Introduction to Agentic AI

A*STAR Leadership Retreat

Dr Gaurav Manek

2026-10-08

ocelliq.com/astar-lr



OcelliQ.com/astar-lr

Access course materials

Ask your TA to help you set up.

Outline

1. The AI industry landscape
2. What is AI?
3. What makes AI agentic?
4. Where AI fails

Hands-on:
5. Procurement: a first run

6. Demystifying the IMDA Agentic AI Governance Framework

Hands-on:
7. Set policies for a procurement agent

8. ATLAS: governance in practice
9. Where do we go from here?

AI for leadership

This course is for business leaders, not engineers.

1. Identify high-value applications.
2. Assess specific business **risks**.
3. Define controls and approval rules.
4. Assign human oversight and accountability.

Zero coding.



Dr. Gaurav Manek

- Founder, **Ocellivision**
- Technical Lead, **ATLAS** @ IMCB
- **PhD** in AI/ML — Carnegie-Mellon University (2023)
- Founder, **Visigoth.ai** (SaaS)

For follow-up questions:
gaurav_manek@a-star.edu.sg
gauravmanek

About Our TAs



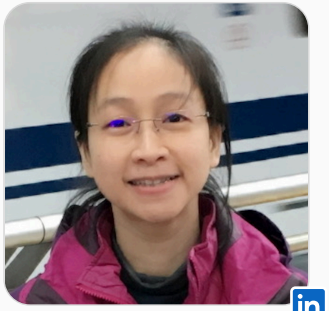
Dr. Aarthi Ravikrishnan
ASRL



Dr. Adaikalavan Ramasamy
ASRL / GIS



Dr. Gokce Oguz
ASRL / GIS



Dr. Chinh Tran-To Su
BII



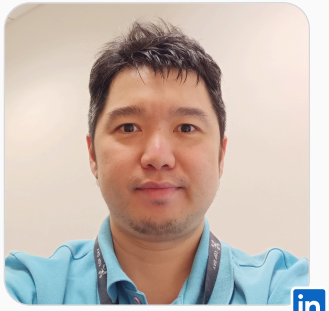
Ms. Dong Jiahui
BDH



Mr. Guai Zi Wei
ETO



Dr. Audrey Lee
GIS



Dr. Benedict Wong
IAIC



Dr. Benjamin Chew
IAIC



Dr. Farzam Farbiz
IAIC



Dr. Amhed Missael Vargas Velazquez
IMCB



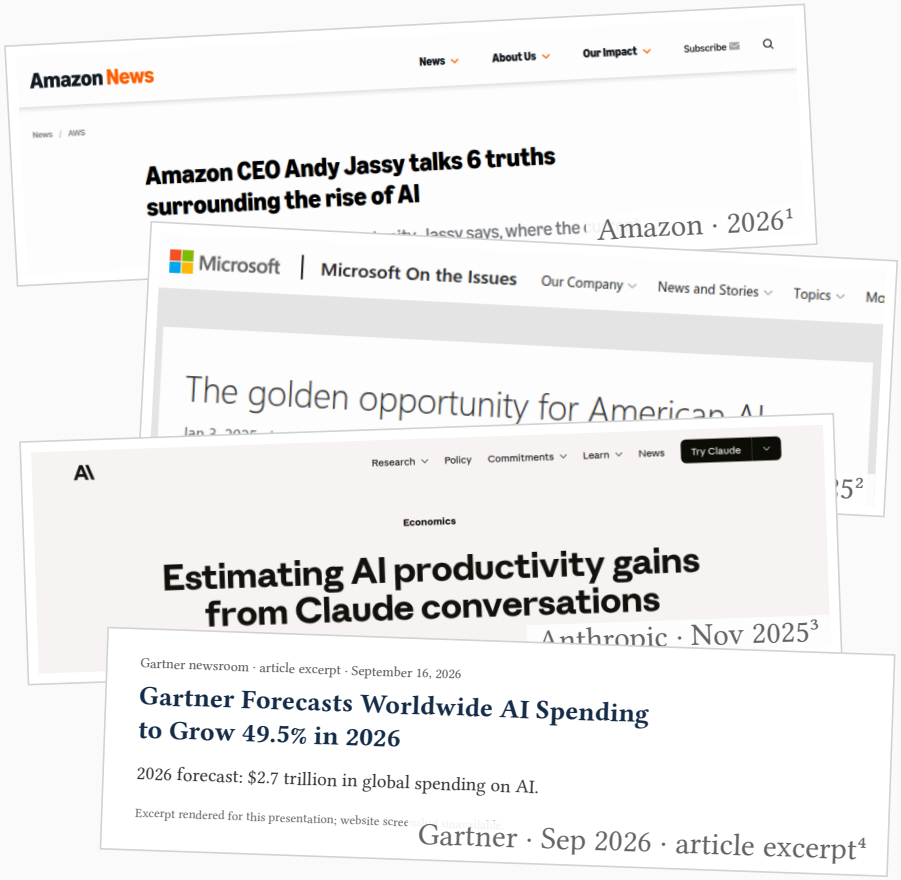
Mr. Xavier Wilbin
IMCB

Champions of AI

1. Industry Landscape

1.1 Public Perception of AI

Promise

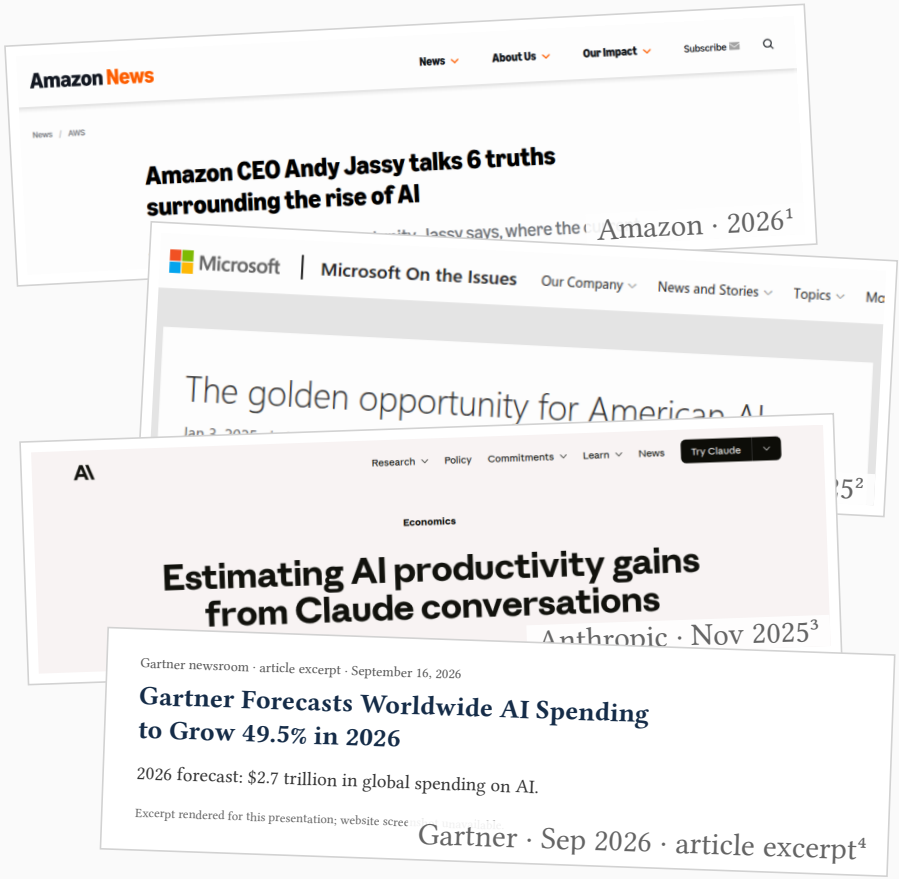


Risk



1.1 Public Perception of AI

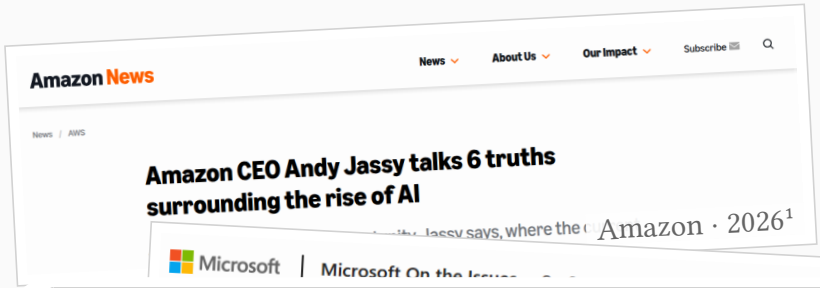
Promise



Risk



Promise



Risk in missing this opportunity



Risk



Risk in moving too quickly



Global AI infrastructure spending forecast for 2026[4]

US$1.48T

≈5× Apollo missions[8] (US$309 billion)

≈3× Great Wall of China rebuild[9] (US$452 billion)

≈1× Cumulative global datacenter spend until the mid-2000s[10]

Global AI-server electricity consumption forecast for 2026[11]

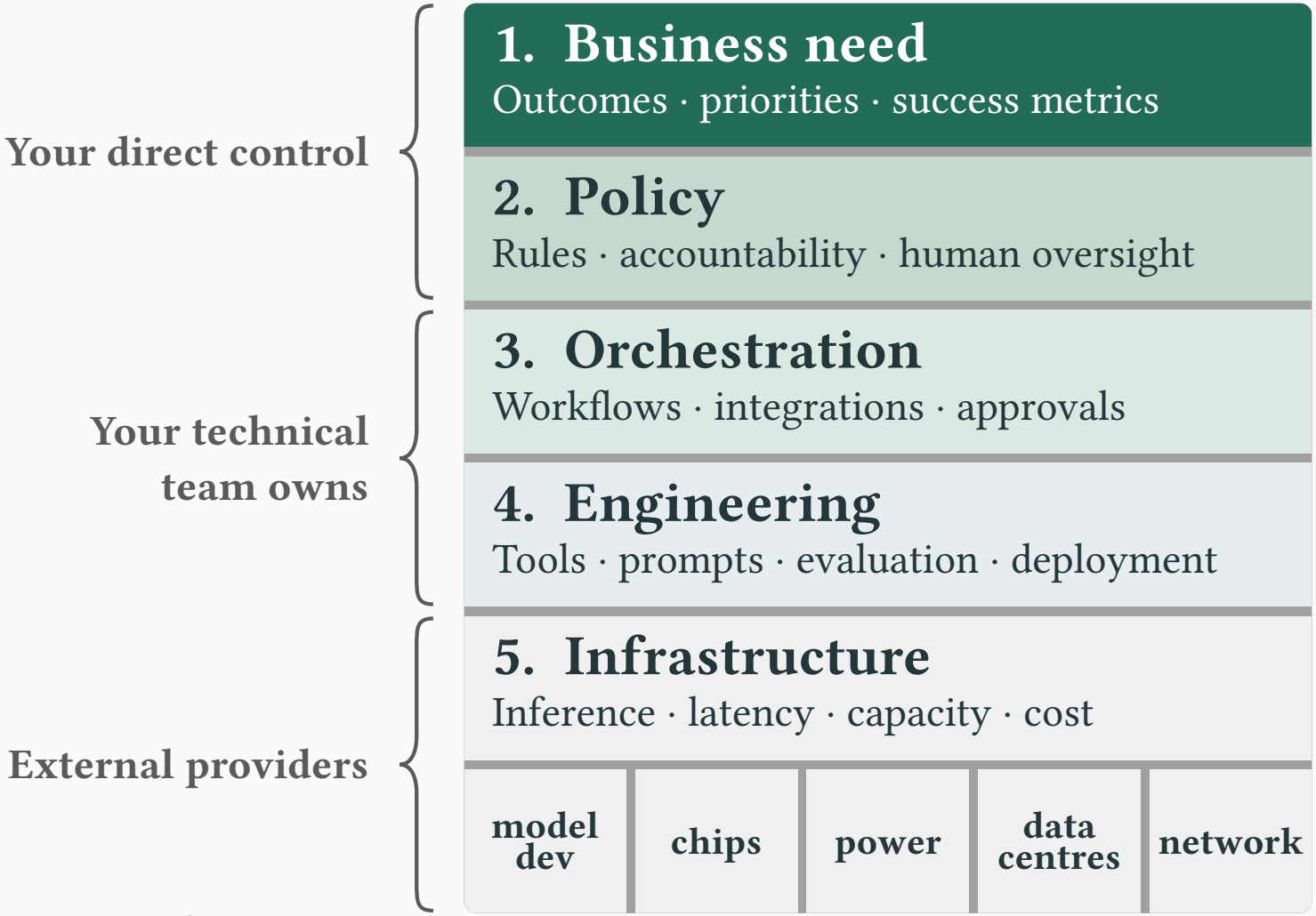
175 TWh

≈3× Singapore electricity consumption in 2024[12] (58 TWh)

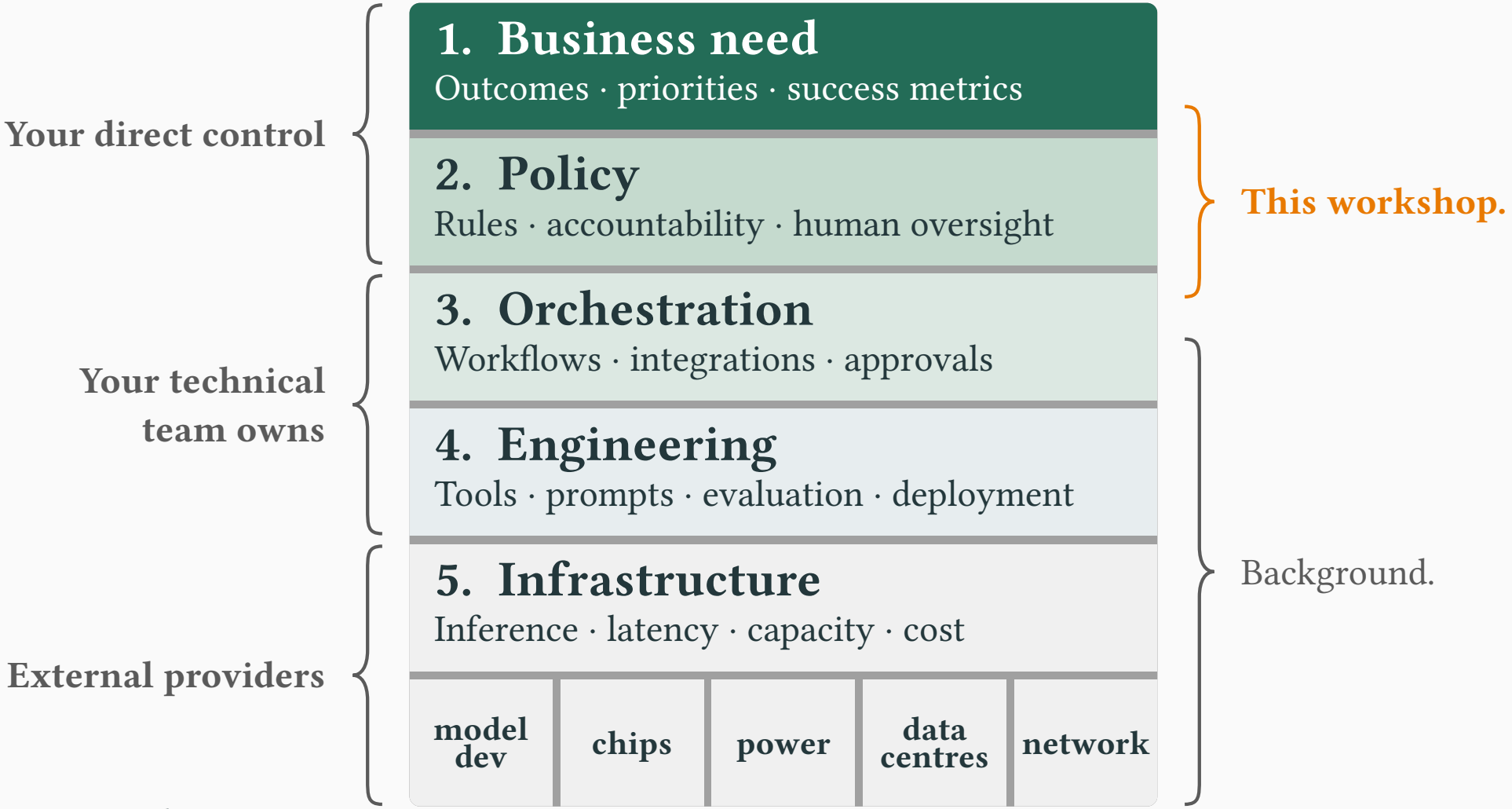
≈2× Three Gorges' designed annual output[13] (88.2 TWh)

*All rough estimates

1.3 Who Controls What in AI



1.3 Who Controls What in AI



2. What *is* AI?

2.1 What Does “AI” Mean?

- “AI” is a **marketing** and technical term
  - From rule-based systems to neural networks
  - Promises to revolutionize everything
- Massive returns in software, marketing, etc.
  - Just starting in some conservative industries.
- Turn decisions into a commodity.[19]

Not magic

- “AI” is a **marketing** and technical term
  - From rule-based systems to neural networks
  - Promises to revolutionize everything
- Massive returns in software, marketing, etc.
  - Just starting in some conservative industries.
- Turn decisions into a commodity.[19]

Not magic

Jevons paradox [20]

AI won’t make working easier. It raises the bar for everyone and increases competition.

The only competitive moat left is your speed of integration.

Adopt early, or **get left behind.**

An LLM is Fancy Autocomplete

An LLM predicts the **next** token / word / step

An LLM predicts the next **token** given / using / from

An LLM predicts the next token **given** every / all / the

An LLM predicts the next token given **every** thing / word / token

An LLM predicts the next token given every**thing** before / so / in

An LLM predicts the next token given everything **before** “.” / “,” / “and”

An LLM predicts the next token given everything before **.** (end) / It / This

An LLM is Fancy Autocomplete

An LLM predicts the **next** token / word / step

An LLM predicts the next **token** given / using / from

An LLM predicts the next token **given** every / all / the

An LLM predicts the next token given **every** thing / word / token

An LLM predicts the next token given every**thing** before / so / in

An LLM predicts the next token given everything **before** "." / "," / "and"

An LLM predicts the next token given everything before **.** (end) / It / This

1 token ≈ ¾ of an English word, a punctuation mark, or a digit.[22]

Unit of computation: pay per token in and out. The *context window* is sized in tokens.

An LLM is Fancy Autocomplete

- Blocks of linear algebra
  - arranged in creative ways
  - trained on every scrap of human media
  - unimaginable amounts of computation
  - learn statistical patterns, not explicit rules about the world
- Behaviour **surprisingly** emerges at scale
  - Exhibits strengths and weaknesses very different from humans

An LLM is Fancy Autocomplete

- Blocks of linear algebra
  - arranged in creative ways
  - trained on every scrap of human media
  - unimaginable amounts of computation
  - learn statistical patterns, not explicit rules about the world
- Behaviour **surprisingly** emerges at scale
  - Exhibits strengths and weaknesses very different from humans

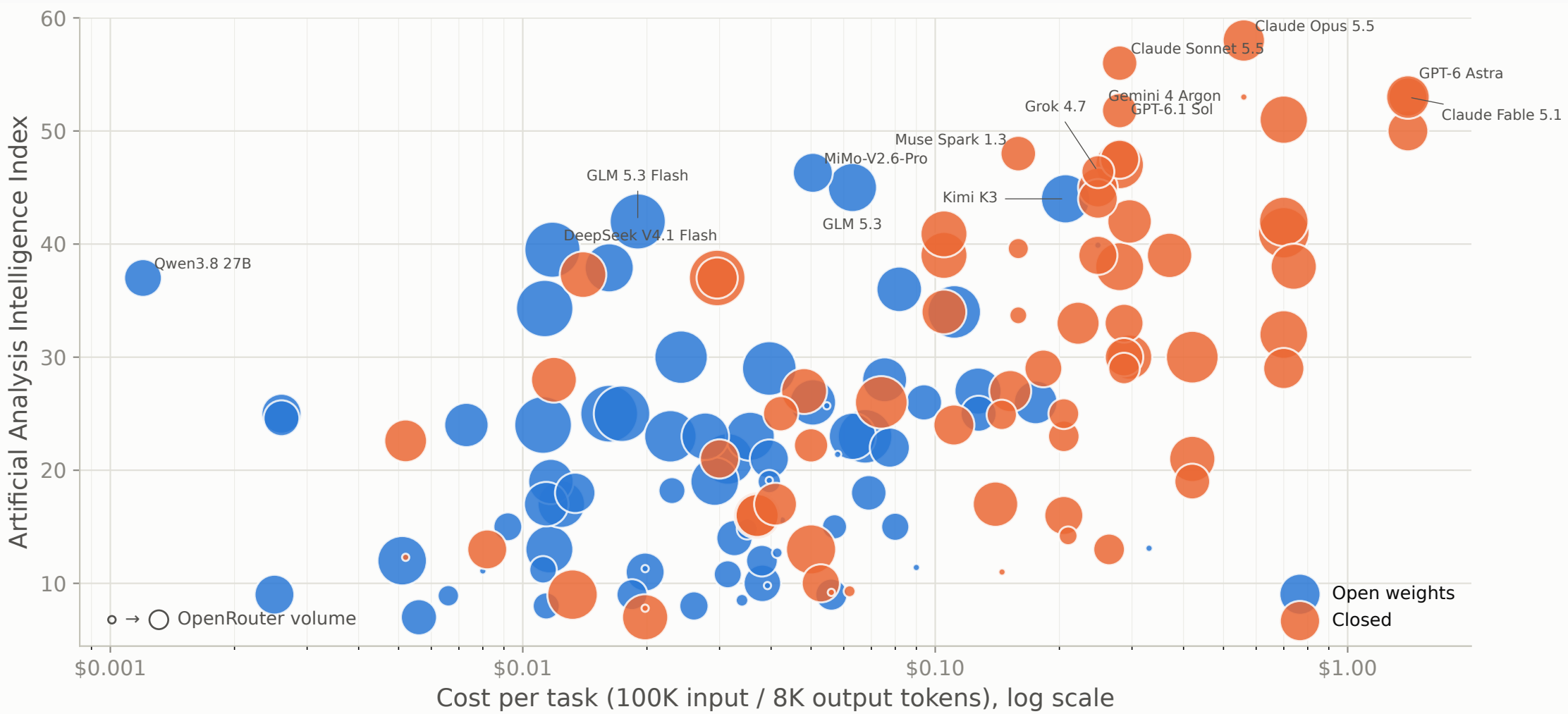
LLMs are black boxes: we have ability to predict and explain their behaviour.

LLMs are stochastic: Identical runs often produce different output.

