



KubeCon



CloudNativeCon



China 2024



Build Interactive Monitoring

-- *Prometheus AI Agent*



Zhihao Liu @ Quwan
Devops Engineer

Current Situation



KubeCon



CloudNativeCon



THE LINUX FOUNDATION
OPEN SOURCE SUMMIT



AI_dev
Open Source Dev & ML Summit

China 2024

Question:

- What are the metrics for this instance resource?
这个实例资源有哪些指标?
- How to write PromQL for this?
如何写的PromQL?
- For this Alert / Event, which metrics should I check?
这个告警 /事件，我应该去看哪些有效指标?
- How can I view the chart of related resources?
如何查看关联资源的面板?
- Can it be viewed quickly on mobile devices?
可以快捷移动端查看吗?



Can LLM solve these problems? bring a new interactive experience?
LLM能解决这些问题吗？能带来新的交互体验吗？

Explore



KubeCon



CloudNativeCon



THE LINUX FOUNDATION
OPEN SOURCE SUMMIT

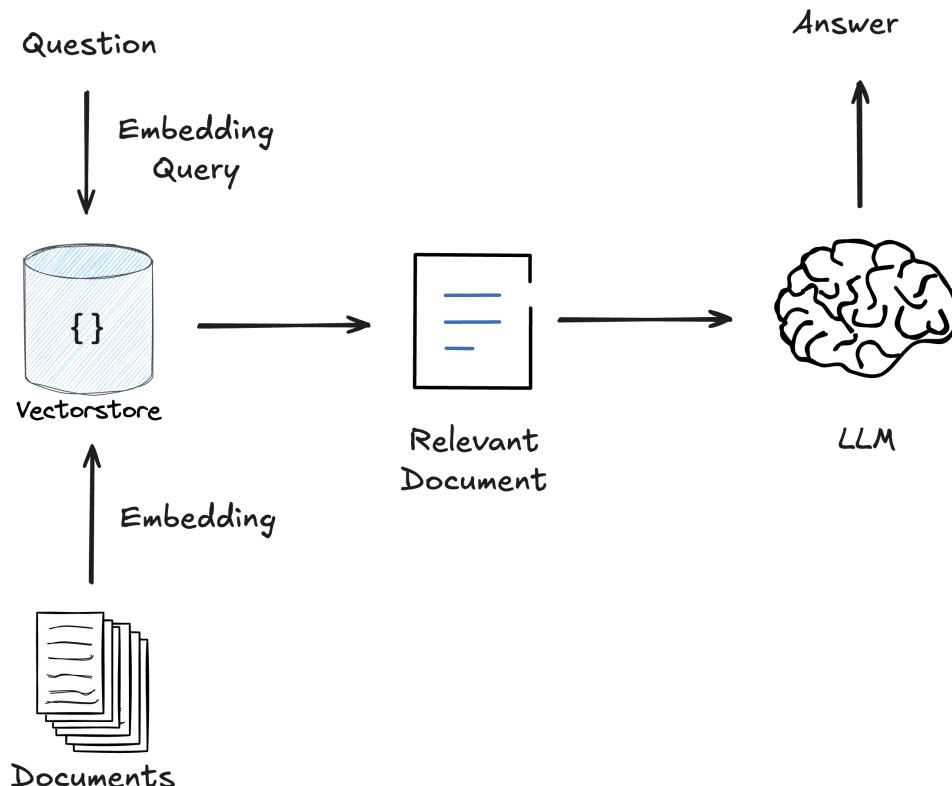


AI_dev
Open Source Dev & ML Summit

China 2024

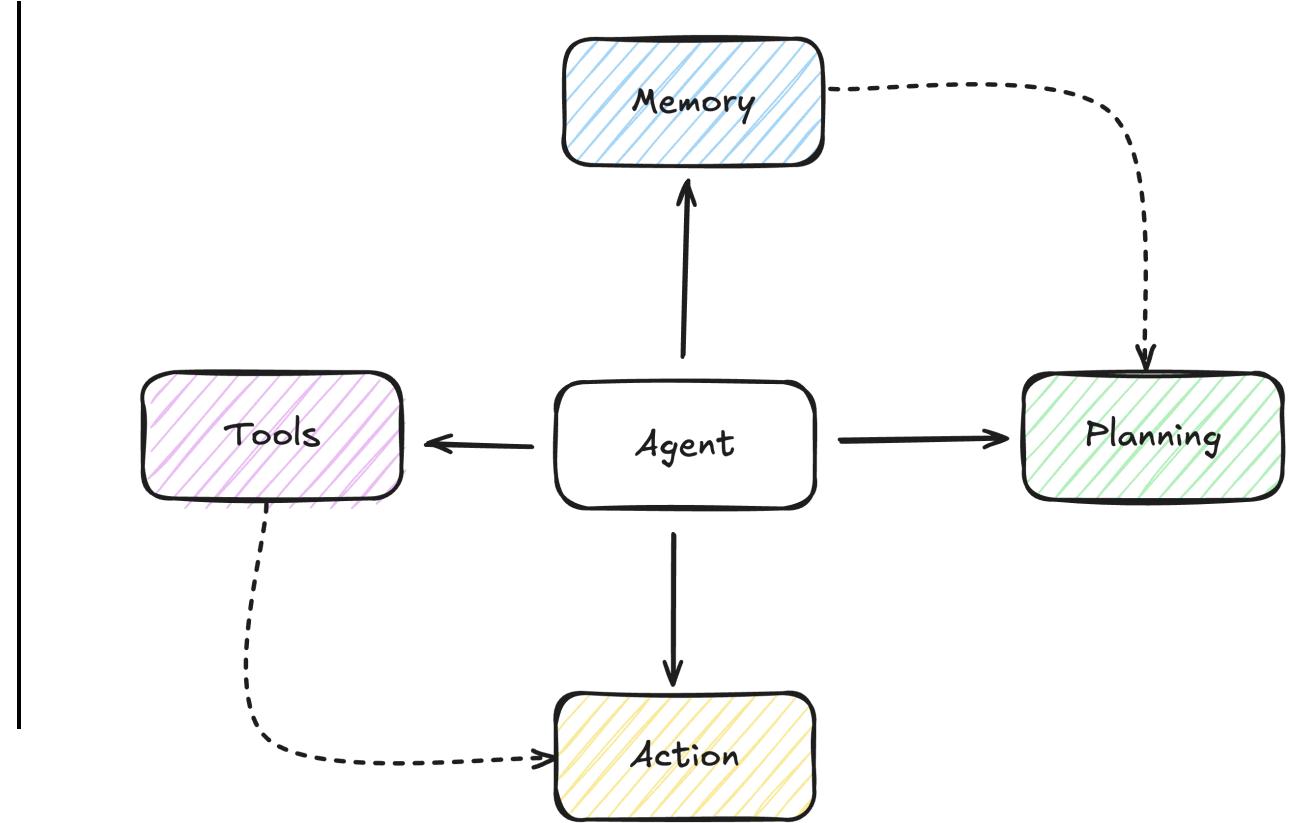
Text-to-PromQL

Retrieval Augmented Generation (RAG)



Like Human

AI Agent



But LLM has its own ideas? (Hallucination)

