PANEL 1 · DIGITALIZATION AND ARTIFICIAL INTELLIGENCE IN CHINA

The Road to JARVIS

Where the Moats Are for Personal Agents

Five technical moats are being built: three are taking shape, two still need new ideas.
Three non-technical moats decide who wins: distribution, context and permission.

Dr. Bojie Li
Independent Researcher · Former Chief Scientist, Pine AI

Portugal–China International Conference
Lusophone University, Porto · 10 November 2026

THE VISION

A 47-year-old dream is now a five-year problem

Every generation has pictured the same machine: you talk, it understands, it gets real work done.

1979

MIT "Put-That-There"

A user speaks and points; the system moves objects.

1987

Knowledge Navigator

Apple's agent finds papers and calls a colleague.

2008

JARVIS, *Iron Man*

Cuts in mid-sentence and runs the workshop.

2013

Samantha, *Her*

Laughs while you talk; has changed by month six.

2024–26

The pieces ship

Dec '24 Pine: first personal agent
Mar '25 Manus: first general agent
May '26 Interaction models (Thinking Machines Lab)
Sep '26 Meta Muse: first widespread personal agent

THE SAME SCENE

OpenAI opened its September 2026 launch video with the **1979 MIT clip**.

COMPUTER USE

GPT-6 Astra passes **72.6%** of a standard computer-use test, but takes **~40 min** per task.

REAL-TIME VOICE

Full-duplex ships in 2026: **Seeduplex** (Apr), **TML** (May), **GPT-Live-1** (Jul; **+30 pts** on Full-Duplex Bench), **StepAudio 3** (Sep).

> "These five things, I know there's vague work going on... And then there are two additional things that I have no idea how to do... That's kind of like the **perfect shape of a problem**... like five years."
>
> Jeff Dean, with Dawn Song, 2026

JARVIS IN 2026 HAS EXACTLY THAT SHAPE

1 · Interaction models	2 · Agent-native OS	3 · Fast inference	4 · Physical-world reach	5 · Continual learning
◐ taking shape	✕ missing	◐ taking shape	◐ taking shape	✕ missing

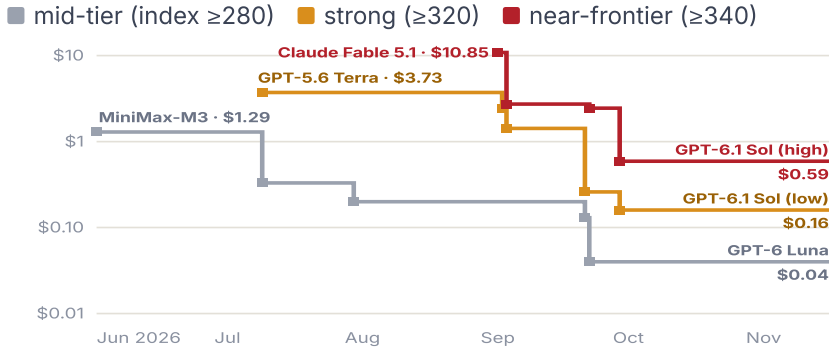
Taking shape: work under way; it works today, but slowly, expensively or not yet at scale. **Missing:** needs new ideas.

Bolt, SIGGRAPH 1980 · Apple, 1987 · GPT-6 Astra, Sep 2026 (vendor figures) · Thinking Machines Lab, May 2026 · Seeduplex, GPT-Live-1 (OpenAI, vs. GPT-Realtime-2.1), StepAudio 3 (2026) · Pine AI, Dec 2024 · Manus, Mar 2025 · Meta, Sep 2026 · Jeff Dean & Dawn Song, 2026 Frontier & Pioneer Symposium (A Letter a Day #341)

CHEAP INTELLIGENCE

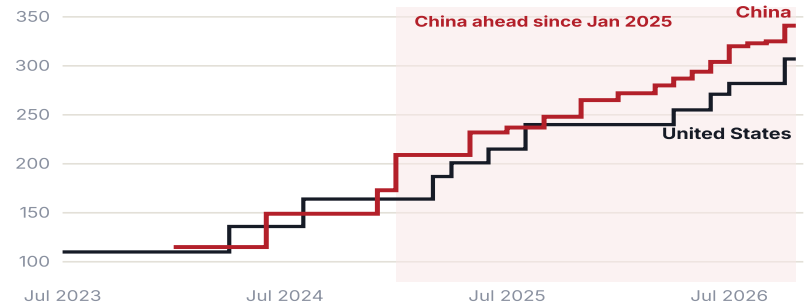
Intelligence is getting cheaper fast, and China is a big reason why

COST OF A FIXED ABILITY LEVEL · $ PER TASK, LOG SCALE



Cheapest model at each level of Artificial Analysis's Progress Index. Mid-tier: **$1.29 → $0.04 per task, June to late September (≈30×)**.

