Observability Day

NORTH AMERICA

Observability Day
NORTH AMERICA

“What’s the Dollar Cost of Slow?”

Answering Business Questions with OTel

PRESENTED BY

Kayla Reopelle & **Hannah Ramadan**

Speakers

Observability Day
NORTH AMERICA



Hannah Ramadan

Software Engineer

New Relic



Kayla Reopelle

Lead Software Engineer

New Relic

Agenda

Observability Day
NORTH AMERICA

01 Why slow costs money, and what the research says

02 Business Observability & turning telemetry into dollars

03 An example: an AI support assistant having a bad day

04 What each signal (metrics, traces, logs) adds to the answer

05 Where OpenTelemetry semantic conventions can help

Observability Day
NORTH AMERICA

What fast feels like

Our expectations haven’t changed

Observability Day
NORTH AMERICA

What “fast” feels like may be biological in nature

1968 → **2010**

The research

Observability Day
NORTH AMERICA

2 seconds

A wait longer than two seconds breaks concentration
and hurts productivity

Robert Miller 1968

The research

The research

Observability Day
NORTH AMERICA

0.1s

Feels “instantaneous”

Jakob Nielsen
1993, 2010

The research

Observability Day
NORTH AMERICA

0.1s

Feels “instantaneous”

Jakob Nielsen
1993, 2010

1s

Keeps our flow of thought seamless

Jakob Nielsen
1993, 2010

The research

Observability Day
NORTH AMERICA

0.1s	1s	10s
Feels “instantaneous”	Keeps our flow of thought seamless	Barely keeps our attention
Jakob Nielsen 1993, 2010	**Jakob Nielsen** 1993, 2010	**Jakob Nielsen** 1993, 2010

The research

Observability Day
NORTH AMERICA

0.1s	1s	10s	>10s
Feels “instantaneous”	Keeps our flow of thought seamless	Barely keeps our attention	Our minds wander
Jakob Nielsen 1993, 2010	**Jakob Nielsen** 1993, 2010	**Jakob Nielsen** 1993, 2010	**Jakob Nielsen** 1993, 2010

> **When companies make us wait instead of providing responsive service, they seem either arrogant or incompetent.**
>
> Jakob Nielsen, “Website Response Times” 2010

The research

Observability Day
NORTH AMERICA

500 ms slower, same content, blind test on mobile

Radware 2013 · via Planet Performance

The research

500 ms slower, same content, blind test on mobile



Standard speed

Radware 2013 · via Planet Performance

The research

500 ms slower, same content, blind test on mobile



Standard speed



+500 ms slowdown

Radware 2013 · via Planet Performance

Fast with AI doesn’t always feel reliable

Observability Day
NORTH AMERICA

Fast with AI doesn’t always feel reliable

Observability Day
NORTH AMERICA

2s

Answers feel less thoughtful. Two seconds is unreliable

Fang-Yi Tan & Nov
2026

Fast with AI doesn’t always feel reliable

Observability Day
NORTH AMERICA

2s

Answers feel less thoughtful. Two seconds is unreliable

Fang-Yi Tan & Nov
2026

9s

The sweet spot for perceived quality of an AI response

Fang-Yi Tan & Nov
2026

Fast with AI doesn’t always feel reliable

Observability Day
NORTH AMERICA

2s

Answers feel less thoughtful. Two seconds is unreliable.

Fang-Yi Tan & Nov
2026

9s

The sweet spot for perceived quality of an AI response.

Fang-Yi Tan & Nov
2026

>20s

Perceptions of reliability drop.

Fang-Yi Tan & Nov
2026

Observability Day
NORTH AMERICA

Fast is wired into our neuropsychology.

The cost to brand reputation is high.

Our perceptions with AI are different.

Slow costs money

Every 100ms in added page load time cost 1% in revenue

2006 Amazon study

The research

Observability Day
NORTH AMERICA



20%

30 search results vs 10

=

0.5 seconds longer load

=

-20% traffic & revenue

Google 2006

The research

Observability Day
NORTH AMERICA



$1,670

Small business with fewer than 25 employees can lose **~$1,670/minute** when a server goes down

ITIC
2006

The research

Observability Day
NORTH AMERICA



$1,670

Small business with fewer than 25 employees can lose **~$1,670/minute** when a server goes down

ITIC
2006



>50%

Mobile site visits are abandoned once load time crosses **3 seconds**

Google
2016

The research

Observability Day
NORTH AMERICA

$1,670

Small business with fewer than 25 employees can lose **~$1,670/minute** when a server goes down