但是大模型常常有它自己的想法? (幻觉)

- User questions are casual and ambiguous.
用户的提问是随性的，模棱两可的
- Hallucinations occur, resulting in low accuracy and poor relevance.
出现“幻觉”，结果准确性低，相关性差
- The Chat method is cumbersome.
对答的方式很费劲
- Writing PromQL is just a means, not the end goal.
写PromQL只是目标不是目的

```
100 -  
avg_over_time(irate(node_cpu_seconds_total{resource_name="hw-bj-zt-quickbi-0004", mode="idle"}  
[1m])[2h:1m]) by  
(instance,cmdb_id,resource_name)*100
```

WANT:

cloud_slow_queries <-----



Documents

Output:

mysql_global_status_slow_queries

Thought



KubeCon



CloudNativeCon



China 2024



| Me | AI Agent |
|--|-----------------------------------|
| Determine the query objective 确定查询目标 | Query Rewrite |
| Search Google and look through communities 搜索谷歌、社区寻找 | MultiRetriever and Re-Rank |
| Reference to write PromQL and execute query 参考编写PromQL，执行查询 | LogsRetriever |
| When errors occur, analyze and adjust PromQL 遇到错误，分析，调整PromQL | ReAct Check And Generated PromQLs |
| View the results 查看结果 | Create Chart |