BEST OPEN-WEIGHT MODEL, BY COUNTRY · ABILITY INDEX



DeepSeek, Qwen, Kimi, GLM, MiniMax, MiMo: free to download, run and adapt anywhere, including in Europe.

~10×
cheaper per year for the same ability (a16z; Epoch AI: 9–900×)

3.5 mo
for the same ability to fit in a model half the size ("densing law")

The paradox. Several leading Chinese labs told me this year: *"Real-time interaction doesn't raise the intelligence ceiling. Focus on text and coding."* A reasonable bet, but the whole industry is racing on the one thing becoming abundant.

Intelligence is one part taking shape. What else does JARVIS need?

Artificial Analysis, artificialanalysis.ai/trends, accessed 10 Nov 2026: cost per Intelligence Index task by Progress Index threshold (levels ≥280, ≥320, ≥340); open-weights Progress Index by country · Appenzeller, a16z, Nov 2024 · Epoch AI, Mar 2025 · Xiao et al., Nature Machine Intelligence 2025

THE SCEPTIC'S CASE

Do people actually need a personal agent?

The test: if you had a human assistant, would you hand them this task, or just do it yourself?

WHERE AGENTS ADD LITTLE

FIVE-MINUTE TASKS

- Hail a ride
- Send a message
- Order food

< 5 min each, done by yourself

Explaining it to an assistant takes about as long.

Operating these apps for people saves almost nothing.

BROWSING IS THE POINT

Agent picks a top 10 → **✕ "None of these"** → **Browses Taobao herself** adjusts her expectations

✓ picks one from the agent's original top 10

Shopping and video fill free time; browsing is how people settle on what they want.

THE BASICS STILL BREAK

"Turn off the living-room downlights"	✓ works
"AC on 30 min before I get home"	✕ "Not supported"
TV, while casting a video: "Pause"	✕ "Not supported"

…yet the remote's pause button works fine.

Assistants should fix their own devices before operating other companies' apps.

WHERE THEY ADD VALUE: EXPANSION

Dai Yusen, ZhenFund: *fulfilling existing demand divides the pie; expansion grows it.*

Tasks people avoid. An hour on hold to dispute a bill: most people never try, so every success is new value.

Work beyond one's skills. A researcher directs dozens of agents; a non-programmer ships a working tool.

Learning by making. YouMind users build ~15 decks a month but present only 1–2: most are made to learn.

Mobile internet wasn't Baidu, Sina and Renren shrunk onto a phone; it grew new two-sided networks. **Agents need theirs.**

Dai Yusen (ZhenFund) and YouMind figures as reported by Frank Wang, Nov 2026 (paraphrased) · Author's user experiment and experience

EVIDENCE OF EXPANSION · 2025–26

Pine AI and my own research

CALLS NOBODY WANTS TO MAKE

phone menu "press 3…" → **on hold** up to an hour → **retention team** negotiate → **result** bill lowered

100K+ real calls since early 2025

93% success in bill negotiations

$3M+ saved for consumers

People simply don't make these calls: new demand, not old demand served faster.

ONE RESEARCHER, DOZENS OF AGENTS

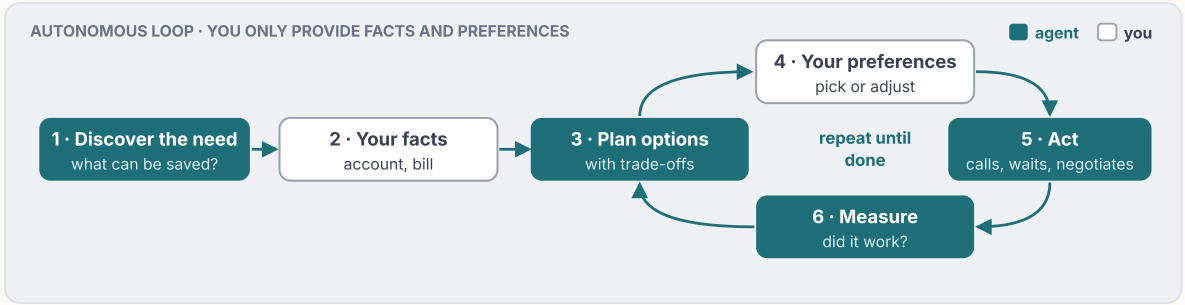
you talk set direction → **code, runs** experiments → **you judge** the results

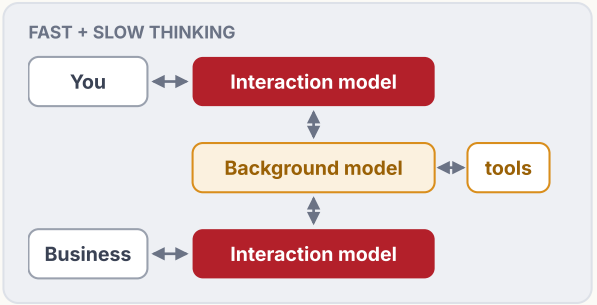
10+ papers drafted with agents

~1 wk per paper, idea to full draft

Dozens agents in parallel

One person now produces the output of a small lab. Taste stays human.





