Engineering Reliable Agents

v20250516

Kam Lasater
@seekayel
kamlasater.com

Goals

1. Frameworks for thinking about your problems

2. Description of the shape of your solutions

3. Inspiration from you / hiring

What are we
talking about?

My Context *

Team Size
- 5 engineers + me

Task Sizing - in Production
- Qualify a new sales lead via email (~100 day)
- Research booked calls - write prep doc (10s / day)
- Source 20 vendors for product [X] (~10 / week)
- Process financial docs and create prequal memo (~10 / week)

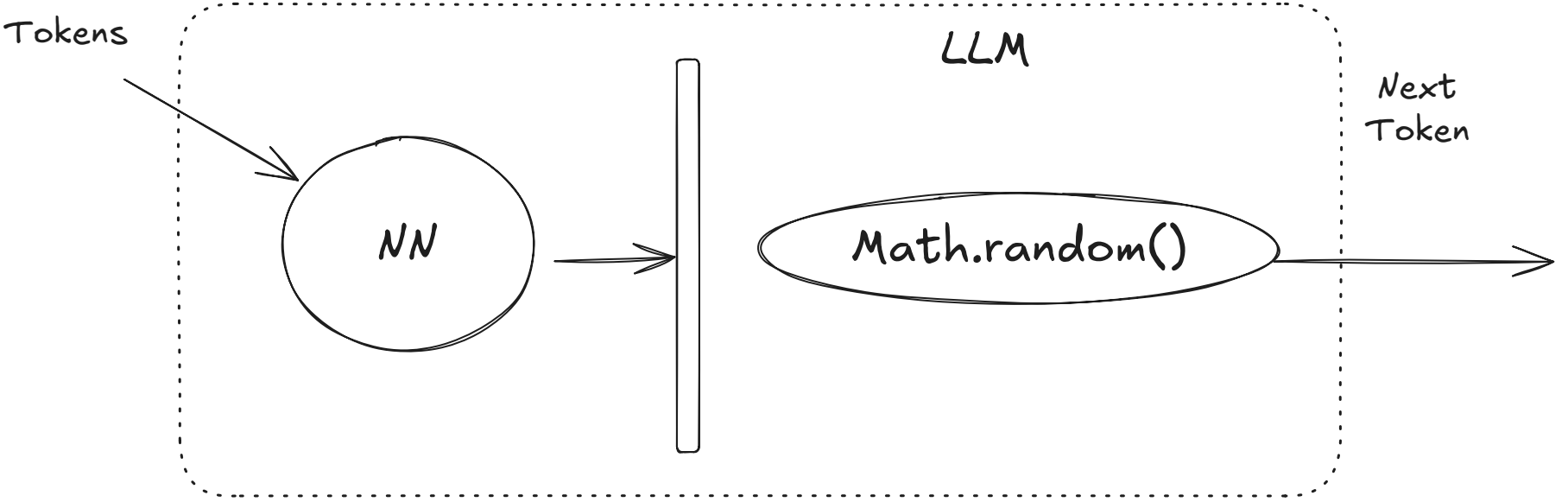
Task Autonomy - in Production
- Email drafting vs sending by email class (90% auto)
- Internal consumers / Internal tools
- Reversible decisions

* YMMV

If an agent can suprise you to the upside,
it can suprise you to the downside.

-me

An Agent is unpredictable by design.



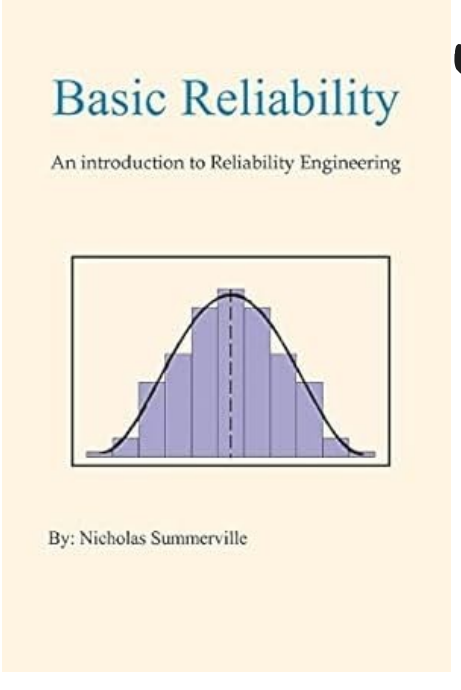
Benchmarks answer if an LLM can.

Not if it will.

