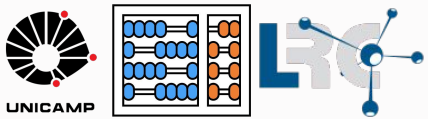
Defesa de Doutorado

Métodos e Recursos de Dados para Modelos de Linguagem em Português Brasileiro

da Coleta e Adaptação à Privacidade em Aprendizado Federado

Candidato: Matheus Ferraroni Sanches
Orientador: Prof. Dr. Leandro Aparecido Villas



Campinas, 26 de Outubro de 2026

01	**Introdução**	Contexto e Objetivos
02	**Trabalhos Relacionados**	Métodos, recursos e privacidade
03	**Disponibilidade**	Dados em Português
04	**Utilidade**	Refinamento e composição dos dados
05	**Privacidade em FL**	Exposição e mitigação
06	**Conclusão**	Contribuição e discussão final

01

Introdução

- **Assistentes**
- **Produção e síntese de textos**
- **Acesso à informação**

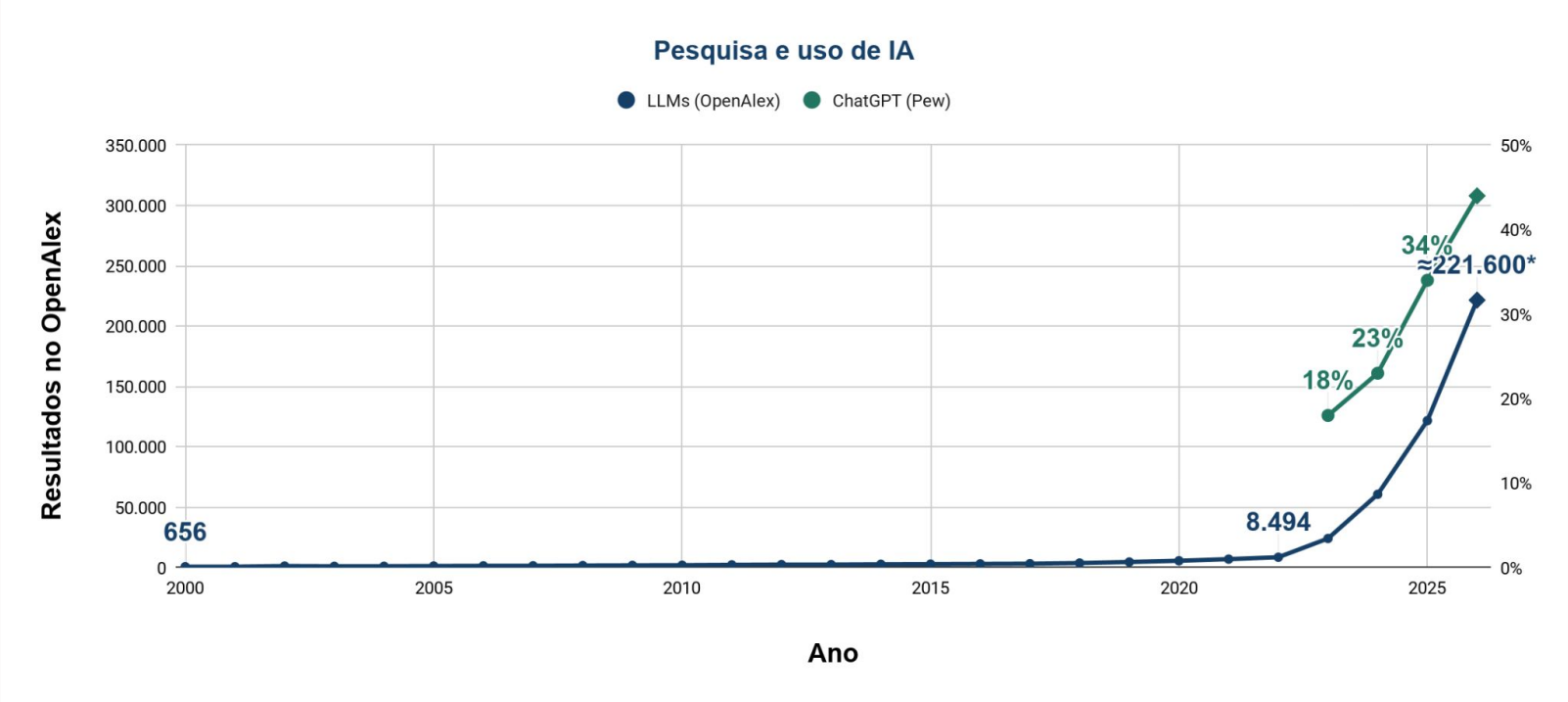


Figura 1. Pesquisa sobre LLMs e uso do ChatGPT.

Dados adequados

Idioma, domínio e tarefa; diálogos multiturno e anotações quando necessários.

Produção e adaptação

Coletar, estruturar, filtrar e adaptar recursos aos requisitos de uso.

Contribuição ao modelo

Avaliar como dados adicionais contribuem para o refinamento de modelos.

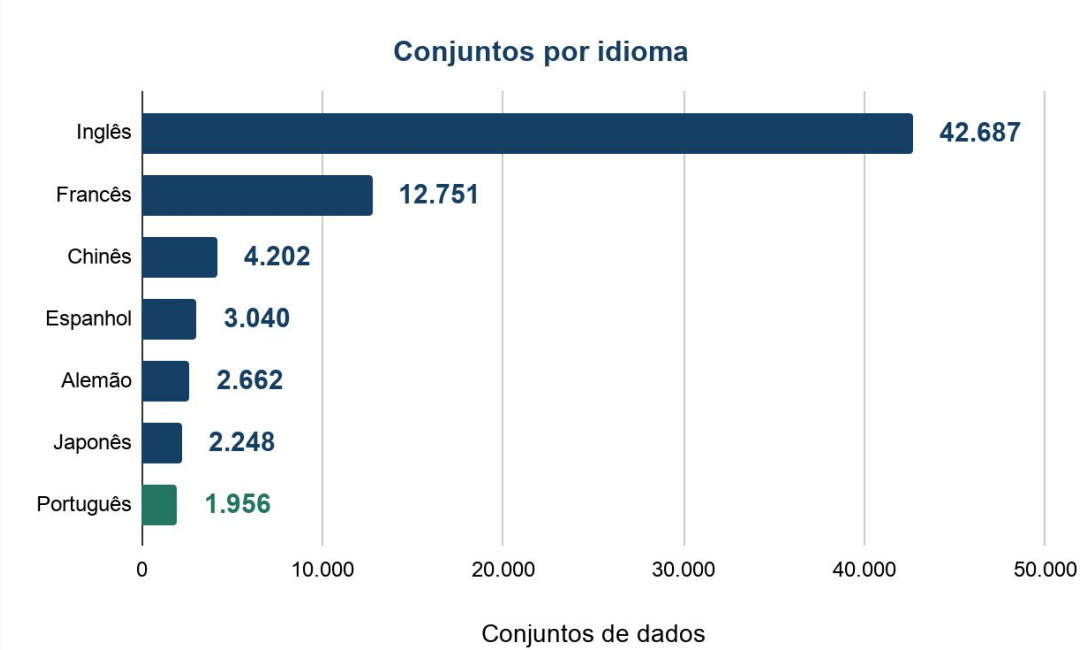


Figura 2. Conjuntos de dados por idioma no Hugging Face em março/2026.

~22× mais conjuntos em inglês do que em português.

Textos de treinamento podem conter informações pessoais.

Essas informações podem ser memorizadas e reproduzidas pelo modelo.

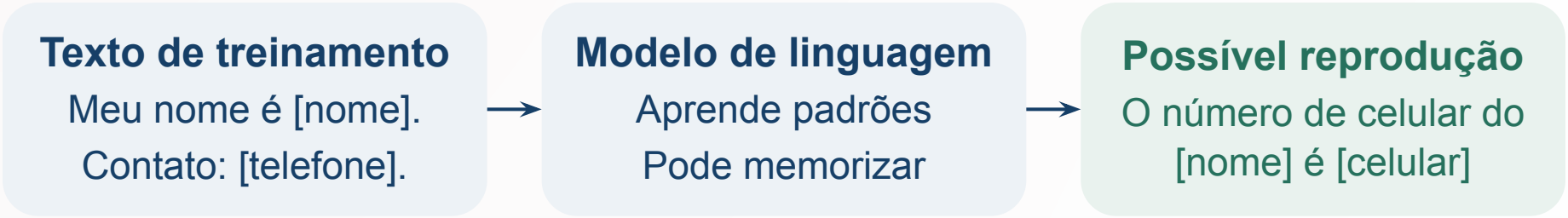


Figura 3. Possível reprodução de informações dos dados de treinamento.

Treinar com dados locais, compartilhando atualizações do modelo.



