

Agentic Discovery & Agentic Engineering

How AI agents change the software development lifecycle: when building gets cheap, the traditional six-phase SDLC collapses into two phases.

Agentic Discovery

learning velocity · a runnable probe to find the right problem

Vision & strategy

Market & customer

Discovery

Design

Requirements & spec

Agentic Engineering

production reliability · production-grade code from a validated spec

Dev

QA

Release & Ops

Tapio Pitkäranta



[longshoreforrest
.github.io/myself](https://longshoreforrest.github.io/myself) →

WHO

25 years	of AI and software agents, since agent research at VTT (2001)
Siili	Head of AI Development, Siili Solutions
4× CTO	retail, healthcare, tax and HR; part of RELEX's growth to a unicorn
D.Sc.	doctoral thesis on generative AI, Aalto University 2026
30+	international publications

HOW I USE AI

Agents every day. Building software, analyses and this very deck.

Consumer apps. Search, writing and summarising daily.

My own apps. I build my own AI applications with AI, often in hours.

The point: this is everyday reality, not the future.

Leena Pitkäranta



LinkedIn
leena-pitkaranta →

WHO

Aalto	Department of Industrial Engineering and Management, Aalto University
Paper	co-author of the Agentic Discovery & Engineering paper
Group	Associate Professor Natalia Vuori's research group: strategy work and AI

JOINT PUBLICATIONS

2026	Agentic Discovery & Engineering ↗ · working paper
2025	HADA ↗ · AGENTICS / IJCCI 2025
2025	AADAR · Springer LNBIP 566
2025	Towards Curated Learning (NordDRG) · PCSI 2025
2024	RAGADA · ICEIS 2024
2024	Software platform of strategic decision-making · PROS 2024 and SMS 2024 · Leena as lead author, with Natalia Vuori

RESEARCH GROUP

Natalia Vuori ↗
Associate Professor, Industrial Engineering and Management, Aalto

Studies how leaders' and teams' cognitive biases and emotions shape strategic decisions and change such as AI adoption. Published in e.g. Strategic Management Journal and Academy of Management Annals; six international SMS and AoM awards.

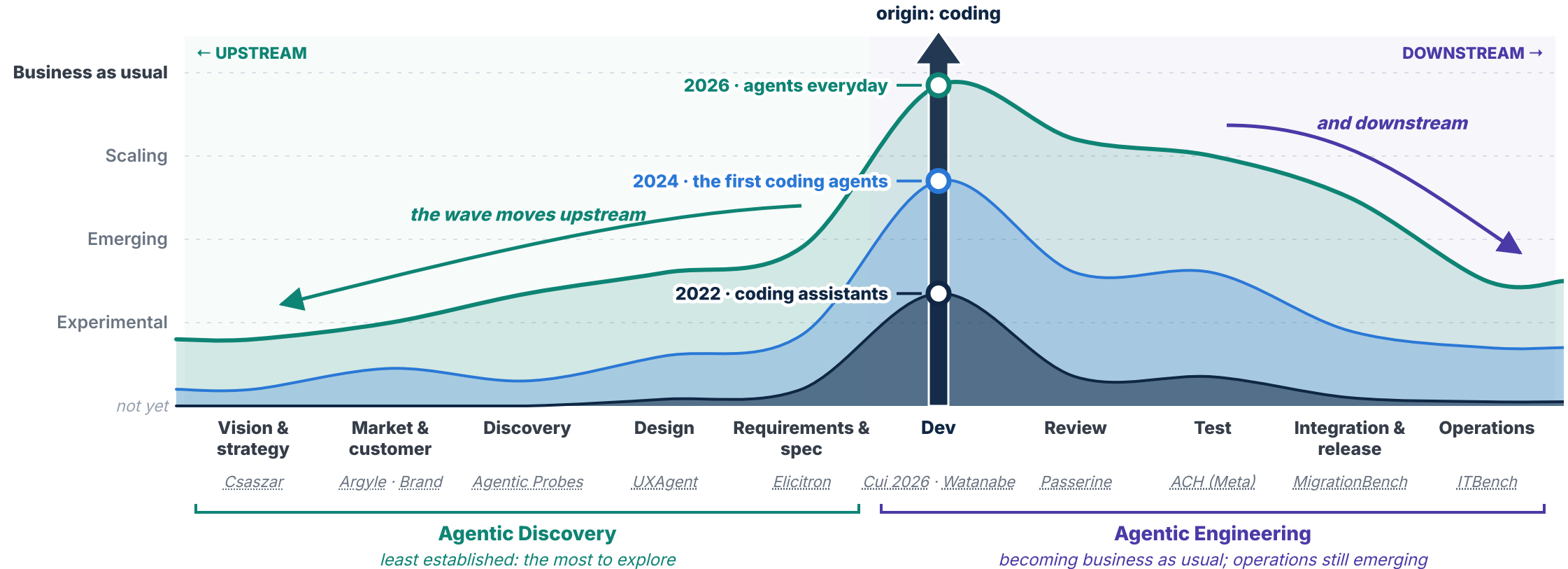
Generative AI in Strategy Work and Firm Performance (AoM 2025, among the best papers) ↗

SYNTHESIS

The AI agentic wave: from Dev, upstream and downstream



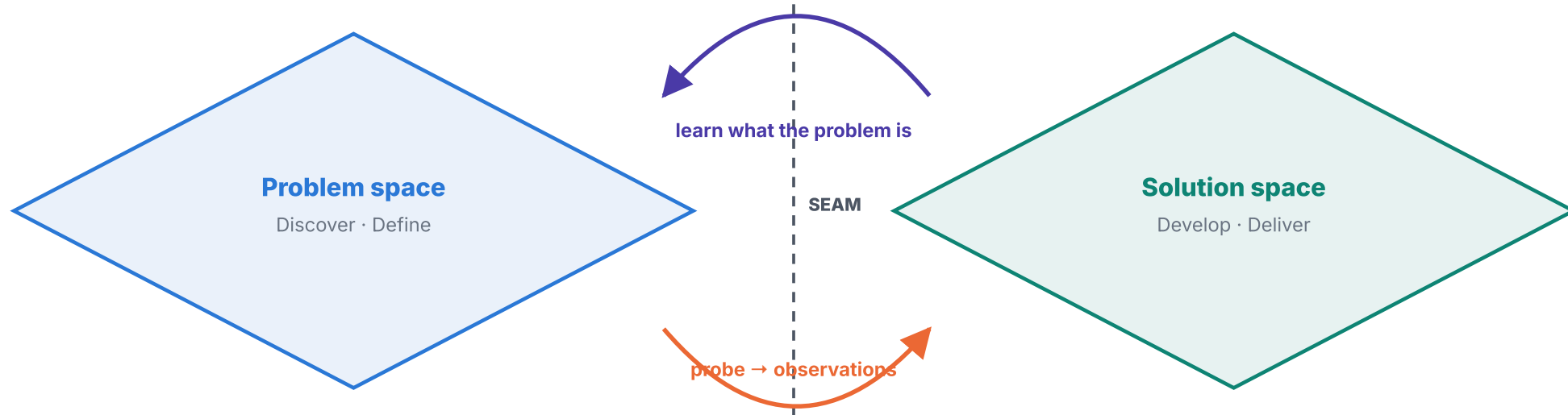
Read and comment



SOURCES Pitkäranta & Pitkäranta 2026

Figure: Pitkäranta & Pitkäranta 2026. Maturity levels are a qualitative estimate, not a measurement; the studies under the phases anchor it (tap a name).

Problem and solution as a joint search



Search both at once

The problem is not fixed before the solution. Discovery looks for the right problem and a fitting solution at the same time.

Learn from the probe

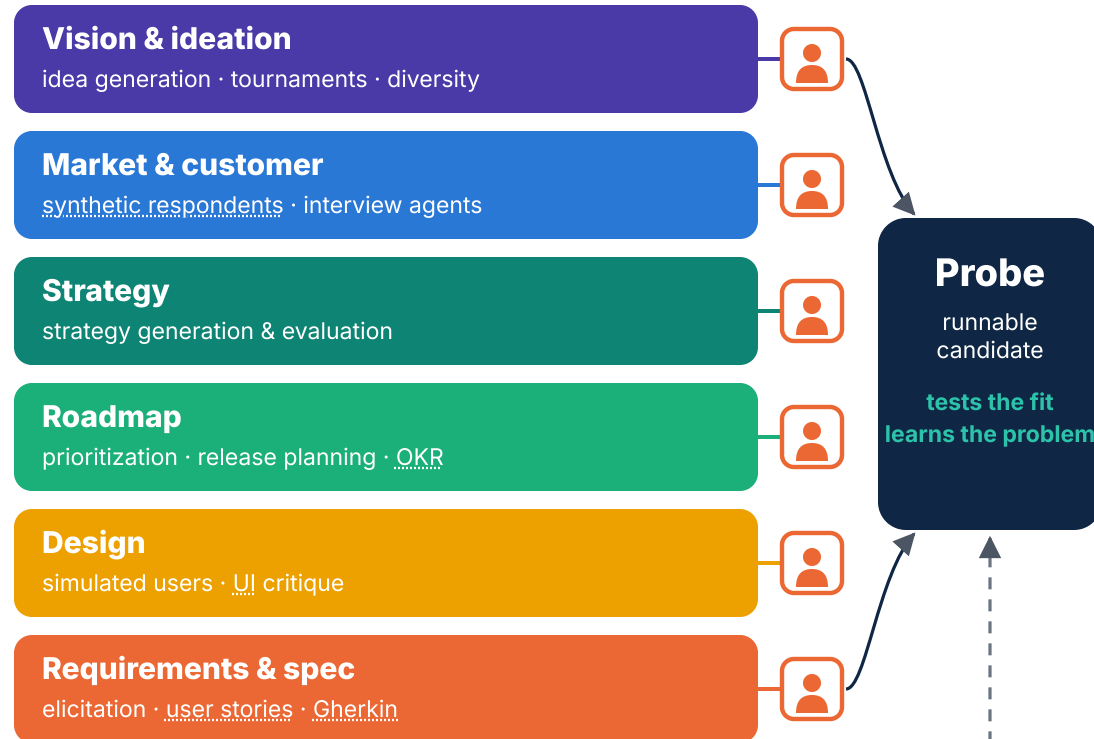
A stakeholder uses the probe, and their feedback shows what the problem really is.