Query Rewrite



KubeCon



CloudNativeCon



THE LINUX FOUNDATION

OPEN
SOURCE
SUMMIT



AI_dev
Open Source Dev & ML Summit

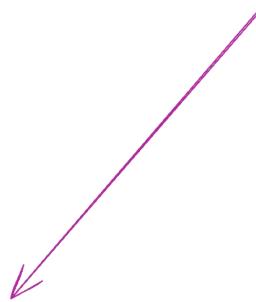
China 2024

query aiagent mysql cpu usage for the past day.



Query the CPU utilization of MySQL resources by aiagent over the past 24 hours.

query documents



Time experts



MultiRetriever and Re-Rank



KubeCon



CloudNativeCon

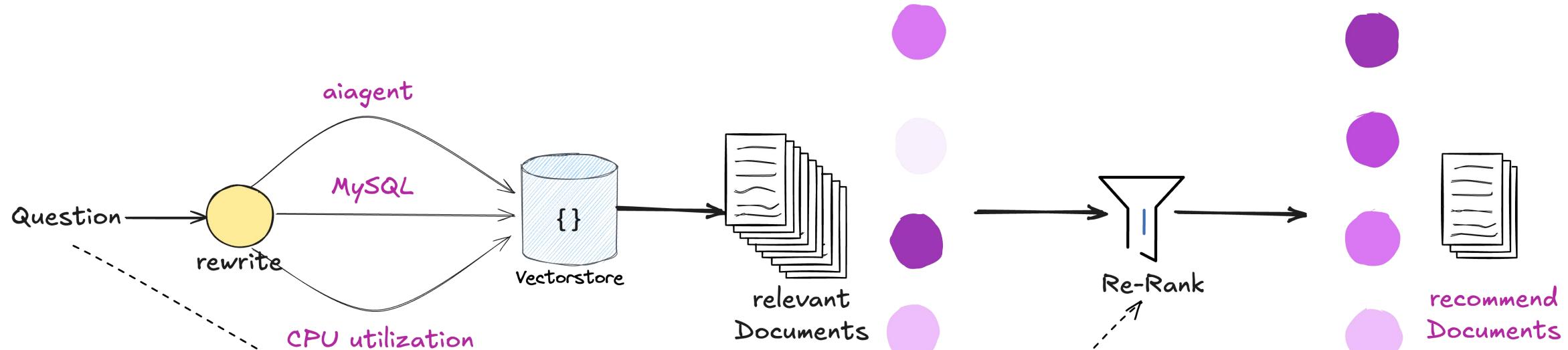


THE LINUX FOUNDATION
OPEN SOURCE SUMMIT



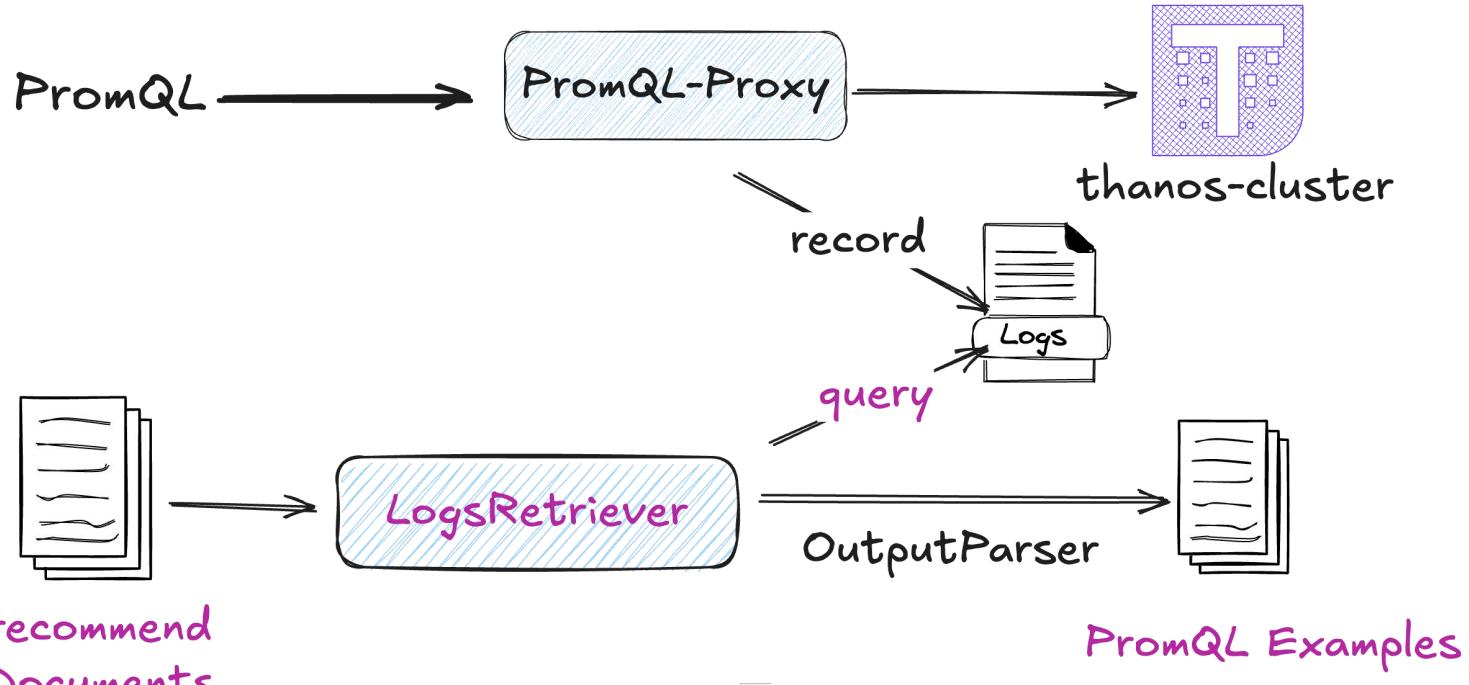
AI_dev
Open Source Dev & ML Summit

China 2024



LogsRetriever

LLM needs examples.
Good Case



The screenshot shows the Grafana interface with the following components:

