

THE TOKEN ECONOMY

OPTIMIZING LLM APPLICATIONS
FOR **PERFORMANCE** AND **COST**



UNDERSTAND TOKENS
from theory to practice



**ENGINEER SMARTER
PROMPTS**
and reduce waste



**BUILD EFFICIENT RAG
AND AGENTIC WORKFLOWS**



**LEVERAGE CACHING
AND ROUTING**
to scale



— **TRAVIS JOHNSON** —

The Token Economy (FREE)

Optimizing LLM Applications for Performance and Cost

Steve Publications

This book is available at <https://leanpub.com/thetokeneconomy>

This version was published on 2026-07-18



This is a [Leanpub](#) book. Leanpub empowers authors and publishers with the Lean Publishing process. [Lean Publishing](#) is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

© 2026 Steve Publications

Contents

- Optimizing LLM Applications for Performance and Cost 1**
- Introduction 2**
- Chapter 1: The Anatomy of a Token 3**
 - The Tokenization Process 3
 - The Multilingual Token Tax 3
 - Context Window Mechanics 3
 - Context Pruning: Strategic Information Removal 4
 - Semantic Compression and Gisting 4
 - Input vs Output Tokens: The Compute Asymmetry 5
 - Implementation Guide: Token Counting and Analysis 5
 - Case Study: Reducing the Multilingual Tax (Llama 2 vs Llama 3) 6
- Chapter 2: Prompt Engineering for Token Efficiency 7**
 - Prompt Minimization and Instruction Density 7
 - Few-Shot Optimization and Dynamic Selection 7
 - Chain-of-Thought (CoT) Trade-offs 8
 - Case Study: Optimizing a Data Extraction Pipeline 9
 - Benchmarking: Naive vs. Optimized Prompting 9
- Chapter 3: Context Engineering and Management 10**
 - The Geometry of Attention: Applying the “Lost in the Middle” Theory 10
 - Semantic Compression 10
 - Priority-Based Windowing Architectures 10
 - Token Budgeting and Enforcement 10
 - Case Study: The Legal Analysis Problem 11
 - Benchmarking: Context Management Strategies 11
- Chapter 4: Efficient Retrieval-Augmented Generation (RAG) 12**
 - Chunking Strategies for Token Efficiency 12

CONTENTS

Reranking for Precision: From Vector Search to Cross-Encoders . . .	12
Contextual Compression and Retrieval	13
Hybrid Search Optimization: Vector + Keyword	13
Implementation Guide: Optimized RAG Pipeline	14
Case Study: The Technical Documentation Bot	14
Benchmarking: Retrieval Precision vs. Token Cost	14
Chapter 5: Managing Conversation Memory	15
The Linear Growth Problem: The Cost of Remembrance	15
Memory Architectures: From Buffer to Summary	15
Entity-Based and Structured Memory: Moving Beyond Text	16
Recursive Summarization for Infinite Horizons	16
Selective Forgetting and Utility-Based Pruning	16
Implementation Guide: Summary-Based Memory Manager	17
Case Study: The Personal AI Assistant	17
Benchmarking: Memory Architecture Efficiency	17
Chapter 6: Agentic Workflows and Tool Use	18
Tool Definition Optimization: Combating Schema Bloat	18
Loop Control and Termination: Preventing Token Runaway	18
State Management: The World Model Pattern	19
Implementation Guide: Agentic Loop with Token Guards	19
Case Study: The Research Agent	20
Benchmarking: Agentic Patterns	20
Chapter 7: Structured Outputs and Format Optimization	21
The Formatting Tax: Analyzing Structural Overhead	21
Constrained Sampling: Eliminating the Retry Loop	21
Eliminating Conversational Bloat	22
Shifting Formatting Logic to Application Code	22
Implementation Guide: Format Optimization	22
Case Study: The High-Volume Extraction Pipeline	22
Benchmarking: Format Efficiency and Reliability	22
Chapter 8: Caching and Inference Optimization	23
Prompt Caching: Architecting for Prefix Consistency	23
KV Cache Fundamentals: The Physics of Prefill and Decode	23
Speculative Decoding: Accelerating the Decode Phase	24
Batching Strategies and Throughput Optimization	24
Implementation Guide: Caching Middleware	24

