a holiday

INFO:main: User query: how many items have been shipped today?

INFO:agent.agent:Generated PromQL Query: increase(demo_items_shipped_total[1d])

INF0:agent.agent:**Prometheus Result:** [{"metric": {"instance": "demo-service-0:10000", "job": "demo"}, "value": [1744152970.506, "455360.05556520226"]}, {"metric": {"instance": "demo-service-1:10001", "job": "demo"}, "value": [1744152970.506, "453045.6537593332"]}, {"metric": {"instance": "demo-service-2:10002", "job": "demo"}, "value": [1744152970.506, "453865.79614516406"]}]

INFO:main: Agent response: Approximately 1,362,271 items have been shipped today

INFO:main: User query: Are requests to the demo api taking longer than usual?

INFO:agent.agent:Generated PromQL Query:

increase(demo_api_request_duration_seconds_sum[5m])>

increase(demo_api_request_duration_seconds_count[5m])*

avg_over_time(demo_api_request_duration_seconds_bucket[5m])

INFO:agent.agent:Prometheus Result: []

INFO:main: Agent response: No data was found

	· · · · · · · · · · · · · · · · · · ·			
	Chat 🕸 S	ettings		
Settings				
LLM Configuration				
Provider				
googleai			~	
Model				
gemini-1.5-flash			~	
Metrics Sources				
Current Sources				
default (prometheus) http://localhost:9090			Remove	
promlab_demo (prometheus) https://demo.promlabs.com			Remove	
Add New Source				
Allas (a name for this source)				
e.g., productorr-ciusidi				
Prometheus				
Prometheus Server URL				
e.g., http://localhost:9090				
Add Metric Source				
		_		
			Save Settings	

Settings		Chat B Settings		
Settings				
LLM Configuration Provider googleal Model gemini-2.0-flash-thinking-exp-01-21 Metrics Sources Current Sources default (prometheus) http://coathast.s000 promabe_demic (prometheus) Remove Add New Source a.g., production-cluster Type Prometheus Type Prometheus Prometheus Current Source Add New Source Add New Source a.g., production-cluster Type Prometheus Vemetheus Source UIL a.g., http://tocathast.s000 Add Metric Source	Settings			
Provider googlaal Model gemini-2.0-flash-thinking-exp-01-21 Metrics Sources Current Sources default (prometheus) http:/flocalhost:9090 Promlab_demo (prometheus) Remove Add New Source Add New Source G., production-cluster Type Prometheus PrometheusServer URL e.g., http://localhost:9090 Add Netric Sources	LLM Configuration			
googleal • Model	Provider			
Model gemini-2.0-flash-thinking-exp-01-21 Metrics Sources Current Sources default (prometheus) http://localhost:9090 promlab_demo (prometheus) http://localhost:9090 Remove Add New Source Aise (a name for this source) e.g., production-cluster Type Prometheus Prometheus e.g., http://localhost:9090	googleai		~	
gemini-2.0-filash-thinking-exp-01-21 Metrics Sources Current Sources default (prometheus) http://localhost.9090 promlab_demo (prometheus) http://demo.promitabs.com Add New Source Aitas (a name for this source) e.g., production-cluster Type Prometheus Prometheus Add Metric Source Add Metric Source Add Metric Source	Model			
Metrics Sources Current Sources default (prometheus) http://localhost:9090 promlab_demo (prometheus) https://demo.promlabs.com Add New Source Alias (a name for this source) e.g., production-cluster Type Prometheus Prometheus e.g., http://localhost:9090 Add Metric Source	gemini-2.0-flash-thinking-e	xp-01-21	~	
Current Sources default (prometheus) http://localhost:9090 promlab_demo (prometheus) http://demo.promlabs.com Add New Source Alias (a name for this source) e.g., production-cluster Type Prometheus Prometheus Server URL e.g., http://localhost:9090 Add Metric Source	Metrics Sources			
default (prometheus) Remove promlab_demo (prometheus) Remove https://demo.promlabs.com Remove Add New Source Remove Alias (a name for this source) e.g., production-cluster Type V Prometheus v Prometheus v Add Metric Source Remove Add Metric Source V Add Metric Source V	Current Sources			
promlab_demo (prometheus) Remove https://demo.promlabs.com Remove Add New Source Alias (a name for this source) e.g., production-cluster Type Prometheus V Prometheus V Add Metric Source Add Metric Source	default (prometheus) http://localhost.9090		Remove	
Add New Source Alias (a name for this source) e.g., production-cluster Type Promethous e.g., http://localhost:9090 Add Metric Source	promlab_demo (prometheus https://demo.promlabs.com	a)	Remove	
Alias (a name for this source) e.g., production-cluster Type Prometheus Prometheus e.g., http://localhost:9090 Add Metric Source	Add New Source			
e.g., production-cluster Type Prometheus Prometheus Server URL e.g., http://localhost:9090 Add Metric Source	Alias (a name for this source)			
Type Prometheus Prometheus Server URL e.g., http://localhost:9090 Add Metric Source	e.g., production-cluster			
Promethous C Promethous Server URL e.g., http://localhost:9090 Add Metric Source	Туре			
Promethous Server URL e.g., http://localhost:9090 Add Metric Source	Prometneus			
e.g., http://deanost:9090 Add Metric Source	Prometheus Server URL			
Add Metric Source	e.g., http://iocainost:9090			
	Add Metric Source			

