LA Convention Center | June 10, 2026

AWS SUMMIT

The Unit Economics of AI Agents

Guille Ojeda 🇦🇷

Principal Innovation Architect @ Caylent

AWS Community Builder

AWS User Group Leader - AI Argentina UG

Andres Moreno 🇲🇽🇺🇸

Principal Architect @ Caylent

AWS Community Builder

AWS User Group Leader - Cloud Del Norte UG

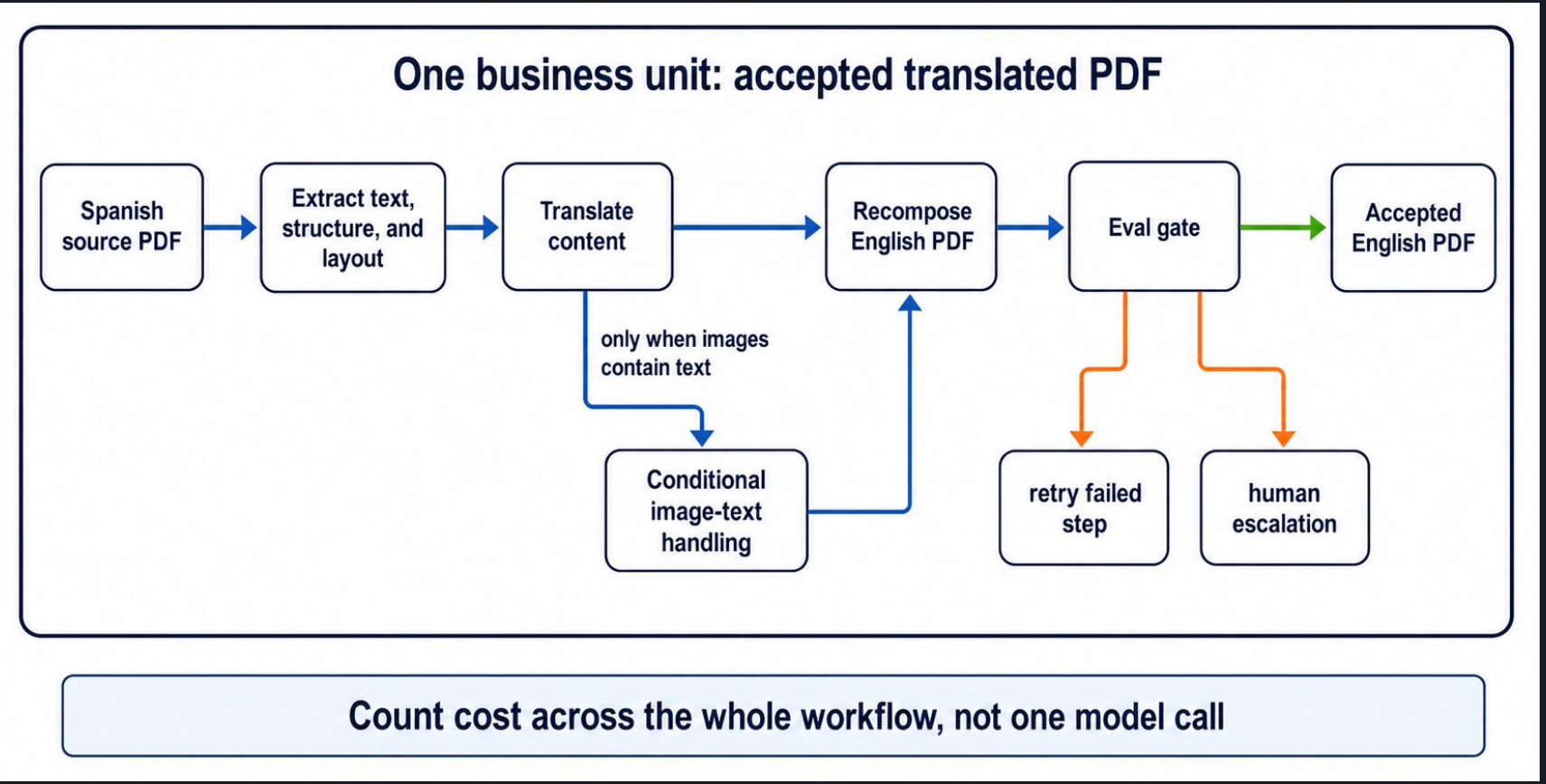
Raise your hand if you have...

Built something impressive with AI

Solved a real problem with AI

Validated that the value is greater than the cost

Spanish PDF to accepted English PDF



Five steps to prove unit economics

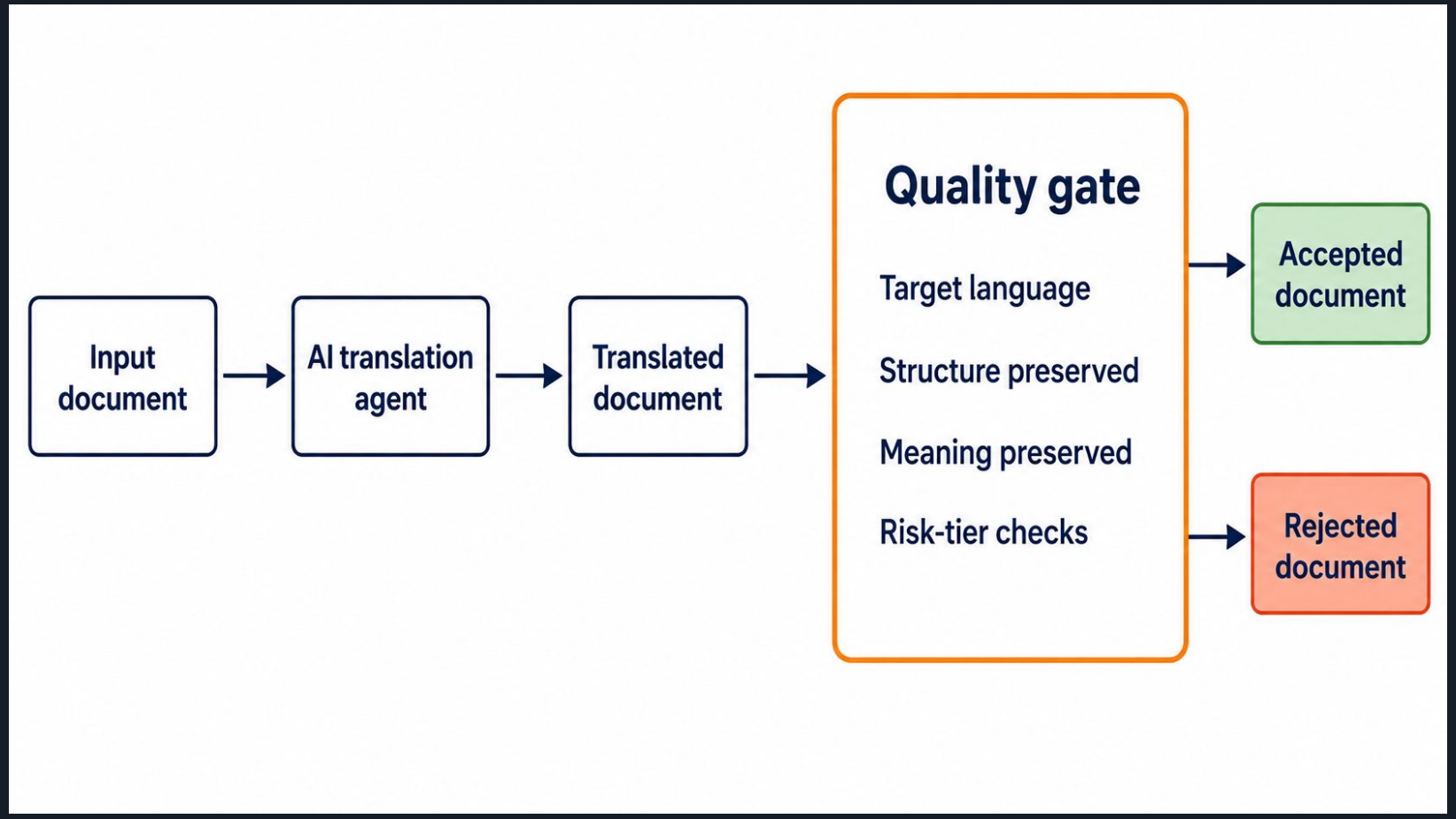
1. Define acceptance criteria
2. Estimate value per unit
3. Build a measurable POC
4. Measure cost per result
5. Iterate for margin

Scale only after proving value > cost.

1. Define acceptance criteria

Before doing a good job, define what "good" means.

When do we accept a translated document?



2. Estimate value per unit

Start conservative. Use the manual baseline first.

What happens without the agent?

Manual work → labor avoided

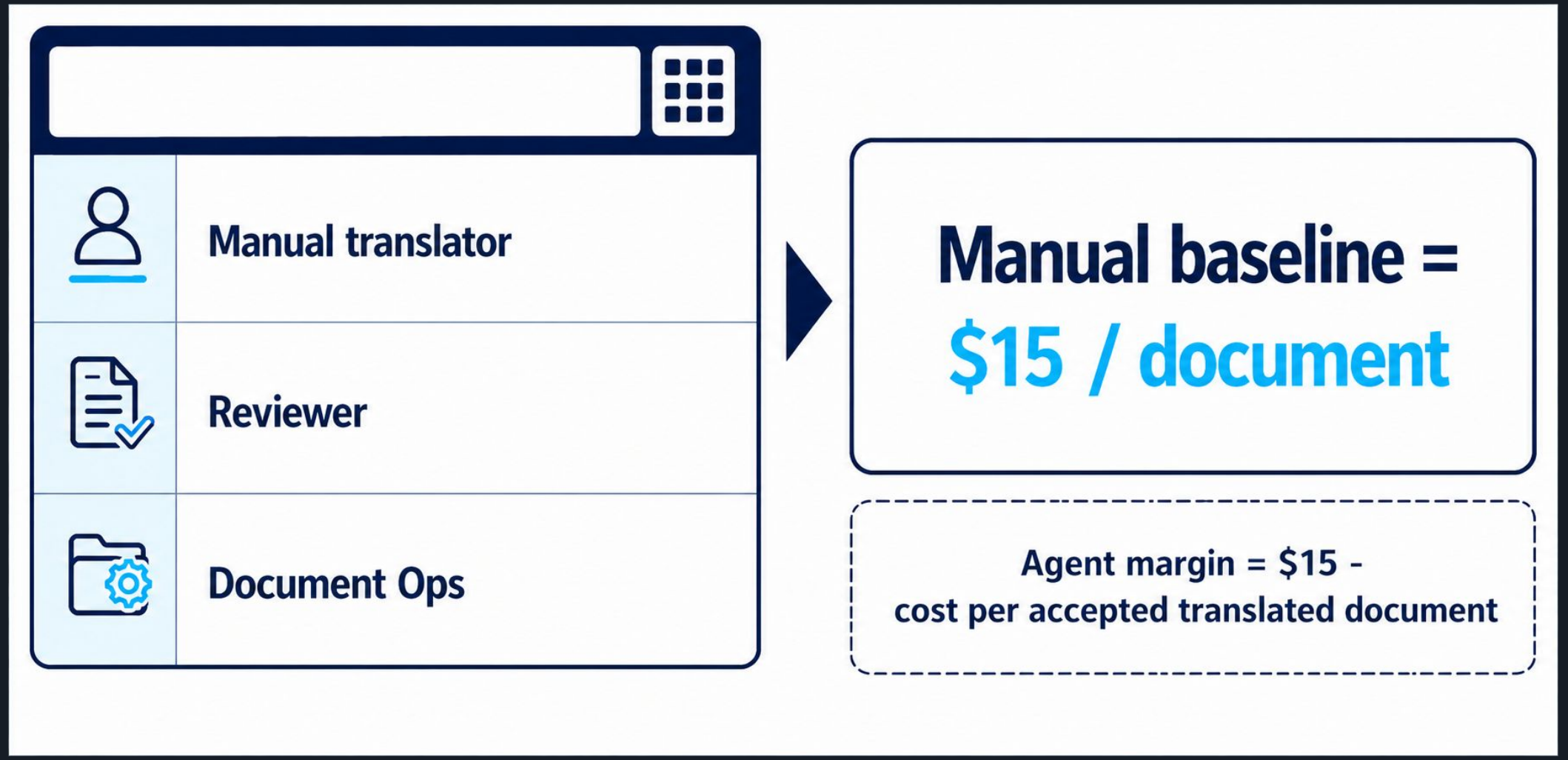
Delayed work → cycle time gained

Unfinished work → throughput unlocked

Inconsistent work → risk reduced

New capability → revenue enabled

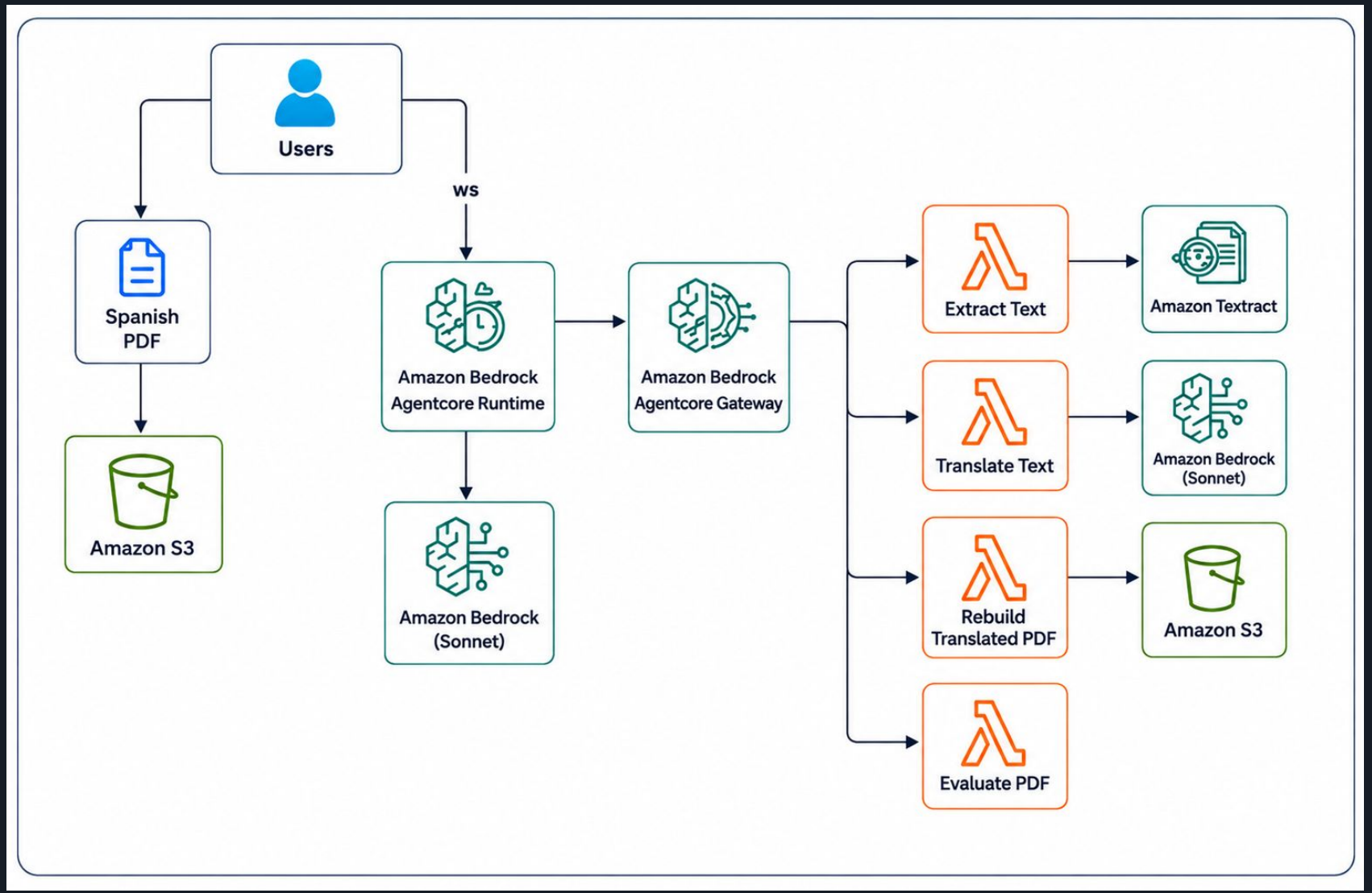
Start with the manual baseline



3. Build a measurable POC

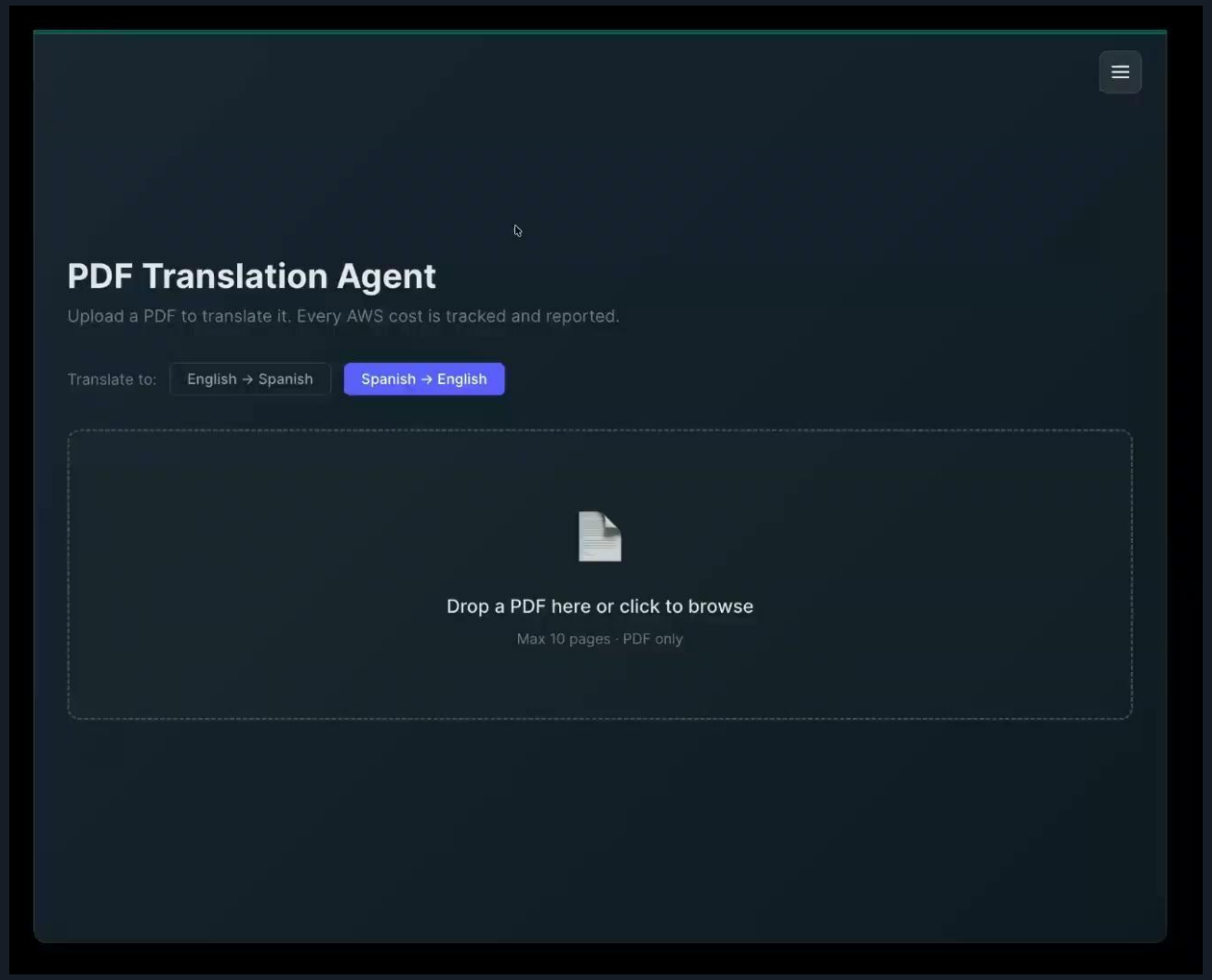
Build the first version, add observability, measure costs

Architecture



DEMO!

One accepted document run, from upload to eval and trace/cost evidence.



4. Measure cost per result

We know the value per translated PDF. Let's see how much it costs.