Benchmark:
Is success in the set of possible outcomes?

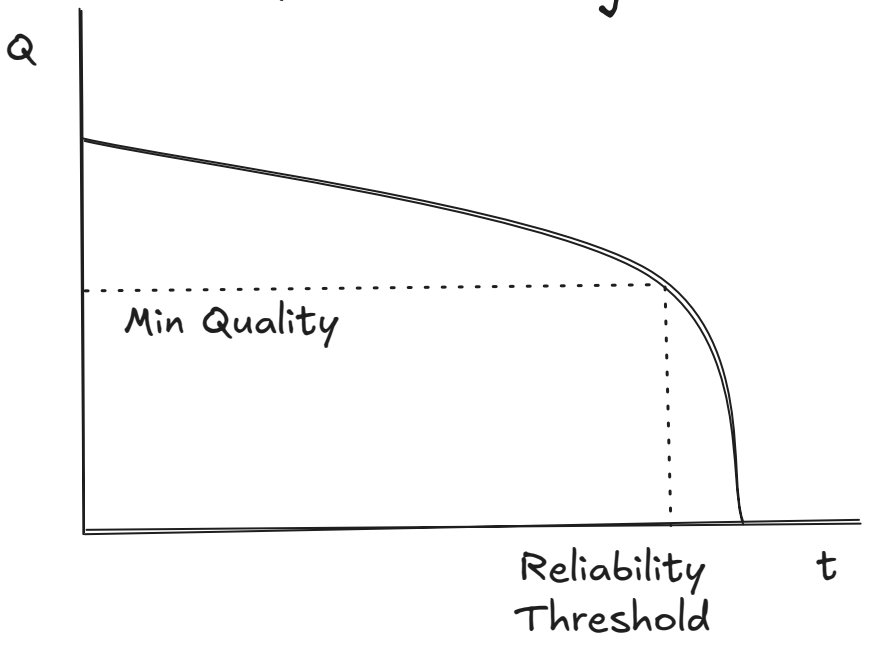
What I want:
What is the likelihood of success?

"Reliability" is a measurement of "Quality" over some period of time.

