

In-Class Exercise — LLM Adoption

(with Complete Answers)

Scenario

Teams of 4–5 role-play as CIOs for three firm types. Each team completes the Adoption Decision Table, selects a quadrant, recommends models, and defines governance + a 90-day pilot.

Adoption Decision Table

Each team completes one row per use case. The "Quadrant" column forces an explicit 2×2 placement before any model is named.

#	Field	Team A (Small)	Team B (Mid)	Team C (Large)
1	Company / size / sector			
2	Use case & business value			
3	Quadrant (1/2/3/4)			
4	Open or closed source?			
5	Model(s) & Adoption Position			
6	Regulation			
7	Volume (tokens/month)			
8	Talent			
9	5-Year TCO			
10	Top Risk & Mitigation			
11	90-Day Success Metric			
12	(A) Trap Avoided (B) Routing Justification (C) Governance Layers			

Routing & Governance Table

Traffic tier	Task examples	Model class	Cost/request	Human-in-loop?
Light (~80%)	Classification, extraction, summaries	Flash / small open-weight	~\$0.002	No
Mid (~15%)	Drafting, standard RAG Q&A	Max / mid-tier	~\$0.01	Review sample
Complex (~5%)	Disputes, compliance explanations	Flagship (Claude/GPT)	~\$0.10	Yes
Frontier (<1%)	Specialized engineering, research	GPT-6 Astra / Gemini Pro	~\$0.50	Yes + approval

Regulations:

- **API** := Application Programming Interface
- **BAA** := Business Associate Agreement
- **DPA** := Data Processing Agreement
- **GLBA** := Gramm-Leach-Bliley Act, the U.S. federal law requiring financial institutions to protect consumers' nonpublic personal information (NPI) and explain their data-sharing practices.
- **HIPAA** := Health Insurance Portability and Accountability Act (1996)
- **PCI-DSS** := Payment Card Industry Data Security Standard
- **PHI** := Protected Health Information — individually identifiable health information that is created, received, maintained, or transmitted by a HIPAA covered entity or business associate.
- **PII** := Personal Identification Information
- **VPC** := Virtual Private Cloud

Team A — Small Firm (12 people, no AI team, low volume)

Chosen Sector: *Healthcare & Pharma (Small Clinic Group)*

#	Field	Recommendation
1	Company / Size / Sector	12-person private dermatology clinic group; Healthcare; PHI under HIPAA.
2	Use Case & Business Value	Internal knowledge assistant for clinical protocols & billing FAQs. Cuts search time 60%; reduces admin overhead.
3	Quadrant	Q1 — Plug-and-Play API (closed, external). No ML talent; low volume; fastest ROI.
4	Open/Closed & Hosted/API	Closed model; external API with zero-retention DPA and BAA.
5	Model(s) & Adoption Position	Claude Haiku 4.5 (HIPAA-compliant via AWS Bedrock). Position 1 (Turnkey SaaS).
6	Regulation	HIPAA; requires BAA, zero-retention, PII masking. No data leaves US perimeter.
7	Volume	~ 5,000 queries/month (low). API cost ~ \$50/mo.
8	Talent	0 ML engineers; admin staff manages via Bedrock console.
9	5-Year TCO	~ \$45K (API + admin time + compliance reviews).
10	Top Risk & Mitigation	<i>Data leakage via prompts.</i> Mitigation: zero-retention endpoint, PII auto-redaction, no patient names in queries.
11	90-Day Success Metric	80% of staff use weekly; 70% "found what I needed" rate; zero PHI in prompt logs.
12	Trap Avoided	Buying a GPU box for "privacy" with no pipelines or talent.

Team B — Mid-Market (350 people, 5-person technical team, millions of tokens/month, PII)