Instrumentation

Orchestration Agent

```
async for event in agent.stream_async(request):
   # result event fires once at the end with full
accumulated usage
   if 'result' in event:
       res = event['result']
       metrics = getattr(res, 'metrics', None)
       usage = getattr(metrics, 'accumulated_usage',
None)
       if usage is not None:
           accumulated_input, accumulated_output =
_extract_usage(usage)

cycles = getattr(getattr(agent,
'event_loop_metrics', None), 'cycle_count', 0)
job_id = _extract_job_id(request)
_write_orchestrator_metrics(job_id, accumulated_in
```

Instrumentation

Text Extraction

```
costs = {
  "textract": {
    "pages": num_pages,
  },
  "lambda": {
      "memoryMb": memory_mb,
      "durationMs": round(elapsed_ms),
      "gbSeconds": round(gb_seconds, 4),
  },
  "s3": {
      "getRequests": 1,
      "putRequests": 1,
  },
}
```

Instrumentation

Translation

```
costs = {
    "bedrock": {
        "model": MODEL_ID,
        "inputTokens": total_input_tokens,
        "outputTokens": total_output_tokens,
        "calls": total_calls,
        "blocks": len(blocks),
    },
    "lambda": {
        "memoryMb": memory_mb,
        "durationMs": round(elapsed_ms),
        "gbSeconds": round(gb_seconds, 4),
    },
    "s3": {
        "getRequests": 1,
        "putRequests": 1,
    },
}
```

Instrumentation

Reassemble PDF

```
costs = {
    "lambda": {
        "memoryMb": memory_mb,
        "durationMs": all_lambda_duration_ms,
        "gbSeconds": all_lambda_gb_seconds,
    },
    "s3": {
        "getRequests": 1,
        "putRequests": 1,
    }
}
```

Instrumentation

Evals

```
eval_costs = {
    "lambda": {
        "memoryMb": memory_mb,
        "durationMs": round(elapsed_ms),
        "gbSeconds": round(gb_seconds, 6),
    },
    "s3": {
        "getRequests": 1,
    },
}
```

Naive Cost Modeling

AI MODEL COST (NAIVE VIEW)

$0.001455

Processing time: **48.9s**

NAIVE
$0.001455 → FULL LEDGER
$0.017105

+$0.015650 **+1076% more**

Naive | Full ledger | Ledger | Unit Economics | Failure Economics | Model Comparison

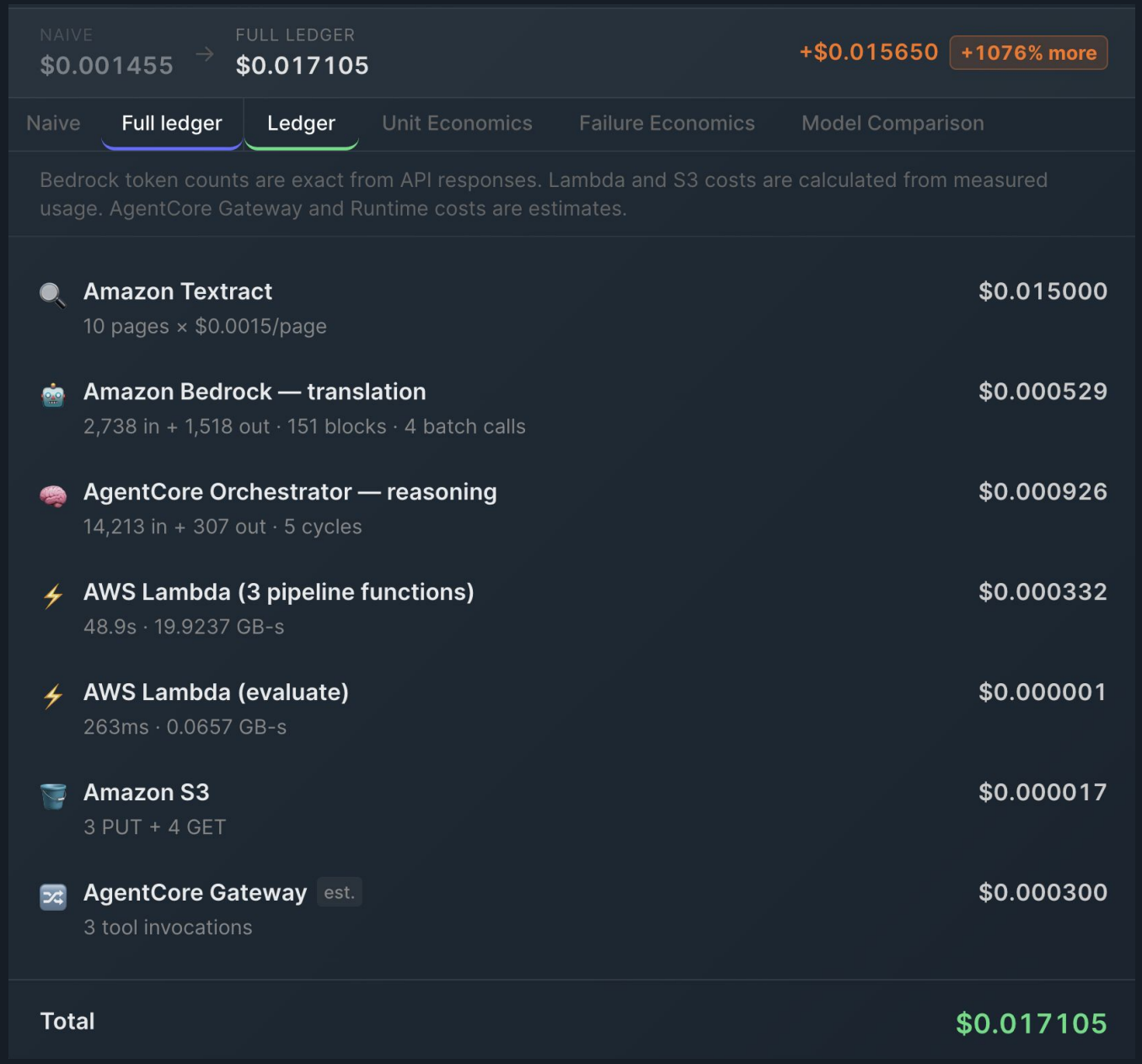
Showing only the AI model cost — the number most people quote.

🤖 **Amazon Bedrock — translation** — $0.000529
2,738 in + 1,518 out · 151 blocks · 4 batch calls

🧠 **AgentCore Orchestrator — reasoning** — $0.000926
14,213 in + 307 out · 5 cycles

Total **$0.001455**

Complete Cost Modeling



Unit Economics

877×

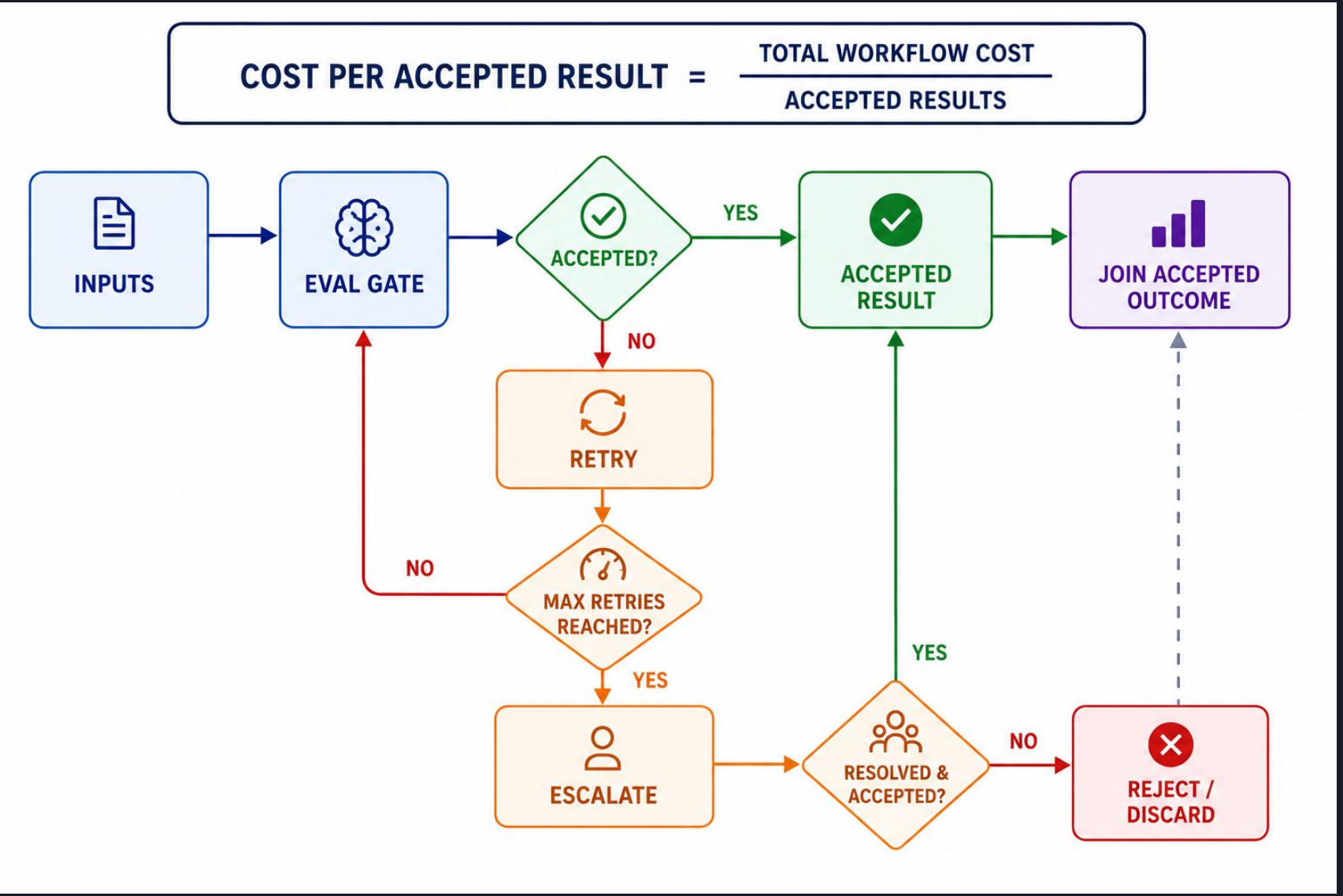
cheaper than manual per translation

Manual translation cost	$15.00
Agent cost per translation	$0.017105
Unit margin	**$14.98 (99.886%)**

AT SCALE illustrative

100 translations	$1.71 cost → **$1,500 value**
1,000 translations	$17.11 cost → **$15,000 value**
10,000 translations	$171.05 cost → **$150,000 value**

What happens when translation fails?



What happens when translation fails?

SCENARIO	RATE	COST PER CASE
First-pass success	95%	$0.017105
Agent retry	4%	2× $0.017105
Human escalation	1%	2× $0.017105 + $7.50 human

Weighted expected cost	$0.092960
Effective unit margin	**$14.91 (99.4%)**

5. Iterate for margin

Keep working on the agent. Use the cost ledger as input.

Cheapest model ≠ cheapest workflow

Claude Haiku 4.5

Cost per run	$0.008262
100 docs × run cost	$0.826240
6 retry runs	$0.049574
3 escalated × $7.50	$22.50

94% pass 3% retry 3% escalated

Total / 100 docs	$23.38
Docs accepted	97
Cost per accepted	**$0.240988**

Claude Opus 4.7

BEST VALUE

Cost per run	$0.154920
100 docs × run cost	$15.49
1 retry runs	$0.154920
1 escalated × $7.50	$7.50

99% pass 1% escalated

Total / 100 docs	$23.15
Docs accepted	99
Cost per accepted	**$0.233807**

Value per unit > full cost to produce it

Before you scale, answer five questions:

1. What's the acceptance criteria?
2. What is the value per unit?
3. What's an upper bound on costs?
4. What parts dominate the cost?
5. What can we improve that brings costs down?

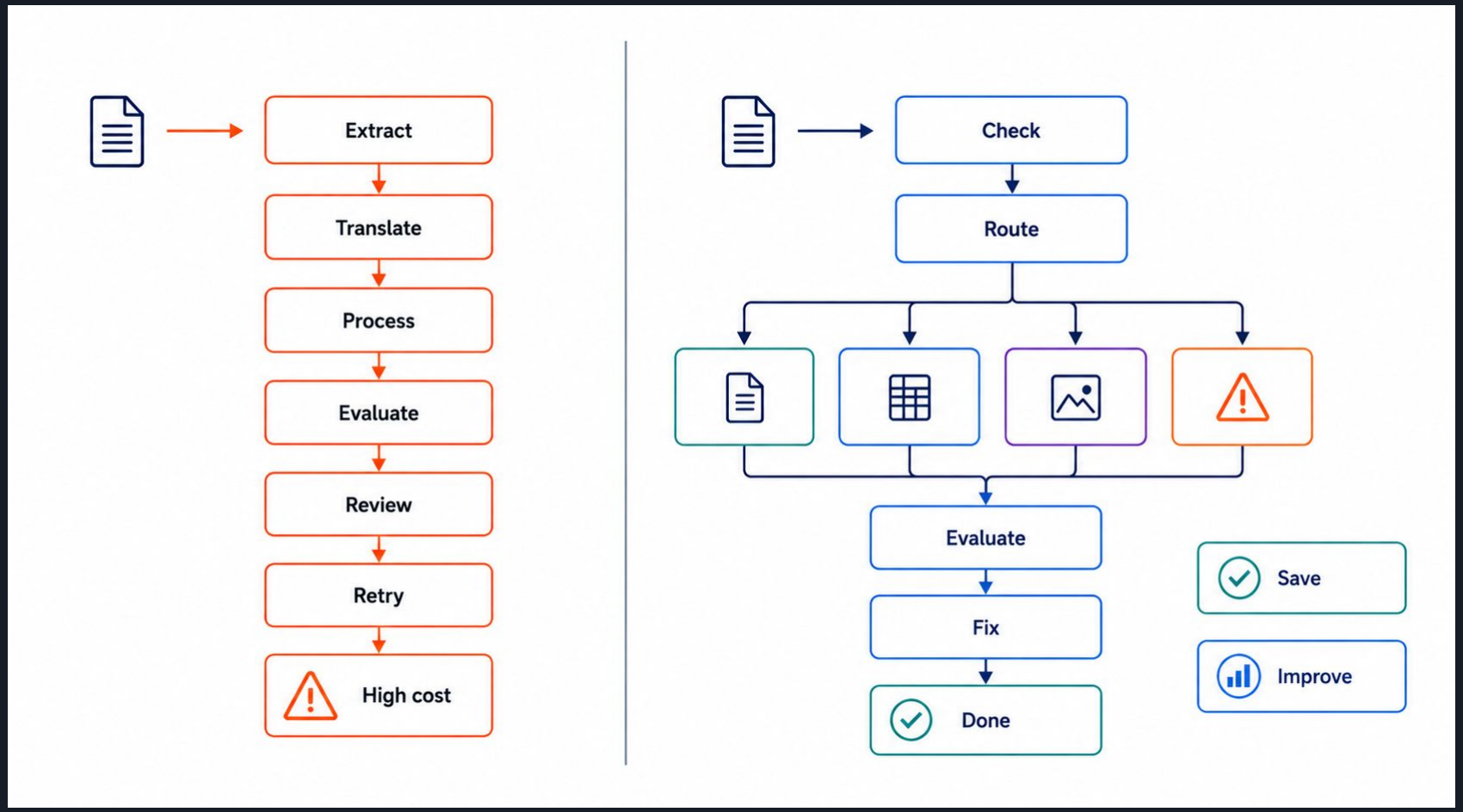
Thank you, let’s connect!

Andres Moreno

Guille Ojeda

Please complete the session survey in the mobile app

How architecture improves margin



4. Measure cost per accepted result

Audience checkpoint

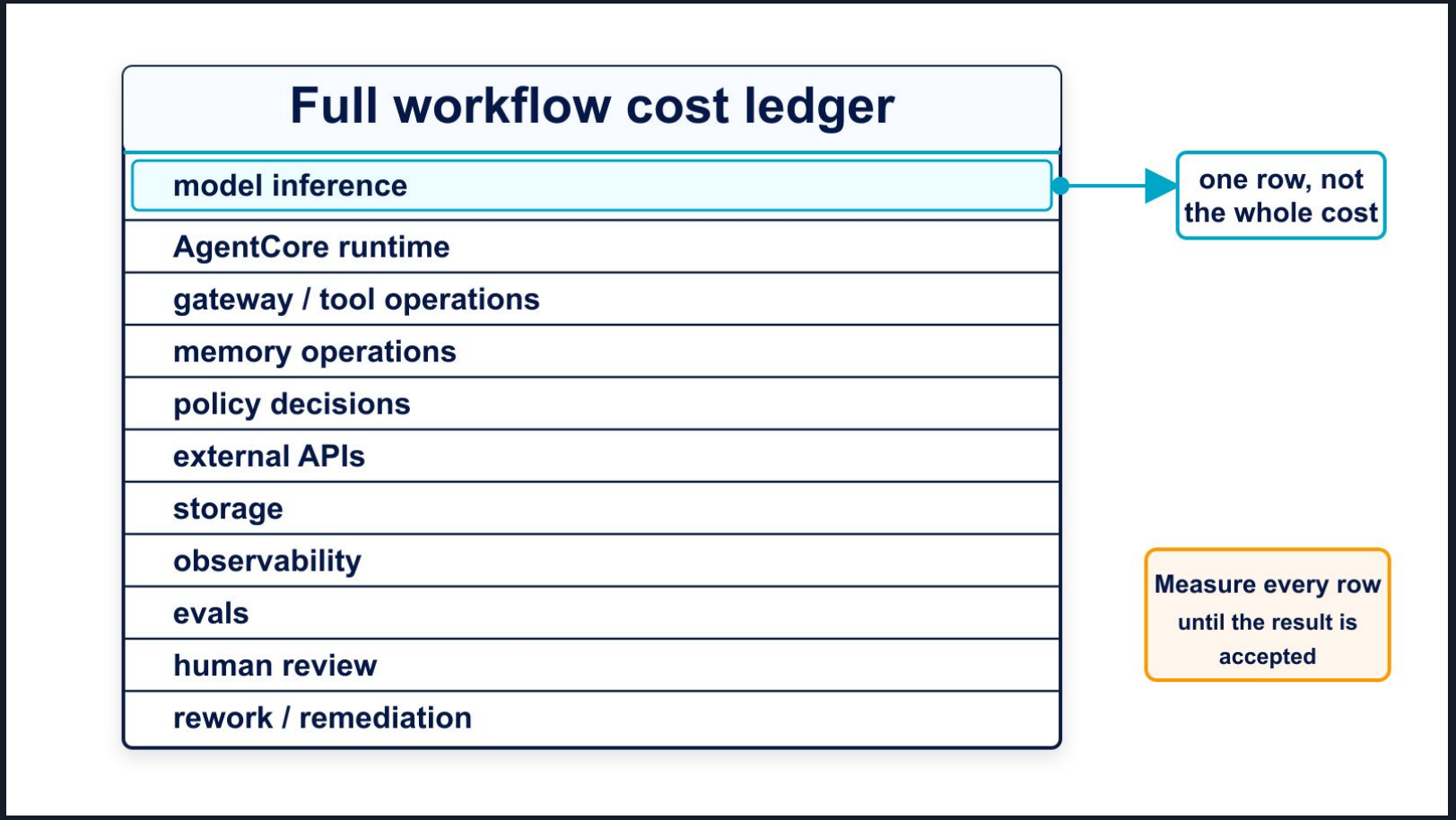
What do you think owns the economics?

Raise your hand if model calls are the biggest cost risk.

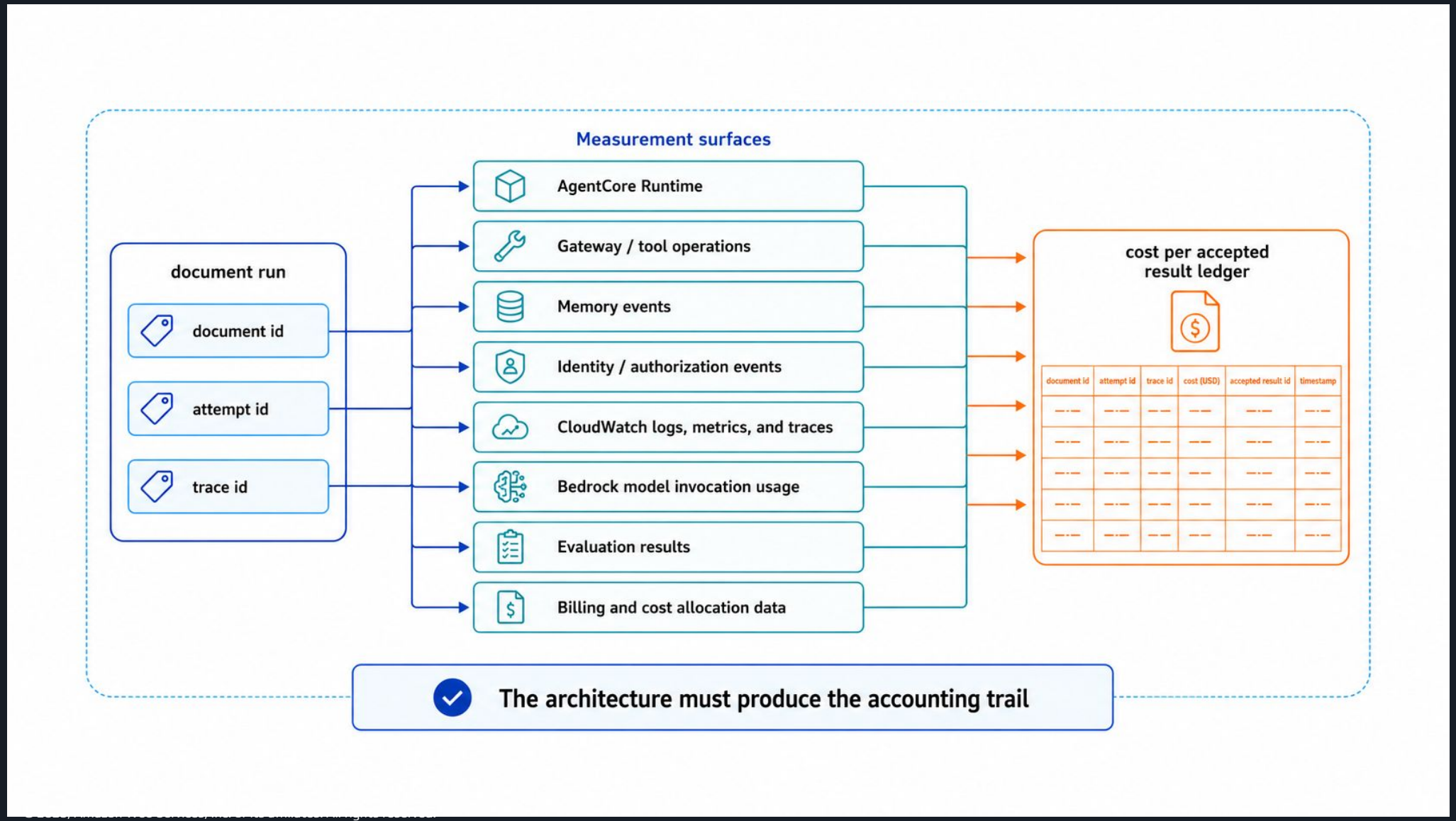
Raise your hand if retries or escalations are more likely to own it.

