OPEN-SOURCE AND LOCAL AI FOR THE OTHER 95%

A Pragmatic Methodology for Academic Library–Led AI Literacy and Research Computing Infrastructure

Ray Uzwyshyn, Ph.D., MBA, MLIS
Director, Research & Technology Services
University of California, Riverside Libraries
raymondu@ucr.edu — rayuzwyshyn.net

Developed from a presentation delivered at the
National Science Foundation–sponsored Research Computing for Smaller Institutions (RCSI) Conference
Swarthmore College, June 16–17, 2026
Companion document: UCR Orbach Science Library STAR Lab AI Transformation Plan (v2)

Abstract

Traditional research computing serves a narrow tier of well-funded laboratories exceptionally well—the grant-supported engineering, biotechnology, and computer-science programs at universities that can sustain proprietary platforms, GPU clusters, cost-recovery models, and the staff to run them. The concentration is structural: the top thirty US universities accounted for forty-two percent of all academic R&D in fiscal 2023, and only 187 of more than five thousand US degree-granting institutions hold Carnegie R1 status[1][2]. Globally, the convergent top-200 cohorts of the major world rankings define a similarly slender band among more than twenty thousand higher-education institutions[3]. Roughly five percent of universities sit at the top of the research-funding pyramid; ninety-five percent do not.

Within that broad outside, the underserved population is larger still: most Ph.D.-level scholars in the humanities, social sciences, arts, business, medicine, and broader non-machine-learning STEM fields—even at well-resourced R1s—lack practical access to research-grade AI compute and the AI-literacy onboarding required to use it. The bottleneck is not headcount but a paradigm shift in thinking, learning, and interdisciplinary synthesis that the cloud-API model neither delivers nor democratizes.

This article presents a replicable, library-led methodology for closing the gap, resting on three pillars: a human-resource architecture pairing AI research librarians with AI/data research scientists; tiered, locally hosted computing built on open-weight models that drives recurring software and API costs toward zero; and a graduated "zero-to-hero" literacy pipeline carrying researchers and students from first exposure to independent deep research. Drawing on the University of California, Riverside Libraries' Research and Technology Services—anchored by the Orbach Science Library STAR Lab—the article supplies the staffing matrix, hardware tiers and costs, model stack, governance posture, phased roadmap, and four worked examples needed to adapt the blueprint to any budget and mission. Annotated appendices survey the leading open-weight models of mid-2026 and the principal supporting sources.

Keywords: *research computing; academic libraries; AI literacy; open-source AI; local large language models; digital scholarship; research data services; interdisciplinary research infrastructure; the other 95 percent; library-led AI*

[1]National Science Foundation, NCSES, Higher Education Research and Development Survey, FY 2023 (NSF 25-313, November 2024). US academic R&D totaled $108.8 billion; the top 30 institutions accounted for 42 percent; the non-S&E share (humanities, education, social work, law, arts) was 6 percent.

[2]Carnegie Classification, 2025 Research Activity Designations (ACE / Indiana University). R1: at least $50 million annual research spending and 70 research doctorates (187 institutions); R2: at least $5 million and 20 doctorates (139); the new Research Colleges and Universities tier: at least $2.5 million (216).

[3]Times Higher Education World University Rankings 2026 ranked 2,191 institutions; QS 2026 ranked 1,501; ARWU / Shanghai Ranking 2025 published its best 1,000 of more than 2,500 reviewed. The three converge on a top-200 cohort as the strict definition of globally research-intensive universities.

Contents

Note: page numbers in this Table of Contents update automatically. In Microsoft Word, right-click the contents listing and choose "Update Field" → "Update page numbers only" to refresh.

1. The Chasm in Academic Computing

Most large research universities operate two parallel computing economies. The first is well understood and well resourced: high-performance computing (HPC) built on proprietary platforms, dedicated GPU clusters, cost-recovery models, and million-dollar federal grant pipelines. It serves engineering, biotechnology, and computer science—disciplines whose funding assumes this infrastructure and whose researchers know how to claim it. Call this the over-served two-and-a-half to five percent.

The second economy is larger, quieter, and structurally underserved. It comprises two overlapping populations: the great majority of medium-sized and smaller institutions, where genuine research happens but capital, GPU access, and operating budgets for cloud APIs are simply absent; and the many researchers within higher-ranked institutions whose questions and data suit AI and machine-learning methods but who need onboarding—untrained in these methods, they lack a basic entry point. The group spans the disciplinary map—humanists and historians, sociologists and anthropologists, artists and architects, business and medical researchers, and engineers whose subfields sit outside the machine-learning mainstream. They arrive at the same methods—text mining, computer vision, geospatial analysis, large language models—but without GPU clusters, grant infrastructure to absorb cloud-API charges, training in the new methodologies, or a single staff member whose job is to help them begin. Priced out by the very pay-per-token services the popular narrative treats as universal access, they are the other ninety-five to ninety-seven percent.

The strategic question for these institutions and departments is not whether to engage AI, but *how—without the budgets, GPU clusters, educational priors, or grant infrastructure that proprietary models presuppose. The answer inverts the usual order: rather than buying access to someone else's cloud, the institution builds a modest, locally owned open-source capability and offers it with human support accessible to the entire university—locating it where the whole campus already comes for neutral, cross-disciplinary help: the research library.*

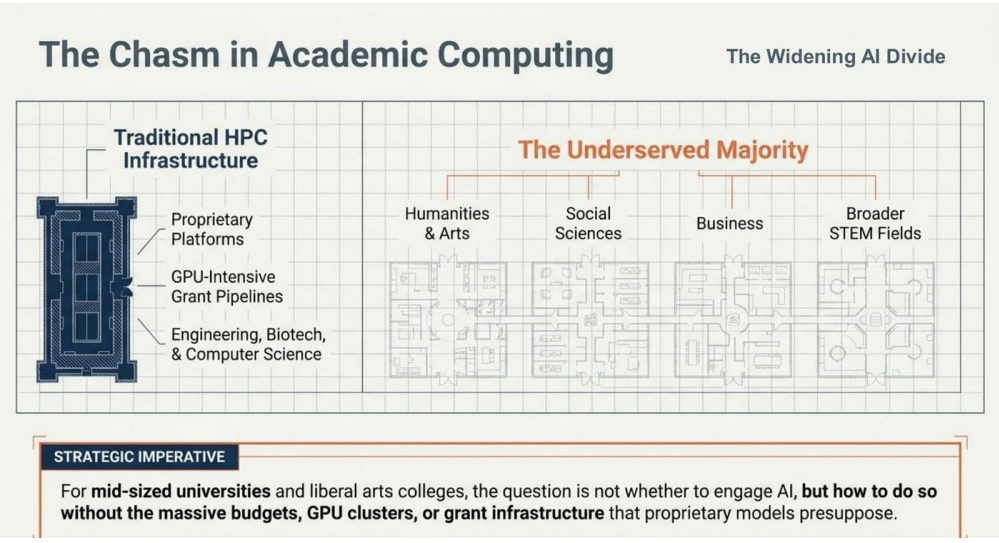

Figure 1. *The structural divide between traditional HPC infrastructure and the underserved majority of research disciplines.*

2. The Core Premise: Why the Academic Library, and Why Local and Open-Source

Two design commitments distinguish this model from the familiar defaults—enterprise licenses for the few, hyperscaler compute for a grant-funded minority, HPC cost recovery that leaves the broader campus outside.