Case Study: The Enterprise Knowledge Base	25
Benchmarking: Caching and Latency Impact	25
Streaming Token Optimization: Reducing Real-Time Waste	25
Provider-Specific API Optimizations	25
Chapter 9: Model Selection and Routing	27
The Model Hierarchy: Maximizing Intelligence per Token	27
LLM Routing Architectures: Dynamic Model Selection	27
Distillation for Token Efficiency: Creating Specialized Models	28
Comparison Metrics: The Quality vs. Cost Curve	28
Implementation Guide: Embedding-Based Router	29
Case Study: The Customer Support Tiering	29
Chapter 10: Measuring, Benchmarking, and Evaluation	30
Token Accounting: Attributing Cost to Value	30
Latency Decomposition: TTFT vs TPOT	30
Designing Quality vs. Cost Sweeps	30
A/B Testing and LLM-as-a-Judge	31
Implementation Guide: Latency and Token Tracker	31
Case Study: Validating a Prompt Compression Strategy	31
Chapter 11: Production Architectures and Security	33
Token-Based Rate Limiting and Quota Management	33
Security: Mitigating Token Exhaustion Attacks	33
The Token Firewall: A Comprehensive Middleware Design	34
Asynchronous Processing and Batch Queues	34
Implementation Guide: Token-Based Rate Limiter	34
Case Study: Mitigating a Token Exhaustion Attack	34
Conclusion: The Future of the Token Economy	35
Appendix: The Prompt Library	36
1. Data Extraction	36
2. Document Summarization	36
3. Complex Reasoning / Logic	36
4. Code Generation	36
Instruction Density Audit Checklist	36
References	37

Optimizing LLM Applications for Performance and Cost

About this book: This comprehensive guide provides developers, AI engineers, and architects with the theoretical foundation and practical strategies needed to optimize token usage in Large Language Model (LLM) applications. It explores the entire lifecycle of a token, from initial tokenization and prompt engineering to advanced RAG architectures, agentic workflows, and production-grade caching and routing strategies.

Introduction

In the era of generative AI, tokens are more than just fragments of text; they are the fundamental unit of cost, latency, and cognitive capacity. For a developer building a simple prototype, the difference between 1,000 and 5,000 tokens in a prompt is negligible. However, for an engineer deploying a system to millions of users, that same difference represents a massive escalation in operational expenditure and a noticeable degradation in response time.

The “Token Economy” refers to the constant trade-off between the amount of information provided to a model and the resources required to process it. While modern models boast context windows of hundreds of thousands or even millions of tokens, the reality is that more data does not always lead to better results. In fact, providing too much irrelevant context can lead to “performance collapse,” where the model ignores critical instructions or fails to retrieve key facts buried in the middle of a prompt.

The goal of this book is to move beyond naive prompting and toward a disciplined approach to token management. We will examine how different tokenizers slice language, why some formats are more expensive than others, and how architectural choices like RAG (Retrieval-Augmented Generation) and model routing can be optimized to maximize “intelligence per token.”

Whether you are fighting the “lost in the middle” phenomenon, trying to slash your API bill, or designing a complex agentic workflow that must remain stable over long conversations, this book provides the blueprints. We will progress from the basic physics of tokens to the sophisticated orchestration of production AI systems, ensuring that every token spent contributes directly to the value of the final output.

A word of caution: optimization is not always the priority. In safety-critical domains (medical diagnosis, legal advice, aviation), clarity and completeness should trump brevity. Over-optimized prompts can become brittle, harder to maintain, and more prone to edge-case failures. The strategies in this book are tools for scaling, not excuses to sacrifice reliability for savings. Know when to optimize, and know when not to.

