

Engenharia de IA

**Construindo aplicações com
modelos de fundação**

Chip Huyen

O'REILLY®
Novatec

Authorized Portuguese translation of the English edition of AI Engineering ISBN 9781098166304 © 2025 Developer Experience Advisory LLC. This translation is published and sold by permission of O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

Tradução em português autorizada da edição em inglês da obra AI Engineering ISBN 9781098166304 © 2025 Developer Experience Advisory LLC. Esta tradução é publicada e vendida com a permissão da O'Reilly Media, Inc., detentora de todos os direitos para publicação e venda desta obra.

© Novatec Editora Ltda. [2026].

Todos os direitos reservados e protegidos pela Lei 9.610 de 19/02/1998. É proibida a reprodução desta obra, mesmo parcial, por qualquer processo, sem prévia autorização, por escrito, do autor e da Editora.

Editor: Rubens Prates

Tradutor: Aldir Coelho Corrêa da Silva

ISBN do impresso: 978-85-7522-996-5

ISBN do ebook: 978-85-7522-997-2

Novatec Editora Ltda.

Rua Luís Antônio dos Santos 110

02460-000 – São Paulo, SP – Brasil

Tel.: +55 11 2959-6529

Email: novatec@novatec.com.br

Site: <https://novatec.com.br>

Twitter: twitter.com/novateceditora

Facebook: facebook.com/novatec

LinkedIn: <https://linkedin.com/company/novatec-editora/>

GRA20260305

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Huyen, Chip
Engenharia de IA : construindo aplicações com
modelos de fundação / Chip Huyen ; tradução Aldir
Coelho Corrêa da Silva ; editor Rubens Prates. --
São Paulo : Novatec Editora, 2026.

Título original: AI Engineering.
ISBN 978-85-7522-996-5

1. Engenharia de prompt 2. Inteligência
artificial 3. Tecnologia I. Prates, Rubens.
II. Título.

26-340930.0

CDD-006.3

Índices para catálogo sistemático:

1. Inteligência artificial 006.3

Eliane de Freitas Leite - Bibliotecária - CRB 8/8415

Sumário

Prefácio	11
CAPÍTULO 1: Introdução à construção de aplicações de IA com modelos de fundação	21
Surgimento da engenharia de IA	22
Dos modelos de linguagem aos grandes modelos de linguagem	22
Dos grandes modelos de linguagem aos modelos de fundação	28
Dos modelos de fundação à engenharia de IA	31
Casos de uso dos modelos de fundação	35
Codificação	38
Produção de imagem e vídeo	40
Escrita	40
Educação	42
Bots de conversa	43
Agregação de informações	44
Organização de dados	44
Automação de fluxo de trabalho	45
Planejando aplicações de IA	46
Avaliação do caso de uso	46
Definindo expectativas	49
Planejamento de etapas	50
Manutenção	51
Pilha da engenharia de IA	52
As três camadas da pilha de IA	54
Engenharia de IA <i>versus</i> engenharia de ML	56
Engenharia de IA <i>versus</i> engenharia full-stack	62
Resumo	63

CAPÍTULO 2: Entendendo os modelos de fundação	65
Dados de treinamento.....	66
Modelos multilíngues.....	67
Modelos específicos de domínio.....	71
Modelando	73
Arquitetura do modelo.....	73
Tamanho do modelo	82
Pós-treinamento	91
Ajuste fino supervisionado.....	93
Ajuste fino conforme as preferências	96
Amostragem.....	100
Aspectos básicos da amostragem.....	100
Estratégias de amostragem	102
Computação em tempo de teste.....	106
Saídas estruturadas	110
A natureza probabilística da IA.....	114
Resumo	120
CAPÍTULO 3: Metodologia de avaliação	122
Desafios na avaliação dos modelos de fundação.....	123
Entendendo as métricas de modelagem de linguagem	127
Entropia.....	128
Entropia cruzada.....	128
Bits-por-caractere e bits-por-byte.....	129
Perplexidade	130
Interpretação e casos de uso da perplexidade.....	130
Avaliação exata.....	133
Exatidão funcional.....	133
Medições de similaridade em relação a dados de referência.....	135
Introdução ao embedding.....	140
IA como juiz.....	143
Por que a IA como juiz?	143
Como usar a IA como juiz.....	145
Limitações da IA como juiz.....	148
Quais modelos podem agir como juízes?	151
Modelos de classificação com avaliação comparativa	154
Desafios da avaliação comparativa.....	157
Futuro da avaliação comparativa	160
Resumo	161
CAPÍTULO 4: Avaliação de sistemas de IA	163
Critérios de avaliação.....	164
Capacidade específica do domínio	165
Capacidade de geração.....	167