2.1 The library as neutral interdisciplinary ground

The research library is the one campus entity positioned to synthesize data science, domain knowledge, and machine learning across every discipline without belonging to any of them. Most academic libraries already run data services, host repositories, and teach digital literacy across the curriculum; they serve every department, mediate a complex ecosystem of databases and digital-scholarship tools, and carry a user-oriented service ethic rather than a grant-recovery mandate or competitive positioning. Where an engineering-housed HPC center necessarily prioritizes engineering, a library lab is genuinely common infrastructure—giving the civil engineer a first exposure to Python, GitHub, and PyTorch while helping an art-history graduate student train an image classifier or an environmental-science postdoc fuse GIS fieldwork with AI methods in ArcGIS. All can walk in, be taken seriously, and find workshops.

2.2 Local execution and open weights

Frontier open-weight models have closed much of the capability gap with proprietary systems, and they can be run on hardware an institution owns outright[4]. They are now available in highly capable Western and Chinese variants—a plural ecosystem in which institutions can match selection to mission and governance. Choosing local, open-weight execution along with cloud APIs produces four compounding advantages:

- **Cost.** recurring per-token and subscription costs for open-source models fall toward zero, converting an unpredictable operating expense into a one-time, planned capital purchase.
- **Privacy.** sessions wipe clean, and institutional intellectual property never leaves campus—safe for the sensitive qualitative and human-subjects data researchers will not send to a third-party API, and low-stakes for learning how these methods apply to one's own research.
- **Control.** the institution is not exposed to opaque model deprecations, sudden price changes, or terms and models being altered or withdrawn by vendors or regulators—a real factor in 2026's volatile proprietary landscape.
- **Equity.** access is decoupled from the federal grants whose cost-recovery accounting would otherwise gate eligibility; participation is defined by curiosity and needs rather than budget line and prestige.

[4] 'Open-weight' models release downloadable parameter weights but typically not training data or code; 'open-source' in the strict OSI sense releases weights, data, code, and recipes. Most leading models here are open-weight; AI2's OLMo family is the most prominent fully open-source release.

This is the sense in which the academic research library, with its centralized budget and campus-wide mandate, becomes the great equalizer and an interdisciplinary hub for research collaboration: *a vendor-neutral, locally hosted, universally accessible, and more democratic AI infrastructure whose reason for being is to serve the university researchers the conventional model leaves out.*

3. Methodology at a Glance: Three Interlocking Pillars

The model is not a hardware purchase, a training series, or a staffing plan in isolation; it is the deliberate integration of all three, and each pillar fails without the others. Hardware with no one to teach it, or to translate complexity into something a Ph.D. in any discipline can grasp, sits idle. AI or digital-literacy programming with no compute on which to experiment and graduate into larger projects stalls at the introductory level. Staff with neither AI tools nor an onboarding curriculum become a help desk. The methodology, summarized below and elaborated in what follows, is to stand up the three pillars together and connect them with a flywheel that turns first-time visitors into independent AI researchers ready for larger, more complex journeys.

Pillar	What it provides	Failure mode if isolated
I. Human (AI teaching and research-support) infrastructure	AI research librarians and AI/data research scientists; consultations, workshops, meetups, and faculty partnerships; brokered, clearly defined relationships with HPC and enterprise IT so responsibilities are shared.	Knowledgeable staff reduced to a ticket queue; potential users alienated and never onboarded.
II. Computing infrastructure	Tiered, locally hosted workshops and workstations running an open-weight AI model stack inside the campus boundary, alongside university-purchased licenses, for both research and learning.	Expensive hardware no non-specialist can approach—high barriers, no friendly orientation.
III. Literacy pipeline	A graduated AI-literacy pathway—entry, skill-building, deep research—with workshops, a certificate, thematic weeks, and camps.	Faculty and student enthusiasm that dissipates at the first barrier and never converts into research output.

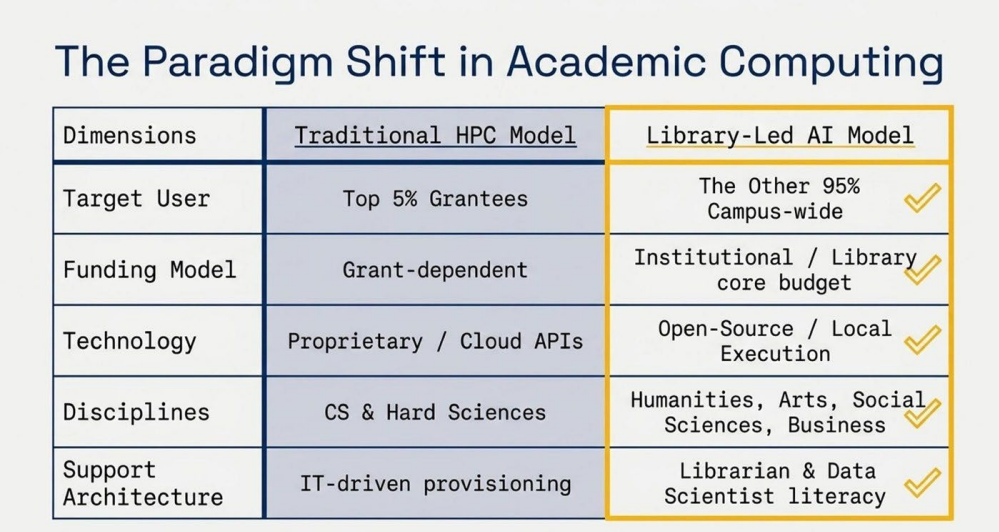

Figure 2. *The paradigm shift: from a grant-dependent HPC model toward a library-led, institutionally budgeted, open-source AI model.*

4. Pillar I — Human Infrastructure: Staffing the AI Ecosystem

Hardware is the cheapest part of this model. The decisive investment is people. We live in a technological society, a technocracy in which literacy itself—and specifically information, digital, and AI literacy—has become an imperative for serious research and workforce readiness in an AI-driven economy. A small, deliberately composed team whose explicit charge is to make emerging-technology methods approachable and workable for non-specialists is therefore not optional: it is the institution's answer to a generational shift in how research is conducted. UCR's Research and Technology Services division, housed in the Orbach Science Library, illustrates one effective composition among several possible.

4.1 The core digital and AI-literacy team

The unit blends established library-technology roles with the two roles that are the engine and center of the AI model. The table marks the pivotal hires.

Established library-technology roles	AI-engine roles (the pivotal hires)
Director of Research & Technology Services Innovative Media Librarian Digital Scholarship Librarian Geospatial Information Systems Librarian Medical & Clinical Outreach Librarian Maker Space / Robotics Lab Services Coordinator Research Services Information Assistant Cyberinfrastructure Manager and staff (Core Libraries IT:	AI Research Librarian (scaling to two) AI / Data Research Scientist (scaling to two) These roles convert tools and compute into literacy, workshops, meetups, consultations, and research output. They are the difference between a server room and a service.

Established library-technology roles	AI-engine roles (the pivotal hires)
programming, software/database support, IT project management, desktop support)	

4.2 A deliberate division of labor

The two pivotal roles are complementary, not redundant. Keeping them distinct lets the lab serve both a humanist who has never written a line of code and an engineer who needs a fine-tuned model on their own data. The AI-engine roles are planned to scale from the current two staff to four or six as demand grows; a backup teaching-and-learning unit is also possible, if staffing allows, for the wider university's more basic AI and information-literacy skills.

Role	Focus
AI Research Librarian	AI literacy and programming; interdisciplinary workshops; qualitative and humanities-facing consultations and integrations; the on-ramp for newcomers. Owns the pedagogy.
AI / Data Research Scientist	Python, R, more complex data modeling, and machine-learning architectures; one-on-one support for specific research-data questions; technical depth for advanced projects; co-leadership of operational lab transformation. Owns the build.
Cyberinfrastructure Manager	Hardware, networking, model deployment, security posture; integration with enterprise IT; operational reliability of the local compute environment.