Chapter 1: The Anatomy of a Token

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Tokenization Process

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The BPE Merge Algorithm in Detail

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Comparative Analysis: BPE vs. WordPiece vs. SentencePiece

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Multilingual Token Tax

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Context Window Mechanics

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Quadratic Bottleneck

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Geometry of Attention: Lost in the Middle

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Context Pruning: Strategic Information Removal

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Heuristic Filtering

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Model-Based Relevance Scoring

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Semantic Compression and Gisting

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Recursive Summarization (The Pyramid Pattern)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Gisting and Soft Prompting

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Input vs Output Tokens: The Compute Asymmetry

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Prefill vs. Decoding

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The KV Cache and Memory Pressure

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Economic Implications

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Implementation Guide: Token Counting and Analysis

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Python Implementation

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

TypeScript Implementation

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Case Study: Reducing the Multilingual Tax (Llama 2 vs Llama 3)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Chapter 2: Prompt Engineering for Token Efficiency

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Prompt Minimization and Instruction Density

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Theory of Instruction Density

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Prompt Pattern Catalog

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Prompt Anti-Patterns: The “Token Leaks”

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Advanced Prompt Compression (LLMLingua)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Few-Shot Optimization and Dynamic Selection

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Dynamic Selection Workflow

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Multi-Language Implementation

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Benchmarking Dynamic vs. Static Few-Shot

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Chain-of-Thought (CoT) Trade-offs

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Cost of Reasoning

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

CoT Strategies and Efficiency

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The “Internal Monologue” Pattern

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Case Study: Optimizing a Data Extraction Pipeline

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Benchmarking: Naive vs. Optimized Prompting

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Chapter 3: Context Engineering and Management

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Geometry of Attention: Applying the “Lost in the Middle” Theory

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Semantic Compression

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Priority-Based Windowing Architectures

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Three-Zone Architecture

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Mermaid Architecture Specification

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Token Budgeting and Enforcement

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Implementing the Budgeter

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Case Study: The Legal Analysis Problem

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Benchmarking: Context Management Strategies

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Chapter 4: Efficient Retrieval-Augmented Generation (RAG)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Chunking Strategies for Token Efficiency

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Failure of Fixed-Size Chunking

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Semantic Chunking: Breaking by Meaning

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Parent-Child Indexing (Small-to-Big Retrieval)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Reranking for Precision: From Vector Search to Cross-Encoders

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Bi-Encoders vs. Cross-Encoders

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Reranking Pipeline

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Contextual Compression and Retrieval

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The “Distillation” Approach

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Contextual Retrieval (The Anthropic Method)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Hybrid Search Optimization: Vector + Keyword

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Power of BM25

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Reciprocal Rank Fusion (RRF)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Implementation Guide: Optimized RAG Pipeline

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Case Study: The Technical Documentation Bot

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Benchmarking: Retrieval Precision vs. Token Cost

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Chapter 5: Managing Conversation Memory

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Linear Growth Problem: The Cost of Remembrance

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Mathematics of Accumulation

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Memory Architectures: From Buffer to Summary

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

1. Buffer Memory (The Naive Approach)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

2. Sliding Window Memory (The Fixed-Cost Approach)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

3. Summary-Based Memory (The Balanced Approach)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Entity-Based and Structured Memory: Moving Beyond Text

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The User Profile Pattern

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Long-Term Memory with Mem0

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Recursive Summarization for Infinite Horizons

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Hierarchical Memory Algorithm

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Selective Forgetting and Utility-Based Pruning

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Utility Scoring Model

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Implementation Guide: Summary-Based Memory Manager

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Case Study: The Personal AI Assistant

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Benchmarking: Memory Architecture Efficiency

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Chapter 6: Agentic Workflows and Tool Use

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Tool Definition Optimization: Combating Schema Bloat

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Anatomy of Schema Bloat

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Dynamic Tool Selection (On-Demand Loading)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Prefix Caching for Tool Definitions

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Model Context Protocol (MCP)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Loop Control and Termination: Preventing Token Runaway

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Implementing Loop Guards

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Reasoning Traces and Debugging

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Parallelism vs Sequentiality: The Token Trade-off

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

State Management: The World Model Pattern

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Observation Summarization

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The World Model Pattern

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Implementation Guide: Agentic Loop with Token Guards