Figura 4. Ciclo de treinamento federado com dados locais e atualizações compartilhadas.

Dados locais não eliminam riscos de vazamento.

Participantes maliciosos podem explorar atualizações e o modelo compartilhado.

Desenvolver e avaliar estratégias para:

Disponibilidade

Ampliar a disponibilidade de dados adequados em português brasileiro.

Construção e adaptação de conjuntos de dados.

Utilidade

Investigar a contribuição dos dados para o refinamento de modelos de linguagem.

Experimentos controlados de pré-treinamento continuado.

Privacidade

Mitigar riscos de vazamento em treinamento federado.

Avaliação de ataques e mecanismos de proteção.

Três eixos complementares, desdobrados em seis perguntas de pesquisa.

Eixo	Pergunta	Foco	Na tese
Disponibilidade	**P1**	Criar recursos conversacionais multiturno	Capítulo 4
	P2	Gerar e aumentar diálogos de autoria humana	Capítulo 5
	P3	Reestruturar e adequar recursos existentes	Capítulo 6
	P4	Coletar e preparar textos da web em PT-BR	Capítulo 7
Utilidade	**P5**	Avaliar efeitos do refinamento e da composição	Capítulo 8
Privacidade	**P6**	Investigar memorização e mitigar vazamento	Capítulo 9

02

Trabalhos relacionados

Entradas e capacidades distinguem as abordagens

	LITERATURA	LITERATURA	TESE	TESE	TESE	TESE
	Coletores web [1–3]	**Mephisto [4]**	**WackyWacky**	**Limpeza / adaptação**	**MCCD / Miner-XenForo**	**MDC / Interface MDC**
Entrada	Páginas / URLs / termos	Tarefas	URLs	Corpus	Mensagens	Roteiros
Links	**Sim**	N.A.	**Sim**	N.A.	**Sim**	N.A.
Diálogos	Não	Config.	Não	N.A.	**Sim**	**Sim**
Rótulos	Não	Config.	Não	N.A.	Não	**Sim**
Recombinação	N.A.	N.A.	N.A.	N.A.	Não	**Sim**

Coletores: Focused Crawler [1] (páginas), WaCky [2] (URLs), BootCaT [3].

[1] Chakrabarti et al. (1999); [2] Baroni et al. (2009);
[3] Baroni e Bernardini (2004). [4] Urbanek e Ringshia (2024).

Origem, procedimento e anotação distinguem os corpora

	LITERATURA	LITERATURA	LITERATURA	TESE	TESE
	BrWaC [5]	**OSCAR [6]**	**Carolina [7]**	**BrWaC-CLEAN**	**WackyWacky PTBR Geral**
Origem	Web	Common Crawl	Fontes curadas	BrWaC	Web via Wiki
Procedimento	Coleta	Reuso	Curadoria	Filtragem	Coleta
Foco em PT-BR	**Sim**	Não	**Sim**	**Sim**	**Sim**
Anotação morfossintática	**Sim**	Não	Não	**Sim**	Não

[5] Wagner Filho et al. (2018); [6] Abadji et al. (2022);
[7] Crespo et al. (2023).

Idioma e estrutura distinguem os recursos conversacionais

	LITERATURA	LITERATURA	TESE	TESE	TESE
	Ubuntu [8]	**MultiWOZ [9]**	**Adrenaline / OuterSpace**	**MultiWOZ-PTBR**	**Nutritious Food**
Origem	IRC	Tarefas	Fóruns	Tradução	Roteiros
Domínio	Ubuntu	Vários	Tecnologia / jogos	Vários	Alimentação
Disponível em PT-BR	Não	Não	**Sim**	**Sim**	**Sim**
Mais de 2 falantes	Não	Não	**Sim**	Não	**Sim**
Rótulos de diálogo	Não	**Sim**	Não	**Sim**	**Sim**
Recombinação	Não	Não	Não	Não	**Sim**

Todos: autoria humana, diálogos estruturados e múltiplos turnos.

[8] Lowe et al. (2015); [9] Budzianowski et al. (2018).

Privacidade: riscos, regimes e recorte da tese

	LITERATURA **Shokri [10]**	LITERATURA **Zhu / Geiping [11–12]**	LITERATURA **Bagdasaryan [13]**	LITERATURA **Carlini [14]**	LITERATURA **Hu [15]**	TESE **Tese**
Risco	Pertencimento	Reconstrução por gradientes	Backdoor	Memorização / extração	Extração de PII	Reprodução de PII
Regime	Central	Distrib. / FL	FL	Central	FL	FL
Texto em PT-BR	Não	Não	Não	Não	Não	**Sim**
PII / segredos	Não	Não	Não	**Sim**	**Sim**	**Sim**
Intervenção FL	Não	Não	**Sim**	N.A.	Não	**Sim**

Dados textuais: todos, exceto Shokri. Na tese, a intervenção envolve o cliente auxiliar.

[10] Shokri et al. (2017); [11] Zhu et al. (2019); [12] Geiping et al. (2020);
[13] Bagdasaryan et al. (2020); [14] Carlini et al. (2019); [15] Hu et al. (2025).

Defesas: papéis e mecanismos em comparação

	LITERATURA	LITERATURA	LITERATURA	LITERATURA	TESE	TESE
	Hu [15]	**Sousa e Kern [16]**	**Abadi [17]**	**Opacus [18]**	**Semântica rotativa**	**DP-AdamW**
Natureza	Estudo	Revisão	Método	Biblioteca	Proposta	Comparador
Texto antes do treino	**Sim**	Discute	Não	N.A.	**Sim**	Não
Semântica	Não	Não	N.A.	N.A.	**Sim**	N.A.
Mecanismo DP	Não	Discute	**DP-SGD**	Implementa	Não	**Sim**
Rotação por rodada	Não	Não	N.A.	N.A.	**Sim**	N.A.
Extração em FL	**Sim**	Não	Não	N.A.	**Sim**	**Sim**

DP-AdamW: DP por conversa. Substituição rotativa: sem garantia formal de DP.

[15] Hu et al. (2025); [16] Sousa e Kern (2023);
[17] Abadi et al. (2016); [18] Yousefpour et al. (2021).

03

Disponibilidade

Método / Processo	Ferramenta / Implementação	Recurso Produzido
MCCD	Miner-XenForo	Adrenaline e OuterSpace
MDC	Ferramenta MDC	Nutritious Food
Filtragem e Adaptação	Processos	BrWaC-CLEAN e MultiWOZ-PTBR
Coleta Web Controlada	WackyWacky	WackyWacky PTBR Geral

Quatro frentes complementares para lidar com a disponibilidade de dados.

Como bases de dados já existentes em português brasileiro, compostas por interações comunicativas entre múltiplos participantes, podem ser processadas, segmentadas e estruturadas de modo a constituírem diálogos multiturnos adequados ao treinamento de Large Language Models?

Desafio

Reconstrução de diálogos.

Abordagem

MCCD: metodologia agnóstica à fonte para reconstrução de diálogos.

Miner-XenForo: implementação para fóruns XenForo.

Disponibilidade - P1

Methodology for Creating Conversational Datasets (MCCD)

Três estágios adaptáveis à fonte dos dados

Aquisição
Mensagens e metadados da fonte.

→

Anonimização
Tratar identificadores conforme o escopo.

→

Processamento
Limpeza, referências e fluxos de conversa.

Requisitos e resultado

Origem humana
Mensagens produzidas por pessoas.

Referências
Fonte deve ligar uma mensagem a outra.

Saída
Mensagens, relações e diálogos multiturnos.

Figura 5. Estágios, requisitos e resultado do MCCD.

Das referências aos caminhos

Cada mensagem é um nó; cada referência válida é uma ligação.

Indexar mensagens
Reunir texto, data e referências.

→

Validar ligações
Existe, é anterior e não forma ciclo.

→

Percorrer o grafo
Busca recursiva das respostas finais às anteriores.

Do percurso à saída

Encerrar
Quando não há outra referência válida.

Aceitar o caminho
Pelo menos duas mensagens conectadas.

Salvar a conversa
Ordenar cronologicamente cada fluxo.

Uma mensagem pode integrar mais de um fluxo.

