GLOBAL SUMMIT AI VANCOUVER · MAIN STAGE

From AI agents to reliable enterprise systems

What it takes to move beyond the demo

GLOBAL SUMMIT
AI VANCOUVER 2026

PRESENTED BY:

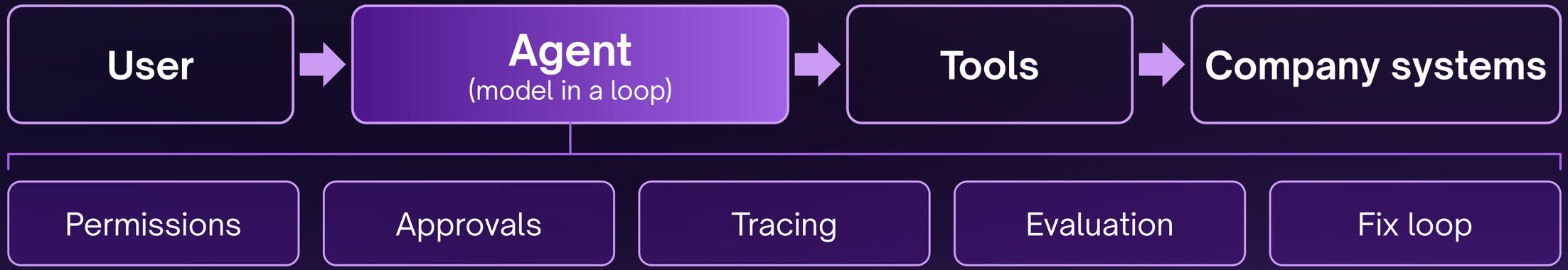
Bas Brekelmans
Microsoft

GLOBAL SUMMIT
AI VANCOUVER 2026

Agenda

1. Why the last 20% takes most of the time
2. Giving agents work
3. Permissions and approvals
4. Tracing and monitoring
5. Evaluation
6. Fixing failures automatically

Everything in this talk is part of this system

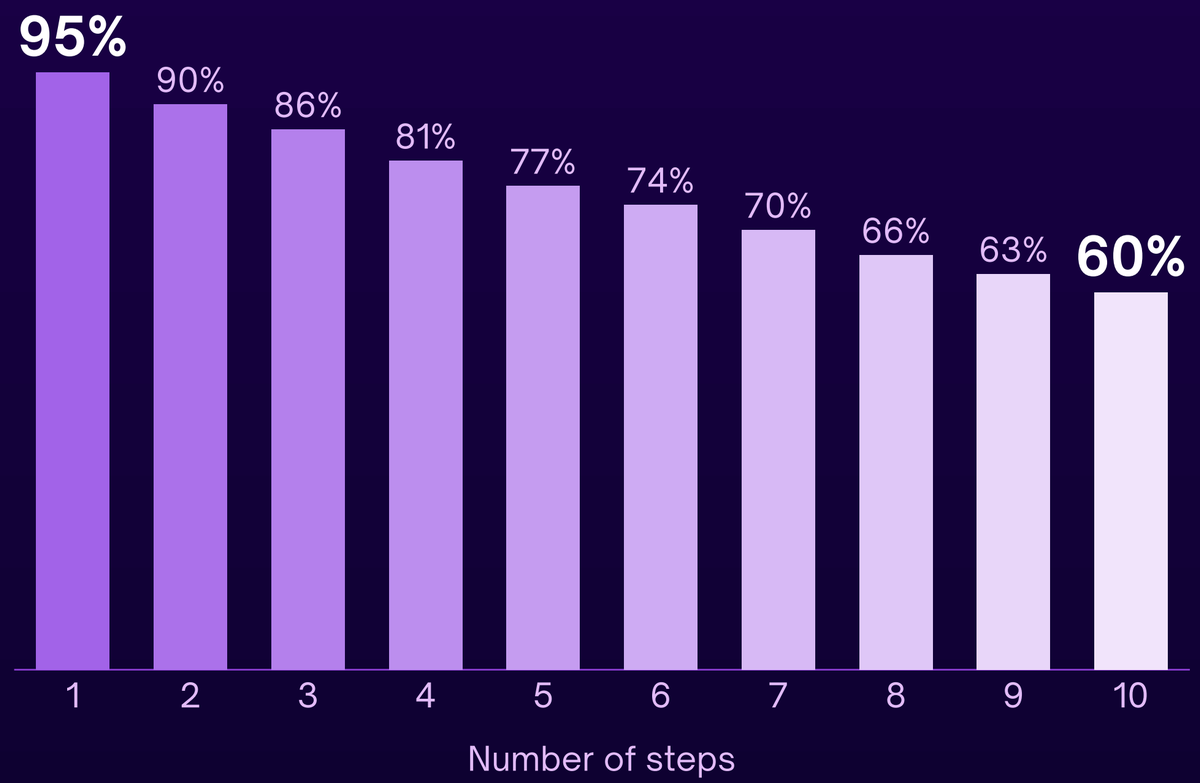


The model is one box. The rest makes it reliable.