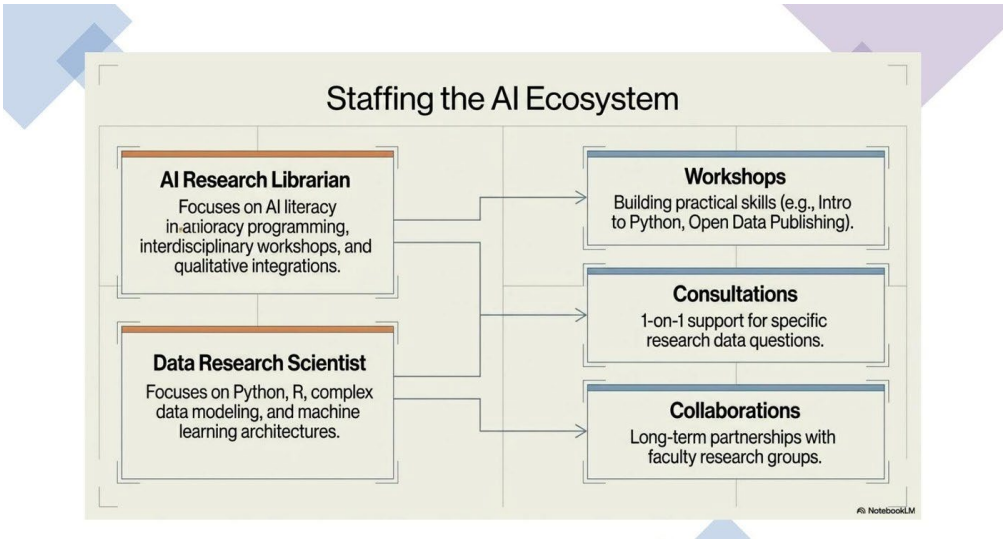

Figure 3. *Staffing the AI ecosystem—complementary roles producing workshops, consultations, and durable collaborations.*

4.3 Service outputs

The team's effort resolves into three durable service lines covering the entire literacy curve:

- **One-on-one research consultations.** deep, project-specific support for individual researchers, faculty groups, and departments—the highest-value, lowest-volume work.
- **AI literacy workshops and curriculum.** a recurring, public-facing program—serving research and teaching faculty, students, staff, and community—that builds practical skills at scale (introductory Python, open-data publishing, literature mapping, AI and deep-research model introductions, computational text analysis, PyTorch, and more).
- **Thematic weeks and events.** recurring campus-wide moments—UC Love Data Week, a summer Robotics Camp, an annual Maker Week, GIS meet-ups—that lower the social barrier to entry and feed newcomers into the pipeline.

4.4 Cooperative relationships, not turf

The library-led lab does not replace HPC or enterprise IT; it completes a three-body division of labor and explicitly hands researchers off when their needs outgrow—or do not rise to—the local lab. Naming these boundaries in advance prevents the political friction that kills cross-unit collaboration. Cooperative relationships are the name of the game[5].

Unit	Owns	Hand-off relationship
Research Library (the AI lab)	AI literacy and researcher support; use of AI tools in research design, data collection, and analysis; data management; model work on regular-sized datasets.	Onboards the 95 percent and refers the most compute-intensive projects upward to HPC.
Enterprise IT Services	University-wide proprietary AI licenses (e.g., Gemini, Vertex, limited OpenAI, Anthropic offerings); broad general training; security and identity.	Supplies the secured proprietary tier for use cases that warrant it.
HPC / Research Computing	High-performance AI compute for the top 2–5 percent of projects; major grant-supported workloads where significant costs are justified and cost recovery is needed; instructor development.	Receives mature projects the library has already onboarded, de-risked, and trained.

Framed this way, the library lab is an onboarding ramp onto the larger clusters rather than a competitor—a point worth making to HPC leadership early: it converts a potential rival into a beneficiary and strong collaborator.

[5]Mihoko Hosoi became University Librarian at UC Riverside in October 2025. The Research and Technology Services division reports to the University Librarian.

5. Pillar II — Computing Infrastructure: Hardware Tiers and the Local Model Stack

The infrastructure is deliberately modest and deliberately scalable: a credible lab can begin at a price a single department can defend and grow without re-architecting. Three reference tiers anchor the budgeting conversation. (This technology is current as of July 2026 and should be updated at procurement. Configurations balance processing power—capacity to hold larger open-source models—and speed.)

5.1 Hardware tiers for any institutional budget

Tier	Indicative cost	Configuration	Best for
Option 1 Baseline — "Starter Kit"	~$27,000 2 workstations	1× Alienware (RTX 5090) 1× DGX Spark (≈200B-parameter inference capacity)	Departmental pilots
Option 2 Targeted — "The Sweet Spot"	~$72,800 4 workstations	2× Dell Pro Max (RTX PRO 6000, 96GB) 1× Alienware (RTX 5090, 32GB) 1× DGX Spark (≈200B capacity)	Mid-sized campus hubs
Option 3 Comprehensive — "The Full Vision"	~$124,800 8 workstations	3× Dell Pro Max (RTX PRO 6000) 3× Alienware (RTX 5090) 1× DGX Spark + 1× Mac Studio (Apple Silicon, up to 512GB unified memory)	Full-scale research centers

Most institutions should begin with Option 2, the targeted expansion, which is the practical sweet spot for a campus-wide hub. If a larger facility is required, the comprehensive tier scales linearly—roughly Option 3 doubled or tripled—without changing the underlying design[6][7]. The DGX Spark deserves a note: NVIDIA rates it for inference up to approximately 200 billion parameters and fine-tuning up to roughly 70 billion. It excels at prototyping; its 273 GB/s memory bandwidth makes a dedicated RTX PRO 6000 the better choice for higher-throughput serving. Pricing through 2026 has been affected by industry-wide GPU and memory supply pressures; treat the figures above as point-in-time floors and obtain current vendor quotations when procuring.

[6]NVIDIA DGX Spark specifications: inference on models up to about 200 billion parameters, fine-tuning up to 70 billion; two units linked via ConnectX-7 networking address about 405 billion. Launched at $3,999; $4,699 as of February 2026.

[7]Hardware prices and street availability through mid-2026 have been affected by a memory and GPU supply squeeze. The figures reported here should be treated as approximate, point-in-time estimates rather than fixed quotations. Institutions are encouraged to obtain current vendor quotations at the moment of procurement.

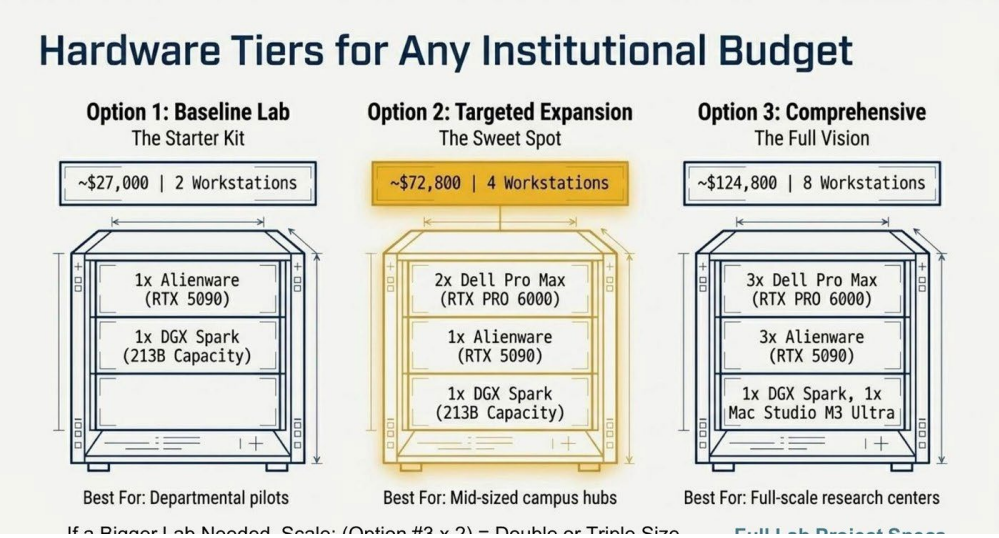

Figure 4. *Three indicative hardware tiers, scalable to any institutional budget; cost and component figures should be verified at the moment of procurement.*