Building is now cheap

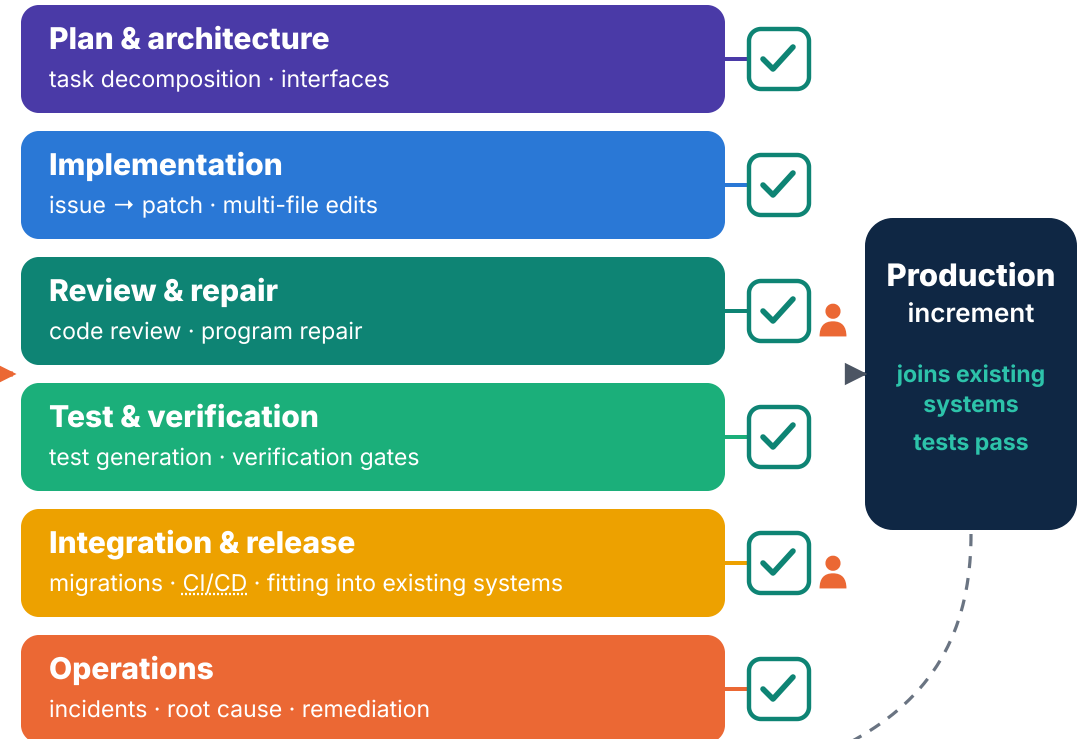
With agents, building a probe costs about as much as describing the problem, so it pays to build in order to learn.

Twelve layers: agents discover, build and run

Agentic Discovery · learning velocity



Agentic Engineering · production reliability



👤 human validation gate ✓ automated gate

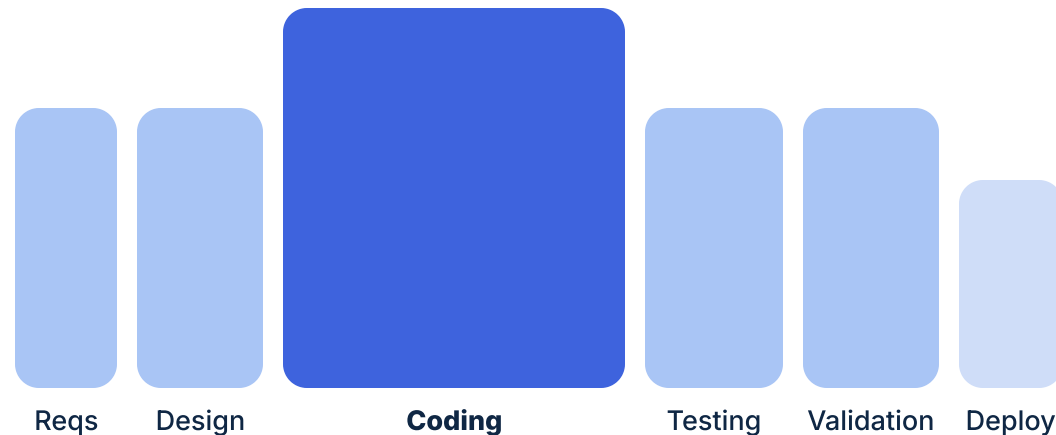
seam: validated spec

usage data: back to Discovery

Coding shrinks; the problem and validation grow

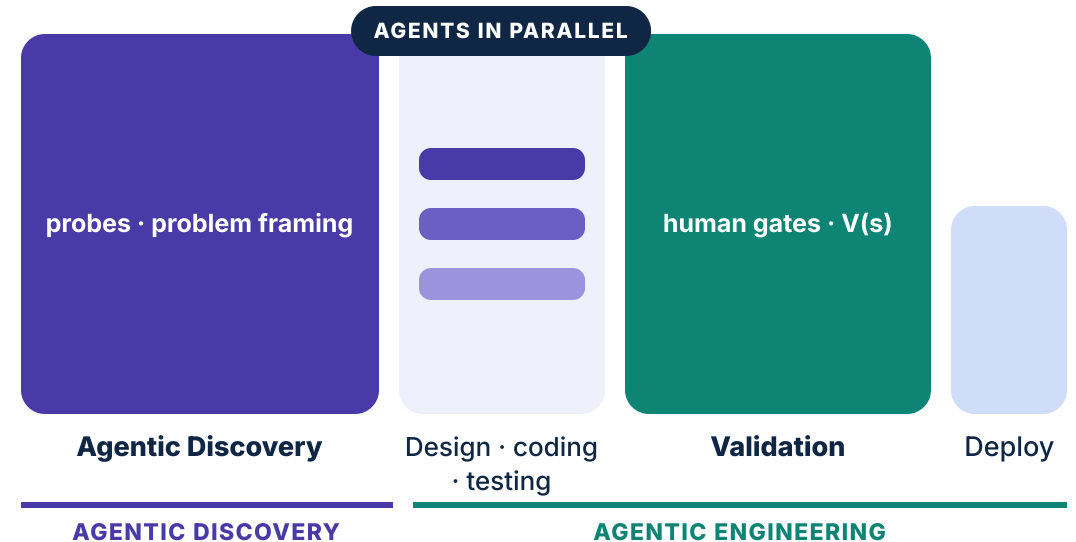
TRADITIONAL SDLC

Most of the time is spent writing code.



AGENTIC SDLC

Time shifts to finding the problem and validating.



Block size shows where human attention goes. When agents do design, code and tests in parallel, human work moves to the two ends: finding the right problem with probes (Agentic Discovery) and validating the result at gates (Agentic Engineering).

INTERLUDE · OCTOBER 2026

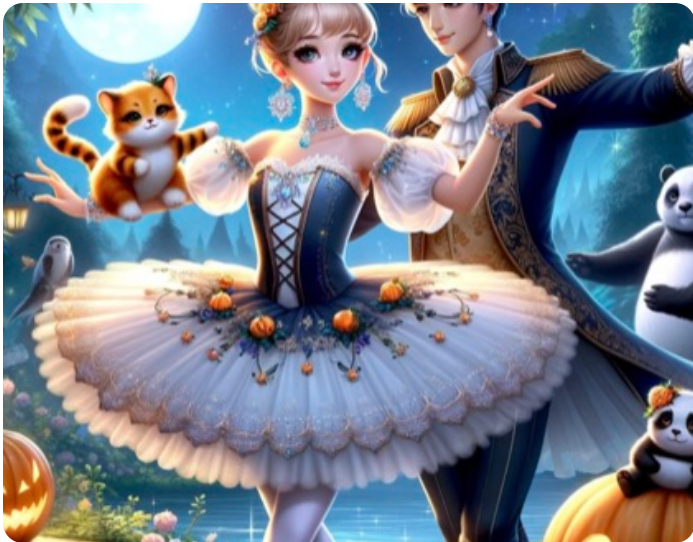
AI status: where we are now

Benchmarks, agents and observations from 2026: what the yardsticks say, where they run out, and what that means for agentic product development.

Figures checked against primary sources on 12 Oct 2026; vendor self-reports are marked.

Look at AI with an open mind, like a child

Children do not ask what AI cannot do. They make things with it.



2023

a 4-year-old makes bedtime stories with AI



2025

an 8-year-old builds a 3D game in hours



2026

an 11-year-old founds a game studio

Software engineering: On a linear scale, almost all the growth is in the last year



The previous slide's log scale shows steady doubling as a straight line. On a linear scale the same exponential is a hockey stick: until 2024 every model stayed under an hour; in spring 2026 the frontier passed the 16-hour measurement ceiling. Vertical bars are METR's confidence intervals; an arrow up = the interval continues beyond the

SOURCES [METR 2026b](#) [METR 2026a](#) [Kwa et al. 2025 \(METR\)](#)

Is the AI shift mostly about programming?



Lana Del Rey - I Hate You (Official Lyrics Video)



Woman in Love | AI Lip Sync Cover - Dana Winner Version #danawinner #coversong #masonmusic

YouTube results: left, the artist's official video; right, an AI-made lip-sync cover (#aimusic).

No.

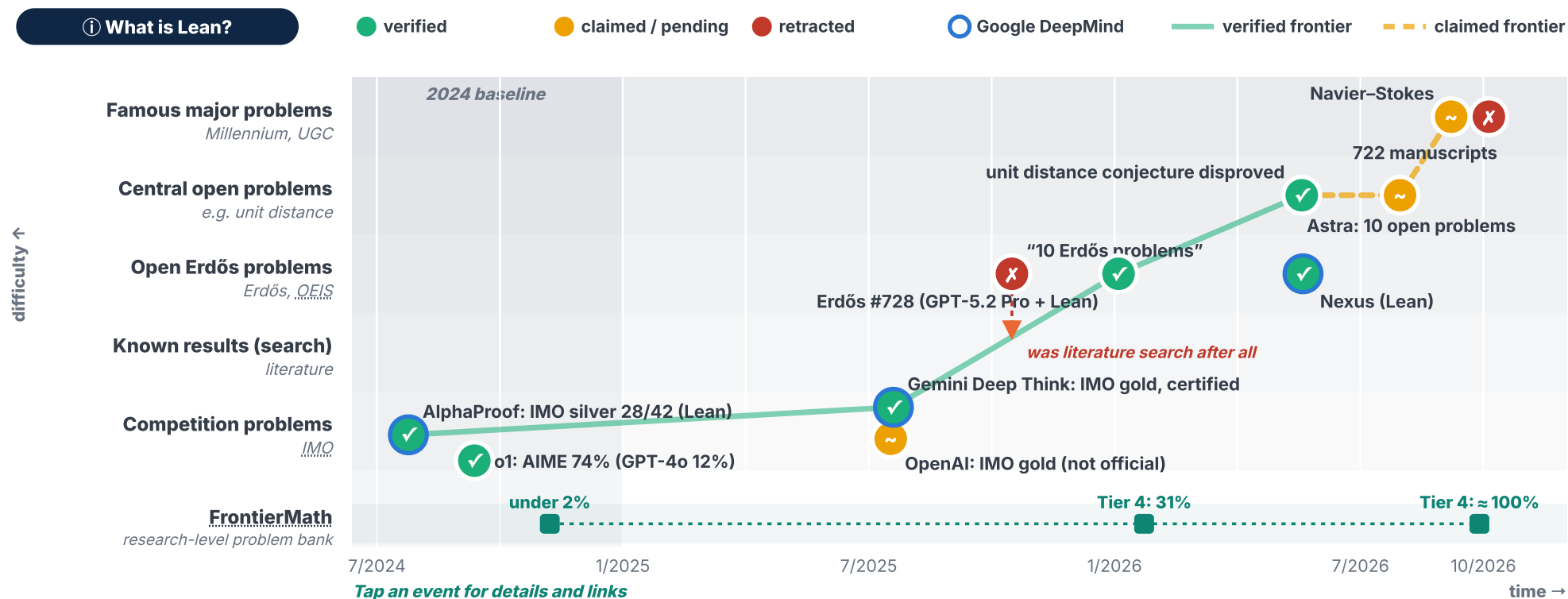
Try it tonight

Search YouTube for your favourite artist with the tag **#aimusic**. AI music is already everyday in the creative industries.