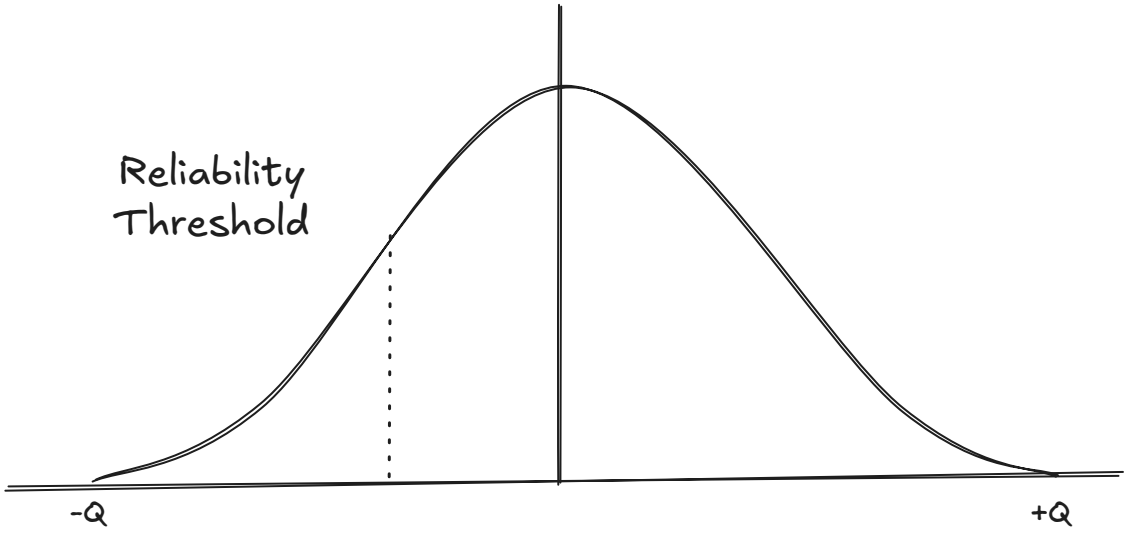


ummerville

Manufacturing



Agent



Infinitely Strong Bridges

Dictionary

Definitions from Oxford Languages · Learn more

en·gi·neer·ing

/ˌenjəˈniriNG/

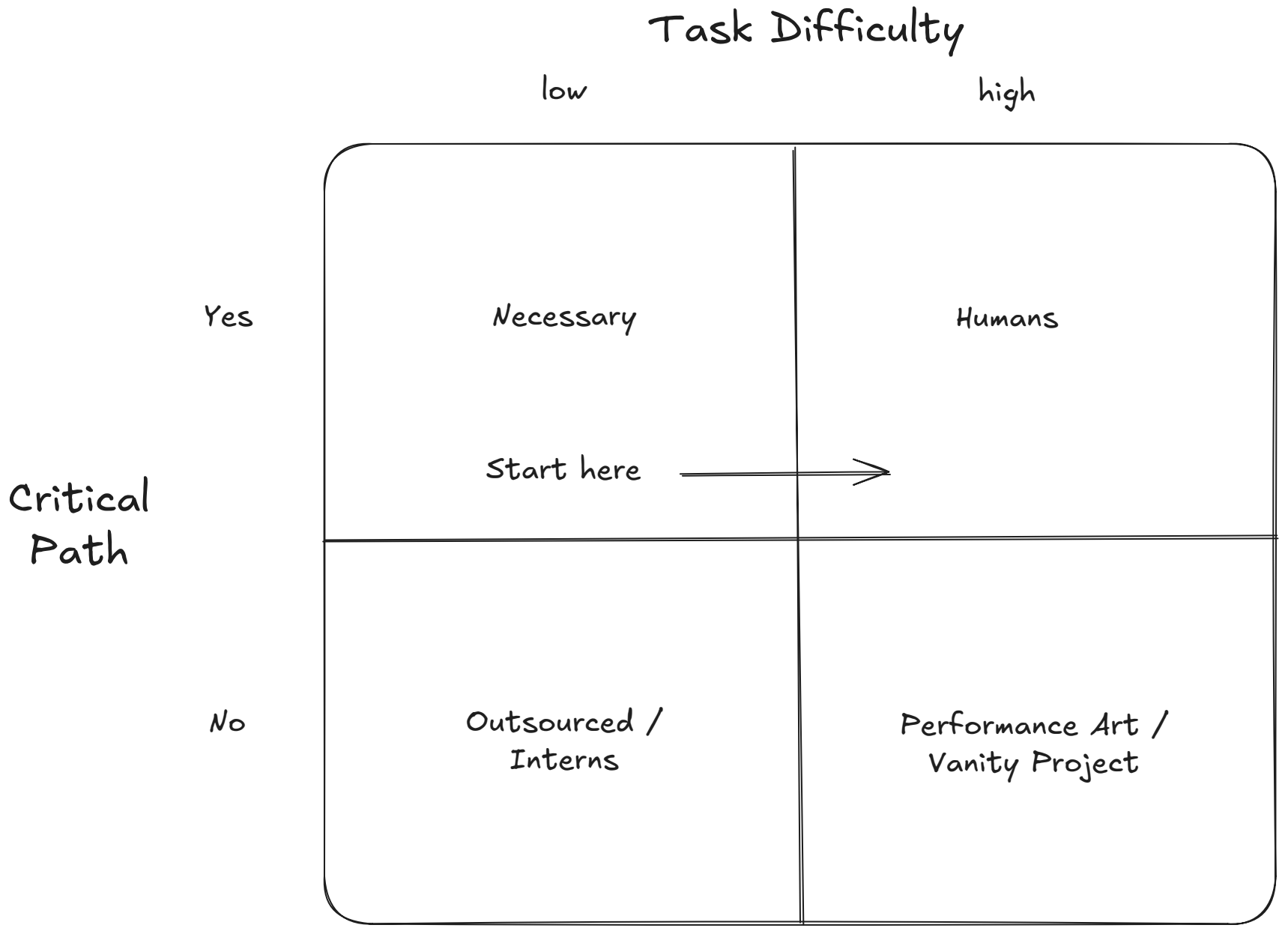
noun

the branch of science and technology concerned with making trade offs.