5.2 The model stack: proprietary alongside open-weight

The hardware exists to run open-weight models locally. The table maps research needs to a proprietary cloud tier (high recurring cost, opaque usage) that enterprise IT can supply where warranted, and a local open-weight tier (zero recurring fees, complete local control) that the lab itself runs. The posture is both/and rather than either/or: the tiers complement one another. The mid-2026 ecosystem includes strong Western and Chinese contributions in every capability class; by a leading independent index, the best open-weight models cluster within a few points of the leading proprietary systems—a meaningful but incremental gap[8]. Every model in the local roster below runs on the Option 1–3 hardware as configured: the DGX Spark addresses roughly 200-billion-parameter models (two linked units, roughly 405 billion), and a 512GB-unified-memory Mac Studio runs 400–700-billion-parameter models at four-bit quantization—quantization keeps near-frontier open models inside the lab. Only trillion-parameter giants remain beyond it, candidates for the Section 4.4 HPC hand-off or metered cloud. Appendix A details each model.

Need	Proprietary / cloud tier	Local open-weight roster (capability class)
Deep reasoning	Frontier proprietary reasoning models (cloud, metered).	OpenAI gpt-oss-120b; OLMo 3-Think 32B; NVIDIA Nemotron 3 Super;

[8]The Artificial Analysis Intelligence Index (AAI) is a weighted composite metric published by Artificial Analysis, an independent AI-benchmarking firm. The current version, v4.1 (June 16, 2026), aggregates nine challenging evaluations across agentic, reasoning, scientific, and coding domains, including GDPval-AA v2 (20 percent), Terminal-Bench 2.1 (16 percent), tau-3-Bench Banking (14 percent), Humanity's Last Exam (12 percent), SciCode (8 percent), GPQA Diamond (6 percent), and others. Scores are reported on a 0-100 scale; frontier models in mid-2026 cluster in the 40s through 60s. The published 95 percent confidence interval is below 1 percent. A 3-6 point gap therefore reflects a meaningful but incremental capability difference, not a step-change. See https://artificialanalysis.ai.

Need	Proprietary / cloud tier	Local open-weight roster (capability class)
		DeepSeek V4 Flash (two linked DGX Sparks).
Heavyweight multimodal	Proprietary multimodal and deep-research services.	Llama 4 Scout (ultra-long context; RTX PRO 6000 tier); Mistral Small 4; Gemma 4 31B (image and video).
Real-world utility	General-purpose commercial assistants.	Qwen3.6-27B and Qwen3.5-30B-A3B; Ministral 3; Microsoft Phi-4.
Lightweight versatility	Smaller hosted endpoints.	Phi-4-mini (3.8B); Gemma 4 small variants; Nemotron 3 Nano; gpt-oss-20b; Ministral 3B.
Fully open-source (training data + code + weights)	(Not typically available.)	Allen Institute for AI: OLMo 3 (7B, 32B; OLMo 3-Think).
Frontier-scale and beyond	Frontier proprietary flagships (cloud, metered).	Quantized on the Mac Studio (512GB): GLM-5.2; Mistral Large 3; Nemotron 3 Ultra; Qwen3.5 397B. Trillion-parameter scale (HPC or cloud): Kimi K2.6; DeepSeek V4 Pro.

A necessary caveat on AI-model release cadence and velocity. The open-weight frontier turns over on a timescale of weeks rather than years. Treat any named model as a placeholder for the current best open-weight option in its capability class, and re-verify the roster at deployment and at each refresh. The architecture, the staffing, current benchmarking, and the literacy pipeline are the durable contribution; the model list is the one layer expected to change at least quarterly, and the local-execution design is precisely what makes swapping models inexpensive. A diplomatic point matters here: the strongest open-weight models in mid-2026 include both Western and Chinese systems, and a credible academic lab is well served by selecting models on capability, license, governance posture, and fit—not on origin alone.

5.3 Data privacy and the campus boundary

The privacy argument is often the single most persuasive point with faculty, IRBs, and provosts, and it follows directly from local execution. In the proprietary cloud pattern, a workstation ships data to an external API, incurring recurring token costs, opaque training usage, and genuine privacy exposure. In the local STAR Lab pattern, the same workstation sends data only to a model executing inside the campus boundary.

- zero recurring costs;
- sessions wiped clean after use;

- institutional intellectual property remains on campus;
- the environment is safe for sensitive qualitative and human-subjects data.

For large swaths of humanities and social-science research, this is not a convenience—it is the precondition that makes using these methods permissible at all. The UCR STAR Lab soft-launches in Fall 2026 as a physical, in-house installation; a terminal-based option for on-campus participants is in active development.

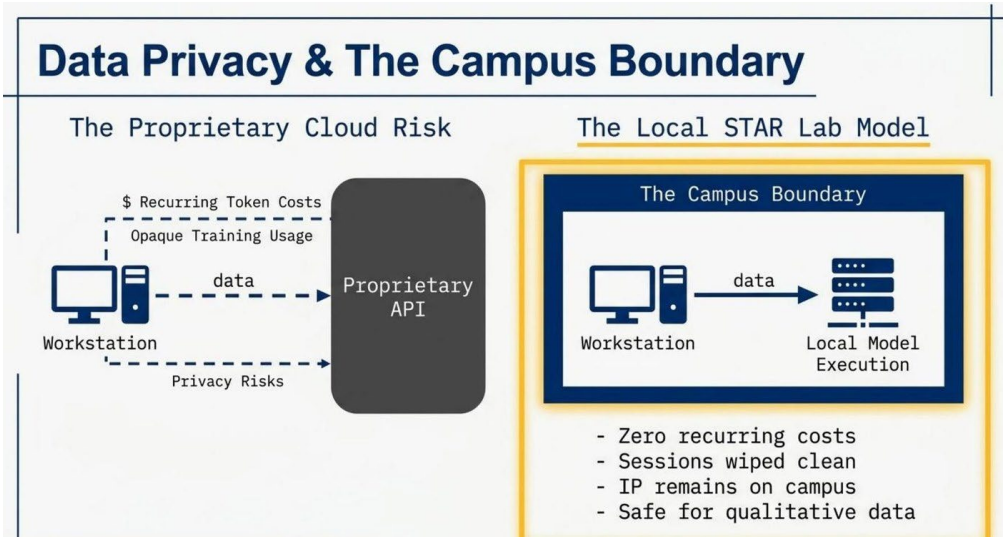

Figure 5. *Data privacy and the campus boundary—locally hosted execution keeps recurring costs, training usage, and intellectual property inside the institution.*

6. Pillar III — The AI Literacy Pipeline: From Zero to Hero

Access without onboarding produces an empty lab. The literacy pipeline carries a curious newcomer, in deliberate stages, from a first low-stakes encounter to running their own models. Each stage has a distinct goal and activities.