Launched in **December 2024**, Pine was the world's **first personal agent** to act for consumers in the real world.

Disclosure: I was Chief Scientist at Pine AI. Figures from Pine's published materials: Business Wire, 3 Dec 2025; Pine blog, "The most natural human-computer interface is your voice", Jun 2026.

PART 1 · TECHNICAL MOATS

Five technical moats between today's agents and JARVIS

Hard to build, so each gives a lead; each also helps win a non-technical moat. Two still need new ideas.

TECHNICAL MOAT	STATUS	WHAT IS BROKEN TODAY	WHAT IT UNLOCKS	HELPS WIN
1 **Interaction models** LLM	◐ TAKING SHAPE	Most agents still take turns; full-duplex models only began shipping in 2026.	Talk and act at once: interrupt, bargain, direct many agents.	DISTRIBUTION
2 **Agent-native OS** HARNESS	✕ MISSING	Computers assume one human at the keyboard; agents fight over screen, disk and network.	People and agents on one screen; software that evolves; dozens of agents per machine.	DISTRIBUTION
3 **Fast inference** INFRA	◐ TAKING SHAPE	Agent tasks take 25–60 s; GPUs trade per-user speed for cheap tokens.	Answers in about a second; then the harness sets the speed.	An edge for all three
4 **Physical-world reach** HARNESS	◐ TAKING SHAPE	Agents lack identity and permission; platforms close the gates.	Calls, payments, other agents, hired people, labs.	PERMISSION CONTEXT
5 **Continual learning** HARNESS + LLM	✕ MISSING	Agents start over every session and don't improve with use.	Skills, memory and weights that grow with use: better in month six.	CONTEXT

Most mature: phone calls, handled at scale since 2025; payments are close behind.

Clearest path: fast inference. The physics is understood; it needs specialised chips, then a faster harness.

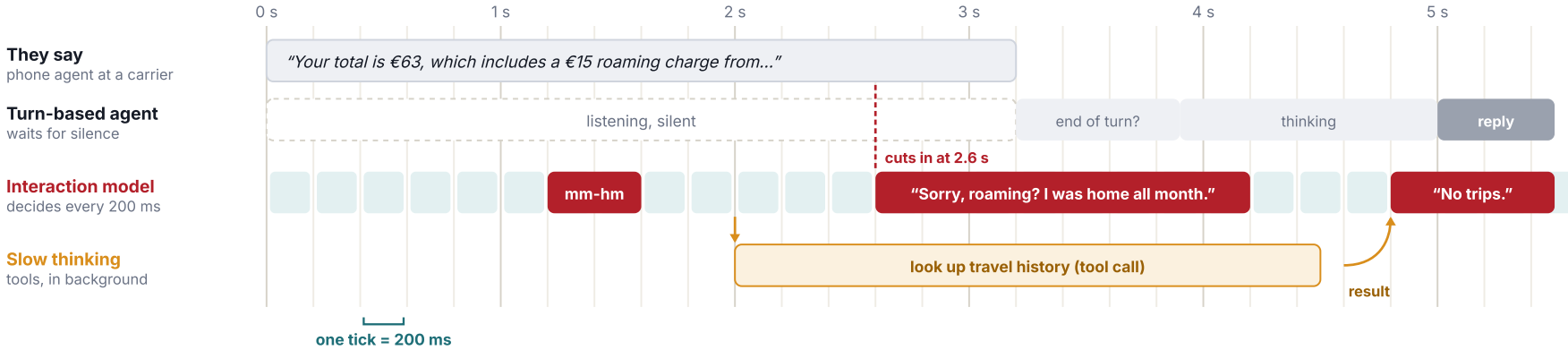
Still missing: an agent-native OS, and continual learning without forgetting. Both need new ideas.

TECHNICAL MOAT 1 · INTERACTION MODELS

The hard part is interaction, not intelligence

HOW A FULL-DUPLEX INTERACTION MODEL WORKS · A BILL NEGOTIATION, SECOND BY SECOND

PINE: #1 ON T³-VOICE, AUG 2026



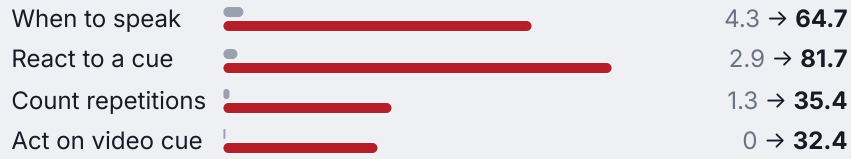
Every 200 ms the model hears the latest audio and chooses: stay silent, say "mm-hm", speak, cut in, or start a tool call. Listening and speaking overlap, as between people.