ITIC
2006

>50%

Mobile site visits are abandoned once load time crosses **3 seconds**

Google
2016

$30,000

High-impact outage costs can reach **~$30,000/minute**

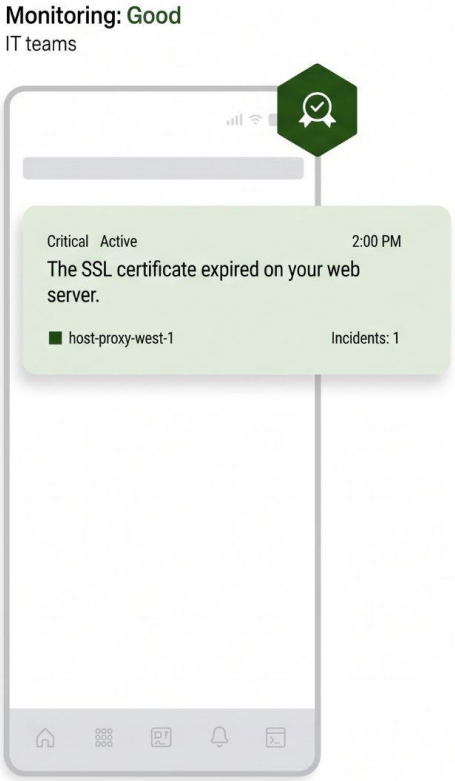
New Relic
2026

Observability Day
NORTH AMERICA

Business Observability

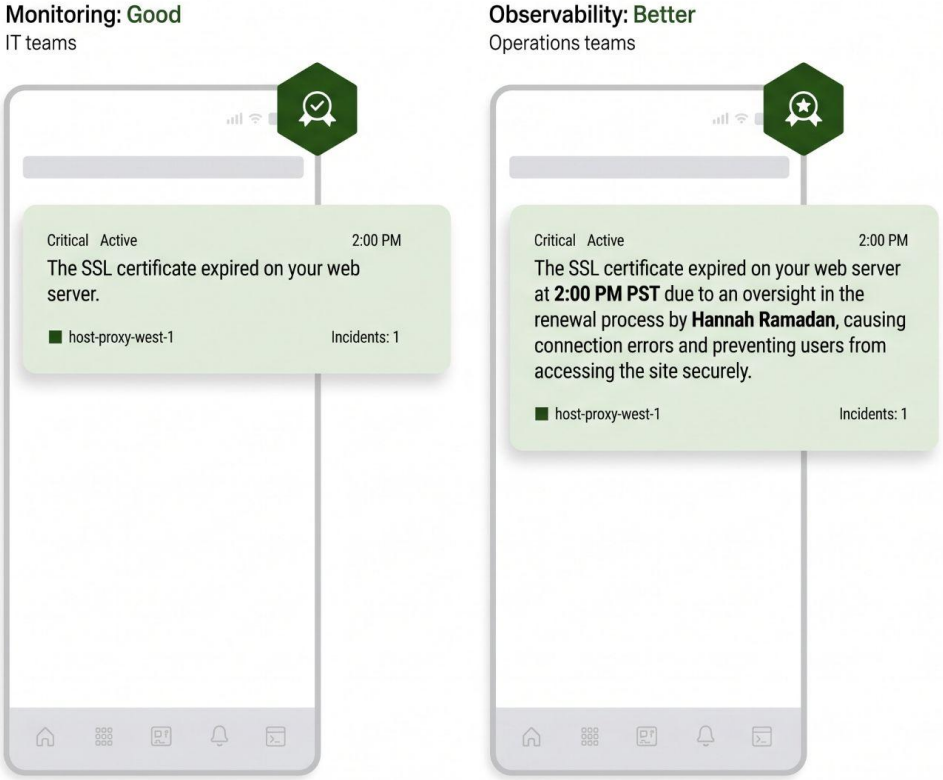
‘Integrating technical signals with business metrics’

Observability Maturity Model



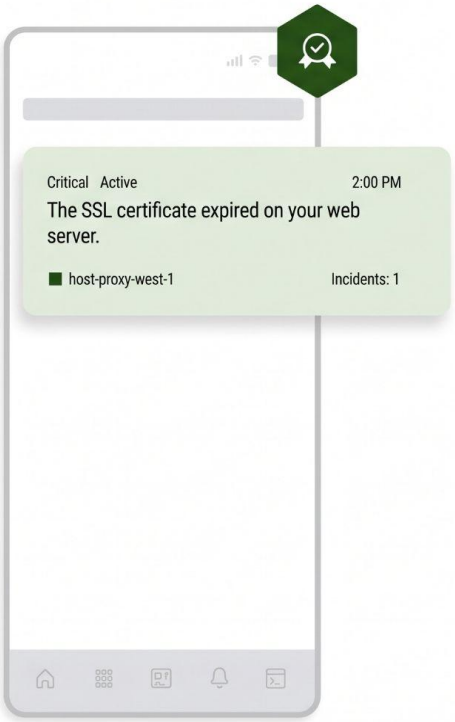
Observability Maturity Model

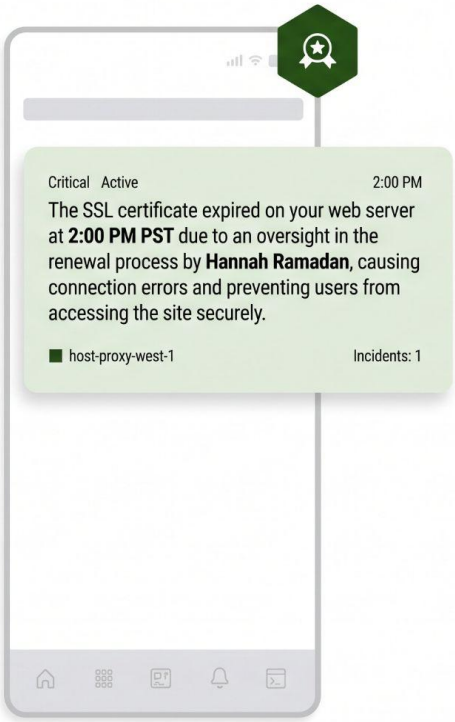
Observability Day
NORTH AMERICA

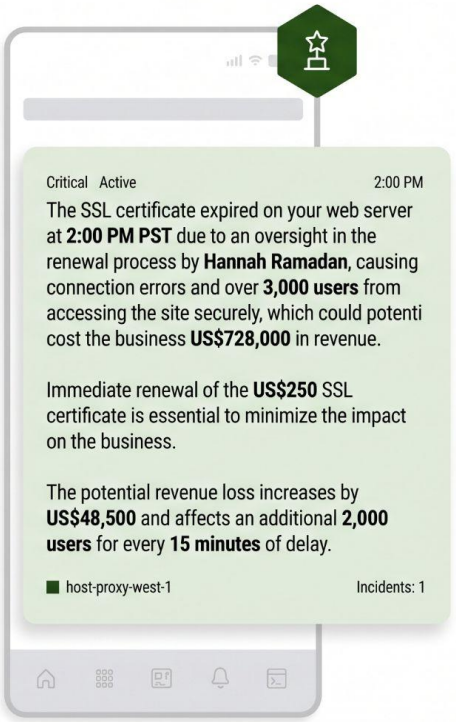


Observability Maturity Model

Observability Day
NORTH AMERICA







Key Performance Indicator

Measures of progress towards a goal that allow businesses to make data driven decisions about their strategy

Business and technology, together since 1979

Observability Day
NORTH AMERICA

1979

John F. Rockart

Director, Center for Information Systems Research
MIT Sloan School of Management

"Critical success factors"

”

[Performance] should be continually measured, and that information should be made available.

Business and technical metrics

Observability Day
NORTH AMERICA

BUSINESS METRICS

- Conversion Rate
- Churn Rate
- Customer Satisfaction
- Support Costs

Business and technical metrics

Observability Day
NORTH AMERICA

BUSINESS METRICS

- Conversion Rate
- Churn Rate
- Customer Satisfaction
- Support Costs

TECHNICAL METRICS

- p99 Latency
- Error Rate
- Apdex Score
- Cache Hit Rates

Building the bridge with two questions

Observability Day
NORTH AMERICA



1. User Flow

Which user flow leads directly a business outcome?

Building the bridge with two questions

Observability Day
NORTH AMERICA



1. User Flow

Which user flow leads directly a business outcome?



2. Measurement

What event or metric lets us measure it?

