SAMPLE

Building the Generative AI of Tomorrow

Building the Technology That Creates

Aiman JoeHanz

Free sample: Introduction & Chapter 1

putrabooks.github.io

Buy the full book:

Carousell carousell.sg/p/building-the-generative-ai-of-tomorrow-ebook-pdf-1465658753

Gumroad

Ko-fi

ElevenReader

Payhip

Contents

Introduction

Chapter 1: The Current State of Generative AI

Chapter 2: Limitations and Failure Modes

Chapter 3: Common Mistakes in Generative AI Adoption

Chapter 4: Foundational Principles of Generative AI

Chapter 5: Architecture and Design of Generative Models

Chapter 6: Implementation Strategies for Production Systems

Chapter 7: Technology Deep Dive: Transformers, Diffusion, and Beyond

Chapter 8: Case Studies: Real-World Deployments

Chapter 9: Scaling and Future-Proofing Generative Systems

Chapter 10: The Future of Generative AI

Chapter 11: Action Plan for Organizations

Chapter 12: Advanced Topics and Emerging Patterns

Chapter 13: Trust, Ethics, and the Road Ahead

Frequently Asked Questions

Glossary

This free sample contains the Introduction and Chapter 1. The complete book contains the full contents above.

Introduction

Generative AI is the most significant technological shift since the internet. It transforms how we create text, images, code, and audio — not by retrieving existing content, but by generating new content from learned patterns. The implications extend far beyond chatbots and image generators. Generative AI is reshaping software development, scientific research, creative industries, education, healthcare, and virtually every domain where knowledge work happens.

This book is for ML engineers, software architects, product managers, and technology leaders who need to build production generative AI systems. It covers the full stack: from mathematical foundations and model architecture to deployment, evaluation, cost optimization, and strategic planning. By the end, you will have a practical framework for building, deploying, and scaling generative AI in your organization.

This book assumes you have a working knowledge of software engineering and a basic understanding of machine learning concepts. You do not need a PhD in mathematics, but you should be comfortable with linear algebra, probability, and neural network fundamentals. If you have never trained a model, Chapter 4 provides the foundational principles. If you have never deployed a model to production, Chapter 6 walks through the implementation strategies. If you are already building generative AI systems and need to scale them, Chapters 7 through 9 address the challenges you are facing today.

The book serves four primary audiences. ML engineers will find detailed coverage of model architectures, training techniques, and inference optimization. Software architects will find patterns for system design, model abstraction, and infrastructure planning. Product managers will find frameworks for evaluating use cases, managing cost-quality trade-offs, and planning roadmaps. Technology leaders will find strategic guidance on build-vs-buy decisions, team building, and organizational adoption.

The book follows a progression from problem to solution to strategy. Part I (Chapters 1-3) examines the current state of generative AI, its limitations, and the common mistakes organizations make when adopting it. Part II (Chapters 4-6) covers the technical foundations — principles, architecture, and implementation. Part III (Chapters 7-9) dives deep into specific technologies, real-world deployments, and scaling challenges. Part IV (Chapters 10-13) looks to the future, providing an action plan, advanced topics, and a framework for trust and ethics.

Each chapter follows a consistent structure: a practical insight at the start,

detailed technical content in the body, and a set of key takeaways at the end. The practical insights are drawn from production deployments and represent hard-won lessons. The key takeaways distill each chapter into its essential points for quick reference.

This book does not teach you to train a frontier model from scratch. Training a model like GPT-4 or Claude requires resources, data, and expertise that few organizations possess. Instead, this book focuses on what most organizations actually do: fine-tuning, adapting, and deploying existing models for production use cases.

This book does not advocate for any specific model, framework, or vendor. The field moves too quickly for any specific recommendation to remain current. Instead, the book provides frameworks for evaluating and selecting the right tools for your specific needs. Vendor names are mentioned for context, not as endorsements.

This book does not cover the mathematics of generative models in exhaustive detail. The mathematical foundations are explained at the level needed to make informed engineering decisions, not at the level needed to publish research papers. For readers who want deeper mathematical treatment, references to the original papers are provided throughout.

This book was written in 2026, a pivotal year for generative AI. The technology, market, and regulatory landscape are all in flux. Some specific details — model names, prices, capabilities — will be outdated within months. The principles, patterns, and frameworks in this book are designed to be durable. They will outlast the specific technologies they describe and remain applicable as the field evolves.

The goal is not to make you an expert in any single model or tool. The goal is to make you capable of building generative AI systems that work in production, scale to meet demand, and adapt to the changes that are coming. The field will look different in 2027, 2028, and 2030. The practices in this book will help you navigate whatever it becomes.

