BOODLEBOX

EXECUTIVE BRIEFING

The Two-Tier Institutional AI Architecture

A cost-responsible alternative for higher education

The bottom line

At a 50,000-FTE research university, an institution-wide commodity foundation-model deployment can model out at roughly $25M to $55M+ per year. A two-tier architecture — a universal collaborative platform for everyone, plus governed agentic tooling for the few who need it — can deliver comparable coverage for about $4.0M per year, frequently 8–15% of the commodity approach, with materially lower governance burden.

This briefing is written for institutional leaders evaluating AI strategy in 2026. It is informed by the enterprise cost events that became public this spring, by emerging higher-education procurement disputes, and by what BoodleBox observes across 1,300+ institutions. Distribution and reproduction encouraged.

For CIOs, Provosts, CFOs & Boards • Published by BoodleBox | 2026

Executive Summary

Agentic AI is not priced like the software universities are used to buying, and that single difference drives everything that follows. Traditional enterprise software is sold by the seat — a fixed, predictable per-user license. Agentic AI is priced by consumption: cost scales with how much work the tool does, not how many people hold a login. That difference breaks the budgeting assumptions institutional procurement is built on, and it is the root of the cost surprises now surfacing across the enterprise market.

Against that backdrop, a recognizable template for institutional AI procurement is taking shape across R1 universities, state systems, and large private institutions. The pattern: enterprise AI RFQs that contemplate institution-wide access to commodity foundation-model platforms — OpenAI, Anthropic, Microsoft, Google — priced primarily on per-million-token consumption, with usage controls and pedagogical governance treated as secondary.

Two things have changed the calculus in 2026. First, the enterprise market has now produced public, named examples of what uncapped token-based AI can cost at scale: budgets exhausted in a third of a year, licenses cancelled mid-deployment, and at least one reported half-billion-dollar monthly bill. Second, higher education has begun its own version of this debate, with multimillion-dollar system contracts drawing faculty petitions and renewal fights even as institutions face budget cuts.

This briefing offers an architectural alternative. It separates the universal need (a collaborative AI environment for teaching, learning, and routine work) from the specialist need (agentic developer tooling for a small population doing code-intensive work), and procures each on terms suited to its risk profile. The result is predictable budgeting, governable usage, and stronger alignment with academic mission. Because this decision spans IT, Academic Affairs, and finance, the sections that follow speak to each in turn.

Three takeaways for the leadership team

- **Token-priced AI does not behave like seats.** Cost scales with how hard the tools are used, not with headcount — the exposure finance and IT models built on per-seat licenses cannot anticipate. (CIOs, CFOs, boards)
- **Buying model access is not an AI strategy.** Native tools are individual productivity apps built to remove friction; teaching and learning depend on it. The pedagogical layer has to be procured as deliberately as the model layer, with Academic Affairs at the table. (Provosts, deans, faculty)
- **Target, don't blanket — and expect two procurements.** Give everyone a governed platform and reserve agentic power tools for those who need them. An IT-led foundation-model layer and an Academic-Affairs-led pedagogical layer will emerge anyway; design them together rather than in sequence. (the whole table)

What the Enterprise Market Just Learned

Over spring 2026, a series of named cases moved the question of agentic-AI cost from cautionary hypothetical to documented pattern, widely discussed across the engineering and CIO communities. Each makes the same point from a different angle: at scale, uncapped token-based access produces bills that arrive faster and larger than the buyer planned for.

These cases center on a single tool — Anthropic's Claude Code — because it scaled fastest and generated the most public reporting. But the dynamic is product-agnostic: the same consumption economics apply to Cursor, OpenAI's Codex, GitHub Copilot's agentic modes, Google's Gemini CLI, and whatever launches next quarter. The lesson is about how agentic tools are priced, not about any one vendor.

Uber: an annual AI budget gone in four months

Uber deployed Claude Code to roughly 5,000 engineers and by April reported having exhausted its full 2026 AI tools budget — about four months into the year. The company's CTO described being “back to the drawing board” on a forecast actual consumption had overrun. Reporting put individual engineer spend between $500 and $2,000 per month, with adoption inside one organization climbing from roughly a third to over four-fifths of engineers in a matter of months. Uber subsequently imposed a $1,500 monthly per-employee cap on agentic coding tools.