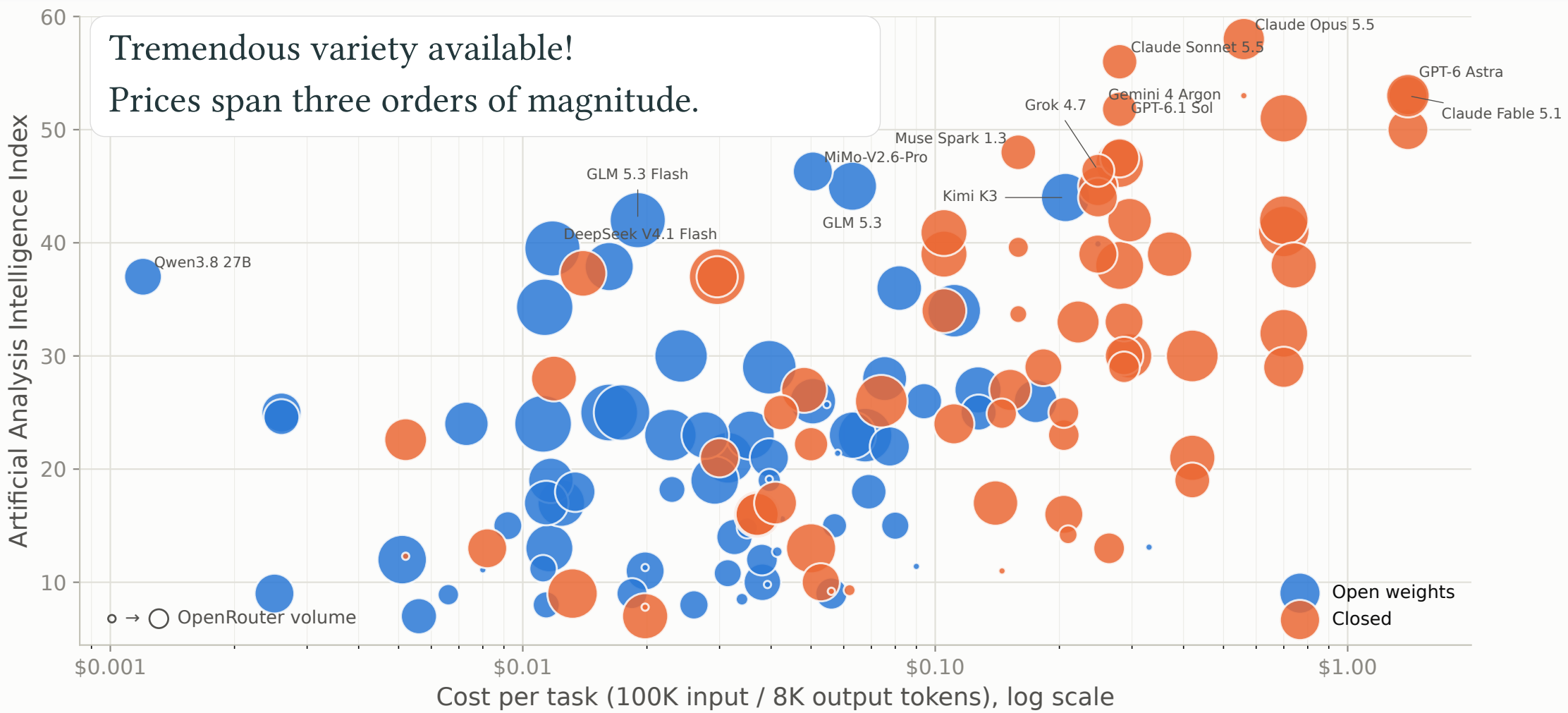
- `Sunny skies ☀, 30°C.`
- `30°C with plenty of sunshine.`
- `It's sunny and 30°C today.`

Agent engineering extracts useful work despite these limits.

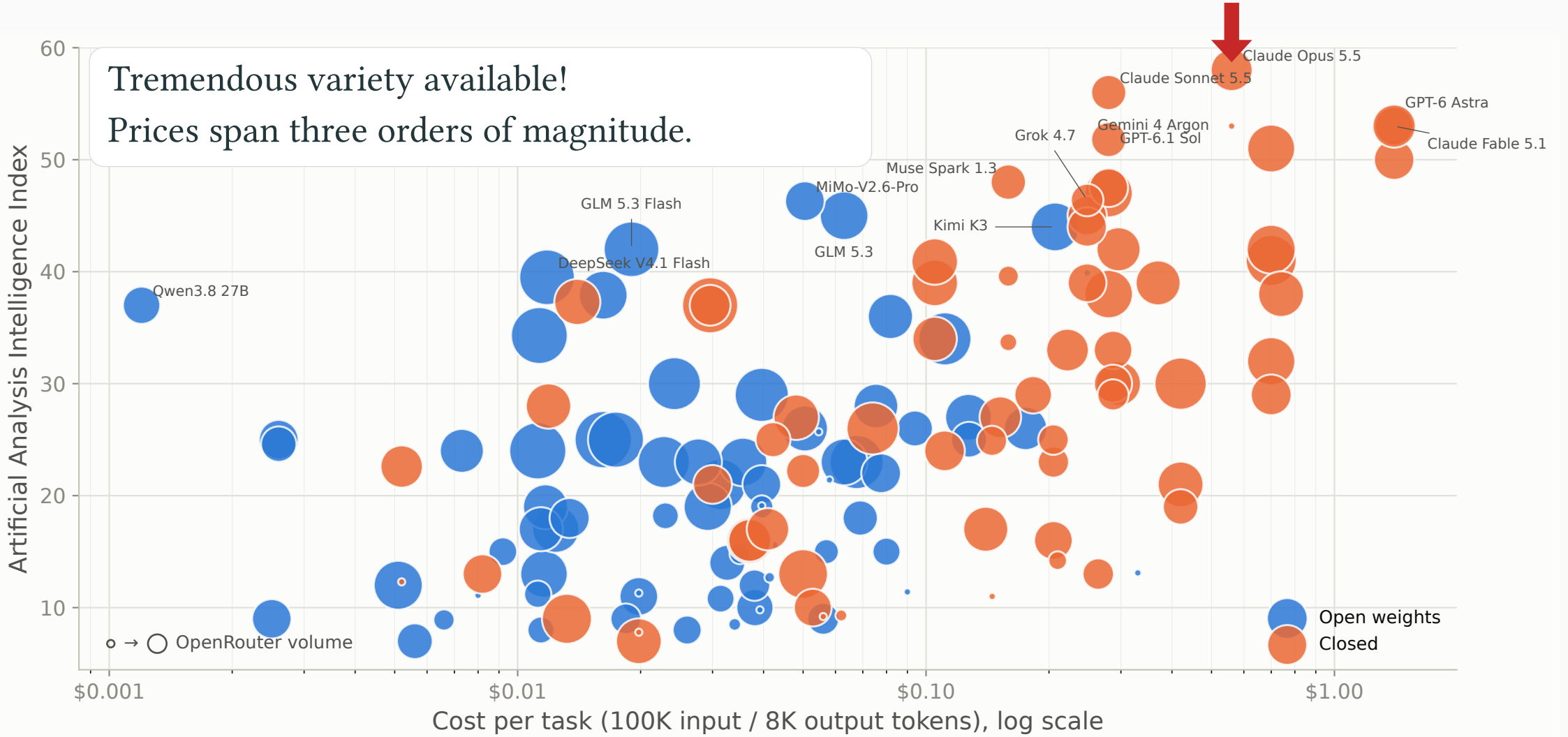
2.4 Cost of AI


Artificial Analysis Intelligence Index
60
50
40
30
20
10
Claude Opus 5.5
Claude Sonnet 5.5
GPT-6 Astra
Gemini 4 Argon
GPT-6.1 Sol
Claude Fable 5.1
Grok 4.7
Muse Spark 1.3
MiMo-V2.6-Pro
GLM 5.3 Flash
Kimi K3
GLM 5.3
DeepSeek V4.1 Flash
Qwen3.8 27B
Open weights
Closed
OpenRouter volume
$0.001
$0.01
$0.10
$1.00
Cost per task (100K input / 8K output tokens), log scale

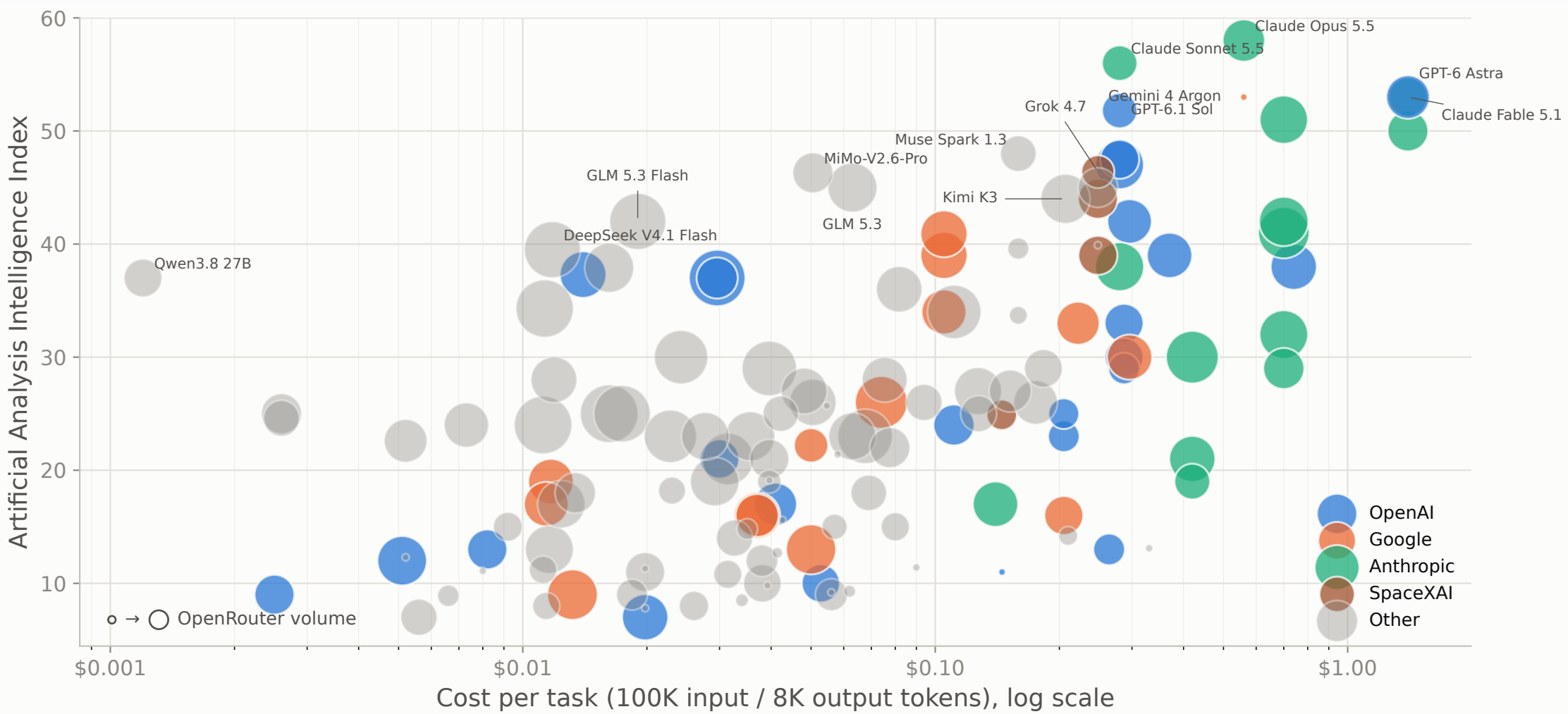
2.4 Cost of AI


Tremendous variety available!
Prices span three orders of magnitude.
Artificial Analysis Intelligence Index
60
50
40
30
20
10
Claude Opus 5.5
Claude Sonnet 5.5
GPT-6 Astra
Gemini 4 Argon
GPT-6.1 Sol
Claude Fable 5.1
Grok 4.7
Muse Spark 1.3
MiMo-V2.6-Pro
GLM 5.3 Flash
Kimi K3
GLM 5.3
DeepSeek V4.1 Flash
Qwen3.8 27B
Open weights
Closed
OpenRouter volume
$0.001
$0.01
$0.10
$1.00
Cost per task (100K input / 8K output tokens), log scale

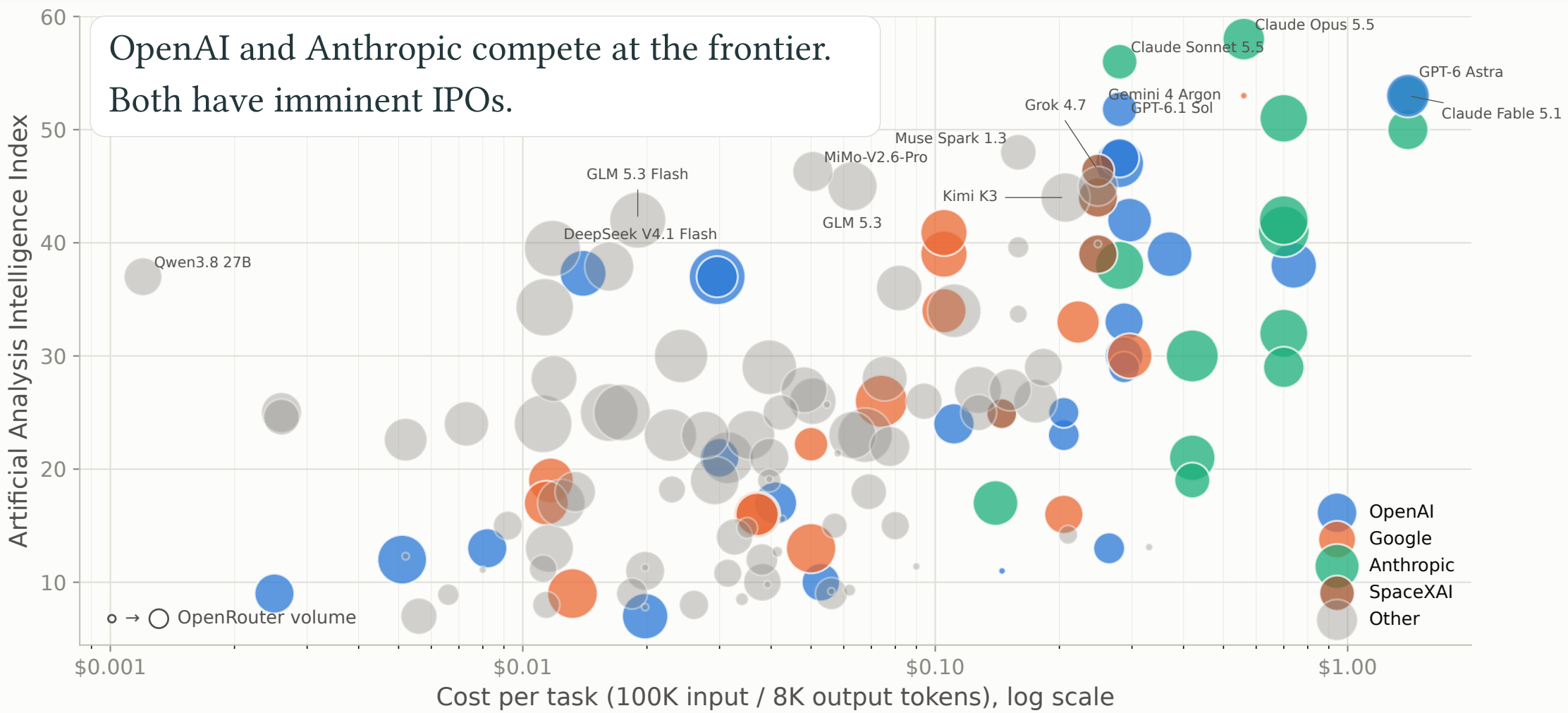
2.4 Cost of AI


Tremendous variety available!
Prices span three orders of magnitude.
Claude Opus 5.5
Claude Sonnet 5.5
GPT-6 Astra
Gemini 4 Argon
GPT-6.1 Sol
Claude Fable 5.1
Grok 4.7
Muse Spark 1.3
MiMo-V2.6-Pro
GLM 5.3 Flash
Kimi K3
GLM 5.3
DeepSeek V4.1 Flash
Qwen3.8 27B
Open weights
Closed
OpenRouter volume
Artificial Analysis Intelligence Index
60
50
40
30
20
10
$0.001
$0.01
$0.10
$1.00
Cost per task (100K input / 8K output tokens), log scale