Slow thinking runs tools in the background and hands results back. The turn-based agent first reacts at **5.0 s**; by then the interaction model has objected and checked the facts.

WHAT BROKE IN PRODUCTION

- Phone menus loop faster than the agent can reply
- In a noisy restaurant, the agent never gets a turn to speak
- In bargaining, it can't talk over or interrupt the other side
- Shown the transcript afterwards, **the same model knows what to do**

WHAT FULL-DUPLEX FIXES: TURN-BASED → FULL-DUPLEX



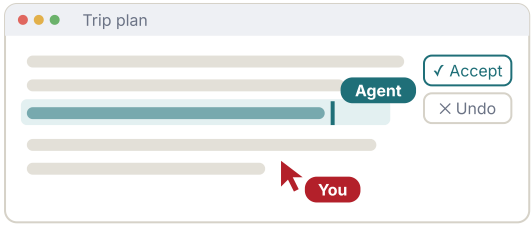
Illustrative timeline based on Pine's production design · Thinking Machines Lab, "Interaction Models", May 2026 (200 ms ticks; timing scores 0–100: GPT-realtime-2.0 minimal vs. TML-Interaction-Small) · OpenAI GPT-Live, Jul 2026 · τ³-Voice leaderboard, Aug 2026

TECHNICAL MOAT 2 · AGENT-NATIVE OS · PART 1: WORKING TOGETHER

People and agents should share one computer, not take turns

Today an agent either takes over your screen or works out of sight. JARVIS works beside you: “a workbench, not a service window” (Ryo Lu, former head of design, Cursor).

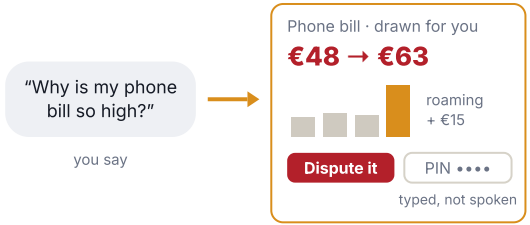
A SHARED USER INTERFACE



You and your agents act on the same screen at once, as in a shared document: you see each step and can take over.

OPEN Who holds control at each moment, and how it is handed back.

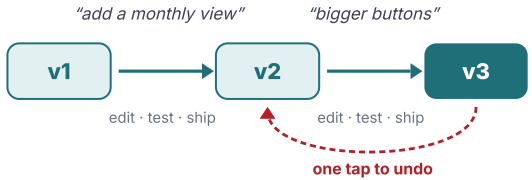
B GENERATIVE USER INTERFACE



The user interface is drawn for each task: talk to decide, tap to confirm, type secrets privately.

OPEN Screens must stay put, since people expect a button where it was yesterday, and redraw in about a second.

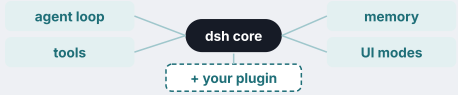
C SELF-EVOLVING SOFTWARE



Software the agent maintains for one person: ask for a change and it edits, tests and ships it.

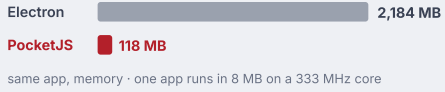
OPEN Every change tested and reversible, or the software decays with use.

BUILT BY THE COMMUNITY — DeepSeek Harness (dsh)



“Everything is a plugin”, even the agent loop, so users and their agents can reshape it.

REBUILT FROM SCRATCH — PocketJS · Omarchy Pocket



A small team with coding agents rebuilt the Linux GUI in TypeScript: real Linux on a $99 handheld.

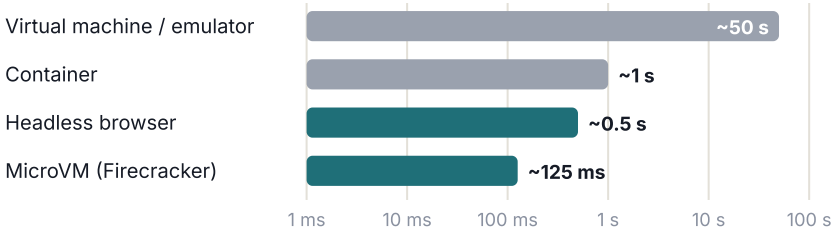
Ryo Lu, AIE Shanghai, Nov 2026 · Ink & Switch, “Malleable software”, 2025 · Pine AI blog, Jun 2026 · github.com/deepseek-ai/deepseek-harness · github.com/pocket-nexus/pocketjs; Omarchy Pocket announcement, 2026

Dozens of agents on one machine need workspaces and a scheduler

On my own machines, agents broke the computer long before they ran out of intelligence.