The same holds for knowledge work

Next slide: AI results on real expert tasks across 44 occupations.

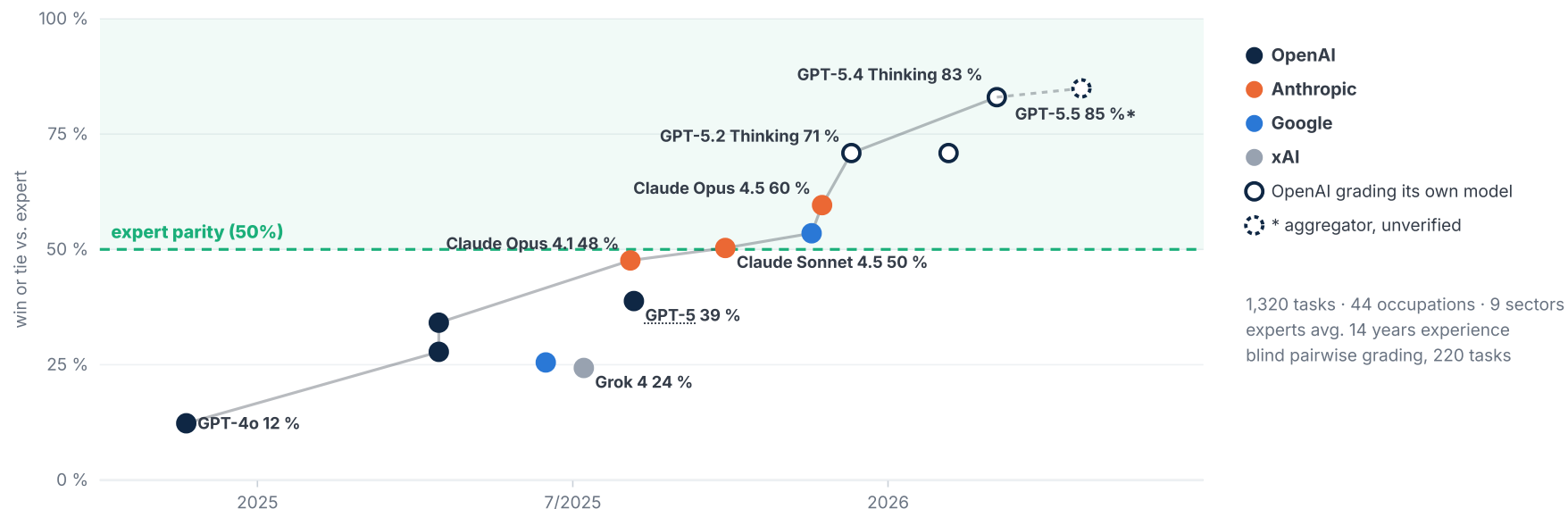
Mathematics: genuine breakthroughs and retractions in the same year



What changed: 5/2026 AI for the first time solved a central open problem of a field, and it survived human verification. FrontierMath Tier 4: 31% → nearly 100% during 2026.

What did not change: the claimed frontier rises faster than the verified one. Verification is the bottleneck: "no human has read it, and no human is going to read it" (Bloom).

Expert work: from 12% to over 80% in eighteen months



Claude Sonnet 4.5 was first above 50% (9/2025). **The metric froze in April 2026:** after GPT-5.5 no new model (GPT-5.6, GPT-6, Claude Opus 5.x) has reported the original win-or-tie figure. Caveats: tasks are precisely specified and one-shot; grader agreement about 71%; 2026 numbers are OpenAI's own; on the harder GDPval+ the top is about 22%.

Successor: GDPval-AA (9/2026)

An independent re-run of the same 220 tasks: agentic harness, LLM judge, Elo scores. Leader **Claude Opus 5.5: 1866 ±26 Elo**. Elo and percentage are different scales, and the scale has been re-anchored three times, so it cannot be drawn on the chart above.

SOURCES

Patwardhan et al. 2025 (GDPval)

Epoch AI 2026c

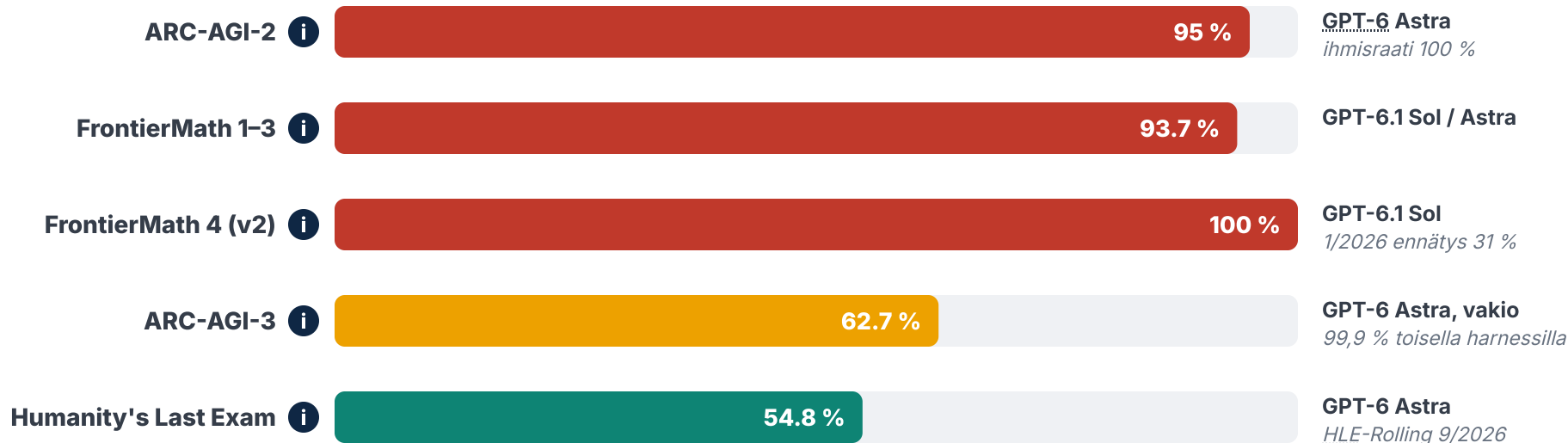
OpenAI 2026

Aggregaattorit 2026

Artificial Analysis 2026b

Snorkel AI 2026

Even the hardest tests are nearing the ceiling



The "hardest" of 2025

ARC-AGI-2, FrontierMath and HLE were the toughest tests of 2025; two are now above 90%.

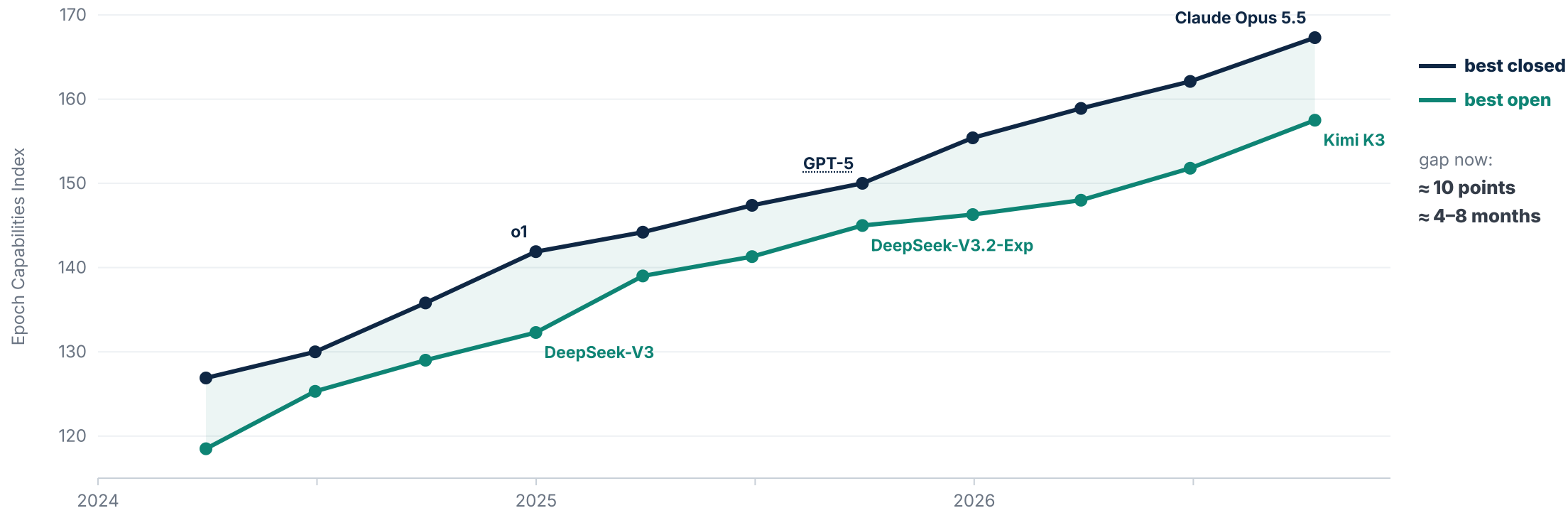
The harness again

ARC-AGI-3: the same model scores 62.7% with the standard harness and 99.9% with another. A model name alone does not tell the result.

