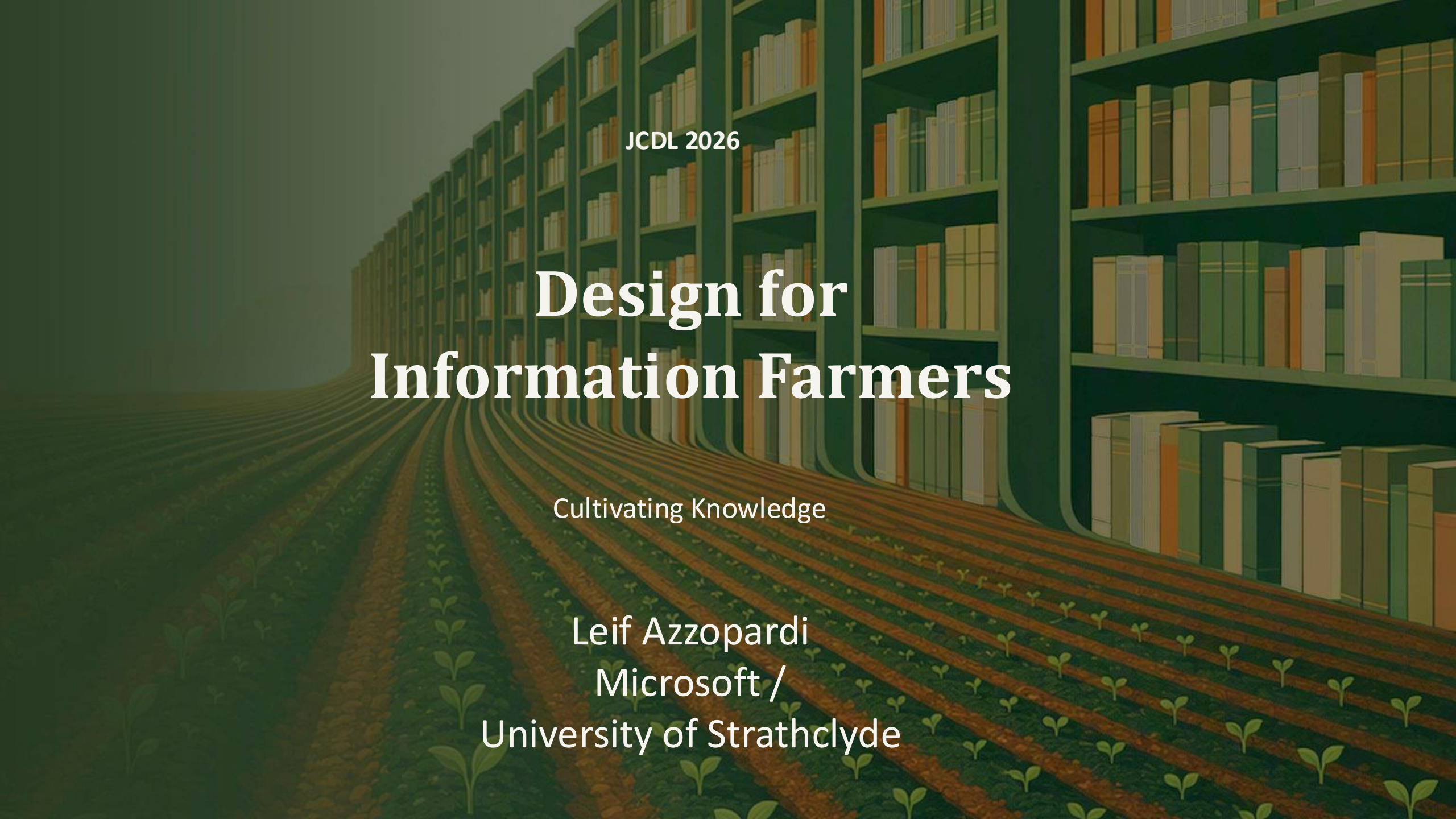




JCDL 2026

Design for Information Farmers

Cultivating Knowledge

Leif Azzopardi
Microsoft /
University of Strathclyde

Two keynotes ask what libraries become; this one asks what users become

Chengxiang Zhai

How do repositories become intelligent knowledge assistants?

Nicholas Belkin

Can digital libraries continue their historical trajectory
under ubiquitous generative AI?

This talk

What do digital libraries become when users are
no longer searchers, but cultivators?

**The distinctive move: the library itself becomes the seed, soil and steward -
not only the field that is searched, but ploughed!**



Foraging discovers the field. Farming changes the state of the field.

The digital library's future value lies not only in what it contains,
but in what can be grown from it.