- you can't always get what you want, but if you try, you might find, you get what you need.
- the art of staying employeed while telling your boss no

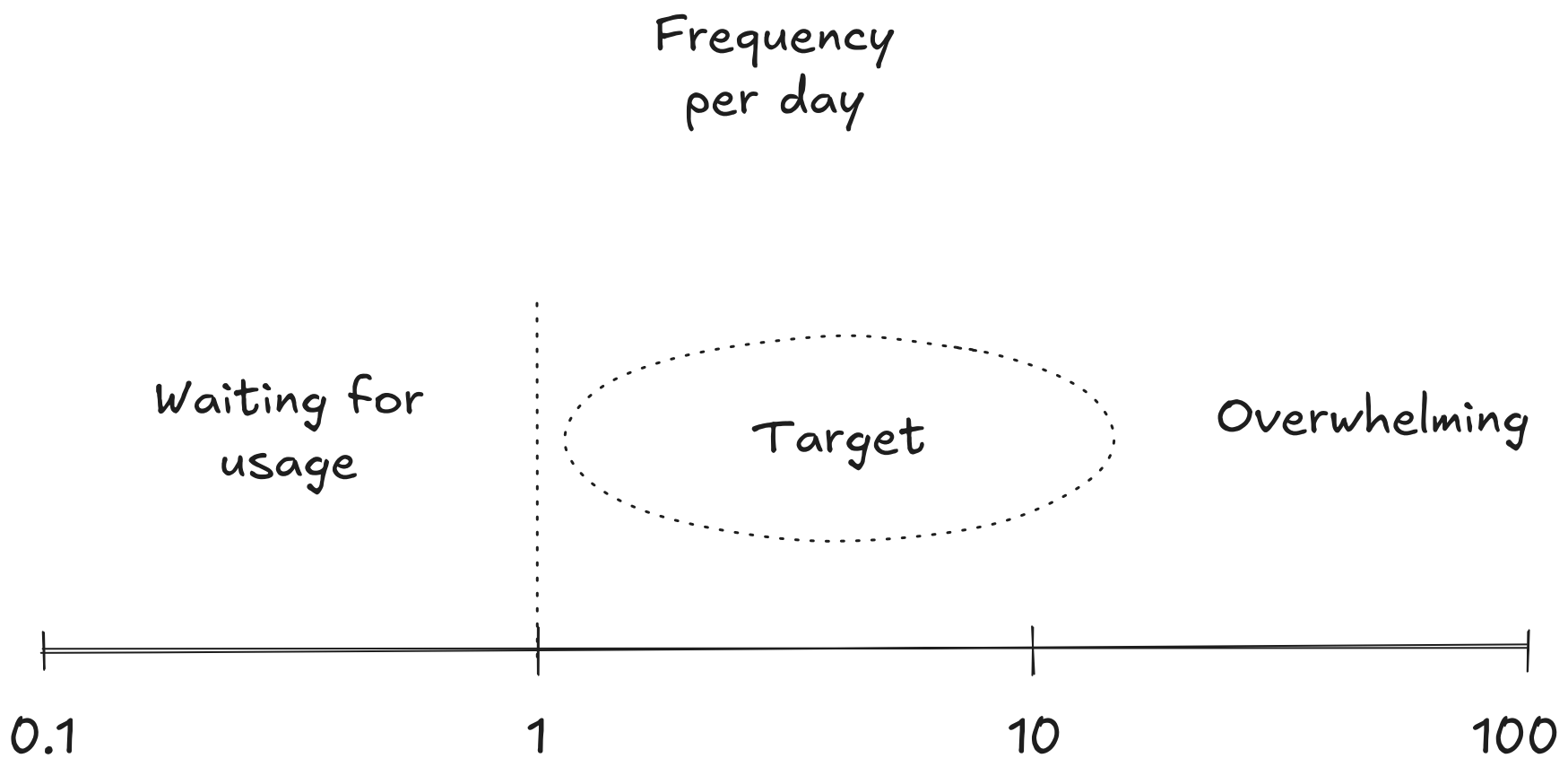
Our first failure...

Task Selection is a Strategic Decision



Our second failure...

Task Cadence is a Strategic Decision



Task Selction Matters

When building applications with LLMs, we recommend finding the simplest solution possible, and only increasing complexity when needed. This might mean not building agentic systems at all. Agentic systems often trade latency and cost for better task performance, and you should consider when this tradeoff makes sense.