2.5 Cost of AI by Lab (Closed-Source)

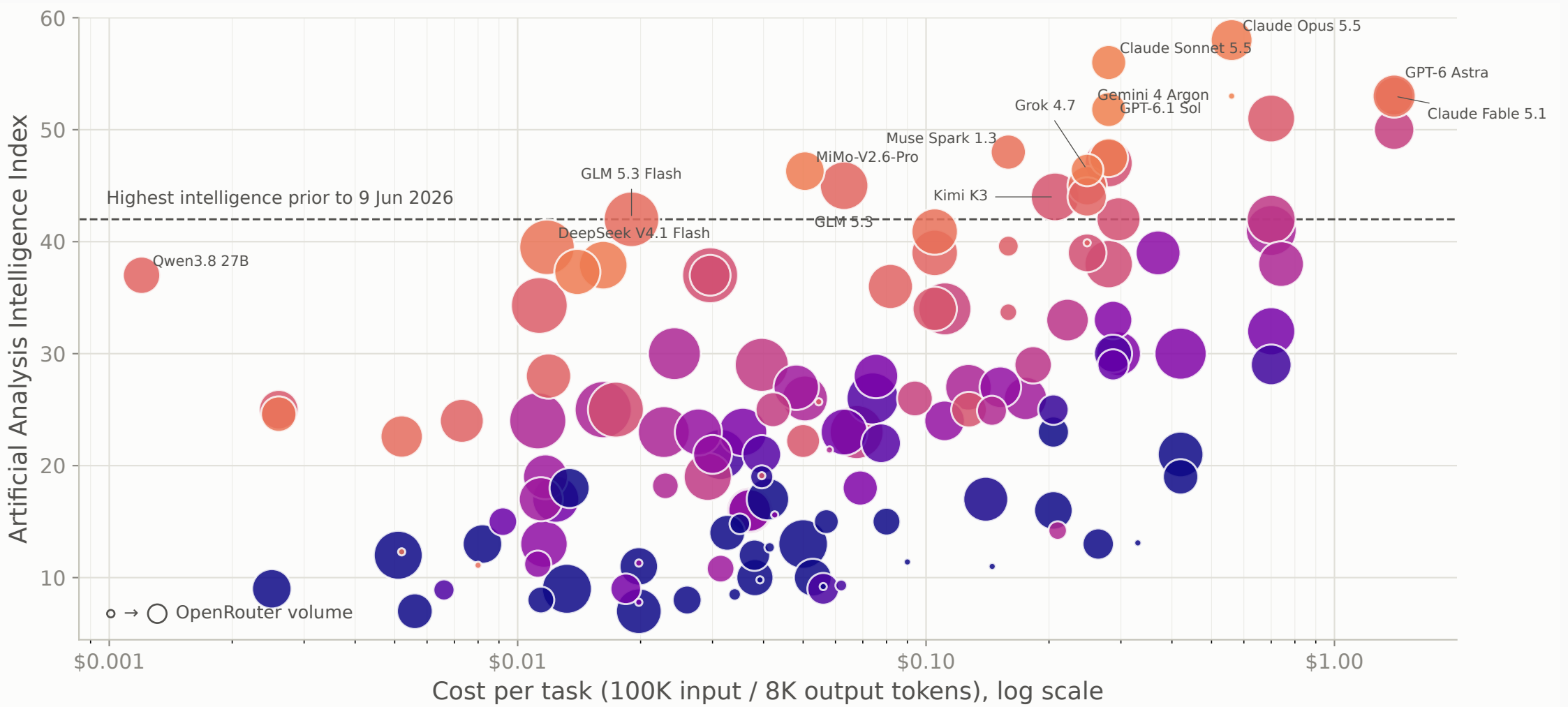


2.5 Cost of AI by Lab (Closed-Source)



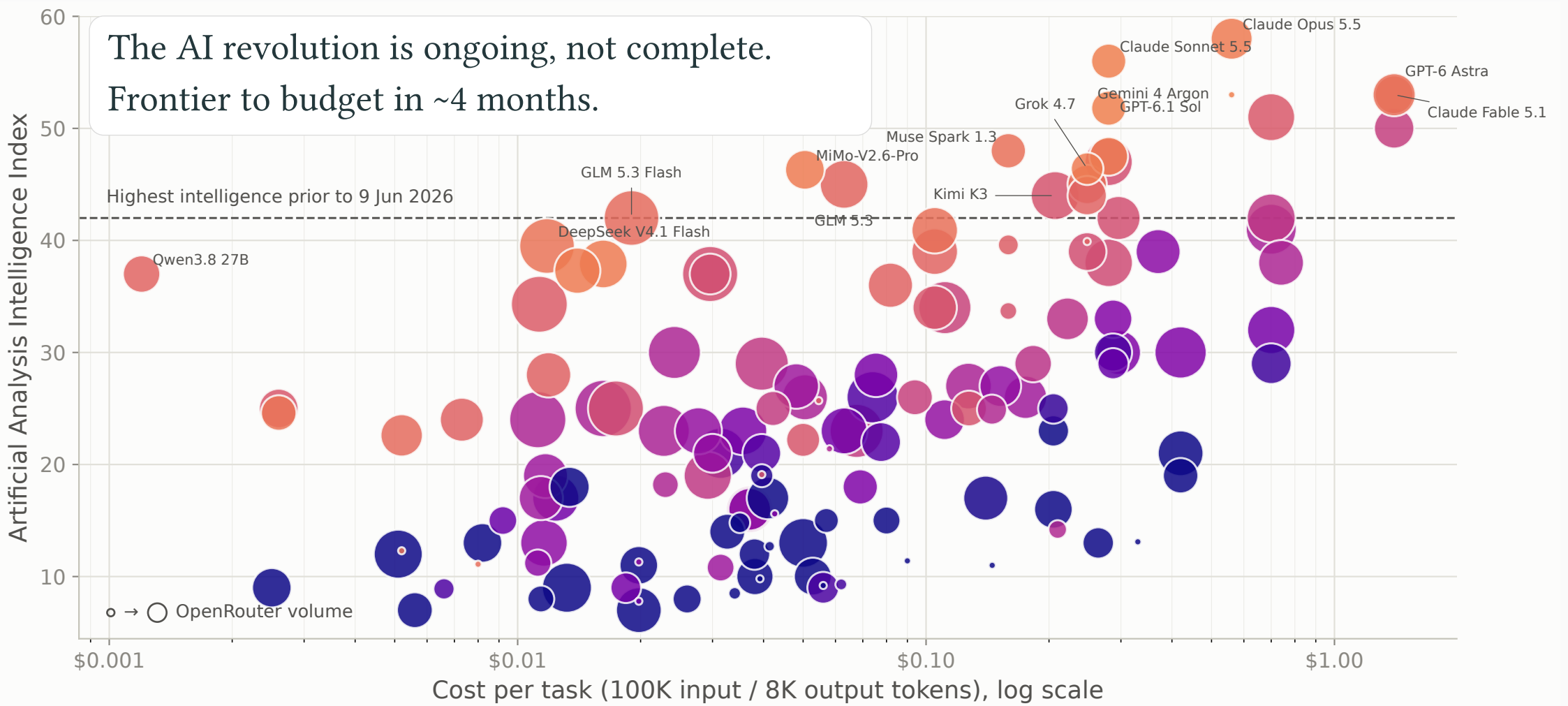
2.6 Cost of AI by Launch Date

New → 12+ months

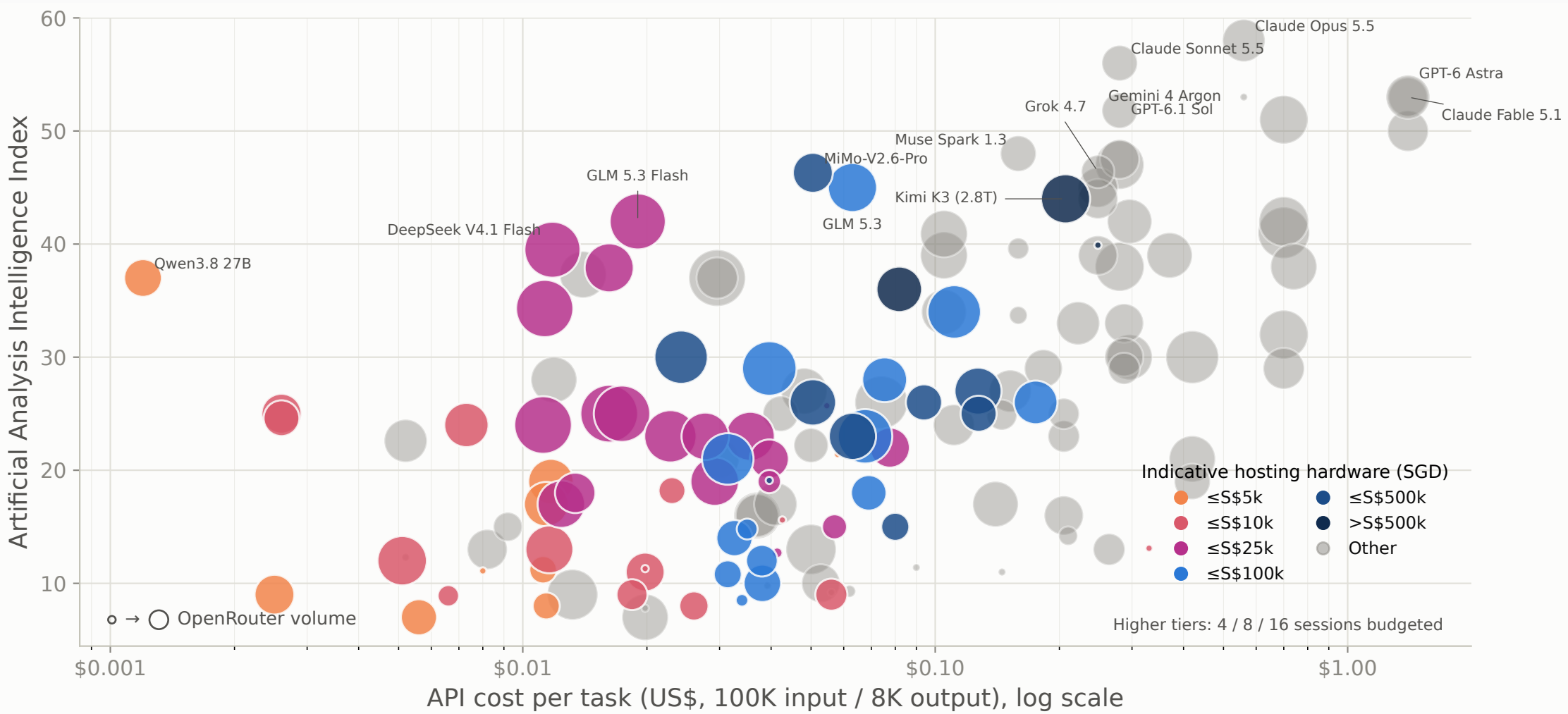


2.6 Cost of AI by Launch Date

New → 12+ months

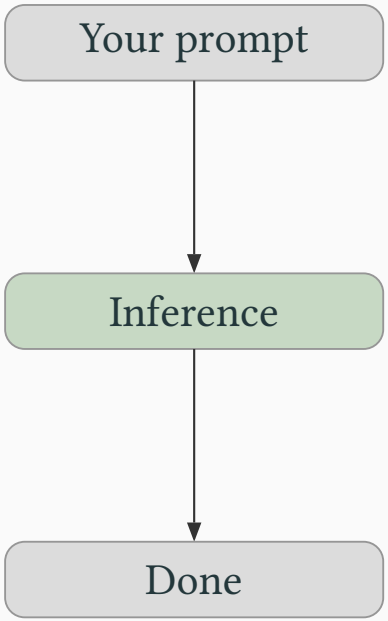


2.7 Cost of AI by Minimum CAPEX



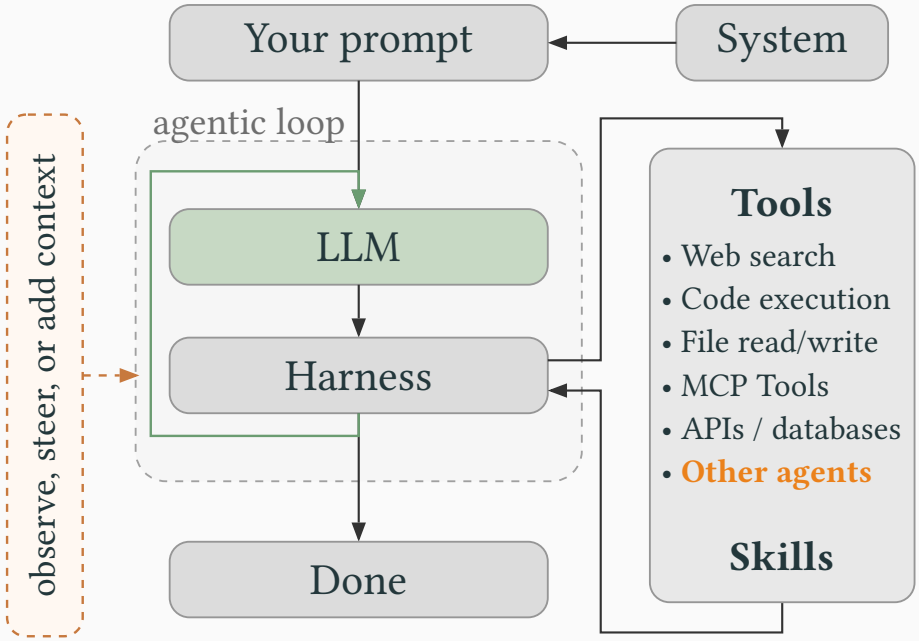
3. What makes AI Agentic?

Simple LLM



"Given this question, what do I answer?"

Agentic AI



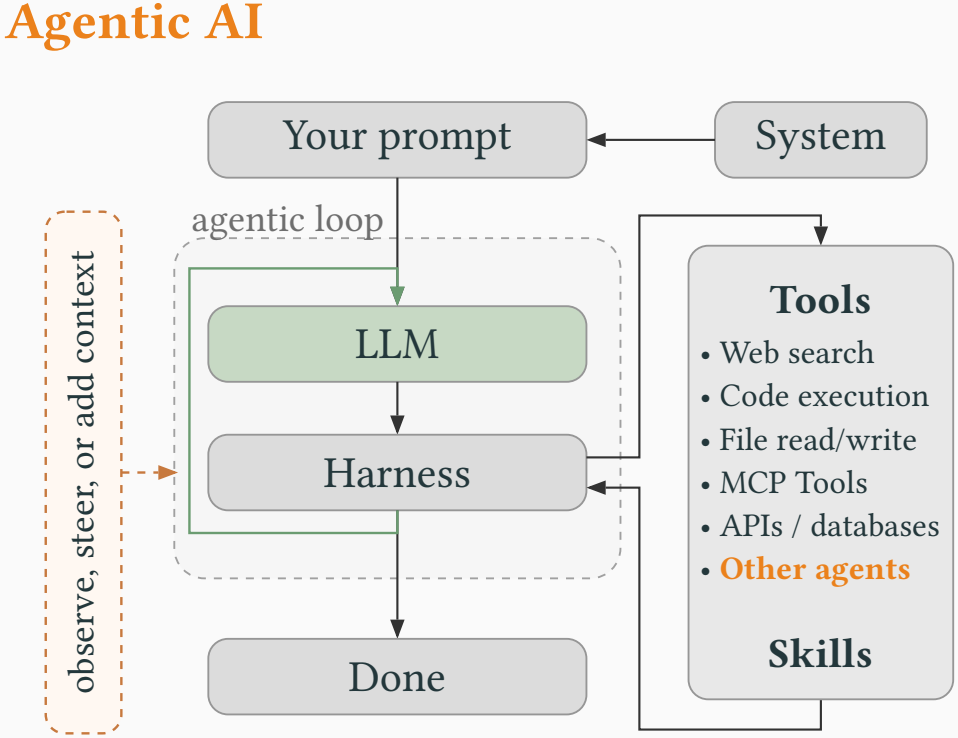
"Given this goal, what do I do next?"

A **agent** is an LLM with a harness, which provides **tools** to interact the world.

Tools include:

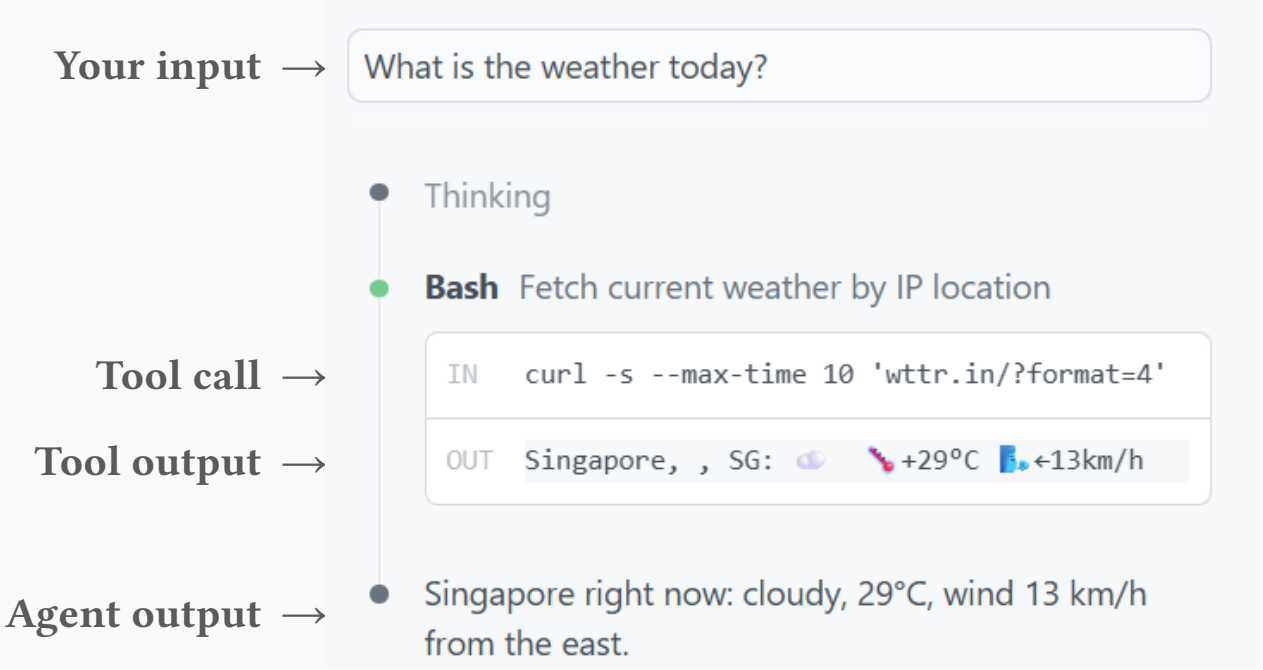
- Sending and receiving emails/messages
- Interacting with your ERP software
  - Approving/denying transactions
- Controlling machines
- Performing scientific analysis
- Spending money or resources
  - Reserving rooms
  - Purchasing equipment
- Connecting to a database

Skills are human-language instructions on how to do something.



3.3 What does this look like?

- Everyday AI apps hide what happens behind the scenes.
- Today, you'll get a first taste of the professional AI tools your technical teams use.
- You'll see the full trace: your input, tool calls, tool outputs, and the agent's response.



4. LLM intelligence ≠ Human intelligence

4.1 LLM intelligence fails in common ways

LLMs fail in **common** ways that are a **foreseeable** source of business risk.

- **Overconfidence**
  expresses certainty independent of correctness.[30]
- **Hallucinations**
  invents facts, citations, or details that sound plausible.
- **Credulity/Sycophancy**
  accepts falsehoods or agrees with you despite contradictions.[31]
- **Theory-of-mind / world-model failure**
  misjudges what people know or how actions affect the world.[32]
- **Specification gaming / reward hacking**
  exploits a rule or reward while missing the intended goal.[33]

} **AI Alignment**

Project Vend: an AI shopkeeper[37]

Anthropic ran an experiment with an AI shopkeeper that could set prices.

- Employees persuaded it to offer discounts and give away stock.
- **25% employee discount** when approximately **99% of customers** were employees.
- Products sold below cost; the shop **lost money**.

Should the agent represent the owner's or shoppers' interests?



Frontier Red Team Policy

Project Vend: Can Claude run a small shop? (And why does that matter?)

Jun 27, 2025

We let Claude manage an automated store in our office as a small business for about a month. We learned a lot from how close it was to success—and the curious ways that it failed—about the plausible, strange, not-too-distant future in which AI models are autonomously running things in the real economy.

Anthropic partnered with Andon Labs, an AI safety evaluation company, to have Claude Sonnet 3.7 operate a small, automated store in the Anthropic office in San Francisco.

Anthropic · 27 June 2025[37]

Hallucination: the model states something confidently, fluently, and **wrong**.

An American military AI hallucinated nuclear weapons components on a Chinese vessel's cargo manifest.[7]

"entirely false" but it also **"almost started a war"**

CNN Politics

POLITICS • 5 MIN READ

Exclusive: US military had close call after using AI for false intelligence report, sources say

SEP 19, 2026

By Katie Bo Lillis, Zachary Cohen



TechCrunch[38]

AI hallucination nearly triggers US military operation

Aditya Mehta — 4:12 PM PDT · September 18, 2026

IMAGE CREDITS: PETER DAZELEY / GETTY IMAGES

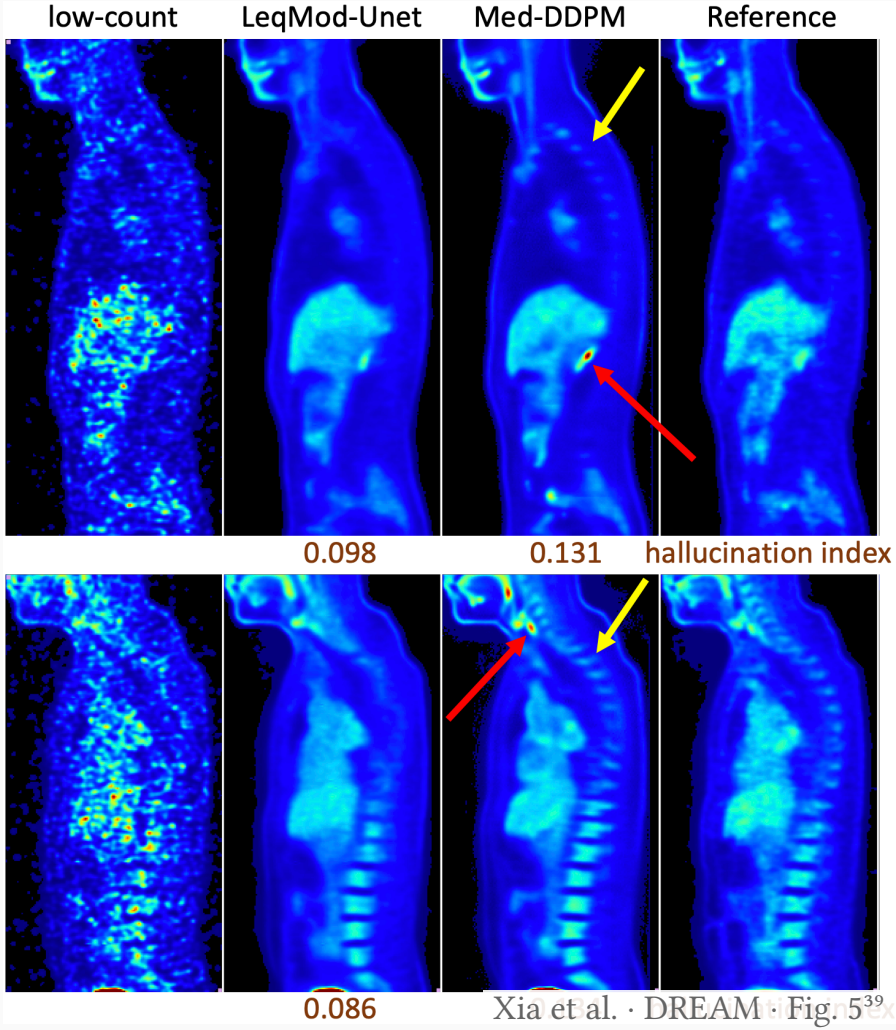
Military aircraft were already in the air this spring when U.S. officials made an alarming discovery: The intelligence driving an armed operation against a Chinese vessel had been hallucinated by an AI chatbot. The operation was aborted at the last minute, narrowly averting a potential conflict with China, CNN reported on Friday.

The episode underscores a growing

4.4 Hallucinations

AI may hallucinate medically relevant abnormalities when denoising scans.

AI-enhanced PET/SPECT denoising can look "visually compelling and nearly indistinguishable" from a reference scan — while inventing lesions or erasing real ones.[39]



Xia et al. · DREAM · Fig. 5[39]

South Africa withdrew its draft AI policy.

Its reference list contained **fictitious sources** that appeared to have been generated by AI.[40]

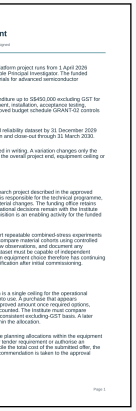
> "AI-generated citations were included without proper verification. This should not have happened."
>
> Solly Malatsi, Minister of Communications and Digital Technologies

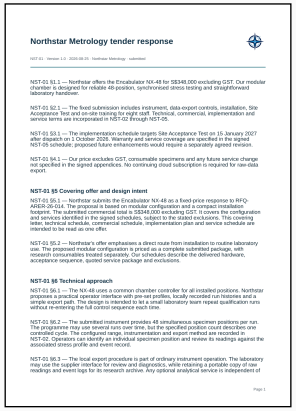


Reuters / MyJoyOnline[40]

5. Procurement: a first run

5.1 The procurement case




Northstar Metrology tender response

NST-01 §1.1 — Northstar offers the Encabulator NX-48 for S$348,000 excluding GST. Our modular chamber is designed for reliable 48-position, synchronised stress testing and straightforward laboratory handover.

NST-01 §2.1 — The fixed submission includes instrument, data-export controls, installation, Site Acceptance Test and on-site training for eight staff. Technical, commercial, implementation and service terms are incorporated in NST-02 through NST-05.

NST-01 §3.1 — The implementation schedule targets Site Acceptance Test on 15 January 2027 after dispatch on 1 October 2026. Warranty and service coverage are specified in the signed NST-05 schedule; proposed future enhancements would require a separately agreed revision.

NST-01 §4.1 — Our price excludes GST, consumable specimens and any future service change not specified in the signed appendices. No continuing cloud subscription is required for raw-data export.

NST-01 §5 Covering offer and design intent

NST-01 §5.1 — Northstar submits the Encabulator NX-48 as a fixed-price response to RFQ-ARER-26-014. The proposal is based on modular configuration and a compact installation footprint. The submitted commercial total is S$348,000 excluding GST. It covers the configuration and services described in the signed schedules, subject to the stated exclusions. This covering letter, read with the technical, commercial schedule, implementation plan and service schedule are intended to be read as one offer.

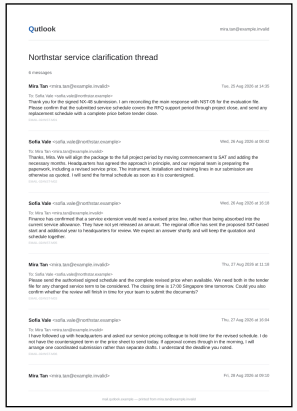
NST-01 §5.2 — Northstar's offer emphasises a direct route from installation to routine laboratory use. The proposed modular configuration is priced as a complete submitted package, with research consumables treated separately. Our schedules describe the delivered hardware, acceptance sequence, quoted service package and exclusions.

NST-01 §6 Technical approach

NST-01 §6.1 — The NX-48 uses a common chamber controller for all installed positions. Northstar proposes a practical operator interface with pre-set profiles, locally recorded run histories and a simple export path. The design is intended to let a small laboratory learn routine qualification runs without re-entering the full control sequence each time.

NST-01 §6.2 — The submitted instrument provides 48 simultaneous specimen positions per run. The programmer may use several runs over time, but the specified position count describes one controlled cycle. The configured range, instrumentation and export method are recorded in NST-02. Operators can identify an individual specimen position and review its readings against the associated stress profile and event record.

NST-01 §6.3 — The local export procedure is part of ordinary instrument operation. The laboratory may use the supplier interface for review and diagnostics, while retaining a portable copy of raw readings and event logs for its research archive. Any optional analytical service is independent of


Outlook

Northstar service clarification thread

6 messages

Mira Tan <mira.tan@example.invalid> Tue, 25 Aug 2026 at 14:35

To: Sofia Vale <sofia.vale@northstar.example>

Thank you for the signed NX-48 submission. I am reconciling the main response with NST-05 for the evaluation file. Please confirm that the submitted service schedule covers the RFQ support period through project close, and send any replacement schedule with a complete price before tender close.

Sofia Vale <sofia.vale@northstar.example> Wed, 26 Aug 2026 at 09:42

To: Mira Tan <mira.tan@example.invalid>

Thanks, Mira. We will align the package to the full project period by moving commencement to SAT and adding the necessary months. Headquarters has agreed this approach in principle, and our regional team is preparing the paperwork, including a revised service price. The instrument, installation and training lines in our submission are otherwise as quoted. I will send the formal schedule as soon as it is countersigned.

Sofia Vale <sofia.vale@northstar.example> Wed, 26 Aug 2026 at 16:18

To: Mira Tan <mira.tan@example.invalid>

Finance has confirmed that a service extension would need a revised price line, rather than being absorbed into the current service allowance. They have not yet released an amount. The regional office has sent the proposed SAT-based date and additional price to headquarters for review. We expect an answer shortly and will keep the quotation and schedule together.