Stage	Activities	Goal / outcome
1. Entry & Exposure	Introductions to AI LLMs and deep-research models; introductions to physical computing and data science; introductions to AI ethics; AI/GIS meet-ups; streaming watch-parties; thematic-week events; graduate-student and junior-faculty AI lecture series.	Low-barrier introductions and wide entry for everyone, regardless of discipline or background, matching existing research interests to AI possibilities.
2. Skill Building	Introduction to Python for data analysis, cleaning, and AI; computational text analysis and software tools; AI-enhanced research data visualization; introduction to	Formalized AI literacy and the confidence to build with code-assistive tools, APIs, open-

Stage	Activities	Goal / outcome
	"vibe coding" and AI-assisted development; Raspberry Pi jams and robotics and Arduino introductions; the Library's Digital Scholarship Certificate.	source models, and agentic-AI possibilities.
3. Deep Research Execution	Running local LLMs; introductory and multi-workshop walkthroughs of PyTorch model development for research; interdisciplinary synthesis of complementary technologies (GIS, data visualization, digital-scholarship publishing) on real projects.	Independent STAR Lab utilization—the newcomer is now an AI-enabled researcher able to advance their work, write competitive grants, and produce AI-related methodological scholarship in their discipline.

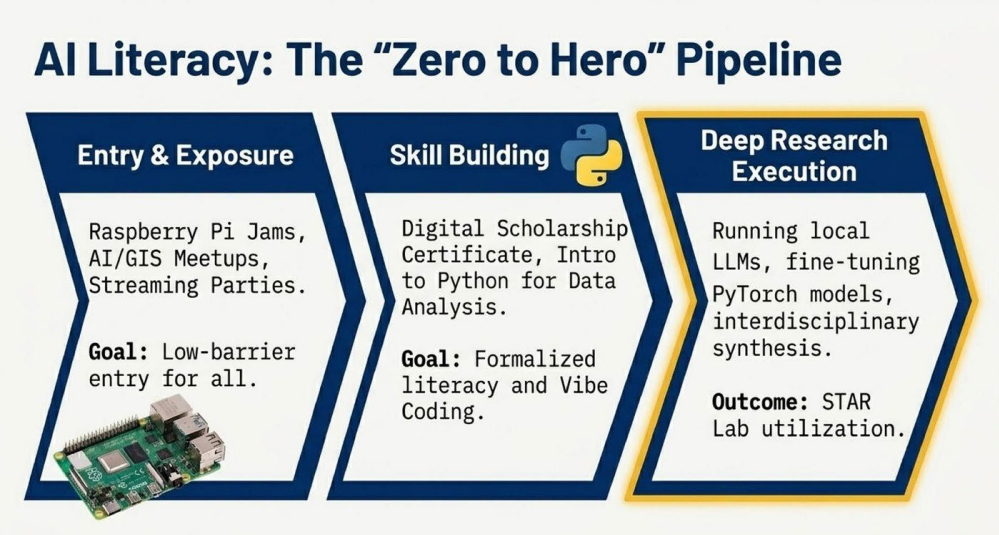

Figure 6. *The AI-literacy pipeline—three deliberate stages from entry and exposure through skill-building to independent deep research.*

6.1 The ecosystem flywheel

These stages are not a one-time funnel but a self-reinforcing loop. Digital and AI literacy plus entry feed research experimentation—lab usage, local model fine-tuning, consultations—which produces deep interdisciplinary research. That research, showcased through faculty presentations and thematic weeks (Makerspace Week, UC Love Data Week, Robotics Camp Week, GIS Week), draws the next cohort—inspired by the visible results—into the entry-level events. Every completed project becomes recruiting material for the next, and the lab's visible output is what justifies its continued resourcing. The flywheel answers the perennial question of how a small team sustains momentum and develops an AI-positive workforce: once the loop is turning, demand no longer needs manufacturing.

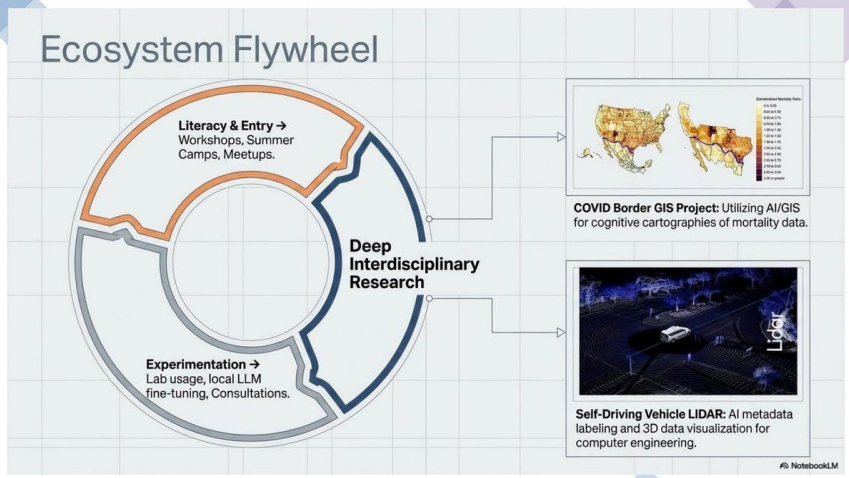

Figure 7. *The ecosystem flywheel—literacy and entry feed experimentation, which produces deep interdisciplinary research that recruits the next cohort.*

6.2 Curricular scaffolding

Two structures give the pipeline durability beyond ad-hoc events:

- **A Digital Scholarship Certificate.** a stackable credential with separate graduate and undergraduate tracks, a defined workshop sequence, and a capstone project. The certificate turns scattered attendance into a recognizable accomplishment that researchers and students can put on a CV and that administrators can count. Introductory digital-scholarship certificates can later be developed for areas from AI-enhanced robotics to AI-enhanced GIS and innovative media. The key is strong foundations, with the instructor providing affordances with AI as the learning foundations are set.
- **A standing workshop series.** a published, recurring digital, information, and AI literacy workshop calendar—machine learning, customizing AI models via API, AI literature mapping, GIS and drones, data cleaning for ML with Python, computational text analysis, deep-research models, deep learning with PyTorch, open-access publishing—so that a researcher who misses one offering knows the next is coming.

7. Why It Works: The Library as Interdisciplinary Nexus

The model succeeds because it sits at an intersection no other campus unit occupies. Three competencies have to meet for interdisciplinary AI work to happen: data science and machine learning; traditional research and domain knowledge; and library technical expertise (digital scholarship, data visualization, geospatial systems, makerspaces, and robotics). Departments hold the first two in isolation. The library is the only entity that can broker all three at once and offer them as neutral ground.

The existing service map already radiates from the traditional research article and monograph into AI-adjacent methods: AI-powered mapping and spatial humanities through GIS; computational humanities

and interactive visualization through digital scholarship; research dashboards and data art through visualization; physical prototypes through the makerspace; intelligent robotics through the robotics lab. The AI lab does not bolt a new function onto the library; it supplies the connective compute and literacy that let these strengths combine. This is why a library is the right host: the interdisciplinary surface area already exists and merely needs a shared engine.

