

Study Aid: Enterprise LLM Adoption

Part 1: Core Concepts & Brief Definitions

Concept	Definition
LLM (Large Language Model)	A pattern-matching prediction engine trained on massive text corpora; probabilistic text generator, not a fact database.
Weights	Weights are learned numerical parameters inside an LLM that determine how it responds to any input.
Closed Model	Proprietary weights; API-only access (e.g., GPT-5.x, Claude, Gemini Pro). No self-hosting.
Open-Weight Model	Downloadable weights; usually no training code/data disclosed. Self-hostable (e.g., Llama 4, DeepSeek V4, Qwen3).
Open-Source AI (OSI/OSAID)	Weights + training code + data published; fully reproducible (e.g., OLMo 2). Almost no frontier model qualifies.
RAG (Retrieval-Augmented Generation)	Retrieves docs from a private knowledge base and injects them into the prompt. Gives LLMs secure access to enterprise data without retraining.
Agentic Workflow	LLM plans steps, calls tools (search, APIs, databases), and iterates autonomously. Highest potential and highest risk.
Fine-Tuning	Retraining a model on enterprise-specific data for a narrow task. Higher upfront cost; ideal for brand voice or legal classifiers.
TCO (Total Cost of Ownership)	Full cost of AI adoption: integration (25–35%), data/retrieval infra (15–25%), governance/audit (10–20%), eval/tuning (5–15%), inference/ops (15–30%). Token cost is only 10–20%.
FinOps for AI	Real-time token metering, per-workflow budgets, circuit breakers, and model routing to control AI spend.
Model-First Pitfall	Spending most effort choosing a model while neglecting data plumbing, context, and governance (~ 80% of implementation time).
Shadow AI	Unauthorized LLM usage (unknown APIs, unapproved agents) creating security, cost, and compliance risks.

Digital Divide (AI)	Resource-rich firms self-host (Q3/Q4 = data ownership, moats); resource-poor rent intelligence (Q1/Q2 = vendor lock-in).
2x2 Adoption Matrix	X-axis: Open vs. Closed model. Y-axis: Self-Hosted vs. External API. Quadrants: Q1 (Plug-and-Play API), Q2 (Open-Source-as-a-Service), Q3 (Enterprise Private Cloud), Q4 (Full-Stack Autonomy).
5-Filter Decision Chain	Data & Regulation → Capability Level → Volume & TCO → Org Capability → Strategic Control.
Hybrid Routing	Sending hard reasoning to frontier APIs; high-volume execution to open-weight self-host; edge/offline to small models.
Hallucination	Model generates plausible but false statements. Mitigated via RAG citations, human review, data grounding.
Context Window	Max text held in memory per request (e.g., 128K tokens ≈ 300-page book). Enables long-document workflows.
Prompting vs. Fine-Tuning	Prompting = query-time instructions (zero extra cost); Fine-tuning = retraining on proprietary data.
EU AI Act	Enforced from Aug 2026; requires transparency, chatbot labeling, deepfake marking, model inventory, human-review assignment.

Part 2: Key Questions & Brief Answers

#	Question	Answer
1	What % of organizations use AI in at least one function?	88% (McKinsey).
2	What % of genAI pilots show no measurable P&L impact?	~ 95% (MIT Media Lab) — the pilot-to-production wall.
4	What is the #1 executive AI problem in 2026?	The ROI gap — individual productivity gains don't automatically reach corporate P&L.
5	What drives Chinese open-weight adoption in enterprises?	TCO + data control (not "free weights"): avoids token margins, keeps PHI/financial data in VPC, protects against vendor price hikes.
6	What is the average number of models per enterprise account?	~ 4.7 models/account (up from 2.1).

	(Q1 2026)?	
7	What share of enterprise AI use cases are bought (not built)?	~ 76% (SaaS, gateway, or API).
8	What is the rule of thumb for API vs. self-hosting cost crossover?	Below ~ 100K queries/day , APIs are typically more cost-effective. High stable volume favors self-hosting.
9	Which vendor leads enterprise LLM spend in 2026 surveys?	Anthropic Claude (~ 32–40% share), ahead of OpenAI in enterprise due to audit trails, data isolation, long-context pricing.
10	What are the four quadrants of the LLM Adoption Matrix?	Q1 Plug-and-Play API (closed, external); Q2 Open-Source-as-a-Service (open-weight, external); Q3 Enterprise Private Cloud (closed, self-hosted); Q4 Full-Stack Autonomy (open-weight, self-hosted).
11	What is the recommended first step in the 12-month roadmap?	Educate (Months 1–2): literacy for leadership, identify 3–5 use cases, distinguish viable ideas.
12	What is the most common enterprise architecture pattern?	Context Injection (RAG) — retrieves from knowledge base, injects into prompt.
13	What are the six core LLM capability families (2026)?	C1 Language Understanding, C2 Knowledge Synthesis, C3 Content Generation, C4 Reasoning & Planning, C5 Tool Use/Agents, C6 Multimodal Perception.
14	What is "vibe-coding" in the education context?	Shortcut-style AI use that simulates the form of inquiry while hollowing out learning substance — to be avoided.
15	What is the strategic frame for 2026–2027?	Adoption Gap → Scaling Gap → ROI Gap → Governance Gap. LLMs treated as infrastructure , not magic.