Microsoft: cancelling a competitor's tool it had just rolled out

After inviting thousands of staff to use Claude Code beginning in December 2025, Microsoft began winding down most direct Claude Code licenses in its Experiences & Devices division — the group behind Windows, Microsoft 365, Outlook, Teams, and Surface — with access ending by June 30, 2026. Engineers are being steered toward GitHub Copilot CLI. The stated rationale was toolchain unification; reporting indicated the economics of token-based billing at deployment scale were a significant factor.

The half-billion-dollar month

Separately, an AI consultant cited by Axios described a client that failed to set usage limits on its Claude licenses and received a reported monthly bill on the order of $500 million. The figure is unconfirmed and the company unnamed, but it has become shorthand for the core risk: ungoverned, uncapped, token-priced access is a financial exposure pointed back at the buyer.

Why this matters for universities: these were sophisticated, AI-forward enterprises with mature engineering discipline and direct P&L accountability. Most universities differ in ways that increase exposure: far larger user populations, weaker per-user budget accountability, agentic tools that consume far more tokens than chat, and limited real-time consumption visibility during early rollout.

The structural lesson

A traditional software deal is denominated in users; a token-priced deal is denominated in how much work the tool does. When agentic workflows become standard and tasks run in parallel, per-user cost widens in ways quarterly budget models built in late 2025 could not anticipate. Anthropic itself moved Claude Code toward usage-based billing for autonomous agents and, in mid-2026, separated programmatic usage onto metered API-style credits — a clear signal that all-you-can-eat agentic pricing was not sustainable for the provider either.

Field Note: A Claude + Claude Code Deployment, One Year In

BoodleBox recently spoke with a college that implemented Claude alongside Claude Code at the institutional level. Their experience is instructive because it captures, in a single deployment, the three cost traps the

enterprise cases illustrate separately: a steep price increase, paid-for capacity that went unused, and a small group of power users generating overage charges.

What the institution reported

- **A significant year-over-year price increase.** They paid $8 per seat per month last year. This year the per-seat price rose to $11 — a ~38% increase before any usage is considered.
- **Paying for shelfware.** Of 500 licenses purchased, only 299 were ever activated, yet the institution paid for all 500. Roughly 40% of contracted seats were never used but billed in full.
- **Overages on top of the seats.** A small minority of users consumed a large share of the tokens, triggering overage charges beyond what the seats covered. The institution paid twice over.

> *"Claude Code is not scalable at an institutional level."*
>
> — The CIO's verdict on this campus-wide deployment

The deployment fails in both directions at once. The seat model punishes the institution for over-provisioning to the whole campus; the 201 unactivated licenses are pure shelfware. The consumption model simultaneously punishes it for the handful of genuine power users whose agentic workflows blow past the seat allotment. Neither problem is a misuse story; both are predictable outcomes of deploying an agentic developer tool as if it were a campus-wide seat license. The instinct behind a broad rollout is the right one — putting capable tools in people's hands is exactly what a forward-leaning IT leader should want; the seat-license mechanism is simply the wrong way to honor it.

This is precisely the failure mode the two-tier architecture is designed to avoid: it stops provisioning agentic tooling to people who will never activate it, and it puts genuine power users on a governed, sponsored track where overages are expected, budgeted, and owned by the department or grant rather than absorbed centrally.

Higher Education Is Already Feeling This

The cost-and-governance debate is not a future risk for higher education; it is already underway, most visibly around large system contracts and campus budget politics.

The California State University – OpenAI contract

CSU — the nation's largest public four-year system — signed an 18-month, roughly $17M contract to provide ChatGPT Edu to all students, faculty, and staff, and in spring 2026 renewed it at about $13M per year for three years, covering more than 470,000 students and 63,000 employees. It is the largest higher-education partnership OpenAI has. The renewal reignited a fight over priorities. In January, faculty delivered a petition urging cancellation, arguing the product was not designed or optimized for education, and the spend drew scrutiny against a system facing roughly $144M in cuts and a multibillion-dollar structural gap.

The CSU episode is a governance cautionary tale as much as a cost one. Faculty and students reported that the rollout arrived ahead of any shared framework for academic integrity or appropriate use, leaving instructors to improvise — some reverting to blue-book exams, others leaning on unreliable AI detectors. The

lesson is not that system-wide access is wrong, but that capability procured ahead of governance produces friction that lands on the academic side of the house.