Observability Day
NORTH AMERICA

Business Observability Formula

BUSINESS OBSERVABILITY FORMULA

TOTAL
cost

=

VOLUME
how many

×

RATE
how much each one costs

Business and technical metrics

Observability Day
NORTH AMERICA

BUSINESS METRICS

- Conversion Rate
- Churn Rate
- Customer Satisfaction
- Support Costs

X

TECHNICAL METRICS

- p99 Latency
- Error Rate
- Apdex Score
- Cache Hit Rates

Calculating cost takeaway

Partner with your colleagues in finance and sales to get the answers to questions asked by business metrics

KPIs + signals

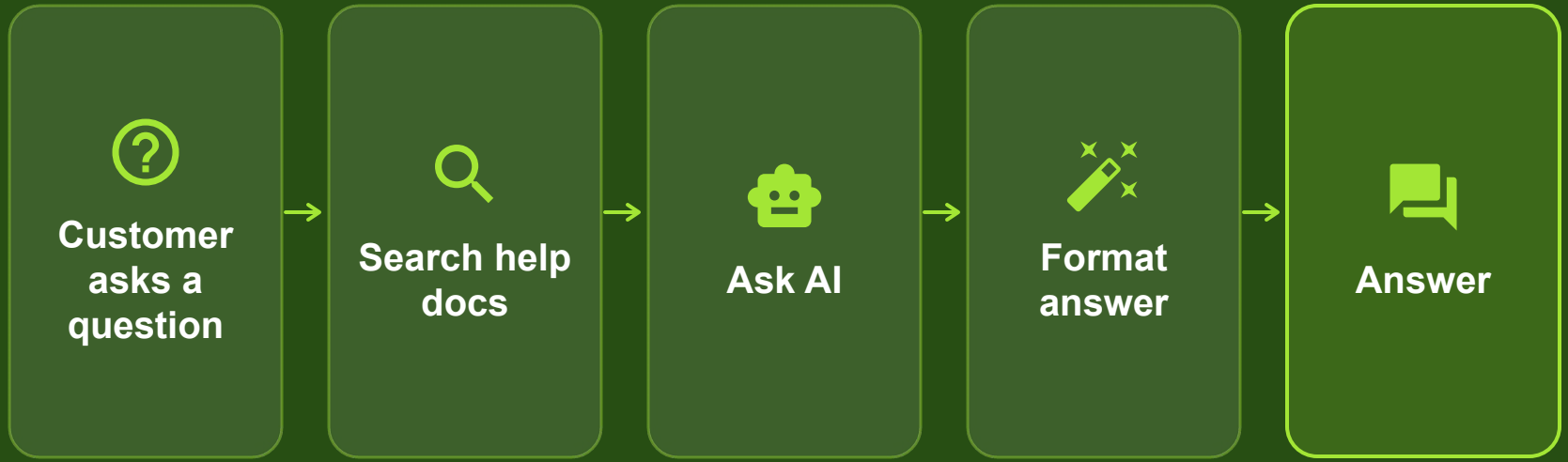
Observability Day
NORTH AMERICA

KPI	How many (telemetry)	How much each (business)
Conversion rate	Checkouts abandoned after a slow page	**Average order value**
Churn	Customers with repeated slow sessions who cancel	**Monthly revenue per customer**
SLA	Minutes over the SLA threshold	**Credit owed per minute**
Support costs	Slow answers that turn into tickets	**Cost per ticket**