Consequence

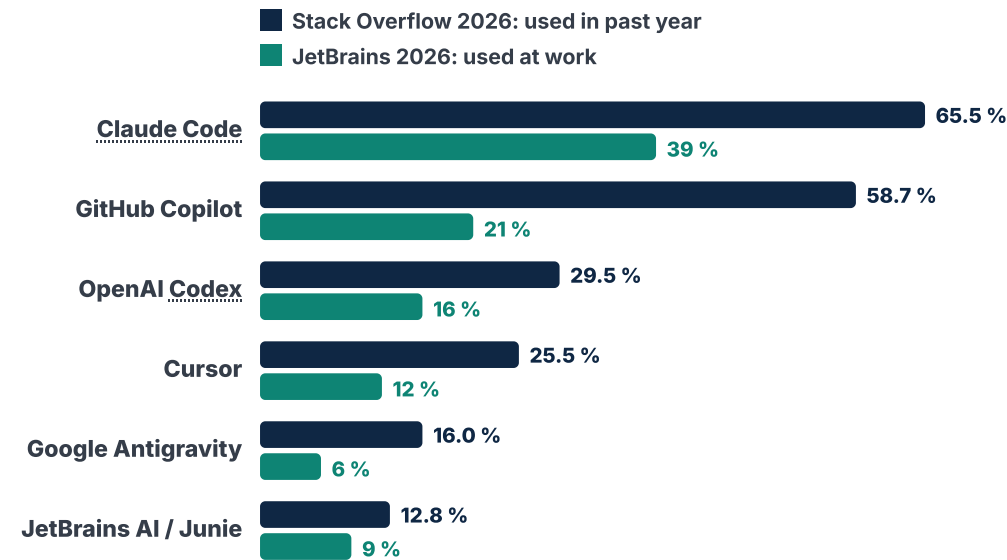
New yardsticks move to real work and long tasks (GDPval, Terminal-Bench, METR's Expenditure Horizon).

Open-weight models follow a few months behind



The open frontier is Chinese (DeepSeek, Qwen, Kimi, GLM); the best Western open models (gpt-oss, Llama 4) are 17–24 points behind. On the Artificial Analysis index the gap shrank in a year from about 13 to 3–6 points, but on the hardest agentic evals it is still about 15.

Coding agents are everyday work, and Claude Code leads in use



Use at work

JetBrains (15,000+ developers, 5–7/2026): 90% use coding agents at work weekly, 68% daily. Claude Code 18% → 39% and Codex 3% → 16% in half a year; Cursor 18% → 12%.

Is the agent's PR accepted?

MSR 2026, 7,156 PRs (2025 data): Codex 77.9%, Cursor 74.5%, Claude Code 71.9%, Copilot 68.0%, Devin 61.6%. Human PRs in popular repos 76.8%.

Surveys measure use, not quality; PR acceptance mixes agent quality with how the agent is used.

PART 0

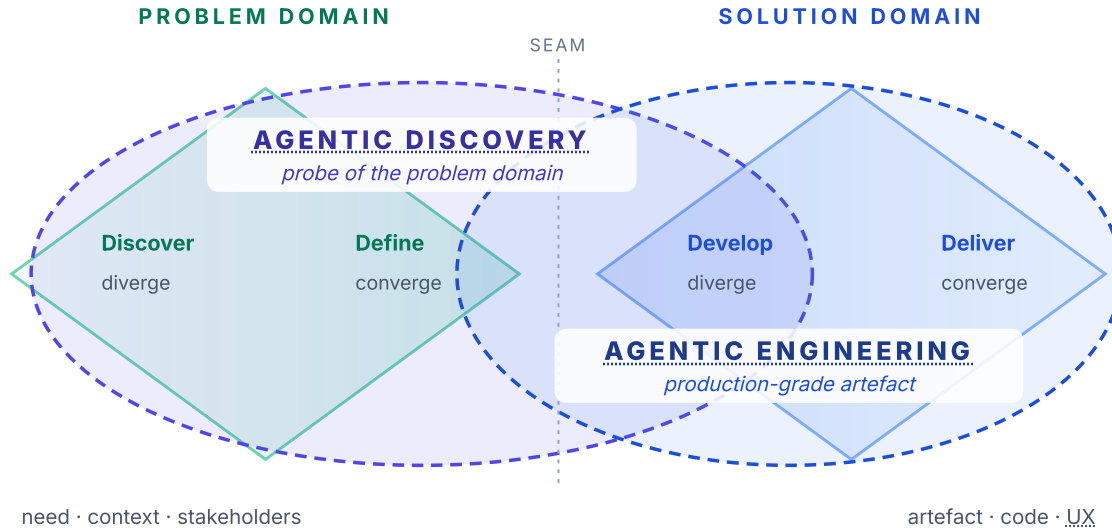
Agentic Discovery & Engineering: the paper's theories, figures and results

Pitkäranta & Pitkäranta (Aalto, 2026). Two-phase SDLC, the Probe, 12 formal frames, 33 traditions and a 69-probe field corpus. Every other part maps onto this frame.

SOURCES

[Pitkäranta & Pitkäranta 2026](#)[Formal sketch \(math.html\)](#)[Traditions gallery](#)[Appendix A.1–A.12](#)

The seam between problem and solution domain becomes permeable



Double Diamond
Design Council 2005

PMF Pyramid
Olsen 2015

Lean Canvas
Maurya 2012

Opportunity Solution Tree
Torres 2021

Problem Frames
Jackson 2001

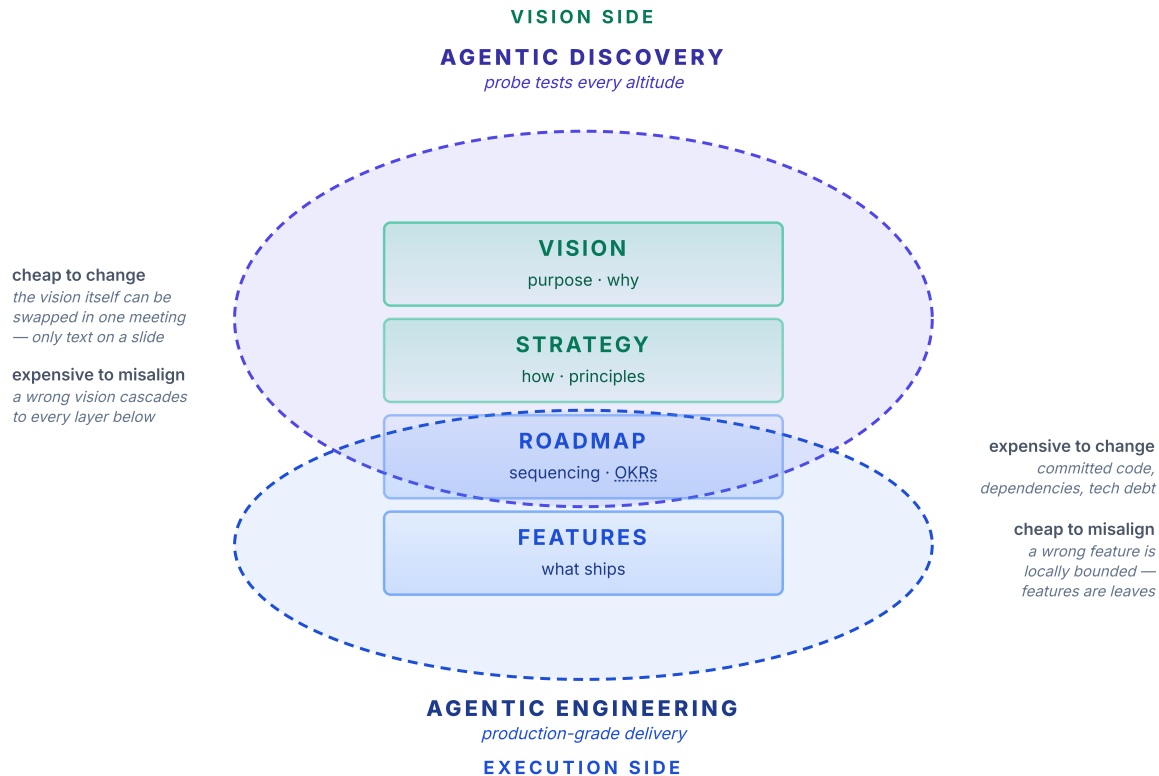
Cynefin
Snowden & Boone 2007

Discovery vs. Delivery
Cagan 2017

Are Your Lights On?
Gause & Weinberg 1990

Hidden premise of all these visuals: the problem can be understood before any artefact exists.

Mistakes at the top cost the most: probe every layer



Golden Circle

Sinek 2009

Vision · Strategy · Roadmap · Backlog

Pichler 2016

Escaping the Build Trap

Perri 2018

Empowered

Cagan & Jones 2020

OKR cascade

Doerr 2018

Hoshin Kanri

Akao 1991

Classical strategy hierarchy

Mintzberg 1994

Vision: cheap to change, expensive to misalign (cascades). Feature: expensive to change, misalignment stays local. A probe tests every altitude.

