

DSPy

in Depth

PROGRAMMING LANGUAGE MODELS FROM ZERO TO PRODUCTION



SIGNATURES

Define tasks declaratively with typed inputs, outputs, and instructions



MODULES

Compose the right modules, from simple predictions to multi-step reasoning pipelines



TOOLS & AGENTS

Build tool-using agents with the ReAct module



RAG SYSTEMS

Implement retrieval-augmented generation with optimized retrieval quality



EVALUATION

Define metrics and measure your program's quality quantitatively



OPTIMIZERS

Automatically tune prompts with GEPA's reflective evolution, MIPROv2's joint search, and BootstrapFewShot's synthesis



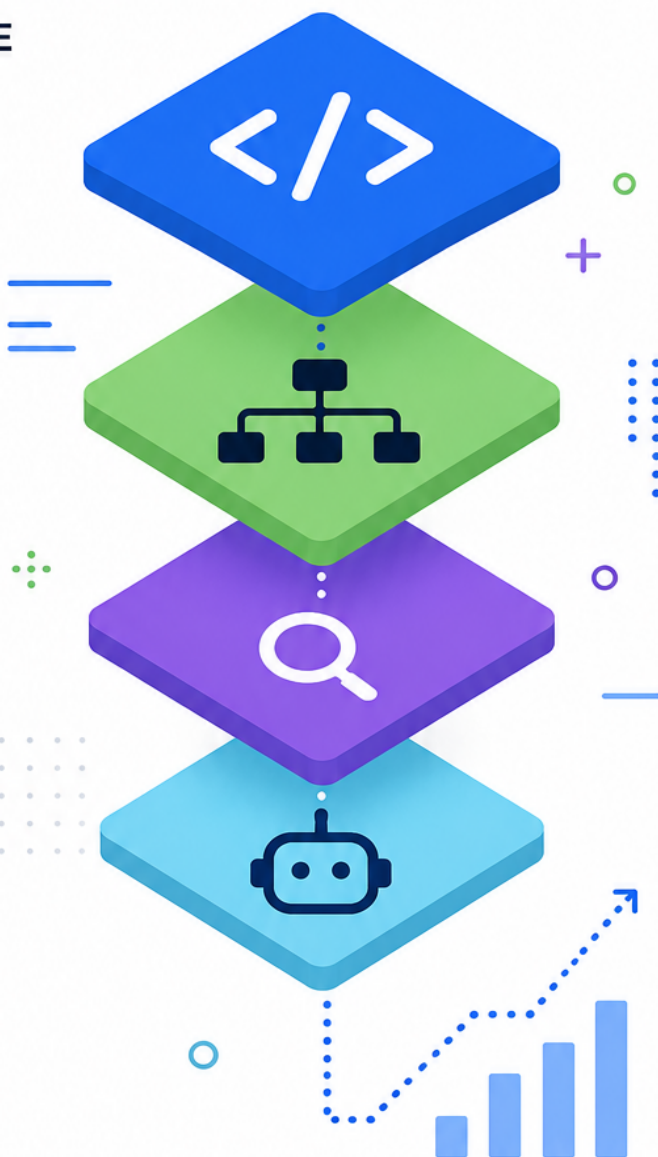
DEPLOYMENT

Deploy with FastAPI, MLflow, and containers



PRODUCTION READY

Caching, streaming, async execution, and performance optimization



STEVE T.

DSPy in Depth

Programming Language Models from Zero to Production

Steve T. Publications

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

This version was published on 2026-07-09



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 T. Publications

Contents

- Programming Language Models from Zero to Production 1**
- Introduction: The End of Prompt Engineering as We Know It 2**
- Chapter 1: What Is DSPy and Why It Matters 3**
 - The Prompting Crisis 3
 - From Prompting to Programming 3
 - The DSPy Philosophy 3
 - How DSPy Fits in the Ecosystem 3
- Chapter 2: Installation and Your First Program 4**
 - Installing DSPy 4
 - Configuring Your Language Model 4
 - Your First Signature and Prediction 4
 - Understanding the Output 4
 - Switching Between Models 4
 - A Complete First Program 4
 - Migration Notes from Older Versions 5
- Chapter 3: Signatures in Depth 6**
 - String-Based Signatures 6
 - Class-Based Signatures 6
 - Input Fields and Output Fields 6
 - Type Annotations and Pydantic Integration 6
 - Signature Mutation and Composition 6
 - Signature Introspection 6
 - Designing Effective Signatures 7
- Chapter 4: Core Modules 8**
 - dspy.Predict: Basic Prediction 8
 - dspy.ChainOfThought: Step-by-Step Reasoning 8