Example

OUR EXAMPLE

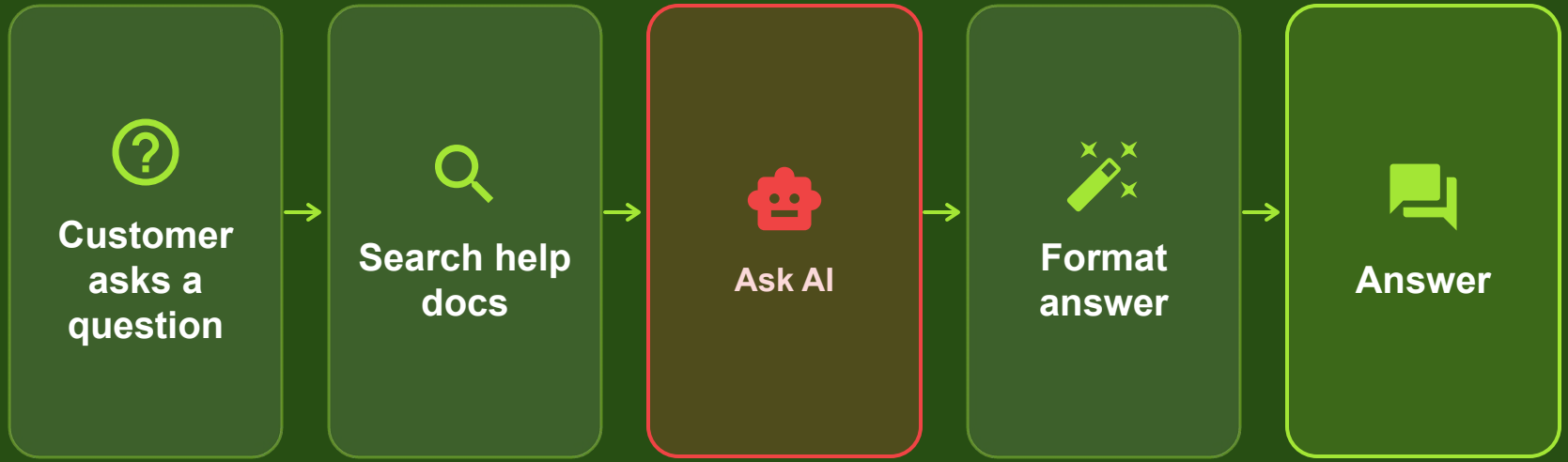
AI Customer Support Assistant



OUR EXAMPLE

AI Customer Support Assistant

Observability Day
NORTH AMERICA



Goal: Reduce costs in support

Observability Day
NORTH AMERICA

KPI

$ Cost per Contact

Telemetry

HOW MANY

% of conversations with a response time <5 seconds

Business

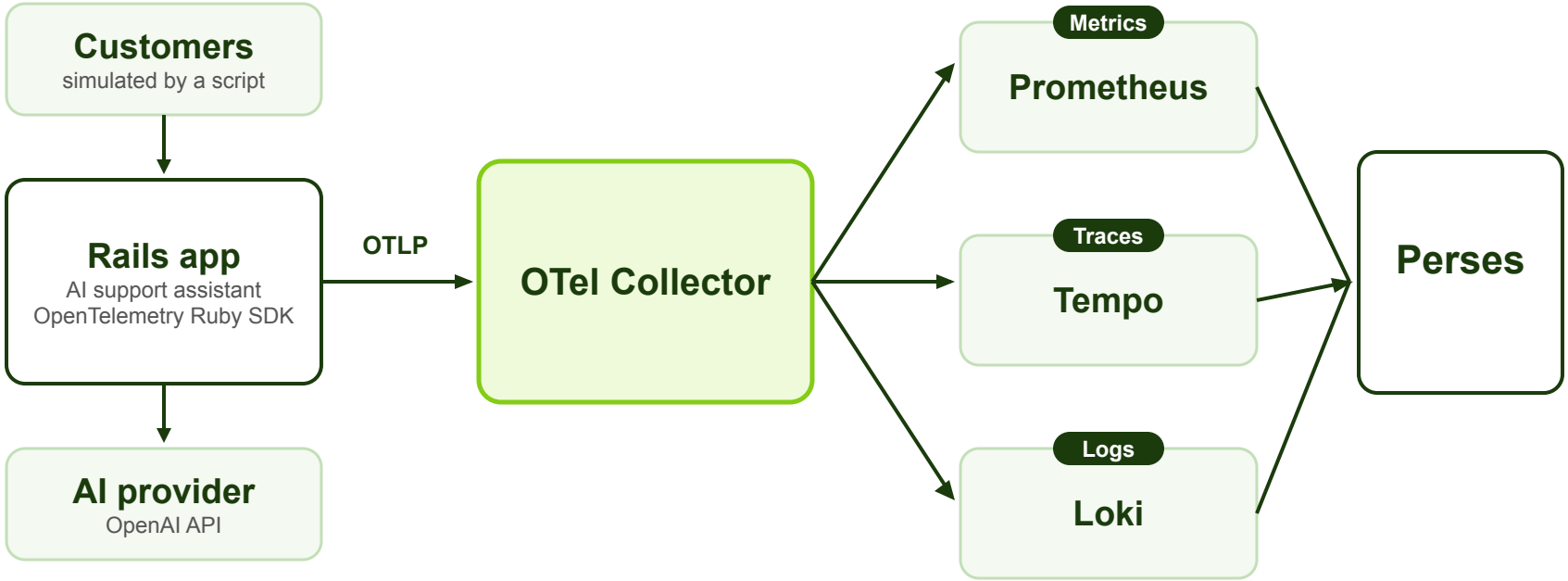
HOW MUCH

$15 per support ticket filed with a human

$0.20 / $1.25 per million input/output tokens

Our example: how the telemetry flows

Observability Day
NORTH AMERICA



Metrics - “How many answers were too slow?”

Traces - “Who was affected, and which step was slow?”

Logs - “What did it actually cost?”

Metrics

“How many answers were too slow?”

Metrics

Observability Day
NORTH AMERICA

Metric: `gen_ai.client.inference.duration`

This metric SHOULD be specified with ExplicitBucketBoundaries of [0.01, 0.02, 0.04, 0.08, 0.16, 0.32, 0.64, 1.28, 2.56, 5.12, 10.24, 20.48, 40.96, 81.92].

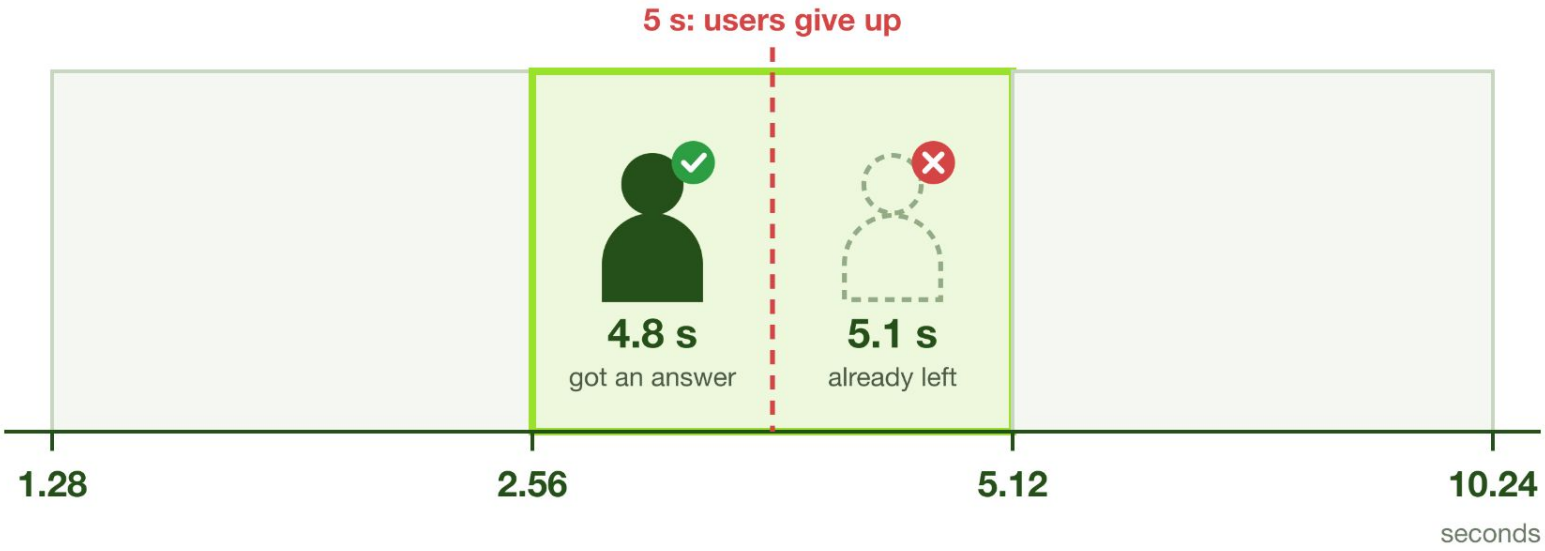
Name	Instrument Type	Unit (UCUM)	Description	Stability	Entity Associations
`gen_ai.client.inference.duration`	Histogram	`s`	GenAI client inference operation duration. [1]	development	

[1]: Represents the end-to-end duration of a model inference operation, measured from when the client issues the inference request to when the response is fully received or the operation ends with an error or cancellation. When this metric is reported alongside a `gen_ai.client.inference` span, the metric value SHOULD be the same as the span duration.

Requirement level: Recommended.

Before: the default buckets

Observability Day
NORTH AMERICA



Same bucket. To the histogram, they're roommates.