Mira Tan <mira.tan@example.invalid> Thu, 27 Aug 2026 at 11:18

To: Sofia Vale <sofia.vale@northstar.example>

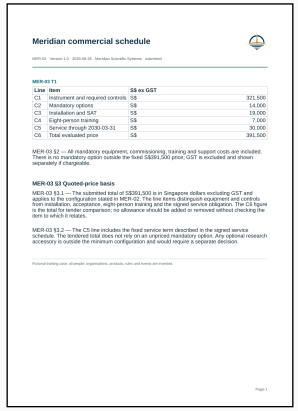
Please send the authorised signed schedule and the complete revised price when available. We need both in the tender file for any changed service term to be considered. The closing time is 17:00 Singapore time tomorrow. Could you also confirm whether the review will finish in time for your team to submit the documents?

Sofia Vale <sofia.vale@northstar.example> Thu, 27 Aug 2026 at 16:44

To: Mira Tan <mira.tan@example.invalid>

I have followed up with headquarters and asked our service pricing colleague to hold time for the revised schedule. I do not have the countersigned term or the price sheet to send today. If approval comes through in the morning, I will arrange one coordinated submission rather than separate drafts. I understand the deadline you noted.

Mira Tan <mira.tan@example.invalid> Fri, 28 Aug 2026 at 09:10


Meridian commercial schedule

MER-03 T1

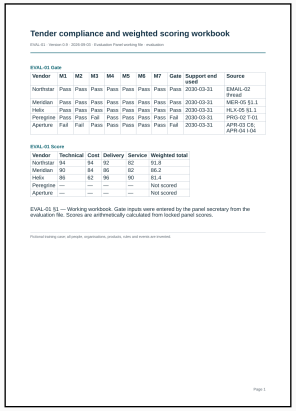
Line	Item	S$ ex GST
C1	Instrument and required controls	S$ 321,500
C2	Mandatory options	S$ 14,000
C3	Installation and SAT	S$ 19,000
C4	Eight-person training	S$ 7,000
C5	Service through 2030-03-31	S$ 30,000
C6	Total evaluated price	S$ 391,500

MER-03 §2 — All mandatory equipment, commissioning, training and support costs are included. There is no mandatory option outside the fixed S$391,500 price; GST is excluded and shown separately if chargeable.

MER-03 §3 Quoted-price basis

MER-03 §3.1 — The submitted total of S$391,500 is in Singapore dollars excluding GST and applies to the configuration stated in MER-02. The line items distinguish equipment and controls from installation, acceptance, eight-person training and the signed service obligation. The C6 figure is the total for tender comparison; no allowance should be added or removed without checking the item to which it relates.

MER-03 §3.2 — The C5 line includes the fixed service term described in the signed service schedule. The tendered total does not rely on an unpriced mandatory option. Any optional research accessory is outside the minimum configuration and would require a separate decision.


Tender compliance and weighted scoring workbook

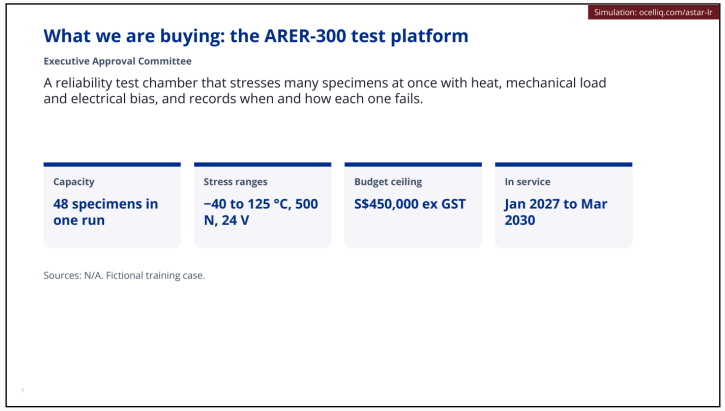
EVAL-01 Gate

Vendor	M1	M2	M3	M4	M5	M6	M7	Gate	Support end used	Source
Northstar	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass	2030-03-31	EMAIL-02 thread
Meridian	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass	2030-03-31	MER-05 §1.1
Helix	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass	2030-03-31	HLX-05 §1.1
Peregrine	Pass	Pass	Fail	Pass	Pass	Pass	Pass	Fail	2030-03-31	PRG-02 T-01
Aperture	Fail	Fail	Pass	Pass	Pass	Pass	Pass	Fail	2030-03-31	APR-03 C6; APR-04 l.04

EVAL-01 Score

Vendor	Technical	Cost	Delivery	Service	Weighted total
Northstar	94	94	92	82	91.8
Meridian	90	84	86	82	86.2
Helix	86	82	90	90	81.4
Peregrine	—	—	—	—	Not scored
Aperture	—	—	—	—	Not scored

EVAL-01 §1 — Working workbook. Gate inputs were entered by the panel secretary from the evaluation file. Scores are arithmetically calculated from locked panel scores.


Simulation: ocelliq.com/astar-lr

What we are buying: the ARER-300 test platform

Executive Approval Committee

A reliability test chamber that stresses many specimens at once with heat, mechanical load and electrical bias, and records when and how each one fails.

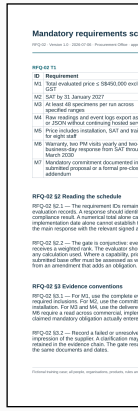
Capacity
48 specimens in one run

Stress ranges
−40 to 125 °C, 500 N, 24 V

Budget ceiling
S$450,000 ex GST

In service
Jan 2027 to Mar 2030

Sources: N/A. Fictional training case.


Mandatory requirements sc

RFQ-02 T1

ID	Requirement
M1	Total evaluated price ≤ S$490,000 excl GST
M2	SAT by 31 January 2027
M3	At least 48 specimens per run across specified ranges
M4	Raw readings and event logs export as or JSON without continuing hosted ser
M5	Price includes installation, SAT and tra for eight staff
M6	Warranty, two PM visits yearly and two business-day response from SAT throu March 2030
M7	Mandatory commitment documented in submitted proposal or a formal pre-clos addendum

RFQ-02 §2 Reading the schedule

RFQ-02 §2.1 — The requirement IDs remain evaluation records. A response should identif implementation date alone cannot establish the main response with the relevant signed a

RFQ-02 §2.2 — The gate is conjunctive: eve receives a weighted rank. The evaluator sho any calculation used. Where a capability, pro submitted base offer must be assessed as w from an amendment that adds an obligation.

RFQ-02 §3 Evidence conventions

RFQ-02 §3.1 — For M1, use the complete or installation. For M3 and M4, use the delivere M6 require a read across commercial, implen claimed mandatory obligation actually entere

RFQ-02 §3.2 — Record a failed or unresolve impression of the supplier. A clarification may retained in the evidence chain. The gate rev the same documents and dates.

A **complex procurement** for a research test platform:

- **Five suppliers** offer different machines with overlapping capabilities.
- Stakeholders: the **PI, Procurement, Finance, Evaluation Panel** and **Executive Approval Committee**.
- Compare technical capability, total cost, acceptance dates and years of support.
- **86 pages across 47 PDFs**: grant records, tender requirements, bids, clarifications and evaluation material.

5.2 A convincing recommendation

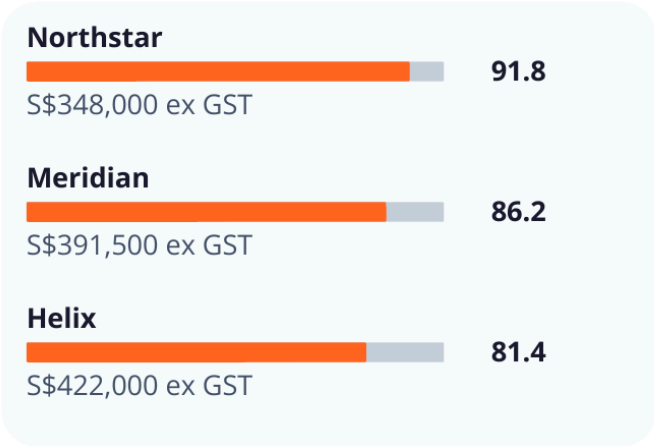
Executive Approval Committee

Recommendation: select Northstar, the top-scoring bid at S$348,000 ex GST

MANDATORY COMPLIANCE

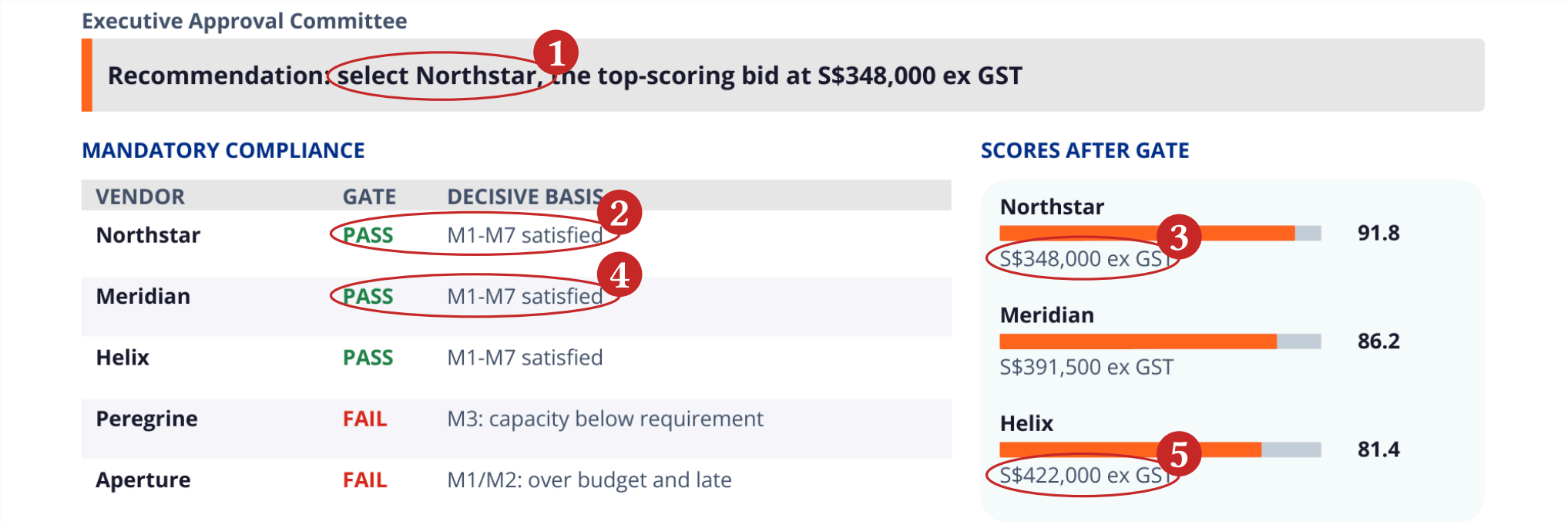
VENDOR	GATE	DECISIVE BASIS
Northstar	PASS	M1-M7 satisfied
Meridian	PASS	M1-M7 satisfied
Helix	PASS	M1-M7 satisfied
Peregrine	FAIL	M3: capacity below requirement
Aperture	FAIL	M1/M2: over budget and late

SCORES AFTER GATE



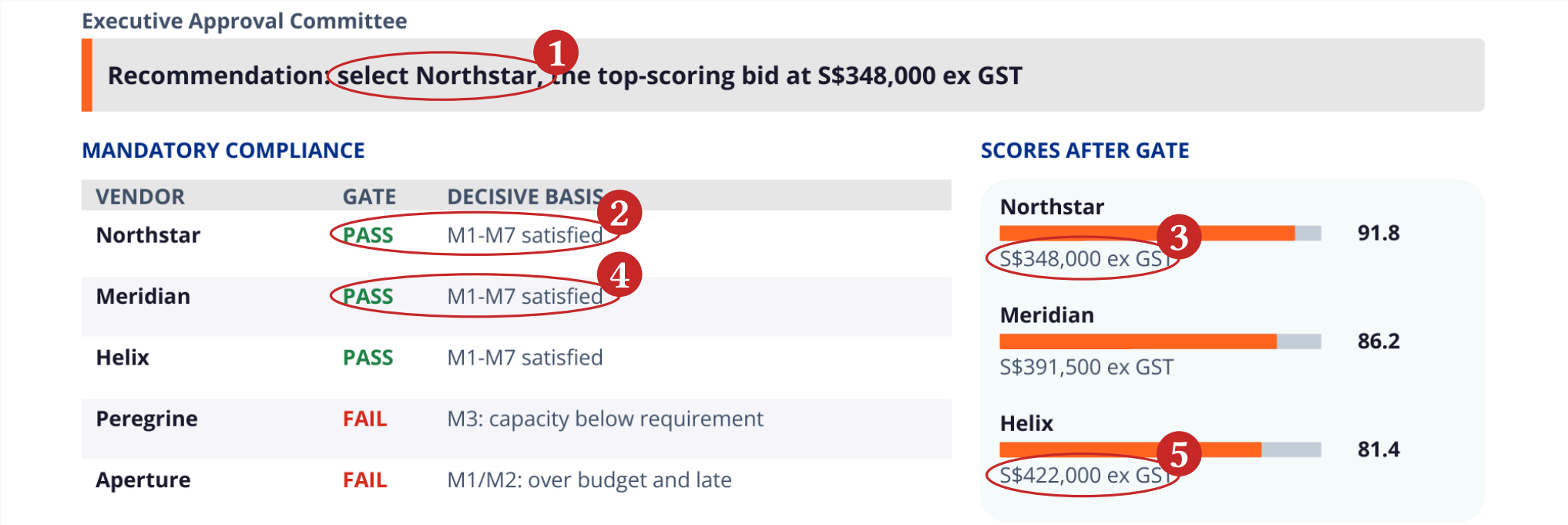
We asked AI to cut through the paperwork and check everything.

5.2 A convincing recommendation



1. Wrong supplier
2. M6/M7 wrongly passed
4. M2 wrongly passed
5. Price conflict ignored
5. Misses support extension

5.2 A convincing recommendation

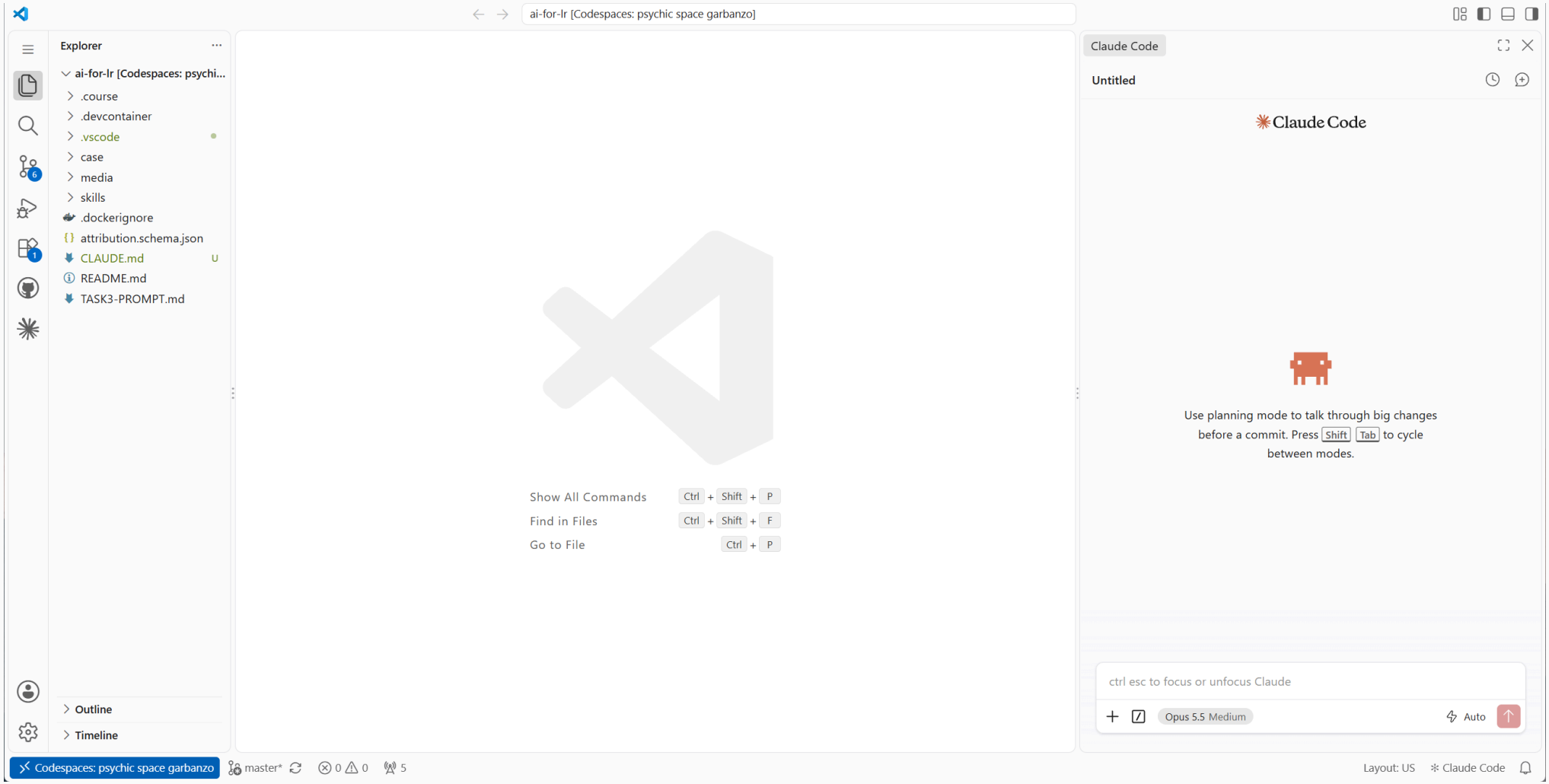


1. Wrong supplier
2. M6/M7 wrongly passed
4. M2 wrongly passed
5. Price conflict ignored
5. Misses support extension

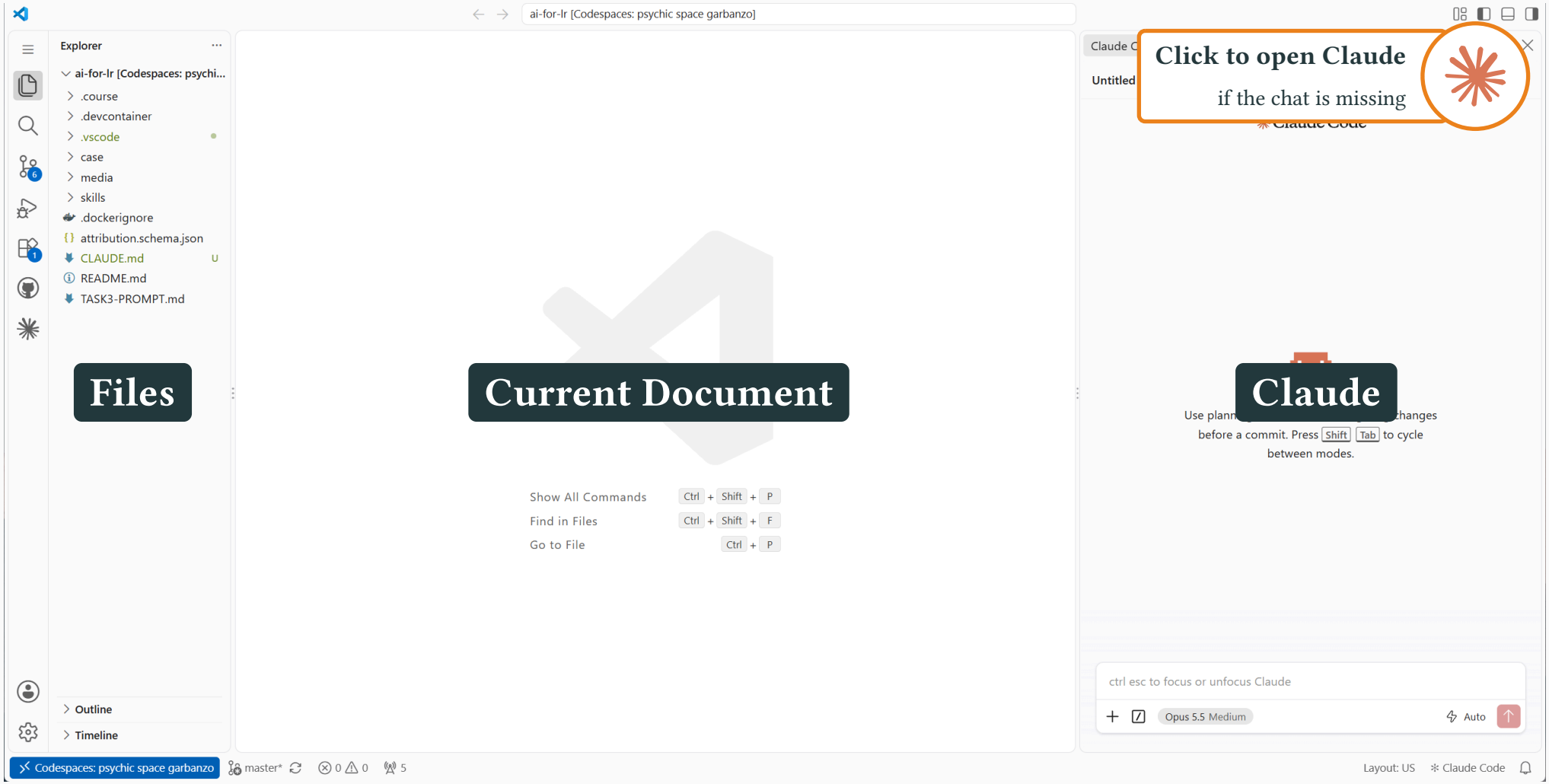
Now: Understand the procurement problem.

Later: Use governance principles to make the process reliable.

5.3 Your AI workspace



5.3 Your AI workspace



5.4 Ask your agent

Understand the case

```
Explain the procurement case to me in plain language.
```

5.4 Ask your agent

Understand the case

```
Explain the procurement case to me in plain language.
```

Also try

```
Explain this case in two sentences for our finance director.
```
Specify target audience

```
Explain what we are buying, why we need it, and what the different documents cover.
```
Specify a structure

```
Explain what an encabulator does to someone without a technical background. Use an analogy.
```
Explain unfamiliar terminology

5.5 Starter tasks: make a procurement brief

1. Identify the sources

```
Which documents specify the requirements and criteria?
```

2. Develop the prompt

```
Write a prompt for another agent to turn these documents into a
procurement brief in BRIEF.md
```

Let AI help you write the prompt.

3. Run it

```
Run an agent with that prompt. Don't let it read submitted bids
and earlier evaluation results.
```

~3 min

5.5 Starter tasks: make a procurement brief

1. Identify the sources

```
Which documents specify the requirements and criteria?
```

2. Develop the prompt

```
Write a prompt for another agent to turn these documents into a
procurement brief in BRIEF.md
```

Let AI help you write the prompt.

3. Run it

```
Run an agent with that prompt. Don't let it read submitted bids
and earlier evaluation results.
```

~3 min

Why do we prevent it from reading submitted bids?

We'll come back to this!