Campus budget politics

At Princeton, a student debate played out in print over whether the university should fund ChatGPT Plus subscriptions amid a budget crunch tied to suspended federal grants — one side framing access as an equity issue, the other arguing scarce funds should not flow to AI subscriptions while campus services were being cut. Multiply that tension by an institution-wide, consumption-priced commitment and the budget exposure becomes a recurring governance question rather than a one-time decision.

The pattern across these cases is clear: the headline cost is only the entry fee. What is revealed over time is that the true costs of AI for institutional adoption are both durable and structural. Institutions find themselves paying for retroactive governance, forced to spend months building policy after deployment has already begun. They inherit academic-integrity disputes that, at their worst, force instructors to abandon modern pedagogy in the face of tools they don't understand. And they cannot leave things this way, so institutions must engage in rounds of follow-on procurement to patch over the gaps the first contract ignored.

Two Rollout Approaches Worth Examining

Most institution-wide commodity deployments take one of two forms. Both can work in the right context; both carry distinct institutional risks worth weighing during procurement design.

Approach A — Deployment without usage caps

Estimated annual cost at a 50,000-FTE R1: **$45–55M+**, depending on adoption velocity and tool mix. The enterprise cases above are the reference points for what uncapped, token-priced access can do at scale.

Consideration	Observed likelihood at peer institutions
Budget pressure within first 12 months	**Notable**
Mid-semester need for service adjustment	Possible
Academic-integrity policy questions	Likely to surface
Compliance and data-handling reviews	Likely to surface

Approach B — Deployment with usage caps

Estimated annual cost at a 50,000-FTE R1: **$25–28M**, depending on cap design. Throttling reduces budget exposure but introduces user-experience friction at the limit, ongoing governance overhead (typically 2–3 FTE plus dashboard and chargeback infrastructure), and the practical difficulty of calibrating caps for agentic tools whose intended workflows consume tokens at rates that make conservative caps functionally restrictive. Uber's $1,500 monthly cap is a real-world example of this retrofit.

Consideration	Observed likelihood at peer institutions
Budget pressure within first 12 months	Possible
User-experience friction at limits	**Common**
Academic-integrity policy questions	Likely to surface
Follow-on procurement within 18 months	Frequently observed

The Hybrid Architecture: Universal Platform + Targeted Agentic Access

Institutional AI need is not uniform. The vast majority of faculty, staff, and students benefit from a collaborative AI environment optimized for teaching, learning, and routine work. A much smaller population — researchers, computational disciplines, and select developer roles — benefits from agentic developer tooling built for code-intensive workflows. These are better served by distinct procurement decisions than by a single institution-wide deployment.

What the “pedagogical layer” means

Throughout this briefing, the pedagogical layer refers to the institution's teaching-and-learning use of AI — how faculty design assignments, how students are permitted to use AI, how academic integrity is maintained, and how AI is embedded in courses — as distinct from the raw model access the IT layer provides. It is the difference between giving someone a model and giving an institution a governed environment for teaching with it. This is the layer commodity foundation-model deals routinely leave unaddressed, and it is where BoodleBox has concentrated since giving academia multi-model access well before agentic tooling existed.

Tier 1 — Universal collaborative AI platform (all users)

A purpose-built institutional platform deployed to faculty, staff, and students. Typical capabilities:

- **Multi-model access** — frontier models from multiple providers in one environment
- **Knowledge Banks** for course materials, syllabi, and institutional content
- **Group collaboration** for AI-mediated faculty and student workflows
- **Assessable chat history** that supports academic-integrity policy
- **Native governance,** FERPA / HIPAA / SOC 2 compliance, and audit trails

This layer covers the substantial majority of institutional AI demand by user count, typically around 95%. Estimated annual cost at a 50,000-FTE institution: **~$2.5M.**

Tier 2 — Targeted agentic developer tooling (initially ~3% of users)

Application-based, departmentally sponsored, instructor-gated access to agentic coding tools (Claude Code, Codex, Copilot, and peers) for the population whose work substantively requires CLI-based, multi-file, Git-integrated agentic workflows: CS and engineering faculty, research-computing staff, graduate students in computational disciplines, and instructor-approved capstone undergraduates.