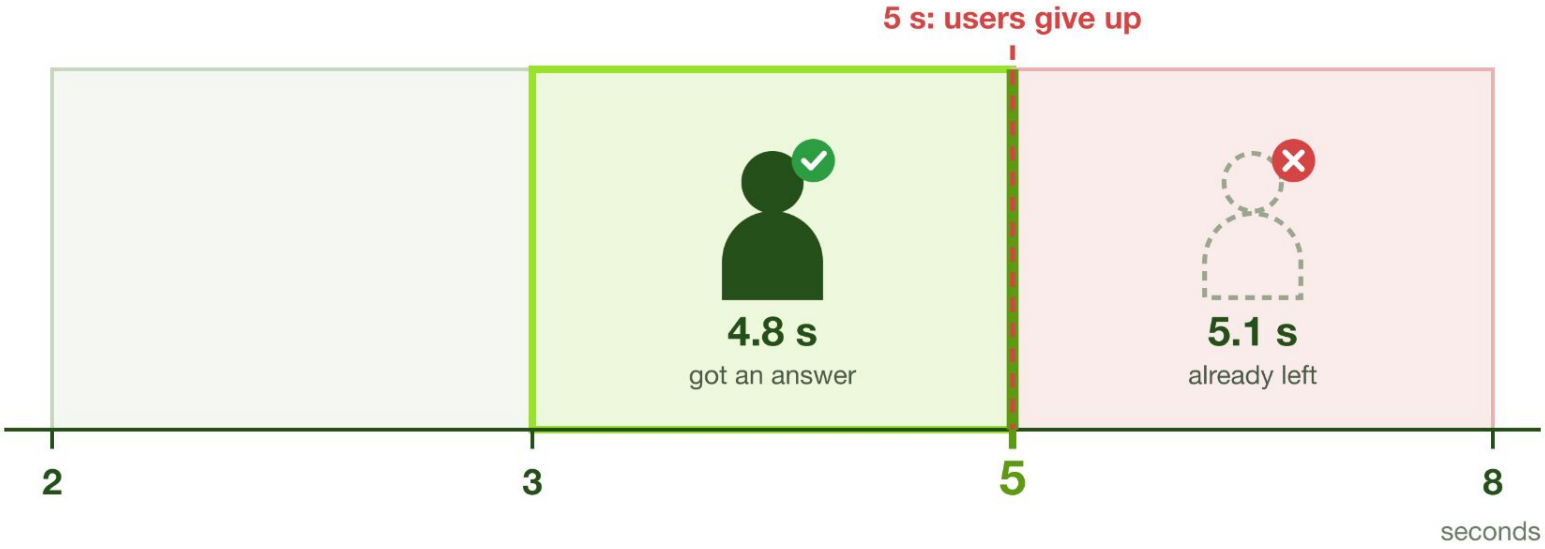
Recommended by the OpenTelemetry GenAI conventions: boundaries that double … 1.28, 2.56, 5.12, 10.24 …

Metrics

Observability Day
NORTH AMERICA

```ruby
OpenTelemetry.meter_provider.add_view(
  "gen_ai.client.inference.duration",
  type: :histogram,
  aggregation: OpenTelemetry::SDK::Metrics::Aggregation::ExplicitBucketHistogram.new(
    boundaries: [0.5, 1, 2, 3, 5, 8, 13]
  )
)
```

After: a boundary at 5 seconds



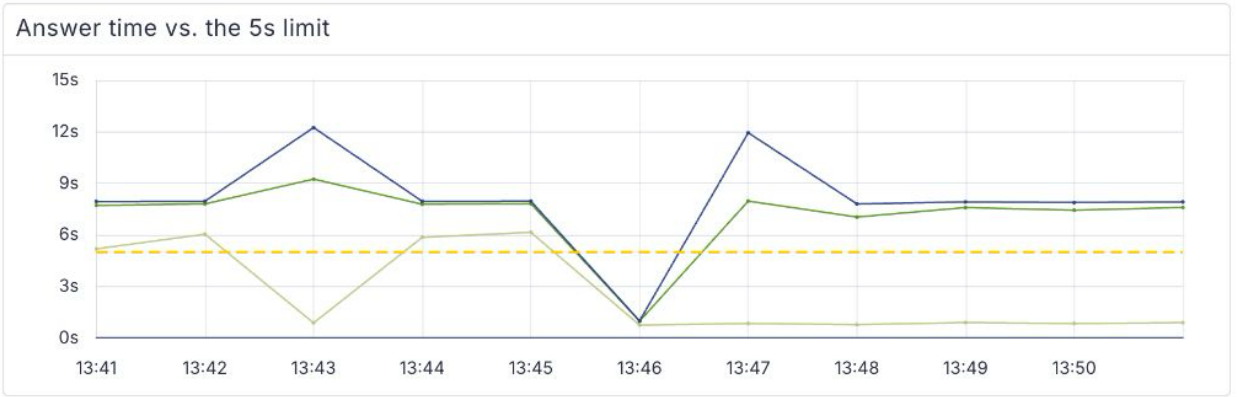
Different buckets. Now we can count who waited too long.

Set with a View: boundaries [0.5, 1, 2, 3, 5, 8, 13]

Metrics - slow day dashboard

Observability Day
NORTH AMERICA





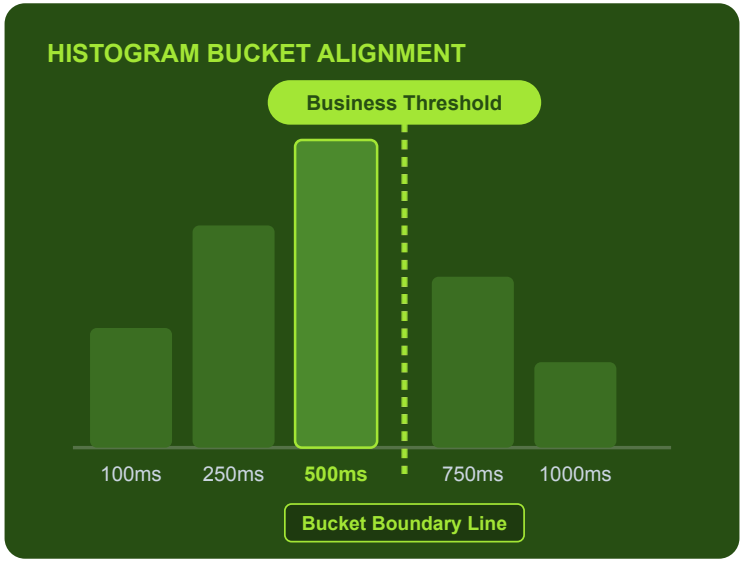
Metrics

Observability Day
NORTH AMERICA

Business Impact & Cost Calculation

Step 1	×	Conversion	×	Unit Cost	=	Total Impact
170		30%		$15		$765
too-slow answers		become tickets		cost per ticket		estimated cost

Metrics takeaway



Put a histogram bucket boundary at the threshold your business cares about

Align histogram buckets with SLA objectives.