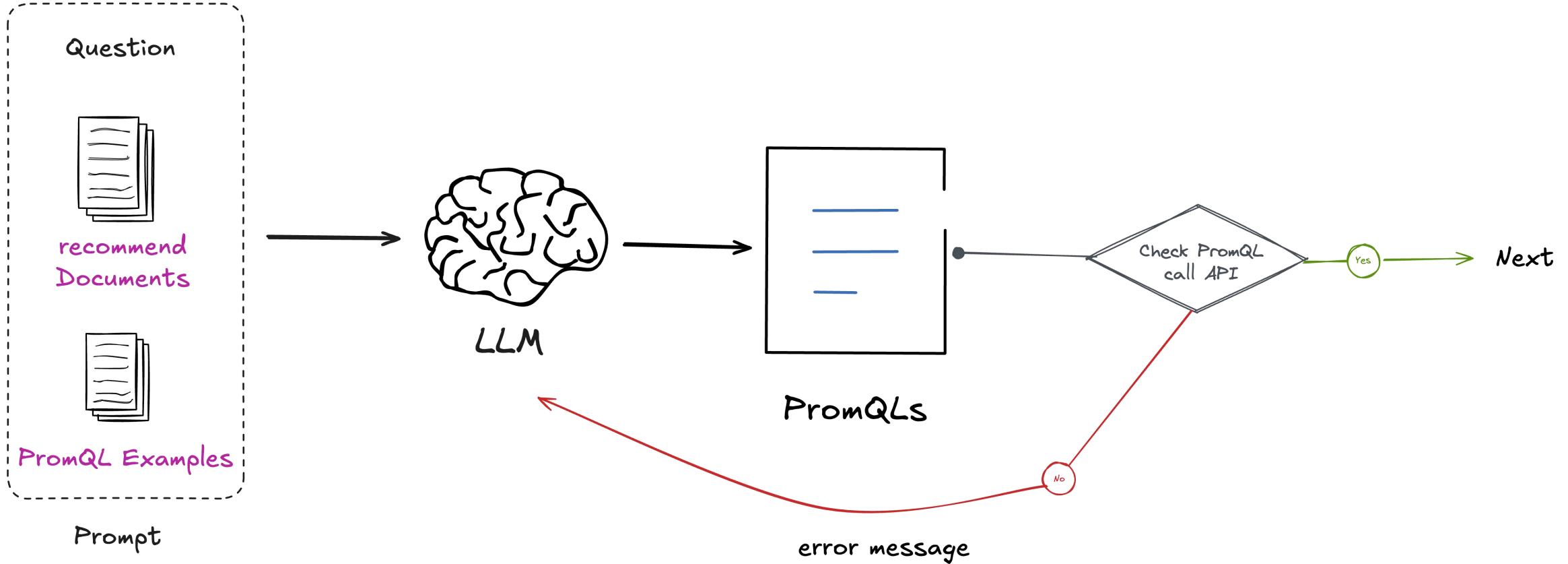
- Top Bar:** Includes "Kick start your query", "Label browser", "Explain query", and "Build" buttons.
- Query Editor:** Displays the query: `{app="promql-proxy",namespace="sre-platform"} | json | line_format "{{ .request_uri }}" |~ "cloud_mysql_cpu" |~ "cloud_cpu_use_rate"`. It also shows "Type: Range" and "Line limit: 50". A note says "This query will process approximately 0 results".
- Bottom Navigation:** Buttons for "+ Add query", "Query history", and "Inspector".
- Logs Volume:** A section labeled "Logs volume" with a "Logs" button.
- Log Filter Options:** Buttons for "Time", "Unique labels", "Wrap lines", "Pretty JSON", "Deduplication", and dropdowns for "None", "Exact", "Numbers", and "Signature".
- Result Order:** Buttons for "Display results", "Newest first", and "Oldest".

```
mmon labels: promql-proxy middleware/log.go:func1:57 k8s-hw-bj-1-yunwei 200 promql-proxy promql-proxy sre-platform/promql-proxy logging.ctx_logger GET request log sre-platform 1 stdout Line limit: 50 (7 returned) Total bytes processed: 18.0 GB