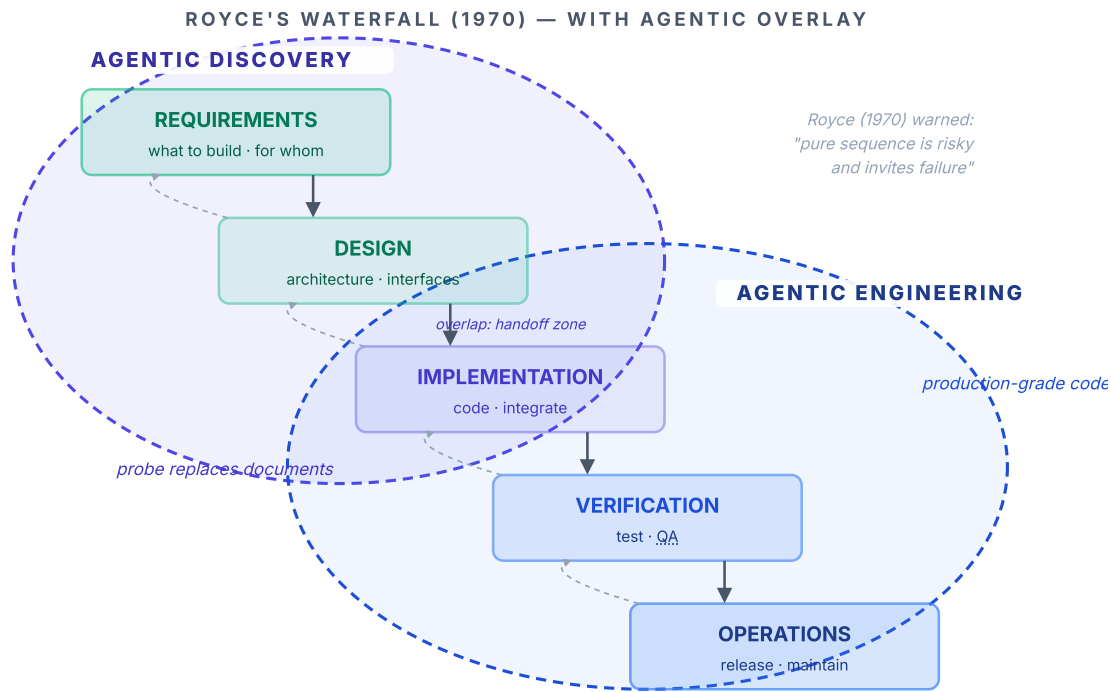
SOURCES

Pichler 2016

Perri 2018

Traditions gallery

Royce's waterfall with an agentic overlay



NATO 1968

Naur & Randell 1969

Waterfall 1970

Royce 1970

Spiral 1988

Boehm 1988

V-Model 1991

Forsberg & Mooz 1991

Scrum 1995

Schwaber 1995

XP 1999

Beck 1999

Agile 2001

Beck et al. 2001

RUP

Kruchten 2003

Lean SD 2003

Poppendieck 2003

Kanban 2010

Anderson 2010

CD/DevOps 2010

Humble & Farley 2010

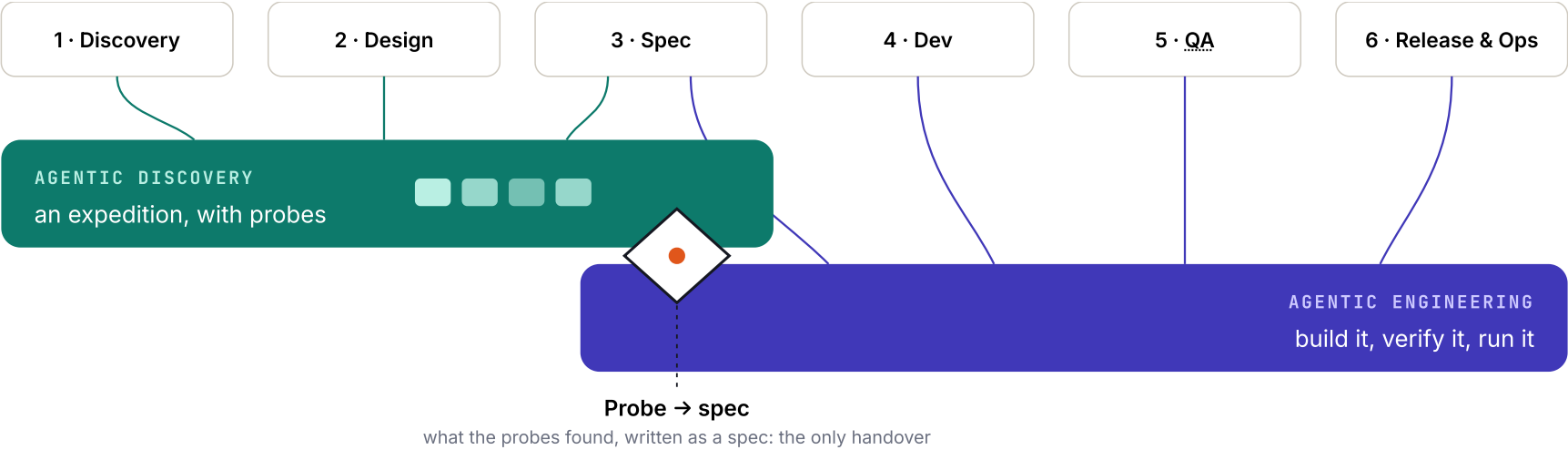
Royce (1970): pure sequence "is risky and invites failure".

SOURCES

Royce 1970

Traditions gallery

The PDLC collapses into two phases



	AGENTIC DISCOVERY	AGENTIC ENGINEERING
Optimises	learning velocity	production reliability
Output	throwaway-grade probe	production-grade code
Spec	extracted from the probe	an input
Cycle	minutes–hours	days–sprints

SECTION

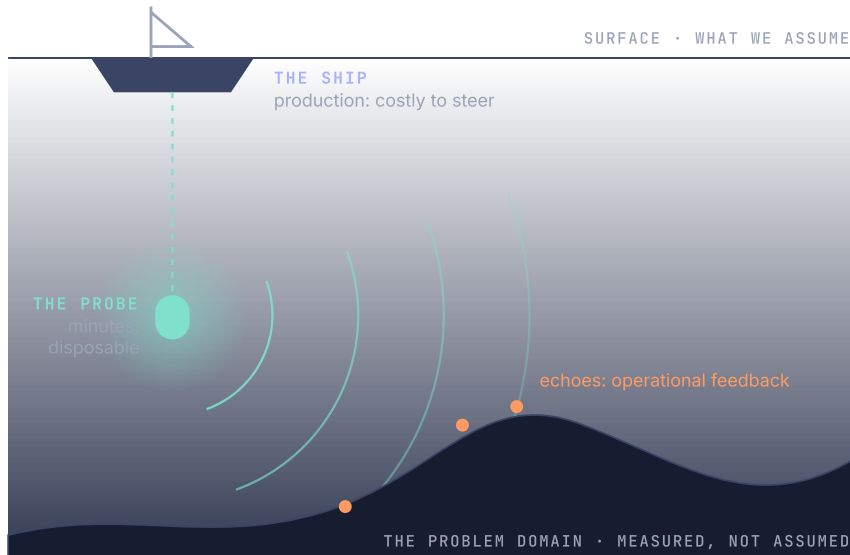
Agentic Probe

The instrument of Agentic Discovery: a runnable, throwaway probe into the problem space, built by an agent from the stakeholder's own words. What it is, why it must be run rather than read, and how it learns about the problem.

SOURCES

Pitkäranta & Pitkäranta 2026

The Probe: a sounding into the problem space



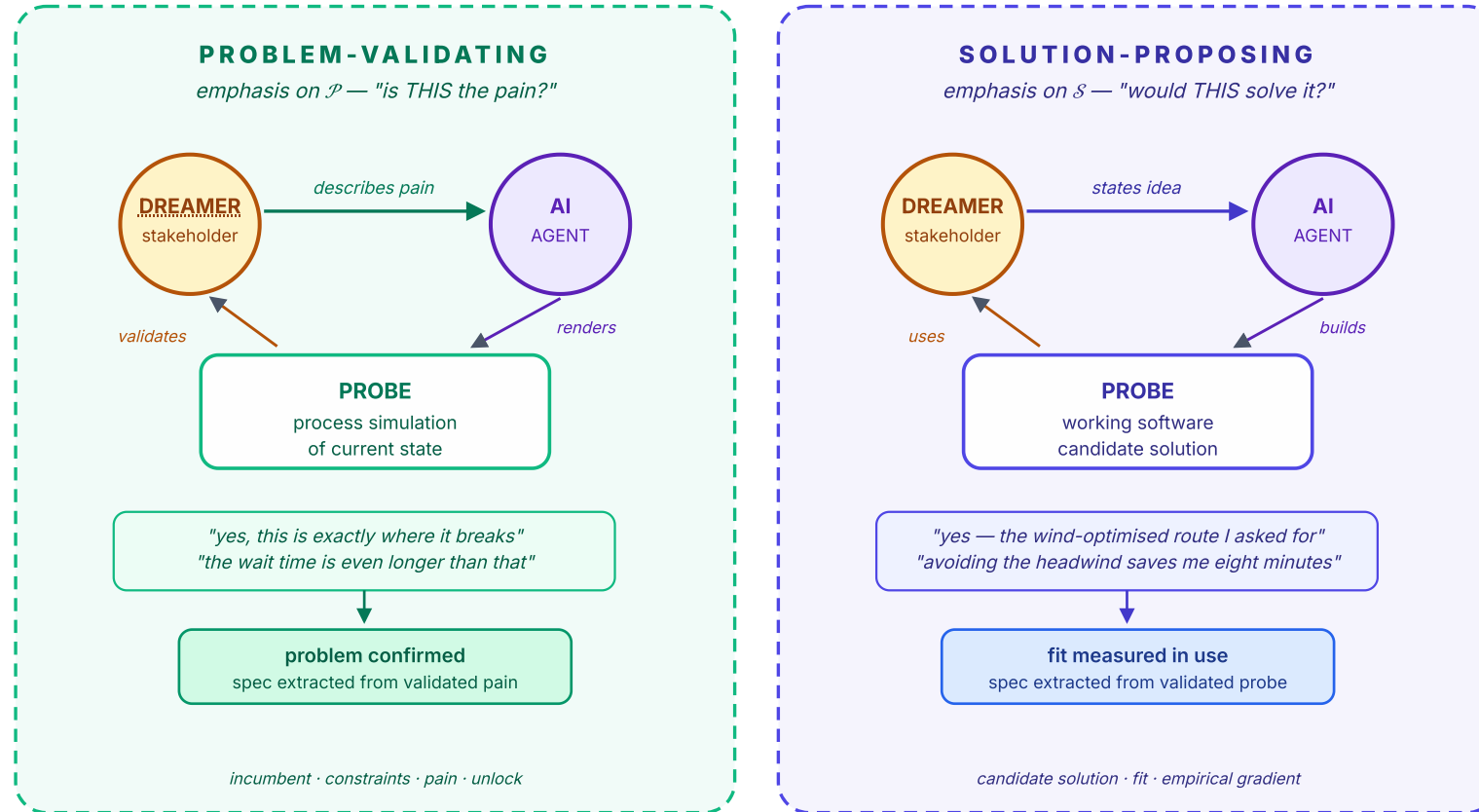
“A **Probe** is a stakeholder-runnable instance of a candidate solution to a stated pain. It is throwaway-grade by design -- its job is to validate or falsify the pain, not to be production-ready.”