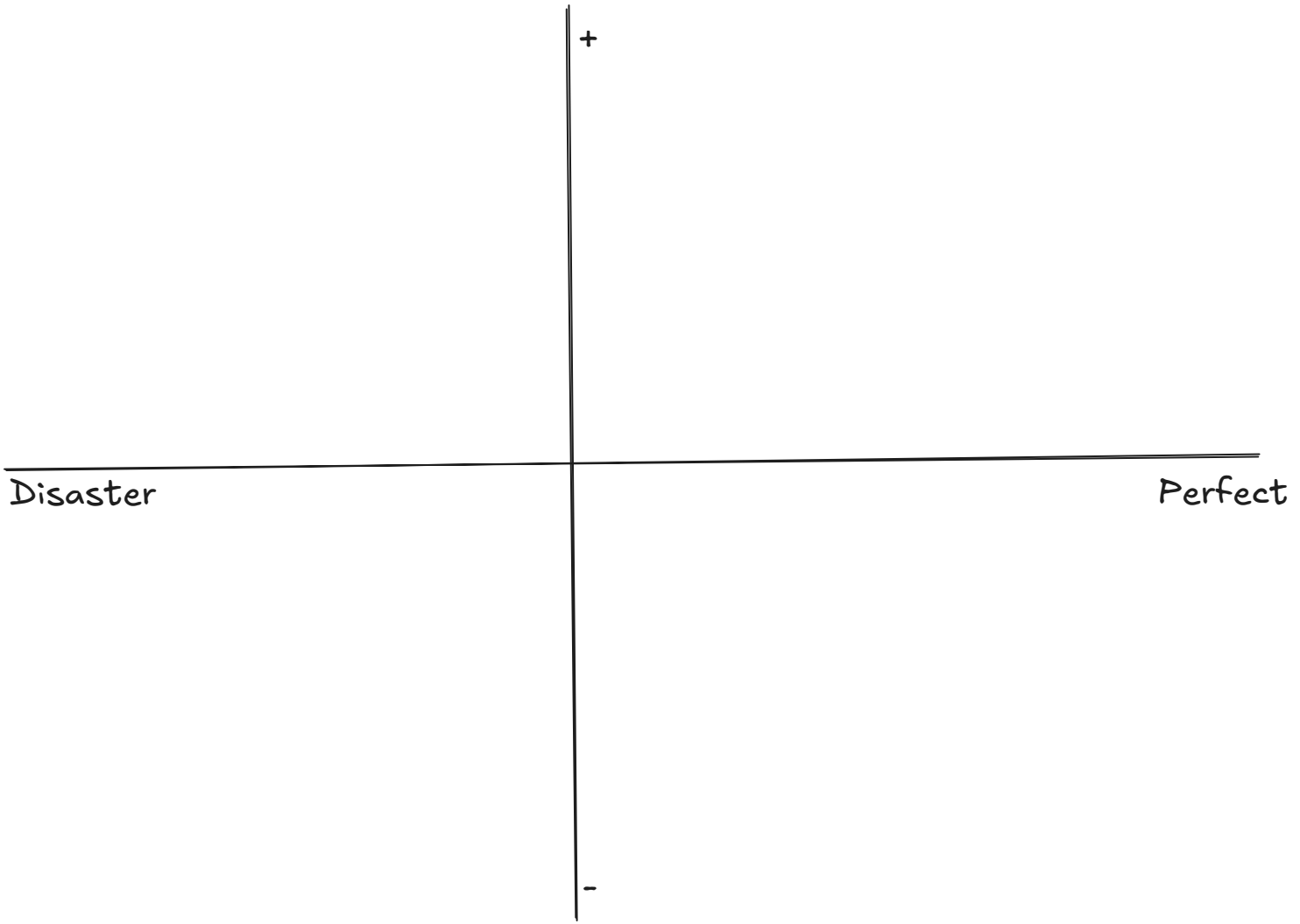
- Building Effective AI Agents
Anthropic Blog

Building agents requires rethinking how your systems make decisions and handle complexity. Unlike conventional automation, agents are uniquely suited to workflows where traditional deterministic and rule-based approaches fall short... Before committing to building an agent, validate that your use case... a deterministic solution may suffice.