2024-08-06 15:25:50 /api/v1/query_range?query=sum%28cloud_mysql_cpu%7Bcluster%3D%22k8s-tc-bj-1-yunwei%22%2Cid%3D%22tencent_bj%22%2Cid_name%3D%22E8%85%BE%8E%AE%F4%BA%91%E5%8C%97%4BA%AC%22%2Cresource_name%3D%22%2Aaiagent.%2A%22%2Csolo_belong%3D%22%5E%AE%9E%4BE%88%22%7D+0+r+cloud_cpu_use_rate%7Bcluster%3D%22k8s-tc-bj-1-yunwei%22%2Cid_id%3D%22tencent_bj%22%2Cid_name%3D%22E8%85%BE%8E%AE%F4%BA%91%E5%8C%97%4BA%AC%22%2Cresource_name%3D%22%2Aaiagent.%2A%22%2Csolo_be1ong%3D%22%5E%AE%9E%4BE%88%22%7D+29+b+4*8resource_name%29&start=172292548&end=172292948&timeout=30s&step=60
2024-08-06 15:24:12 /api/v1/query_range?query=sum%28cloud_mysql_cpu%7Bcluster%3D%22k8s-tc-bj-1-yunwei%22%2Cid%3D%22tencent_bj%22%2Cid_name%3D%22E8%85%BE%8E%AE%F4%BA%91%E5%8C%97%4BA%AC%22%2Cresource_name%3D%22%2Aaiagent.%2A%22%2Csolo_belong%3D%22%5E%AE%9E%4BE%88%22%7D+0+r+cloud_cpu_use_rate%7Bcluster%3D%22k8s-tc-bj-1-yunwei%22%2Cid_id%3D%22tencent_bj%22%2Cid_name%3D%22E8%85%BE%8E%AE%F4%BA%91%E5%8C%97%4BA%AC%22%2Cresource_name%3D%22%2Aaiagent.%2A%22%2Csolo_be1ong%3D%22%5E%AE%9E%4BE%88%22%7D+29+b+4*8resource_name%29&start=172292548&end=172292948&timeout=30s&step=60
2024-08-06 15:21:14 /api/v1/query_range?query=sum%28cloud_mysql_cpu%7Bcluster%3D%22k8s-tc-hi-1-yunwei%22%2Cid%3D%22tencent_bj%22%2Cid_name%3D%22E8%85%BE%8E%AE%F4%BA%91%E5%8C%97%4BA%AC%22%2Cresource_name%3D%22%2Aaiagent.%2A%22%2Csolo_bel
```

