

LLMs

As Partes Difíceis

**Soluções open source de IA para
as armadilhas mais comuns**

Thársis T. P. Souza
Jonathan K. Regenstein Jr.

Apresentação de Simon Guest

O'REILLY®
Novatec

Authorized Portuguese translation of the English edition of Large Language Models: The Hard Parts ISBN 9798341622524 © 2026 Thársis T.P. Souza and Jonathan K. Regenstein, Jr. 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 Large Language Models: The Hard Parts ISBN 9798341622524 © 2026 Thársis T.P. Souza e Jonathan K. Regenstein, Jr. 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

Tradução: Aldir Coelho Corrêa da Silva

ISBN do impresso: 978-65-83913-08-1

ISBN do ebook: 978-65-83913-09-8

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: <https://twitter.com/novateceditora>

Facebook: <https://facebook.com/novatec>

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

GRA20260721

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

Souza, Thársis T P

LLMs : as partes difíceis : soluções open source de IA para as armadilhas mais comuns / Thársis T P Souza, Jonathan K Regenstein Jr ; tradução Aldir Coelho Corrêa da Silva. -- São Paulo : Novatec Editora, 2026.

Título original: Large language models: the hard parts

ISBN 978-65-83913-08-1

1. Inteligência artificial 2. Linguagem
3. Python (Linguagem de programação para computadores) 4. Tecnologia I. Regenstein Jr, Jonathan K. II. Título.

26-378406.1

CDD-006.3

Índices para catálogo sistemático:

1. Inteligência artificial 006.3

Camila Aparecida Rodrigues - Bibliotecária CRB -
SP-010133/O

Sumário

Apresentação.....	13
Prefácio	15
CAPÍTULO 1: Princípios básicos: o que considerar antes de construir com LLMs	21
Por que open source?	22
Considerações estratégicas	23
Requisitos empresariais.....	23
Requisitos relacionados aos stakeholders	23
Requisitos de conformidade e segurança	24
Requisitos de desempenho	24
Requisitos operacionais.....	24
Requisitos de integração	25
Requisitos de gerenciamento de dados	25
Estruturas organizacionais de IA.....	25
Estrutura centralizada	25
Estrutura descentralizada	26
Estrutura federada	26
A importância dos dados.....	27
Tipos de modelo	29
Modelos base	29
Modelos ajustados por instrução	30
Modelos adaptados ao domínio	31
Características dos modelos.....	32
Tamanho do contexto.....	32
Controle da saída	32
Caching	33
Limite de tokens de saída.....	33

Custo e velocidade.....	33
Licenciamento.....	35
Personalização.....	36
Pequenos modelos de linguagem	36
Conclusão	37

CAPÍTULO 2: A lacuna da avaliação.....38

A natureza não determinística dos LLMs	39
Fonte do não determinismo	39
Temperatura	40
Testes de temperatura no 10-K de uma LLMBAs	41
O desafio das avaliações	44
Projetando um framework de avaliação de LLMBAs	45
Design conceitual: uma única LLMBAs	45
Design conceitual: várias LLMBAs.....	46
Metodologias de avaliação	48
Métricas quantitativas	48
Codificando o BLEU e o ROUGE	50
Codificando um LLM como juiz	58
Limitações do LLM como juiz	63
Avaliando os avaliadores.....	64
Uma breve introdução aos benchmarks de LLM	68
O ARC-AGI e o prêmio.....	69
Conclusão.....	72

CAPÍTULO 3: Ferramentas de avaliação para aplicações baseadas em LLMs.....73

LangSmith	74
BLEU com o LangSmith	74
Criando um dataset padrão-ouro	76
Calculando pontuações BLEU.....	77
Gere um resumo.....	78
Execute a avaliação.....	79
Ampliando o LLM como juiz com o LangSmith.....	84
Promptfoo	87
Avaliando modelos com o Promptfoo	88
Avaliando prompts	92
LightEval.....	95
Configuração do LightEval.....	96
Avaliação econométrica com o MMLU	99
Comparando múltiplos modelos em econometria.....	99
Comparação de frameworks	101
Conclusão	102

CAPÍTULO 4: Dos dados ao contexto	103
Pré-processando documentos.....	105
PyPDF2	105
Markdown	106
Docling.....	106
RAG.....	107
Processo do RAG	108
Chunking de documentos	109
Estratégias de chunking	109
Estudo de caso de chunking	110
Gerando conteúdo longo.....	111
Etapa 1: chunking do conteúdo	112
Etapa 2: mantendo o contexto com um prompt base.....	115
Etapa 3a: opção 1 — processamento sequencial	117
Etapa 3b: opção 2 — construindo parâmetros de prompt dinâmico	122
Etapa 4: combinando a saída – geração do relatório	124
Etapa 5: aplicar ao 10-K.....	125
Relatório final	126
Nossa avaliação do resumo	128
Embeddings vetoriais	130
Bancos de dados vetoriais	131
Busca vetorial.....	132
Reclassificação (reranking).....	136
Geração de relatório	138
Desafios e limitações do RAG.....	140
Modelos de contexto longo: o fim do RAG?.....	142
Estudo de caso de LCM: geração de quiz com citações.....	144
Implementação.....	145
Prompting de Corpus-in-Context.....	146
Cache de prompt	148
Geração do quiz.....	148
Exemplo de uso	149
Discussão e limitações.....	152
Conclusão	152
 CAPÍTULO 5: Geração de saídas estruturadas	 154
Por que a saída estruturada importa.....	155
Melhorando a eficiência e o fluxo de trabalho do desenvolvedor.....	156
Atendendo a requisitos de UI e de produto	156
Melhorando a confiança e a experiência do usuário.....	157
Por que é difícil? A arquitetura Transformer	157
Técnicas de restrição de tempo de treinamento	159