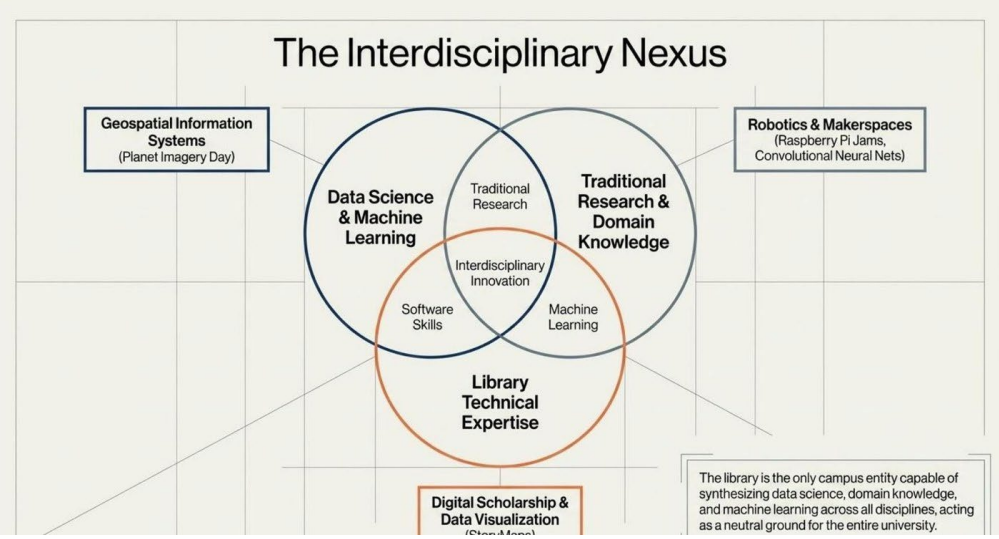

Figure 8. *The interdisciplinary nexus: data science, domain knowledge, and library technical expertise meet on neutral ground.*

8. Infrastructure in Action: Four Worked Examples

The model is best understood through the research it has already enabled at UCR. Each example pairs the same infrastructure inputs—hardware, a local model, data-scientist support, and a domain researcher—with a tangible outcome, and each comes from outside the conventional HPC client base.

8.1 The COVID Border GIS / Data Science Project

This project began with the School of Medicine dean and a consultation among the data research scientist, the digital scholarship librarian, and the GIS librarian. A tiger team was then drawn from the four, with biostatistician and graduate-student support from medicine, to plan infrastructure for mapping mortality data along the US–Mexico border—fusing statistical and qualitative inputs with data-assisted analysis to produce cognitive cartographies of standardized mortality ratios. Work of this kind sets a foundation of trust on a complex, squarely interdisciplinary project that would have been difficult to seat in any single department's computing facility. The project, now complete with publications pending, is prepped with team and faculty for next-level AI-insight experimentation.

8.2 Autonomous-Vehicle LIDAR: 3D Data Support and Metadata Labeling

For computer engineering, the lab supported both AI LIDAR metadata labeling for later model training and post-training three-dimensional visualization and archiving of point-cloud data for download and review. The project demonstrated that the same modest infrastructure that serves a humanist also serves a hard-

engineering use case—lowering the barrier for students who are not yet HPC-credentialed and opening doors for faculty without major HPC grant support who nonetheless need basic AI preparatory work and the post-work of archiving valuable self-driving-vehicle and Bluetooth experimental 3D LIDAR data.

8.3 AI-Enhanced Robotics Summer Camp

A zero-to-hero camp taught students the basics of physical computing and robotics, adding a prototype of AI to the robots (cars or cats, letting students flex their creativity through 3D printing). Students then learned to train a basic convolutional neural network to recognize objects and colors (red, yellow, and green) from images and to integrate that model into a working robot—covering training and real-world deployment, with results benchmarked. Notably, attendance ran roughly three-quarters women and spanned an unusually broad range of majors, from electrical engineering and biochemistry to creative writing and political science—direct evidence that a low-barrier, ungated on-ramp reaches populations the default, more strictured engineering model misses.

8.4 Planet Imagery Day (AI + GIS)

In partnership with the USDA-NIFA Artificial Intelligence for Sustainable Agriculture project, the library hosted Planet Imagery Day on May 2, 2025, featuring vendor-led sessions from Planet and Esri on using satellite imagery in Google Earth Engine and ArcGIS, alongside faculty from environmental sciences and computer science presenting examples from precision agriculture. Postdoctoral researchers presented alongside the vendors, showing how drones combined with satellite imagery can mitigate everything from soil-moisture erosion to blight in California's billion-dollar Central Valley agriculture industry—one of the world's largest economies in this respect. The workshop illustrated the lab's convening, interdisciplinary role: the library brokered industry, funded research, and cross-departmental faculty into one accessible event of mutual learning and cutting-edge research practice.

9. A Phased Implementation Roadmap

The blueprint is best deployed in deliberate phases. The UCR sequence below generalizes into a larger template: establish the foundation, soft-launch with a single contained pilot, expand the curriculum across disciplines, grow the ecosystem, work out bottlenecks, and only then pursue institution-scale impact. Phasing paces ambitions: a small, under-the-radar team given a few years to work out its infrastructures can grow, is protected from over-commitment, and gives leadership visible milestones and evidence against which to fund larger projects.

Phase	Focus
Foundation	Stand up the workshop classroom, smaller lab spaces, baseline infrastructure, and core emerging-technology staff, beginning with digital and data literacy. For AI, secure the two pivotal hires and integrate them with the larger team toward AI/data centrality.

Phase	Focus
Soft launch	Activate the dedicated AI lab on a small scale, deploy the baseline open-source model stack, test with staff, and run a contained graduate-student pilot to surface possibilities and problems before scaling. Begin with the two-workstation infrastructure and expand to four and eight.
Curriculum expansion	Integrate disciplinary AI workshops with the other emerging technologies (GIS, digital scholarship, innovative media, makerspace, and robotics lab), then expand disciplinary reach across the humanities, social sciences, and STEM. This is where the "other 95 percent" framing becomes operational.
Ecosystem growth	Launch signature initiatives—Maker Week, UC Love Data Week, Open Access Week, GIS Week, a Women in AI and Robotics colloquium—and expand into adjacent areas such as AI-enhanced robotics, GIS, and innovative media, organically and simultaneously.
Institutional impact	Serve the other 95 percent at scale, contribute to workforce development, pursue R&D and workforce grants, and position the lab as a transferable national prototype.

10. Governance, Risk, and Sustainability

A pragmatic methodology has to name its own risks. Five recur, with their mitigations:

1. **Model churn.** the open-weight roster will date quickly. Mitigate by treating models as swappable and budgeting for periodic quarterly or semester refresh rather than one-time perfection. Assign managers and desktop support to monitor open-weight benchmarks and test suitable models for these updates.
2. **Key-person dependency.** a two-person engine is fragile. Mitigate by documenting workflows, cross-training the maker/robotics and information-assistant roles, and scaling the pivotal hires as soon as demand is demonstrated.
3. **Turf friction.** HPC or enterprise IT may read the lab as encroachment on their territory. Mitigate by defining cooperative boundaries publicly and early, and by framing the lab explicitly as an onboarding ramp that feeds the clusters—and as strong help to central ITS in covering the workshopping, teaching and learning, and AI onboarding that fall outside enterprise-licensed proprietary models.
4. **Governance and ethics.** local execution removes API privacy exposure but does not remove responsibility. Mitigate by coordinating with university policies on clear data-handling norms, and by reserving the local tier for sensitive data while routing other work to the appropriate proprietary tier as university or state policies specify.

5. **Funding durability.** capital purchases age. Mitigate by anchoring the lab to core library IT budget expansion rather than soft money, while pursuing R&D and workforce-development grants for innovation rather than for survival.

Anchoring the operating model to the institutional and library core budget—and negotiating with central IT for this role rather than depending on grants—is itself the central sustainability decision, because it keeps access decoupled from any individual researcher's funding and the lab open to the 95 percent. The library is already the largest computing lab on most university campuses; that fact extends naturally to the current AI paradigm shift.