Capacidade de seguir instruções.....	175
Custo e latência	180
Seleção do modelo.....	181
Fluxo de trabalho de seleção de modelos	182
Construir <i>versus</i> comprar o modelo	183
Navegue em benchmarks públicos	193
Projete seu pipeline de avaliação	201
Etapa 1: Avalie todos os componentes de um sistema	201
Etapa 2: Crie uma diretriz de avaliação.....	202
Etapa 3: Defina métodos e dados de avaliação.....	204
Resumo	209
CAPÍTULO 5: Engenharia de prompt	211
Introdução ao prompting	212
Aprendizado no contexto: zero-shot e few-shot	213
Prompt do sistema e prompt do usuário	215
Tamanho e eficiência do contexto	217
Melhores práticas da engenharia de prompt.....	219
Escreva instruções claras e explícitas	219
Forneça contexto suficiente.....	222
Divida tarefas complexas em subtarefas mais simples.....	223
Dê tempo ao modelo para pensar.....	225
Itere em seus prompts	227
Avalie ferramentas de engenharia de prompt	228
Organize e versione os prompts.....	231
Engenharia de prompt defensiva.....	233
Prompts proprietários e engenharia de prompt reversa	234
Jailbreaking e injeção de prompt	236
Extração de informações.....	241
Defesas contra ataques de prompt.....	245
Resumo	248
CAPÍTULO 6: RAG e agentes	249
RAG.....	250
Arquitetura RAG.....	252
Algoritmos de recuperação.....	253
Otimização da recuperação.....	263
RAG não é somente para textos.....	268
Agentes.....	271
Visão geral dos agentes.....	271
Ferramentas	273
Planejamento	277
Modos de falha do agente e avaliação	292
Memória	295
Resumo	298

CAPÍTULO 7: Ajuste fino	300
Visão geral do ajuste fino	301
Quando fazer o ajuste fino	304
Razões para realizar o ajuste fino	304
Razões para não realizar o ajuste fino	305
Ajuste fino e RAG	309
Gargalos de memória	312
Retropropagação e parâmetros treináveis	313
Cálculo de memória	314
Representações numéricas	317
Quantização	319
Técnicas de ajuste fino	323
Ajuste fino eficiente dos parâmetros	323
Fusão de modelos e ajuste fino multitarefa	336
Táticas de ajuste fino	344
Resumo	348
CAPÍTULO 8: Engenharia de dataset	350
Curadoria de dados	352
Qualidade dos dados	355
Cobertura dos dados	356
Quantidade dos dados	359
Aquisição e anotação de dados	363
Aumento e síntese dos dados	366
Por que usar a síntese de dados	367
Técnicas tradicionais de síntese de dados	368
Síntese de dados alimentada por IA	372
Destilação de modelos	380
Processamento de dados	381
Inspeção dos dados	381
Desduplicação dos dados	383
Limpe e filtre os dados	385
Formate os dados	385
Resumo	387
CAPÍTULO 9: Otimização da inferência	389
Entendendo a otimização da inferência	390
Visão geral da inferência	390
Métricas de desempenho da inferência	395
Aceleradores de IA	401
Otimização da inferência	408
Otimização do modelo	409
Otimização do serviço de inferência	421
Resumo	427

CAPÍTULO 10: Arquitetura da engenharia de IA e feedback do usuário.....	429
Arquitetura da engenharia de IA.....	429
Etapa 1: melhore o contexto	430
Etapa 2: insira guardrails.....	431
Etapa 3: adicione um roteador de modelo e um gateway.....	436
Etapa 4: reduza a latência com caches.....	440
Etapa 5: adicione padrões de agentes	443
Monitoramento e observabilidade.....	444
Orquestração de pipeline de IA	451
Feedback do usuário	452
Extraindo feedback conversacional.....	453
Planejamento do feedback.....	458
Limitações do feedback.....	466
Resumo	468
Epílogo	470
Índice remissivo	473