Forms

working software · process simulation · walkable future scenario · mixed renderable

Problem-validating and solution-proposing probes

TWO MODES OF AGENTIC DISCOVERY

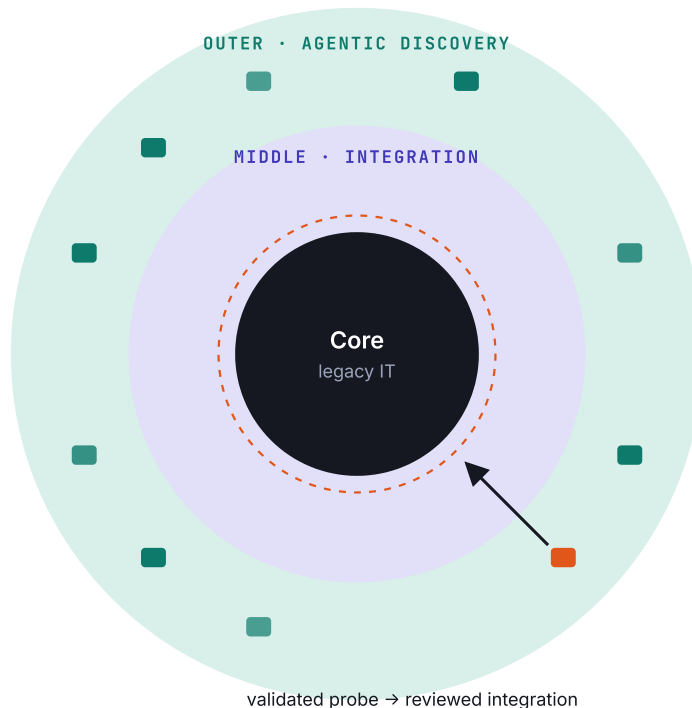


SOURCES

Pitkäranta & Pitkäranta 2026

Formal sketch (math.html)

The probe maps constraints outside-in, inside-out or in between



Outside-in

Start with a new AI process in the outer ring. The probe shows which integrations and core parts are really needed.

Middle-out

Start from the integration layer (APIs, connectors). The probe tests what can change without touching the core.

Inside-out

Interview and reflect the current process as a simulation of the core, then propose paths outward; in regulated settings this eases adoption.

A constraint is a part that cannot or should not change: legacy IT, regulation, data. The probe reveals which constraints are real, i.e. which parts to keep and which to let go. A failed probe costs only in the outer ring.

LITERATURE REVIEW · OCTOBER 2026

Agentic Search & Agentic Information Retrieval

How the concept is defined in the academic literature

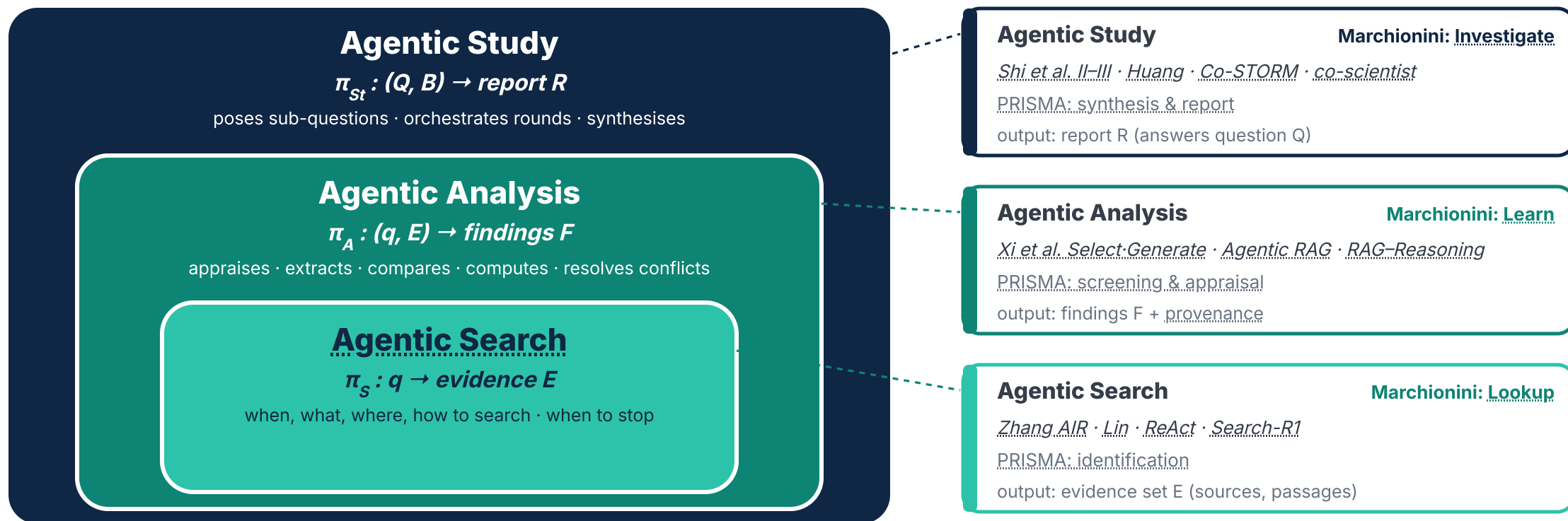
Zhang et al. 2024 · Huang et al. 2025 · Xi et al. 2025 · Lin et al. 2025 · Shi et al. 2025 · Search-R1 · ReAct · ...

SOURCES

Zhang et al. 2024 (Agentic IR)

Huang et al. 2025 (Deep Research Agents)

Agentic Study $\supset$ Agentic Analysis $\supset$ Agentic Search



Each outer layer calls the inner one as a subroutine and decides itself when to call it again. The layers match Marchionini's (2006) three search activities and the PRISMA 2020 review stages; the layering is our synthesis, not any single source's terminology.

SOURCES

Marchionini 2006

Page et al. 2021 (PRISMA 2020)

Shi et al. 2025 (Deep Research survey)

Xi et al. 2025 (search agents)

Zhang et al. 2024 (Agentic IR)

Huang et al. 2025 (Deep Research Agents)

Three nested policies, each with its own output

$$E = \pi_S(q), \quad F = \pi_A(q, E), \quad R = \pi_{St}(Q, B) = \text{synth}\left(\{\pi_A(q_i, \pi_S(q_i))\}_{i=1}^n\right)$$

?

LAYER	DEFINITION	OUTPUT	MARCHIONINI 2006	SHI ET AL. 2025	XI ET AL. 2025 LOOP	PRISMA 2020
Agentic Search	Decides over many steps when, what, where and how to search, and when to stop.	evidence E	Lookup	I · Agentic Search	Plan · Act · Reflect	identification
Agentic Analysis	Appraises, extracts, compares and combines search results; may search again if evidence is insufficient.	findings F + <u>provenance</u>	Learn	I "extracting" + II synthesis	Select · Generate	screening, appraisal
Agentic Study	Poses sub-questions q_i , runs many search and analysis rounds within budget B, synthesises and reports.	report R	Investigate	II Integrated Research · III <u>AI Scientist</u>	the whole loop, repeated	synthesis, reporting

Boundaries. Search ↔ Analysis: does a step change the evidence set E (search) or the interpretation F (analysis)? Analysis ↔ Study: is the question given (analysis) or does the agent pose its own sub-questions and report (study)? Shi et al. do not separate analysis as a phase; that is this synthesis’s addition.

SOURCES

Marchionini 2006

Shi et al. 2025 (Deep Research survey)

Xi et al. 2025 (search agents)

Page et al. 2021 (PRISMA 2020)

Huang et al. 2025 (Deep Research Agents)

Y. Li et al. 2025 (RAG–Reasoning)

SECTION

Agentic Discovery: experiments

What we did and what we measured: the research design (DSRM), the session protocol, eight field engagements with two organisations (Experiment A, 69 probes) and process simulations in healthcare (Experiment B).

A pain in, a runnable probe out, in minutes

ONE SESSION ROUND

1 DREAMER

State a pain or an idea

An employee says in their own words what hurts in their work or what they wish for. About 34 words on average.

2 ORCHESTRATOR

Shape and check

The researcher turns the wish into a prompt, checks it is safe and keeps the queue in order.

3 AGENT

Build and deploy

Claude Code builds a working probe and deploys it to its own URL. Minutes; about 1.7 min per probe in a burst.

4 DREAMER

Use it and react

The stakeholder opens the probe on their own device in the room. Feedback is operational ("save twice fails"), and the next round starts.

Experiment A · field

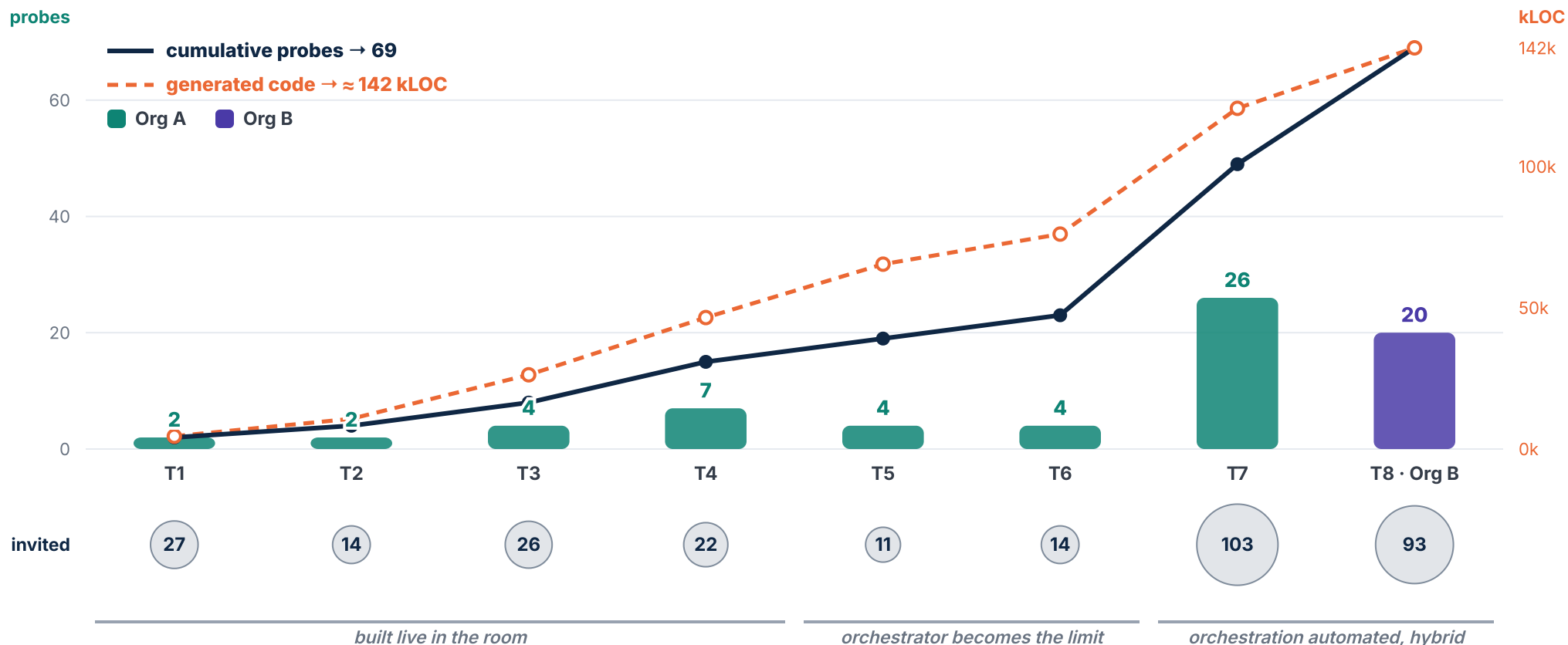
8 engagements in four weeks: a tour of Organisation A's seven offices (49 probes) and a one-day workshop at Organisation B (20 probes). 69 probes and ~142,000 lines of code in total.