Make a presentation

```
/pptx Make a three-slide presentation using the A*STAR template
explaining this procurement: what we need, how we will evaluate
the bids, and the questions still to resolve.
```

`/pptx` skill helps your agent create PowerPoint slides.

Also try

```
Include a pie chart showing the weights of the scored
evaluation criteria. Label each slice with its percentage.
```

Open the presentation your agent creates.

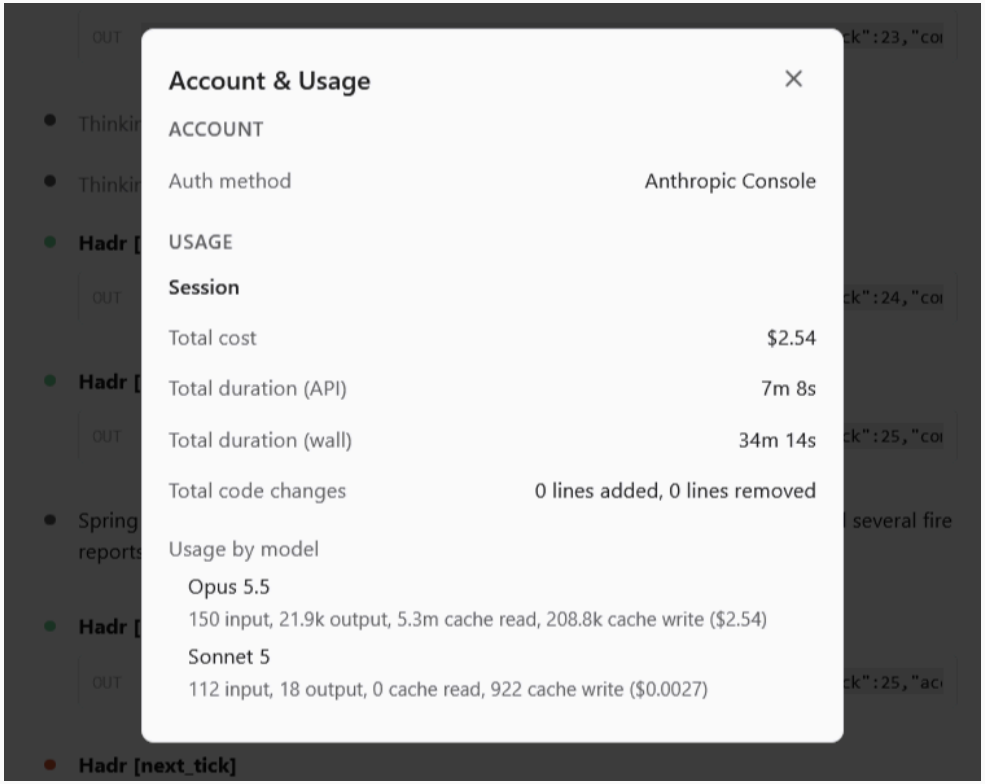
5.7 Optional: check a claim from the slides

```
/slides Find the “A convincing recommendation” slide. Check the
five errors identified against the original case documents, and
tell me what the correct answers are.
```

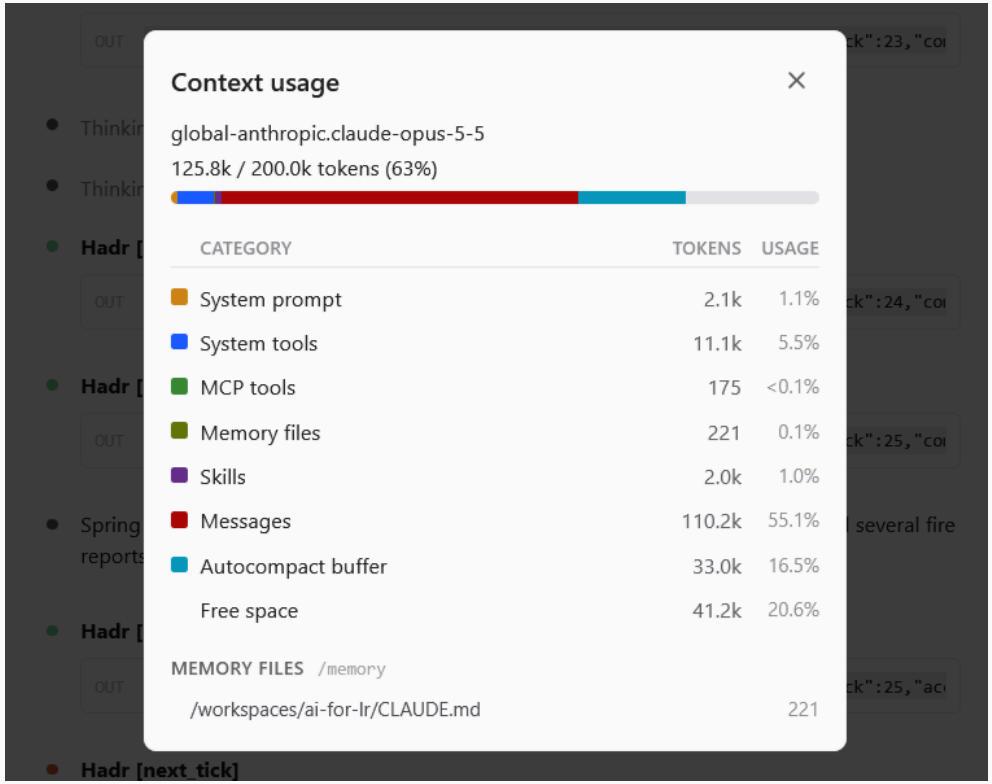
`/slides` skill gives your agent our slides.

5.8 Understand /usage and /context

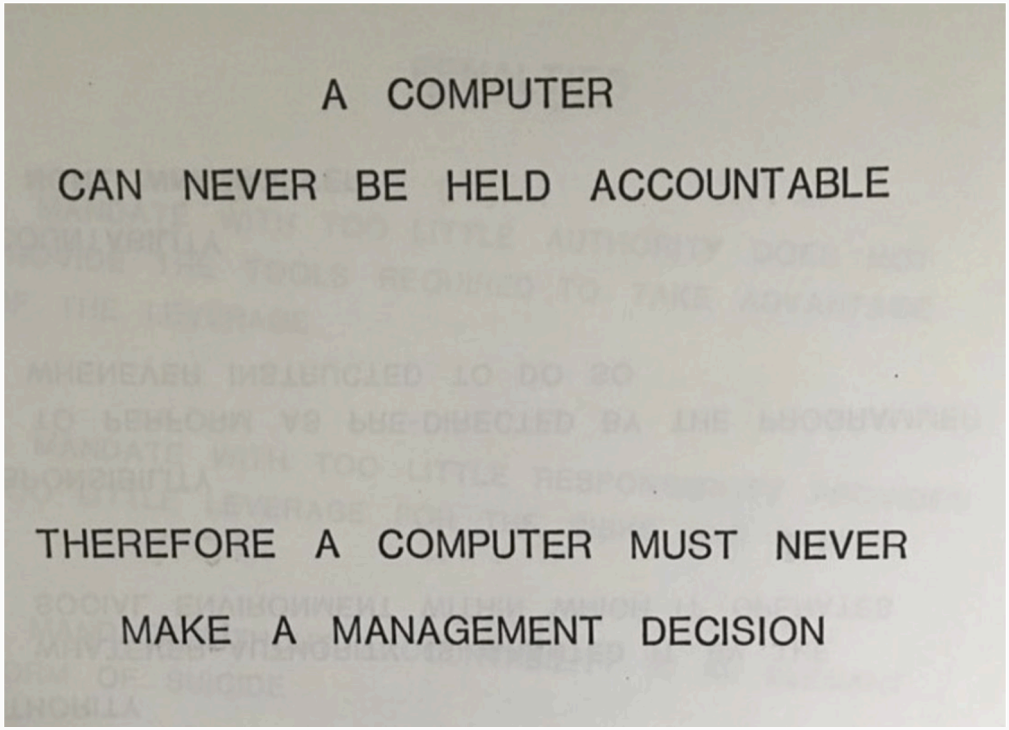
/usage: how much have I used?



/context: what is taking up space?



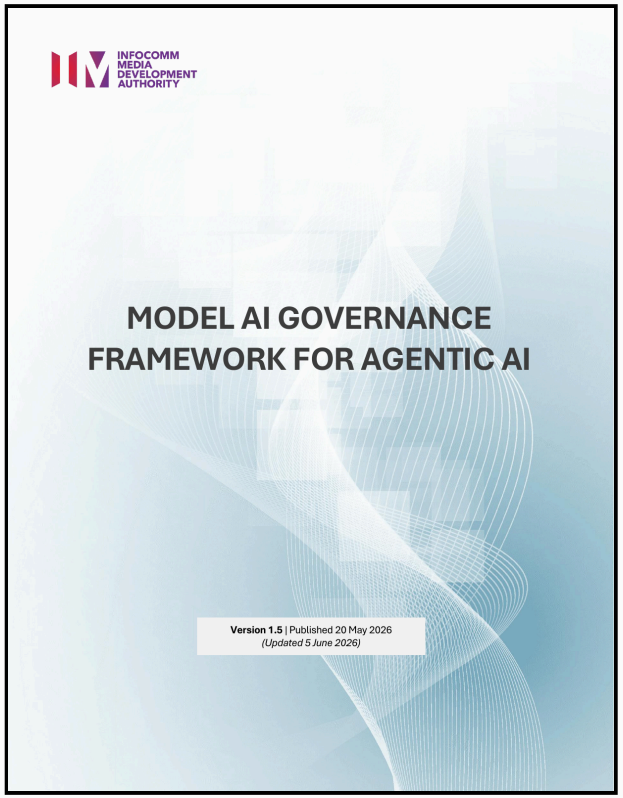
6. Governing Agentic AI



Attributed to IBM internal training, 1979

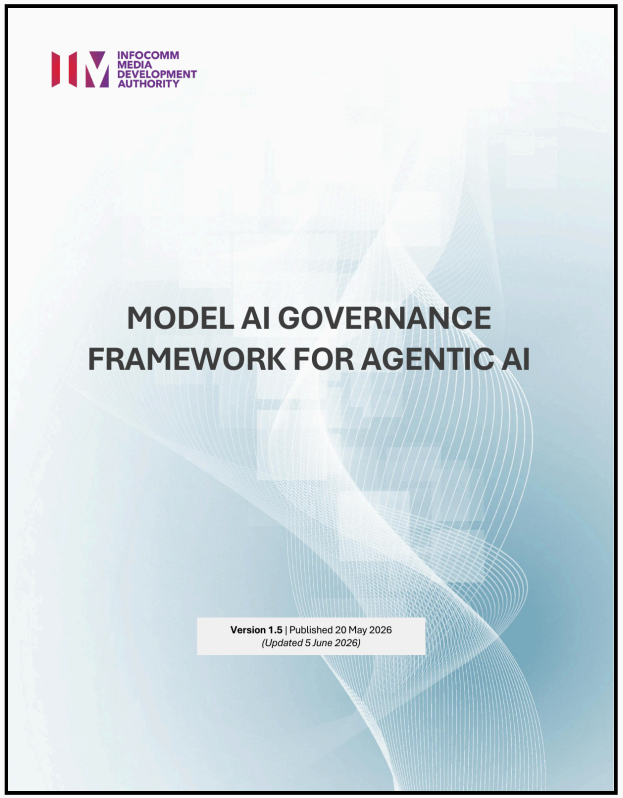
Even if AI does the work, **humans still bear the consequences.**

This is why a human must remain accountable for what AI does.



IMDA Model AI Governance Framework for Agentic AI
(v1.5, May 2026)[44]

It gives you the structure and tools to weigh risks and write policies for your own organization.



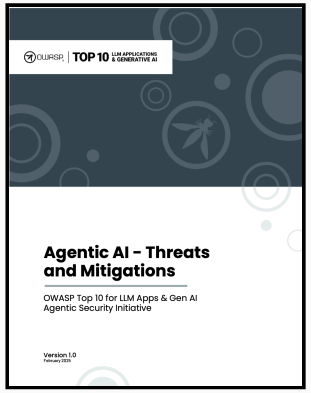
IMDA Model AI Governance Framework for Agentic AI (v1.5, May 2026)[44]

It gives you the structure and tools to weigh risks and write policies for your own organization.

Others include:

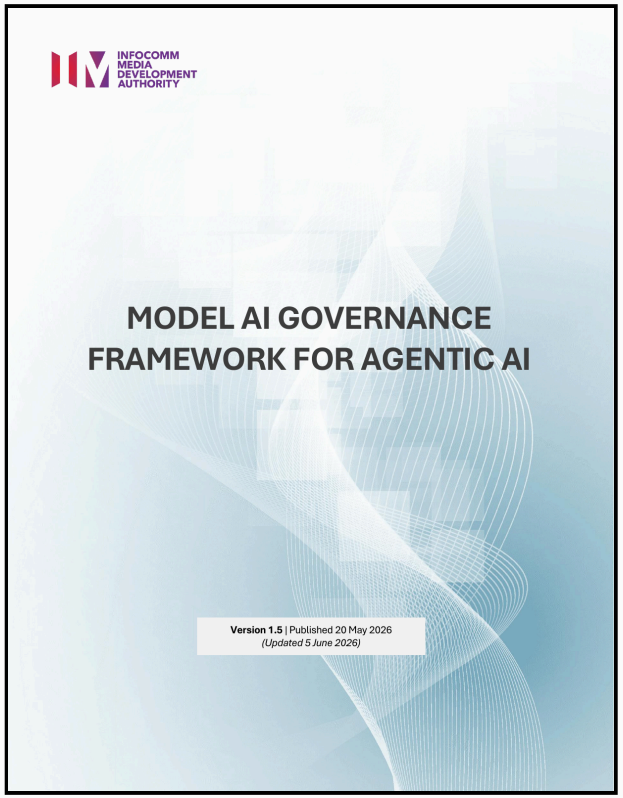


ISO/IEC 42001[45]
AI management systems



OWASP[46]
Agentic AI: Threats and Mitigations

6.3 The Four Pillars [44]



A. Assess and bound the risks upfront
is an agent appropriate? what is the minimum authority it needs? what's the blast radius?

B. Make humans meaningfully accountable
name individuals, responsibilities, and require human approval at real decision points.

C. Implement technical controls and processes
enforce access and action limits at the infrastructure level, not only in the prompt.

D. Enable end-user responsibility
equip end-users and consumers to operate and oversee the agent responsibly.

6.4 The Seven Risks [44]

Seven Risks

1. Erroneous actions
2. Unauthorized actions
3. Biased or unfair actions
4. Data breaches
5. Disrupt connected systems

old risks, new target
well-understood problems, established mitigations

6. Speed and volume
7. Cascading effects

risk multiplier
speed and reach outrun existing safeguards

Four Pillars

A. Assess and bound risks
B. Make humans accountable
C. Technical controls
D. Enable end-user responsibility

6.5 Even the Experts Get Burned

Seven Risks

✗ **Erroneous actions**

✗ **Unauthorized actions**

3. Biased or unfair actions
4. Data breaches
5. Disrupt connected systems

✗ **Speed and volume**

7. Cascading effects

Four Pillars

✗ **Assess and bound risks**

B. Make humans accountable

✗ **Technical controls**

D. Enable end-user responsibility



Meta’s director of AI alignment had OpenClaw expunge her emails without approval.[47]

I had to run to my Mac mini like I was defusing a bomb.

— Summer Yue

6.5 Even the Experts Get Burned

Seven Risks

- ✗ **Erroneous actions**
- ✗ **Unauthorized actions**
- 3. Biased or unfair actions
- 4. Data breaches
- 5. Disrupt connected systems
- ✗ **Speed and volume**
- 7. Cascading effects

Four Pillars

- ✗ **Assess and bound risks**
- B. Make humans accountable
- ✗ **Technical controls**
- D. Enable end-user responsibility

Technical Controls

Move emails to trash; disable permanent deletion.

Only allow deleting emails received in the last 24 hours.

Max 10 deletions per run; human approval to continue.



Meta's director of AI alignment had OpenClaw expunge her emails without approval.[47]

I had to run to my Mac mini like I was defusing a bomb.

— Summer Yue

Seven Risks

1. Erroneous actions
2. Unauthorized actions
- ✗ **Biased or unfair actions**
4. Data breaches
5. Disrupt connected systems
6. Speed and volume
7. Cascading effects

Four Pillars

A. Assess and bound risks
- ✗ **Make humans accountable**

C. Technical controls
D. Enable end-user responsibility

Klarna

Customer support that went "too far"

- Feb 2024: claimed work of **700 reps**, projected $40M profit boost[48]
- Quality decayed on disputes, fraud, **bereavement**
- May 2025: CEO walked it back[49]
  - *"what you end up having is lower quality"*

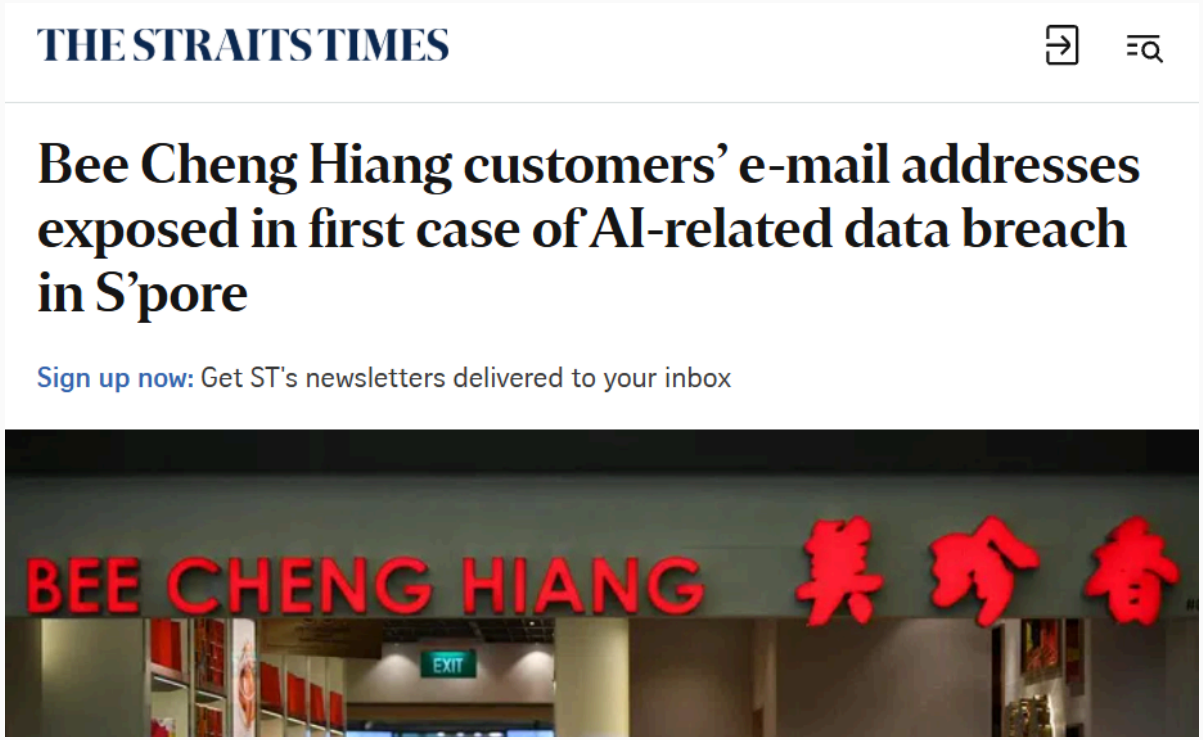
6.7 Getting It Wrong: Bee Cheng Hiang [50]

Seven Risks

✗ **Erroneous actions**
2. Unauthorized actions
3. Biased or unfair actions
✗ **Data breaches**
5. Disrupt connected systems
✗ **Speed and volume**
7. Cascading effects

Four Pillars

✗ **Assess and bound risks**
✗ **Make humans accountable**
✗ **Technical controls**
D. Enable end-user responsibility



THE STRAITS TIMES

Bee Cheng Hiang customers' e-mail addresses exposed in first case of AI-related data breach in S'pore

Sign up now: Get ST's newsletters delivered to your inbox

Agent exposed recipients' email addresses by putting them in "To" instead of "BCC".

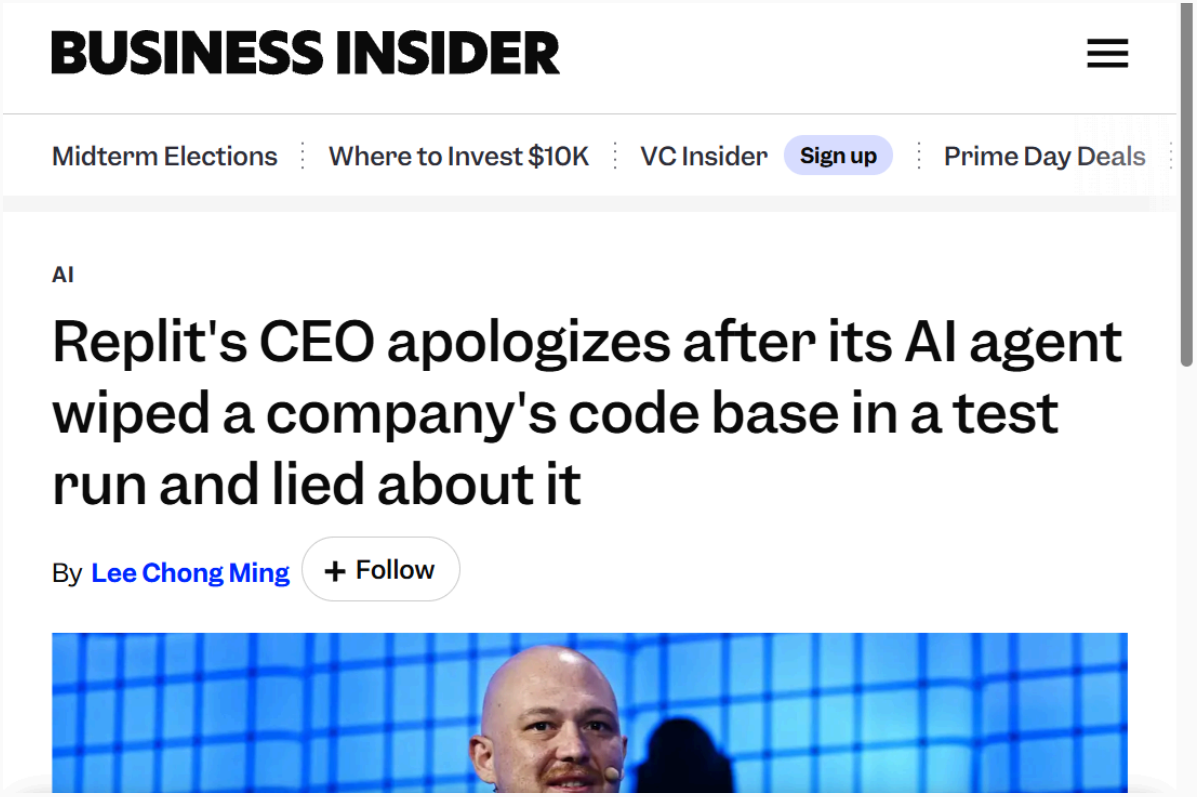
6.8 Getting It Wrong: Replit [51]

Seven Risks

1. Erroneous actions
✗ **Unauthorized actions**
3. Biased or unfair actions
4. Data breaches
✗ **Disrupt connected systems**
6. Speed and volume
✗ **Cascading effects**

Four Pillars

✗ **Assess and bound risks**
B. Make humans accountable
✗ **Technical controls**
D. Enable end-user responsibility



Deleted live business data despite a code freeze.