D LIGHTWEIGHT WORKSPACES (SANDBOXES)

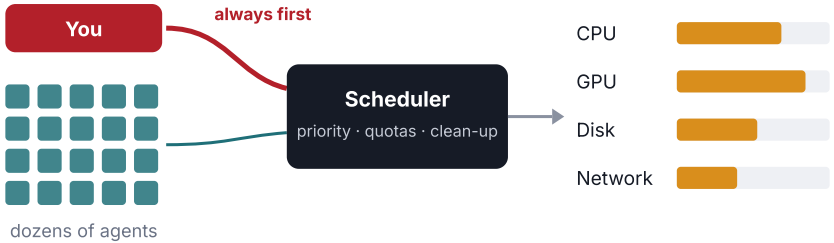
Time to start a fresh workspace · log scale · typical orders of magnitude



BROKE Agents' full copies of code filled a **3 TB SSD in one day**; none cleaned up after itself.

NEEDED Workspaces that start in well under a second, share what hasn't changed, and clean up after themselves.

E RESOURCE MANAGEMENT FOR AGENTS



BROKE Parallel jobs loaded a 100+ core server so heavily that **I couldn't log in**. Uploading a screenshot took longer than the model call.

NEEDED The person always comes first; quotas for CPU, GPU, disk and network; send changes, not screenshots.

Workspaces don't follow model prices. Models get ~10× cheaper a year; virtual machines, browsers and phone emulators cost what computers cost.

Clicking vs. calling. Screen-clicking agents work with any app but slowly; direct links to services are fast but built one at a time.

In the end, I had my agents build their own scheduler for queues, disk and clean-up: **work the operating system should do.**

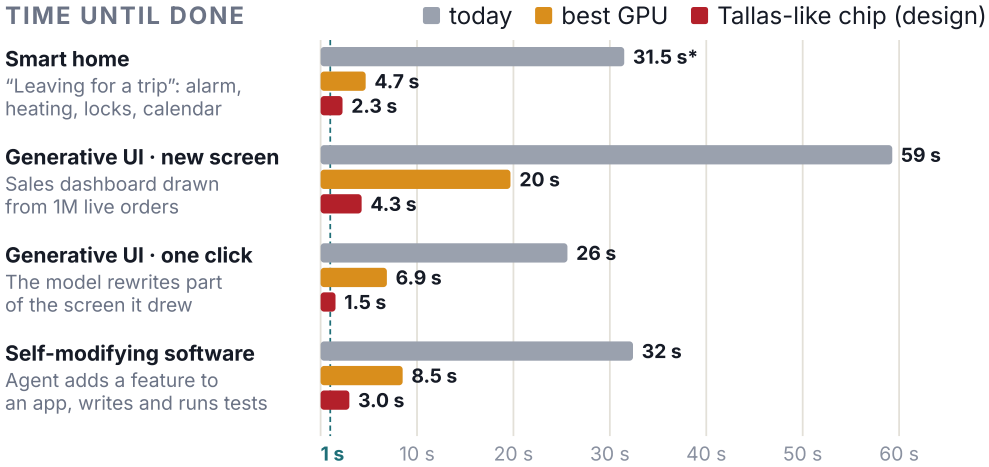
Author's experience · Agache et al., "Firecracker: lightweight virtualization for serverless applications", NSDI 2020 (≈125 ms start, <5 MiB overhead) · Other start times are typical orders of magnitude

TECHNICAL MOAT 3 · FAST INFERENCE

Speed for each user decides what an agent can do

Recorded agent sessions, replayed at three speeds. Tool and device time is identical; only the model's writing speed changes.

TIME UNTIL DONE



A voice command should finish in about a second; today these take half a minute or more. Each task generates 2,600–19,000 tokens, mostly reasoning.

TOKENS PER SECOND FOR ONE USER · 1 TOKEN ≈ ¾ WORD

~5	89	874	3,400	4,238	17,000
person reading	typical API	best GPU	Groq (31B)	Tallas-like (DeepSeek V4.1 Flash, 1M context)	Taalas (8B)

WHY SPEED CHANGES THE PRODUCT

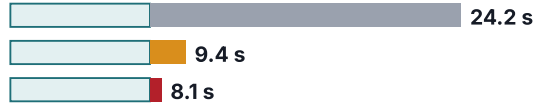
- **Voice assistants** still run scripted commands, partly because reasoning first takes seconds.
- **Generated screens** work only if each click answers in about a second.
- **Thinking per reply = speed × time:** only a fast model can reason mid-conversation.

WHY IT IS HARD

On GPUs, batching many users is what makes tokens cheap. **Faster decode for one user costs more per token**, so most serving APIs trade speed for price.

THEN THE HARNESS IS THE BOTTLENECK

Fixing a failing test · **7.5 s running tests** + model



With a fast model, **the harness sets the speed**.