What we measured

Time to first probe (minutes), deployability (69/69), spec to code (1 : 56), re-engagement (7/8 sites), safety (0 incidents) and where the bottleneck sits.

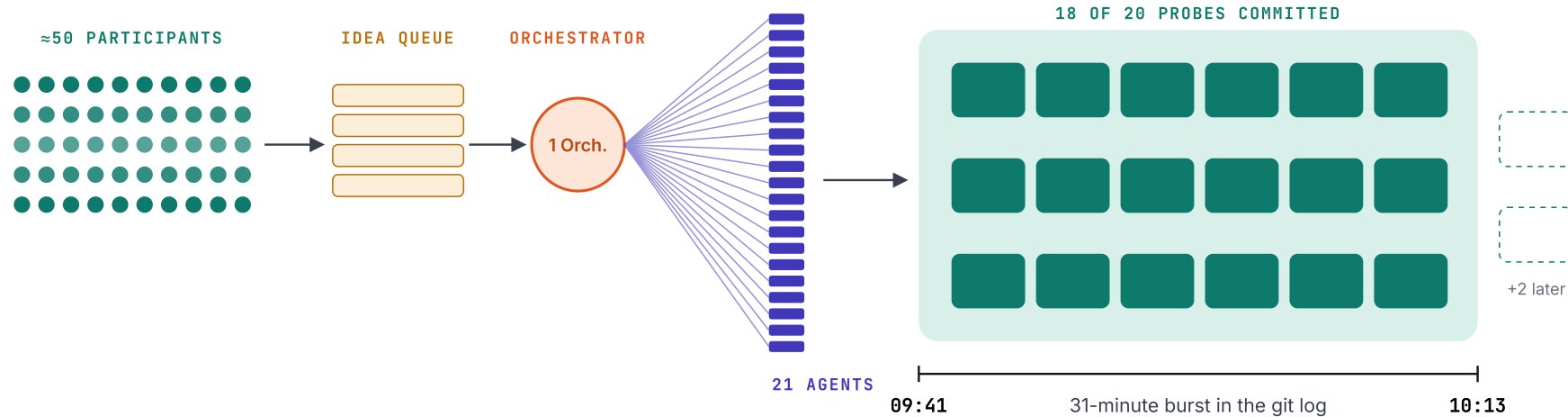
The research design is design science (DSRM): we built the way of working, tried it in real organisations and evaluated it against objectives set in advance. Details on the following slides.

69 probes and ~142,000 lines of code in four weeks



Bar = probes per engagement, solid line = cumulative probes, dashed line = cumulative generated code (kLOC). Circle = people invited (≈ 310 in total; invited is not attended: about 50 were in the room at T8). The two largest engagements produced 46 of the 69 probes.

Getting rid of the bottleneck with orchestrated agentic engineering



Events 1–4

2–4 parallel requests, all built live: production capacity binds.

Event 5

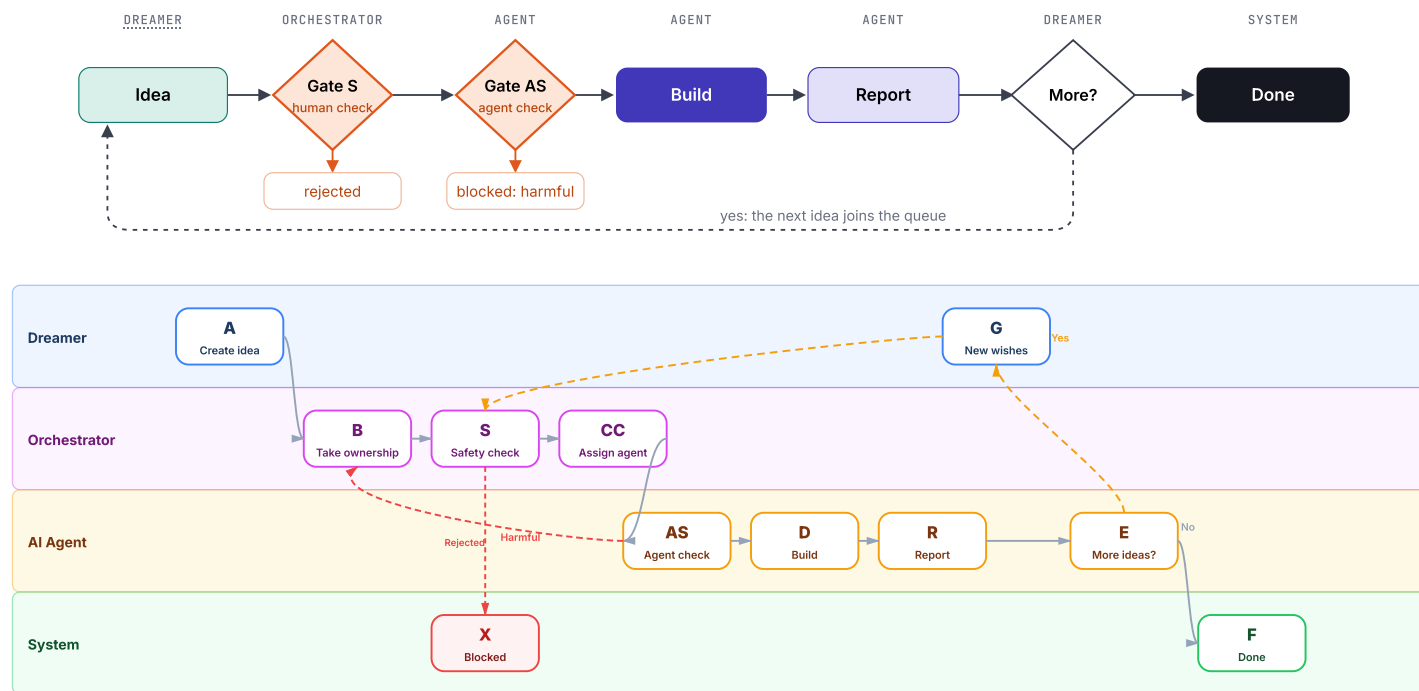
7 requests: the Orchestrator becomes the bottleneck.

Largest

~100 invited; Org B 18/20 probes in 31 min: idea supply and validation bind.

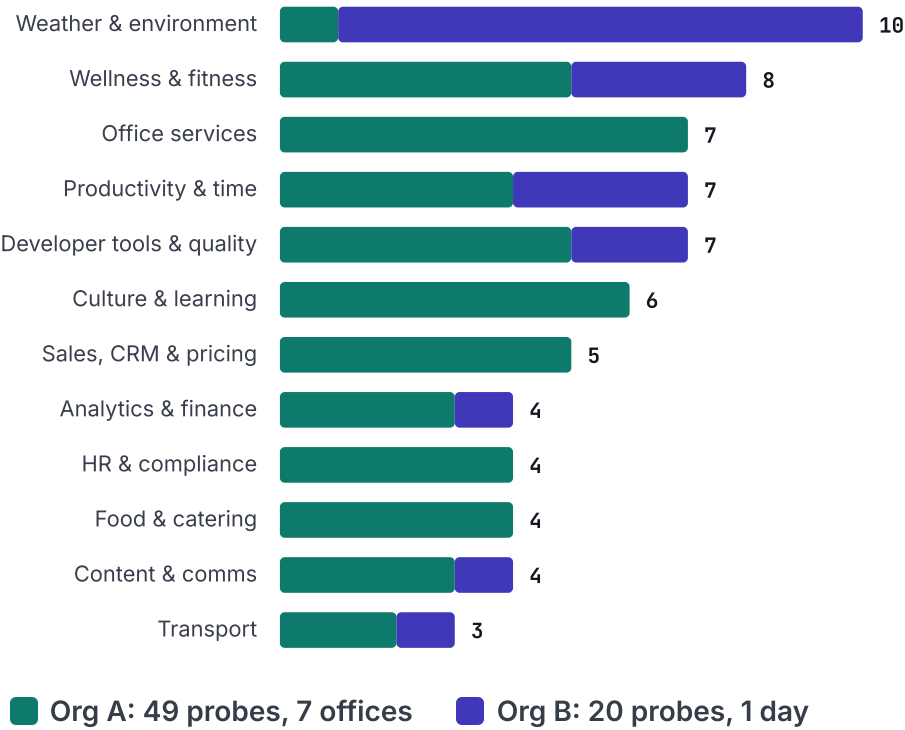
Speed multiplier ~570× on work time (37–65× with a 10× quality discount); ~2,370 person-days traditional vs ~33 h of prompt writing.

"Probe factory": Dreamer, Orchestrator, Agent and two safety gates



Dreamer: stakeholder, natural language or voice · **Orchestrator:** shapes the prompt, safety gate S, curates (skipping it is an anti-pattern) · **AI Agent:** self-check (gate AS, AGENT_SAFETY.md), build, deploy, report. 0 safety incidents in 69 generations.

69 probes, 2 organisations



	ORG A	ORG B	TOTAL
Probes	49	20	69
Spec lines	~1 430	~540	~1 970
Ratio	1:68	1:39	1:60
Span	~4 wk	1 day	

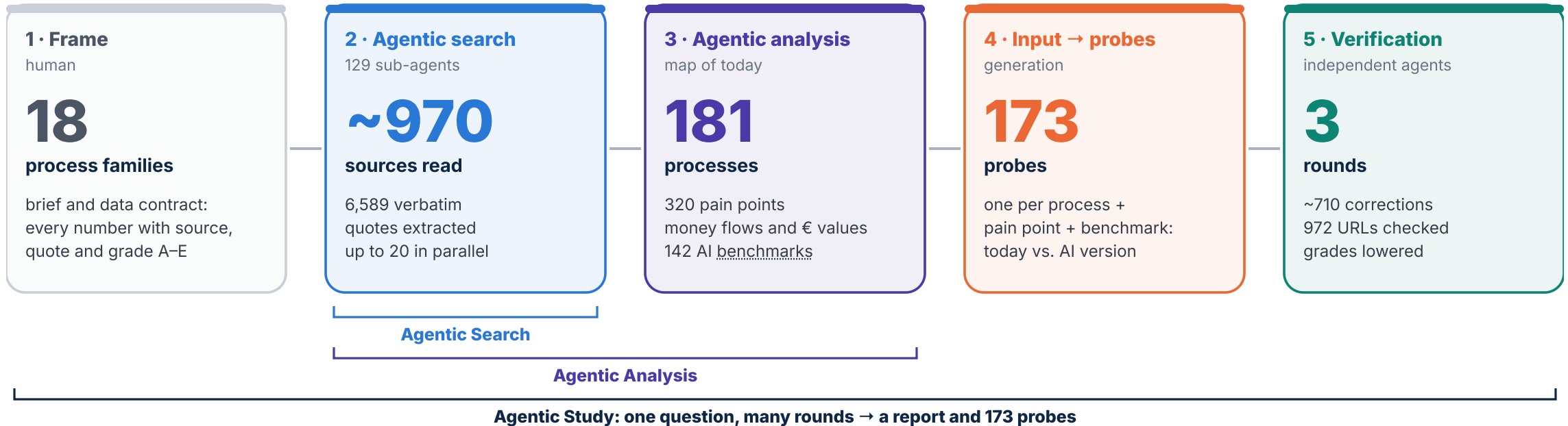
Categories: weather 10, wellness 8, office services 7, productivity 7, dev tools 7... Weather is 45% of Org B.