Farming inherits six decades of theory - and moves the locus of interaction



Berrypicking

Bates, 1989 -
an evolving, bit-by-bit need



Information Foraging

Pirolli & Card -
gain against time and cost



Search as a process

Kuhlthau -
uncertainty, affect, meaning



Sensemaking

Russell et al., 1993 -
useful representations



Exploratory search

Marchionini, 2006 -
finding to learning

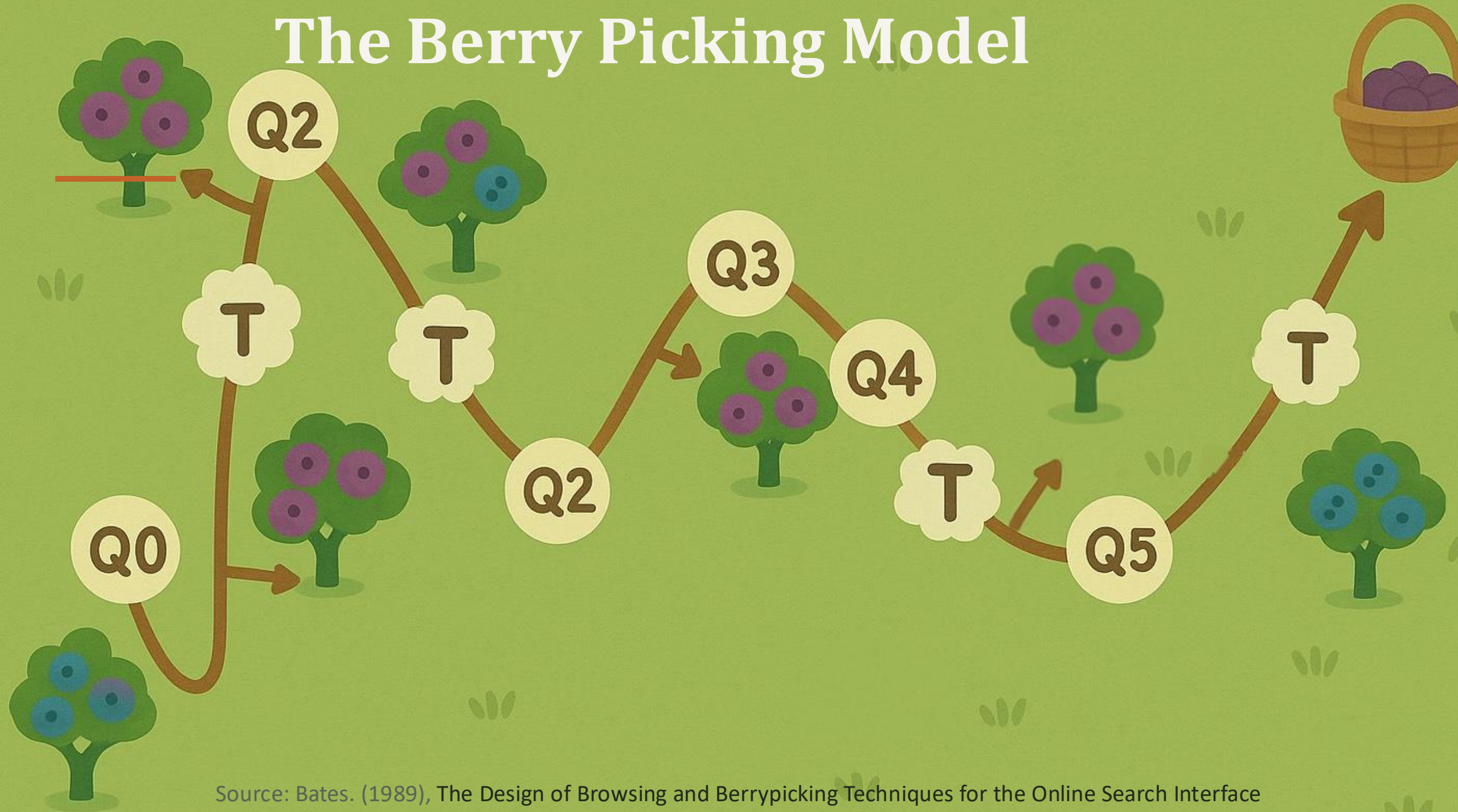


Knowledge Building

Scardamalia & Bereiter -
improvable ideas

Farming moves the locus: users seed a responsive medium, then grow it.

The Berry Picking Model



The Information Landscape Is Changing



A Historical Analogy: The Neolithic Revolution

PHYSICAL WORLD

Hunter-Gatherer

- Move between patches
- Collect what's available
- Adapt to the landscape

Agrarian Farmer

- Tend a fixed plot
- Cultivate and shape the yield
- Compound gains over seasons

INFORMATION WORLD

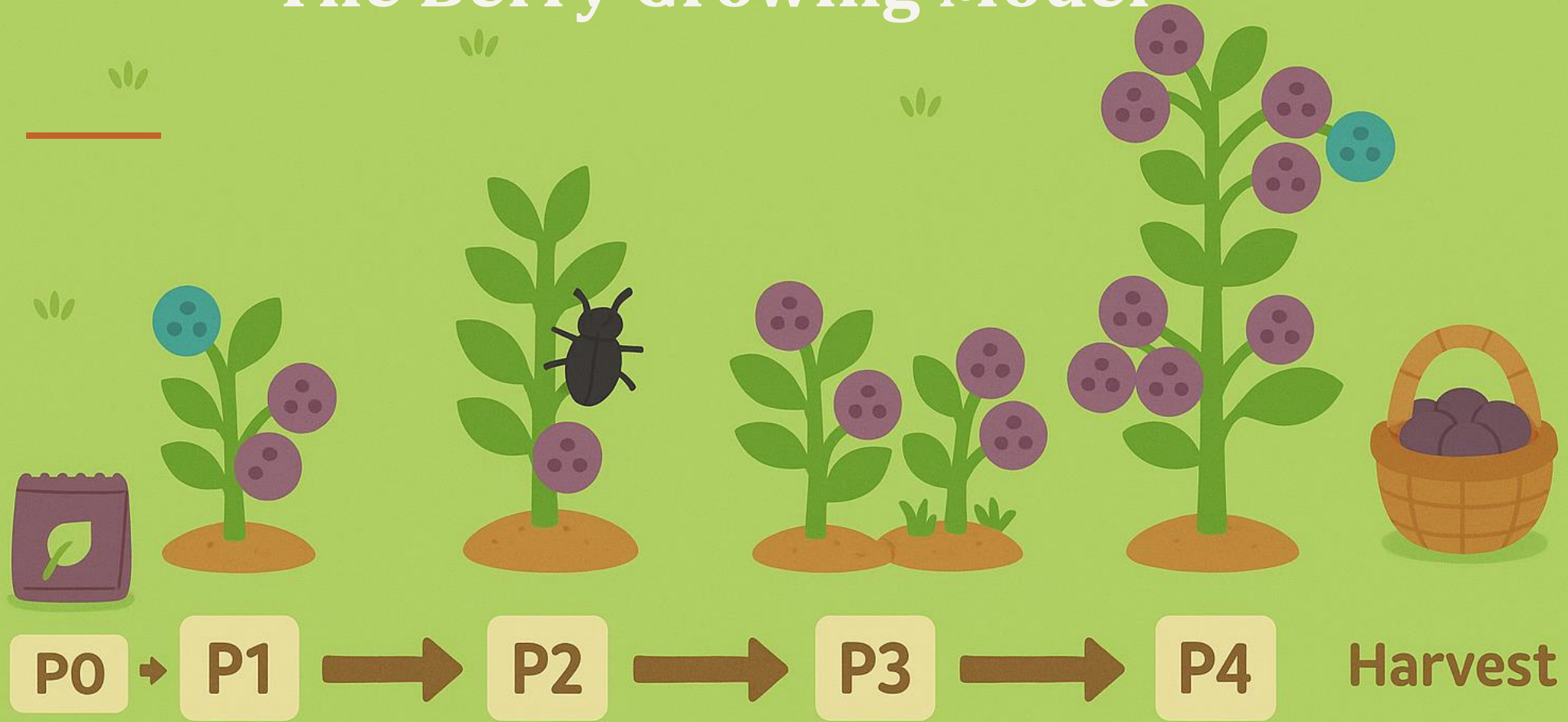
Berry Picker / Info Seeker

- Navigate between sources
- Collect relevant snippets
- Adapt queries to results

Berry Grower / Info Producer

- Cultivate in a single plot
- Shape and refine outputs
- Compound via seed-making