A common control: API consumption capped at a board-approved ceiling, with overages funded through departmental budget transfer or grant funding rather than central-IT absorption. This is the structural answer to the field-note college's problem. Power users belong on a sponsored, owned track, not a campus-wide seat pool. That 3% is a bottom-up estimate, not a guess: it reflects the slice of the institution whose work genuinely requires CLI-based, multi-file, Git-integrated agentic workflows — CS and engineering faculty and their computational graduate students, research-computing and developer staff, and a subset of capstone undergraduates. Computer science and engineering together account for only about 14% of four-year undergraduate enrollment nationally, and only a fraction of even those students do the code-intensive

agentic work Tier 2 is built for; set against a total headcount that also includes every other student, faculty, and staff member, the genuine power-user share nets down to the low single digits — consistent with the field-note college, where a small minority of users drove most token consumption. It is a floor, not a ceiling: the sponsored track is designed to grow as demand grows, so broadening access becomes a budgeted decision rather than an uncapped surprise, and each institution should calibrate the figure against its own program mix.

Estimated annual cost at a 50,000-FTE institution: **~$1.5M.**

> **Combined architecture: approximately $4.0M / year**

How the architecture controls cost

The savings are not a negotiated discount; they are structural. Three mechanisms do the work:

- **Right-sized model routing.** Rather than sending every query to the most expensive frontier model — the default that maximizes a vendor's token revenue — the platform routes each task to the least-expensive model capable of handling it, reserving premium models for the work that genuinely needs them.
- **One governed environment instead of raw API access.** Multi-model access, knowledge banks, and collaboration sit inside a managed platform rather than billing straight against per-token API consumption. Under typical institutional usage this runs roughly 96% below direct foundation-model API cost.
- **Targeting instead of blanket provisioning.** Agentic tooling goes only to the ~3% who need it, on a sponsored track — so the institution stops paying for capacity the other 97% will never activate; as that population grows, the sponsored track grows with it.

Why targeting also improves adoption

The targeting in Tier 2 is not only a cost control; it tracks how technology adoption actually works. Research on technology acceptance consistently finds that voluntary, need-driven (“pull”) adoption sustains engagement, while mandated, top-down (“push”) rollouts tend to produce resistance and disuse — the same dynamic visible in the field note's 201 unactivated licenses. Giving broad, low-friction access to the universal platform while letting genuine power users opt into agentic tooling aligns provisioning with demand, rather than pushing expensive capability onto people who did not ask for it.

A Cost Comparison Worth Modeling

Architecture	Annual cost (50K-FTE R1)	Coverage	Governance burden
Commodity foundation-model, no caps	$45–55M+	Full institution	Significant
Commodity foundation-model, with caps	$25–28M	Full institution	Moderate; institution-built
Two-Tier Hybrid Architecture	**~$4.0M**	**Full institution + power users**	**Largely native to platform**

Under these assumptions, the hybrid architecture may deliver comparable coverage at a meaningfully lower cost, frequently 8–15% of the commodity foundation-model approach. Institutional context will vary; each institution should model these figures against its own enrollment, adoption assumptions, and existing infrastructure.

The hybrid architecture also tends to align with common award criteria (lowest-priced, responsive, and responsible supplier), particularly where institutions weigh sustainability and governance alongside upfront capability:

- **Cost:** substantially lower under most modeling assumptions.
- **Responsiveness:** aligned with needs across teaching, learning, research, and administration.
- **Responsibility:** designed for predictable budget, governable usage, and mission alignment. Lower token consumption also means a smaller computational footprint — a secondary but real benefit as sustainability enters more procurement criteria.

Governance in Practice: A Hypothetical Case Study

Governance is the most institution-specific part of this architecture. No two campuses divide authority over information technology and Academic Affairs the same way, so a single prescribed org chart is as likely to clash with an institution's conventions as to fit them. Rather than propose one structure and try to account for every variation, the scenario below shows how one composite institution might approach the problem — a familiar reference point to adapt, not a template to adopt.

The scenario

Riverside State is a mid-sized public university that has decided to deploy the two-tier architecture. Leadership agrees on one point immediately: because the decision touches teaching, research, budget, and compliance at once, no single office should own it alone. They want a standing body — an AI Leadership Council — to define access criteria, recommend the budget split between tiers, and maintain institutional AI policy.