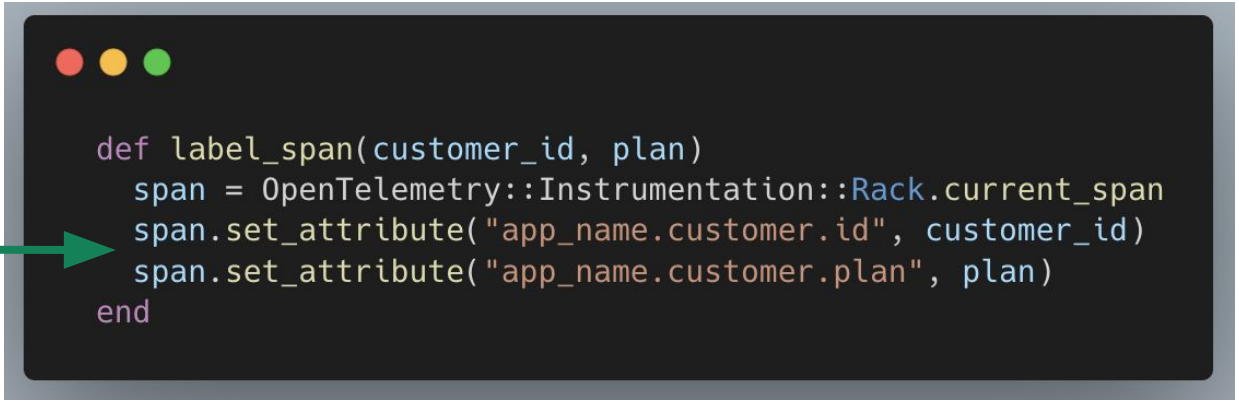
Traces

“Who was affected, and which step was slow?”

Traces

Totals go on metrics, details go on spans

Add customer id and customer plan to spans



```ruby
def label_span(customer_id, plan)
  span = OpenTelemetry::Instrumentation::Rack.current_span
  span.set_attribute("app_name.customer.id", customer_id)
  span.set_attribute("app_name.customer.plan", plan)
end
```

Traces

Now we can create a query



Traces

Connect spans to business moments

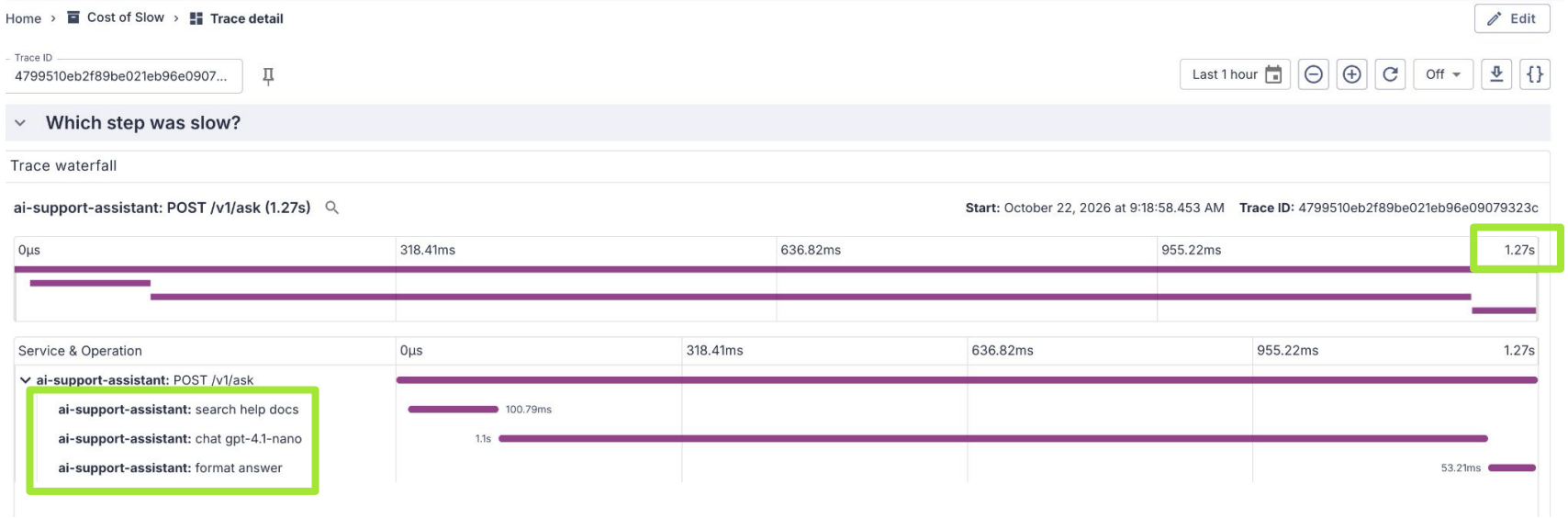
```ruby
# Initialize a tracer
TRACER = OpenTelemetry.tracer_provider.tracer("support_assistant")

# Add named span ‘search docs’ to trace
def call(question)
 TRACER.in_span("search help docs") do
   # important help doc searching code
  end
end
```

```ruby
# Add named span ‘format answer’ to trace
def call(text)
  TRACER.in_span("format answer") do
    #important format answer code
  end
end
```