11. Transferability: Adapting the Blueprint to Your Institution

The library's interdisciplinary value proposition for academic institutions rests on three transferable claims. First, **affordable scale**: recurring software and API costs drop toward zero through local open-weight execution, turning an operating liability into a one-time capital decision even as token costs skyrocket. Second, **true inclusion**: the model is designed from the outset for the 95 percent of researchers outside traditional HPC grant fields, as well as for the broader band of global institutions sitting outside the strict top-200 research-intensive cohort[9]. Third, **a transferable blueprint**: the hardware, software, and staffing matrix presented here is already present in fundamental form—as digital and data literacy—at most academic and research libraries, and then scales to AI budgets, from an additional AI starter kit to a full research center.

A practical adoption sequence:

1. **Site it.** locate the local AI lab in the research library to claim neutral, cross-disciplinary ground.
2. **Minimally staff the engine first.** prioritize the AI research librarian and the AI/data research scientist over additional hardware; the team is the support, the workshops, the value added, and the differentiator.
3. **Buy a tier you can defend.** default to Option 2 if budget allows, Option 1 if it does not, and plan the upgrade path as clientele increases and staff get their legs, rather than over-buying.
4. **Stand up local models.** deploy a current open-weight stack locally and document the swap procedure and its results.
5. **Open the pipeline.** build the three-stage digital, data, and AI literacy ladder and at least one signature recurring event before chasing the eight-workstation AI lab scale.
6. **Broker the partnerships.** agree tacitly on cooperative, overlapping boundaries with HPC and enterprise IT through the circulation of documented plans and/or MOUs in writing if needed.

[9]Association of American Universities, founded 1900, comprises 71 leading research universities, 69 in the United States and 2 in Canada. UC Riverside was elected to membership in 2023. The Association of Research Libraries (ARL) is the parallel North American consortium of major research libraries; the UCR Library is an ARL member.

7. **Prove it with one project.** implement and document at least one to three worked faculty and graduate-student research examples, publicize them, and let the flywheel recruit the next cohort and/or build out.

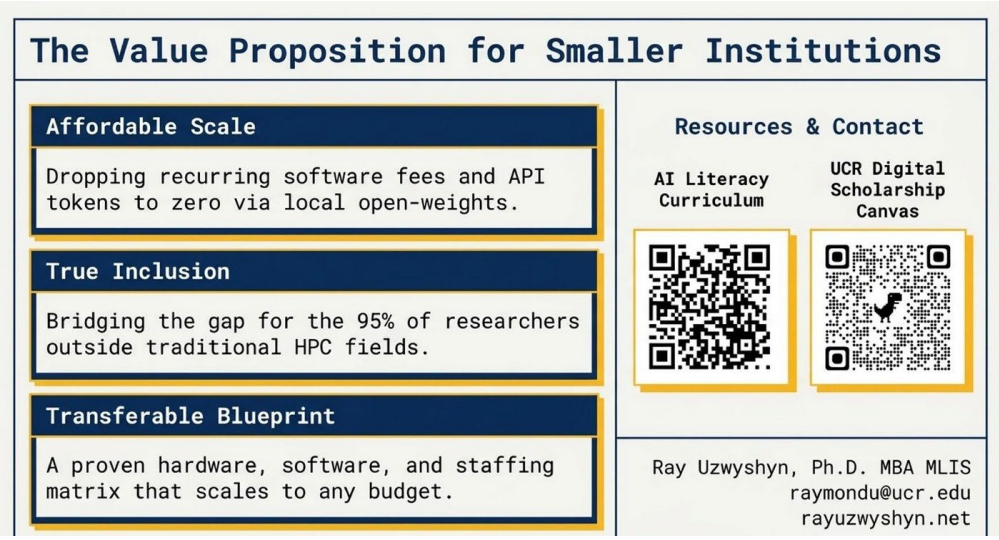

Figure 9. *Three transferable claims for smaller institutions—affordable scale, true inclusion, and a transferable blueprint.*

12. Conclusion: Toward a Paradigm Shift in Interdisciplinary Discovery

Traditional research computing answered the question of who gets to use advanced infrastructure, and answered it narrowly. A library-led, locally hosted, open-source AI lab changes the question. By pairing a small, deliberate human team with affordable owned hardware and a graduated literacy pipeline—and by siting all three in the one place on campus that already belongs to everyone—an institution can extend genuine AI capability to the researchers the conventional model leaves out, at a cost it can sustain from its own budget.

The deeper payoff is not access for its own sake. It is the paradigm shift in thinking, learning, and interdisciplinary synthesis that AI literacy makes possible. When a sociologist can co-design a text-mining analysis with the AI research librarian, when a historian can train an image classifier on archival photographs in an afternoon, when an environmental-science postdoc can fuse satellite imagery with social-survey data inside the campus boundary—these are the moments at which AI tools translate into discovery rather than spectacle. The model proposed here is, finally, an argument that the institutions best positioned to lead the next era of interdisciplinary discovery are the ones that make these methods ordinary.

The question is no longer who gets to use AI, but what our institutions will discover when everyone can.

Appendix A: Leading Open-Weight Models (mid-2026)

This brief identifies the leading open-weight models of mid-2026 across the Chinese and Western/international ecosystems. The Artificial Analysis Intelligence Index (AAI), cited throughout, is an independent 0-100 composite benchmark (v4.1, June 16, 2026)[10]: frontier proprietary models score in the high 50s and low 60s, the leading open-weight models in the mid-40s through low 50s—meaningfully behind, but well within research-useful range. *Open-weight* (downloadable weights) is distinguished from *fully open-source* (weights plus training data and code) model by model[11]. Facts verified against the June–July 2026 research pass (Artificial Analysis v4.1 and vendor documentation); the roster will date—re-verify at deployment.

A.1 Chinese ecosystem (the open-weight performance leaders)

1. Kimi K2.6 — Moonshot AI

One trillion total / 32 billion active MoE, 256K context, multimodal (MoonViT), modified MIT license (April 2026). At or near the top of the open-weights AAI; strongest on long-horizon agentic work. Deployment: trillion-parameter scale—beyond the lab tiers; an HPC or cloud candidate.

2. DeepSeek V4 (Pro and Flash) — DeepSeek

V4 Pro: 1.6 trillion total / 49 billion active. V4 Flash: 284 billion / 13 billion. One-million-token context, MIT-style license. Pro is among the top open reasoning models on AAI v4.1; Flash is the deployable variant. Deployment: Pro exceeds the lab tiers; Flash runs on two linked DGX Sparks or the Mac Studio.

3. GLM-5.2 / GLM-5.1 — Z.ai (Zhipu)

753 billion total / roughly 40 billion active, one-million-token context, MIT license (June 2026). Among the strongest text-only open models on AAI v4.1; a leader in agentic coding. Deployment: four-bit-quantized on the Option 3 Mac Studio (512GB); smaller GLM variants run on any tier.

4. Qwen3.5 / Qwen3.6 family — Alibaba

Qwen3.5: 397 billion / 17 billion active, 201 languages, Apache 2.0, 262K context; dense 27B and 30B-A3B variants run on a single consumer GPU. The flagship Max versions are API-only—use the open variants. Deployment: small variants on every tier; the 397B flagship on the Mac Studio or two linked Sparks.

5. MiNiMax M3 — MiNiMax

Released June 2026; the first open-weight model combining frontier coding, a one-million-token context, and native image and video multimodality. At or near the top of the open-weights AAI v4.1. Deployment: frontier-scale—verify memory fit at deployment.

6. Xiaomi MiMo V2.5 Pro — Xiaomi

One trillion total / 42 billion active, one-million-token context; ties the open-weight leaders on AAI with strong token efficiency. Some variants are closed—verify the checkpoint. Deployment: trillion-parameter scale—beyond the lab tiers.