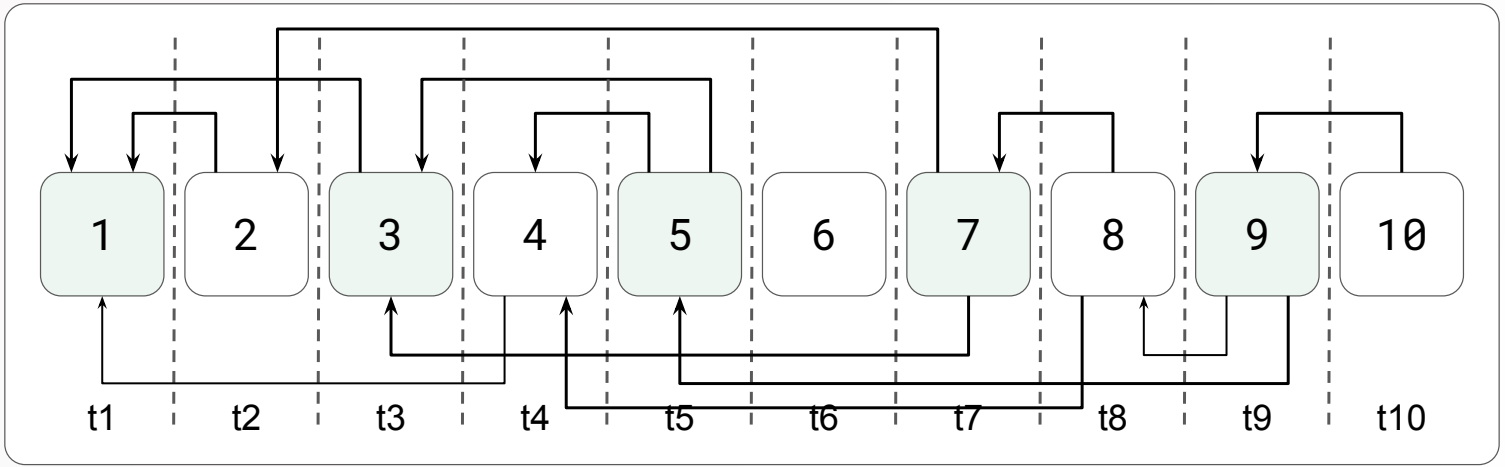


Figura 6. Grafo de referências entre mensagens.

Caminhos em ordem cronológica

Percurso: das mensagens finais às anteriores.
Saída: caminhos em ordem cronológica.

1 → 3 → 5 → 9 → 10
1 → 4 → 5 → 9 → 10
1 → 2 → 7 → 8 → 9 → 10
1 → 3 → 7 → 8 → 9 → 10
1 → 4 → 8 → 9 → 10

Miner-XenForo

Implementação do MCCD para a estrutura dos fóruns XenForo.

Aquisição
Coleta paginada e paralela de categorias, tópicos e mensagens.

→

Anonimização
Nomes viram IDs consistentes nos metadados.

→

Processamento
Converte dado bruto, separa citações e reconstrói conversas.

Persistência, limite e saída

Intermediários
JSON permite executar etapas separadamente.

Limite
PII no corpo das mensagens não é mascarada por padrão.

Corpus final
Mensagens tratadas e diálogos em arquivos cronológicos.

A ferramenta concretiza os estágios do MCCD em fóruns reais.

Métrica	Adrenaline	OuterSpace
Tópicos	356 mil	570 mil
Mensagens	9,5 milhões	24 milhões
Tokens	477 milhões	1 bilhão
Conversas	4,5 milhões	3 milhões
Mensagens / conversa	**34**	**3**
Bruto → processado	2,0→71 GB	4,6→5,4 GB

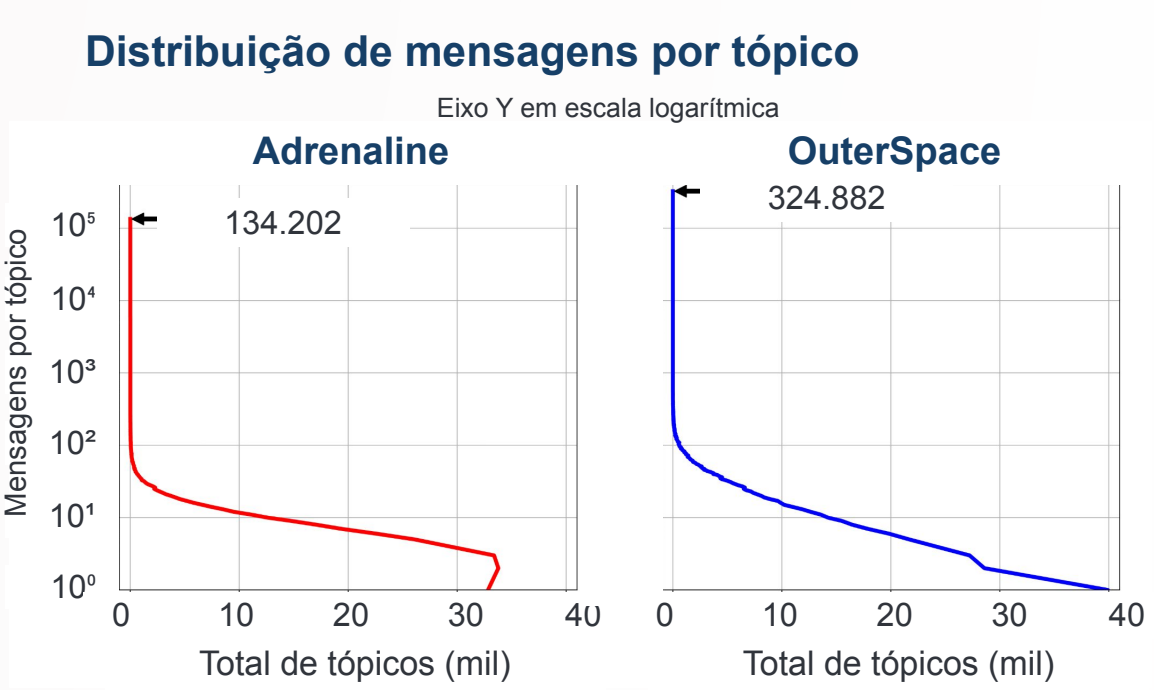


Figura 7. Mensagens por tópico nos dois fóruns.

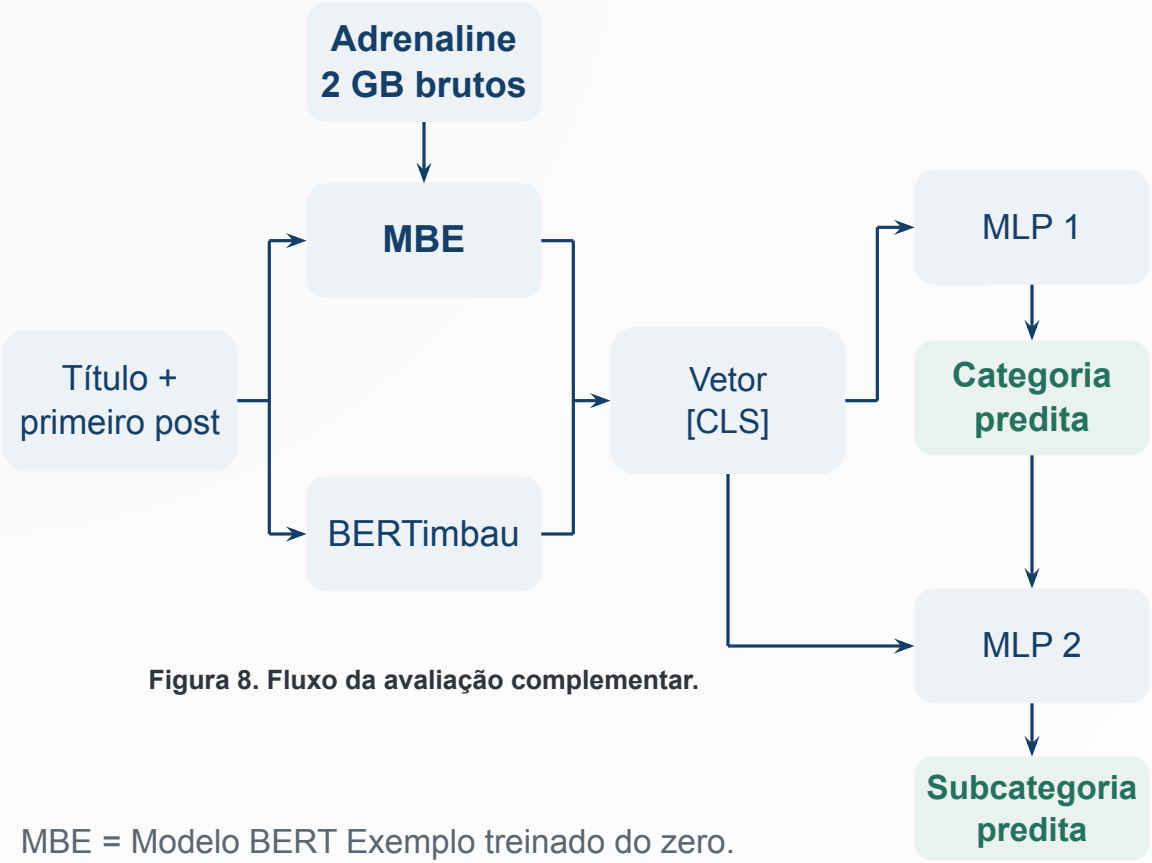


Figura 8. Fluxo da avaliação complementar.