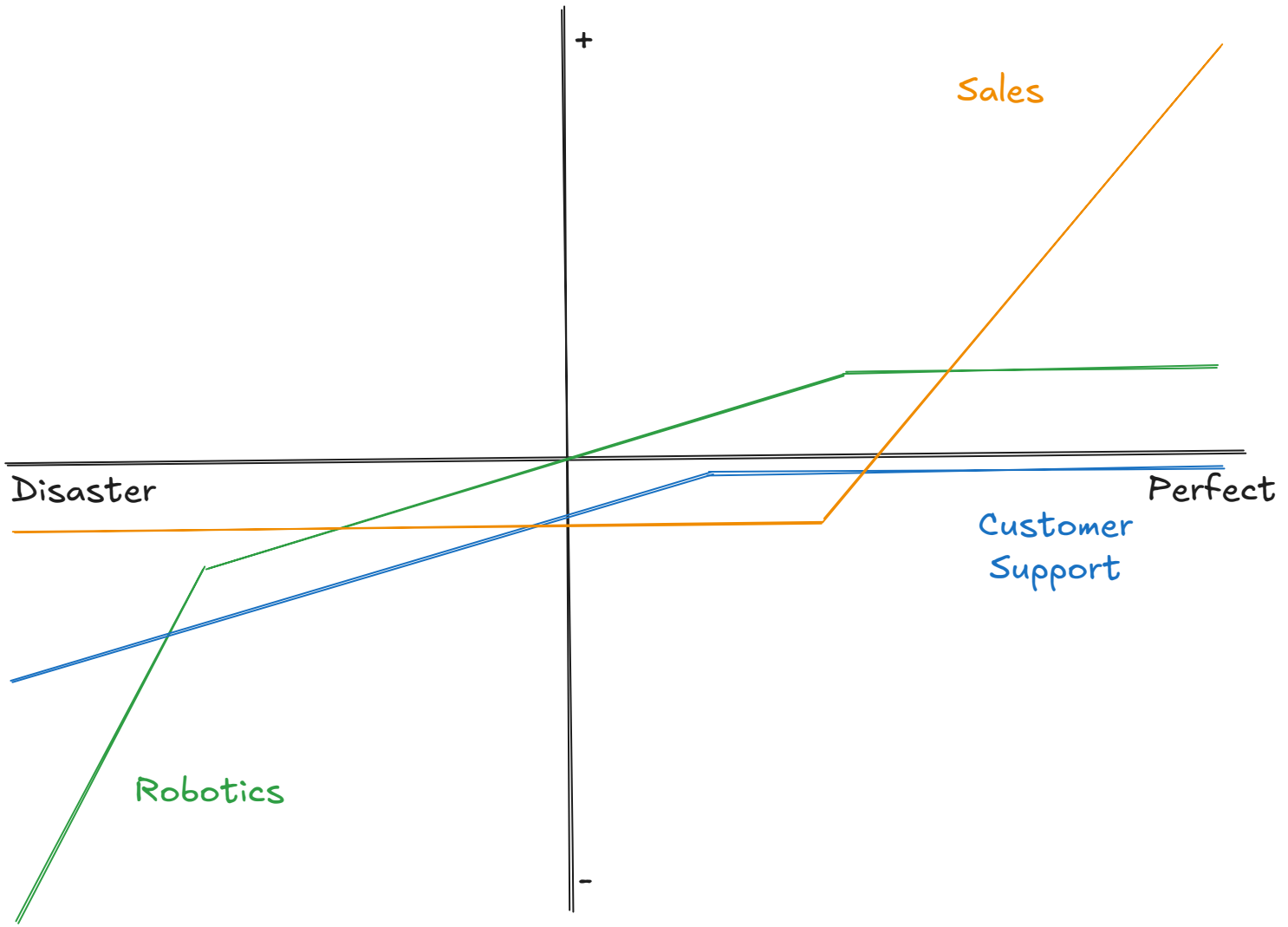
- A Practical Guide to Building Agents
OpenAI

Thinking About Trade offs

Economic Value to Response Quality



Economic Value to Response Quality



Additional Value Streams

Reduction in Lead times
- flywheel of value similar to CICD automation
- reduced transaction cost, reduced batch size, faster feedback, faster learning, better decisions