Chosen Sector: *Financial Services (Regional Credit Union)*

#	Field	Recommendation
1	Company / Size / Sector	350-employee regional credit union; Financial Services; PII, GLBA, PCI-DSS.
2	Use Case & Business Value	Hybrid: (1) Customer support triage (public Q1), (2) Internal loan-doc RAG (private Q2/Q4). Reduces loan processing time 40%.
3	Quadrant	Q2 → Q4 migration. Start Q2 (Together AI / Replicate hosting Llama 4), migrate high-volume RAG to Q4 self-hosted.
4	Open/Closed & Hosted/API	Mix: Frontier API (Claude Sonnet 5) for complex disputes; self-hosted open-weight (Llama 4 Maverick) for volume RAG.
5	Model(s) & Adoption Position	Claude Sonnet 5 (complex, <5% traffic) + Llama 4 Maverick (high-volume, ~ 80% traffic) + DeepSeek V4 Flash (commodity classification, ~ 15%). Position 3 (API + Orchestration).
6	Regulation	GLBA, PCI-DSS; PII must stay in VPC; audit trails mandatory.
7	Volume	~ 8M tokens/month. High stable volume → self-host GPU cluster cost-effective for RAG.
8	Talent	5-person technical team (backend engineers, no dedicated MLOps). Use vLLM + managed GPU.
9	5-Year TCO	~ 680K (API~45K/yr + GPU lease ~8K/mo+engtime ~120K/yr + audit/compliance).
10	Top Risk & Mitigation	<i>RAG ACL leakage across permission groups.</i> Mitigation: role-based access control, document-level permissions, eval harness, circuit breakers.
11	90-Day Success Metric	Pilot on 50k loan docs; 95% retrieval accuracy; \$0.01/query avg cost; zero PII leakage in tests.
12	Routing Justification	<5% complex → frontier API; ~ 80% extraction/classification → open-weight self-host; ~ 15% commodity → Flash API.

Team C — Large Regulated (10,000+ staff, strict compliance, global)

Chosen Sector: *Public Sector (Federal Government Agency)*

#	Field	Recommendation
1	Company / Size / Sector	12,000-employee federal agency; Public Sector; classified-adjacent data, FedRAMP, ITAR, GDPR.
2	Use Case & Business Value	Multi-layer: (1) Public chatbot (Q1, gov cloud), (2) Internal policy RAG (Q3/Q4), (3) Document processing (Q4). Improves citizen response time 50%.
3	Quadrant	Q3 + Q4. Private cloud for sensitive; full-stack autonomy for core operations.
4	Open/Closed & Hosted/API	Closed frontier via FedRAMP High private cloud (Q3) for public tier; self-hosted open-weight (Q4) for internal.
5	Model(s) & Adoption Position	GPT-6 Sol (FedRAMP, public-facing, <1% traffic) + Qwen3-27B (self-hosted, MIT, internal RAG, ~ 85%) + Phi-4 (edge/offline, ~ 14%). Position 5 (Full Private Stack).
6	Regulation	FedRAMP High, ITAR, GDPR, EU AI Act (transparency). No PRC-cloud API; self-host only on US/EU VPC.
7	Volume	~ 50M+ tokens/month. Self-host clearly wins on TCO at this scale.
8	Talent	25-person platform team (MLOps, security, SRE).
9	5-Year TCO	~ 4.2M(GPU cluster ~ 80K/mo + platform team 3.5M + audit/compliance 3.5M+audit/compliance 400K + API reserve \$100K).
10	Top Risk & Mitigation	<i>Geopolitical risk of Chinese open-weight (Qwen).</i> Mitigation: self-host only on air-gapped US GovCloud; weight provenance audit; no PRC-cloud egress; board-level geo-risk review.
11	90-Day Success Metric	Pilot on 100k policy docs; 99% uptime; 100% audit trail coverage; zero data egress outside perimeter.
12	Governance Layers	Model gateway → eval harness → human-in-loop (disputes) → rollback policy → FinOps dashboard.

Exercise Takeaway

- State the quadrant before the model — architecture precedes vendor.
- Route by task difficulty, not price alone.
- Quantify 5-year TCO, not 1-month API sticker price.
- Build a model-agnostic gateway so you swap models, not architecture, when the leaderboard flips.

Debrief: What Good Recommendations Have in Common

- They state the quadrant before the model — architecture precedes vendor.
- They match deployment mode to data class, not to price alone.
- They route by task difficulty rather than sending every request to the flagship.
- They quantify TCO over 5 years, not API sticker price over 1 month.
- They identify the digital-divide constraint — the capital threshold to move from Q1/Q2 to Q3/Q4 — and propose how the firm will cross it (external funding, managed open endpoints, or cooperative infrastructure).
- They plan the pilot as a business experiment with a baseline and a go/no-go, not as a technology demo.

Final thought for students: do not bet on one model. Build a model-agnostic gateway with an eval harness and intelligent routing. When the leaderboard flips next quarter—and it will—you swap a model, not your architecture.