Now let us measure it.

What belongs in the ledger?

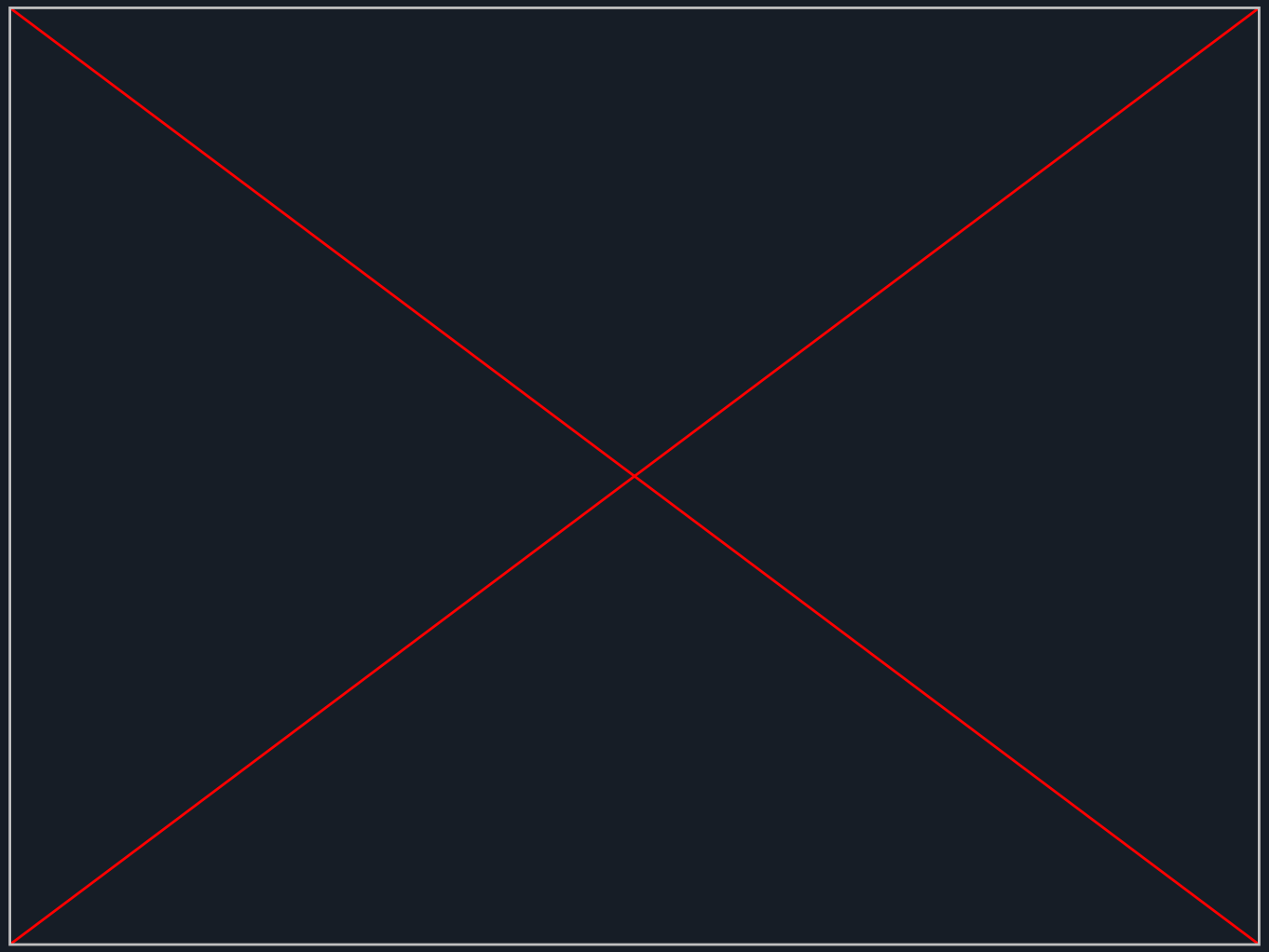


How do we instrument it?

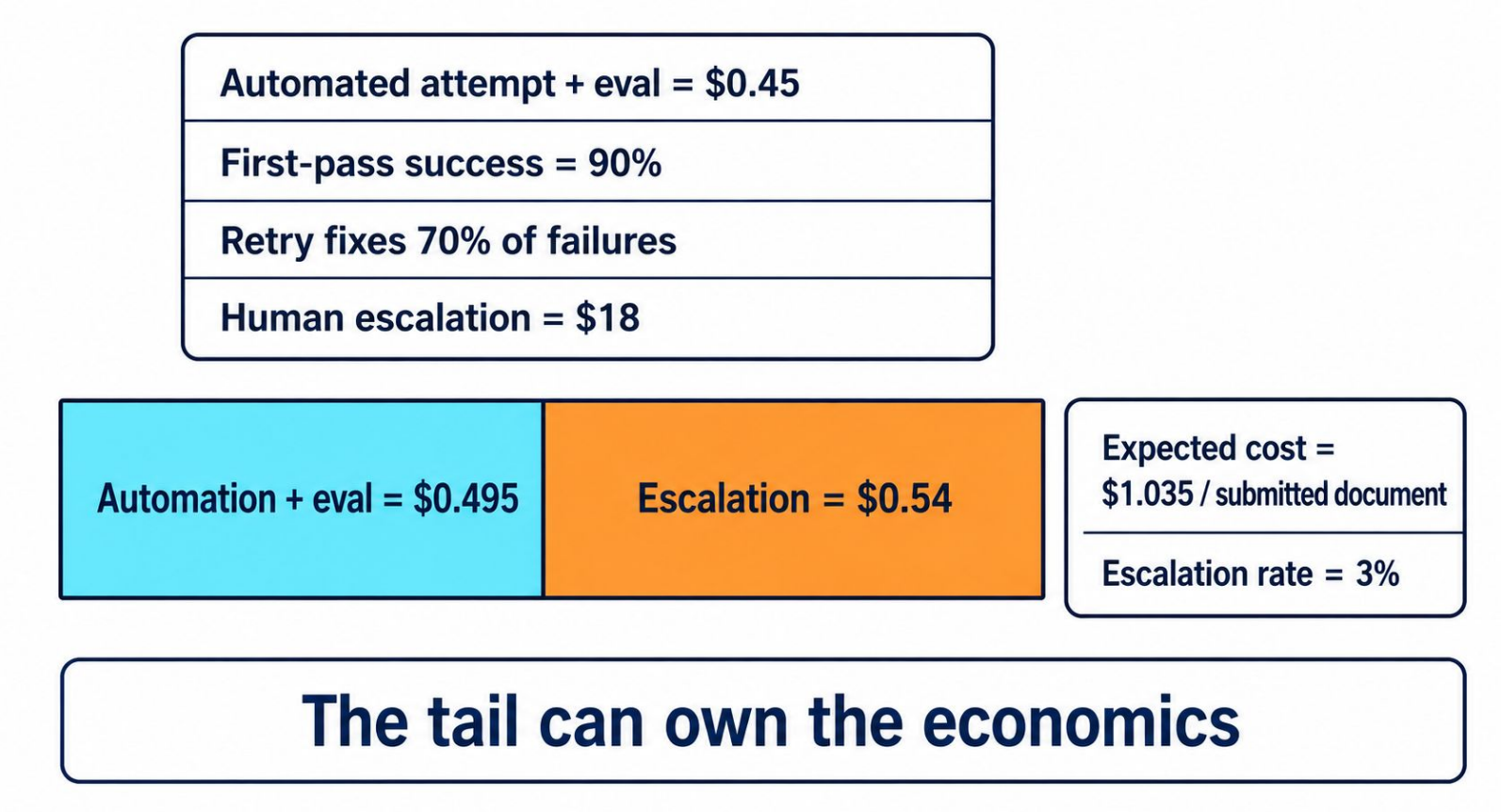


Recorded demo: show failure economics

Show what changes when attempts fail, retry, or escalate.

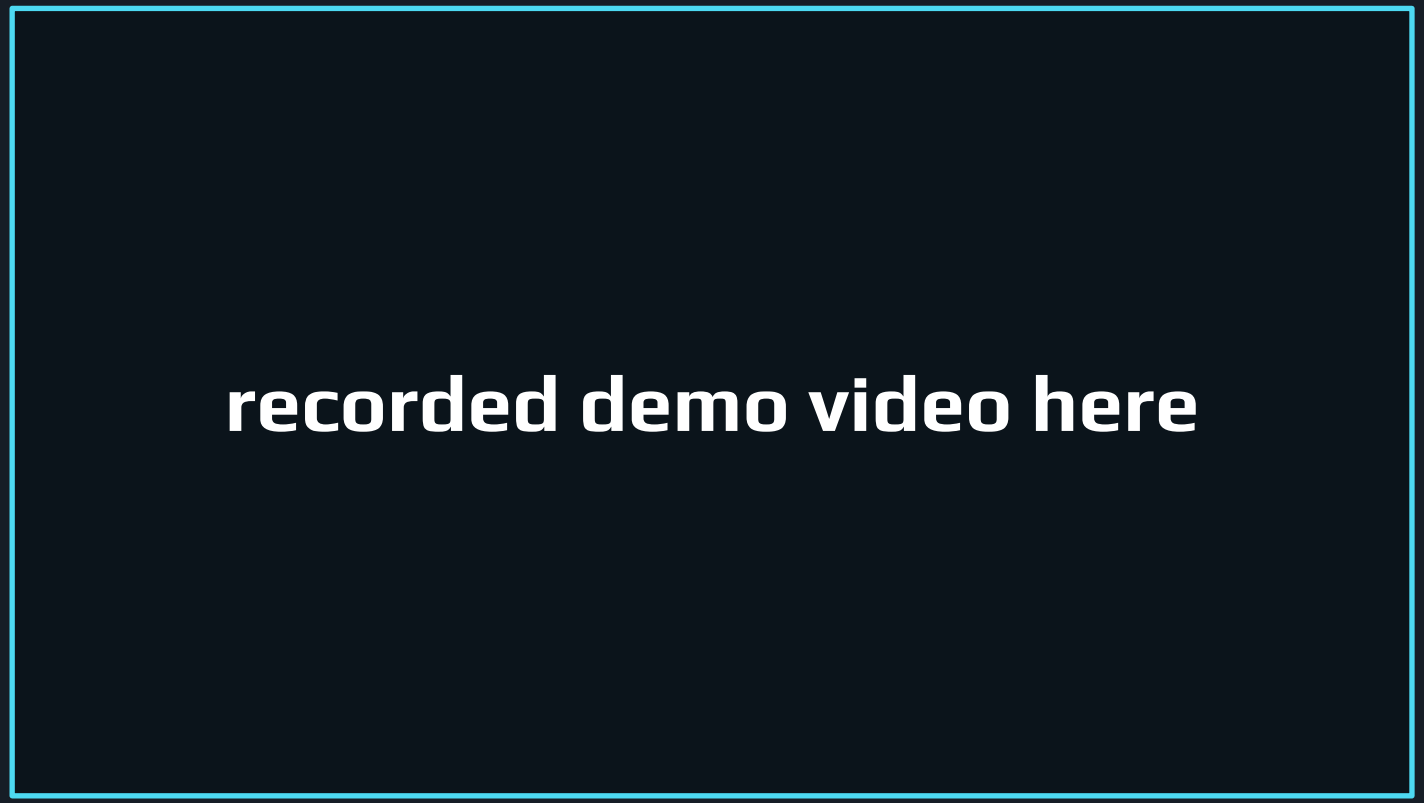


Why escalation can dominate cost



Recorded demo: showing two model paths

Cheapest model call can lose to the cheaper workflow.

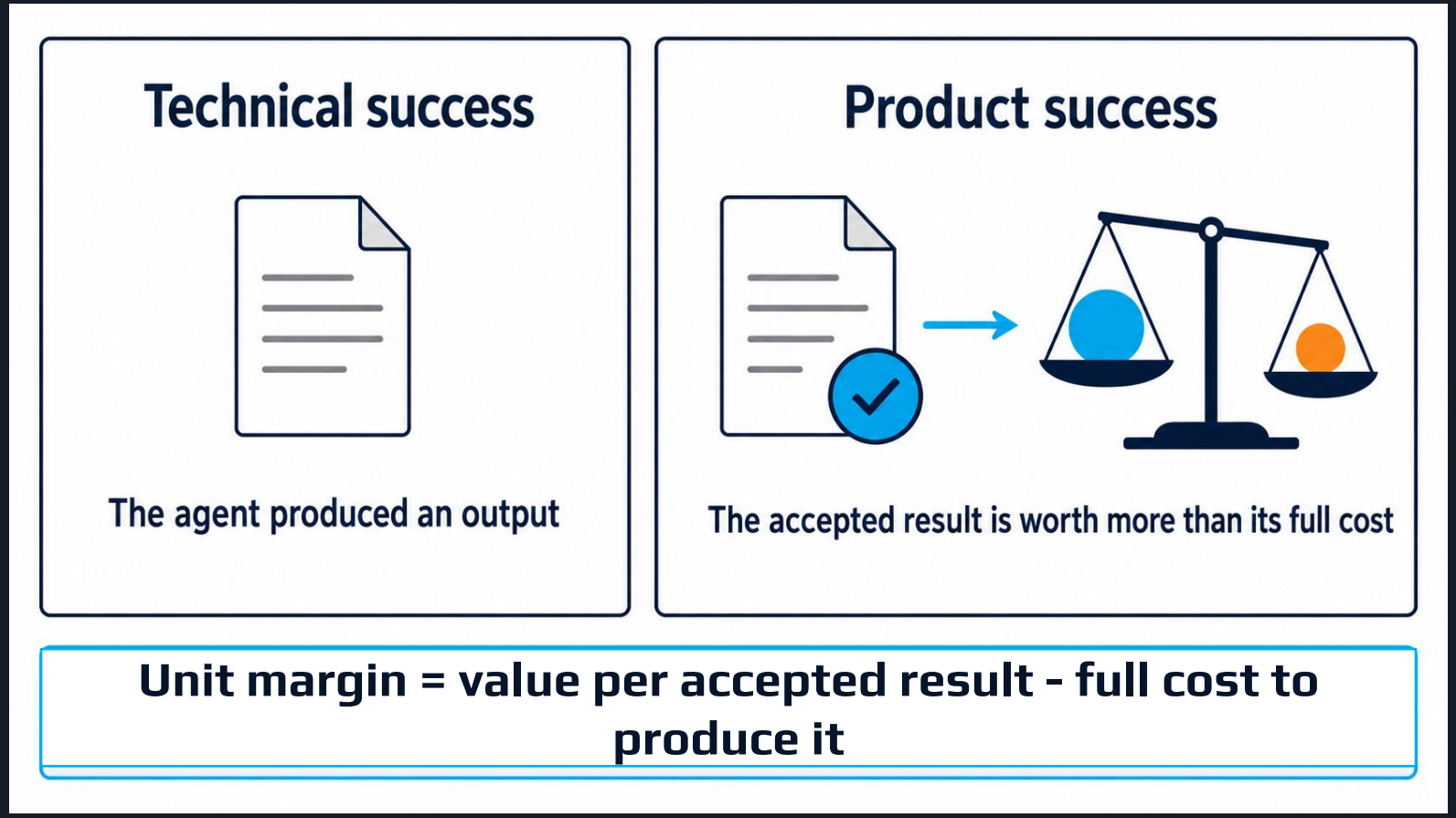


The solved workflow is the product

Cost has to be measured at the accepted-result level, not at the model-call level.

Value per accepted result > Full cost to produce it

Impressive is not the same as viable

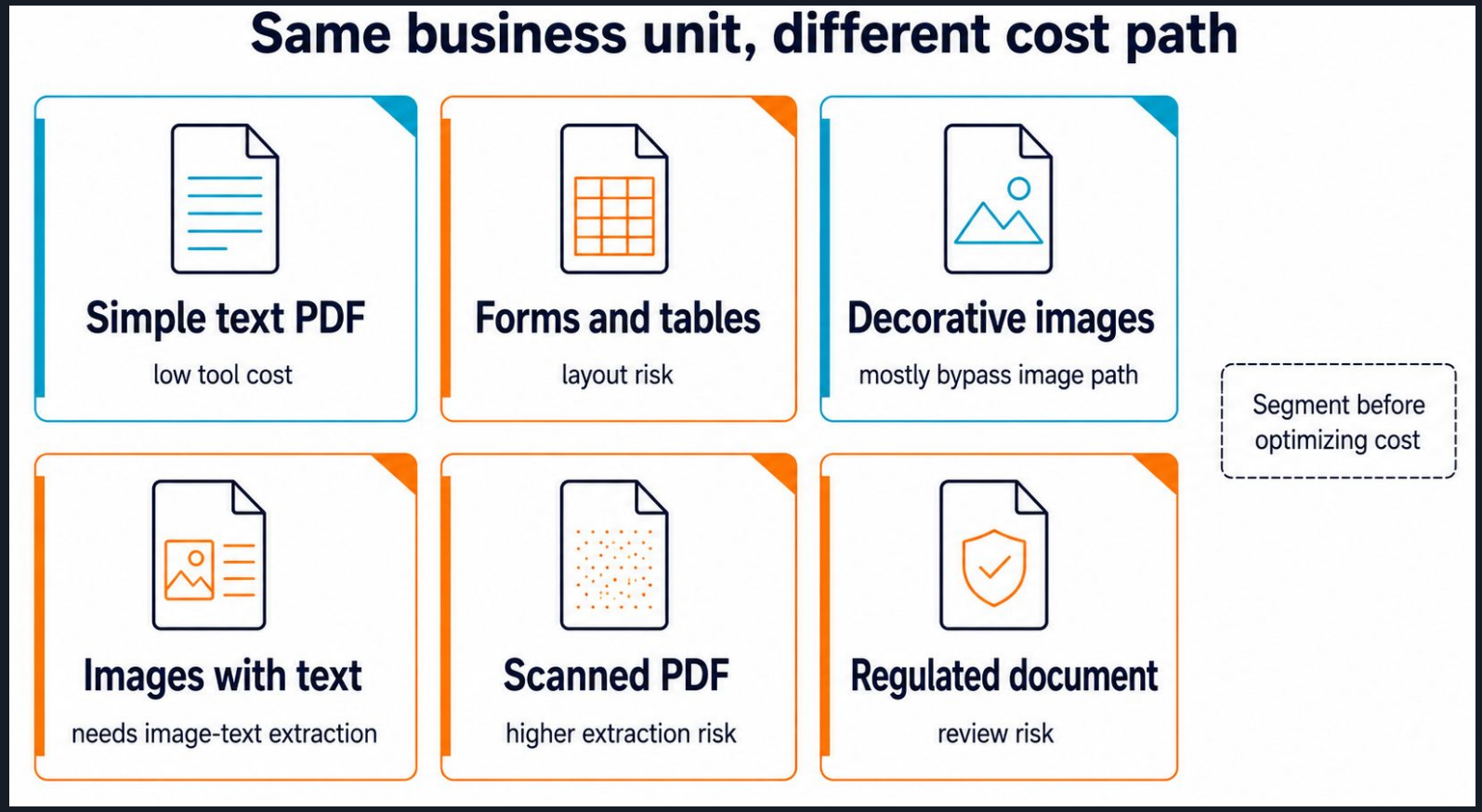


A token is not the business unit

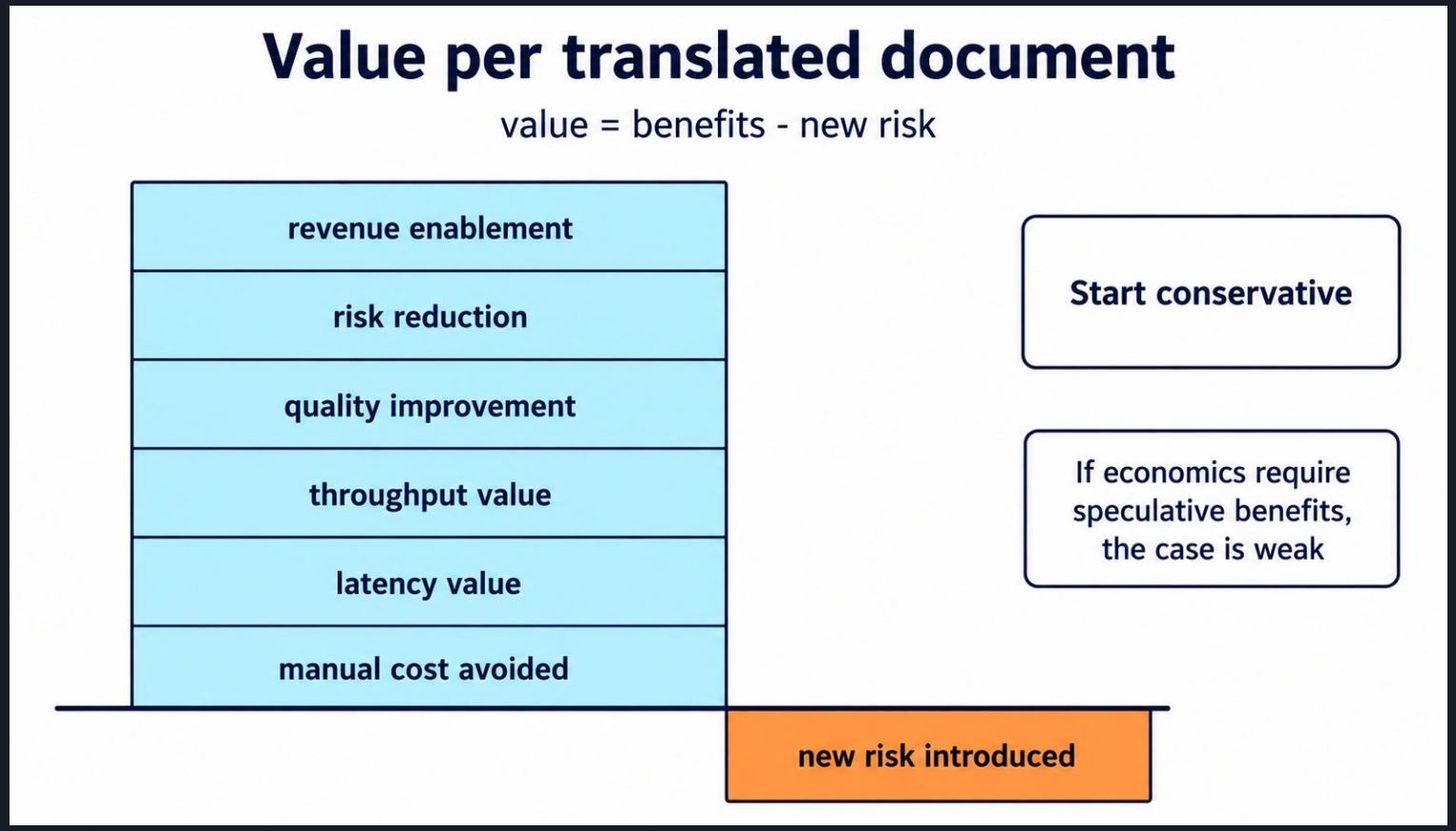
Unit	What it measures	Problem
Token	Model metering	Too low-level
Page	Complexity proxy	Not the completed task
Chunk	Engineering unit	Internal detail
Attempt	System activity	Attempts can fail
Accepted document	**Business outcome**	**Correct unit**

Tokens are metered.
Outcomes create value.