Read this book in the order that matches your current needs. If you are new to generative AI, start at the beginning and read straight through — each chapter builds on the previous. If you are already building generative AI systems and need specific guidance, jump to the chapter that addresses your current challenge.

Engineers implementing their first production system should focus on Chapters 4 through 7. Architects designing systems for scale should focus on Chapters

6, 9, and 10. Product managers evaluating use cases should focus on Chapters 1, 3, and 11. Technology leaders making strategic decisions should focus on Chapters 1, 10, and 13.

The Action Plan in Chapter 11 is designed to be used as a standalone reference. It provides a phased approach to generative AI adoption that can be followed independently of the rest of the book. If you read only one chapter, make it that one.

Throughout this book, certain conventions are used. Practical Insights appear in blockquotes at the start of major sections — these are field-tested lessons that can be applied immediately. Key Takeaways appear at the end of each chapter — these are the essential points that should be remembered. Technical terms are defined on first use and collected in the Glossary at the end of the book.

Code examples are kept minimal. The focus is on patterns and principles, not on specific implementations. Frameworks change; principles endure. When a specific framework or tool is mentioned, it is for illustration, not prescription. The right tool for your system depends on your specific requirements, constraints, and context.

The book uses the term "generative AI" broadly to include all models that generate new content — text, images, audio, video, code, and structured data. While large language models receive the most attention (and the most pages), the principles and patterns apply across modalities. If you are working on image generation, audio synthesis, or video creation, the same architectural patterns, evaluation frameworks, and scaling strategies apply.

Chapter 1: The Current State of Generative AI

Generative artificial intelligence represents the most significant shift in computing since the introduction of the graphical user interface. In less than three years, generative models have progressed from laboratory curiosities to production systems serving billions of queries daily. The technology that once struggled to produce coherent paragraphs now writes code, composes music, designs molecules, and generates photorealistic images indistinguishable from human-created work.

The numbers tell a story of unprecedented adoption. ChatGPT reached 100 million users in two months—a faster adoption curve than any consumer technology in history. Enterprise spending on generative AI tools exceeded $20 billion in 2025, with projections suggesting the market will surpass $300 billion by 2030. GitHub Copilot now assists with 46% of code written by its users.

Adobe Firefly has generated over 10 billion images. Stable Diffusion runs on consumer hardware, enabling anyone with a laptop to create professional-grade artwork.

Yet beneath these headline numbers lies a more complex reality. Most organizations deploying generative AI are struggling. Production systems hallucinate at rates that make them unreliable for critical tasks. Training costs for state-of-the-art models exceed $100 million per run. The talent gap is severe—there are fewer than 50,000 researchers globally with deep expertise in training large generative models. And the gap between what generative AI can do in demonstrations and what it can reliably do in production remains the single greatest challenge facing the field.

This book is written for engineers, architects, product managers, and technology leaders who need to build, deploy, and maintain generative AI systems. It covers the full stack: from the mathematical foundations of generative models through architecture decisions, implementation patterns, operational considerations, and strategic planning. The goal is to provide a comprehensive guide that bridges the gap between research papers and production systems.

The generative AI stack has evolved rapidly and now consists of several distinct layers, each with its own characteristics, players, and challenges.

Foundation models form the base of the stack. These are large-scale models trained on massive datasets that capture broad knowledge and capabilities. The training process requires thousands of GPUs running for months, consuming millions of dollars in compute. The key players include OpenAI

(GPT-4, DALL-E), Google (Gemini, Imagen), Anthropic (Claude), Meta (Llama), and a growing open-source ecosystem centered around Hugging Face. Foundation models are characterized by their parameter count (ranging from 7 billion to over 1 trillion), training data size (trillions of tokens), and the breadth of tasks they can perform.

Fine-tuning and adaptation sit above foundation models. Most production use cases require adapting a general-purpose model to specific domains, styles, or tasks. Techniques include supervised fine-tuning (SFT), reinforcement learning from human feedback (RLHF), parameter-efficient fine-tuning (PEFT) methods like LoRA, and prompt engineering. This layer is where most organizations spend their effort, as it determines the quality and reliability of the final system.

Application frameworks provide the infrastructure for building generative AI applications. LangChain, LlamaIndex, Semantic Kernel, and similar frameworks offer abstractions for chaining model calls, managing context, integrating external tools, and handling the operational aspects of production deployments. These frameworks have matured rapidly but remain in flux, with new approaches and best practices emerging monthly.

Vector databases and retrieval systems enable the storage and retrieval of embeddings—the vector representations that allow generative models to access external knowledge. Pinecone, Weaviate, Chroma, Qdrant, and pgvector provide the infrastructure for retrieval-augmented generation (RAG), the dominant pattern for grounding generative models in domain-specific data.