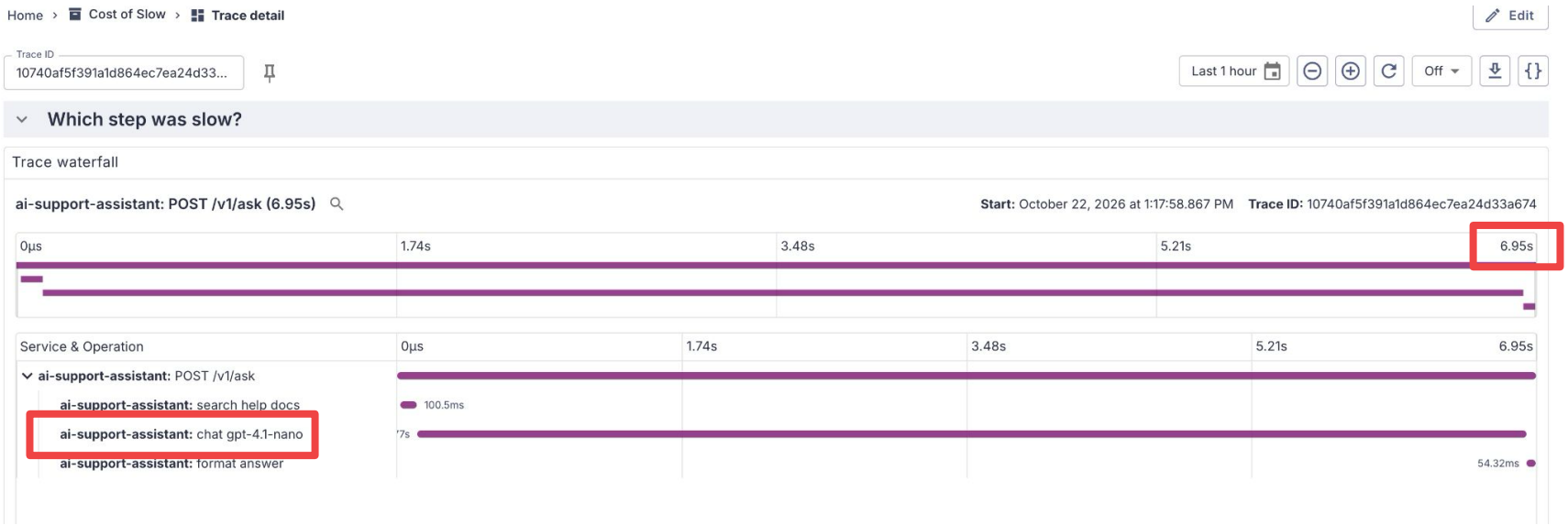
Traces - below threshold

Observability Day
NORTH AMERICA



Traces - above threshold

Observability Day
NORTH AMERICA



Traces takeaway

Observability Day
NORTH AMERICA

Attach business attributes, like plan or account_id, to your spans



Plan

`customer.plan: "enterprise"`



Account

`account.id: "acc_98412"`

Logs

“What did it actually cost?”

Two key log events

Observability Day
NORTH AMERICA

`app.assistant.user_gave_up`

The customer left before the answer arrived. It records the tokens we'd already paid for.

```ruby
OpenTelemetry.logger_provider.logger(name: "support_assistant").on_emit(
  timestamp: Time.now,
  event_name: "app.assistant.user_gave_up",
  attributes: {
    "app.customer.id" => entry[:customer_id],
    "app.customer.plan" => entry[:plan],
    "gen_ai.usage.input_tokens" => entry[:tokens][:input],
    "gen_ai.usage.output_tokens" => entry[:tokens][:output],
    "event.name" => "app.assistant.user_gave_up"
  }
)
```

Two key log events

Observability Day
NORTH AMERICA

`app.assistant.user_gave_up`

The customer left before the answer arrived. It records the tokens we'd already paid for.

```ruby
OpenTelemetry.logger_provider.logger(name: "support_assistant").on_emit(
  timestamp: Time.now,
  event_name: "app.assistant.user_gave_up",
  attributes: {
    "app.customer.id" => entry[:customer_id],
    "app.customer.plan" => entry[:plan],
    "gen_ai.usage.input_tokens" => entry[:tokens][:input],
    "gen_ai.usage.output_tokens" => entry[:tokens][:output],
    "event.name" => "app.assistant.user_gave_up"
  }
)
```

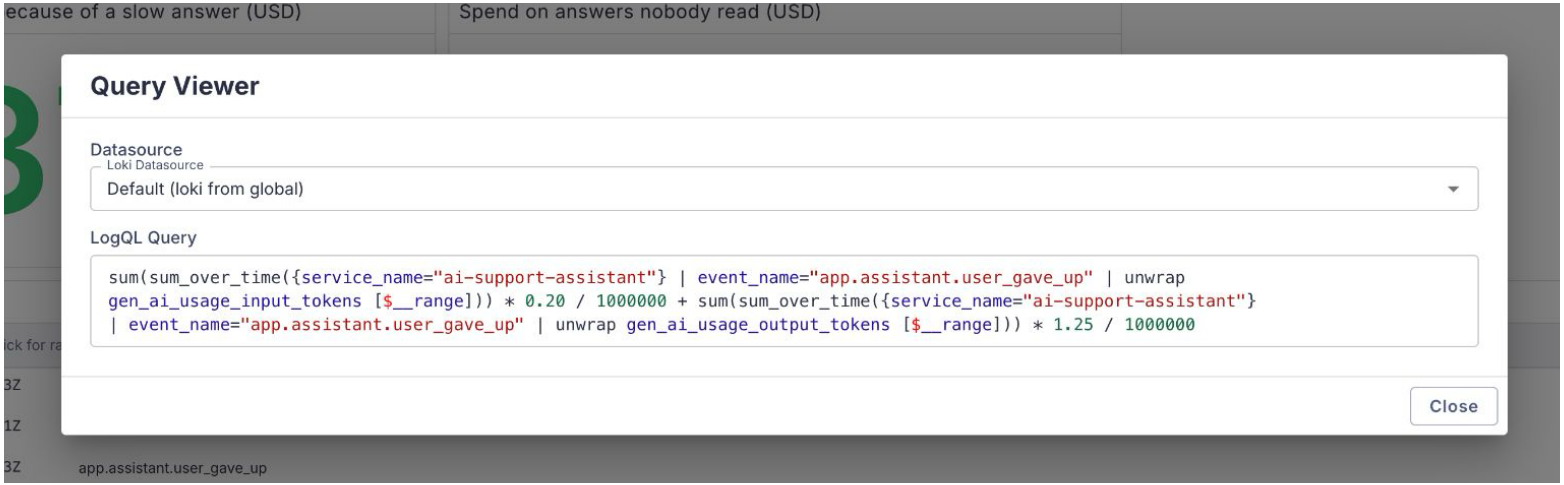
`app.support.ticket_opened`

The customer asked for a human, with a reason attached.

```ruby
OpenTelemetry.logger_provider.logger(name: "support_assistant").on_emit(
  timestamp: Time.now,
  event_name: "app.support.ticket_opened",
  attributes: {
    "app.customer.id" => customer_id,
    "app.customer.plan" => plan,
    "app.ticket.reason" => reason,
    "event.name" => "app.support.ticket_opened"
  }
)
```