The Berry Growing Model



Farming gives us an operational vocabulary - and a lifecycle to evaluate

Seeding

Growing

Pruning

Weeding

Fertilising

Breeding

Cross-pollinating

Harvesting

Packaging

Seed-making

Composting

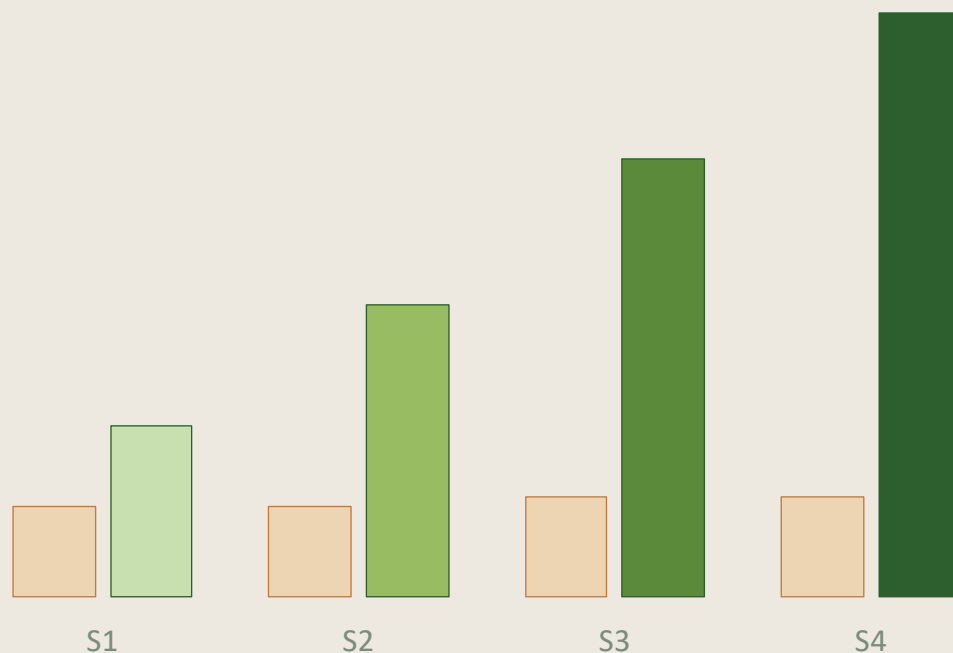
Evaluation must cover the whole lifecycle, not a single query-response pair.

User Interaction	Information Farming
Seeding	“Make a list of laws relevant to library information science”
Growing	“For each law, provide a detailed description, including citations”
Pruning	“Shorten the description for each law, so that we have one sentence ...”
Weeding	“The math for Bradford’s Law is not correct, it should be. ...”
Fertilizing	“Here is the paper on the Least Effort in Finding (LEIFs) Law
Breeding	“Phrase the law on fairest in ten different ways, then compare pairs...”
Cross Pollinating	“Take Newton’s laws of motion and re-interpret each in the LIS context”
Harvesting	“Take all the laws, and preset them in order of appearance...”
Packaging	“Create a table that contains the list of laws in LIS, include citations and bibtex”
Seed Making	“Review the conversation so far, and create a prompt to reproduce the content”
Composting	“Take my notes and flesh them out...”

Information Farming Affords Many Benefits & Opportunities

Yield per session

Foraging
Farming



Higher Yield

Richer, denser, more tailored outputs.
Compound gains from re-planting.

Precision Control

Users specify quality, format, depth — far beyond keyword search.

Create New Knowledge

Cross-pollination produces information that didn't previously exist.

Scale via Agents

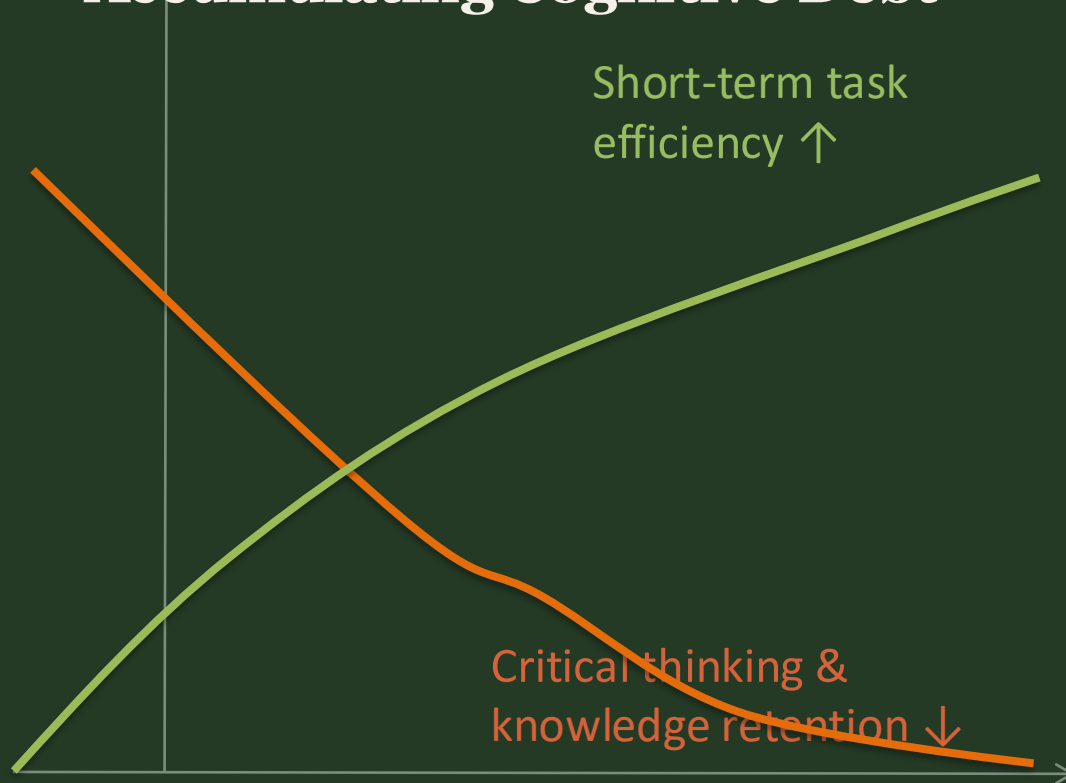
Persistent, unattended growth. Farmers set things growing and return for harvest.

How does Foraging differ from Farming

	Information Foraging	Information Farming
Interaction Style	Opportunistic & exploratory	Deliberate & iterative
Control over Output	Accept what's found	Shaped, pruned, bred for quality
Effort	High-cost navigation + post-processing	Prompt design & cultivation
Yield	Variable by patch density	Higher, more predictable
Innovation	Recombine found information	Create genuinely new information
Scalability	Constrained by user attention	Scalable via agentic systems

Information Farming also comes with Risks & Costs

Accumulating Cognitive Debt



Cognitive Debt

Over-reliance erodes critical thinking and independent inquiry.

Confirmation Bias

Iterative refinement amplifies existing beliefs at scale.

Information Health

Hallucination and misinformation spread through re-planted prompts.

Feudal Dynamics

Platform owners extract disproportionate value from farmers' work.

Farming does not eliminate foraging - it changes what foraging is for

Retrieval supplies the material: evidence, counter-evidence and serendipity.

Farming converts that material into a maintained epistemic product.

Serendipity, citation chasing and direct reading remain load-bearing.
A library that optimises only for synthesis will starve the very process that feeds it.