Inference infrastructure handles the serving of generative models at scale. This includes optimized inference engines (vLLM, TensorRT-LLM, TGI), model serving platforms, and the hardware that runs them. The economics of inference are critical: a model that costs $0.03 per query can become prohibitively expensive at billion-query scale.

Observability and evaluation tools monitor model behavior in production. Traditional software monitoring is insufficient for generative systems, which produce non-deterministic outputs. Tools for detecting hallucinations, measuring response quality, tracking drift, and ensuring safety are still maturing.

The generative AI market is evolving at a pace that makes traditional market analysis obsolete. The competitive dynamics differ across each layer of the stack.

At the foundation model layer, competition is intense but concentrated. Training a frontier model requires capital, compute, and talent that few

organizations possess. The barrier to entry is approximately $500 million for a competitive model, and the costs are rising. OpenAI, Google, Anthropic, and Meta dominate this space, with a growing number of well-funded challengers including Mistral, Cohere, AI21 Labs, and xAI. The open-source ecosystem, led by Meta's Llama releases, is compressing margins and commoditizing the foundation layer.

At the application layer, the market is fragmenting rapidly. Thousands of startups are building generative AI applications for specific use cases: code generation, document analysis, creative tools, customer service, legal research, medical diagnosis, and hundreds more. The key question is whether these applications create defensible value or whether they are thin wrappers around foundation models that will be disintermediated as foundation models improve.

The infrastructure layer is where the most sustainable value may accrue. Companies providing vector databases, inference engines, evaluation tools, and orchestration frameworks are building infrastructure that every generative AI application needs. This layer resembles the database and cloud infrastructure markets, where a few winners captured most of the value.

Five challenges define the current state of generative AI and shape the decisions that organizations must make.

Reliability. Generative models produce outputs that are often correct, sometimes brilliant, and occasionally catastrophically wrong. The non-deterministic nature of these models means that the same input can produce different outputs on different runs. For consumer applications, this unpredictability is tolerable. For enterprise applications in healthcare, legal, financial, or safety-critical domains, reliability requirements are far higher than what current systems can deliver.

Cost. The economics of generative AI are complex. Training costs are enormous but amortized over millions of users. Inference costs are the ongoing expense that determines whether a generative AI application is viable. A chatbot costing $0.05 per conversation is profitable at $0.10 per user per month but loses money at $10 per user per month. The cost trajectory is favorable—model prices have dropped 10x in 18 months—but the absolute costs remain significant for high-volume applications.

Evaluation. Traditional software has clear pass/fail criteria. Generative AI outputs exist on a spectrum of quality, and the definition of "good" is subjective and context-dependent. Automated evaluation tools are improving but remain unreliable. Human evaluation is expensive and slow. The lack of

robust evaluation methods is a structural barrier to deploying generative AI in production.

Safety. Generative models can produce harmful content—misinformation, deepfakes, biased outputs, security vulnerabilities, and privacy violations. Mitigating these risks requires a combination of model-level interventions (alignment training, output filtering), application-level controls (guardrails, content policies), and operational measures (monitoring, incident response). The regulatory landscape is shifting fast, with the EU AI Act, sector-specific rules, and US executive orders all adding new requirements.

Talent. The demand for generative AI expertise vastly exceeds supply. Experienced ML engineers command salaries above $500,000. The skills needed span multiple disciplines—mathematics, software engineering, distributed systems, domain expertise, and product thinking. Organizations are investing heavily in training programs, but building a capable team takes 12-18 months even with aggressive hiring. PRACTICAL INSIGHT Before committing to a model provider, run a side-by-side benchmark on your own data, not published benchmarks. A model that tops leaderboards may underperform on your specific domain, language, or document formats. Spend two weeks on this evaluation; it saves months of rework later.

The economics of foundation models determine who can participate in the generative AI revolution and on what terms. Understanding these economics is essential for strategic planning.

Training costs. Training a frontier model costs $50-200 million in compute alone. The cost includes GPU rental (approximately $2-4 per GPU-hour for A100/H100), data acquisition and cleaning, research engineering time, and the opportunity cost of compute that could be used for other purposes. These costs are rising as models get larger, but the cost per parameter is falling as hardware improves and training techniques become more efficient.

Inference economics. The inference cost for a 70B parameter model is approximately $0.01-0.05 per 1K tokens, depending on the serving infrastructure. At scale, the annual inference cost for a popular model can exceed $100 million. The economics of inference are improving rapidly—new hardware (NVIDIA H100, B200), software optimizations (vLLM, TensorRT-LLM), and architectural innovations (MoE, quantization) are all reducing inference costs.

Pricing dynamics. Model API prices have fallen by roughly an order of magnitude in eighteen months: flagship-model pricing that once measured several cents per thousand tokens now measures fractions of a cent, and mid-

tier models have fallen further still. This price compression is driven by competition, efficiency improvements, and the commoditization of the foundation layer. The implication for application developers is that cost assumptions made at the start of a project may be 10x too high by the time the project launches.