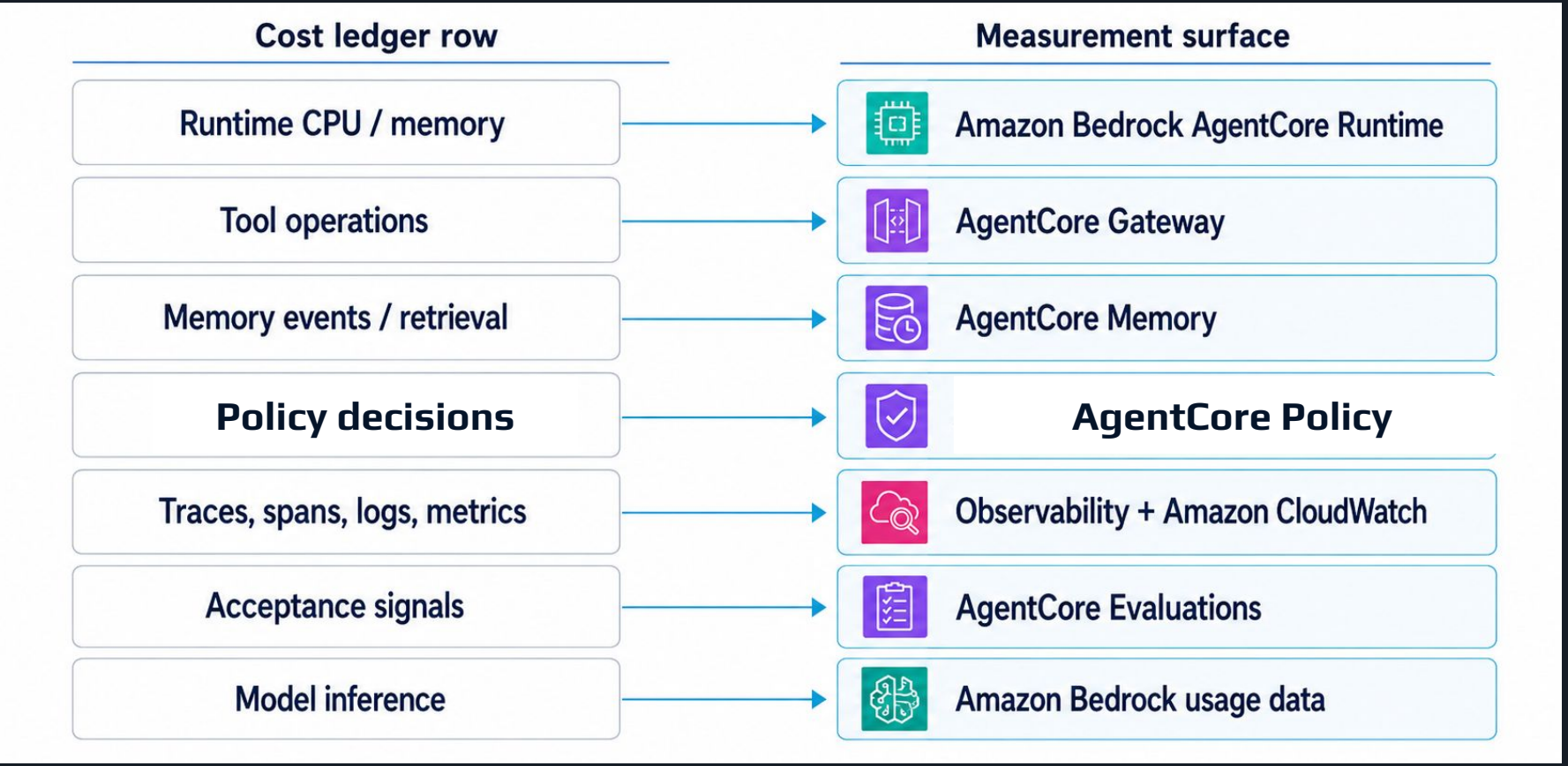
Segment the unit



Start with value, not model price



AgentCore measurement surfaces



Measurement contract before scaling

Use measured POC, production, or controlled dry-run data only.

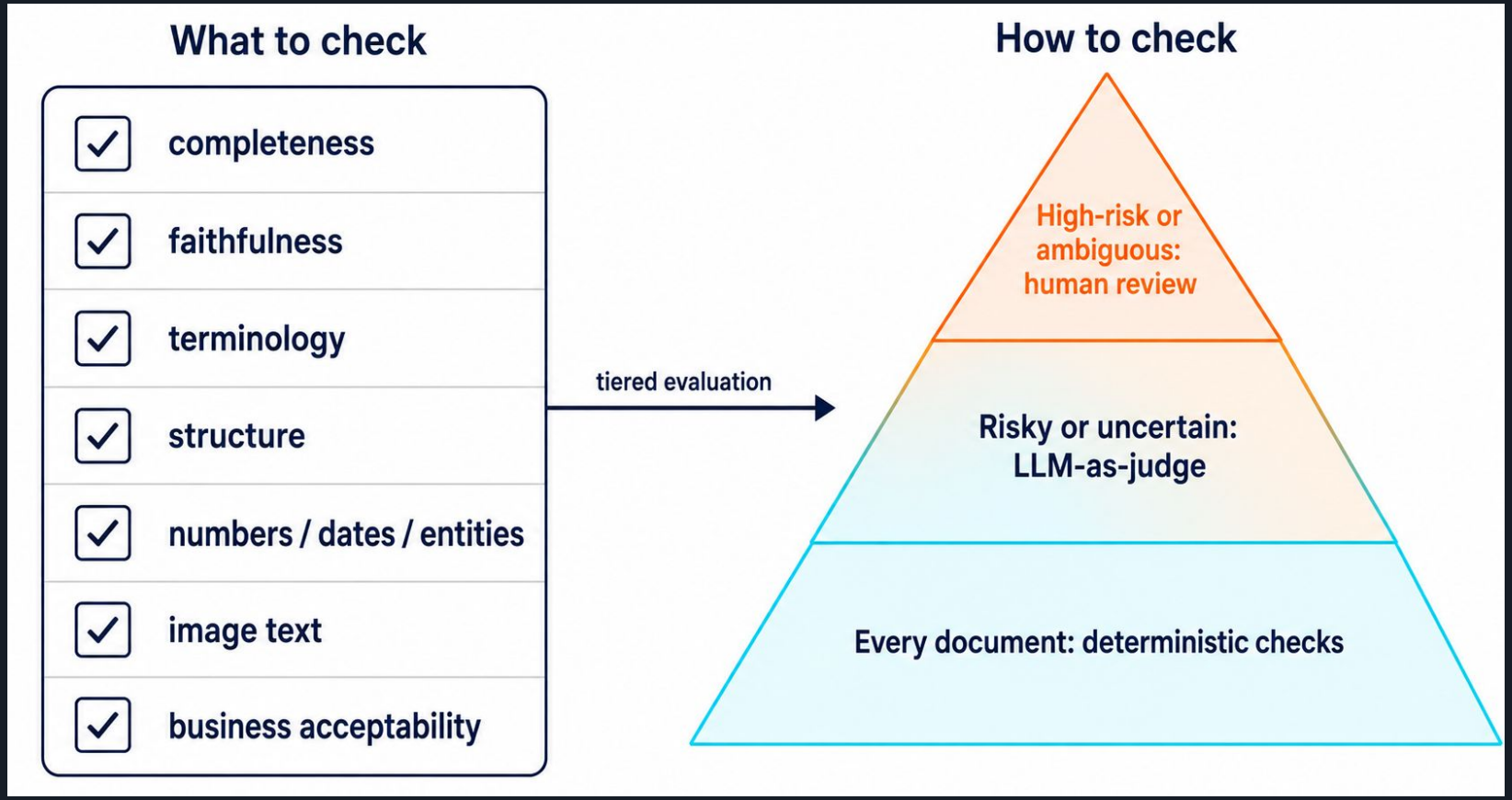
Capture per document run

document_id + run_id
workflow_version
eval_version
model_id / version
segment
attempt_count
eval + escalation outcome
total workflow cost

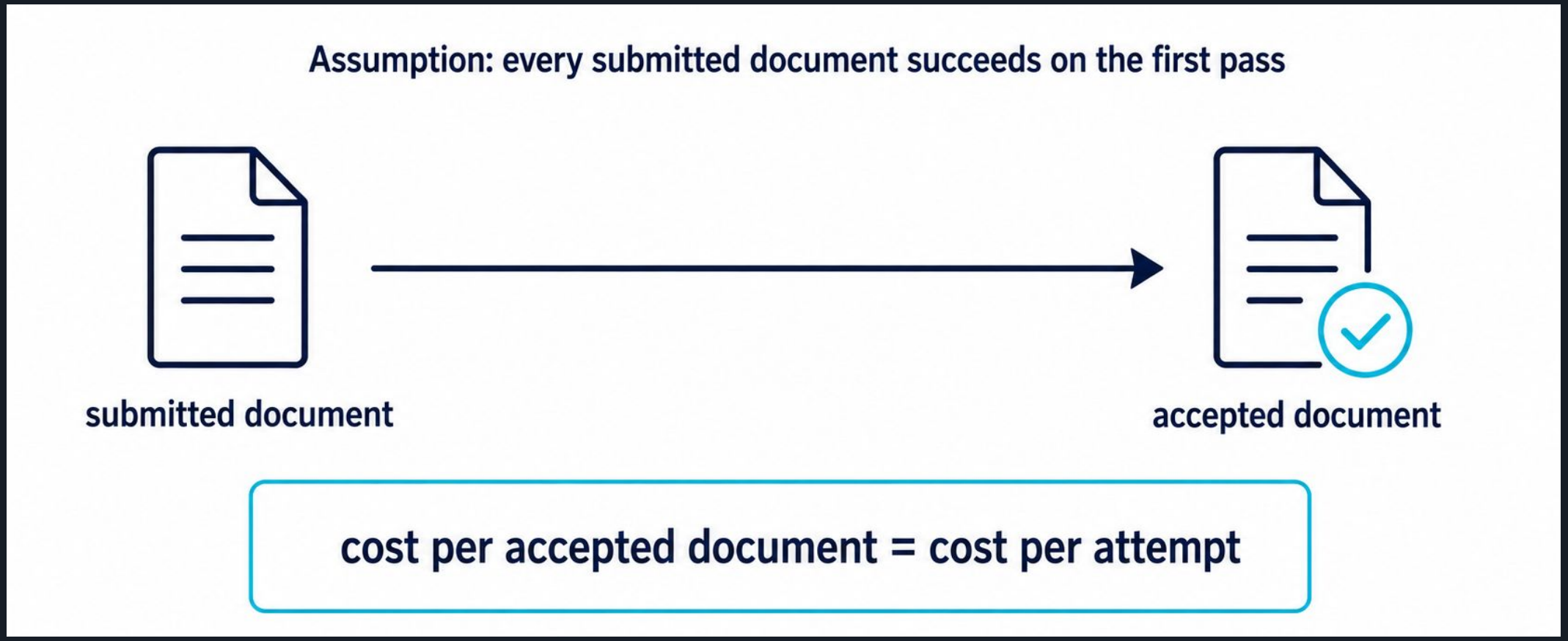
Report by segment

docs
avg attempts
first-pass accept rate
escalation rate
review cost
total workflow cost
cost / accepted result

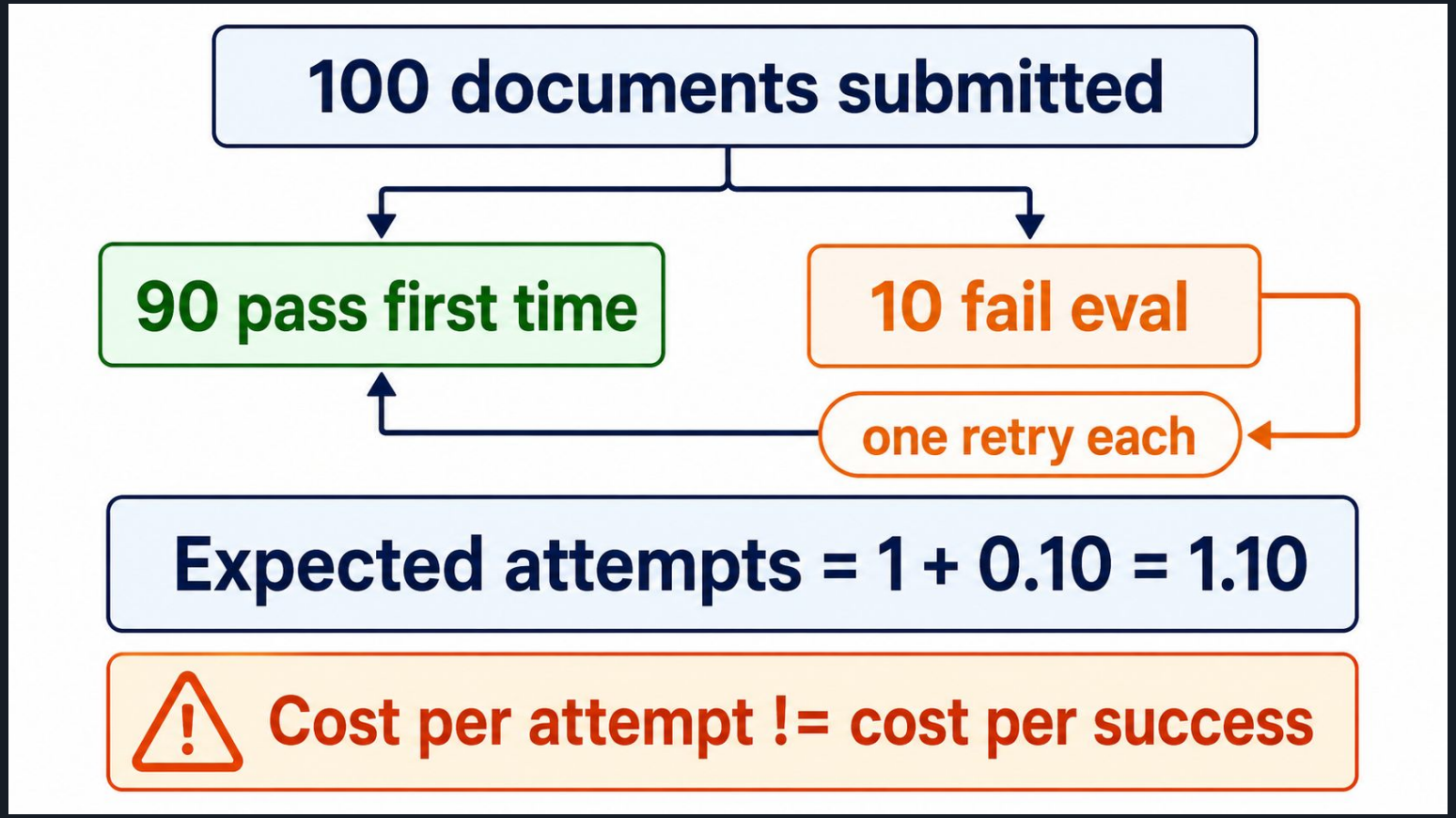
Translation eval strategy



Cost model v1: 100% success



Cost model v2: 90% first-pass success



A retry policy is a cost policy

Caught failure: retry / review cost

Uncaught failure: remediation / risk cost

Failure	Response
Timeout	Retry
Missing extraction	Reprocess
Low faithfulness	Stronger model or prompt
Broken layout	Recompose
Image text missed	Route to image path
High-risk ambiguity	Human review
Bad output accepted	Remediation / risk cost

The goal is not to build the cheapest agent.

Build agents whose accepted results are worth more than their cost

Value per accepted result > Full cost to produce it

About this template

This presentation template is built for 2026 AWS Summits. It will serve as our main template for all non-keynote event presentations. This will help bring brand consistency across our event.

WHAT'S INSIDE:

Dark slide options

Cover and divider slides

Content slides

Examples of application

Designing for events

When creating your presentation, consider how it will be viewed. If the PowerPoint will be viewed by many, in a presentation setting, keep slide content to a minimum and make sure font size is large enough to be read from the furthest reasonable distance.

Your production partner should have a minimum font size recommendation based on accessibility and screen specs/event space and audience seating.

If the PowerPoint will be viewed on individual devices, you can include more content at a smaller size per slide. We still recommend keeping content as minimal as possible. Less is more.

CONSIDERATIONS FOR DESIGN:

Think big

Distance, size, and lighting affect legibility

Reduce for quick understanding

A slide may only be on screen for seconds – make sure slide content can be understood in that time

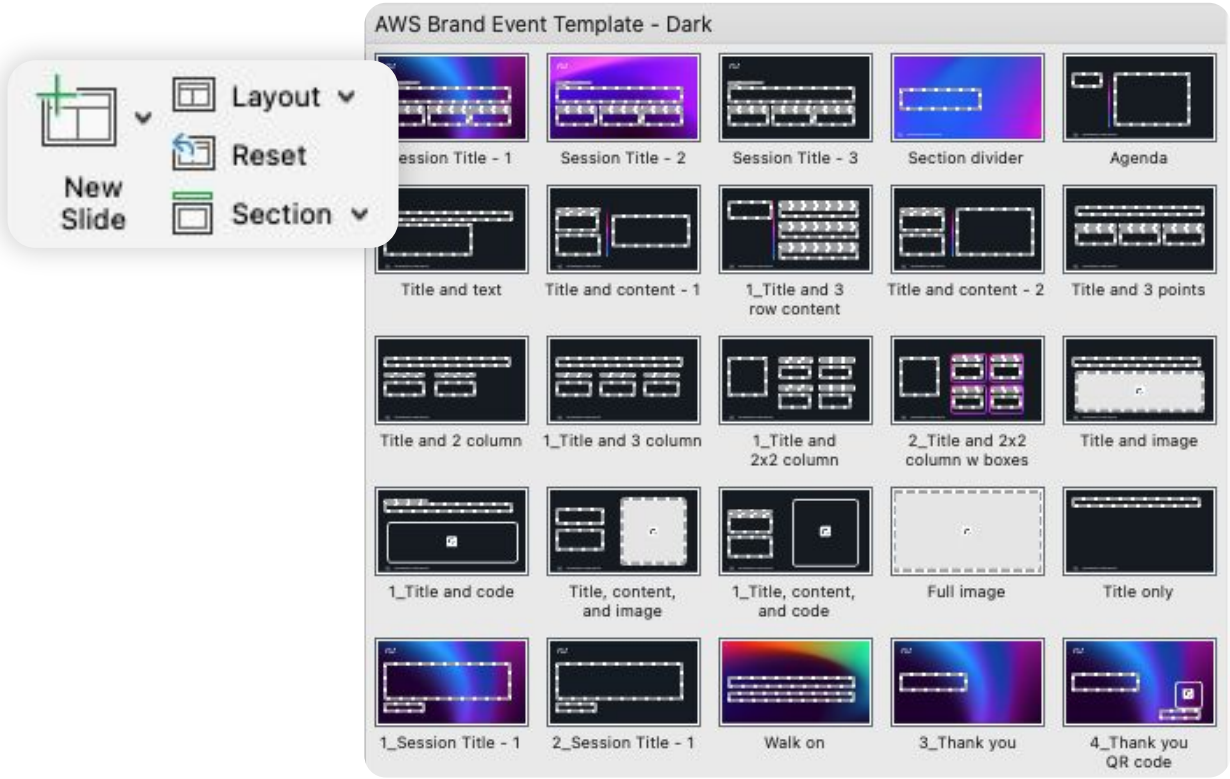
Less content, more slides

Don’t be afraid to break your content out across more slides – slide count doesn’t matter, comprehension does

How to add new content

Follow these steps to use this template to create a new slide deck:

1. Enter your content on the blank slides and delete any unneeded or hidden slides.
2. If you want to add more slides, on the Insert tab, select the little arrow under New Slide, and then select your desired layout from the options displayed (shown right). Fill out your slide content using the preformatted context placeholders.
3. End the deck with the Thank You slide and then the Survey Reminder slide layouts.