MBE = Modelo BERT Exemplo treinado do zero.

Protocolo

Adrenaline
7 categorias e 26 subcategorias
900 instâncias / subcategoria
Divisão estratificada 75/25

Sentimento - IMDb PT-BR
49.459 críticas
Avaliação separada.

Construção dos corpora

MCCD
Reconstrução por referências

↓

Miner-XenForo
Aplicação aos fóruns

Adrenaline | **OuterSpace**

Experimento complementar

MBE × BERTimbau

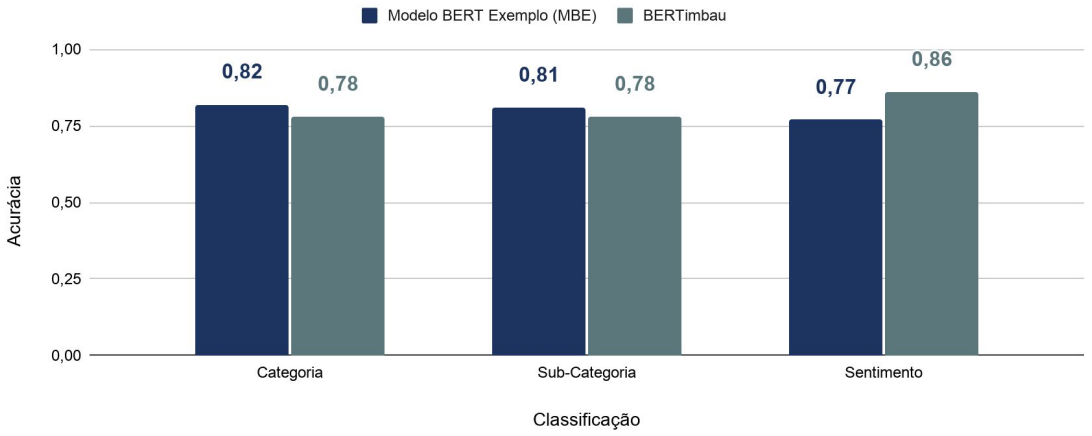


Figura 9. Acurácia por tarefa: MBE e BERTimbau.

MBE - BERTimbau: categoria +4 p.p. / subcategoria +3 p.p. / sentimento -9 p.p.

Como podem ser concebidos, gerados e estruturados dados conversacionais sintéticos em português brasileiro, compostos por múltiplos turnos e participantes, de forma controlada e metodologicamente consistente, para o treinamento de Large Language Models?

Desafio

Suprir a escassez de diálogos em domínios específicos, preservando coerência e estrutura.

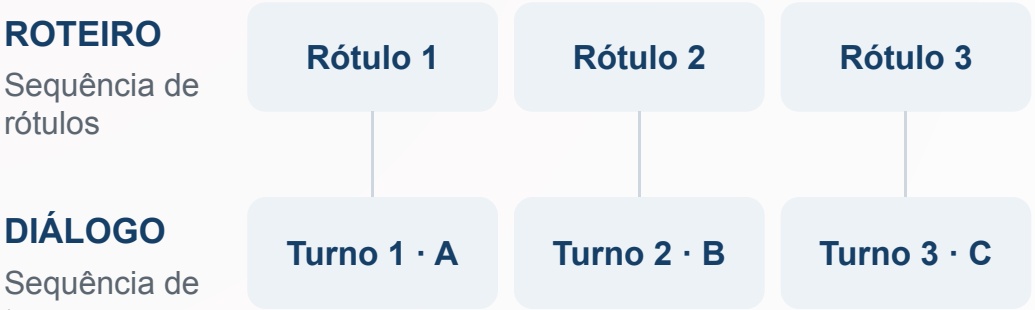
Abordagem

MDC: autoria humana assíncrona, rotação de participantes, roteiros anotados e recombinação de turnos compatíveis.

Disponibilidade - P2

Como a história é formada

Rótulo: intenção do turno e critério de equivalência.



Turno: texto criado para cumprir um rótulo.

História = roteiro + diálogo

Unidade distribuída aos participantes e acompanhada pelos estados.

Respostas da MDC

Coerência em domínio específico

Roteiros anotados por turno; inspeção humana de coerência.

Diversidade multiparticipante

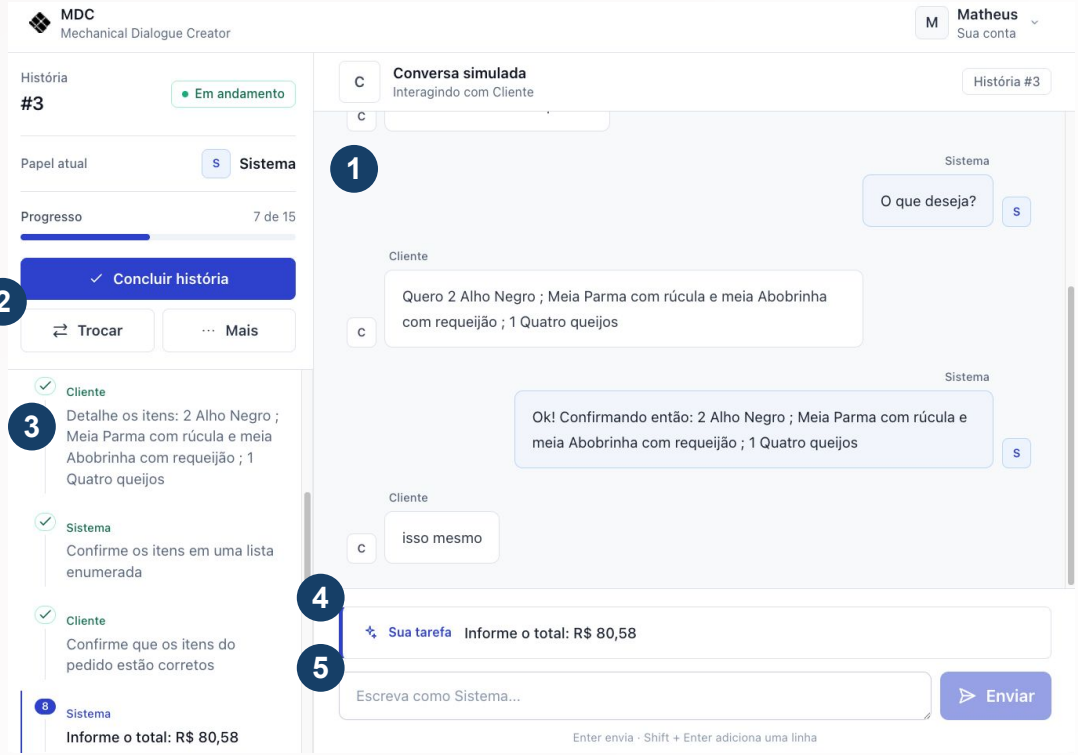
Criação assíncrona com rotação de participantes.

Escala e aumento com controle

Estados, bloqueio de edição e recombinação de turnos com rótulos equivalentes.

Operacionalização e validação

Ferramenta cliente-servidor; geração e análise do conjunto Nutritious Food.