Reliability degrades with complexity

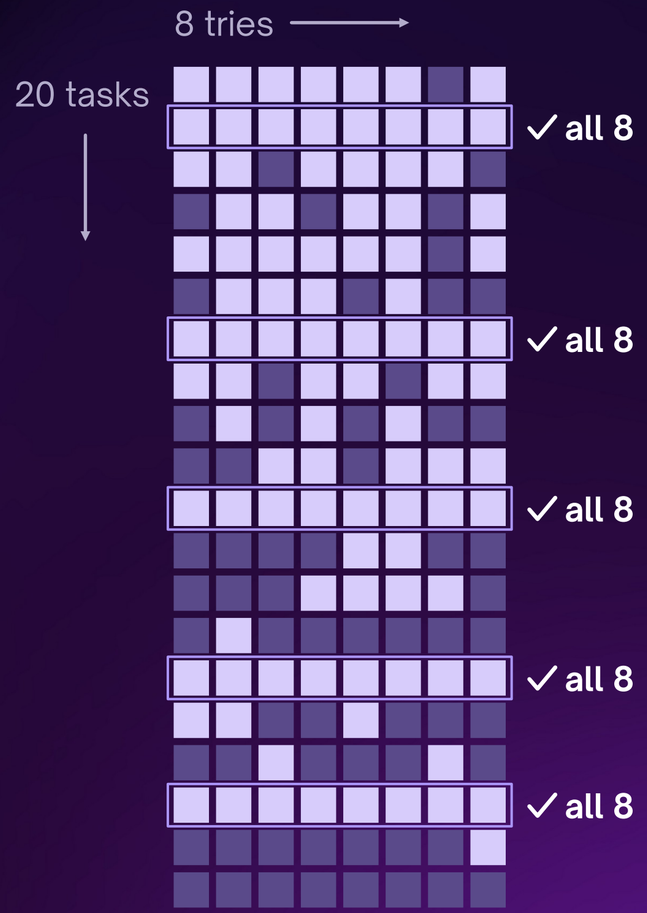
Each step is right 95% of the time.

After 10 steps, the whole task succeeds about 60% of the time.



GLOBAL SUMMIT
AI VANCOUVER 2026

Consistency is not guaranteed



Try succeeded

Try failed

60%

of single tries succeed

25%

of tasks succeed on all 8 tries

Invisible errors persist

Wrong record

The user has two accounts. The agent picked one.

Invented data

A tool returned nothing. The agent filled the gap.

Retry loop

A call kept failing. The agent kept retrying. Cost grew.

False success

The agent reported done but skipped a step.

Bad output

The text had characters the database rejects.

Give a goal and tools, not step-by-step instructions

INSTEAD OF

1. Open the CRM.
2. Search for the customer.
3. Copy the last three emails.
4. Summarize them.
5. Paste into the doc.

Breaks on any case you did not script.

DO

Goal:	Prepare me for the Thursday customer review.
Context:	Account folder, notes, open support cases.
Limits:	Do not contact the customer.
Done when:	One page. Every number has a source.

The agent knows the context and the goal

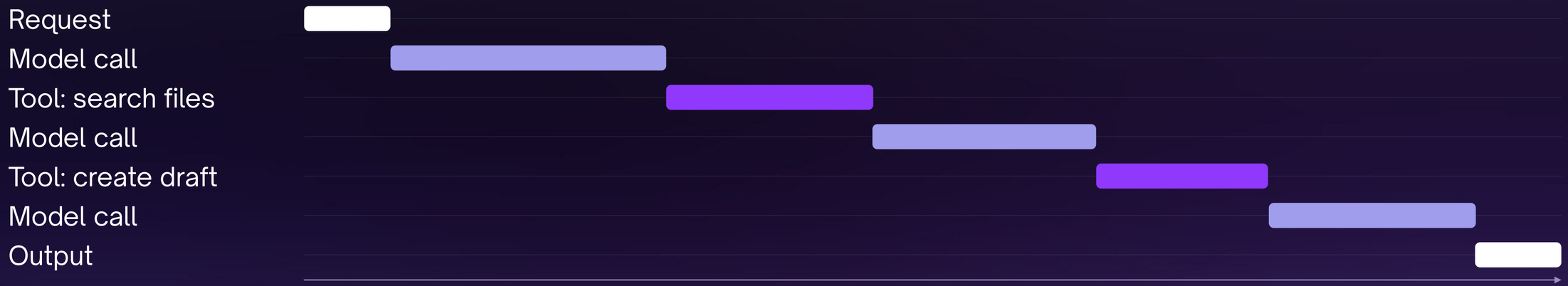
Permissions and approvals

The user's identity comes from the server, not from the model.