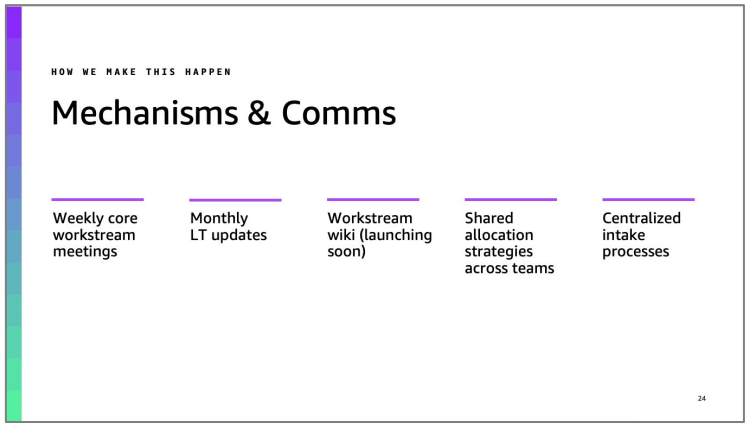
How to use existing slide content

There are two ways we recommend bringing existing content into this deck.

Bring content in individually

Similar to how you would bring new content in from an outline, find a template slide that works for your content, and cut and paste individual sections of copy into the empty content placeholders.

This will make for less restyling and better consistency within the template.



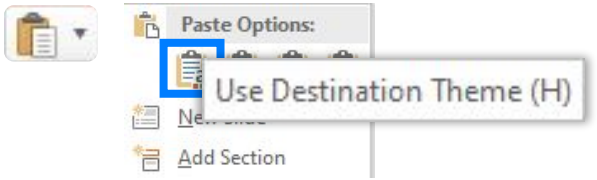
Previously designed slide in different template

Newly designed slide with content migrated over

How to use existing slide content

Copy existing slides into the deck

1. Open the deck you've already created and select the slides you want to move in the left-hand panel. If you want to move all the slides of a deck over, you can select all slides by selecting any slide on the left-hand panel and pressing (Ctrl + A) and copy them (Ctrl + C).
2. In the template file, click in the left-hand pane, paste in your slides (Ctrl + P) and click Use Destination Theme (use this same method to transfer individual slides).



3. For each slide in your deck, if needed, apply the appropriate layout (Home tab > Layout > Select appropriate Layout).
4. Some content won't fit like a glove, and that's okay! First try to reset the slide (Home tab > Reset)
5. If formatting still seems off you might need to copy-paste content into the correct placeholders, separate content by adding additional textboxes
6. Don't forget to save your file with a new file name.

Troubleshooting Tips

When restyling a slide, there may be extra empty text boxes from the template that don't match your current layout, you can delete those.

If you're copying a slide with a complex diagram, and are experiencing trouble restyling, you may want to cut and paste the diagram separately into a blank or title only template slide and select "Keep Source Formatting" then manually restyle elements to fit within brand style.

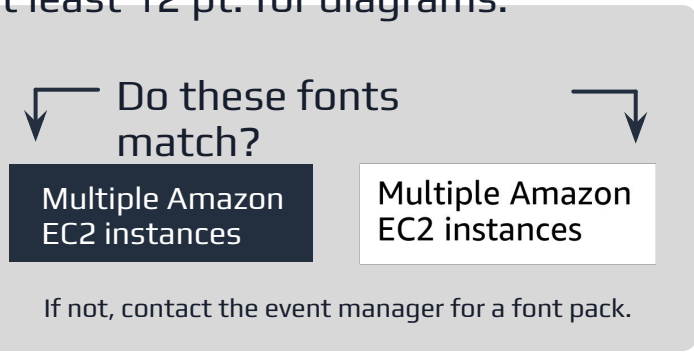
When restyling between dark and light mode slide designs, make sure all slides are in dark or light mode (whichever you are using) then check that all text boxes and elements have properly changed color, select all elements on the slide to make sure none are hidden.

Font guidance

Use AWS brand fonts, Amazon Ember Display.

Do not underline text that is not a hyperlink.

Use font sizes that are at least 20 pt. for slide text and at least 12 pt. for diagrams.



Eyebrow
Amazon Ember Mono

LOREM IPSUM

Headline
Amazon Ember Display, Regular

Dolore eu fugiat nulla aute irureol consequat

Subhead
Amazon Ember Display, Regular

Dolore eu fugiat nulla aute irureol consequat tu

Body copy
Amazon Ember Display, Regular

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmodtempor incididunt ut labore et dolore magna aliqua.

Code
Lucida Console

```
Lorem ipsum dolor sit amet, consectetur adipiscing elit,
sed do eiusmodtempor incididunt ut labore et dolore
```

See this slide example for typography in use

Accessibility

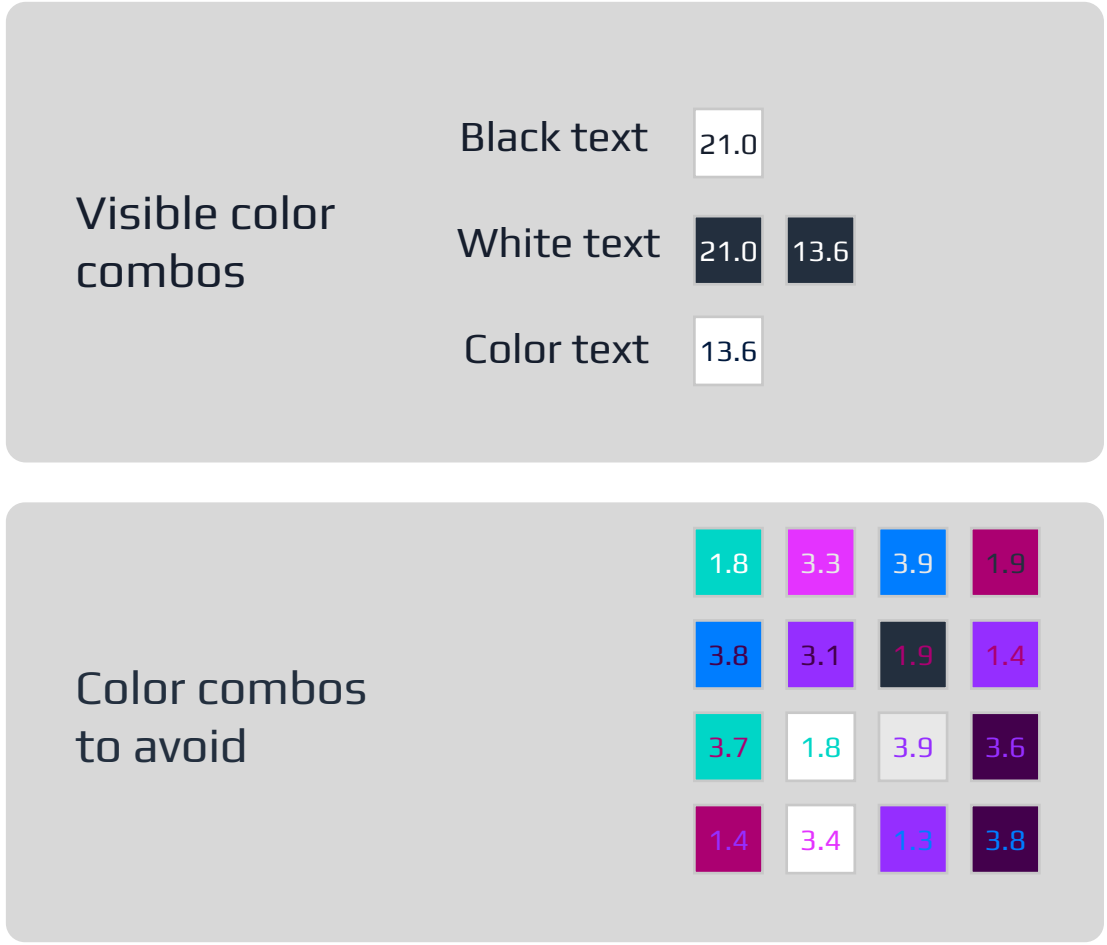
For visibility, use a minimum 20 pt for slide text and 12 pt for diagrams. For lines, use a minimum 1.5 pt line weight.

Color also affects visibility. Text (or lines) and the background should have a contrast ratio of at least 4.5:1. Generally, white or black on a strongly contrasting background is the best choice. Avoid using colored text in your presentations.

Add descriptive alt text to images, charts, and other visual assets. Use the Accessibility Checker under Review > Check Accessibility.

To make it easier for screen readers to identify a layered image, like a picture with callout lines, group them into a single image.

Read more on PowerPoint accessibility from Microsoft.



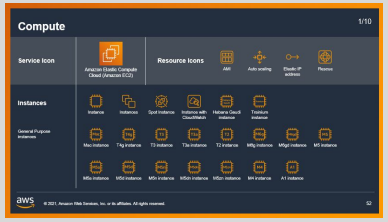
Brand elements

Only use logos, icons, photos, screenshots, and illustrations that your organization has permission (a license) from the copyright owner to use (this includes open source service logos).

Keep slide diagrams as simple as possible; split content over several slides rather than forcing it all onto one.

Use simple, subtle animations as needed to direct audience attention and clarify relationships. Avoid dramatic transitions or animations (such as “Random Bars,” “Dissolve,” or “Spin”).

AI generation: Until further notice, you should not use any AI-generated imagery for AWS marketing materials. As of July 2025, we lack the tools and governance to produce reliably on brand imagery at scale. Please check the AWS Brand Portal for more information.



Download AWS architecture icons from the AWS website

Download illustrations, fonts, and the AWS logo from the brand portal

Less is more. Use components with high judgement.

For licensed stock photography request access to Adobe Stock

Content development

Customer logos and references

For any mentions of AWS customers, you must ensure they are a referenceable customer and that the use case you want to use is approved. You can check on permissions for references using Reference Finder. If the customer is not in Reference Finder, you cannot include that customer or use case.

This database also contains customer logo permissions and AWS must have permission to use the logo in order to display it in a presentation.

AWS offering names

Make sure that all AWS service names are referred to correctly in the presentation by checking the offering names list or the AWS website. For example, you should not refer to just S3; Amazon S3 is the correct use.

If you cannot find a product or service on the website, assume it has not launched and that you cannot include it in your presentation. Remember that products or services may not be referenced in a presentation until they have launched publicly.

Copyrighted content

Review all of the images, videos, logos, and music in your presentation to ensure that you have obtained a corporate license or have permission for use. This includes logos and images for open source software, customers, competitors, and other products.

Content from movies or TV shows or that includes references to popular characters and famous figures carries a high level of copyright and trademark infringement risk and should be removed from your presentation.

As a speaker, you are responsible for the content you present. For additional resources on content and presentation development, please visit your speaker portal or the Speaker Programs Highspot page.

Additional Brand resources

AWS Brand Portal

Photography guidance and AWS photo selects

Illustration guidance and downloads

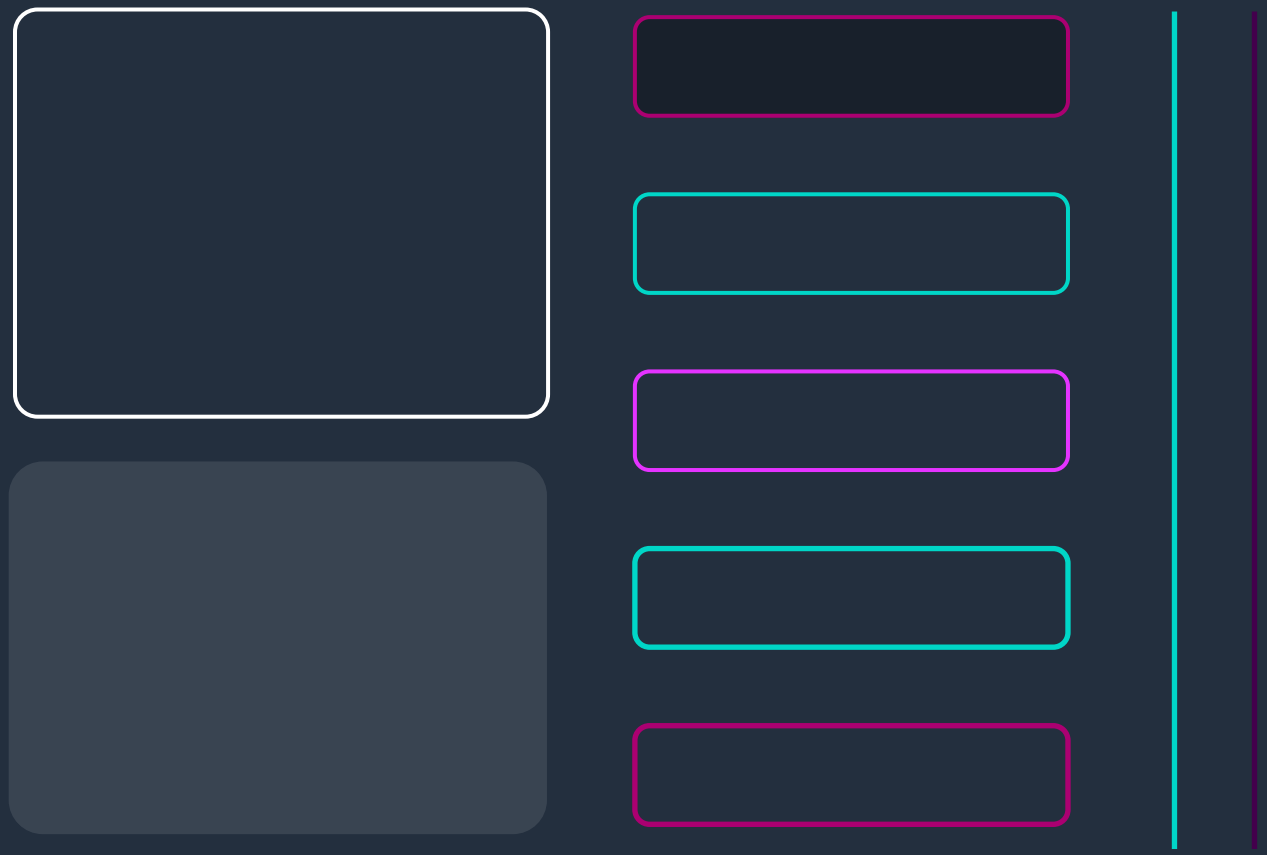
Architecture iconography guidance and downloads