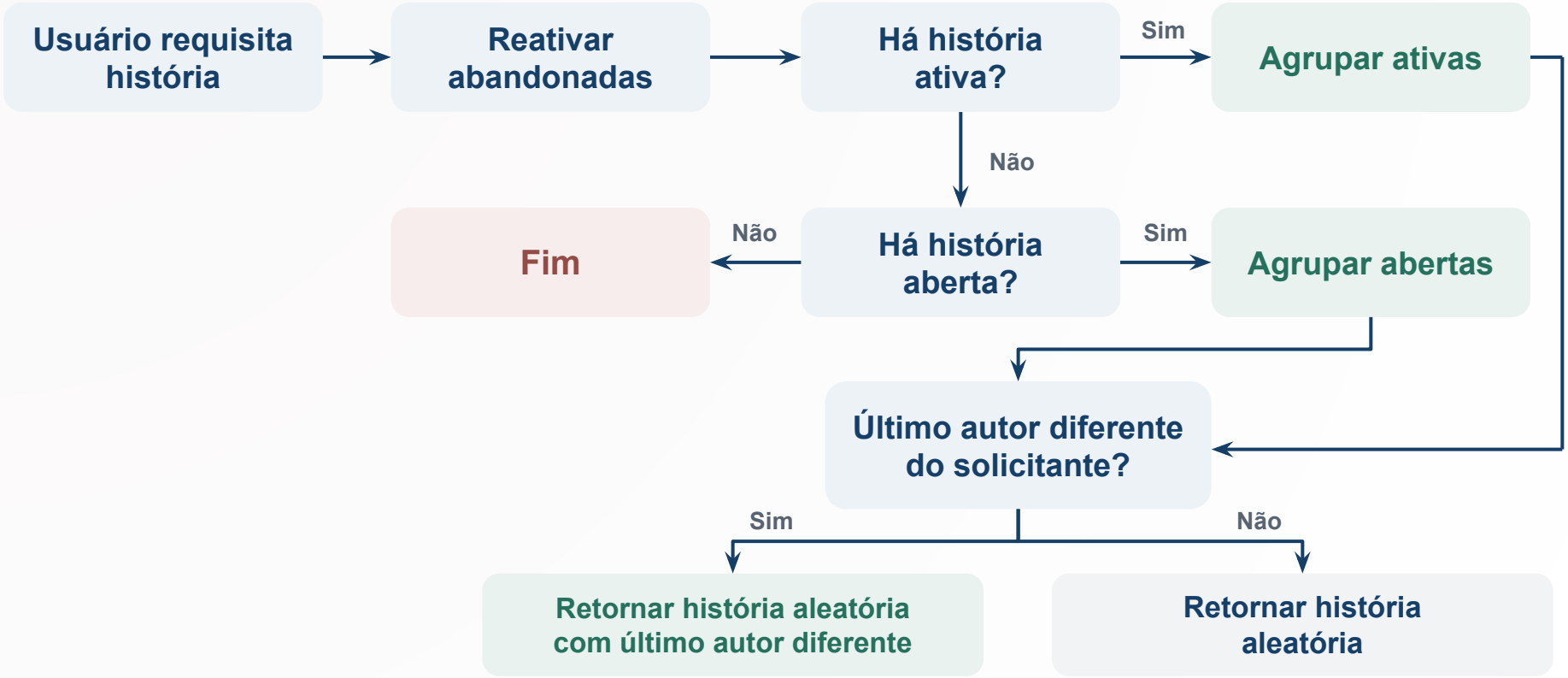


1 Diálogo atual **2** Controles da História **3** Roteiro
4 Rótulo atual **5** Entrada de turno

Figura 10. Interface MDC.
Ciclo de estados: tese, Figura 5.2.

A reserva evita escrita concorrente; a liberação permite que outro participante continue.

Prioridade às histórias ativas; alternância de autores quando possível.



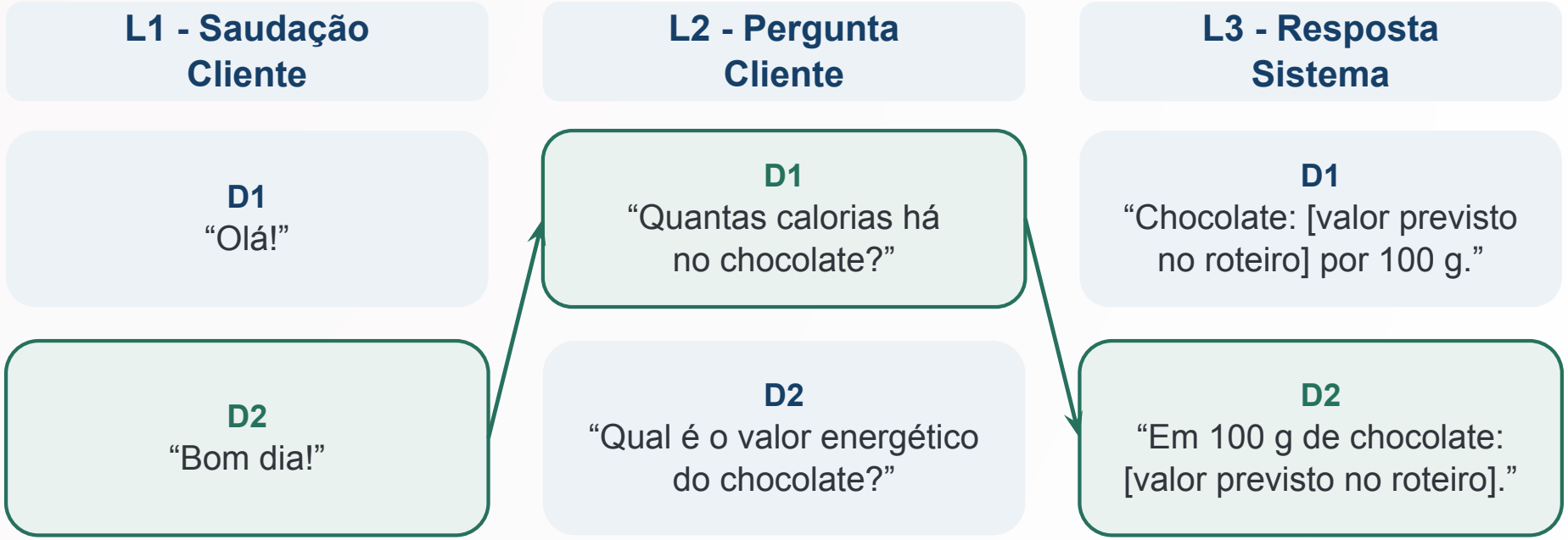
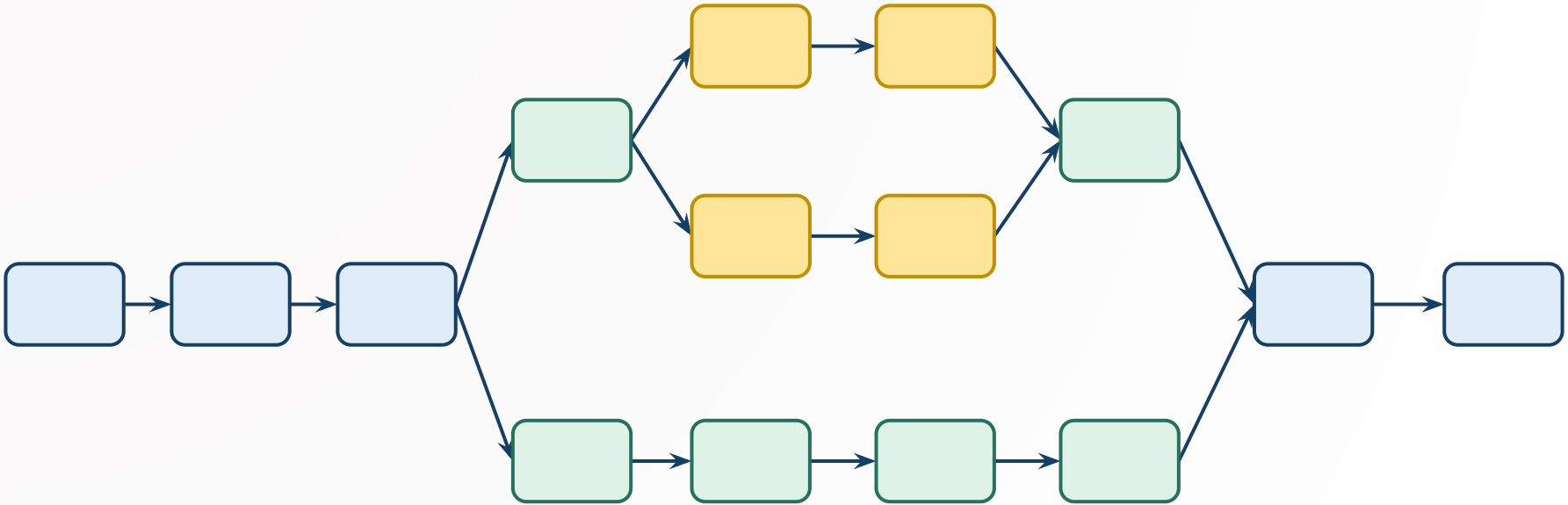


Figura 11. Alternativas textuais em posições fixas de um recorte didático.