ReAct Check And Generated PromQLs



China 2024



Create Chart



KubeCon



CloudNativeCon



THE LINUX FOUNDATION

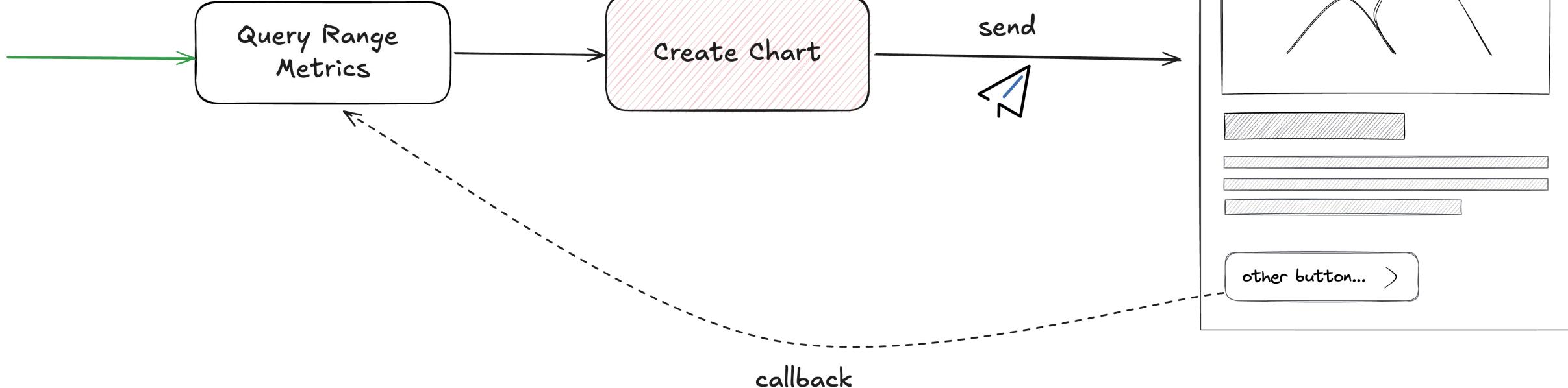
OPEN SOURCE
SUMMIT



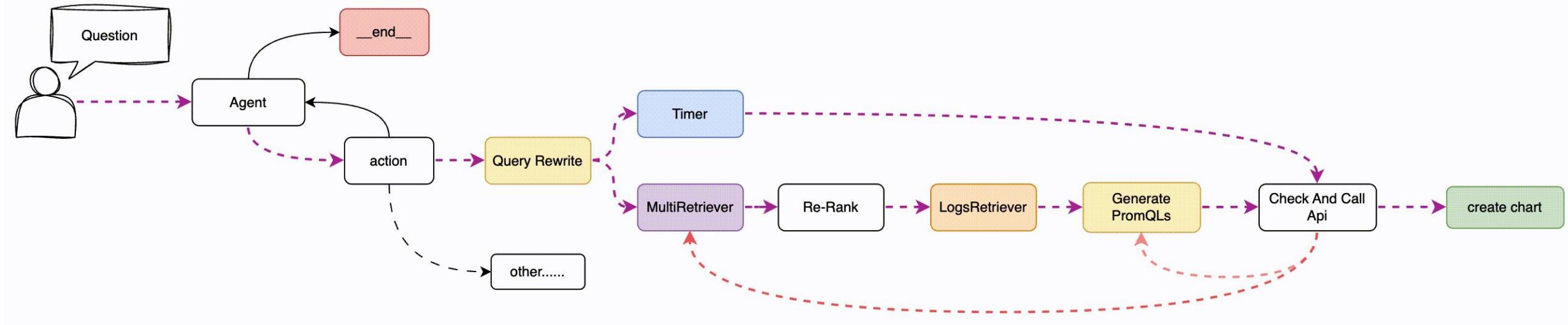
AI_dev