Additional data visualization guidance and downloads

Voice & tone guidance and writing tips

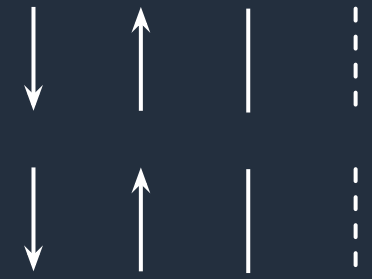
AWS Brand support hub

Kit of parts

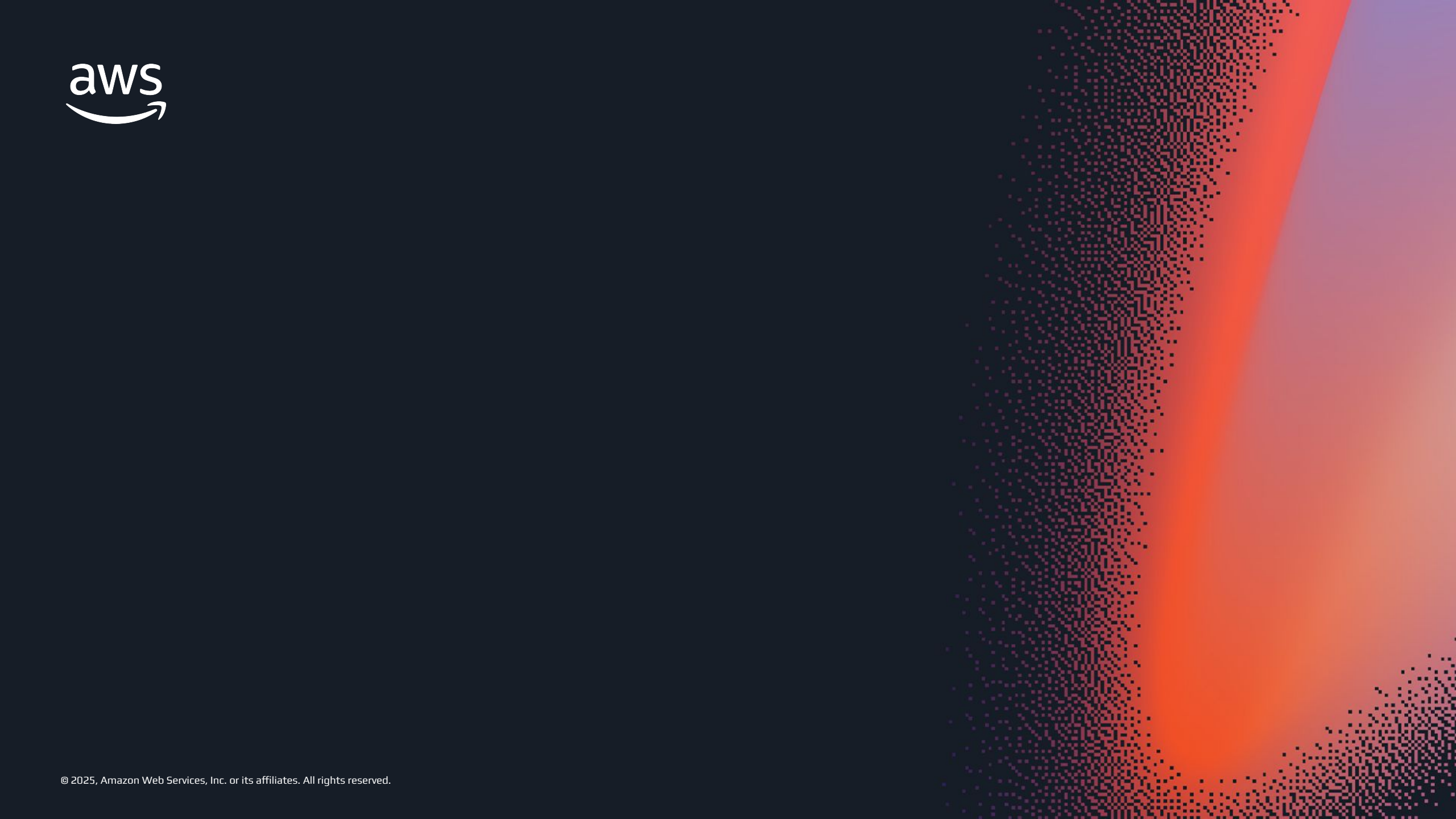


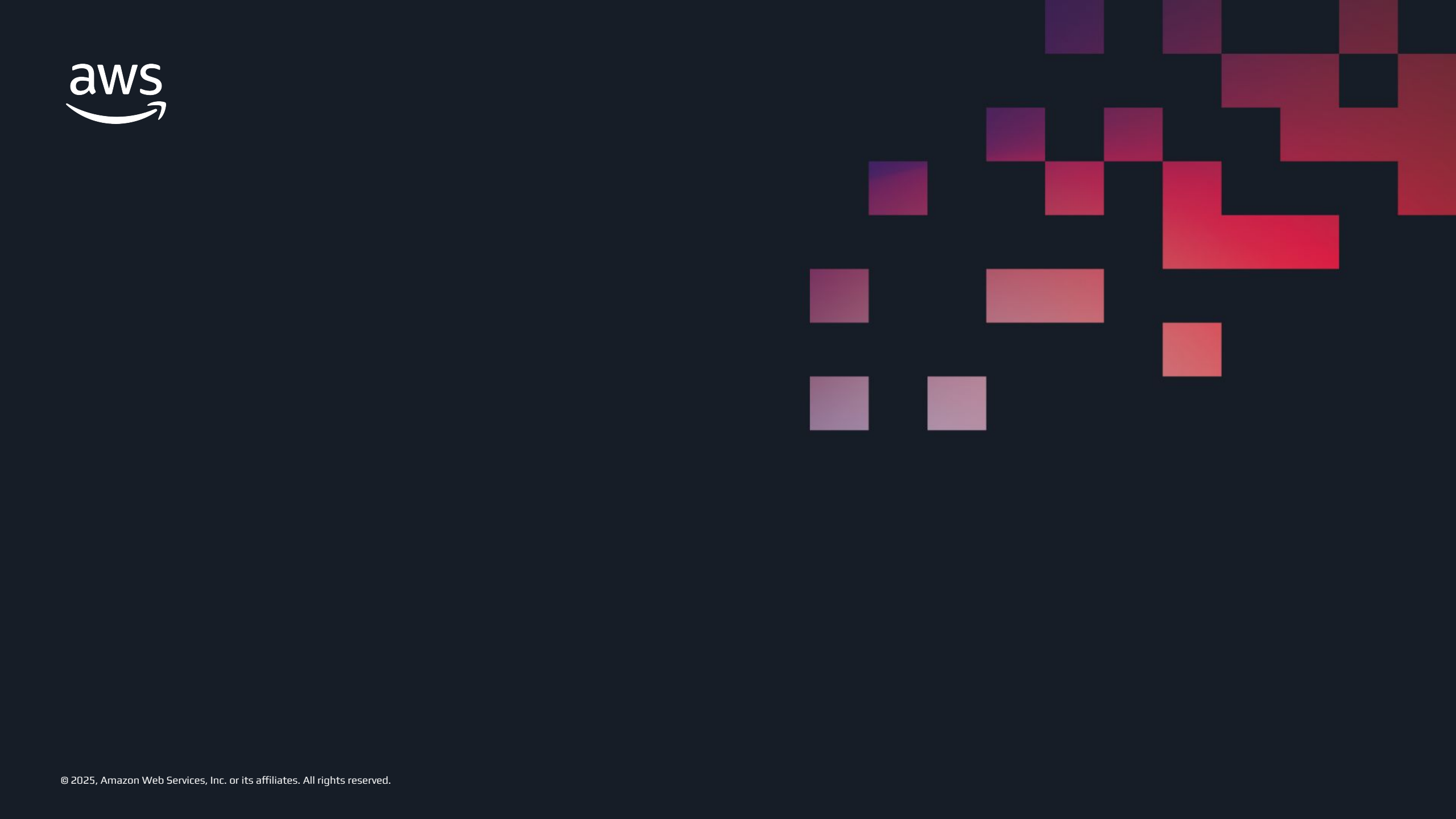
CORNER ARROWS

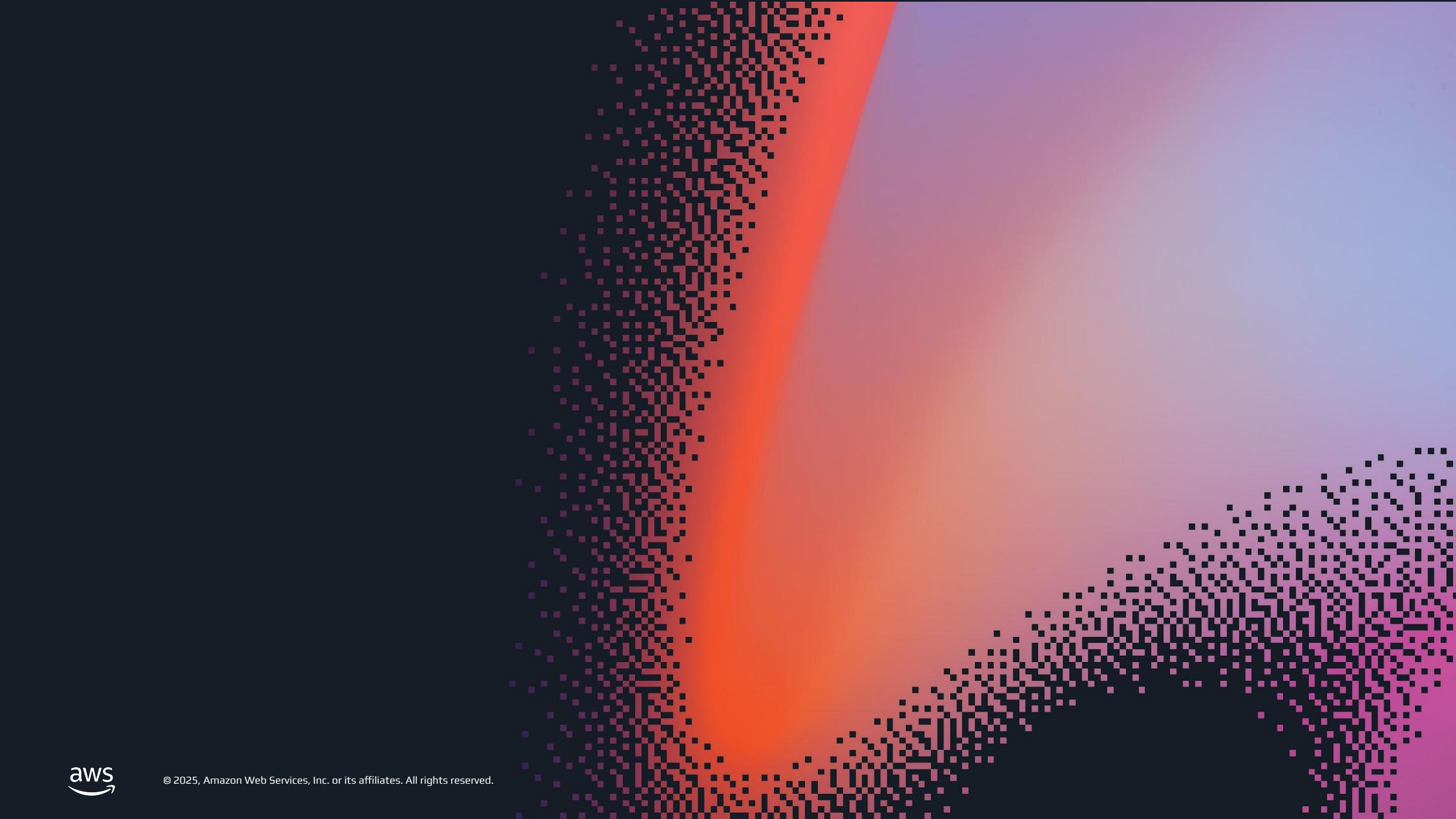
VERTICAL CONNECTORS

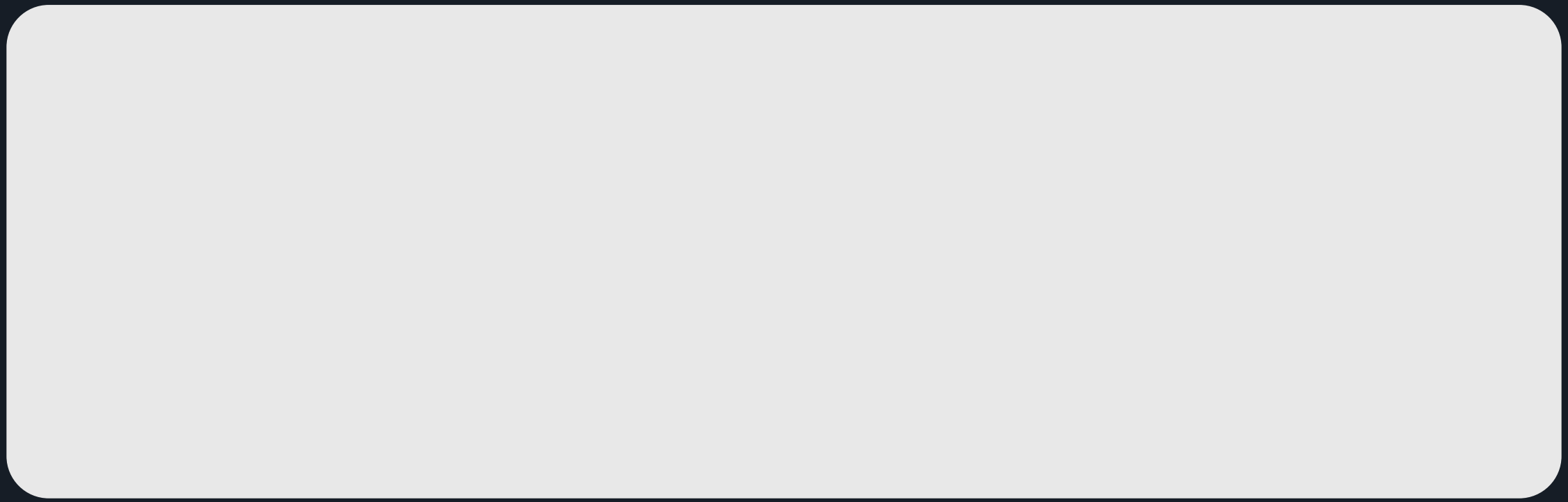


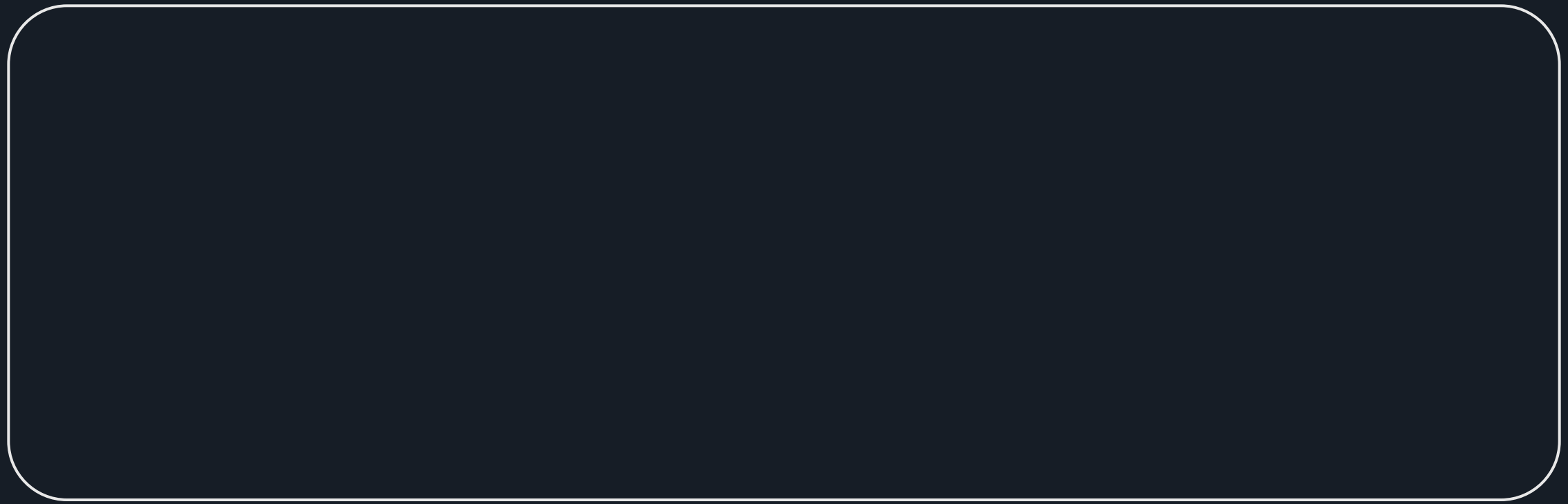
HORIZONTAL CONNECTORS

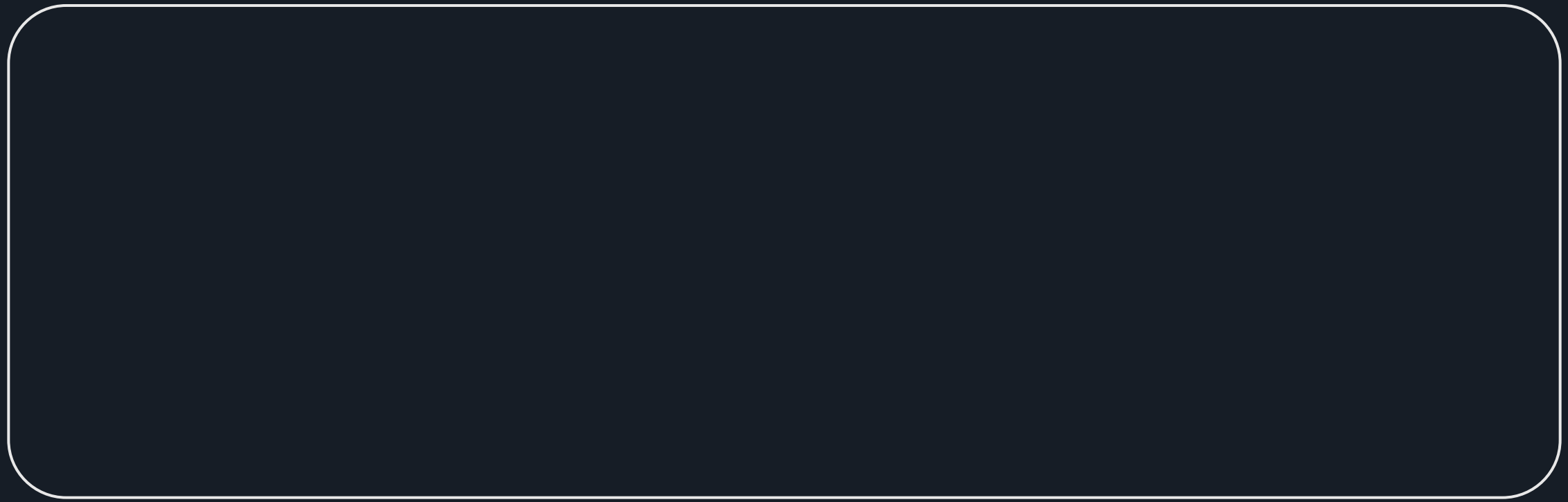


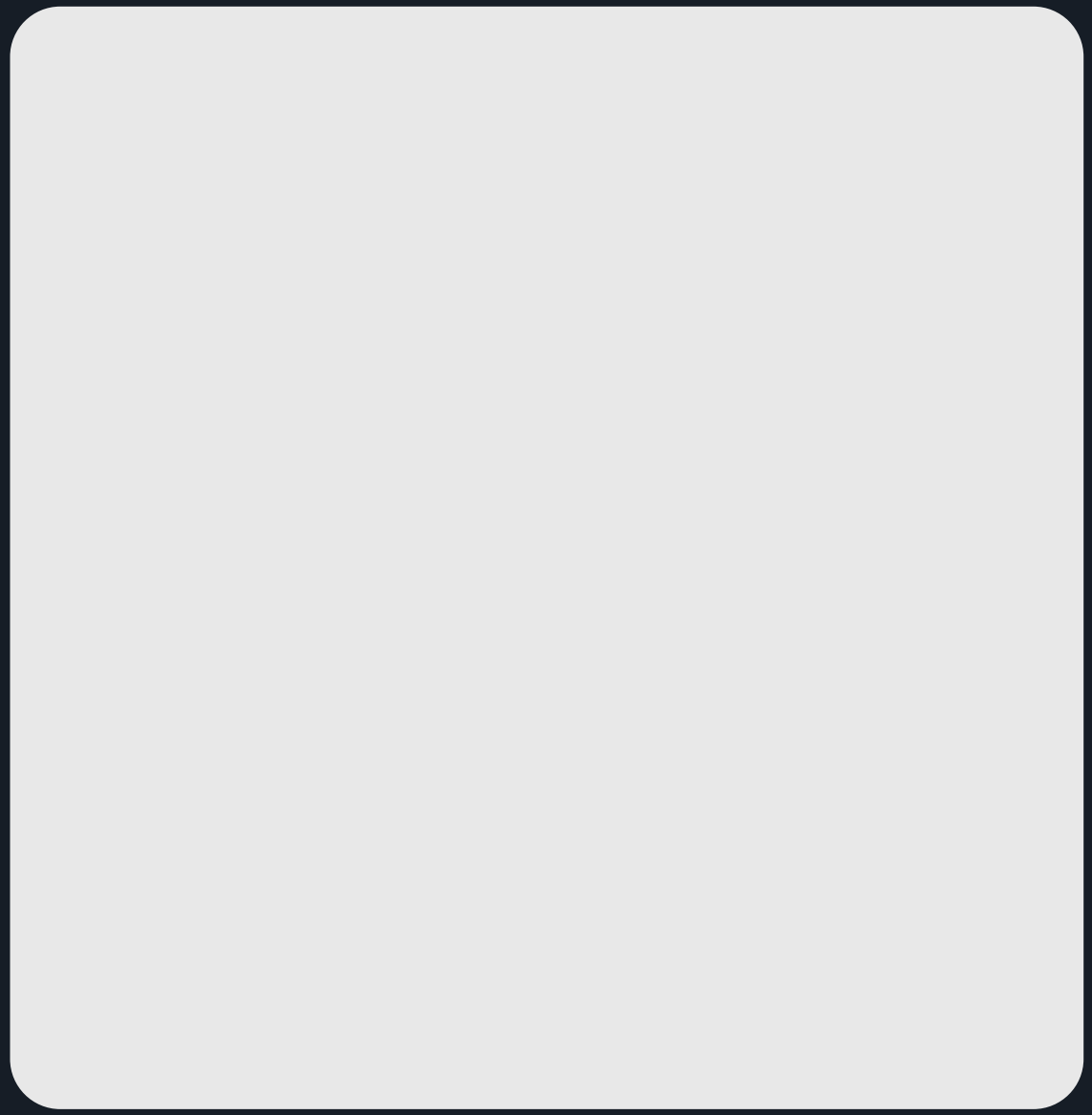


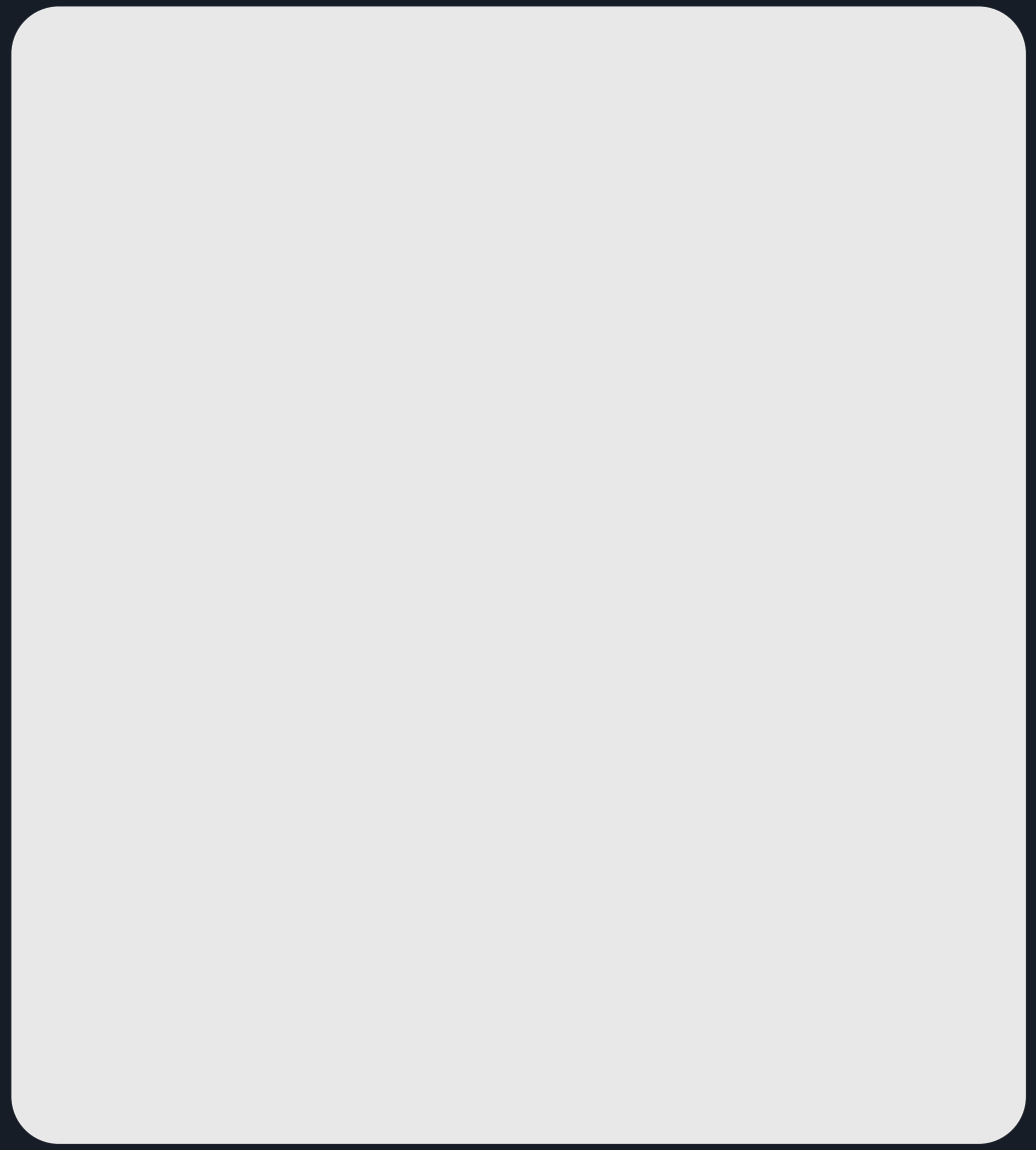


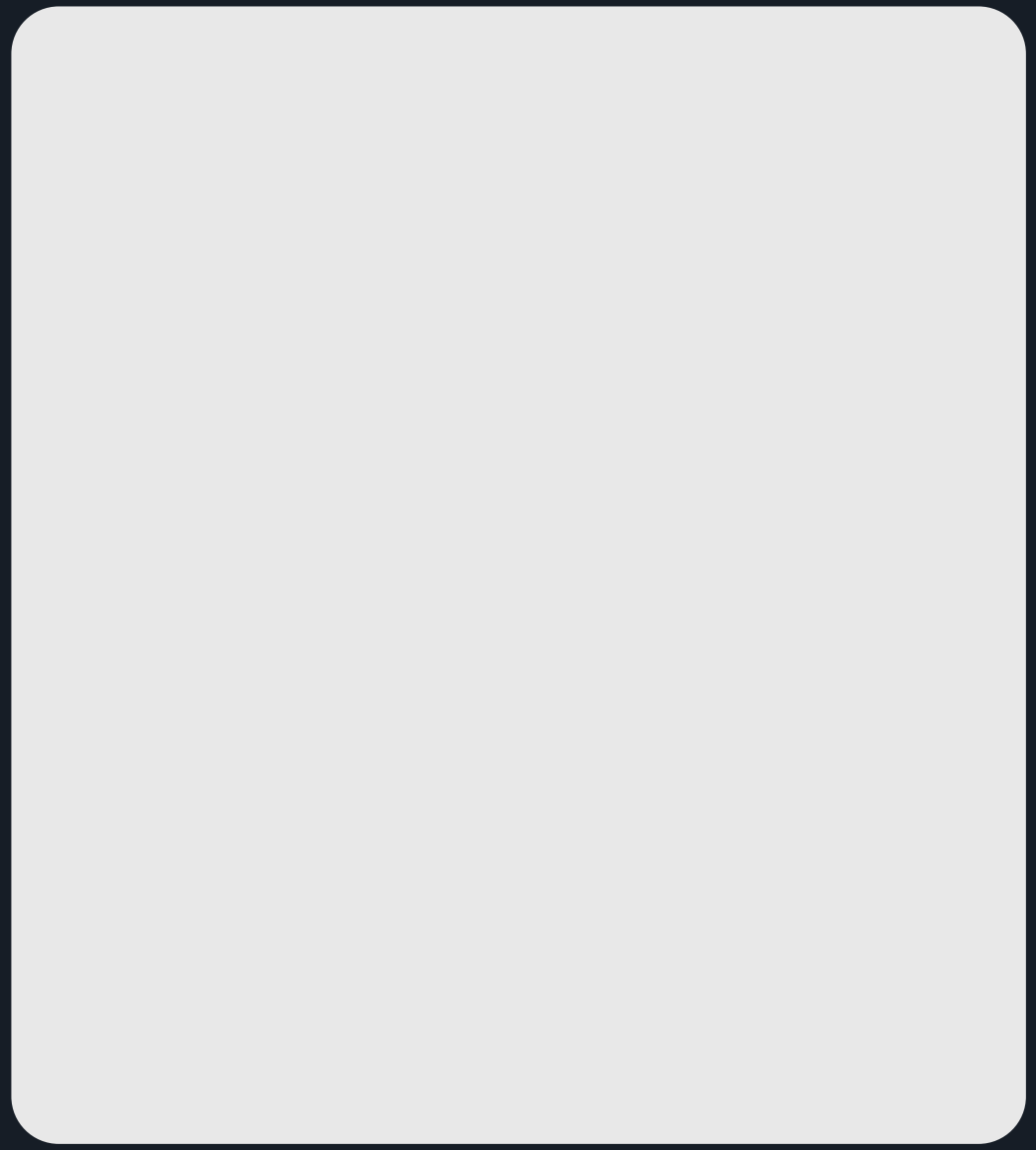


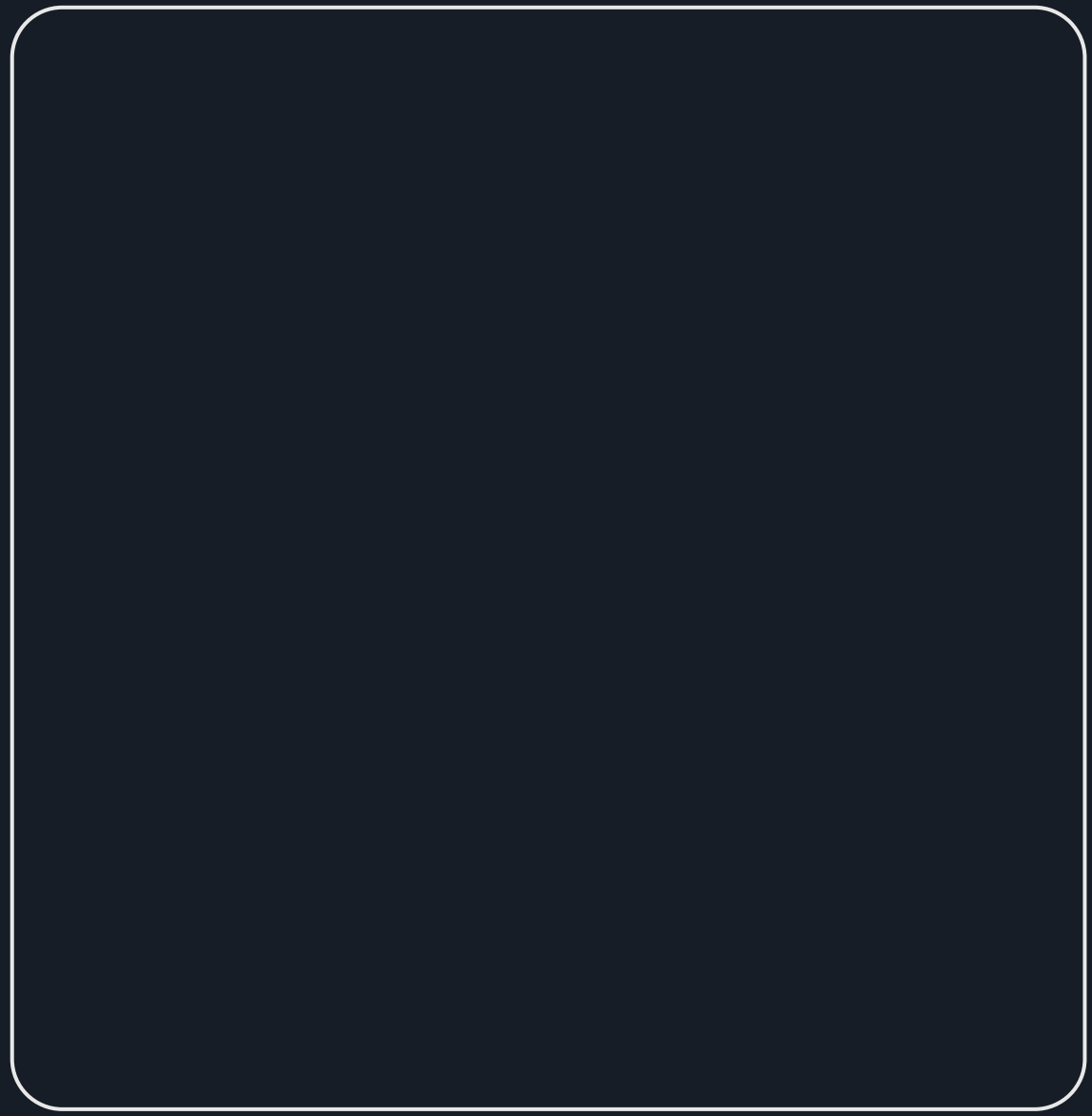




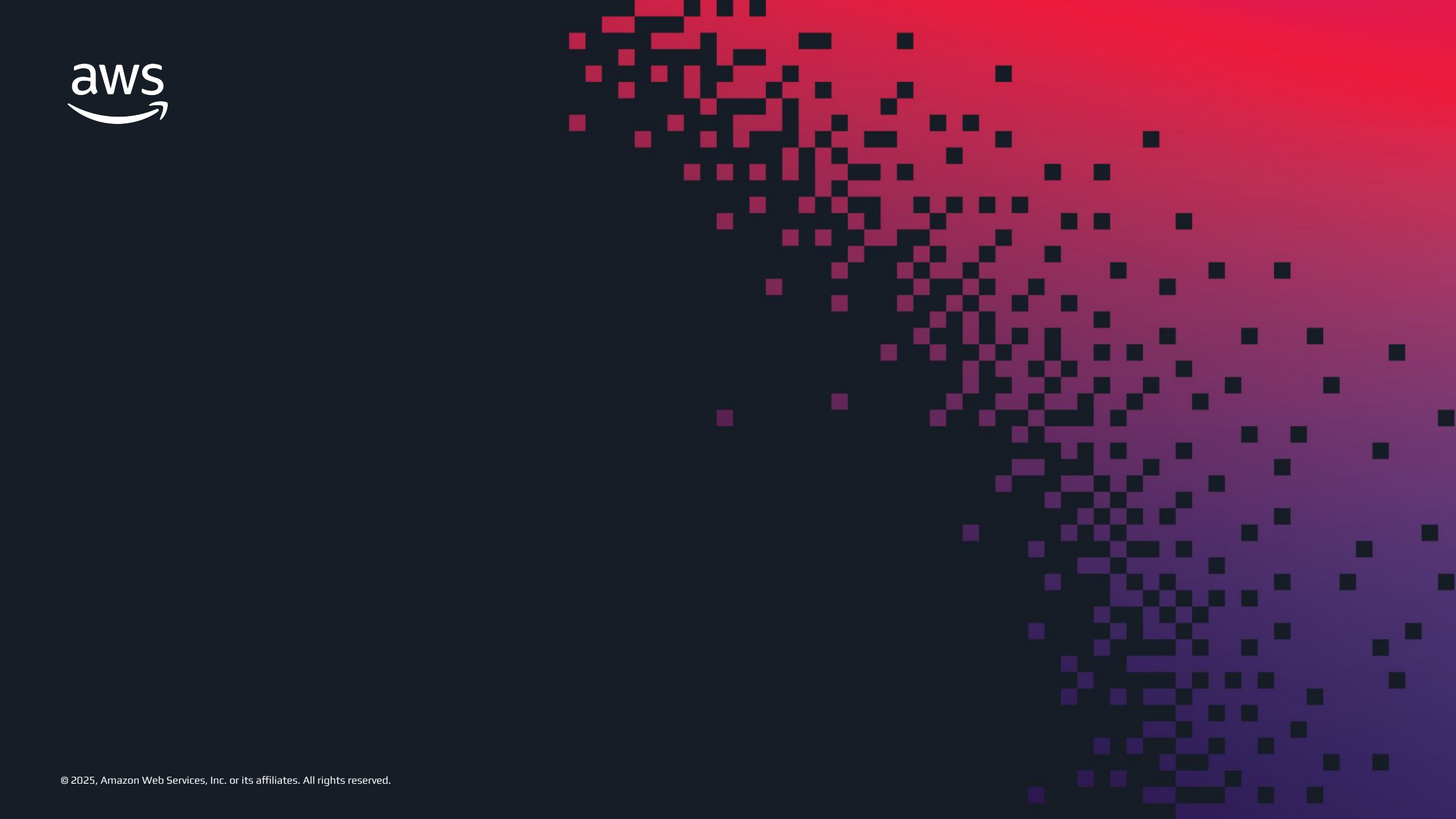


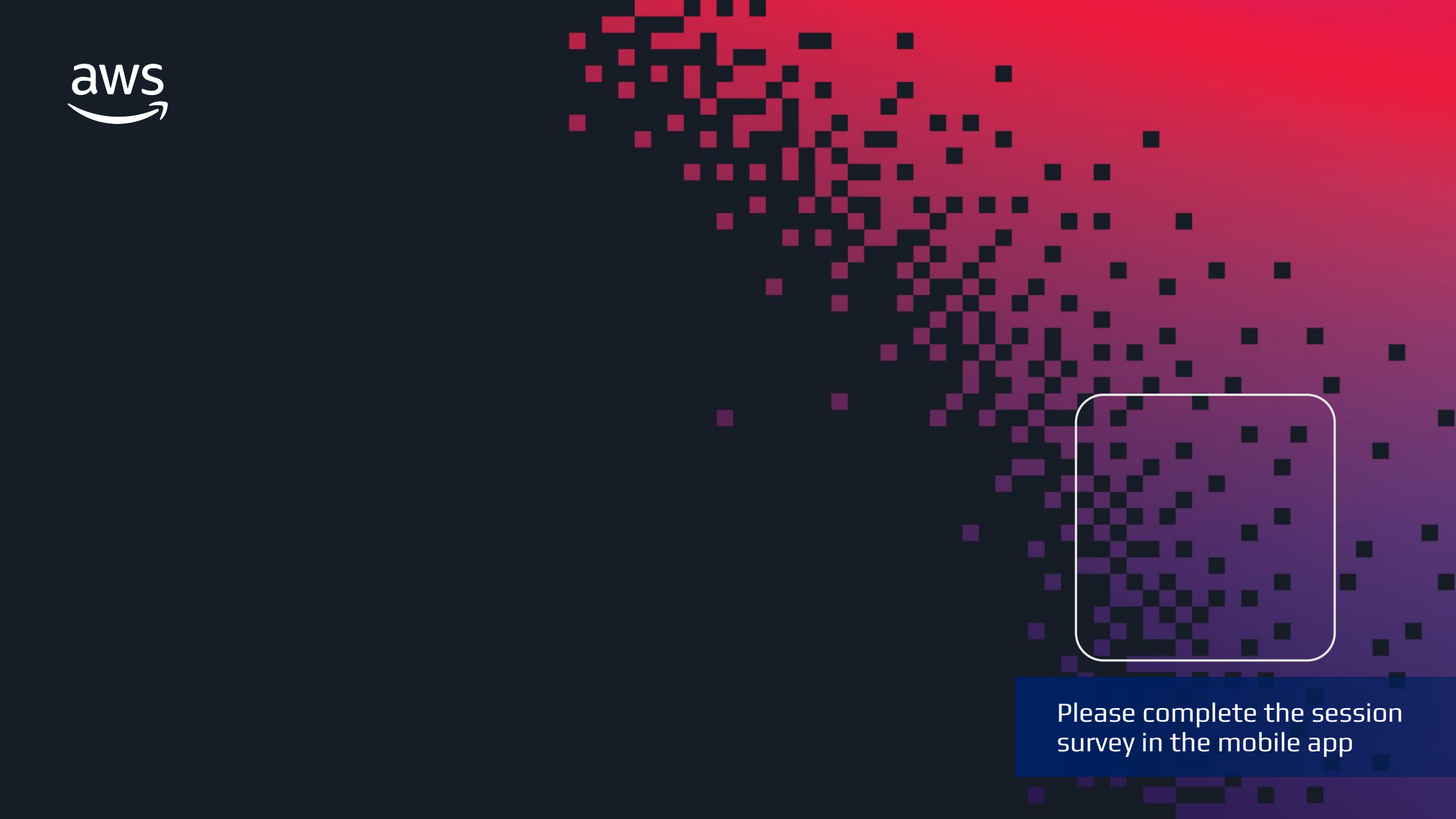


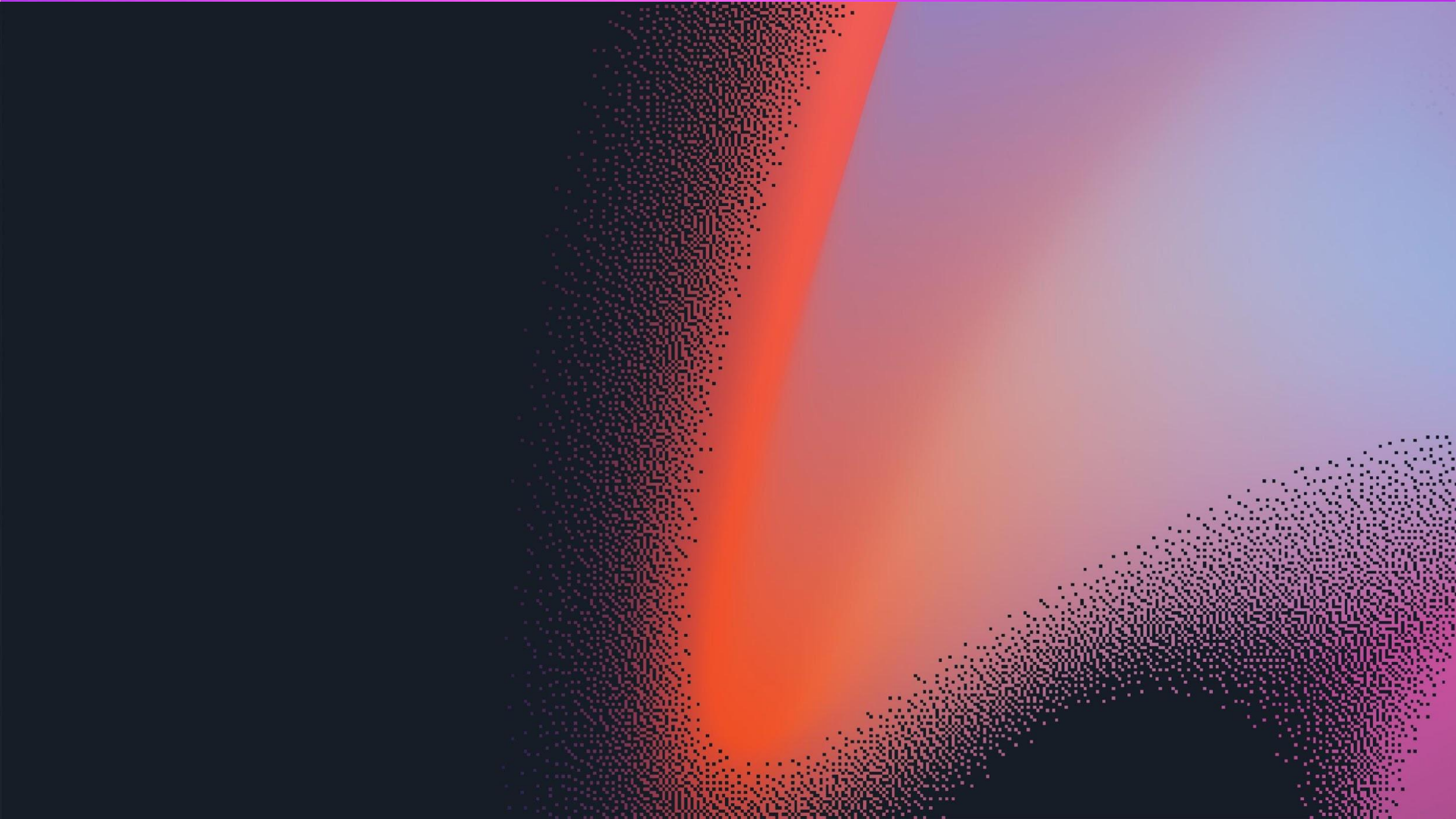




```
Code text
```




Please complete the session
survey in the mobile app



NEW

The next set of slides are example designs

These example slides are examples of the usage of the template’s layouts. Feel free to use and take inspiration from them!

Agenda

- Overview
- About the workstream