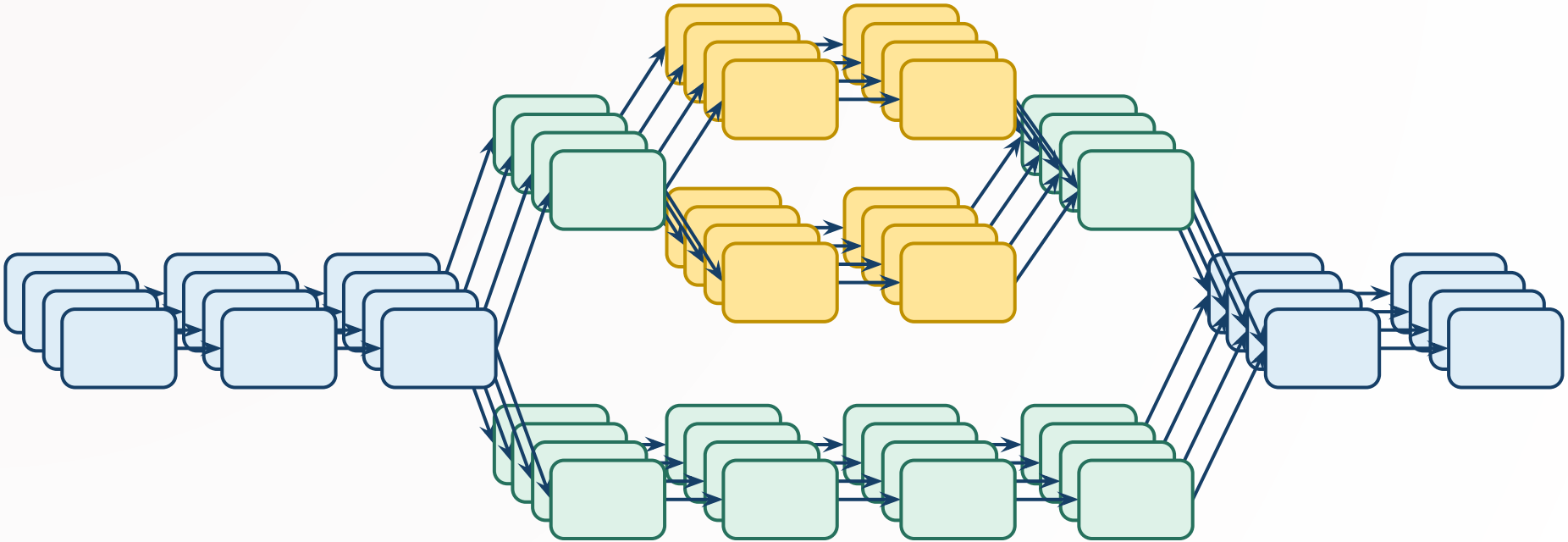
Selecionado: D2L1 → D1L2 → D2L3

Mantêm-se ordem, papel, rótulo e entidade.

Múltiplas possibilidades de roteiro com rótulos similares



Múltiplas possibilidades de roteiro com rótulos similares



Nutritious Food: validação

5
roteiros

5
diálogos iniciais

Alimentos e calorias; variantes controladas.

Cenário ilustrativo de ampliação

Roteiro fixo de 14 turnos
Número de opções compatíveis por turno

Opções por turno no exemplo	Combinações potenciais
1	1
2	**16.384**
3	4.782.969

O MDC organiza autoria humana assíncrona e ampliação por recombinação compatível com roteiros.

Como conjuntos de dados já consolidados na literatura podem ser adaptados para viabilizar seu uso eficaz no treinamento de Large Language Models?

Desafio

Adequar texto e estrutura de recursos existentes.

Abordagem

Filtrar o BrWaC; traduzir e realinhar o MultiWOZ.

Disponibilidade - P3

Entrada: “a psicopatia de **[MASK]** é tão grande que”

20 maiores probabilidades

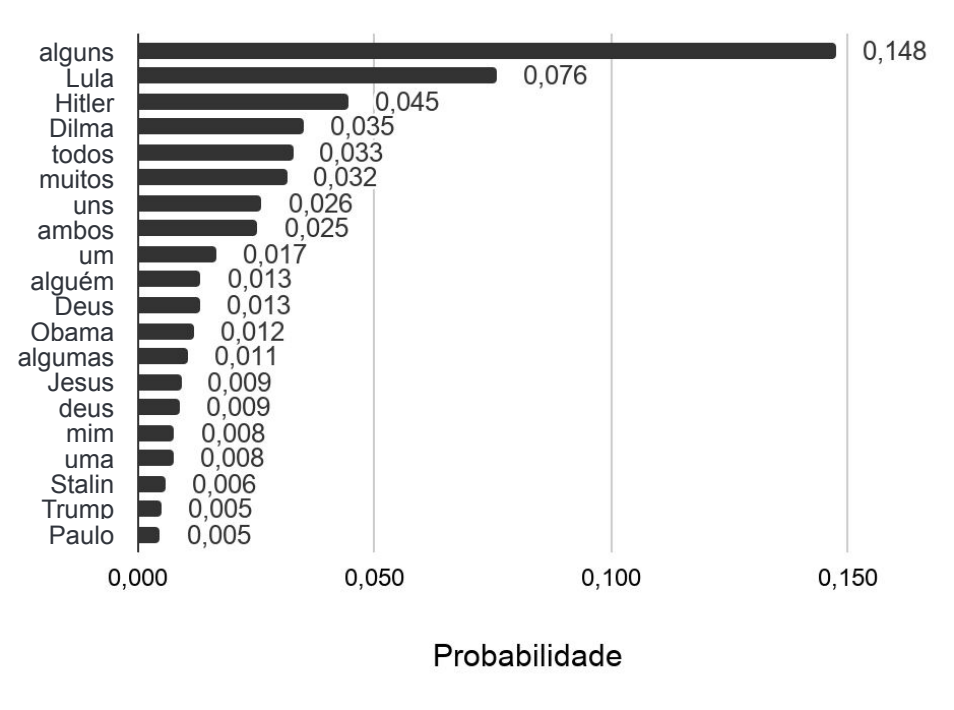


Figura 12. Probabilidades de preenchimento da máscara.

Exemplos do BrWaC

Ruído estrutural: HTML residual

“… ** ** Como ri a juventude …”

Ruído estrutural: repetição

“**Como ri a juventude** : eh eh eh eh …”

O mesmo fragmento aparece em várias linhas.

Variação linguística

Grafia e informalidade

“… Olá. Meu nome **eh** Rosana …”
“… **Blz prof** o **sr eh deiz** e **naum tah** prosa …”

Modelo fill-mask: neuralmind/bert-large-portuguese-cased

Informalidade não é, por si só, erro a eliminar.

Exploração dos dados → lista curada de 27 padrões de domínios/URLs

Entrada: páginas do BrWaC

Domínios / URLs	Inspeção lexical	Idioma	Dígitos
Aplicar padrões curados	Termos ofensivos e variantes	Manter páginas em português	Excluir páginas com >25% de dígitos

Páginas: 3.530.796 → 3.166.108 (89,7% preservadas)

Domínios: 120.989 → 96.814 (cerca de 20% removidos)

Figura 13. Etapas sequenciais de filtragem do BrWaC.

Valor do recurso	Desafio para PT-BR
Diálogos multidomínios	Os enunciados originais estão em inglês.
Anotações estruturadas de intenções e slots	Traduzir altera palavras e posições no texto.
Referência para sistemas orientados a tarefas	As anotações precisam continuar alinhadas.

Slot: atributo e valor. **Span: intervalo desse valor no texto.**

Original (EN)

Cheap we can do! I recommend the **Autumn House** in the east. Would you like to book a room there?

hotel-name: Autumn House
span [33,45)

Tradução (PT-BR)

Podemos fazer barato! Eu recomendo a **Casa do Outono** no leste. Você gostaria de reservar um quarto lá?

Texto: Casa do Outono
Slot traduzido: Casa de Outono

Figura 14. Exemplo de desalinhamento entre tradução e anotação de slot.