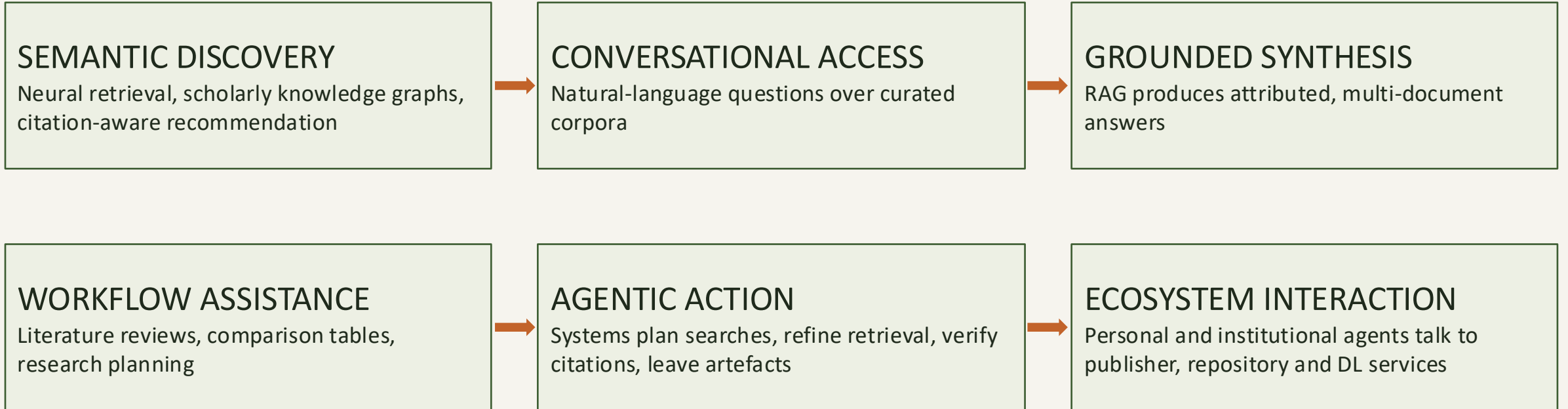
Design for both: a persistent plot, and an open path back into the field.

Farming is a co-evolutionary loop, and it inverts three familiar defaults

Query, system and output develop together: a generative, co-evolutionary loop.

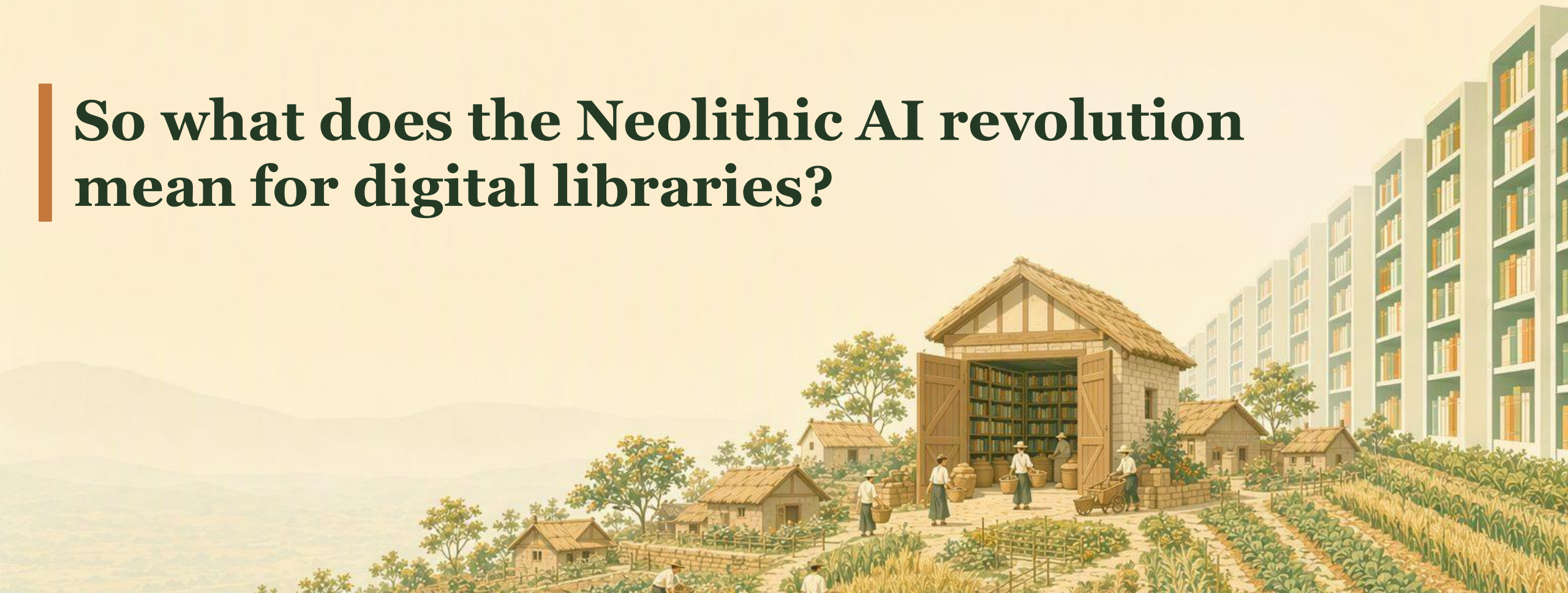


2021 to 2026: from documents, to answers, to work

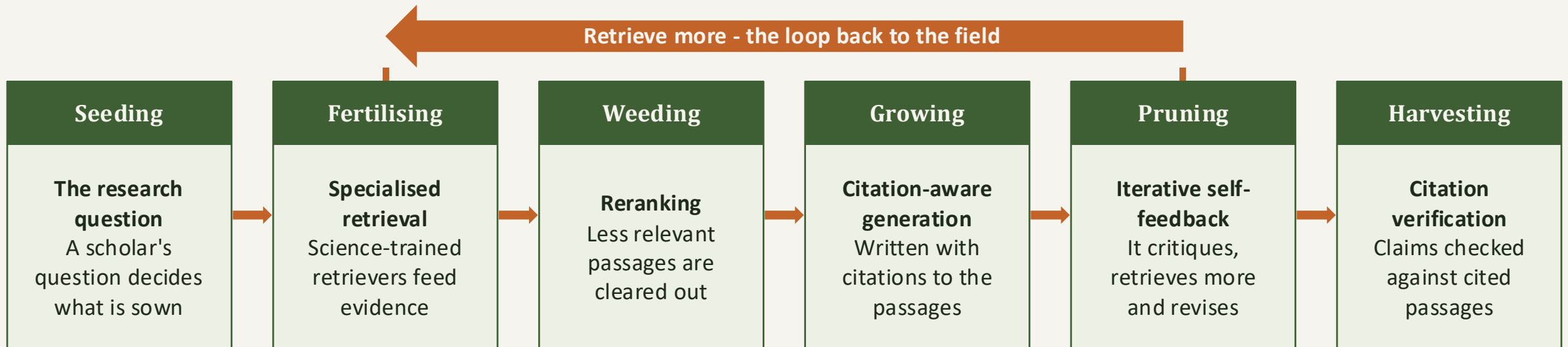


Each step moves the deliverable further from the document and closer to the finished task.