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Case Study: The Research Agent

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Benchmarking: Agentic Patterns

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Chapter 7: Structured Outputs and Format Optimization

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Formatting Tax: Analyzing Structural Overhead

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Quantifying the Tax

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Compact Formats: TOON and Beyond

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Constrained Sampling: Eliminating the Retry Loop

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

How Constrained Decoding Works

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Trade-offs: Flexibility vs. Rigidity

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Eliminating Conversational Bloat

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Zero-Bloat Strategy

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Shifting Formatting Logic to Application Code

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Implementation Guide: Format Optimization

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Case Study: The High-Volume Extraction Pipeline

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Benchmarking: Format Efficiency and Reliability

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Chapter 8: Caching and Inference Optimization

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Prompt Caching: Architecting for Prefix Consistency

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Comparative Analysis: Anthropic vs. OpenAI Caching

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

KV Cache Fundamentals: The Physics of Prefill and Decode

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Prefill Phase vs. The Decode Phase

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Memory Requirements and the KV Cache Bottleneck

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

PagedAttention: Solving Memory Fragmentation

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Speculative Decoding: Accelerating the Decode Phase

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Sequential Bottleneck

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Speculative Approach

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Batching Strategies and Throughput Optimization

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Continuous Batching

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Throughput vs. Latency

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Implementation Guide: Caching Middleware

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Case Study: The Enterprise Knowledge Base

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Benchmarking: Caching and Latency Impact

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Streaming Token Optimization: Reducing Real-Time Waste

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Early Termination and Content Monitoring

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Cost of Abandoned Streams

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Streaming vs. Batching: The Throughput Trade-off

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Provider-Specific API Optimizations

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

OpenAI: Precision Tuning and Automatic Caching

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Anthropic: Explicit Control and Breakpoints

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Google Gemini: Long Context and Multimodal Efficiency

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Tool Call Overhead Comparison

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Chapter 9: Model Selection and Routing

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Model Hierarchy: Maximizing Intelligence per Token

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Metric: Intelligence per Token (IpT)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Comparison: Frontier vs. Flash Performance

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

LLM Routing Architectures: Dynamic Model Selection

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Routing Strategies

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Routing Decision Matrix

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Open-Source vs. Closed-Source: The Token Trade-off

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Distillation for Token Efficiency: Creating Specialized Models

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Distillation Pipeline

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Result: Extreme Token Efficiency

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Comparison Metrics: The Quality vs. Cost Curve

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Finding the “Elbow Point”

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Cost-Adjusted Accuracy

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Implementation Guide: Embedding-Based Router

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Case Study: The Customer Support Tiering

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Chapter 10: Measuring, Benchmarking, and Evaluation

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Token Accounting: Attributing Cost to Value

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Token Attribution Framework

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Latency Decomposition: TTFT vs TPOT

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Time to First Token (TTFT)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Time Per Output Token (TPOT)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Designing Quality vs. Cost Sweeps

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Evaluation Matrix

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Finding the “Elbow Point”

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

A/B Testing and LLM-as-a-Judge

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The LLM-as-a-Judge Framework

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The A/B Testing Pipeline

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Implementation Guide: Latency and Token Tracker

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Case Study: Validating a Prompt Compression Strategy

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Chapter 11: Production Architectures and Security

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Token-Based Rate Limiting and Quota Management

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Token Bucket Algorithm

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Tiered Quotas

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Security: Mitigating Token Exhaustion Attacks

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The “Infinite Loop” Attack

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Context Stuffing and Guardrail Bypass

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

The Token Firewall: A Comprehensive Middleware Design

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Detailed Architecture Flow

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Asynchronous Processing and Batch Queues

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Implementation Guide: Token-Based Rate Limiter

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Case Study: Mitigating a Token Exhaustion Attack

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Conclusion: The Future of the Token Economy

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Appendix: The Prompt Library

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

1. Data Extraction

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

2. Document Summarization

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

3. Complex Reasoning / Logic

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

4. Code Generation

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

Instruction Density Audit Checklist

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.

References

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/thetokeneconomy>.