SECTION

Agentic Discovery in healthcare

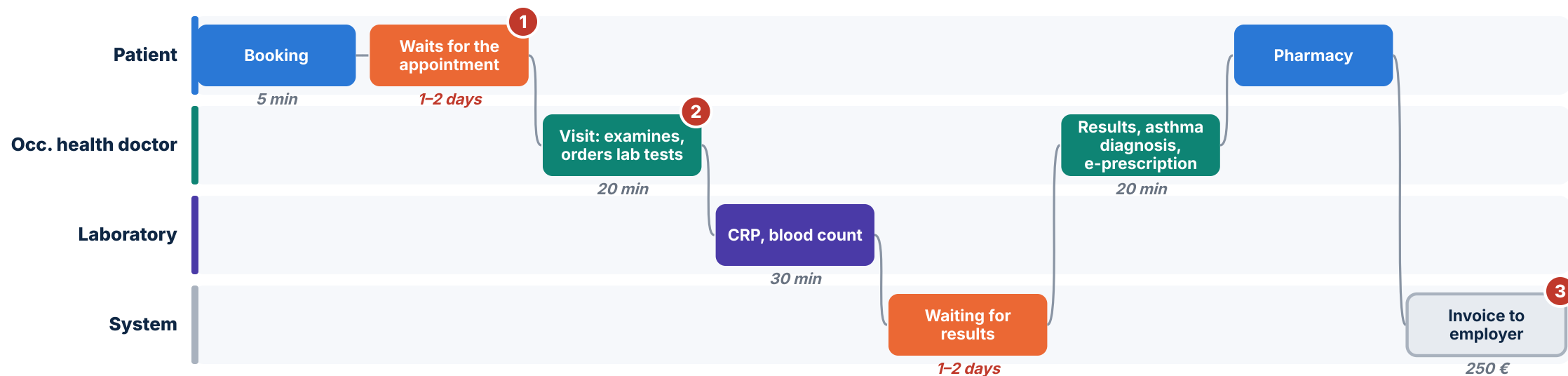
We pointed the method at Finnish public healthcare. Agentic search and analysis built the input, and the input generated one probe per simulation: 173 simulations of today's process and its AI version.

From agentic search to 173 probes in one night



The first complete version was ready in about two hours, the final one with its complement wave and verification in about seven (12–13 Oct 2026). The layers match the concept map in the Agentic Search part.

A familiar condition and a familiar drug, still four days and €250



1 Two waits

the appointment and the lab results, 1-2 days each; the patient is ill throughout.

2 The doctor starts from scratch

earlier bronchitis, the asthma diagnosis and the treatment response are already on record but are not used to prepare.

3 The price of a routine

€250 (doctor €160, lab €90) to the employer for a case whose treatment is the same each time.

≈ 4 days

total duration

250 €

cost

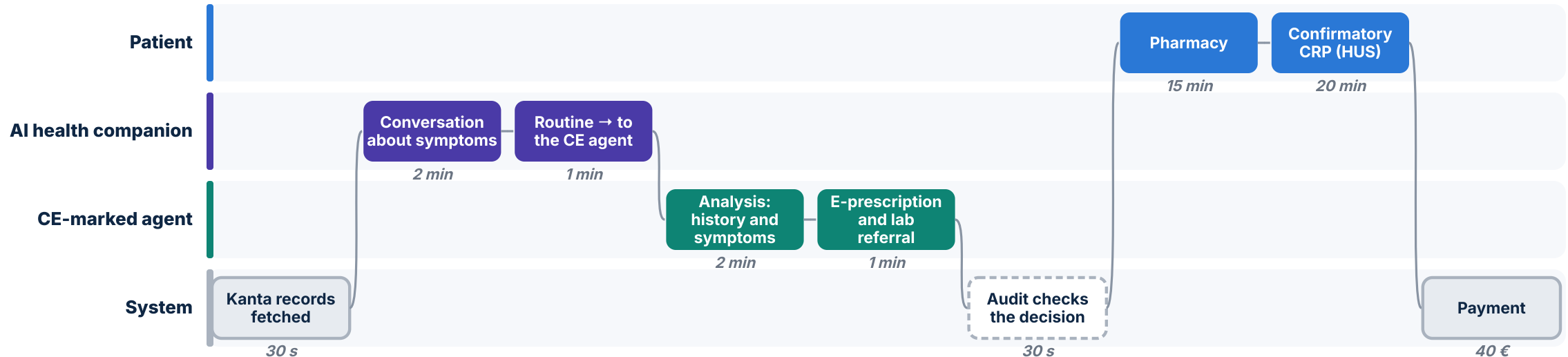
5

steps done by people

0

steps where the history prepares the case

Fully agentic: 42 minutes and €40, two errands for the patient



WHAT WAS RELAXED (K → S*)

Regulation: a CE-marked software agent (a medical device) may renew a prescription for a known, diagnosed condition. **Data:** the Kanta history is available to the agent before the conversation. **Billing:** a Kela agent service and a public lab across the sector boundary. **Safety gate:** certified quality assurance audits every decision; an anomaly routes to a doctor.

This is a possible future, not today's practice: in Finland a prescription is renewed by a doctor today, not by a software agent. That is exactly why it is a probe: the stakeholder sees which constraints κ are real and what relaxing them would be worth, $V(p; s_0)$.

≈ 4 days → 42 min

total duration

250 € → 40 €

cost

Agentic search in numbers

129

sub-agents launched, up to 20 in parallel

~970

sources read; 924 ended up cited

6,589

verbatim quotes backing the numbers

~4,400

data points graded A–E

~710

corrections in three verification rounds

~7 h

start to finish; first version ~2 h

How strong is each data point?



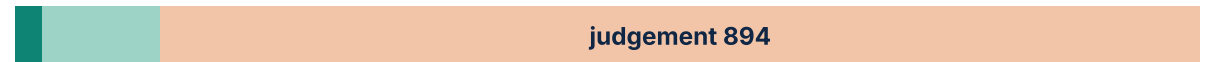
D 200 · E 317 · V 24 (A = strongest, E = weakest, V = changed by a verifier)

Did the sources respond when checked?



972 URLs: 39 blocked bots, 4 did not respond

Where do today's step durations come from?



1,019 steps: source 23 · estimate 102 · judgement 894; step timings are not published

The biggest gap shows in the bottom bar: today's durations are mostly judgement. That is exactly what a problem-validating probe collects from people.

SECTION

Summary

The deck in two slides: what we claim, on what evidence, and what it means for product development.

Building got cheap, so finding the right problem is the new bottleneck

1 Where we are

The agentic wave ignited in writing code and spreads from there both ways. Downstream, **Agentic Engineering** is maturing fast: coding agents are everyday tools, and review, test and release are following. Upstream, **Agentic Discovery** is the least established and has the most to explore.

2 Core claim

When execution gets cheap, the SDLC collapses into two phases: **Agentic Discovery** finds the right problem (learning velocity) and **Agentic Engineering** builds it right (production reliability). The seam between them is the validated spec.

3 Instrument

The **agentic probe** is a runnable, throwaway instrument into the problem space: problem-validating or solution-proposing. It must be run, not read, and it also reveals the constraints.

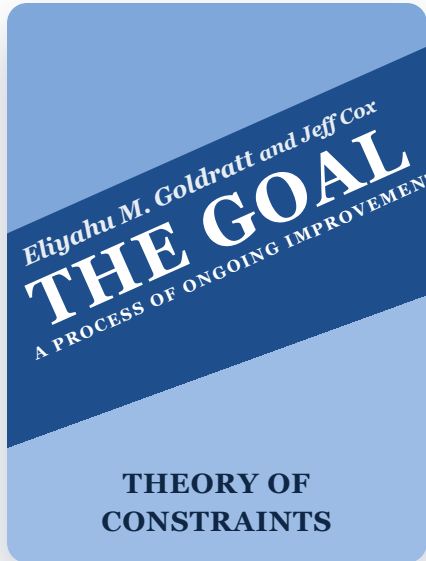
4 Evidence

8 field engagements, 2 organisations, 69 probes and ~142,000 lines of code in four weeks. Intent-to-code 1 : 56 on average (up to 1 : 467). The bottleneck moved from production to orchestration, then to idea supply and validation.

5 Engine

Discovery is agentic search over $\mathcal{P} \times \mathcal{S}$. Agentic Search $\subset$ Analysis $\subset$ Study; the human acts as an ask-or-act gate, and autonomy is a design choice, not a capability.

The bottleneck moves, and we have only scratched the surface



Goldratt & Cox 1984: a constraint is not removed, it moves. [About the book ↗](#)

BEFORE

Writing code

Agents made building cheap, and this bottleneck opened up.



WHAT WE SAW

Two new bottlenecks

1 The supply of ideas

Agents built faster than ideas arrived: 18 of 20 probes in 31 minutes.

2 The operator's context switching

One person steering many agents at once, and their attention set the pace.



NEXT

The wave reaches Agentic Discovery

Upstream, many more bottlenecks wait to be found and solved. **We just scratched the surface.**

SOURCES

Goldratt & Cox 1984

Pitkäranta & Pitkäranta 2026

TAPIO PITKÄRANTA & LEENA PITKÄRANTA · AALTO UNIVERSITY · 2026

Thank you

Everything is openly online. Read and comment.

 The paper in a commentable reader

 This long slide deck and its PDFs



[longshoreforrest.github.io/
publications/agentic-discovery/](https://longshoreforrest.github.io/publications/agentic-discovery/)

Building got cheap. The bottleneck is now the right problem.