So what does the Neolithic AI revolution mean for digital libraries?



OpenScholar as a farming agent: the pipeline is the farming loop



Seed stock - data store: 45 million open-access papers

Same flow as the farm: retrieve, grow, prune, check -
and go back to the field when the crop is thin.

System: Asai et al. (2026), Nature. Mapping to farming actions: this talk's reading.

OpenScholar is the strongest research demonstration we have

45M

open-access papers in the
corpus

+6.1%

OpenScholar-8B correctness
over GPT-4o

+5.5%

OpenScholar-8B correctness
over PaperQA2

70%

OpenScholar-GPT-4o preferred
over expert-written answers

**Same model, different system:
the gain came from the corpus and the loop, not from scale.**

OpenScholar-8B alone was preferred in 51% of comparisons - roughly on par with experts.
Unretrieved GPT-4o fabricated most of the paper titles it cited.

Source: Asai et al. (2026), Nature; benchmark: ScholarQABench.

But this is not autonomous scholarship - and the authors say so

Imperfect retrieval

The most relevant work is not always found.

Unsupported claims

Some statements still lack support.

Benchmark scope

One slice of scholarly work.

Licensed literature

Rights-protected work unresolved.

The system is an instrument, not an author. But note where the gain came from.

Source: Asai et al. (2026), Nature - all four limitations are stated in the paper's own Limitations section.

**Curated corpora, specialised retrieval, iteration and verification
matter more than model scale alone.**

Commercial and open infrastructure are converging on the same shape

Scopus with AI

Cited summaries, conversational history, foundational documents, concept maps, emerging themes. Deep Research plans, searches, revises and reports.

Web of Science Research Assistant

A research question becomes a connected workflow: topic exploration, literature review, author influence, trend interpretation.

Wiley AI Gateway / MCP

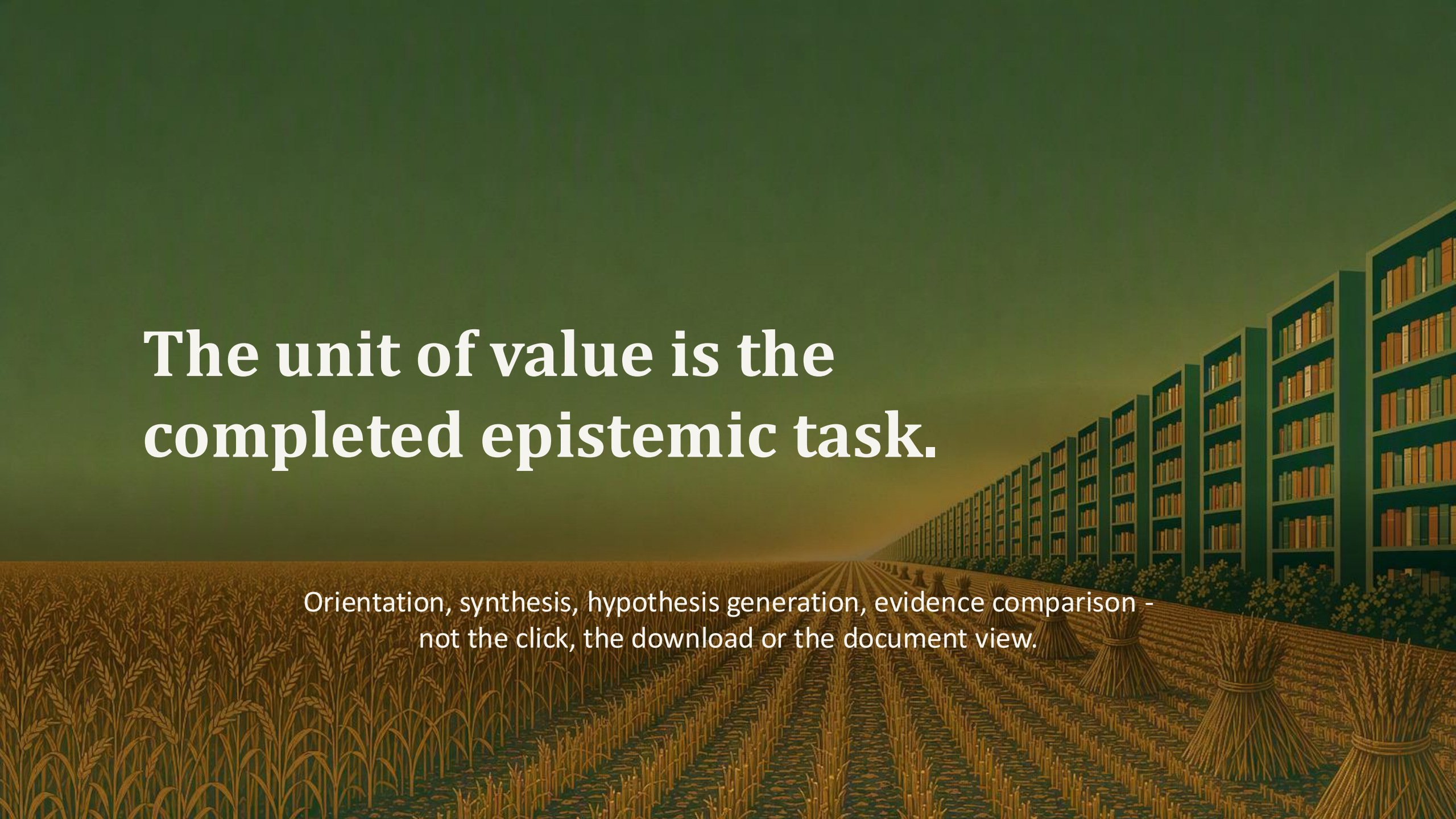
Licensed peer-reviewed content placed inside external AI environments, rather than behind a publisher's own interface.

OpenAlex

'Soil for the scholarly ecosystem': CC0 data, API, snapshots, full text, CLI, explicit agent support - a non-profit counter-model.

The unit of value is the completed epistemic task.