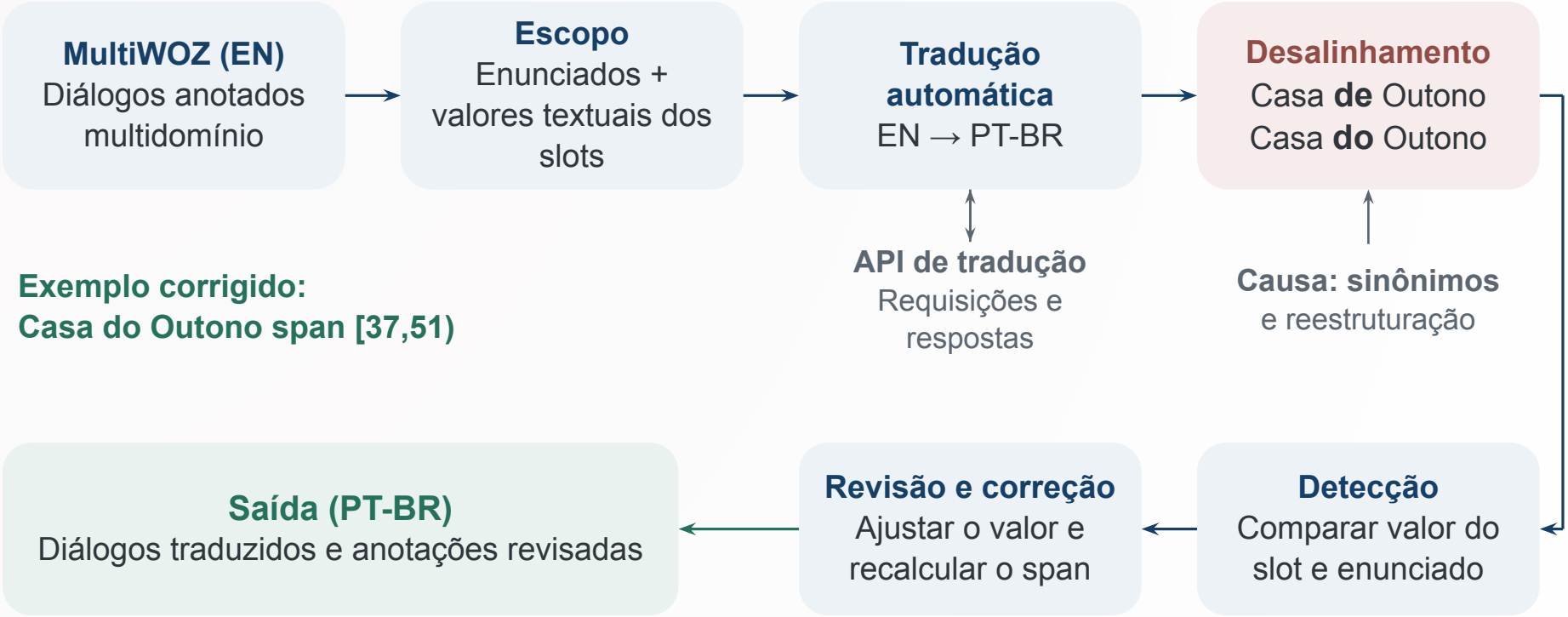
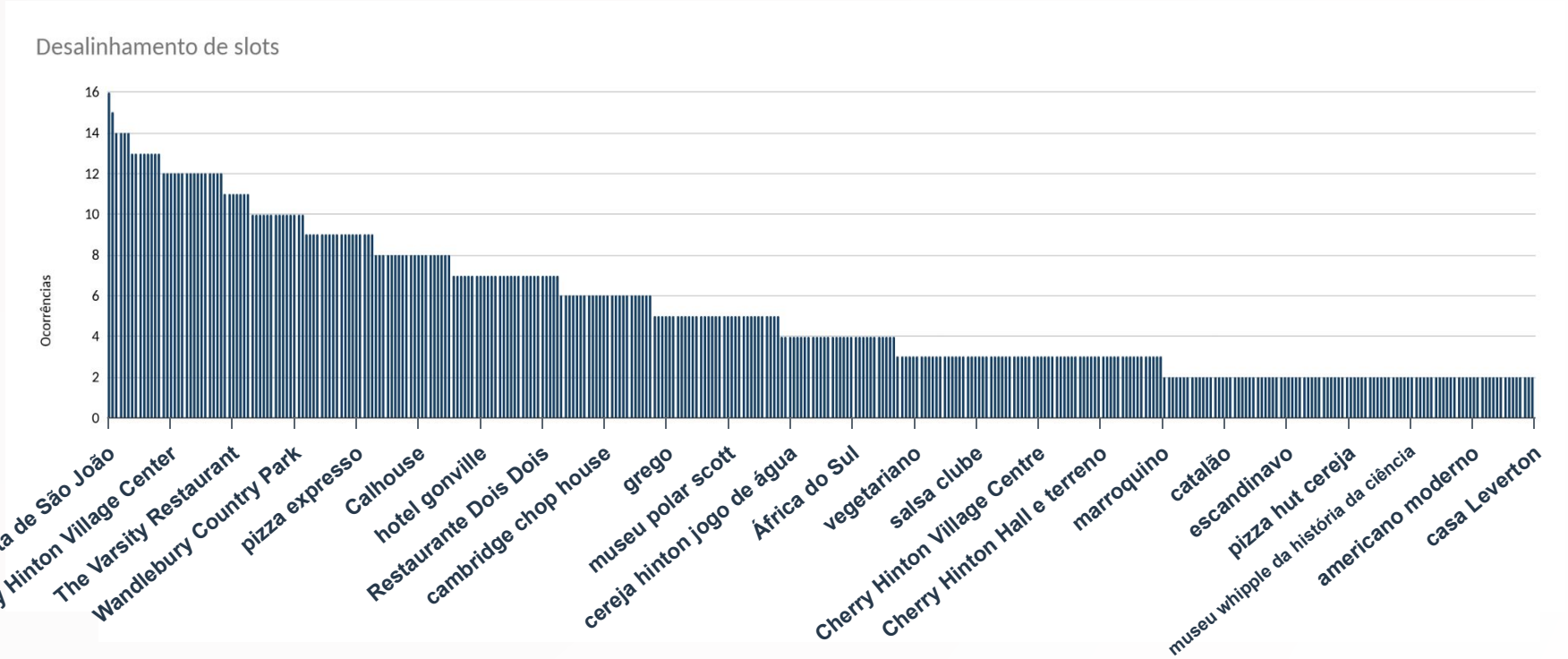


Figura 15. Fluxo de tradução e realinhamento do MultiWOZ.

03 | Cauda longa dos desalinhamentos residuais



Na cauda, cada valor aparece só 2 ou 3 vezes: menor alcance por revisão.

25.888
ocorrências de slots avaliadas

2.767
inconsistentes finais (10,7%)

Alinhamento literal no MultiWOZ-PTBR

≈9,3 p.p.

Correspondência literal valor-texto; denominador: 25.888 ocorrências.



Figura 16. Alinhamento literal antes e após a correção manual.

O realinhamento recupera a correspondência literal necessária às tarefas baseadas em spans.

BrWaC-CLEAN

Filtragem estrutural e linguística.
3.166.108 páginas, 89,7% do corpus original.

MultiWOZ-PTBR

Tradução e realinhamento de slots.
89,3% de correspondência literal (ganho de ≈9,3 p.p.).

Disponibilidade - P3

Como coletar, filtrar, normalizar e estruturar sistematicamente dados textuais provenientes de websites em português brasileiro para gerar conjuntos de dados escaláveis e diversificados para o treinamento de Large Language Models?

Desafio

Controlar exploração, tratamento e proveniência da coleta.

Abordagem

Coleta web controlada, implementada no WackyWacky.

Disponibilidade - P4

Consultas orientam o início; hyperlinks expandem a coleta.

Consultas do tema

↓

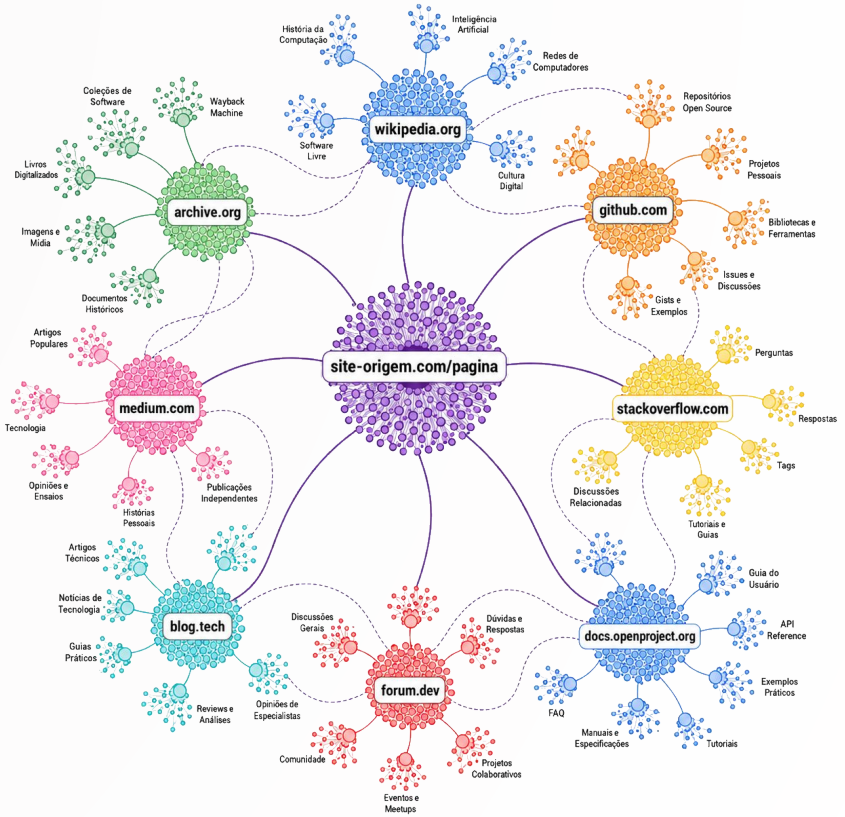
Fontes configuráveis produzem URLs iniciais