The first instinct is to have the council report jointly to the CIO and the Provost, reflecting how genuinely shared the work is. The cabinet quickly sees the problem: a body reporting to two cabinet-level executives has no clear line of accountability. When the two offices disagree — over a Tier 2 access denial, say, or who absorbs an overage — dual reporting offers no way to break the tie, and the council stalls exactly when a decision is needed. Split reporting tends to invite turf friction rather than resolve it.

Riverside settles it the way many institutions do, by giving the council a single reporting line. Because it has chosen to treat AI as pedagogical infrastructure rather than an IT purchase, the Provost chairs and the council reports through the Provost to the President's cabinet; the CIO serves as vice-chair, giving IT genuine co-ownership of the agenda without a second reporting line. The council is advisory: recommendations that carry budget or policy weight go to the full cabinet for decision. This is not the only workable arrangement — an institution where AI is led out of IT might reasonably invert the chairs — but the principle holds either way: one accountable line, with shared ownership built in, avoids the paralysis that dual reporting invites.

With that settled, the rest falls into place. Riverside's council looks like this:

Composition

- Provost or designee (chair)
- Chief Information Officer (vice-chair)
- Center for Teaching & Learning director
- Faculty Senate AI Committee chair
- Graduate School representative
- Student Government representative
- Vice President for Research
- General Counsel (ex officio)

Charter

- Define Tier 2 access criteria and review applications
- Maintain institutional AI policy (academic integrity, research use, data handling)

- Recommend the annual budget allocation between tiers
- Monitor usage patterns and recommend adjustments
- Liaise with peer institutions on emerging practices

The specific seats and duties matter less than the two choices underneath them: that AI is treated as pedagogical and institutional infrastructure rather than an IT procurement, and that authority runs along a single accountable line even as ownership is genuinely shared. A body built on those two principles holds up better than any fixed org chart as the technology — and the institution's use of it — evolves.

Considerations by Role

Role	Consideration
CIOs	Institution-wide commodity deployment may carry cost and governance exposure worth modeling carefully before commitment. A collaborative platform deployed broadly, with agentic tooling reserved for the population that needs it, is a pattern peer institutions have found workable. Crucially, this targets cost, not capability: your power users still get the frontier agentic tools they want — on a sponsored, governed track rather than a campus-wide pool.
Provosts / Deans / CTL directors	This is an academic decision as much as an IT one, and Academic Affairs should co-own it. Establish AI governance before deployment — not after — and make sure Academic Affairs has a seat in any enterprise AI procurement. A consumption-based foundation-model contract answers the infrastructure question but leaves teaching, learning, and academic integrity unaddressed; frame AI as pedagogical infrastructure and procure the pedagogical layer as deliberately as the model layer.
CFOs / Boards	Per-token pricing introduces commodity exposure worth modeling against institutional risk tolerance. Hard ceilings, orchestration layers, and demonstrable cost predictability are common features of more durable architectures.
Faculty Senates	Governance participation prior to deployment, plus academic-integrity and student-use frameworks developed in parallel with rollout, support smoother adoption.
Procurement officers	Distinguish foundation-model access from institutional AI platforms during specification design. They are different categories of product; treating them as one produces specs that are hard for non-hyperscaler vendors to answer.

For Provosts and Deans: Keeping Learning at the Center

Enterprise AI is usually pitched as an IT or infrastructure purchase, and the pressure to buy often arrives from two directions at once: a consumption-based, native foundation-model deal presented as the obvious next step, and a group of enthusiastic faculty asking for the most powerful frontier tools. Both pressures are legitimate, and both come from people doing their jobs well. Neither, on its own, answers the question that sits with academic leadership — what does this do to teaching, learning, and the student experience?

A consumption-based foundation-model contract answers the access question, not the academic one

Native foundation-model access — ChatGPT Edu, Claude, Copilot, Gemini — is powerful and genuinely useful. But it is pedagogically unopinionated by design: it provides raw capability, not the scaffolding teaching and learning require. Out of the box it typically lacks assessable interaction history, course- and curriculum-level integration, guardrails for appropriate student use, and the academic-integrity supports faculty depend on. It optimizes for what a model can do, not for what a course is trying to achieve. Procuring it and calling it an AI strategy leaves the pedagogical layer — the part that actually touches the mission — unaddressed.

Native AI is built to remove friction. Education runs on it.