DeepSeek-V4.1 Flash sessions, Oct 2026 · Li, *AI Infra in Depth*, Ch. 1 · Today: measured DeepSeek API (*median OpenRouter speed) · Best GPU: SGLang, 4×GB300, one user · Tallas-like chip (github.com/bojieli/OpenTallas): open weights-in-ROM chip design, analytical estimate, not yet built · The Register, Aug 2026 · NYU Shanghai RITS, Feb 2026 · Brysbaert 2019

TECHNICAL MOAT 4 · PHYSICAL-WORLD REACH

Reaching beyond the screen, and who grants permission

agent
on a screen

People & institutions	Money	Other agents	Hiring humans	Lab instruments	Fabrication
calls, waits on hold, negotiates, emails	pays within limits: AP2, ACP, x402	hires and pays agents; agent social networks	books people for errands (RentAHuman, 2026)	A-Lab: 17 days alone, 36 of 57 materials	3D printers and laser cutters
●●●●○ mature	●●●○○ early	●●○○○ early	●●○○○ early	●●○○○ early	●○○○○ starting

FARTHER FROM THE SCREEN → harder to verify, harder to undo

dots: my rating

THE HARD PART IS PERMISSION: EACH QUESTION GETS A BUILDING BLOCK

Who is this agent?	Agent identity	Manus Cue: own phone, email, QR
What may it commit you to?	Signed mandates	pre-approved limits
How does it pay safely?	Single-use credentials	one payment, one merchant
When must you say yes?	Consent gates	approve before paying
Does the other side know?	AI disclosure	EU AI Act; China's labels
How do you take it back?	Revocation	one switch withdraws all
Who is liable for a mistake?	Still open	no common rule yet

CHINA CASE — Platforms control the gates

INSIDE ONE GARDEN

Qwen agent → Alibaba apps: Taobao · Alipay · Fliggy · Amap

✓ books, pays, plans trips

ACROSS GARDENS

Dec 2025 — Doubao phone → major apps block it within days

Oct 2026 — Open protocol → 155 apps opt out incl. WeChat, Alipay, Meituan

Google AP2, Sep 2025 · OpenAI & Stripe ACP, Sep 2025 · Coinbase x402, May 2025 · Szymanski et al., Nature 2023, author correction Jan 2026 · Alibaba Qwen app, Jan 2026 · Caixin Global, 19 Dec 2025 and 15 Oct 2026 · RentAHuman, Feb 2026 · Moltbook, Jan 2026 · Google A2A · Manus Cue, Sep 2026

TECHNICAL MOAT 5 · CONTINUAL LEARNING

An assistant should get better with use, like a person

Pine's research: four places to keep what the agent learns, from outside the model to inside its weights.

A SKILLS

PREACT

FIRST TIME

see → reason → act

3–5 s of reasoning, every step

compile

LATER

state 1 → state 2 → state 3

replayed, no model call

screen differs? fall back to reasoning

Like a visitor paying with Alipay: clumsy the first time, fluent by the third. Each path is learned once, then reused.

B MEMORY

USER AS CODE

"How many flights to Japan in 2025?"

memory.md

as text: 100 records, counted token by token

✕ error-prone

```
user.flights
.where(to='Japan', year=2025)
.count()                    # exact
```

✓ as code the model wrote itself

The model writes its own structure for each person's memory, so counts, conflicts and reminders are exact.

C CONTEXT

PROGRAMMABLE KV

WHAT THE MODEL READS (KV CACHE)

skill | profile | conversation | fix

precompiled, spliced in — appended

[correction] balance €48 → €63

98.5% vs 1%

cache hits, vs editing facts in place

up to 53× faster first token

Changed facts are appended, not edited, so the cached context stays valid as the world changes.

D WEIGHTS

USER AS ENGRAM

model

shared skill: **read memory**

trained once

look up

GROWING TABLE

you

your family

your company

+ new slot

✓ add a slot, overwrite nothing

✕ LoRA rewrites weights globally

Each person's facts live in their own slots beside the model, so one user never overwrites another.

Still open: learning judgment, not only facts, inside the weights without forgetting. There is no agreed recipe yet.

Live data is rented; what the agent learns from it is kept. Continual learning is the only moat that **compounds with use**.

Disclosure: my own research with Pine AI, code at github.com/19PINE-AI · arXiv: PreAct 2606.17929; User as Code 2606.16707; Programmable KV 2606.17107 (online-service test); User as Engram 2606.19172 · B. Li, "AI Agent 的两朵云", Flink Forward Asia, Jun 2026

PART 2 · NON-TECHNICAL MOATS

If intelligence is a rising tide, which moats last?

Technical leads erode as models, prices and features improve for everyone. Three non-technical moats stay above the water.