Loop iterations and cost per run are capped.

Action	Rule
Read data	Runs, with the user's permissions
Reversible change	Runs, and is logged
Irreversible action: send, pay, delete, release	**Needs a person's approval**
Decisions about people, money, legal terms	**Not delegated**

Tracing: one trace per run, one span per step



RECORDED FOR EACH RUN

Input · sources read · tool calls and arguments · results · model and prompt version · tokens · cost · time · outcome

Monitoring: most failures do not return an error

HTTP 200 OK

{ "status": "done",
"total": "$4,120" }

Correct total: $3,980

WHAT WE WATCH

- Quality score on a sample of real runs
- Steps, cost and retries per run
- Tool error rates
- Changes after a model or prompt update

Evaluation: build the test set from real failures

1

Read 50 to 100 real traces. Write down what went wrong.

2

Group the failures. Fix the most common first.

3

Turn each failure into a pass/fail test.

4

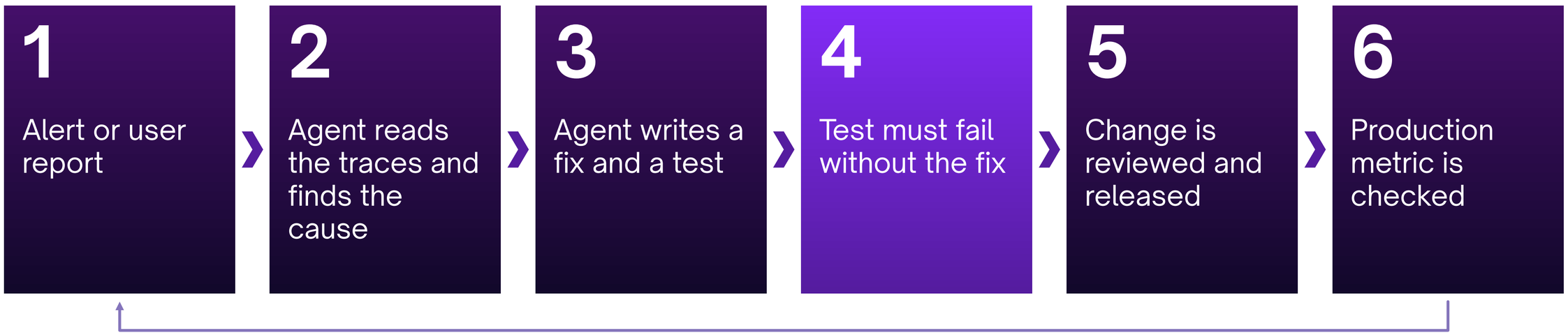
Use code checks where possible. Use a model as judge where not, and check it against human labels.

5

Run the tests before every change. Score a sample of production runs.

Method: Hamel Husain and Shreya Shankar. Judges need checking: Shankar et al., UIST 2024.

Fixing failures automatically



Each fixed failure adds a test, so it is caught if it returns.

Checks on the fix loop

1

The agent that wrote a fix does not decide whether it worked.

2

A test must fail when the fix is removed.

3

A fix is done when the problem stops in production, not when the code is merged.

4

Each agent has one named owner who can turn it off.

Models are unreliable at correcting their own work without an outside signal (Huang et al., ICLR 2024).

Most of our human effort goes into the system around the model.

How our team works now

ENGINEERS

- Write specs
- Review plans
- Own permissions, tests and alerts

AGENTS

- Write and fix code
- Investigate alerts
- Add tests

- [ ] The goal and "done when" are written down.
- [ ] It has the right permissions. Irreversible actions need approval.
- [ ] Loop count and cost are capped.
- [ ] Every run is traced.
- [ ] Quality is measured on sampled production runs.
- [ ] Failures become tests.
- [ ] One named owner can turn it off.

Before an agent goes to production

GLOBAL SUMMIT
AI VANCOUVER 2026

Questions