Native AI apps are, at their core, individual productivity tools, and their entire design goal is to make work as frictionless as possible — fewer steps, faster output, less effort. That is exactly what you want when drafting an email or summarizing a report. But education is not individual productivity work. It is a collaborative activity, and much of its value comes from productive struggle — the effortful thinking, revision, and dialogue through which understanding is actually built. In a learning context, friction is not a defect to engineer away; it is frequently the point. A tool optimized to eliminate effort, dropped into a course without pedagogical design, can quietly remove the very struggle the course exists to create — yielding more polished output and less learning. This is the deepest reason the pedagogical layer matters: its job is to put the right friction back in — supporting productive struggle, collaboration, and assessable process — rather than optimizing all of it away.

The fallout lands on the academic side of the house

When capability arrives ahead of governance, the cost is paid in the classroom, not the data center. CSU is the cautionary case: faculty and students reported that the tools arrived before any shared framework for academic integrity or appropriate use, leaving instructors to improvise — some reverting to blue-book exams, others leaning on unreliable AI detectors. A Provost or CIO who signs a capability contract without an accompanying academic framework is, in effect, outsourcing course-level policy to whichever vendor won the procurement.

Where the pressure comes from — and what it leaves out

- **From the market:** a consumption-based native deal is operationally clean and straightforward to execute, and the instinct is sound for the foundation-model layer. Many CIOs are themselves wary of its open-ended pricing — the same concern this briefing raises. But on its own it is not a teaching-and-learning plan, and its usage-based pricing can expose academic budgets to costs no one can forecast.
- **From faculty:** some of your most capable faculty and researchers will — rightly — want the best frontier and agentic tools. That need is real, and it is a Tier 2 need. Provisioning the entire institution to satisfy a relatively small group of power users is precisely the failure the field-note college lived through.

The two-tier architecture is what lets you say yes to both without buying an institution-wide consumption contract: power users get governed, sponsored access to the advanced tools they are asking for, while everyone else gets a pedagogically designed platform built for teaching and learning. This is not Academic Affairs versus IT — it is the two working from one design, with IT owning the foundation-model layer and Academic Affairs owning the pedagogical layer, exactly the two-procurement pattern described next.

Six questions to ask before the institution signs

- Does this support academic integrity — assessable history, coaching toward appropriate student use — or does it just produce raw output?
- Who owns academic-use policy once it is deployed: Academic Affairs, or the vendor's defaults?
- Is there a governance body with faculty and Academic Affairs at the table, established before rollout rather than after?
- Does the pricing expose academic budgets to unpredictable, usage-based cost that competes with instructional priorities?
- Does this serve the ~95% whose need is teaching and learning, or chiefly the ~3% who are power users?
- What happens to equity if access — or cost — falls unevenly across departments and students?

The thoughtful move for CIOs, Provosts, and Deans

Insist that the pedagogical layer be procured as deliberately as the foundation-model layer, and that Academic Affairs and IT co-own both procurement decisions. The goal is not to slow AI adoption; it is to ensure the institution adopts AI in a form that serves learning, protects integrity, distributes access equitably, and keeps cost predictable enough that it does not crowd out the academic mission it is meant to support.

The Two-Procurement Pattern

What BoodleBox observes consistently across the 1,300+ institutions in our footprint is that two procurements tend to emerge once a campus begins deploying AI seriously:

- **The foundation-model layer,** typically led by IT/CIO with Academic leader input, focused on enterprise access to LLMs, agentic tooling, and API consumption.
- **The pedagogical and collaborative layer,** typically led by the Provost's office, a Center for Teaching & Learning, or Academic Affairs, with input from IT/CIO, focused on how the institution teaches, learns, and governs AI use at scale.

Institutions that recognize this duality early can design coherent procurements the first time, rather than addressing the pedagogical and governance layer in a follow-on cycle. The hybrid architecture in this briefing is one operational expression of that recognition.

A Choice Available Now

Institutions weighing institution-wide AI deployment have meaningful options. One path emphasizes commodity foundation-model access at scale; another emphasizes a layered architecture that separates universal collaborative use from targeted agentic capability. Both can work in the right context. Our observation, reinforced by the enterprise cost events of 2026 and by early higher-education procurement disputes, is that the layered approach tends to produce more predictable budgets, smoother governance, and stronger alignment with academic mission.

The math, the peer-institution experience, and the structural realities of higher-education adoption all point toward this being a choice worth modeling carefully now, before institution-wide deployment decisions become difficult to revisit.