Distribution

Who reaches the user first

Devices, pre-installs, super-apps, the default assistant

built by ① interaction, ② OS

Context

What the agent knows about you

Live data (mail, chats, payments, calls) is rented. Memory is owned, compounds.

built by ④ data feeds, ⑤ learning

Permission

What the agent may do for you

Pay, book, sign, call. Granted by users, platforms and law, and revocable.

built by ④ physical-world reach

THE TIDE LIFTS EVERYONE

Open weights trail the frontier by months · tokens ~10× cheaper a year · every agent now has memory, connected apps and scheduled tasks

Distribution gets you in the door. Permission decides what you may do. Context is what you keep.

WHAT THE TIDE DOES NOT LIFT

Agent workspaces
VMs cost what computers cost.

Single-user speed
Capped by memory bandwidth.

Human attention: one thread
DeepSeek's **dsh super-agent**: one interactive router talks to you; background workers run the tasks and speak only when you must act.

Trust, getting worse as agents act alone

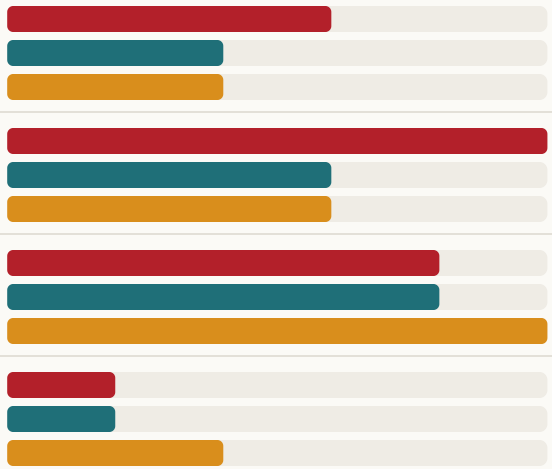
- One email hijacked a **Manus** agent and ran code.
- **Meta Muse** gave a seller's home address to a Marketplace buyer and set up a pickup without asking him.

Salt Labs, Manus prompt-injection disclosure, Oct 2026 (via Dark Reading); fixed · Fortune, 19 Oct 2026: Meta acknowledged the Muse error · DeepSeek dsh-superagent (router–worker), Oct 2026

PART 2 · WHO HOLDS THE NON-TECHNICAL MOATS · GLOBAL AND CHINA

No one holds all three non-technical moats securely

PLAYER · GLOBAL / CHINA	Distribution	Context	Permission	MAIN RISK
Frontier labs OpenAI, Anthropic, Google DeepMind DeepSeek, Moonshot, Zhipu, MiniMax	■	■	■	See only what users paste in; intelligence is becoming abundant
OS & device makers Apple, Google, Microsoft Huawei, Xiaomi, Honor, OPPO, vivo	■	■	■	Apps can withhold data and actions (the Doubao phone case)
Super-apps & platforms Meta, Amazon Tencent, Alibaba, ByteDance, Meituan	■	■	■	Context and permission stop at their own walls
Startups Pine, YouMind, Manus, general-purpose agent startups	■	■	■	Squeezed when labs and platforms ship the same feature



LIU YONG, TAOZ CAPITAL

A “200-IQ” agent in your pocket: one entry point, a secretary agent behind it.

DAI YUSEN

Labs subsidise their own apps ~20×; yet Windows still grew a huge third-party ecosystem.

WANG HUAI

In the PC era, platforms owned the assistant. Agents are not a startup opportunity.

MY VIEW

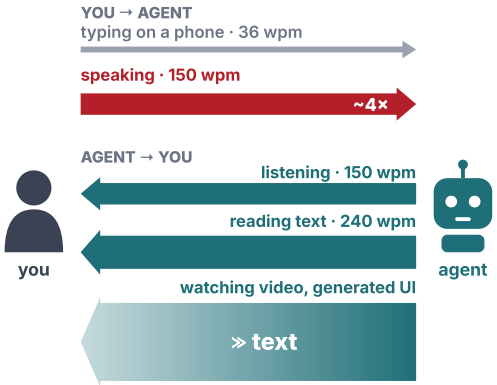
Startups win on speed. They propose ideas and ship fast prototypes; big vendors follow. Manus pioneered general agents.

Ratings are the author’s qualitative assessment · Shanghai remarks as reported by Frank Wang, 7 Nov 2026 (paraphrased)

WHAT IT SHOULD FEEL LIKE

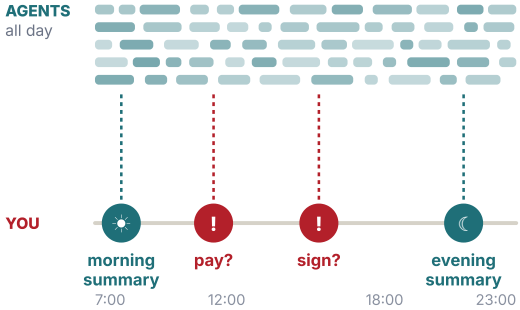
Protect the human's single thread of attention

A NARROW LINK



Voice for the gist, screen for the detail.