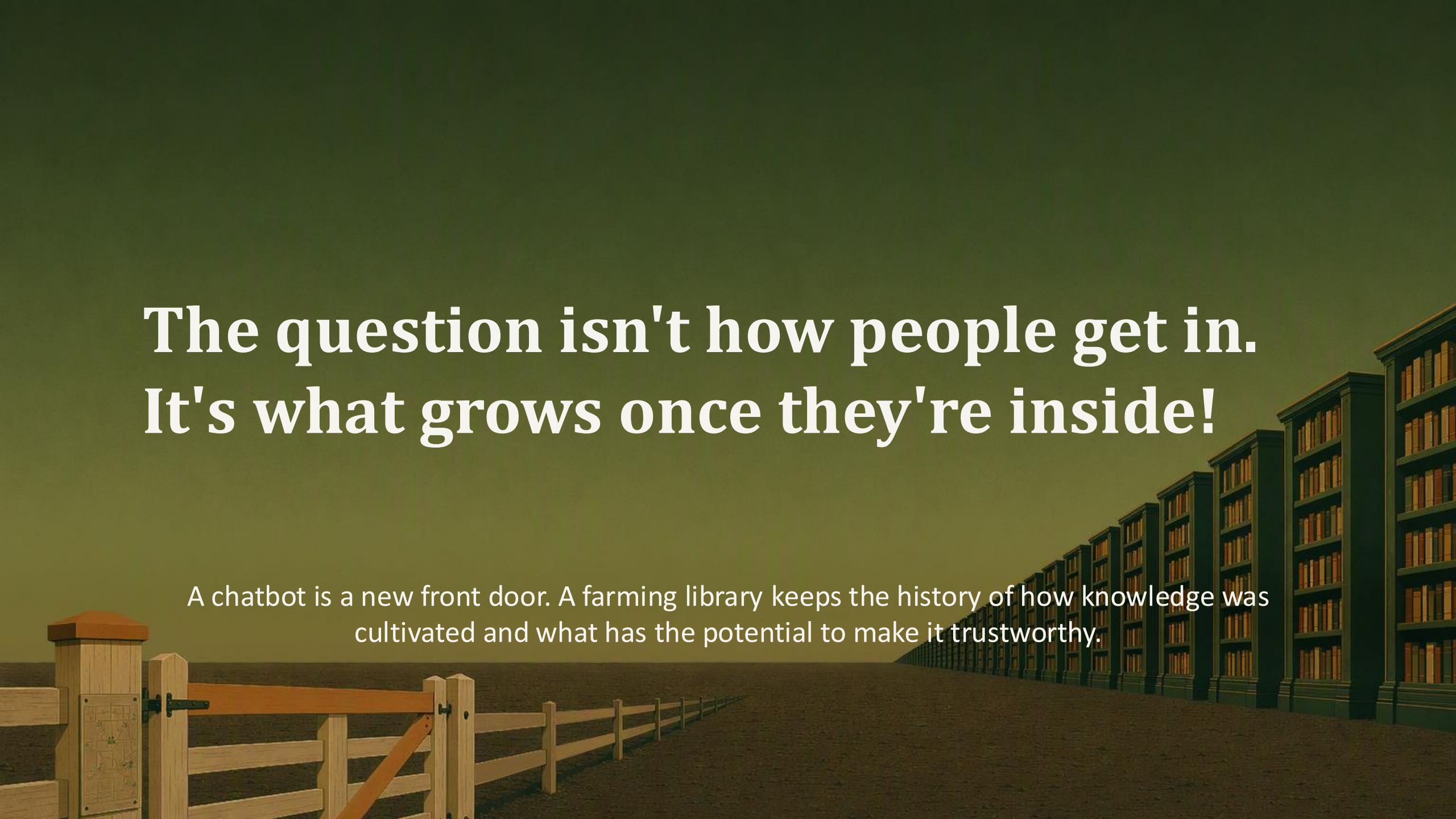
Orientation, synthesis, hypothesis generation, evidence comparison -
not the click, the download or the document view.



Six layers a farming digital library has to operate

Evaluation	Coverage, entailment, provenance, correction cost, reuse, learning
Stewardship	Preservation, correction propagation, audit, versioning, human review
Gateways	APIs, MCP endpoints, identity, entitlement, quotas, delegation
Memory	Saved questions, representations, corrections, decisions, seeds
Cultivation	Retrieval, extraction, comparison, synthesis, contradiction detection
Seeds and soil	Seeds: full text, data. Soil: metadata, citations, identifiers, licences

Read it from the soil upwards: nothing above the soil layer is trustworthy without it.



The question isn't how people get in. It's what grows once they're inside!

A chatbot is a new front door. A farming library keeps the history of how knowledge was cultivated and what has the potential to make it trustworthy.

Knowledge Forum is the historical bridge - this did not begin with GenAI

KNOWLEDGE FORUM, LONG BEFORE LLMS

Community knowledge as a persistent space of editable notes, links, annotations, scaffolds and higher-order 'rise-above' syntheses - aimed at improving public ideas, not individual learning.



WHAT GENERATIVE AI CHANGED

The cultivation operations became cheap, generative and potentially agent-scalable - so what a small, disciplined community could sustain, a whole field might.

The open question is no longer whether communities can cultivate knowledge together.
It is who operates the plot, at what scale, and under whose terms.

Information Farming is not the claim that GenAI invented knowledge creation.

Right now machine patrons look like an attack - and we respond as if they are



66

repositories responding to the COAR global
survey

90%+

reported aggressive bot traffic

80%

reported service disruption

**Mitigations - IP blocking, rate limiting, robots.txt, firewalls, CAPTCHAs -
also block benign services and legitimate users.**

Yet machine users have always been essential:
search engines, aggregators and scholarly indexes are what make a repository discoverable at all.

Source: COAR (2025), The Impact of AI Bots and Crawlers on Open Repositories.

The real design problem is not human versus bot

We are asking the wrong question at the door.

Authenticated, not anonymous
Authorised, not extractive
Attributable, not opaque
Rate-bounded, not abusive
Retrieval for a task, not bulk training
Licence-compliant, not rights-indeterminate

Six binaries we can actually implement - none of which is 'is this a person?'

Three working answers already exist, at very different points on the spectrum

Wiley and Anthropic, 2025 pilot

An MCP interface exposes
subscribed journal content inside
Claude, preserving citations,
attribution and institutional
access.

Elsevier APIs

Scopus metadata and abstracts,
ScienceDirect full text, OAuth and
entitlement checking, tied to
organisational subscriptions.

OpenAlex

The open end: CC0 data, an API,
bulk snapshots, synchronisation
and agent-oriented access by
design.

Governed connectors, not indiscriminate crawling - the question is who governs them.

Subscriptions must evolve from reading rights to delegation rights.

From 'a human may read this interface' to 'a user may delegate bounded scholarly operations to an identified agent' - with delegated credentials, purpose declarations, content-level rights, quotas, auditability, source-return requirements and revocation.

Not by pretending the agent is human.

No stable licensing model has emerged - and price is the least of it

Arrangements observed so far

- One-time training licences
- Recurring content feeds
- Dual subscriptions
- Per-use compensation
- Content bundled into enterprise AI products

Unresolved, and not about price

- Author consent
- Correction and retraction propagation
- Provenance and citation links
- Whether generated synthesis devalues the version of record

We are licensing a product whose unit of consumption we have not yet defined.

The danger and the leverage are two sides of the same coin

The danger

If agent answers replace document visits, conventional usage metrics can make a valuable library look unused.

The leverage

Curated corpora, rights expertise, metadata, institutional identity and trusted provenance become scarce assets.

Watkinson: the question is not whether bot usage is bad, but whose machine, for what purpose, on what terms.

Institutional agent entitlement

Metered evidence retrieval

Public-interest open infrastructure

Licensed grounding services

Remuneration for attributed use

Library-built trusted agents

Information feudalism.

Researchers cultivate the outputs.
Libraries supply the trusted soil.
Platform owners control the plot and the harvest.