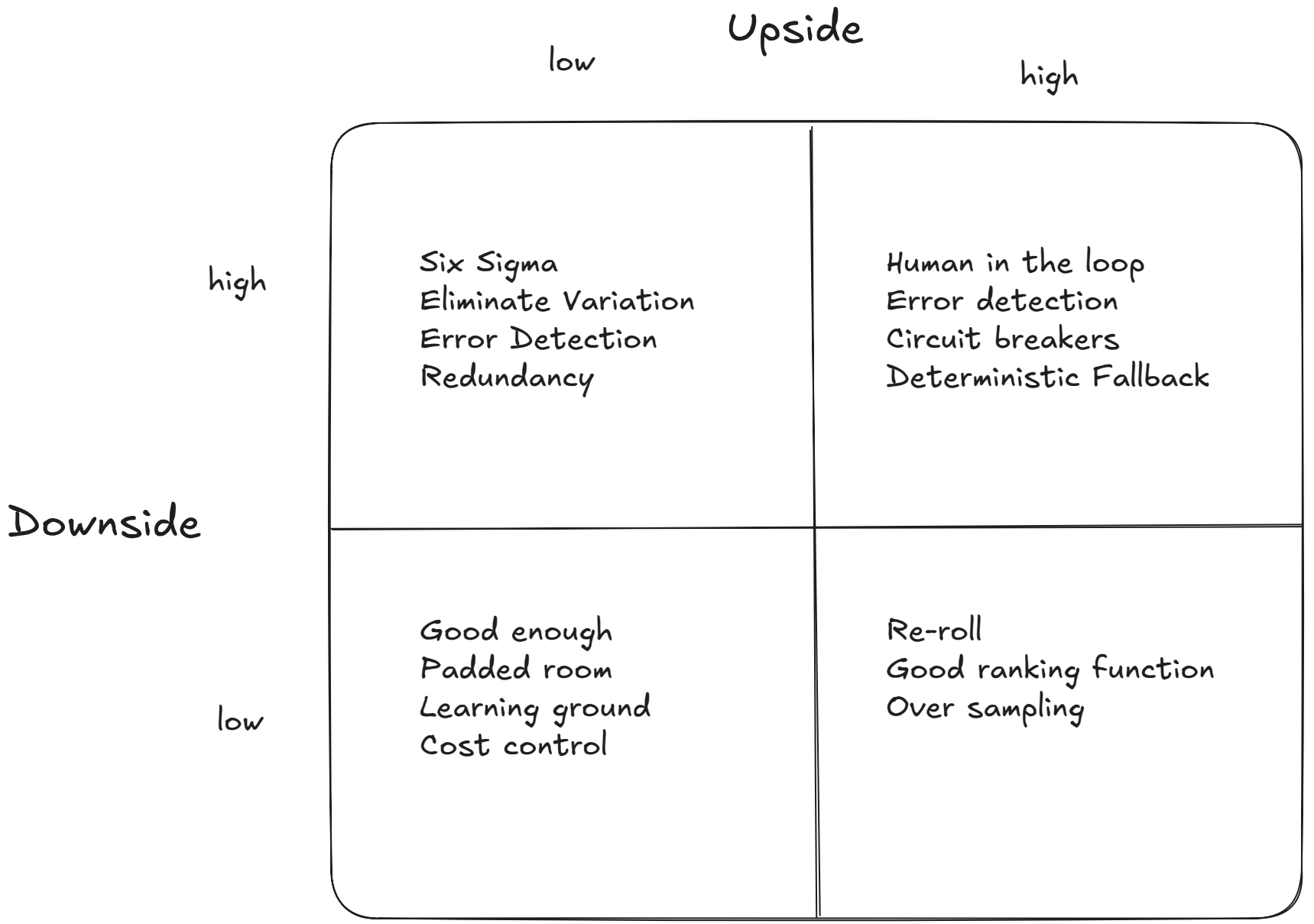
Doing the work that wouldn't get done
- improved process adherence

No task fatigue
- easier to be an editor than writer
- scale capacity of higher skill/knowledge people
- scale quality of lower skill/knowledge people

Some Patterns

(that might help)

Trade-offs Matched to Value Function



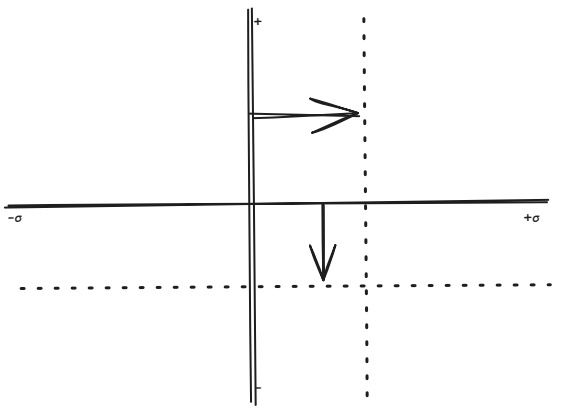
Over sampling

What:
Run same/similar prompt N times.
Rank result.
Pick (a) top result.

Requires:
Ranking function
10 < N < 100+

Trade off:
N times as expensive.
Fixed additional cost.