CONTENTS

dspy.ProgramOfThought: Code-Assisted Reasoning	8
dspy.MultiChainComparison: Comparing Outputs	8
dspy.BestOfN and dspy.Refine: Output Refinement	8
dspy.Parallel: Concurrent Execution	8
Choosing the Right Module	9
Chapter 5: Composing Custom Modules	10
The dspy.Module Base Class	10
Building Your First Compound Module	10
Forward Method Design Patterns	10
Nested Modules and Sub-modules	10
Module Inspection and Introspection	10
Deep Copy and State Management	10
Chapter 6: Language Models and Configuration	12
The dspy.LM Interface	12
OpenAI, Anthropic, and Google Providers	12
Local Models and OpenAI-Compatible Endpoints	12
Context Management with dspy.context()	12
Caching and Performance	12
Tracking Usage and Costs	12
Adapter Selection	13
Chapter 7: Evaluation and Metrics	14
Writing Metric Functions	14
The dspy.Evaluate Class	14
Built-in Metrics	14
LLM-as-a-Judge Evaluation	14
Debugging with inspect_history()	14
Composite Metrics	14
Evaluation Best Practices	15
Chapter 8: Optimizers (Teleprompters)	16
The Optimization Paradigm	16
BootstrapFewShot Family	16
COPRO: Instruction Optimization	16
MIPROv2: Joint Instruction and Demo Search	16
GEPA: Reflective Prompt Evolution	16
SIMBA and Other Specialized Optimizers	16
Choosing the Right Optimizer	17

- Saving and Loading Compiled Programs 17
- Chapter 9: Retrieval-Augmented Generation (RAG) 18**
 - The dspy.Retrieve Module 18
 - ColBERTv2 Integration 18
 - Building a Basic RAG Pipeline 18
 - Multi-Hop RAG 18
 - Optimizing RAG Programs 18
 - Custom Retriever Integration 18
- Chapter 10: Agents and Tool Use 20**
 - The ReAct Paradigm 20
 - dspy.ReAct Module 20
 - Defining Tools as Functions 20
 - Building Production Agents 20
 - Advanced Tool Patterns 20
- Chapter 11: Structured Outputs and Type Safety 21**
 - Typed Predictors 21
 - Pydantic Models in Signatures 21
 - JSON Schema Generation 21
 - Output Validation Strategies 21
 - Complex Nested Structures 21
- Chapter 12: Advanced Patterns and Workflows 22**
 - Assertions and Constraints (Refine Module) 22
 - Multimodal Programs with Images and Audio 22
 - Conversation History Management 22
 - Async and Streaming Support 22
 - MCP Integration 22
 - Caching Strategies for Production 22
- Chapter 13: Production Deployment 24**
 - Saving and Loading Optimized Programs 24
 - FastAPI Deployment 24
 - MLflow Integration 24
 - Caching Strategies 24
 - Monitoring and Observability 24
- Chapter 14: Performance Optimization and Best Practices 25**

Cost Optimization	25
Latency Reduction	25
Error Handling Patterns	25
Testing DSPy Programs	25
Common Pitfalls and Anti-Patterns	25
Chapter 15: Real-World Projects	26
Project 1: Intelligent Customer Service Agent	26
Project 2: Multi-Document Research Assistant	26
Project 3: Automated Code Review System	26
Conclusion: The Future of LLM Programming	27
References	28

Programming Language Models from Zero to Production

About this book: DSPy is the framework that transforms how we build AI systems. Instead of hand-crafting brittle prompts, you write compositional Python code and let DSPy teach your language model to deliver high-quality outputs automatically. This book takes you from zero to production, covering every concept in depth with complete, runnable code examples. You will learn to define signatures, compose modules, build retrieval-augmented pipelines, construct tool-using agents, optimize programs with state-of-the-art algorithms like GEPA and MIPROv2, and deploy your systems reliably. Whether you are a software engineer new to LLM development or an experienced practitioner looking for a systematic approach, this book gives you the practical knowledge to build AI applications that work.