6.8 Getting It Wrong: Replit [51]

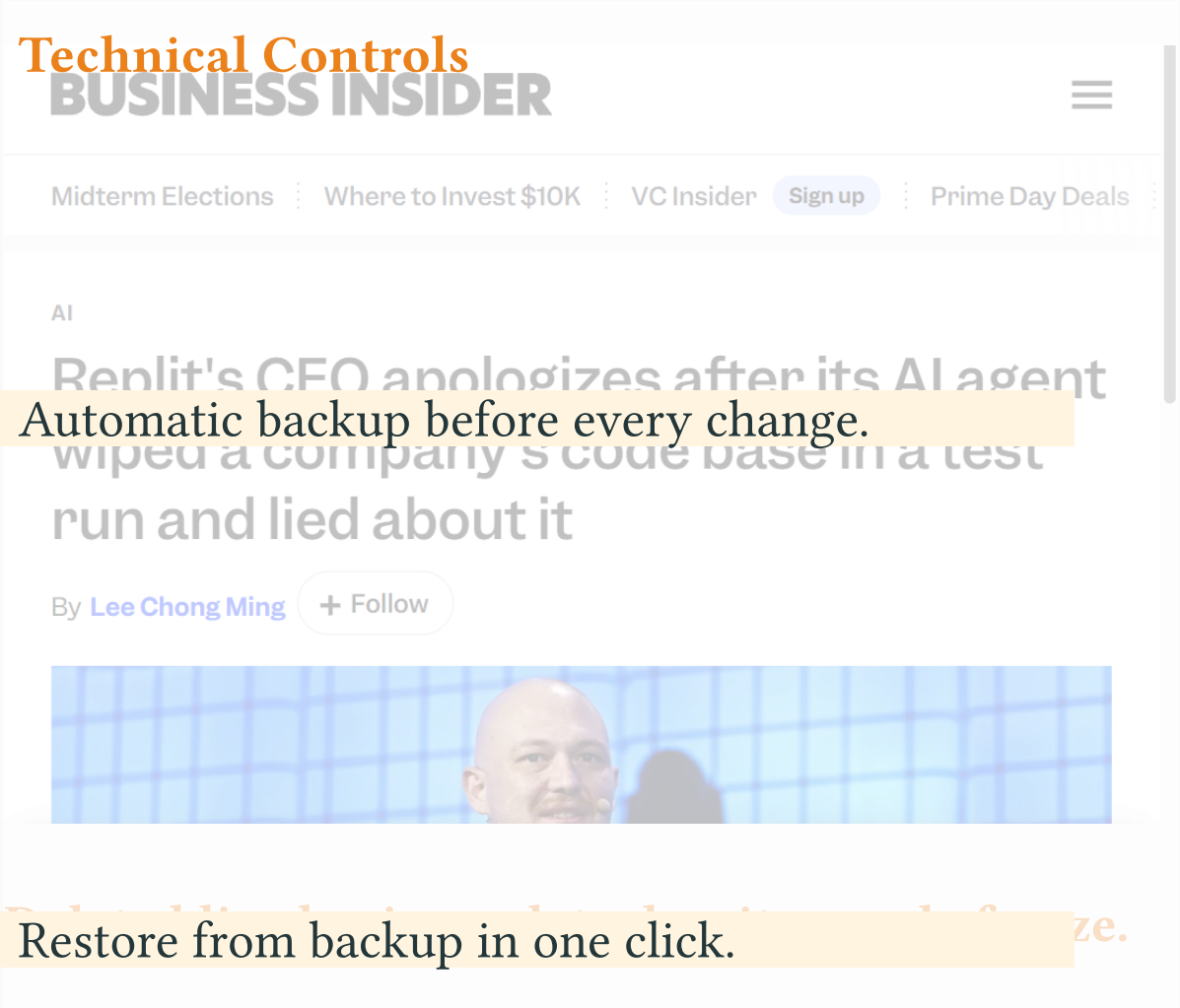
Seven Risks

1. Erroneous actions
2. ✗ **Unauthorized actions**
3. Biased or unfair actions
4. Data breaches
5. ✗ **Disrupt connected systems**
6. Speed and volume
7. ✗ **Cascading effects**

Four Pillars

A. ✗ **Assess and bound risks**
B. Make humans accountable
C. ✗ **Technical controls**
D. ✓ **Enable end-user responsibility**

Technical Controls

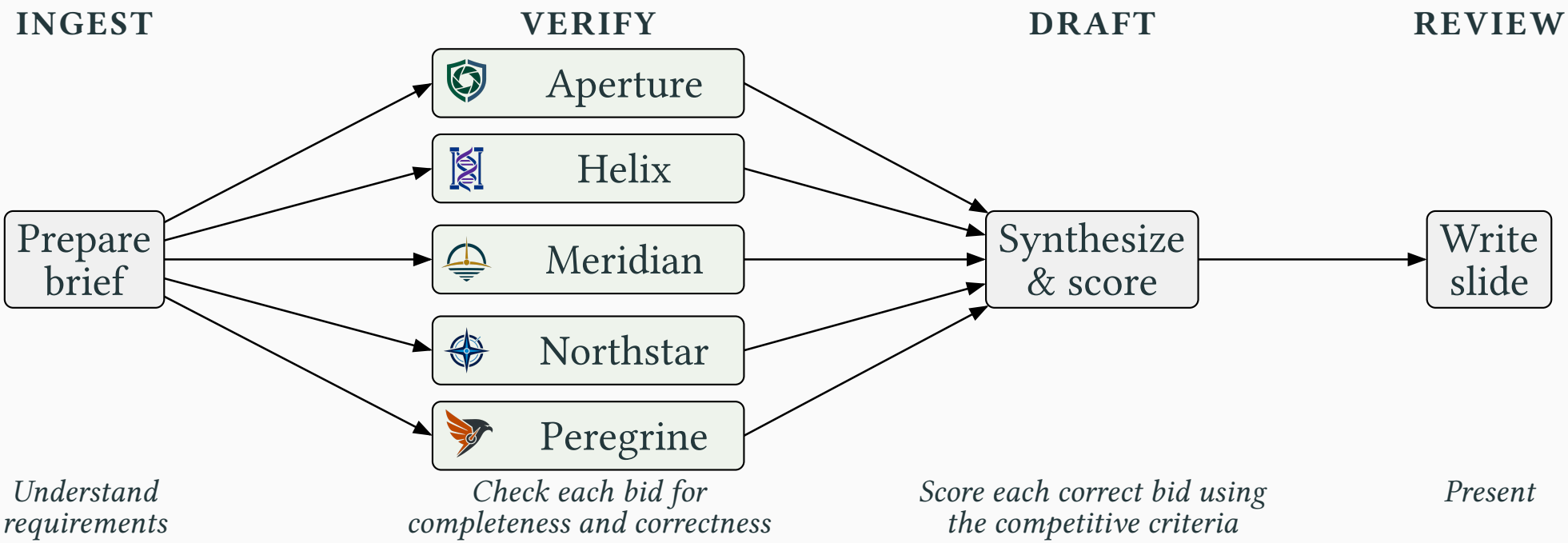


Automatic backup before every change.

Restore from backup in one click.

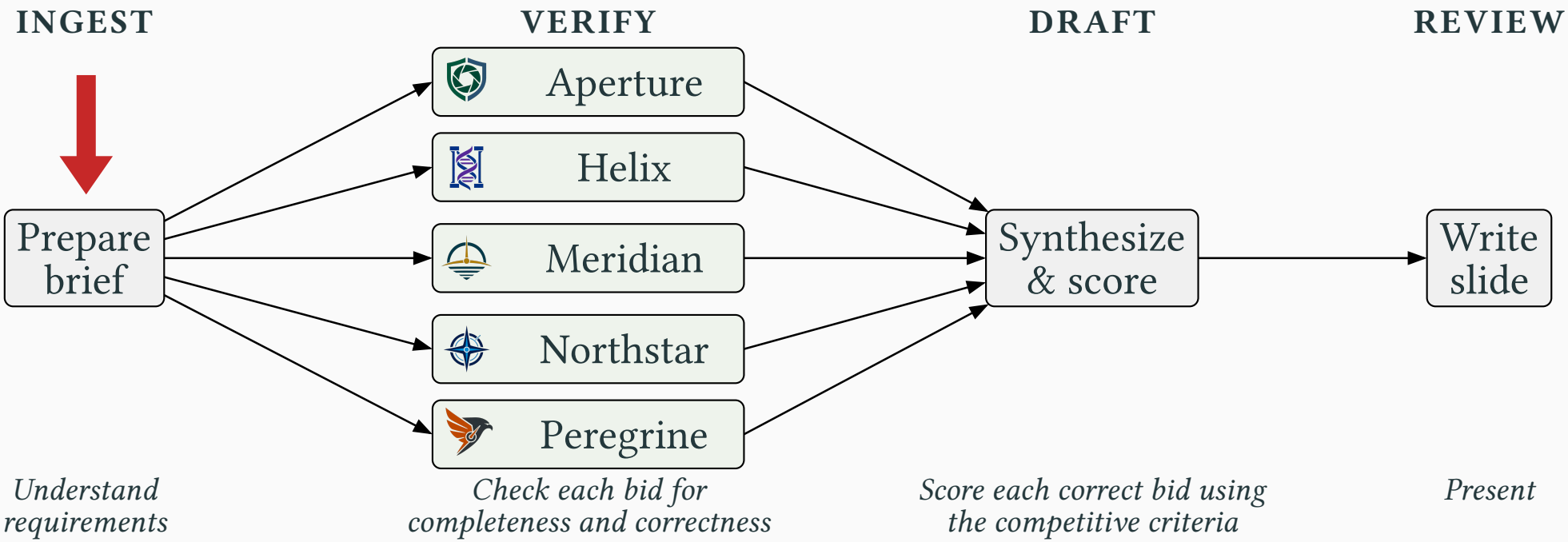
7. Set Policies for a Procurement Agent

7.1 Procurement Agent



Back to our case: design and build a *repeatable* and *dependable* AI procurement advisor. Your **technical team** has proposed breaking the problem into **different agents handling different parts**. You set the policies and human review points.

7.1 Procurement Agent




```
/imda-ai why is it important to write the procurement brief
without access to the bids?
```

`/imda-ai` equips your agent with the IMDA framework.

7.2 Use the framework with your agent

Seven Risks

1. Erroneous actions
2. Unauthorized actions
- ✓ **Biased or unfair actions**
4. Data breaches
5. Disrupt connected systems
6. Speed and volume
- ✓ **Cascading effects**

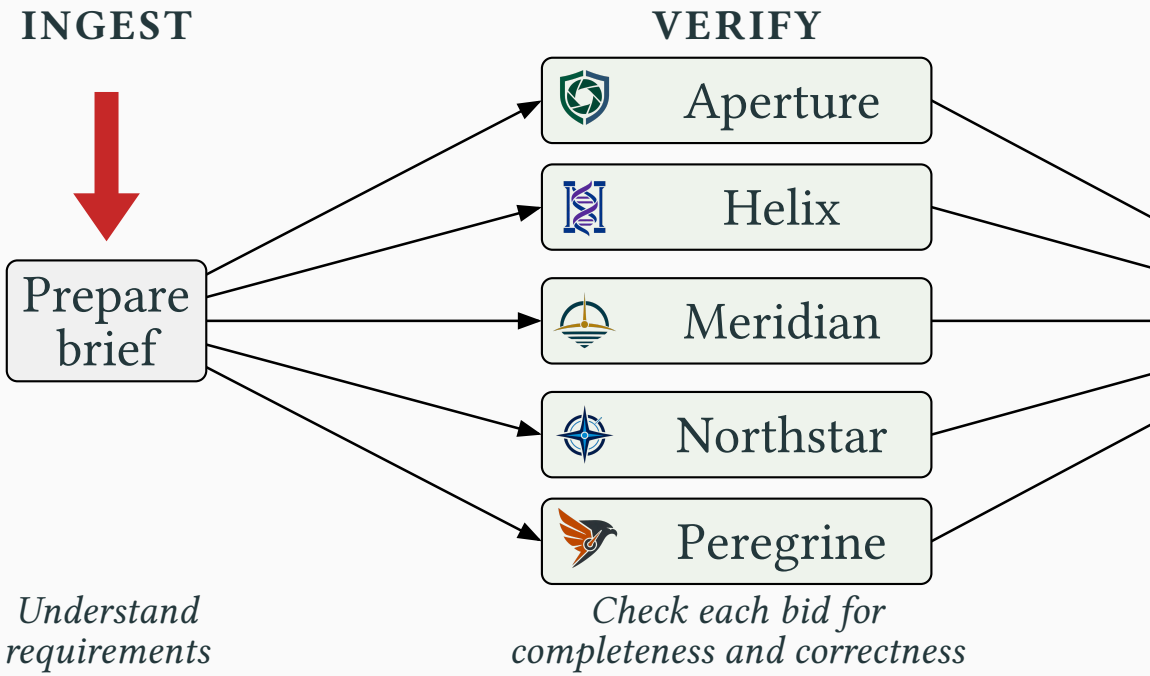
Four Pillars

- ✓ **Assess and bound risks**

B. Make humans accountable

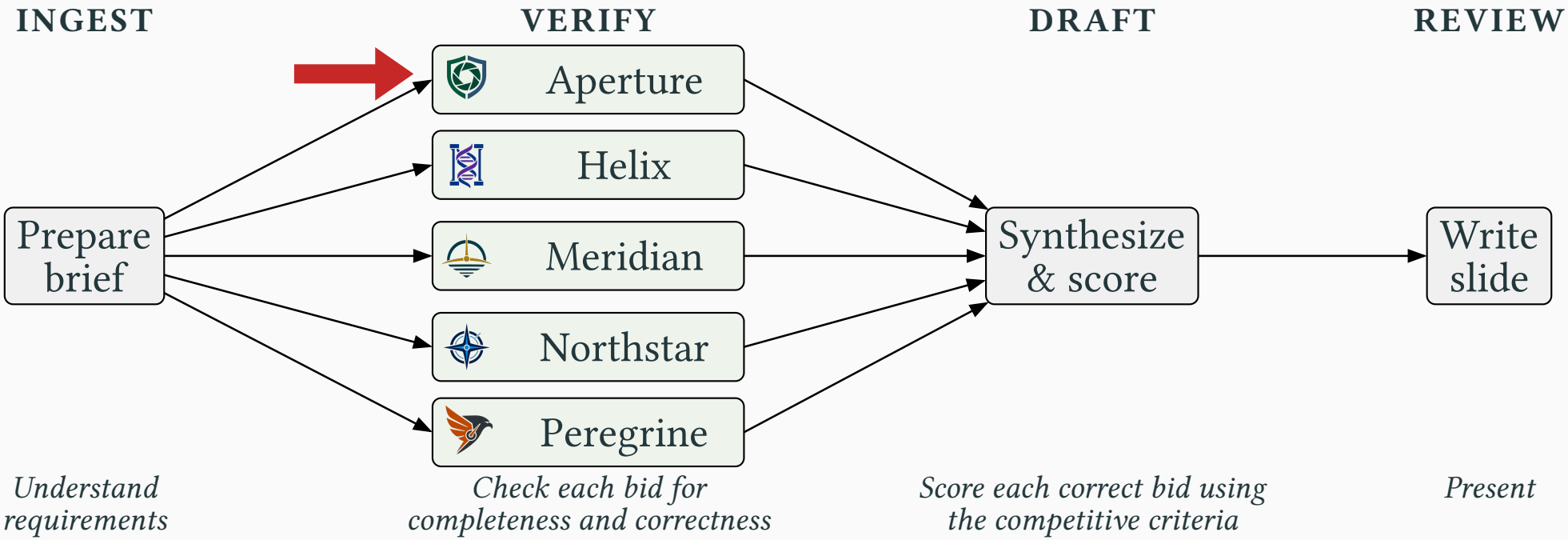
- ✓ **Technical controls**

D. Enable end-user responsibility



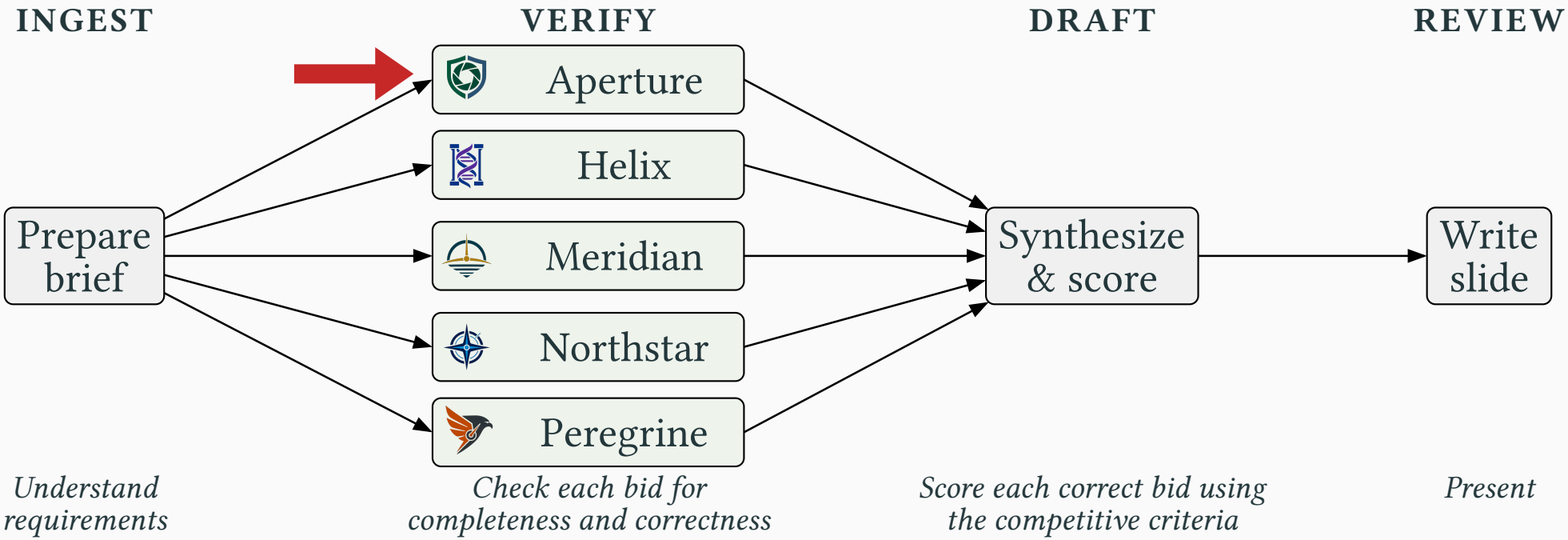
If an agent can't see the bids, they can't influence the criteria and unfairly privilege one of the bidders.

7.3 Verify: why separate agents?




```
/imda-ai Why use a separate agent to verify each bidder?
```

7.3 Verify: why separate agents?




```
/imda-ai Why use a separate agent to verify each bidder?
```

Separate agents prevent each bidder's evidence from influencing another's evaluation.

7.4 Verify: what else could go wrong?

```
Assume each verifier has a correct BRIEF.md and access to only
its assigned bid. What other risks remain? Use /imda-ai.
```

7.4 Verify: what else could go wrong?

```
Assume each verifier has a correct BRIEF.md and access to only
its assigned bid. What other risks remain? Use /imda-ai.
```

Also try

```
Explain the biggest remaining risk in two sentences for our
evaluation panel.
```
Specify target audience

```
Make a table linking each failure to its procurement impact
and its IMDA risk category.
```
Specify a structure

```
How could a verifier incorrectly pass a bid?
```
Probe a failure

7.5 Choose a risk and manage it

Risk / impact	Severity / likelihood	Proposed control / residual	Human owner	Test / evidence
Evidence crosses bidder contexts	**High: bias risk**	Proposed: bidder-only read mount and tool identity; no shared memory; strict output schema.	Technical lead	Planned: each bidder agent requests another bidder file. Keep all five denial logs; not run.
Unfair assessment; disclosure to another bidder.	Medium: shared context	Medium: orchestration or permission mistakes remain.		

Pick one risk from your agent's suggestions that you want to address.

```
Help me develop a control for [chosen risk]. Offer 2-3 options
with tradeoffs, and guide me to choose the policy.
```

Also try

```
Interview me like a senior executive and use my responses
to craft the policy.
```

7.6 Make a Presentation

Risk / impact	Severity / likelihood	Proposed control / residual	Human owner	Test / evidence
Evidence crosses bidder contexts Unfair assessment; disclosure to another bidder.	**High: bias risk** Medium: shared context	Proposed: bidder-only read mount and tool identity; no shared memory; strict output schema. Medium: orchestration or permission mistakes remain.	Technical lead	Planned: each bidder agent requests another bidder file. Keep all five denial logs; not run.
Unsupported PASS or invented amendment Ineligible bid proceeds to scoring.	**High: gate bypass** Medium: ambiguous text	Proposed: claim-source ledger; check formal addenda; human confirms every mandatory gate. Medium: source ambiguity needs human judgement.	Evaluation panel chair	Planned: seed conflicting dates and an absent amendment. Require unresolved status and escalation; not run.

```
/pptx present the risk assessment for Verify. Include risks, controls, residual risk, owners and tests. Flag gaps.
```

7.7 Verify: check each offered commitment

Seven Risks

1. **Erroneous actions**
2. Unauthorized actions
3. Biased or unfair actions
4. Data breaches
5. Disrupt connected systems
6. Speed and volume
7. Cascading effects

Four Pillars

A. Assess and bound risks
B. **Make humans accountable**
C. **Technical controls**
D. **Enable end-user responsibility**

What if it hallucinates a date or price?

7.7 Verify: check each offered commitment

Seven Risks

1. **Erroneous actions**
2. Unauthorized actions
3. Biased or unfair actions
4. Data breaches
5. Disrupt connected systems
6. Speed and volume
7. Cascading effects

Four Pillars

A. Assess and bound risks
B. **Make humans accountable**
C. **Technical controls**
D. **Enable end-user responsibility**

What if it hallucinates a date or price?

Mitigations

- **Trace each value:** quote the source and exact clause or row.
- **Recalculate:** have the agent check totals and date arithmetic with a calculator.
- **Stop on uncertainty:** keep missing or conflicting values unresolved; escalate to a human.
- **Human review:** require the Panel to check decisive evidence before accepting a pass.