Our CHIIR paper puts it plainly: platform owners may extract disproportionate value from the farmers' work.

The profession has already written the requirements - twice

ALA, 2026 guidance

Adoption, pilots, delay, refusal, removal and discontinuation are all legitimate outcomes of a values-based review. It calls for human oversight, non-AI access routes, privacy review, explainability, labour consultation and environmental assessment.

ARL, 2026 workshop

Adoption is accelerating in staff training, distinctive collections and campus governance - while operational controls for vendor AI, provenance, IP compliance and agent authorisation remain underdeveloped.

Principled adoption, not adoption versus refusal: convert caution into requirements.

Three horizons: assistants, then gateways, then institutions

**Next 3 years: assistants
become normal**

1

AI embedded in discovery layers and scholarly databases. Libraries inventory vendor features, negotiate AI clauses, teach AI literacy and pilot collection-grounded assistants.

**3-5 years: agent gateways
and delegated access**

2

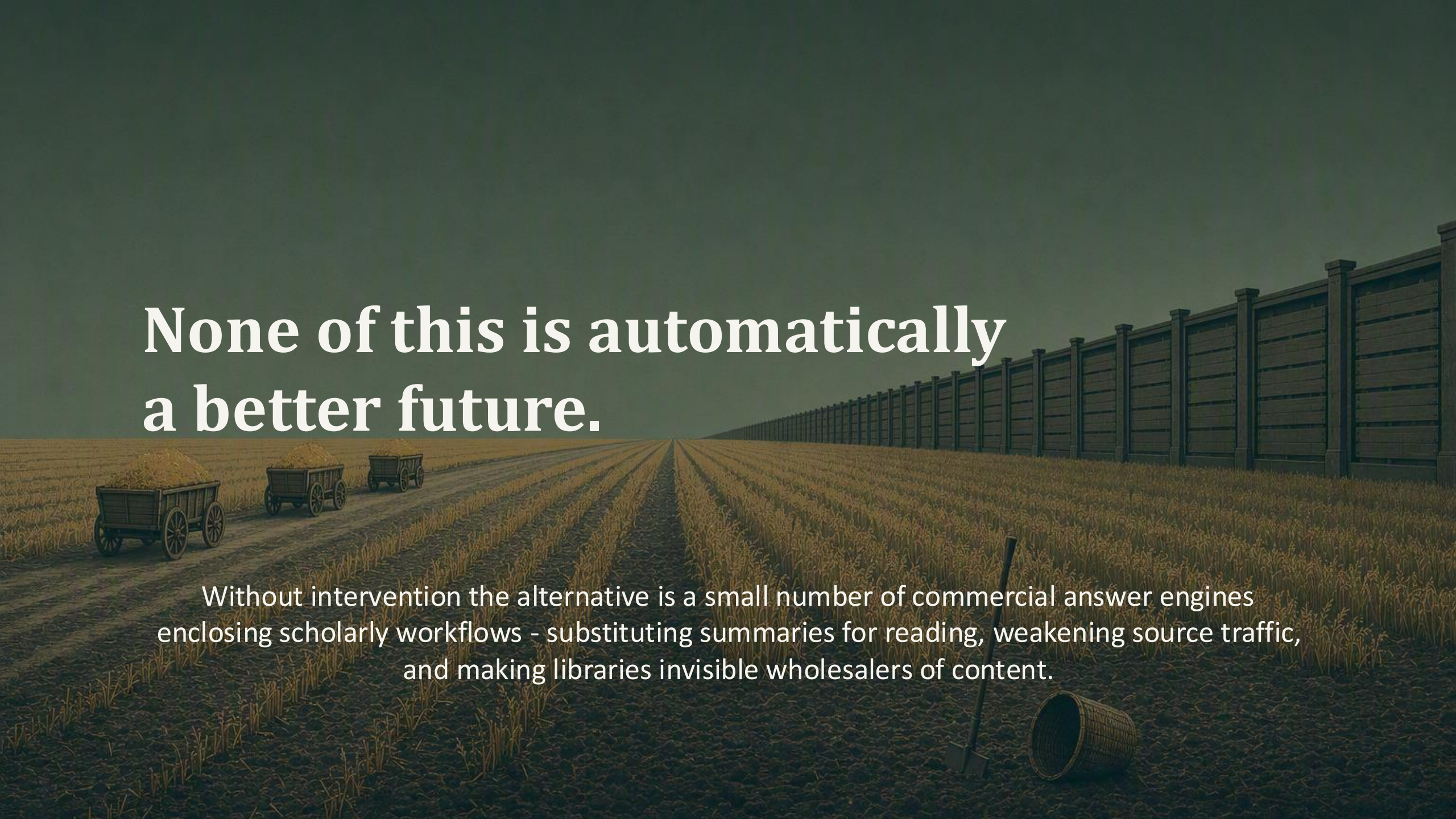
Standardised machine endpoints, entitlement-aware retrieval, agent identities, source-level citation requirements, new usage metrics. Open graphs become the common substrate.

**5-10 years: digital libraries
as epistemic institutions**

3

Less like websites, more like governed knowledge environments: preserving sources, exposing them computationally, maintaining scholarly memory, hosting institutional agents.

ICOLC: AI licence clauses should permit lawful non-commercial research and teaching, remain revisable, and impose neither unenforceable restrictions nor new user liabilities.

The background image is a surreal landscape. On the right side, a long, dark wooden fence with vertical posts stretches from the foreground into the distance. To the left of the fence is a wide, straight dirt road that also leads into the distance. Three wooden carts filled with golden hay are parked on the left side of the road. The ground on either side of the road is covered in a field of golden wheat. In the foreground, on the right side, there is a dark, tilled patch of earth with a pitchfork and a woven basket lying on it. The sky is a dark, uniform grey, suggesting a twilight or overcast setting.

None of this is automatically a better future.

Without intervention the alternative is a small number of commercial answer engines enclosing scholarly workflows - substituting summaries for reading, weakening source traffic, and making libraries invisible wholesalers of content.

Design principles for the farming library

Preserve roots with harvests

Every claim retains links to the passages, versions, licences and provenance it grew from.

Store cultivation history

Prompts, retrieval sets, corrections, annotations, user decisions, model and tool versions.

Make knowledge improvable

Versioning, challenge, branching and community maintenance, with corrections that propagate into saved crops and downstream agents.

Support foraging and farming

Preserve browsing and direct access to the document, with visible uncertainty and a route to a human expert.

Authorise agents explicitly

Delegated identity, scoped permissions, quotas, etc as reading, mining and training are not equivalent acts.

Measure epistemic outcomes

Coverage, entailment, provenance, reuse, learning and correction cost, not only speed.

Six things to take away

01

The document is no longer always the unit of access. It is evidence inside a cultivated knowledge object.

02

The future competitor to a DL is not another repository. It is the agent that becomes the user's primary interface.

03

The answer to bad bots is not a human-only web. It is governed machine access.

