User → Inference App → OTEL →

Metrics (DCGM Exporter) → [Prometheus] → [Grafana]

Traces → [Tempo] → [Grafana] → [Phoenix] (prompt/response)

Spanmetrics → Prometheus (exemplars + metrics) → Grafana

App + LoadGen → OTEL Collector → Prometheus + Tempo → Grafana → Phoenix

LLM Trace IDs link back to Phoenix

Exemplar	×
2025-09-24 10:16:03	
Value	2166
__name__	traces_span_metrics_calls_total
exported_job	demo-app
instance	otel-collector:8889
job	otel-collector
llm_model_name	llama3.2:3b
llm_request_type	completion
otel_scope_name	spanmetricsconnector
service_name	demo-app
span_id	97985d3013528bbb
span_kind	SPAN_KIND_CLIENT
span_name	llm.generate
status_code	STATUS_CODE_UNSET
trace_id	5b3caa920eb1536a42456f58fb30988f

Latency p50 / p95 / p99 (s) by Service

Name
- demo-app p50
- demo-app p95
- demo-app p99