Open Source DevOps & ML Summit

China 2024

Card



General



Show Case



KubeCon



CloudNativeCon

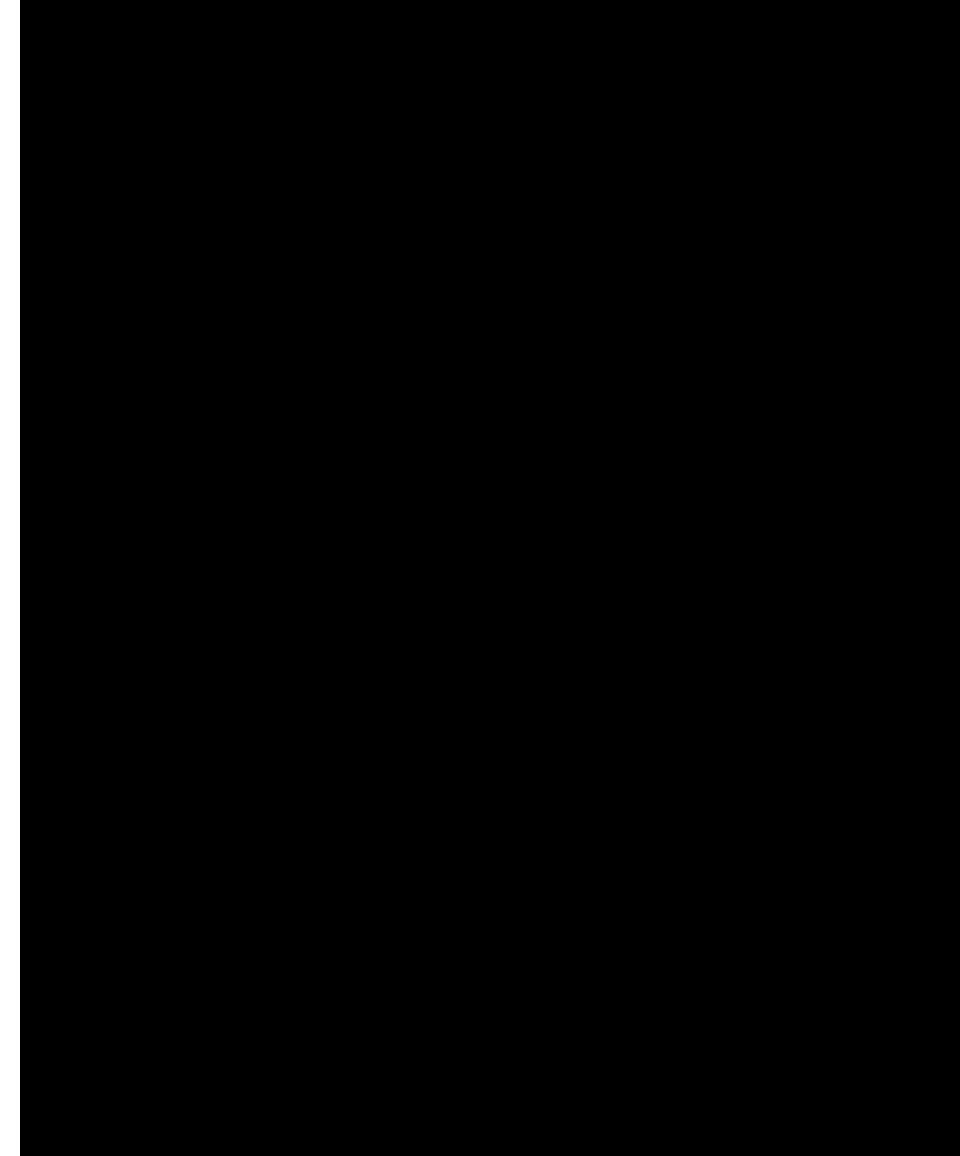


THE LINUX FOUNDATION
OPEN SOURCE SUMMIT



AI_dev
Open Source DevOps & ML Summit

China 2024



The End !