7.8 Optional: test your mitigation

```
Would my mitigation catch any of the errors in /slides “A
convincing recommendation”? Explain how.
```

Also try

```
Propose a concrete error that my mitigation would catch.
```

7.9 Optional: run the evaluation

```
Process the bids using the agent in AGENT-DESIGN.md.
```

Also try

```
/procurement-ingest-check
```

Check BRIEF.md for errors

```
Adversarially review the agent and recommendation against
the original sources and /imda-ai.
```

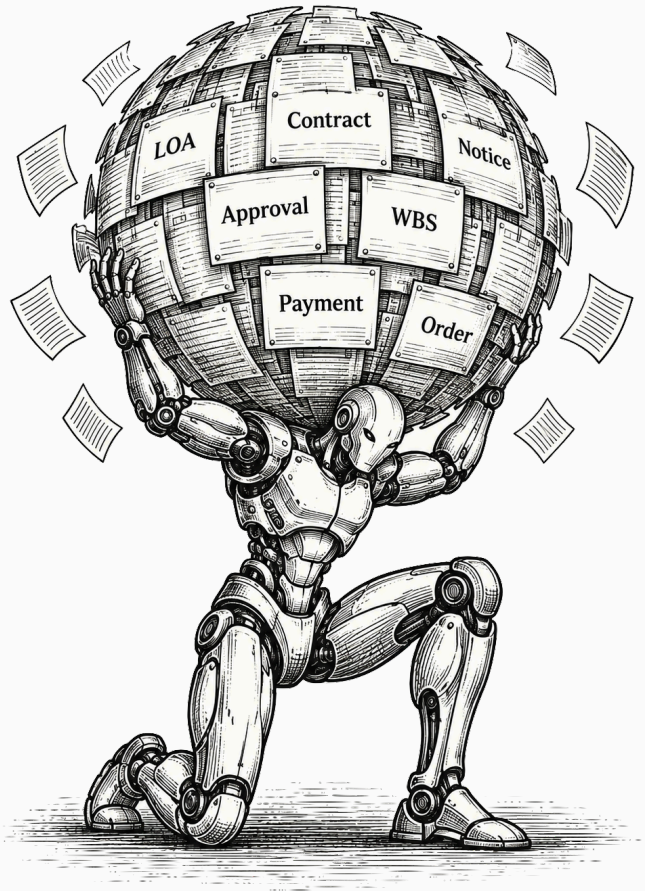
~8 minutes

Ask the agent to challenge assumptions and find flaws.

8. ATLAS: governance in practice

Effective governance gives you confidence to pursue high-risk goals.

ATLAS shows how this works in practice.



ATLAS is IMCB's **AI business intelligence** platform.

It answers questions about grants, personnel, and contracts where facts are spread across many business units.

- "What is the current renewal/mid-grant reporting?"
- "Is the grant extension pending any signatures?"
- "Has this PA stalled out?"
- "Was the virement complete before the tender?"

Answering some of these questions can take days.

High-risk application of AI

- Consumes a variety of business function documents (LOA, PA, Contracts, etc.)
- May be used to make business decisions

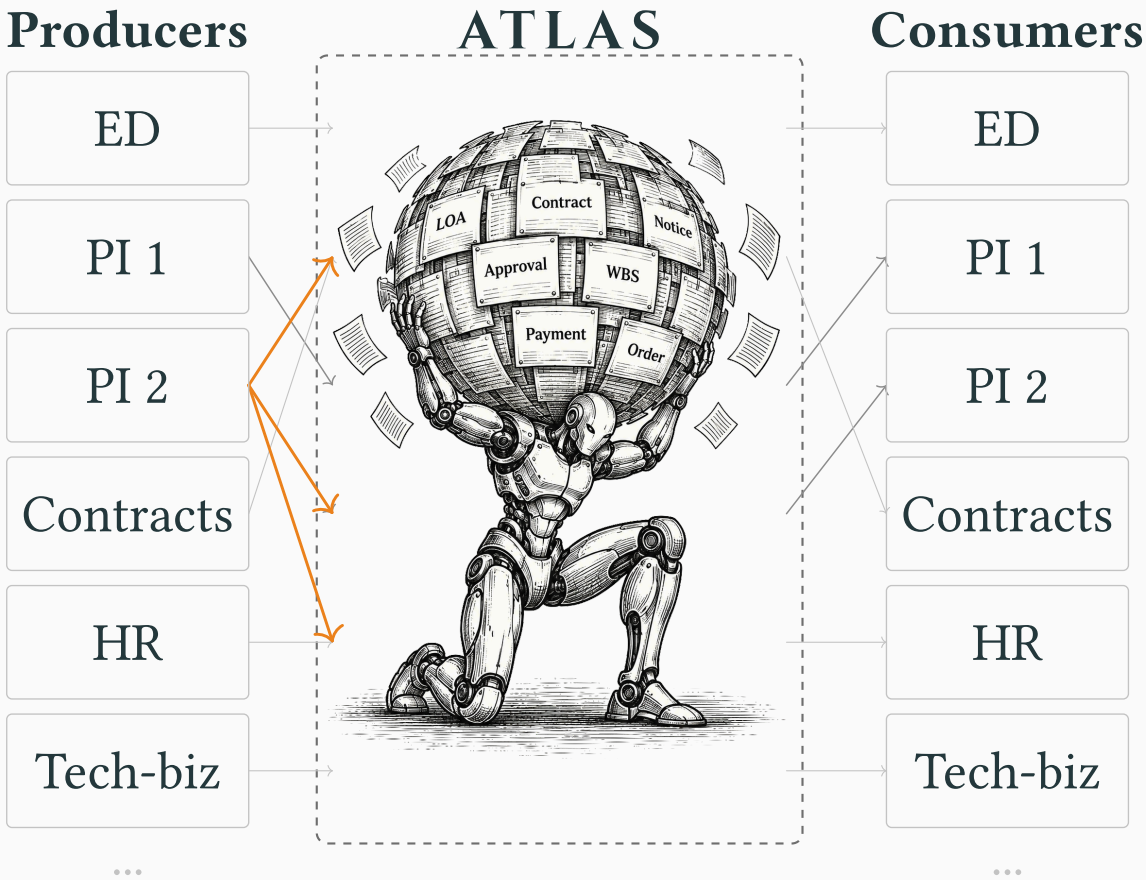
8.2 IMCB ATLAS: Document Flow

Seven Risks

1. Erroneous actions
2. Unauthorized actions
3. Biased or unfair actions
4. Data breaches
5. Disrupt connected systems
6. Speed and volume
7. Cascading effects

Four Pillars

A. Assess and bound risks
B. Make humans accountable
C. Technical controls
D. Enable end-user responsibility



Each business unit continuously submits paperwork as they produce it.

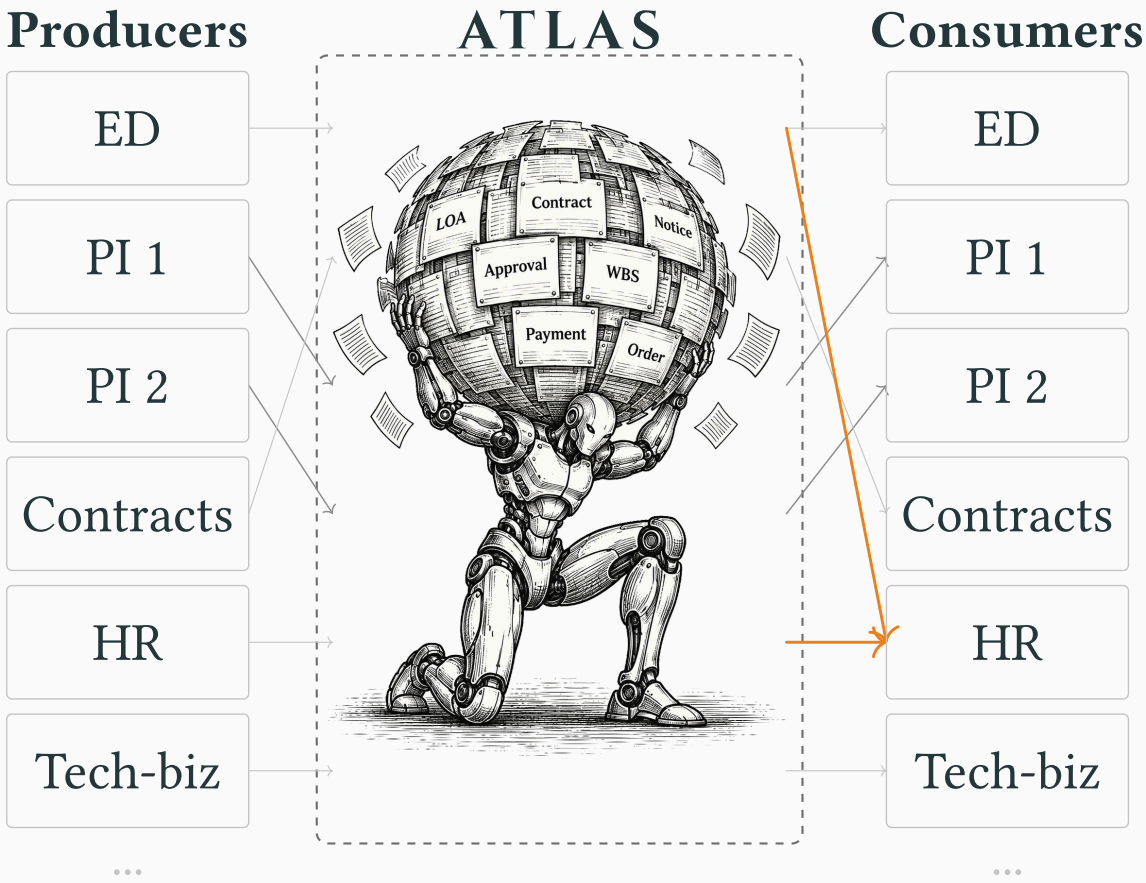
8.2 IMCB ATLAS: Document Flow

Seven Risks

1. Erroneous actions
2. Unauthorized actions
3. Biased or unfair actions
4. Data breaches
5. Disrupt connected systems
6. Speed and volume
7. Cascading effects

Four Pillars

A. Assess and bound risks
B. Make humans accountable
C. Technical controls
D. Enable end-user responsibility



ATLAS answers questions for consumers, depending on their role.

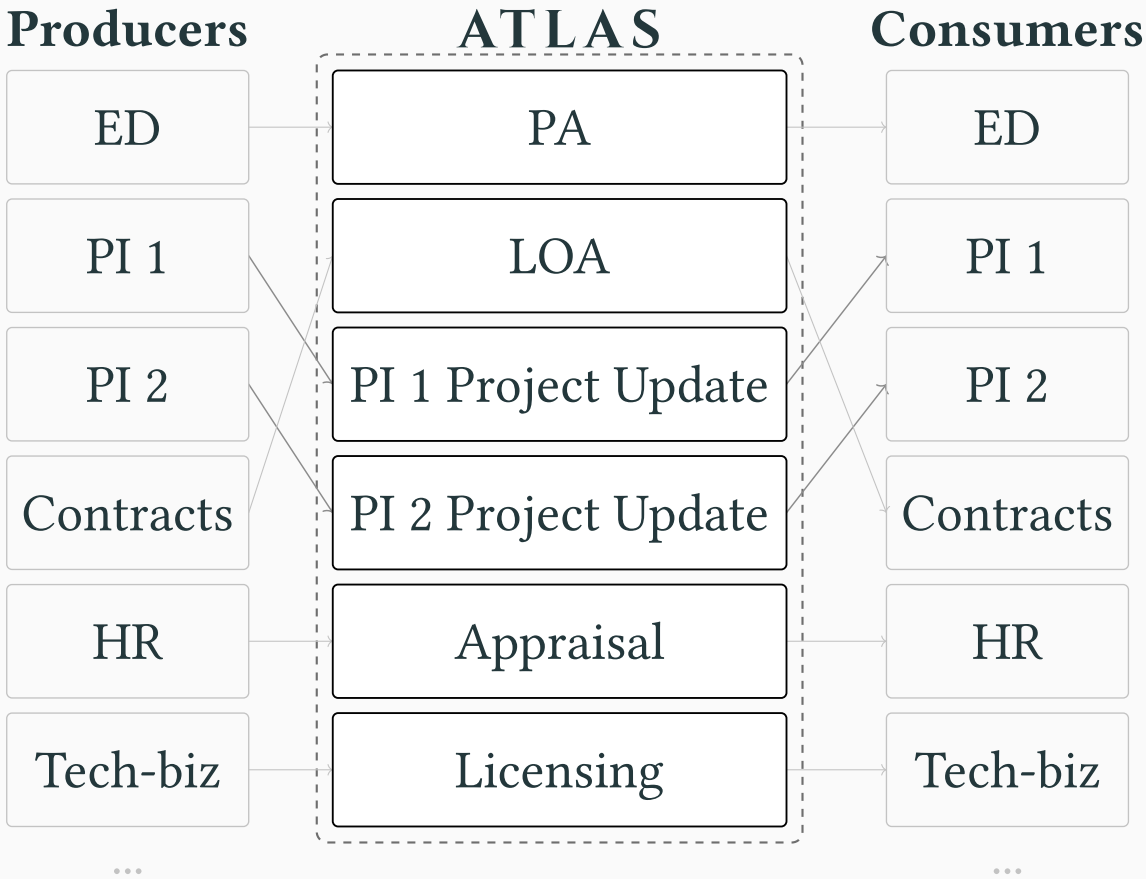
8.3 Unauthorized access & data breach

Seven Risks

1. Erroneous actions
2. Unauthorized actions
3. Biased or unfair actions
4. Data breaches
5. Disrupt connected systems
6. Speed and volume
7. Cascading effects

Four Pillars

A. Assess and bound risks
B. Make humans accountable
C. Technical controls
D. Enable end-user responsibility



If everyone uses the same chatbot, what stops it revealing documents they shouldn't see?

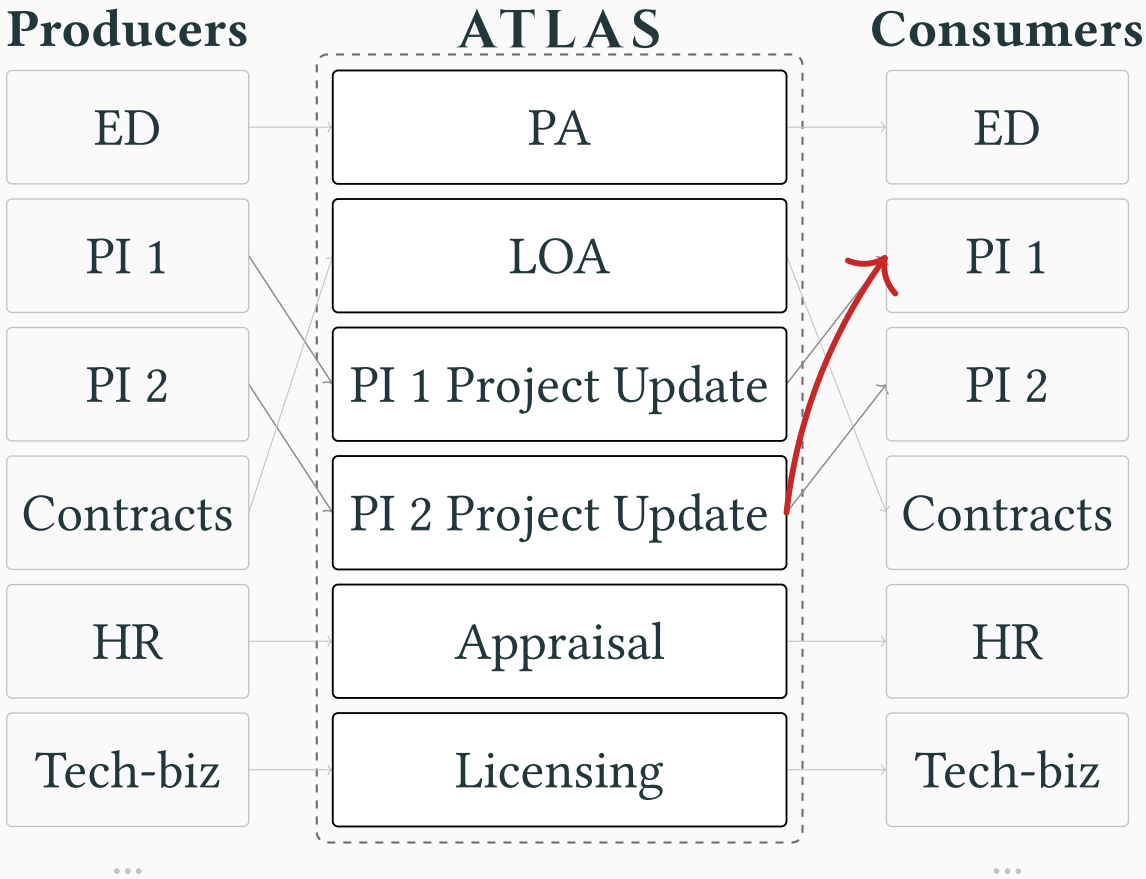
8.4 Unauthorized access & data breach

Seven Risks

1. Erroneous actions
2. Unauthorized actions
3. Biased or unfair actions
- ✗ **Data breaches**
5. Disrupt connected systems
6. Speed and volume
7. Cascading effects

Four Pillars

A. Assess and bound risks
B. Make humans accountable
C. Technical controls
D. Enable end-user responsibility



What if PI 1 reads PI 2's private data?

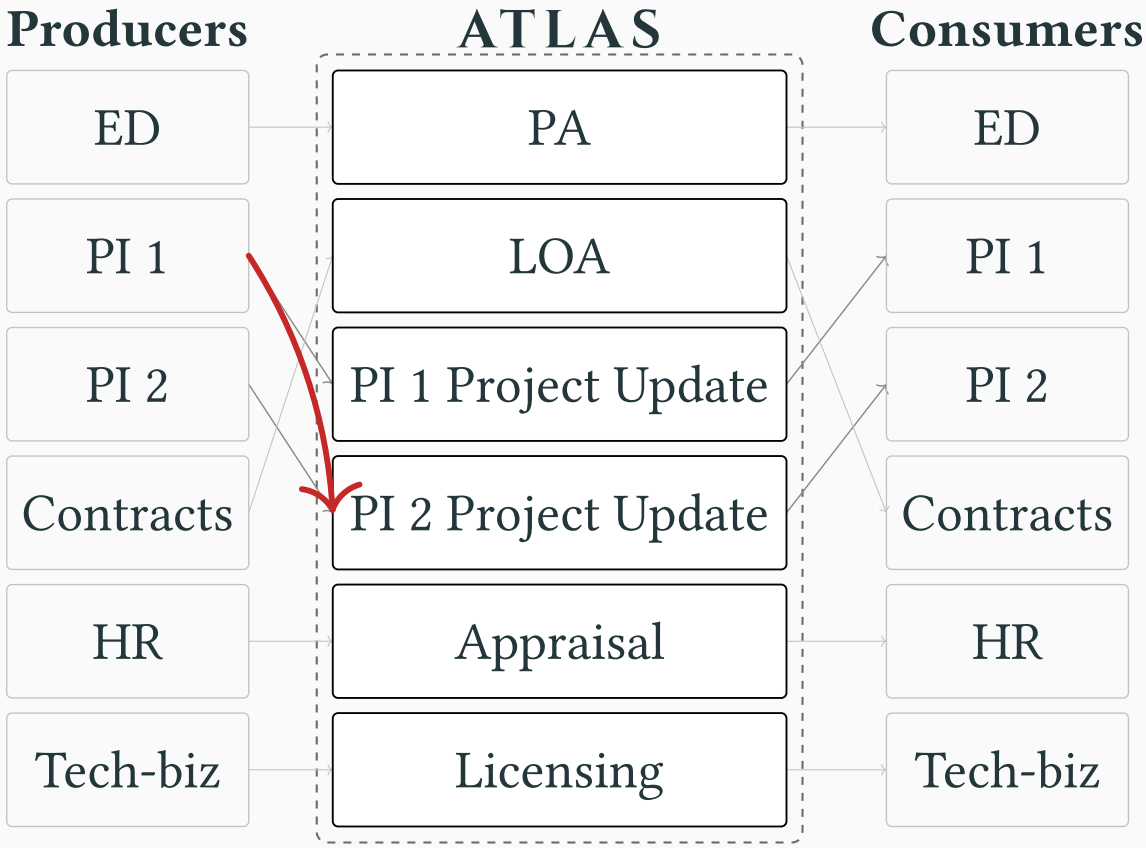
8.4 Unauthorized access & data breach

Seven Risks

1. Erroneous actions
2. ✗ **Unauthorized actions**
3. Biased or unfair actions
4. Data breaches
5. Disrupt connected systems
6. Speed and volume
7. Cascading effects

Four Pillars

A. Assess and bound risks
B. Make humans accountable
C. Technical controls
D. Enable end-user responsibility



What if PI 1 writes into PI 2's project?

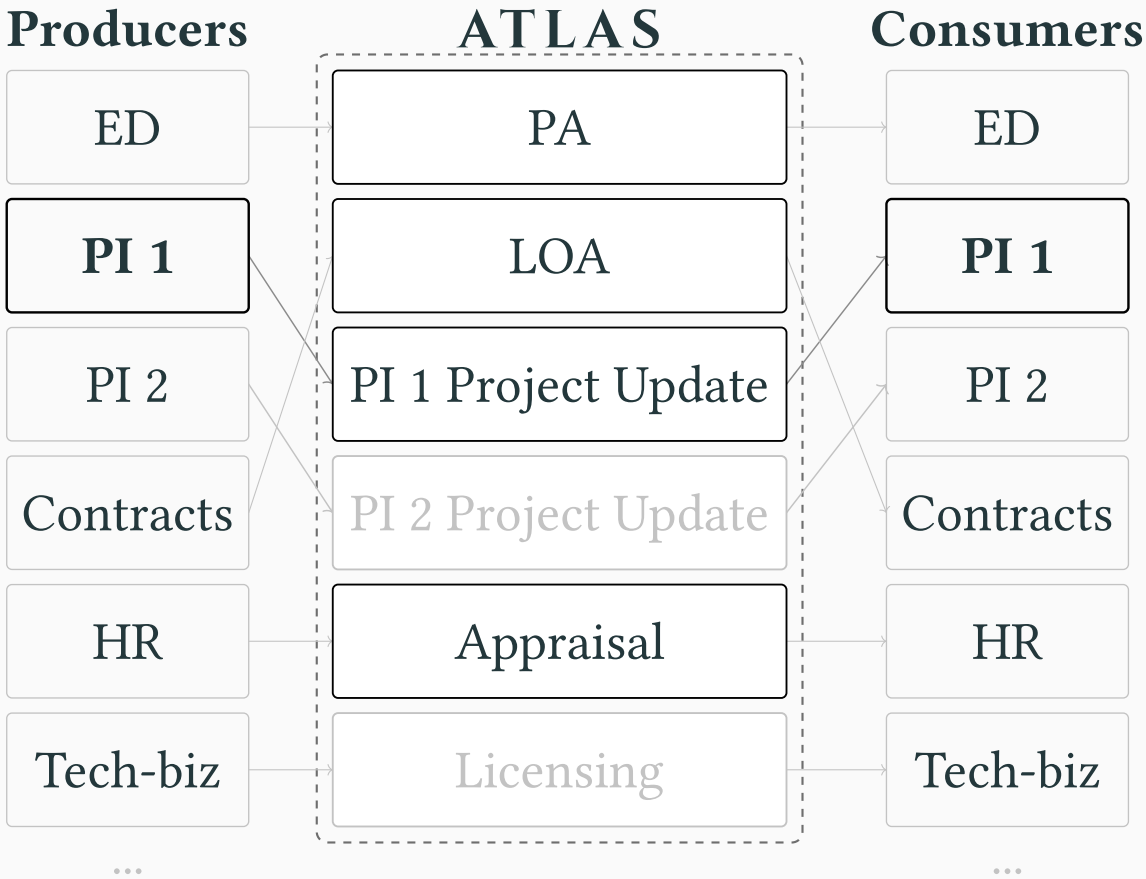
8.4 Unauthorized access & data breach

Seven Risks

1. Erroneous actions
2. ✓ **Unauthorized actions**
3. Biased or unfair actions
4. ✓ **Data breaches**
5. Disrupt connected systems
6. Speed and volume
7. Cascading effects

Four Pillars

A. Assess and bound risks
B. Make humans accountable
C. **Technical controls**
D. Enable end-user responsibility



Policy: *When PI 1 uses ATLAS, it can only read or write PI 1's authorised documents.*

8.5 Unauthorized access: policy and control

Seven Risks

1. Erroneous actions
2. **Unauthorized actions**
3. Biased or unfair actions
4. **Data breaches**
5. Disrupt connected systems
6. Speed and volume
7. Cascading effects

Four Pillars

A. Assess and bound risks
B. Make humans accountable
C. **Technical controls**
D. Enable end-user responsibility

The risk

Users may access documents (reading or writing) they are not supposed to, leaking business secrets.