- Key players
- Cadence for comms
- Actions and resources

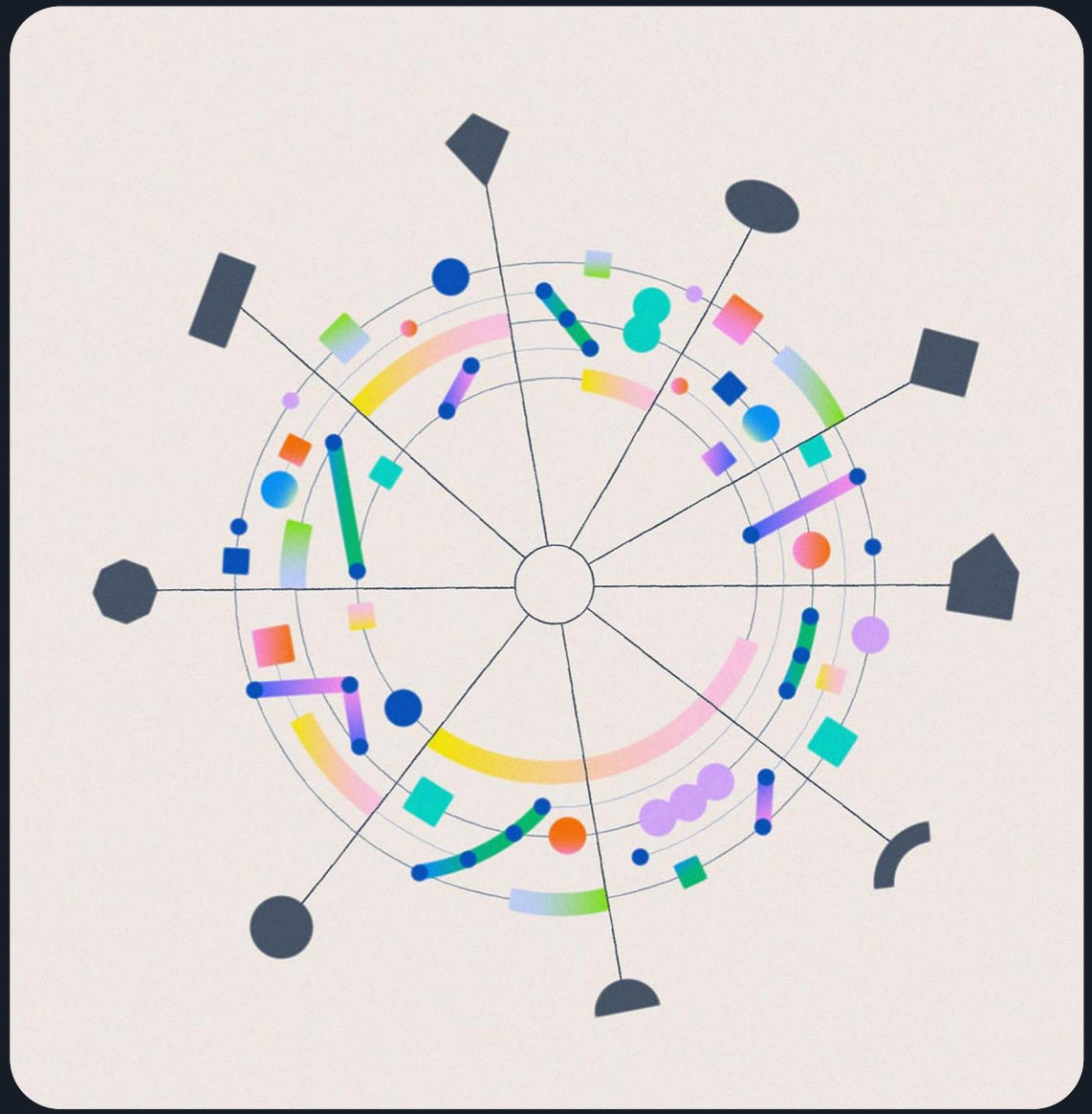
Actions and resources

Actions and resources

First internal comms – next week

Category strategy and templates

Individual workstreams connecting with POCs



HOW WE MAKE THIS HAPPEN

Mechanisms and Comms

Aligning teams through structure, transparency, and shared tools

- Weekly core workstream meetings
- Monthly LT updates
- Workstream wiki (launching soon)
- Shared allocation strategies across teams
- Centralized intake processes

Generative AI

Amazon Q Developer

The most capable generative AI–powered assistant for software development

Artificial Intelligence (AI)

Amazon SageMaker

Build, train, and deploy machine learning models at scale

Compute

Amazon Elastic Compute Cloud (EC2)

Virtual servers in the cloud


Explore the universe with AWS

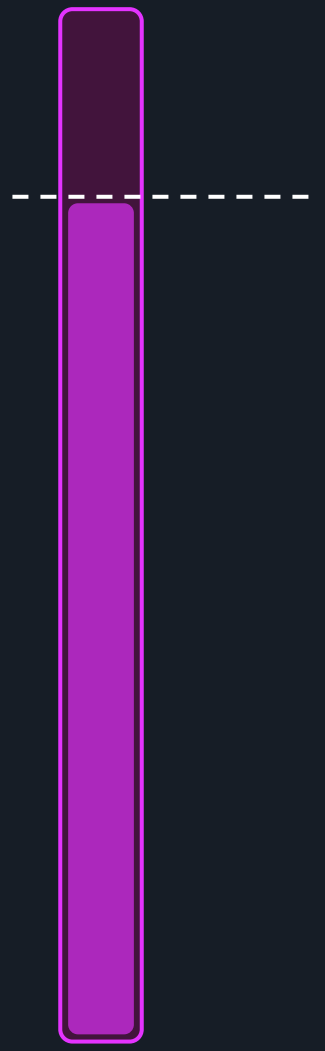
AMAZON BEDROCK

Build responsible AI applications with Guardrails

Amazon Bedrock Guardrails provides configurable safeguards to help safely build generative AI applications at scale.