Introduction: The End of Prompt Engineering as We Know It

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

Chapter 1: What Is DSPy and Why It Matters

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

The Prompting Crisis

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

From Prompting to Programming

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

The DSPy Philosophy

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

How DSPy Fits in the Ecosystem

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

Chapter 2: Installation and Your First Program

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

Installing DSPy

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

Configuring Your Language Model

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

Your First Signature and Prediction

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

Understanding the Output

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

Switching Between Models

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

A Complete First Program

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

Migration Notes from Older Versions

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

Chapter 3: Signatures in Depth

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

String-Based Signatures

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

Class-Based Signatures

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

Input Fields and Output Fields

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

Type Annotations and Pydantic Integration

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

Signature Mutation and Composition

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

Signature Introspection

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

Designing Effective Signatures

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

Chapter 4: Core Modules

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

dspy.Predict: Basic Prediction

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

dspy.ChainOfThought: Step-by-Step Reasoning

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

dspy.ProgramOfThought: Code-Assisted Reasoning

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

dspy.MultiChainComparison: Comparing Outputs

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

dspy.BestOfN and dspy.Refine: Output Refinement

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

dspy.Parallel: Concurrent Execution

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

Choosing the Right Module

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

Chapter 5: Composing Custom Modules

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

The `dspy.Module` Base Class

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

Building Your First Compound Module

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

Forward Method Design Patterns

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

Nested Modules and Sub-modules

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

Module Inspection and Introspection

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

Deep Copy and State Management

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

Chapter 6: Language Models and Configuration

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

The dspy.LM Interface

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

OpenAI, Anthropic, and Google Providers

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

Local Models and OpenAI-Compatible Endpoints

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

Context Management with `dspy.context()`

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

Caching and Performance

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

Tracking Usage and Costs

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

Adapter Selection

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

Chapter 7: Evaluation and Metrics

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

Writing Metric Functions

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

The dspy.Evaluate Class

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

Built-in Metrics

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

LLM-as-a-Judge Evaluation

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

Debugging with inspect_history()

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

Composite Metrics

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

Evaluation Best Practices

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

Chapter 8: Optimizers (Teleprompters)

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

The Optimization Paradigm

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

BootstrapFewShot Family

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

COPRO: Instruction Optimization

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

MIPROv2: Joint Instruction and Demo Search

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

GEPA: Reflective Prompt Evolution

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

SIMBA and Other Specialized Optimizers

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

Choosing the Right Optimizer

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

Saving and Loading Compiled Programs

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

Chapter 9: Retrieval-Augmented Generation (RAG)

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

The dspy.Retrieve Module

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

ColBERTv2 Integration

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

Building a Basic RAG Pipeline

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

Multi-Hop RAG

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

Optimizing RAG Programs

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

Custom Retriever Integration

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

Chapter 10: Agents and Tool Use

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

The ReAct Paradigm

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

dspy.ReAct Module

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

Defining Tools as Functions

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

Building Production Agents

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

Advanced Tool Patterns

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

Chapter 11: Structured Outputs and Type Safety

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

Typed Predictors

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

Pydantic Models in Signatures

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

JSON Schema Generation

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

Output Validation Strategies

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

Complex Nested Structures

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

Chapter 12: Advanced Patterns and Workflows

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

Assertions and Constraints (Refine Module)

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

Multimodal Programs with Images and Audio

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

Conversation History Management

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

Async and Streaming Support

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

MCP Integration

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

Caching Strategies for Production

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

Chapter 13: Production Deployment

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

Saving and Loading Optimized Programs

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

FastAPI Deployment

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

MLflow Integration

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

Caching Strategies

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

Monitoring and Observability

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

Chapter 14: Performance Optimization and Best Practices

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

Cost Optimization

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

Latency Reduction

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

Error Handling Patterns

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

Testing DSPy Programs

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

Common Pitfalls and Anti-Patterns

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

Chapter 15: Real-World Projects

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

Project 1: Intelligent Customer Service Agent

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

Project 2: Multi-Document Research Assistant

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

Project 3: Automated Code Review System

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

Conclusion: The Future of LLM Programming

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

References

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