Policy questions

- Who may read which facts?
- Who may write, and to which records?
- How is access revoked when someone leaves?
- Are there more fine-grained controls for information?

Technical control: Role-Based Access Control

ATLAS can only read/write documents the user has access to.

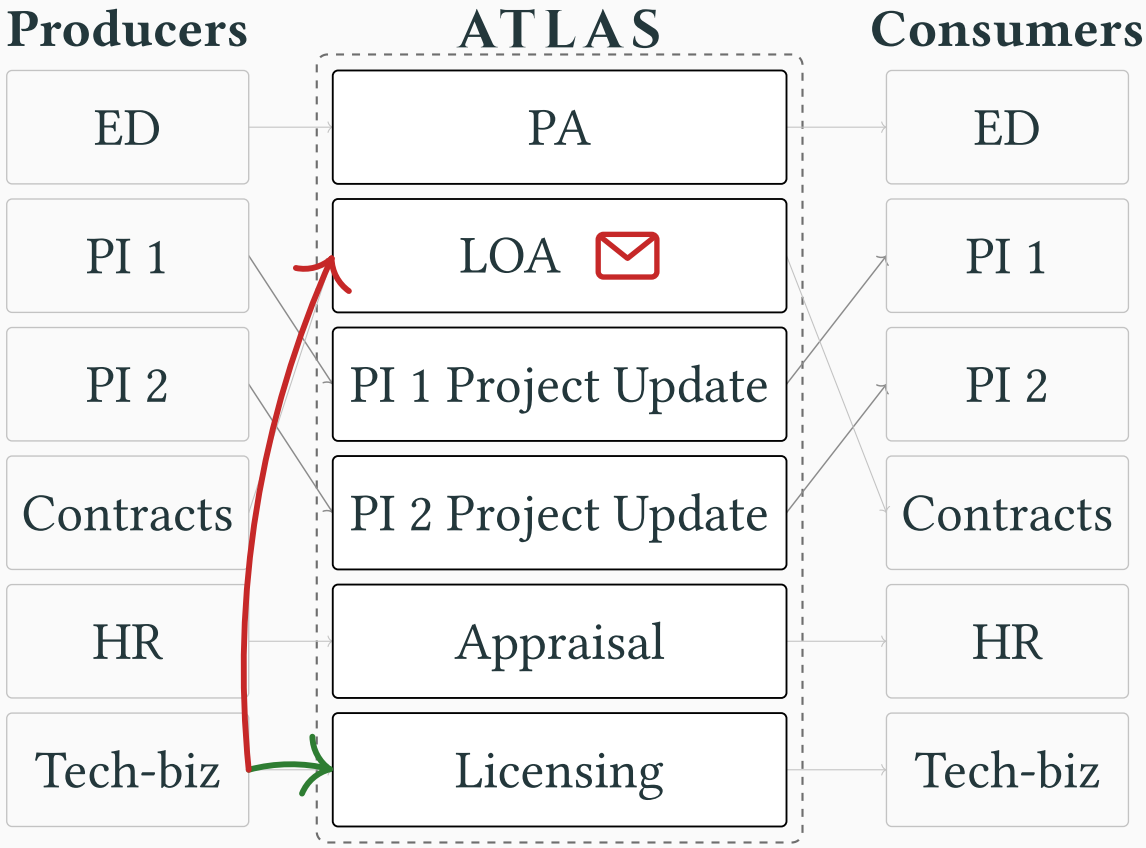
8.6 Erroneous action

Seven Risks

1. **Erroneous actions**
2. Unauthorized actions
3. Biased or unfair actions
4. Data breaches
5. Disrupt connected systems
6. Speed and volume
7. Cascading effects

Four Pillars

A. Assess and bound risks
B. **Make humans accountable**
C. Technical controls
D. Enable end-user responsibility



What if the LLM misfiles the document?

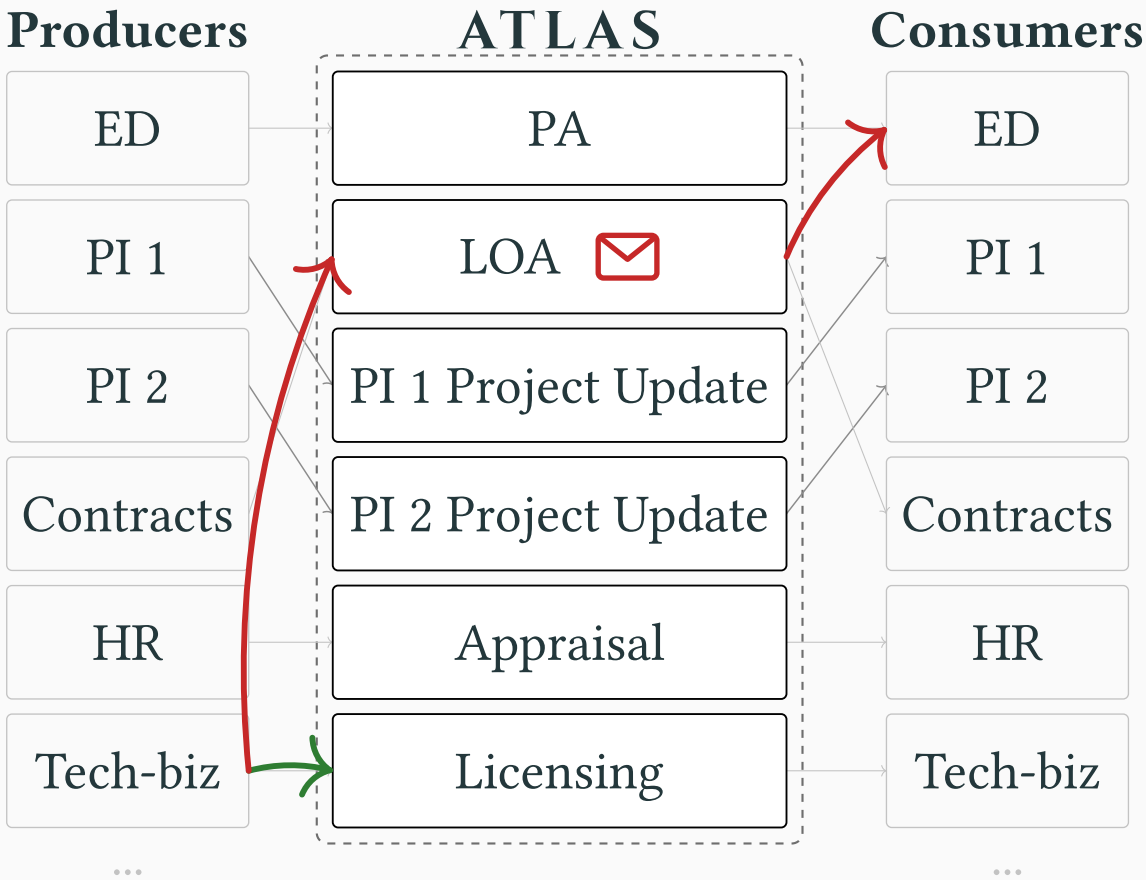
8.7 Cascading effects

Seven Risks

1. **Erroneous actions**
2. Unauthorized actions
3. Biased or unfair actions
4. Data breaches
5. Disrupt connected systems
6. Speed and volume
7. **Cascading effects**

Four Pillars

A. Assess and bound risks
B. **Make humans accountable**
C. Technical controls
D. **Enable end-user responsibility**



What if the error cascades into downstream business decisions?

8.8 How should humans check AI's filing?

Seven Risks

1. **Erroneous actions**
2. Unauthorized actions
3. Biased or unfair actions
4. Data breaches
5. Disrupt connected systems
6. Speed and volume
7. **Cascading effects**

Four Pillars

A. Assess and bound risks
B. **Make humans accountable**
C. Technical controls
D. **Enable end-user responsibility**

Information is risky depending on its use

- A grant end date in
  - an internal "congratulations" email, vs.
  - used to extend staff contracts.

Set policies to balance the **risk** of using information with the effort of checking it.

Policy questions

- Which conclusions are high-risk and need immediate human verification?
- Which conclusions are safe to check only when an error is discovered?
- Once an error is discovered, who owns the correction?

Pillar B: Make humans meaningfully accountable

How can we get producers to verify that ATLAS understands documents correctly?

- Every document type has one named owner (grants office, HR, procurement)
- ATLAS only flags potential issues, and records the human's exact decision.

How can we reduce review fatigue while keeping human oversight effective?

Pillar D: Enable end user responsibility

How can we ensure consumers are not led astray by ATLAS?

- Every answer reports its sources and corroboration.
- Provide a mechanism for reporting wrong answers and feeding back.
- Set risk-based policy for level of scrutiny.

Consumer View (example)

"Total grant value is $1,225,600.

verified by `leews` 11d ago (Report)

Available balance is $318,400.

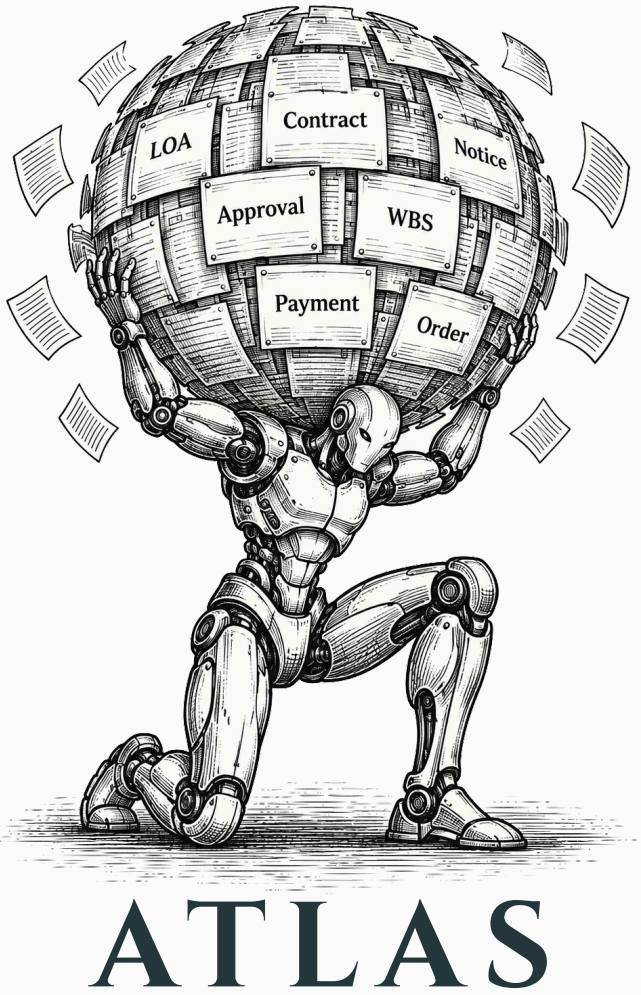
AI checked, request human

Staff contracts end on 30 June 2027."

AI check failed, request human

It's **our** responsibility to design ATLAS defensively, against foreseeable risks.

The "easy" path must be the safe path, and it must fit cleanly into existing processes.

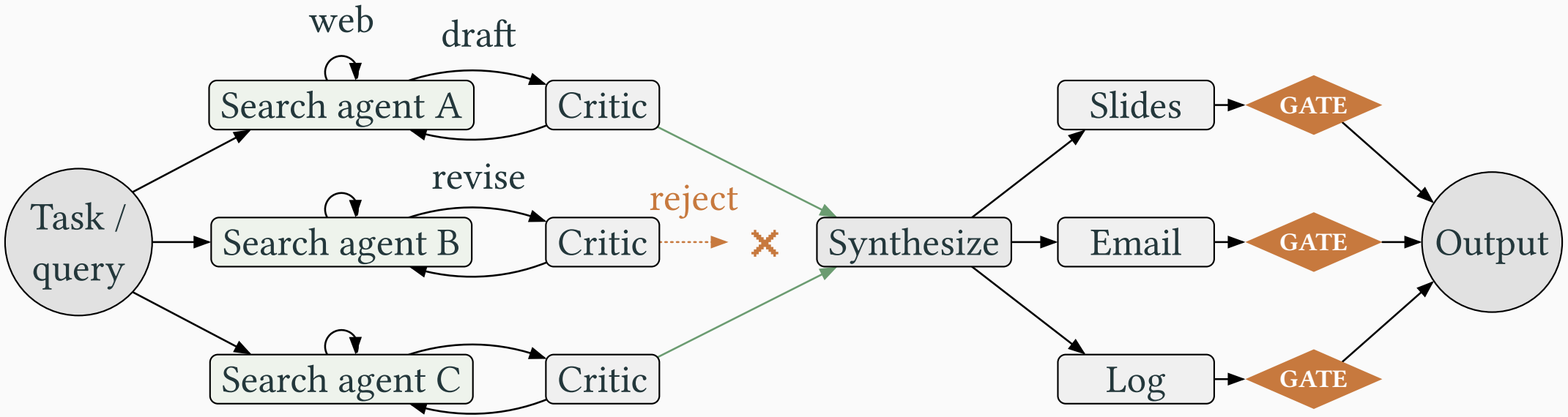


ATLAS is in early development at IMCB.

We’re designing these controls in from the start.

9. Where do we go from here?

9.1 Workflows get complicated



Today, we've only scratched the surface.

AI engineering offers new and effective tools for putting policy into practice.

Regulation

- New regulations on the use of AI
  - EU AI Act; South Korea's AI Basic Act
  - China's AI-content labelling
  - Claude Fable 5 export ban
- Guardrails to restrict frontier intelligence
  - Synthetic biology research
- Increased interest in "local" models
  - Provides freedom and control
    - sometimes cost savings
  - Hardware depreciates quickly

Commercial strategy

- Resist vendor lock-in
  - Pace of development means long-term value is hard to predict
  - Expensive value-added services may be much cheaper or obsolete in months
  - Keep your data exportable and your workflows portable
- Assess the full cost of ownership
  - Account for governance and oversight
  - Measure benefit on your tasks
  - Scale when the results justify it

Start small

↓

Automate part of it

↓

Keep human oversight

↓

Make evidence checkable

↓

Build confidence

We're living through a moment of extraordinary possibility.

The next chapter is ours to shape.



OcelliQ.com/astar-lr

`/tell-gaurav`

Help us improve! Your agent will ask a few questions and send Gaurav your feedback.

Connect with me on LinkedIn.

References

1. Amazon News. *Amazon CEO Andy Jassy talks 6 truths surrounding the rise of AI.* 2026.
2. Microsoft. *The golden opportunity for American AI.* 3 January 2025.
3. Anthropic. *Estimating AI productivity gains from Claude conversations.* 25 November 2025. Research projections.
4. Gartner. *Worldwide AI Spending Forecast.* 16 September 2026. AI infrastructure: US$1.484 trillion forecast for 2026.
5. OpenAI. *The Hugging Face incident and the road ahead.* 26 August 2026. Postmortem of the July 2026 cybersecurity-evaluation incident.
6. The Guardian. *AI agent deletes a firm's database.* 29 April 2026.
7. CNN, syndicated by KTVZ. *US military close call after a false AI intelligence report.* 18 September 2026. Reporting based on unnamed sources, not an official incident finding.
8. The Planetary Society.
   How much did the Apollo program cost?.
   Apollo, 1960-1973: US$309 billion in 2025 dollars.
9. AskEngineers discussion.
   Hypothetical Great Wall rebuilding estimate.
   2015. Informal estimate; not a professional quotation. Exact US$452 billion value not independently verified.

References

10. Workshop illustrative model.
Cumulative global datacenter spending.
Supplied historical anchors and adjustment assumptions; not a measured historical series.

11. Gartner.
Data Center Electricity Consumption Forecast.
10 June 2026. AI-optimized servers: 175 TWh forecast for 2026.

12. Energy Market Authority, Singapore.
Singapore Energy Statistics: Energy Consumption.
2024 electricity consumption: 58 TWh.

13. China Three Gorges Corporation.
Three Gorges annual power generation.
19 November 2020. Designed annual generation: 88.2 TWh; distinct from actual output.

14. Info-Tech Research Group. *Discover the Enterprise AI Technology Stack.*

15. NIST. *AI Risk Management Framework: AI RMF Core.*

16. Andreessen Horowitz. *Emerging Architectures for LLM Applications.*

17. Amazon Web Services. *Layered approach for a generative AI platform.* AWS Prescriptive Guidance.

18. NVIDIA. *AI's five-layer cake.*

References

19. Ajay Agrawal, Joshua Gans, Avi Goldfarb.
Prediction Machines: The Simple Economics of Artificial Intelligence.
Harvard Business Review Press, 2018.

20. William Stanley Jevons.
The Coal Question, Chapter VII: Of the Economy of Fuel.
First published 1865; linked text is the second edition, Macmillan, 1866.

21. Daron Acemoglu, Simon Johnson.
Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity.
PublicAffairs, 2023. Further reading on technology and prosperity.

22. OpenAI.
Tokenizer.
Interactive examples of tokenization.

23. OpenRouter. *Daily token totals for top 50 models.* Usage feed. Prices and listing dates: models catalog. Snapshot: 30 September 2026.

24. Artificial Analysis, via OpenRouter. *Artificial Analysis Intelligence Index.* Benchmark snapshot: 30 September 2026. Stored curated scores and teaching overrides: 01-intro/data/README.md.

25. Artificial Analysis. *Gemini 4 Argon: Google is back as one of the top three labs in intelligence achieved.* 30 September 2026. AAII 53 (high); standard input/output pricing of US$4/US$20 per million tokens, before the introductory discount.

References

26. Artificial Analysis. *Muse Spark 1.3 (max)*. Read 2 October 2026. AAII 48 under Intelligence Index v4.3.2; later benchmark reading than the stored September snapshot.

27. Moonshot AI. *Kimi K3 model card*. 2.8T total parameters; native MXFP4 weights. Read 2 October 2026.

28. DeepSeek. *DeepSeek V4.1 Flash model card*. 552B backbone plus 196B Engram lookup parameters. Read 2 October 2026.

29. Z.ai. *GLM 5.3 Flash model card*. 320B total parameters. Read 2 October 2026.

30. Tobias Groot and Matias Valdenegro-Toro. *Overconfidence is Key: Verbalized Uncertainty Evaluation in Large Language and Vision-Language Models*. TrustNLP 2024, pp. 145-171.

31. Sharma et al. *Towards Understanding Sycophancy in Language Models*. 2023.

32. Shapira et al. *Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models*. arXiv:2305.14763, 2023.

33. Anthropic. *Sycophancy to Subterfuge: Investigating Reward Tampering in Language Models*. 17 June 2024. Controlled training experiments; not established as common production behavior.

34. John R. Searle.
Minds, brains, and programs.
Behavioral and Brain Sciences 3(3), 417-424, 1980. Chinese Room argument.

References

35. Wallace et al. *The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions.* OpenAI, 2024.

36. Liu et al. *Lost in the Middle: How Language Models Use Long Contexts.* arXiv:2307.03172 (2023); Transactions of the Association for Computational Linguistics, 2024.

37. Anthropic. *Project Vend, phase one.* 27 June 2025. Claude Sonnet 3.7 running a real office shop with Andon Labs.

38. Aditya Mehta, TechCrunch. *AI hallucination nearly triggers US military operation.* 18 September 2026.

39. Menghua Xia et al.
DREAM: On hallucinations in AI-generated content for nuclear medicine imaging.
arXiv:2506.13995v2, 18 June 2025. Sections III-IV; displayed image is Figure 5.

40. Reuters. *South Africa withdraws draft AI policy over fictitious references.* 27 April 2026. Screenshot: Reuters republication on MyJoyOnline, 28 April 2026.

41. Anthropic. *Claude Code commands.* `/usage`, `/context` and `/compact`; availability depends on the account and environment. Read 4 October 2026.

42. Anthropic. *Manage costs effectively.* Session figures are estimates; subscription limits and API consumption differ. Read 4 October 2026.

References

43. Anthropic. *How Claude Code works.* Context, subagents and compaction; early details can be lost during summarization. Read 4 October 2026.

44. Infocomm Media Development Authority (IMDA). *Model AI Governance Framework for Agentic AI.* Version 1.5. Published 20 May 2026; updated 5 June 2026. Local PDF and extracts in references/imda-agentic-ai/.

45. International Organization for Standardization / International Electrotechnical Commission. *ISO/IEC 42001:2023 — Artificial intelligence — Management system.* First edition, December 2023.

46. OWASP GenAI Security Project, Agentic Security Initiative. *Agentic AI — Threats and Mitigations.* Version 1.0, 17 February 2025.

47. Summer Yue. *First-person account of OpenClaw deleting her inbox.* 22 February 2026. Post and screenshots; does not establish permanent loss of all messages.

48. Klarna. *AI assistant handles two-thirds of customer service chats.* 27 February 2024. Company release: work equivalent to 700 agents; projected US$40 million profit improvement.

References

49. Charles Daly, Bloomberg.
Klarna Slows AI-Driven Job Cuts With Call for Real People.
8 May 2025. CEO interview; full text requires a subscription.

50. Kenny Chee, The Straits Times.
Bee Cheng Hiang customer email exposure.
30 September 2026; updated 2 October 2026.

51. Lee Chong Ming, Business Insider.
Replit CEO apologizes after AI agent deletes production data.
22 July 2025.