AMAZON Q

An industry-leading enterprise AI solution customized for your organization

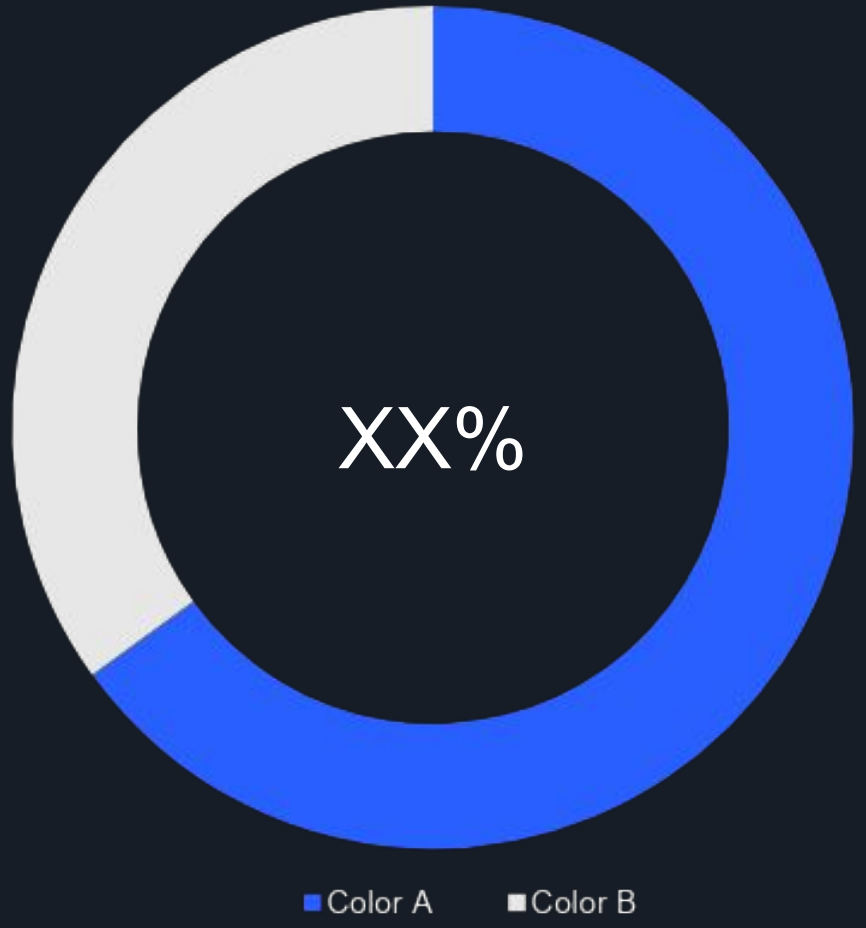


80%

Acceleration in coding tasks with Amazon Q's advanced agents

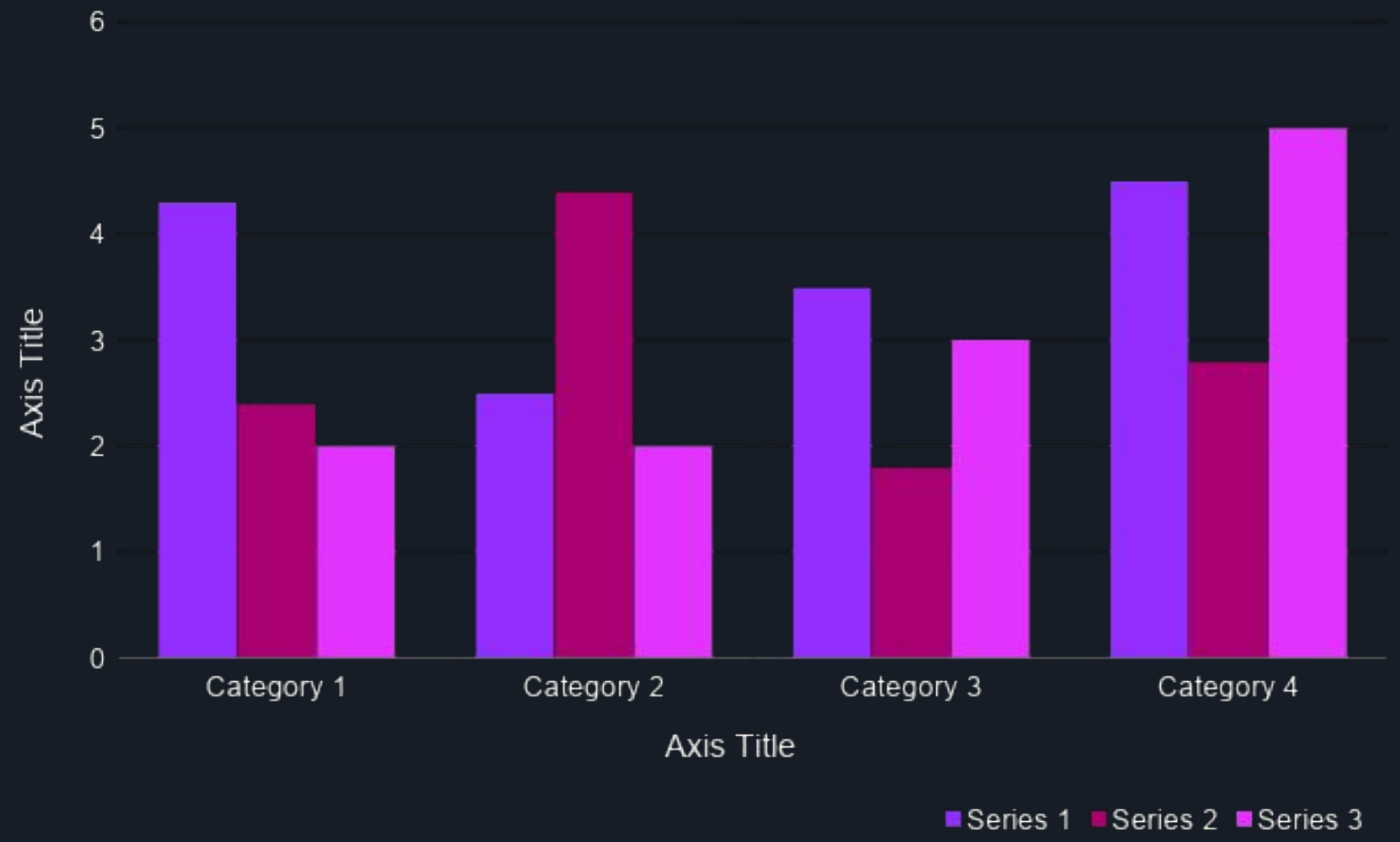
AMAZON Q

An industry-leading enterprise AI solution customized for your organization



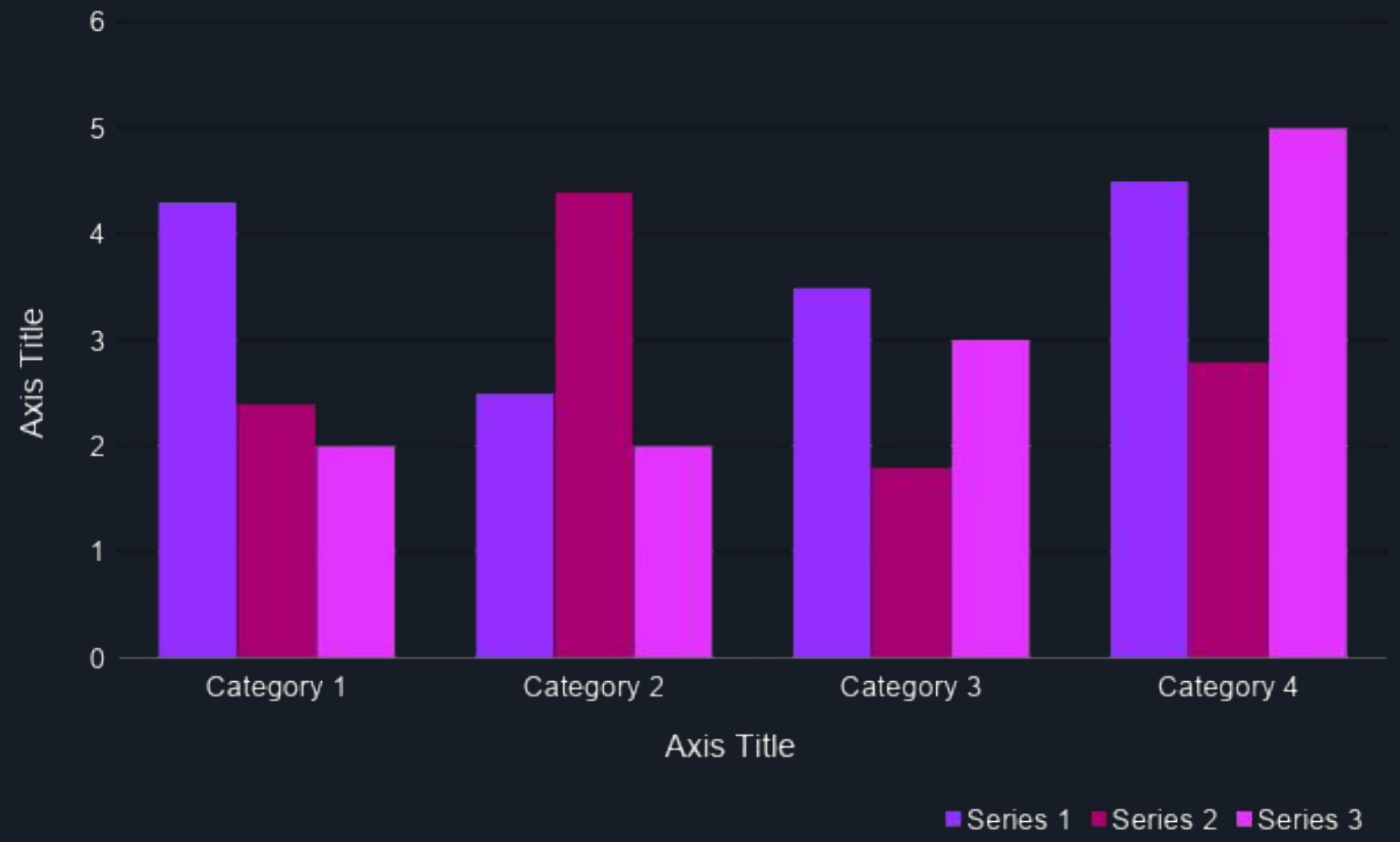
AMAZON Q

An industry-leading enterprise AI solution



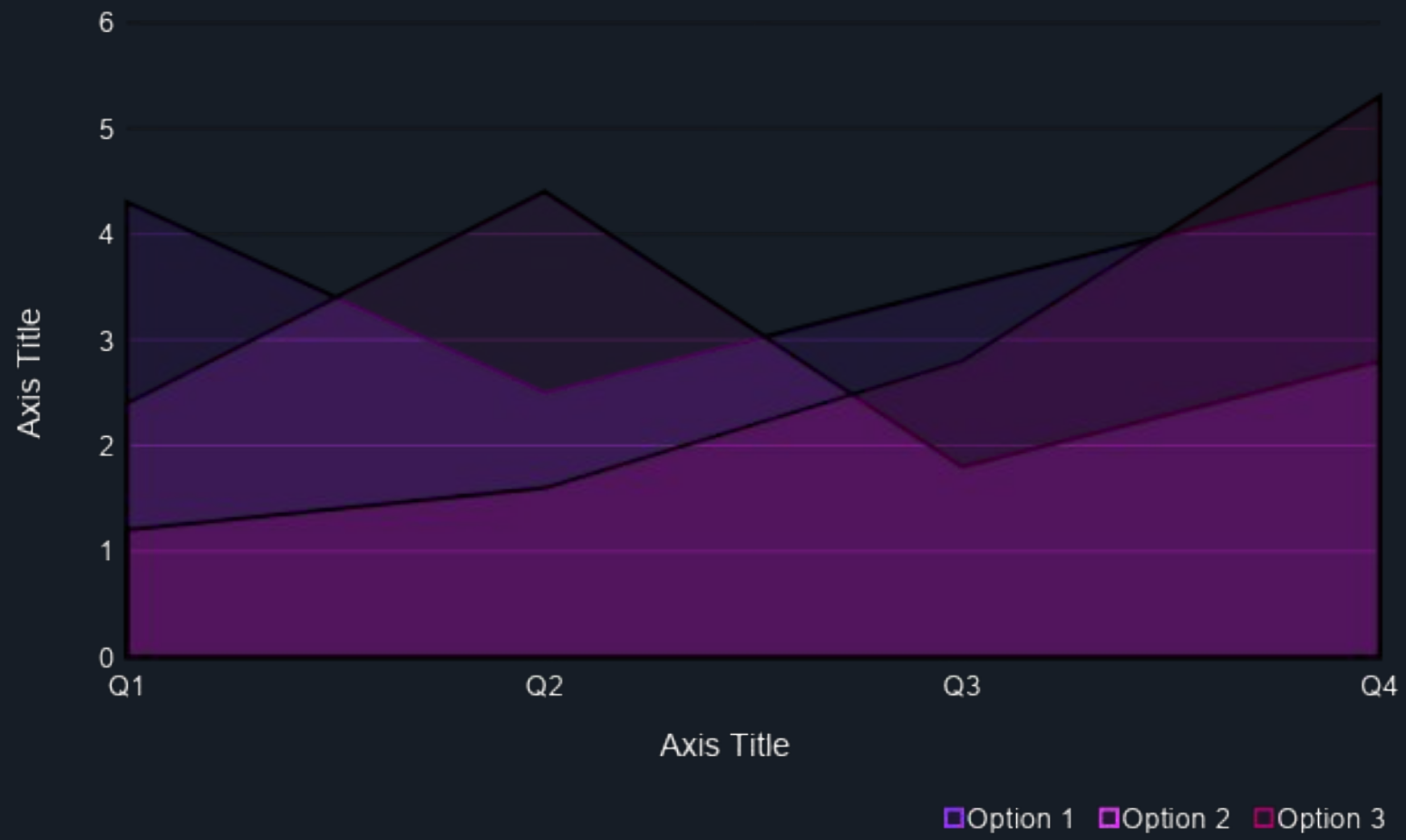
AMAZON Q

An industry-leading enterprise AI solution



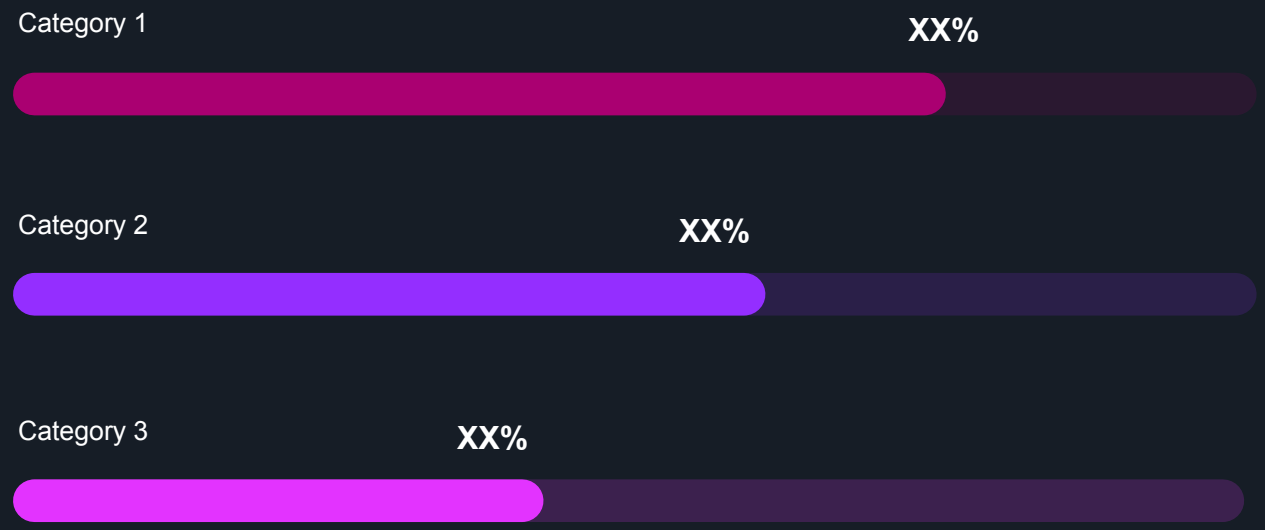
AMAZON Q

An industry-leading enterprise AI solution



AMAZON Q

An industry-leading enterprise AI solution customized for your organization



Label A

Lorem ipsum dolor sit amet, consectetur adipiscing elit

Label B

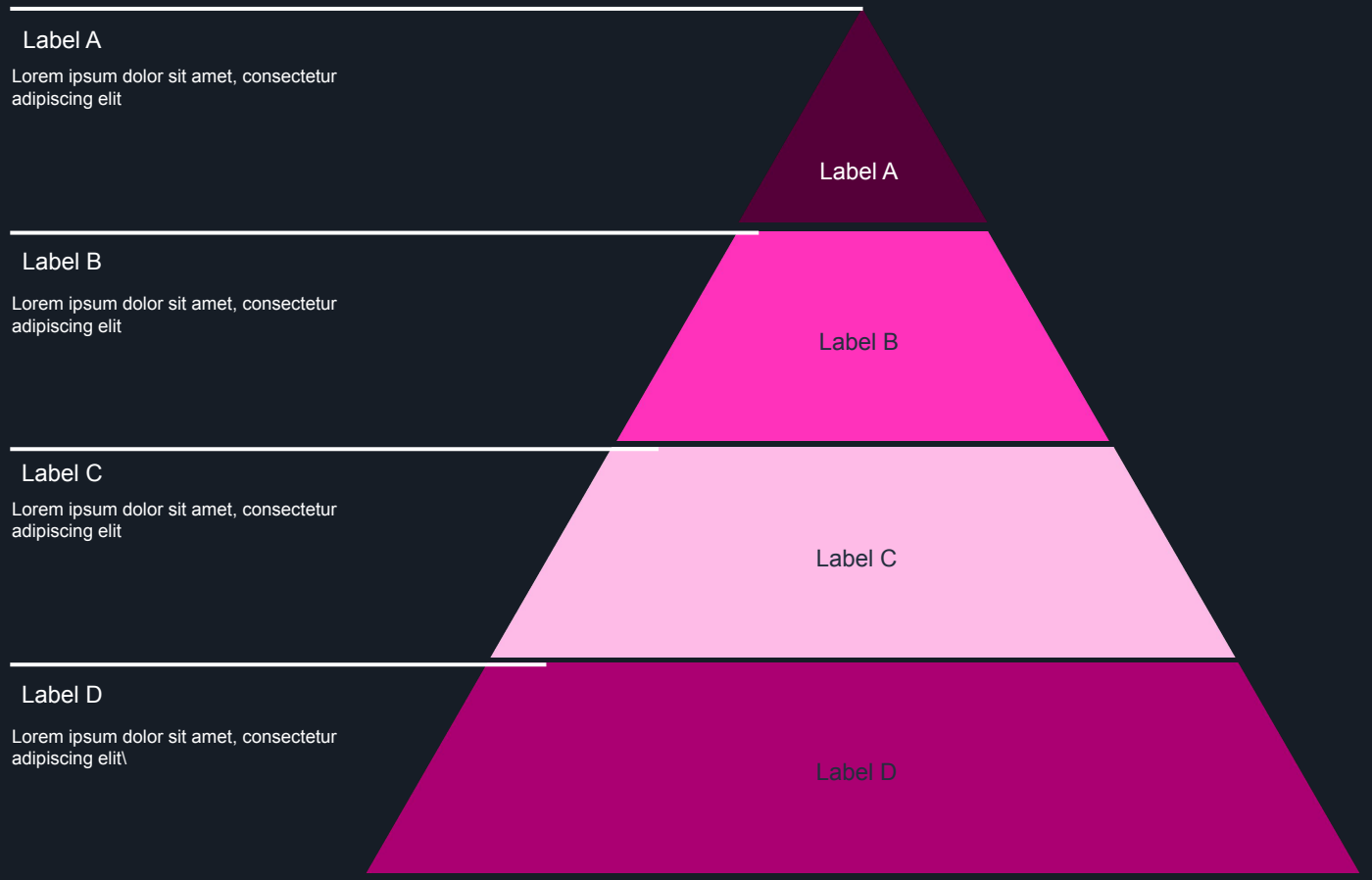
Lorem ipsum dolor sit amet, consectetur adipiscing elit

Label C

Lorem ipsum dolor sit amet, consectetur adipiscing elit

Label D

Lorem ipsum dolor sit amet, consectetur adipiscing elit\



Bayer Crop Science scales regenerative agriculture with AWS generative AI

Bayer Crop Science is empowering its data scientists to innovate faster and easier using generative AI.

Amazon Q Business

Up to 70%

reduction in employee onboarding time with Amazon Q Business

Amazon Q Developer

Up to 30%

increase in developer productivity with Amazon Q Developer