V) 🛛 🔞 🗉 🗋 🖙
Chat 🕸 Settings	
Chat with Metrics	Clear History
Hello! I'm your Prometheus metrics assistant. How can I help you too Is too	lay? lay an holiday?
Yes, 'demo-service-2' reports that today is a holiday.	hinned today
Approximately 1,362,271 items have been shipped today.	
Are requests to the demonstrating long No data was found.	
The 99th percentile of request durations for the '/api/bar' path, especially POST requests resulting in 500 errors, is significantly high (around 0.2 seconds) compared to other paths like '/api/foo' or '/api/nonexistent', which are often below 0.05 seconds. This sugges that requests to '/api/bar' are taking longer than requests to other A paths.	ier man Usual?
Ask about cpu usage	۲

INFO:main: User query: Are requests to the demo api taking longer than usual?

INFO:agent.agent:Generated PromQL Query: histogram_quantile(0.99,

rate(demo_api_request_duration_seconds_bucket[5m]))

INF0:agent.agent:**Prometheus Result:** [{"metric": {"instance": "demo-service-2:10002", "job": "demo", "method": "GET", "path": "/api/nonexistent", "status": "404"}, "value": [1744153995.69, "0.0000990000000000001"]}, {"metric": {"instance": "demo-service-1:10001", "job": "demo", "method": "POST", "path": "/api/bar", "status": "500"}, "value": [1744153995.69, "0.21946694418525692"]}, ...]

INFO:main: Agent response: The 99th percentile of request durations for the `/api/bar` path, especially POST requests resulting in 500 errors, is significantly higher (around 0.2 seconds) compared to other paths like `/api/foo` or `/api/nonexistent`, which are often below 0.05 seconds. This suggests that requests to `/api/bar` are taking longer than requests to other API paths.



Result

- Lessons
- Limitations
- Future Plans
- Contributing

RESULT SUMMARY

Lessons from implementing the PoC

LESSONS

- LLMs are capable of writing correct PromQL without the need for additional training or fine tuning
- One short prompting to avoid response validation
- Thinking models write better PromQL

FUTURE IMPROVEMENTS

- Support more complex queries
- Support more metrics sources
- Learn from user interactions

LIMITATIONS

- Labels missing in metadata documentation
- Inconsistency on the generated queries

CONTRIBUTING

- Source code is available on github at <u>https://github.com/neslahg/promchat</u>
- Issues, PRs are welcomed.
- Live demo available at: <u>https://promchat.nesla.co</u>

THANK YOU



Mail: hello@davidasamu.com

Blog: https://doa.sh

X: @atechbruv

LinkedIn: https://www.linkedin.com/in/david-as amu/

David Asamu