About BoodleBox

BoodleBox is a collaborative AI platform purpose-built for higher education and AI-native workteams, serving 100,000+ learners across 1,300+ institutions. The platform provides:

- Multi-model access to 38 frontier AI models in a single environment
- Custom AI bots tied to specific courses, departments, or workflows
- Knowledge Banks for institutional content and course materials
- Assessable chat history supporting academic integrity, plus Coach Mode for pedagogically aligned student–AI interaction
- Group collaboration enabling multi-human, multi-AI workteams
- FERPA, HIPAA, GDPR, SOC 2 Type II, HECVAT Full, and VPAT AA compliance; no training on institutional data
- Approximately 96% lower token cost — and a correspondingly lower computational footprint — than direct foundation-model API consumption under typical institutional usage
- Soon available as a Microsoft Transactable Partner on Azure Marketplace

For a conversation about how peer institutions are deploying the two-tier architecture or to have us model the two-tier cost against your own enrollment and adoption assumptions, please **schedule a demo with our team**.

This executive briefing is offered as market reference material for higher-education leaders evaluating institutional AI strategy. Distribution and reproduction encouraged.

References

Figures are as reported in public sources from spring 2026; readers are encouraged to verify against primary sources before relying on them for procurement decisions. The field-note college is anonymized at its request.

Angelo, J. (2026, May 26). Uber's COO says it's getting harder to justify the company's AI spend. Fortune. https://fortune.com/2026/05/26/uber-coo-ai-spending-tokens-claude-code/

Bay City News Service. (2026, May 25). California State University renews controversial systemwide contract with OpenAI. SFGate. https://www.sfgate.com/news/bayarea/article/edsource-california-state-university-renews-22276261.php

Corzo, A. (2026, May 1). Cal State struck a deal with OpenAI. Some students and faculty refuse to use it. CalMatters. https://calmatters.org/education/2026/05/california-state-university-open-ai-chatgpt-contract/

Ghoshal, A. (2026, May 14). Anthropic puts Claude agents on a meter across its subscriptions. InfoWorld. https://www.infoworld.com/article/4171274/anthropic-puts-claude-agents-on-a-meter-across-its-subscriptions.html

Langford, D. S. (2026, April 2). ANU–Anthropic Claude AI partnership. AcademicJobs.com. https://www.academicjobs.com/au/higher-education-news/anu-anthropic-claude-ai-partnership-or-australian-uni-innovation-11366

Mills, M. (2026, May 28). AI sticker shock hits corporate America. Axios. https://www.axios.com/2026/05/28/ai-spending-roi-enterprise-costs

National Student Clearinghouse Research Center. (2026, January). Clearinghouse final fall enrollment trends (Fall 2025). https://nscresearchcenter.org/final-fall-enrollment-trends/

Palmer, K. (2026, March 27). Faculty push back against OpenAI deals. Inside Higher Ed. https://www.insidehighered.com/news/tech-innovation/artificial-intelligence/2026/03/27/faculty-push-back-against-openai-deals

Ropek, L. (2026, June 2). Uber caps employee AI spending after blowing through budget in four months. TechCrunch. https://techcrunch.com/2026/06/02/uber-caps-employee-ai-spending-after-blowing-through-budget-in-four-months/

Stanciuc, A.-M. (2026, May 25). Microsoft's quiet Claude Code retreat and the real cost of enterprise AI. The Next Web. https://thenextweb.com/news/microsoft-claude-code-retreat-ai-cost

Styris, G. (2025, October 23). In a moment of budgetary crisis, don't waste funds on ChatGPT Plus. The Daily Princetonian. https://www.dailyprincetonian.com/article/2025/10/princeton-opinion-opguest-response-chatgpt-plus-equity-ai-models-research-funding-student-employment-budget-cuts

University World News. (2026, March 28). Faculty in US universities push back against OpenAI deals. https://www.universityworldnews.com/post.php?story=20260328123358633

Warren, T. (2026, May 14). Microsoft starts canceling Claude Code licenses. The Verge. https://www.theverge.com/tech/930447/microsoft-claude-code-discontinued-notepad

Zeff, M. (2025, July 28). Anthropic unveils new rate limits to curb Claude Code power users. TechCrunch. https://techcrunch.com/2025/07/28/anthropic-unveils-new-rate-limits-to-curb-claude-code-power-users/