The open-source dynamic. Meta's decision to release Llama models for free (with commercial restrictions) has fundamentally changed the market. Open-source models are now within 10-20% of proprietary models on most benchmarks, and the gap is closing. This compression is forcing proprietary providers to compete on quality, features, and ecosystem rather than price alone.

The talent shortage in generative AI is the single biggest constraint on the field's growth. Understanding the talent landscape is essential for organizations building generative AI capabilities.

Research talent. Researchers who can develop new architectures, training methods, and alignment techniques are extremely rare—perhaps 5,000 globally. They are concentrated at a few organizations (OpenAI, Google DeepMind, Anthropic, Meta FAIR) and command compensation packages above $1 million annually.

Engineering talent. Engineers who can build and deploy production generative AI systems are more numerous but still scarce—perhaps 50,000 globally. They need expertise in ML, distributed systems, and software engineering. Compensation ranges from $200,000 to $500,000+ depending on experience and location.

Applied talent. Applied practitioners who can use generative AI tools to solve business problems are the largest group and the fastest growing. They need domain expertise, prompt engineering skills, and basic ML knowledge. Organizations can develop this talent internally through training programs.

Training and education. Universities are rapidly adding generative AI courses, but the curriculum lags behind industry practice by 1-2 years. Online courses (Coursera, fast.ai, DeepLearning.AI) provide more current content. Internal training programs, mentorship, and hands-on projects are the most effective way to build team capability.

The foundation model market is the most strategically important layer of the generative AI stack. Understanding the competitive dynamics helps organizations make informed decisions about model selection and vendor relationships.

OpenAI. The market leader, with first-mover advantage and the strongest brand recognition. GPT-4 and its variants are widely used across consumer and enterprise applications. Key strengths: model quality, developer ecosystem, brand trust, and capital resources (backed by Microsoft). Key challenges: commoditization pressure from open-source, regulatory scrutiny, and the need to continuously innovate to maintain leadership.

Google (DeepMind). The deepest research bench and integration with Google's infrastructure. Gemini models compete directly with GPT-4 on quality. Key strengths: research talent, compute infrastructure (TPUs), integration with Google Workspace, and distribution through Google Cloud. Key challenges: slower commercialization than competitors, internal cultural challenges, and antitrust scrutiny.

Anthropic. Focused on safety and reliability, with strong enterprise adoption. Claude models are competitive on quality with strong emphasis on alignment. Key strengths: safety-first approach, enterprise customer relationships, and backing from Amazon. Key challenges: smaller than competitors, less consumer brand recognition, and the cost of safety research.

Meta. The open-source champion, with Llama models that have democratized access to high-quality generative AI. Key strengths: open-source strategy, research talent, and integration with Meta's social platforms. Key challenges: monetization model is unclear, and open-source models may cannibalize proprietary model revenue.

Mistral. A European challenger focused on efficient, open-weight models. Key strengths: technical innovation (mixture of experts), European regulatory compliance, and competitive quality at lower cost. Key challenges: limited resources compared to US competitors, and the need to differentiate beyond cost.

The open-source ecosystem. Hugging Face, EleutherAI, and individual contributors provide models, datasets, and tools that complement and compete with proprietary offerings. The open-source ecosystem is the primary driver of commoditization, compressing margins and democratizing access.

Market trajectory. The foundation model market is in a phase of rapid innovation and intense competition. Over the next 2-3 years, expect: consolidation among smaller providers, increasing quality from open-source models, specialization by domain, and pricing pressure. Organizations should avoid lock-in to any single provider and maintain the ability to switch models as the market evolves.

KEY TAKEAWAYS - Generative AI has achieved unprecedented adoption but

production deployment remains challenging - The technology stack consists of six layers: foundation models, fine-tuning, application frameworks, vector databases, inference infrastructure, and observability - Five key challenges define the field: reliability, cost, evaluation, safety, and talent - The market is concentrated at the foundation layer, fragmenting at the application layer, and consolidating at the infrastructure layer - Organizations must address all five challenges to deploy generative AI successfully in production

End of free sample. The full book is available at putrabooks.github.io

ABOUT THE AUTHOR

Aiman JoeHanz writes about the technologies that are quietly reshaping the world. His Building of Tomorrow series examines the infrastructure behind artificial intelligence -- from blockchain and cloud computing to agentic systems and token optimization -- and asks how these technologies will evolve over the coming decade.

His books are built on technical research and engineering documentation, written for readers who want to understand systems rather than trends. He covers what is being built today and what it will take to build what comes next.

He believes that understanding infrastructure is the difference between watching the future happen and helping to create it.