Examples
- Data transform
- Data extract



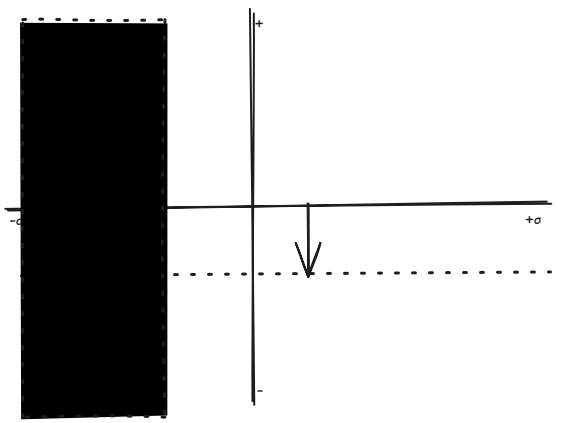
N-way Redundancy

What:
Run same node/process N times
Vote on result

Requires:
Voting mechanism (categorical)

Trade off:
N times as expensive.
Smaller 2 < N < 5
Fixed additional cost.

Examples
- Spam classifier
- Quality assesment



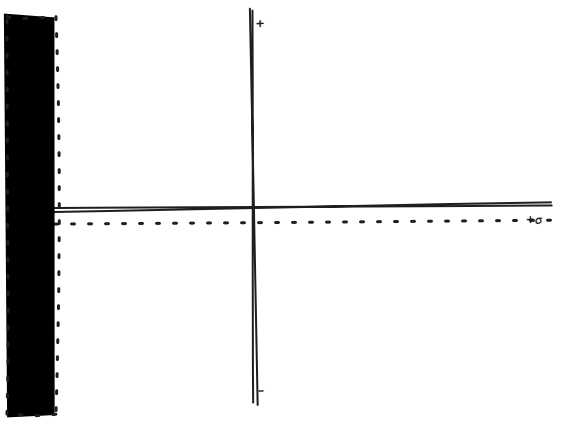
Reroll

What:
Run some node
Measure quality
If not acceptable repeat

Requires:
Quality measure (continuous)

Trade off:
Small additional cost
Known minimum threashold

Examples
- Messaging



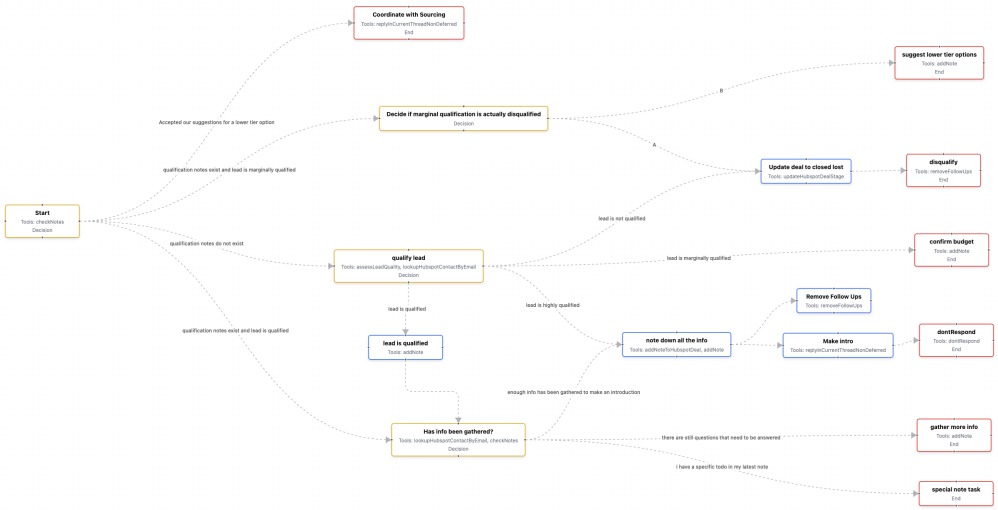
Blend Deterministic-Non-Deterministic

What:
Structured workflow
shared context window

Requires:
Framework support

Trade off:
Higher compliance
Higher human design

Examples
- Lead qualification
email responses



Encapsulate Context

What:
Everything is a "tool"
Prune context on return

Requires:
Framework support

Trade off:
Higher human design
Lose some "magic"

Examples
- Comms layer on outside
- Verify X business

Goals

1. Frameworks for thinking about your problems

- GET ON THE CRITICAL PATH

- Select a task frequency that works for you

- Draw your Value to Quality graph

- Analogous industry for inspiration

Goals

2. Description of the shape of your solutions

- Some patterns

- How to characterize the value function

- Matching pattern to value function

Goals

3. Inspiration
- Let me know your thoughts!
- Questions?

Join Us - We are hiring!

kam@manufactured.com or Discord

But the model
providers will save us

For Model Labs Quality is Job #1