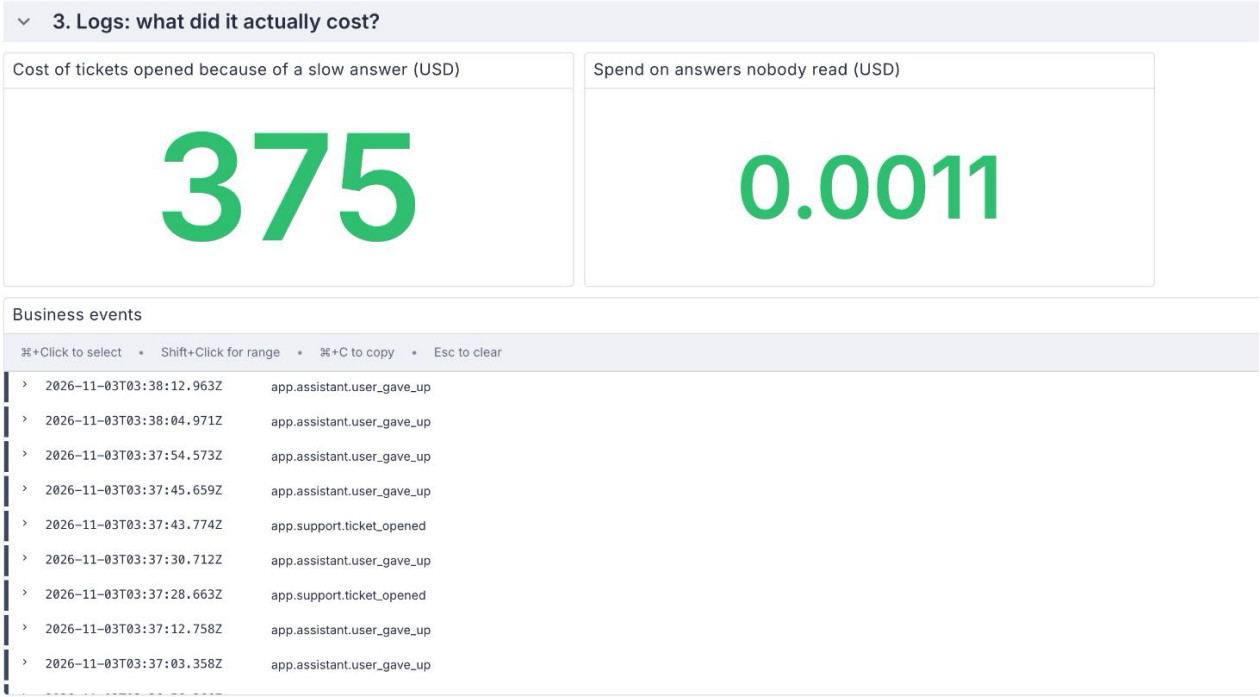
Logs - query

Observability Day
NORTH AMERICA



Logs dashboard

Observability Day
NORTH AMERICA



Logs

Observability Day
NORTH AMERICA

SLOW ANSWER IMPACT

[$375]

[25] tickets opened because of a slow answer × [$15] = [$375]

WASTED TOKENS

[$0.001]

Tokens paid for answers nobody read, but combining input ($0.2/M) and output ($1.25/M) tokens

Logs takeaway

Emit an event that closes the gap between your existing telemetry and your business question.

Observability Day
NORTH AMERICA

Semantic Conventions
“What do the conventions say about cost?”

Cost in the conventions

Observability Day
NORTH AMERICA

How many

Well covered: what you're billed for

gen_ai.usage.input_tokens
tokens sent to an AI model

gen_ai.usage.output_tokens
tokens the model sent back

aws.dynamodb.consumed_capacity
database capacity used

faas.invocations
serverless function calls

faas.invoke_duration
how long each call ran

faas.mem_usage
memory each call used

faas.net_io
network traffic each call used

Cost in the conventions

Observability Day
NORTH AMERICA

How many

Well covered: what you're billed for

gen_ai.usage.input_tokens
tokens sent to an AI model

gen_ai.usage.output_tokens
tokens the model sent back

aws.dynamodb.consumed_capacity
database capacity used

faas.invocations
serverless function calls

faas.invoke_duration
how long each call ran

faas.mem_usage
memory each call used

faas.net_io
network traffic each call used

Who pays

Just started

service.instance.cost_center.id
the part of the business that

service.instance.cost_center.name
pays for a service (new in v1.45.0)

cloud.account.id
the account on the cloud bill

Proposed

service.business_unit.id
which business unit

service.owner
who owns a service

A tenant or customer attribute
which customer it was for

Cost in the conventions

Observability Day
NORTH AMERICA

How many

Well covered: what you're billed for

gen_ai.usage.input_tokens
tokens sent to an AI model

gen_ai.usage.output_tokens
tokens the model sent back

aws.dynamodb.consumed_capacity
database capacity used

faas.invocations
serverless function calls

faas.invoke_duration
how long each call ran

faas.mem_usage
memory each call used

faas.net_io
network traffic each call used

Who pays

Just started

service.instance.cost_center.id
the part of the business that

service.instance.cost_center.name
pays for a service (new in v1.45.0)

cloud.account.id
the account on the cloud bill

Proposed

service.business_unit.id
which business unit

service.owner
who owns a service

A tenant or customer attribute
which customer it was for

How much

Proposals only

gen_ai.usage.cost.amount
the cost of one AI call

gen_ai.usage.cost.currency
for example, USD

Cloud cost conventions
an open issue since 2024

Observability Day
NORTH AMERICA

Wrapping up

Three questions, one for each signal

Observability Day
NORTH AMERICA

Metrics	How many answers were too slow?

Three questions, one for each signal

Observability Day
NORTH AMERICA

Metrics	How many answers were too slow?
Traces	Who was affected, and which step was slow?

Three questions, one for each signal

Observability Day
NORTH AMERICA

Metrics	How many answers were too slow?
Traces	Who was affected, and which step was slow?
Logs	What did it actually cost?

Questions to start your investigation

Observability Day
NORTH AMERICA

01 Which user flow is the direct path to a business outcome?

02 What application event or metric lets us measure it?

03 What information do you need from sales or finance to close the gap?

04 What data is missing from your metrics / traces / logs to answer this question?

05 Is there anything in the semantic conventions that might help?

Thank you

Observability Day
NORTH AMERICA

GITHUB REPOSITORY

github.com/
hannahramadan/
2026_observabilityday_cost_of_slow

- ✓ Slide deck
- ✓ Example application
- ✓ Resource links & citations

FEEDBACK

"What's the Dollar Cost of Slow?": Answering Business Questions with OTel