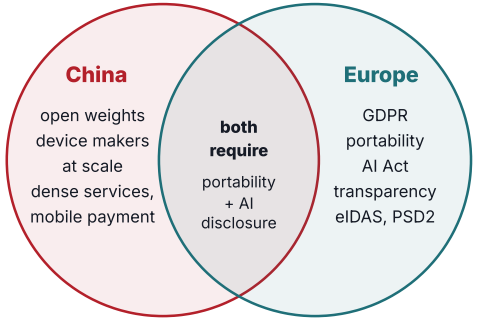
MANY AGENTS, ONE HUMAN



Dozens of agents, four interruptions a day:
only money, commitments, irreversible steps.

Delegate the work, not your attention.

EUROPE AND CHINA



A personal agent needs both: **the ability to act** and **the trust to be allowed to**.

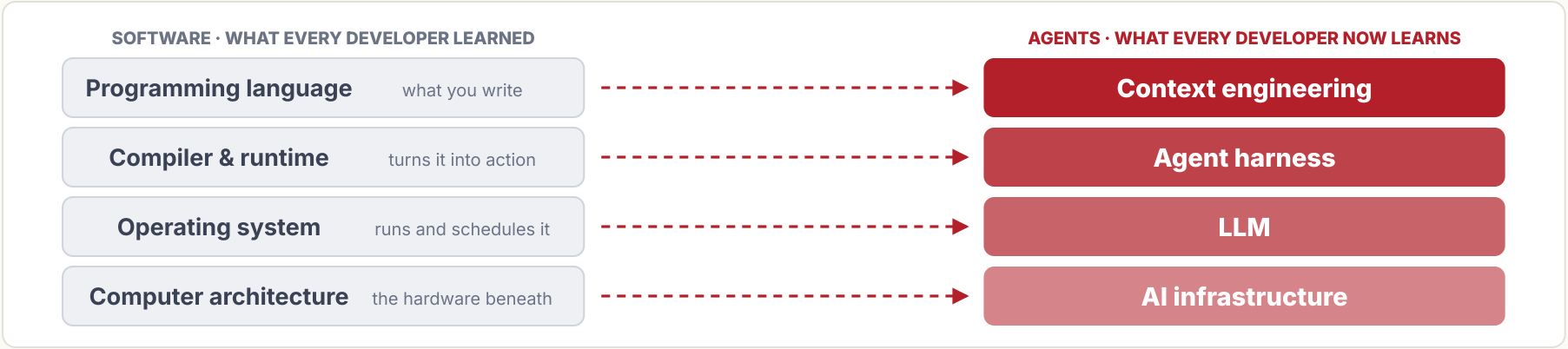
"You can outsource your thinking, but you can't outsource your understanding." Andrej Karpathy, 2026

Andrej Karpathy, 2026 (post on X, widely quoted) · GDPR Art. 20 · EU AI Act Art. 50 · PIPL (2021) Art. 45 · CAC AI-content labelling rules (2025) · Palin et al. 2019 (phone typing ~36 wpm) · Ruan et al. 2016 (speech vs. typing) · speech ~150 wpm · Brysbaert 2019 (reading ~240 wpm)

FOR DEVELOPERS AND UNIVERSITIES

Agent development is becoming a basic skill, like C or Python

Compilers, operating systems and architecture taught a generation to build software. Agents, LLMs and AI infrastructure are next.



AI Agent

CONTEXT + HARNESS

AI Agent in Depth

by Bojie Li · open source

Context, memory, tools, evaluation, learning

github.com/bojieli/ai-agent-book

LLM

LLM

Hands-On Large Language Models

Alammar & Grootendorst · my Chinese translation 《图解大模型》

How LLMs work, visually

AI Infra

AI INFRASTRUCTURE

AI Infra in Depth

by Bojie Li · open source

Accelerators, inference, training, scheduling

github.com/bojieli/ai-infra-book

Agent, LLM and infrastructure: three core courses, three books

TAKEAWAYS

Personal agents: imagined in 1979, now a five-year problem. Five technical moats to build; three non-technical moats to hold.

NON-TECHNICAL MOATS applications

- **Distribution** — gets you in the door
- **Permission** — decides what you may do
- **Context** — is what you keep

TECHNICAL harness

- **2 Agent-native OS** — ✕ missing
  one screen for people and agents
- **4 Physical-world reach** — ◐ taking shape
  limited by permission, not ability
- **5 Continual learning** — ✕ missing
  rented data becomes owned context

TECHNICAL LLM

- **1 Interaction models** — ◐ taking shape
  timing is a skill: fast thinking up front, slow thinking behind

TECHNICAL infra

- **3 Fast inference** — ◐ taking shape
  from a minute's wait to about a second: specialised chips

Agents that only **fulfil** divide the pie. Agents that **expand** what one person can do grow it. That is the road to JARVIS.

Thank you

Dr. Bojie Li · 01.me