Técnicas de restrição de tempo de inferência	159
Engenharia de prompt.....	160
Modo JSON (ajustado).....	162
Combinando o modo JSON com o Pydantic.....	164
Pós-processamento de logits	166
Outlines.....	171
Ollama	175
Ollama com o Pydantic.....	177
LangChain.....	179
Comparando ferramentas.....	181
Conclusão	182
CAPÍTULO 6: Considerações de segurança dos LLMs	183
Riscos gerais de segurança da IA.....	184
Riscos de segurança específicos dos LLMs	185
Jailbreaking	185
Injeção de prompt (prompt crafting)	185
Edição furtiva	186
Considerando os riscos de segurança das LLMBAs	187
Prevenção contra desinformação.....	187
Prevenção contra conselhos não qualificados.....	188
Detecção de viés	189
Proteção de privacidade.....	189
Diretrizes de laboratórios de IA e centros de políticas.....	190
OpenAI.....	190
Anthropic	191
Google	192
Benchmark de segurança de IA da MLCommons	193
Critérios do Centre for the Governance of AI	195
Duas abordagens específicas para a segurança de LLMs	196
Red Teaming.....	196
IA Constitucional.....	198
Redução de danos por meio da autocrítica.....	198
Equilíbrio entre utilidade e segurança.....	199
Melhoria da transparência e da escalabilidade.....	200
Conclusão	200
CAPÍTULO 7: Avaliando a segurança dos LLMs.....	201
Avaliação de segurança com datasets de benchmark	203
SALAD-Bench.....	203
TruthfulQA	209
HarmBench.....	219
Guardrails e moderação no tempo de execução.....	224
NeMo Guardrails	226

Guardrails do TruLens	227
Llama Guard	229
API de moderação da Mistral	232
API de Moderação da OpenAI	233
Moderação personalizada com LLM como juiz	235
Estudo de caso: implementando um filtro de segurança para alunos do ensino fundamental e médio	236
Dataset de avaliação	236
Amostras ruins	237
Amostras boas	239
Filtros de segurança	242
Usando a API de Moderação da Mistral	243
Usando a API de Moderação da OpenAI	244
Usando um validador personalizado baseado em juiz	245
Benchmarking	247
Conclusão	252
CAPÍTULO 8: Alinhamento do LLM: um estudo de caso	253
Por que o alinhamento é necessário?	254
RLHF: um avanço decisivo que tornou os LLMs mais úteis e controláveis	255
Otimização de política proximal	257
DPO	258
Estudo de caso: alinhando um LLM a uma política	260
Ferramentas de alinhamento	261
Política educacional da Acme	262
Dataset de preferências — geração de dataset sintético	264
Prompts de usuário	266
Respostas rejeitadas	269
Respostas escolhidas	273
Geração de dataset de DPO	276
Otimização baseada em DPO	278
Preparação dos dados	278
Ajuste fino	279
Escolhas para o DPO e o ajuste fino	280
Ajuste fino em ação	281
Resultados do treinamento	282
Opção de salvar os resultados na Hugging Face	284
Vibe check	285
Avaliação do alinhamento: LLM como juiz	286
Amostra de prompts e respostas	288
Chamando o LLM juiz	289
Testando o LLM como juiz	291
Sucesso?	293
Composição do dataset de DPO	295

A escolha do modelo base	296
A metodologia de avaliação.....	296
O processo de ajuste fino e a escolha de parâmetros	297
Projetando um plano de segurança e alinhamento.....	297
Fase 1: definição de políticas	297
Fase 2: Pesquisa de usuários e identificação de riscos.....	298
Fase 3: Framework de avaliação	299
Fase 4: design da arquitetura de segurança	300
Fase 5: implementação e seleção de ferramentas.....	300
Fase 6: resposta a incidentes	301
Armadilhas comuns	302
Conclusão	303
CAPÍTULO 9: Epílogo: LLMBAs na era da queda dos custos de inferência	305
APÊNDICE: Ferramentas para implantação local de LLMs	308
Ollama.....	309
llama.cpp	312
GPT-Generated Unified Format.....	313
Exemplo do llama.cpp.....	314
llama-server	315
llamafire	318
Quantização	319
Perplexidade e divergência KL.....	321
Dataset de prompts.....	322
Exemplo de quantização	323
Resultados	324
Ferramentas locais, poder real	327
Índice remissivo	331