A.2 Western and international ecosystem (the open-weight pluralists)

7. NVIDIA Nemotron 3 (Ultra / Super / Nano) — NVIDIA

Ultra: 550 billion / 55 billion active—the most intelligent US open-weights model on AAI v4.1. Super: 120 billion / 12 billion hybrid Mamba-Transformer, one-million-token context, notably fast. Nano: 30 billion. NVIDIA Open Model License with open training data. Deployment: Super and Nano on a single lab workstation; Ultra quantized on the Option 3 Mac Studio.

8. OpenAI gpt-oss (120B / 20B) — OpenAI

OpenAI's first open weights since GPT-2 (August 2025). The 120B MoE (5.1 billion active, Apache 2.0) runs on a single high-end workstation in native MXFP4; the 20B runs on a 16GB GPU. Deployment: every lab tier.

9. Google Gemma 4 (31B and smaller) — Google DeepMind

Dense multimodal 31B, Apache 2.0, 256K context, from the Gemini 3 lineage. Arguably the easiest serious model to stand up locally—image and video inputs on a single 24GB GPU at four-bit quantization. Deployment: every lab tier.

10. Meta Llama 4 (Scout / Maverick) — Meta

Scout: 109 billion / 17 billion active MoE with a ten-million-token context—the differentiator for corpus-scale work. Its benchmark standing was overtaken in 2026, but installed base and tooling keep it relevant. Deployment: the RTX PRO 6000 tier at four-bit quantization.

A.3 Strong international additions worth including for institutional planning

11. Mistral 3 / Mistral Large 3 — Mistral AI (France)

Apache 2.0 family. Large 3: 675 billion / 41 billion active (December 2025). Ministral 3 (14B/8B/3B) runs from a 16GB GPU down to laptops; Small 4 (119B / 6B active) unifies reasoning, multimodal, and coding. EU data residency suits jurisdiction-sensitive institutions. Deployment: Ministral on every tier; Small 4 on the RTX PRO 6000 tier; Large 3 quantized on the Mac Studio.

12. Microsoft Phi-4 family — Microsoft Research

MIT-licensed family, 3.8B–15B, including a reasoning-vision variant (March 2026). Small models trained on curated synthetic data that match models several times their size—the "other 95 percent" hardware thesis in miniature. No "Phi-5" exists as of mid-2026. Deployment: every tier, down to laptops and Raspberry Pi.

13. OLMo 2 / OLMo 3 — Allen Institute for AI

The Western standard-bearer for strict open source: weights, training data, code, logs, and checkpoints all released. OLMo 3 (7B, 32B; November 2025) includes OLMo 3-Think 32B, the strongest fully open-source reasoner. The most defensible choice where reproducibility and auditability rule. Deployment: every lab tier.

Honorable mentions: Mistral Devstral (code), smaller GLM variants, Yi, DBRX, and Falcon for niche deployments.

Appendix B: Annotated Bibliography

The following are the principal sources for the empirical claims and the conceptual framing of this article. Annotations identify why each source matters to the methodology.

B.1 Research-funding concentration and the institutional landscape

National Science Foundation, National Center for Science and Engineering Statistics. *Higher Education Research and Development Survey, Fiscal Year 2023* (NSF 25-313). Alexandria, VA: NSF, 2024. — The primary source for the concentration argument: top-thirty institutions captured forty-two percent of FY 2023 academic R&D, of a $108.8 billion total, with a non-S&E share of just six percent. The strongest single-citation foundation for the "other 95 percent" claim.

American Council on Education and Indiana University Center for Postsecondary Research. *2025 Carnegie Classifications: Research Activity Designations*. 2025. — Defines R1 (187 institutions), R2 (139), and the new Research Colleges and Universities tier (216). Provides the operational denominator for any "what fraction of US institutions is research-intensive" claim.

Times Higher Education, QS, and ShanghaiRanking. *World University Rankings, 2025–2026 editions*. — The three major rankings converge on a top-200 cohort: the strict frame for globally research-intensive universities. The IAU/UNESCO World Higher Education Database supplies the global denominator of more than 20,800 institutions.

B.2 The compute and AI-literacy landscape

Stanford Institute for Human-Centered AI. *Artificial Intelligence Index Report 2025*. Stanford, CA: Stanford HAI, April 2025. — The standard annual review: academia leads highly cited AI research while about ninety percent of notable models now originate in industry; the open-versus-proprietary benchmark gap fell to under two percent.

Artificial Analysis. *Artificial Intelligence Index v4.1*. artificialanalysis.ai, June 16, 2026. — The independent composite benchmark referenced throughout; retrieve current scores at the moment of use.

National Science Foundation. *National Artificial Intelligence Research Resource (NAIRR) Pilot*. 2024–2026. — The federal pilot for broadening AI research access: 600+ projects and 6,000 students by 2026, with total compute estimated by Georgetown CSET at roughly 5,000 H100-equivalents—the scale-of-the-problem anchor[12].

B.3 Library, higher-education, and research-computing community

[12]National Artificial Intelligence Research Resource (NAIRR) Pilot, launched January 2024 by NSF and OSTP. By 2026 it had supported more than 600 projects and 6,000 students across all 50 states. Georgetown CSET estimated its initial compute at roughly 3.77 exaFLOPS—about 5,000 NVIDIA H100 GPUs.

Baytas, C., and D. Ruediger. *Making AI Generative for Higher Education*. Ithaka S+R, May 1, 2025 (DOI 10.18665/sr.322677). — Multi-institution study of generative-AI adoption across 19 US and Canadian universities; cited for the institutional-readiness landscape.

Georgieva, M., and J. Stuart. "*Ethics Is the Edge*." EDUCAUSE, 2025. — The annual EDUCAUSE AI Landscape Study; the institutional-policy and AI-literacy-gap context.

Research Computing for Smaller Institutions (RCSI) Conference. "Research Computing at Smaller Institutions, 3rd edition." Swarthmore College, June 16–17, 2026. — The NSF-supported convening from which this article develops; explicitly targets non-R1 institutions.

Uzwyshyn, R., et al., eds. *New Horizons in Artificial Intelligence in Libraries*. IFLA Publications Series. Berlin: Walter De Gruyter, January 2025. — The author's edited volume situating AI in international library practice.

UCR Research Services. *STAR Lab AI Lab Transformation Planning Document (v2)*. University of California, Riverside Libraries, June 2026. — The full operational build specifications for the UCR case study; companion document at rayuzwyshyn.net/UCRiverside2026.

B.4 Local-LLM deployment in academic contexts

"FLEXI: Open-Source Local LLM Deployment in an Academic Setting" (arXiv:2407.13013, 2024) and "LLM Right-sizing: Data Sovereignty and Strategic Autonomy in Academic Computing" (arXiv:2504.13217, 2025). — Applied academic precedents for local hosting on data-protection, sovereignty, and cost grounds.

B.5 Companion presentations by the author

Uzwyshyn, R. "Harnessing AI for Research: Generative AI, Large Language Models, AGI and Reasoning, Multimodal Models, and Deep Research with Autonomous Agents." March 2025. Available via ResearchGate.

Uzwyshyn, R. "Deep Research AI Tools for Research and Discovery." May 2025. Available via ResearchGate.

Uzwyshyn, R. "Vibe Coding for AI-Driven Development: An Introduction for Research and Learning." November 2025. Available via ResearchGate.

Uzwyshyn, R. Vibe-coding and generative-AI workshop demonstrations. UCSF Generative AI Office Hours Colloquium, University of California, San Francisco, March 2026.

Author profile and full publication list: rayuzwyshyn.net — linkedin.com/in/rayuzwyshyn