04

A trusted corpus is necessary but insufficient. Next-gen trust requires visible provenance and cultivation history.

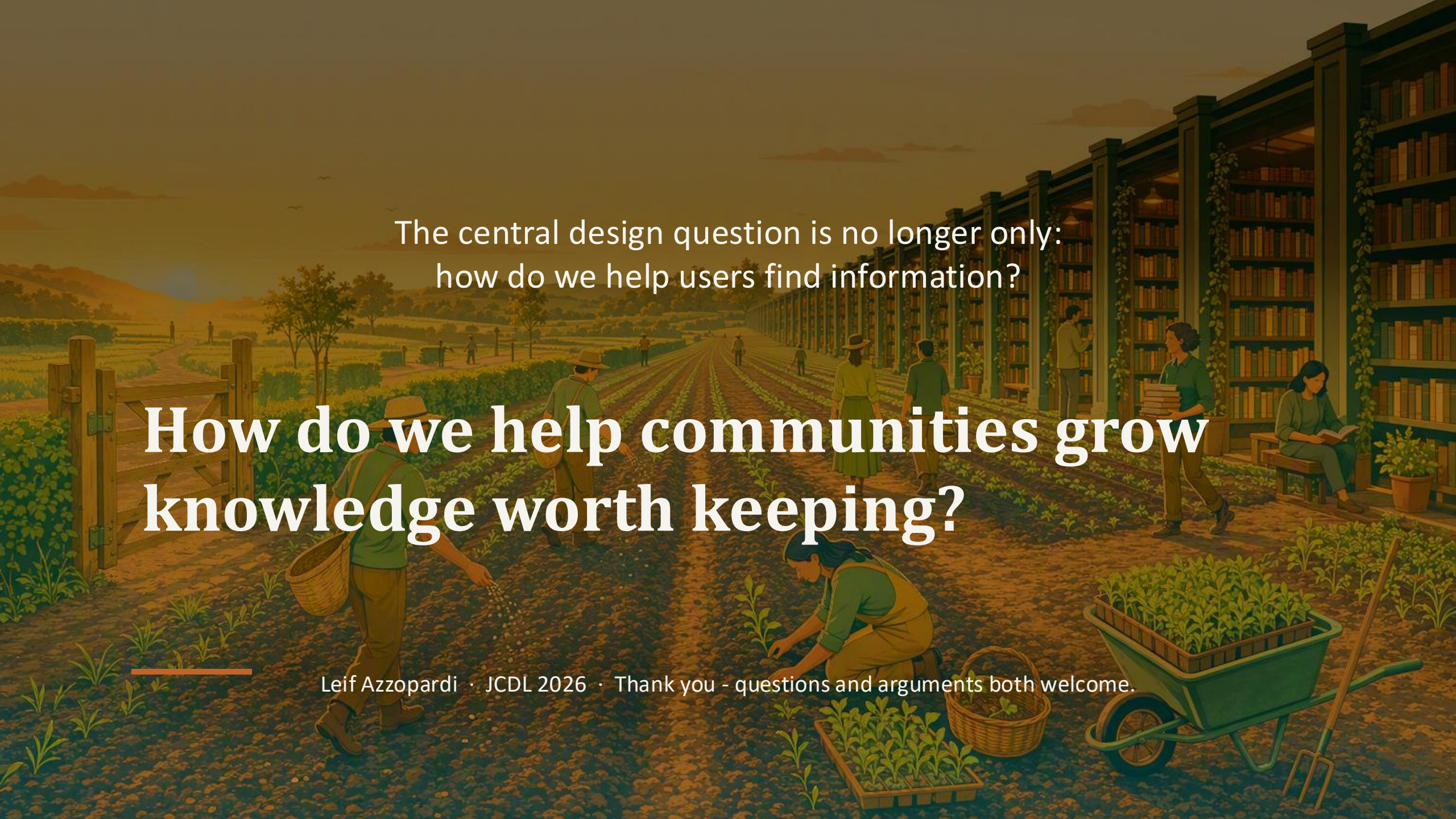
05

If DLs measure only clicks and downloads, they may record declining value precisely when their value is going up.

06

The opportunity is to make DLs the seed bank, the trusted soil and the steward of human-agent knowledge creation.

Seed, soil and steward - that is the institutional role worth claiming.



The central design question is no longer only:
how do we help users find information?

How do we help communities grow knowledge worth keeping?

Leif Azzopardi · JCDL 2026 · Thank you - questions and arguments both welcome.

Selected references

Azzopardi, L. & Roegiest, A. (2026). Information Farming: From Berry Picking to Berry Growing. CHIIR 2026. arXiv:2601.12544.

Bates, M. (1989). The Design of Browsing and Berrypicking Techniques for the Online Search Interface.

Pirolli, P. & Card, S. Information Foraging in Information Access Environments.

Kuhlthau, C. Information Search Process.

Russell, Stefik, Pirolli & Card (1993). The Cost Structure of Sensemaking.

Marchionini, G. (2006). Exploratory Search: From Finding to Understanding.

Scardamalia, M. & Bereiter, C. Knowledge Building and Knowledge Creation.

Borgman, C. et al. (2015). Knowledge Infrastructures in Science.

Asai, A. et al. (2026). Synthesizing Scientific Literature with Retrieval-Augmented Language Models. Nature 650(8103), 857-863.
doi:10.1038/s41586-025-10072-4

Lewis, P. et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020.

Asai, A. et al. (2024). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. ICLR 2024.

Gao, T., Yen, H., Yu, J. & Chen, D. (2023). Enabling Large Language Models to Generate Text with Citations. EMNLP 2023.
doi:10.18653/v1/2023.emnlp-main.398

Skarlinski, M. D. et al. (2024). Language agents achieve superhuman synthesis of scientific knowledge. arXiv:2409.13740.

Selected references

Walters, W. H. & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports* 13, 14045. doi:10.1038/s41598-023-41032-5

Barba, L. et al. (2026). Artificial Intelligence, Authentic Reactions.

Estlund & Hudson Vitale (2026). Strategic Implications of AI Futures for Research Libraries. ARL.

COAR (2025). The Impact of AI Bots and Crawlers on Open Repositories.

ALA (2026). Guidance on the Use of Artificial Intelligence in Libraries.

ICOLC (2024). Statement on AI in Licensing.

Ithaka S+R. Generative AI Licensing Agreement Tracker (last updated 7 August 2026). <https://sr.ithaka.org/our-work/generative-ai-licensing-agreement-tracker/>

Schonfeld, R. C. (2024). Tracking the Licensing of Scholarly Content to LLMs. *The Scholarly Kitchen*, 15 October. <https://scholarlykitchen.sspnet.org/2024/10/15/licensing-scholarly-content-llms/>

OpenAlex. Open scholarly graph: CC0 data, API, snapshots.

Wiley-Anthropic MCP pilot (2025).

Scopus with AI and Deep Research, Elsevier.

Web of Science Research Assistant, Clarivate.

Harvard public-domain corpus; ARL auditable shared corpora.