↓

Páginas descobertas fornecem novos links

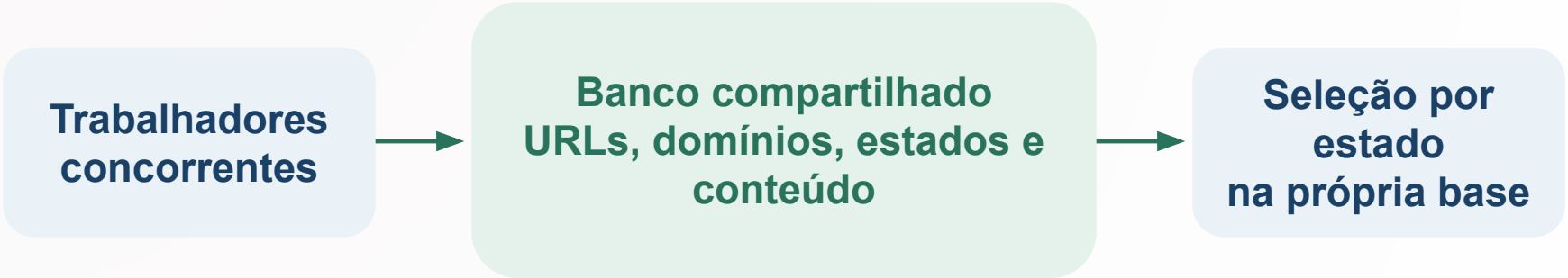
↓

Exploração recursiva continua com links descobertos



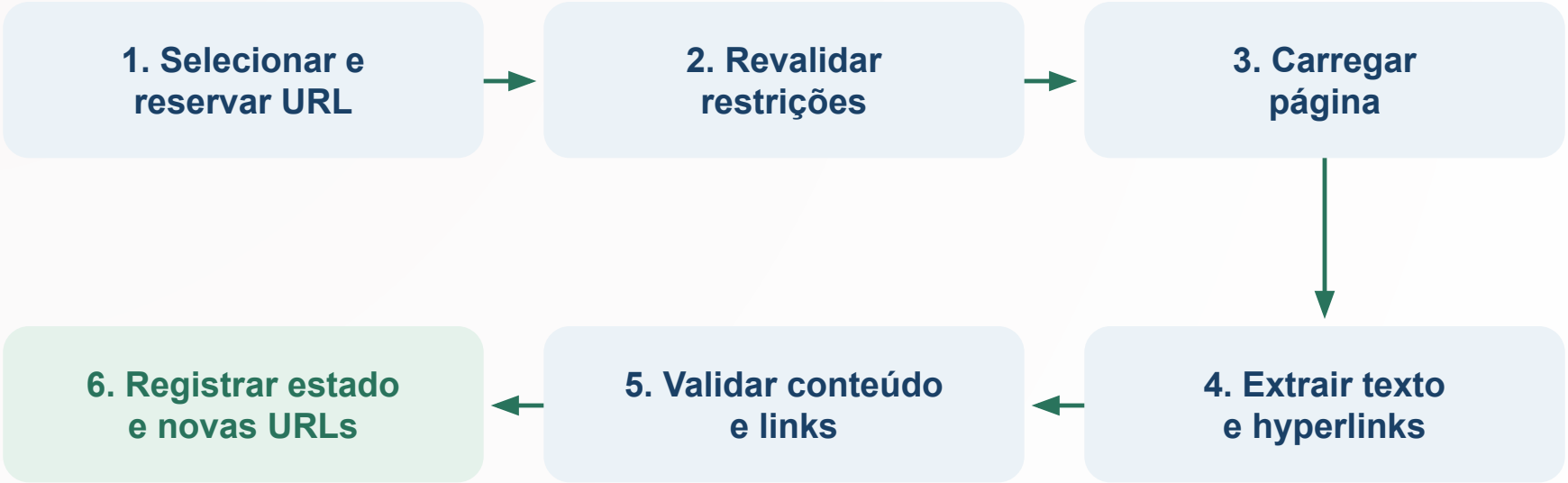
Páginas = nós; hyperlinks = arestas.

O banco mantém o estado e seleciona trabalho elegível.



- Estado global
- Reserva exclusiva
- Fila elegível
- Proveniência
- Retomada

Cada trabalhador decide, registra e volta a buscar uma URL elegível.



Uma página inválida permanece registrada e não expande a coleta.

Toda URL é reavaliada no momento do acesso concorrente.

Alcance

Profundidade da página e transições entre domínios.

Ritmo

Intervalo mínimo entre acessos ao mesmo domínio.

Retentativas

Máximo de tentativas e espera após falha.

Elegibilidade

Estado da URL, domínio bloqueado e limites de processamento.

O resultado da página decide se seus links podem expandir a coleta.

Conteúdo

Idioma e critérios de qualidade.
Origem inválida não gera links elegíveis.

Redundância

URL e domínio normalizados por hash.
Texto idêntico identificado por hash exato.

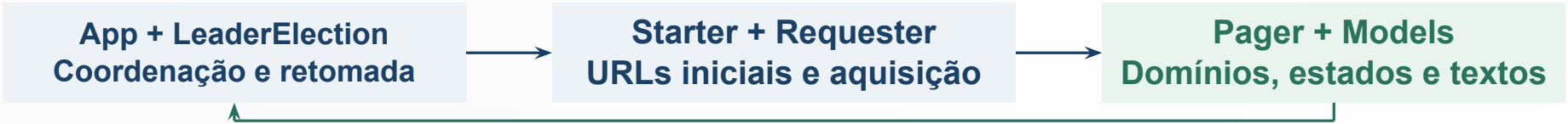
Proveniência

Relações pai e filho e same_as preservam percurso e duplicatas.





Figura 17. Trabalhadores concorrentes e persistência compartilhada.



14/03/2026 a 21/08/2026: exportação parcial, coleta em andamento.

318 termos-semente com StarterWikipedia

10 s	3 tentativas	500 ms
timeout por requisição	60 s entre tentativas	mínimo por domínio

Profundidade máxima de recursão: 4
Profundidade máxima de recursão por página: 4

83.590.843 URLs inventariadas no recorte



64,91% bloqueadas por profundidade 26,25% pendentes 6,32% concluídas 2,52% demais

5,29 mi
registros concluídos

2,65 mi
textos válidos

2,62 mi
documentos finais

2,72 bilhões de palavras no corpus analisado.

Nível	Domínios	Requisições	Requisições (%)
0	**1**	**5.977.393**	**65,32**
1	445.727	2.025.474	22,13
2	104.726	1.103.935	12,06
3	62.922	44.561	0,49
4	3.053	0	0,00

87,45% das 9.151.363 requisições ocorreram nos níveis 0 e 1.

616.429 domínios.

Domínio	Requisições	Textos válidos	Palavras após deduplicação
pt.wikipedia.org	**5.977.393**	**2.104.784**	**2.401.010.397**
pt.wikiversity.org	166.275	27.569	15.060.740
pt.wikisource.org	55.229	21.308	14.493.009
planalto.gov.br	18.146	6.858	13.373.498
pt.wikibooks.org	85.163	23.929	12.398.211

A Wikipédia responde por cerca de 88% das 2,72 bilhões de palavras.

Método e ferramenta

Exploração recursiva, fila implícita, controles e proveniência.
WackyWacky executa e retoma a coleta.

Recorte e limites

2,62 mi documentos e 2,72 bi palavras.
Concentração na Wikipédia e resíduos textuais.

Disponibilidade - P4

04

Utilidade dos dados no refinamento

Como o pré-treinamento continuado e a proximidade de domínio dos dados afetam o desempenho de um modelo de linguagem em português brasileiro quando o volume de dados, o orçamento computacional e os demais hiperparâmetros são mantidos constantes?

Desafio

Examinar treinamento adicional e composição dos dados em contrastes distintos.

Abordagem

Base e duas variantes de CPT; condições equivalentes entre Geral e Fórum/Tec.

Utilidade - P5

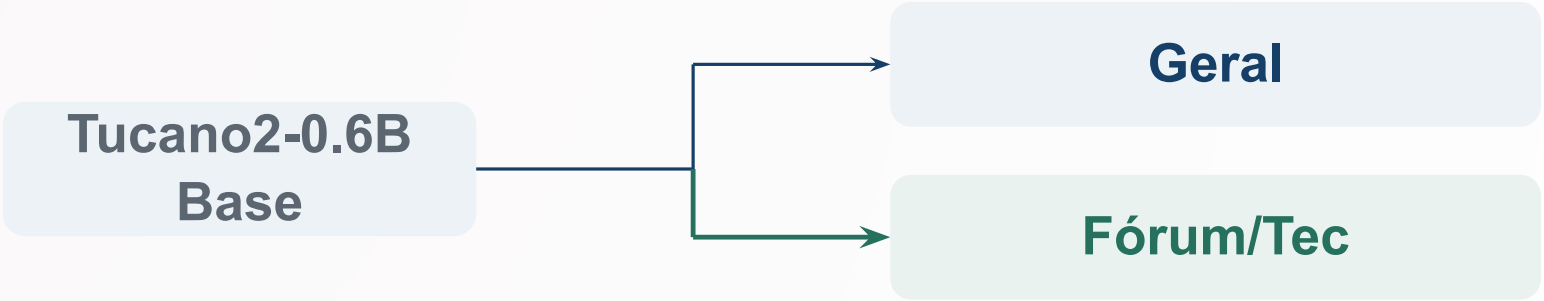
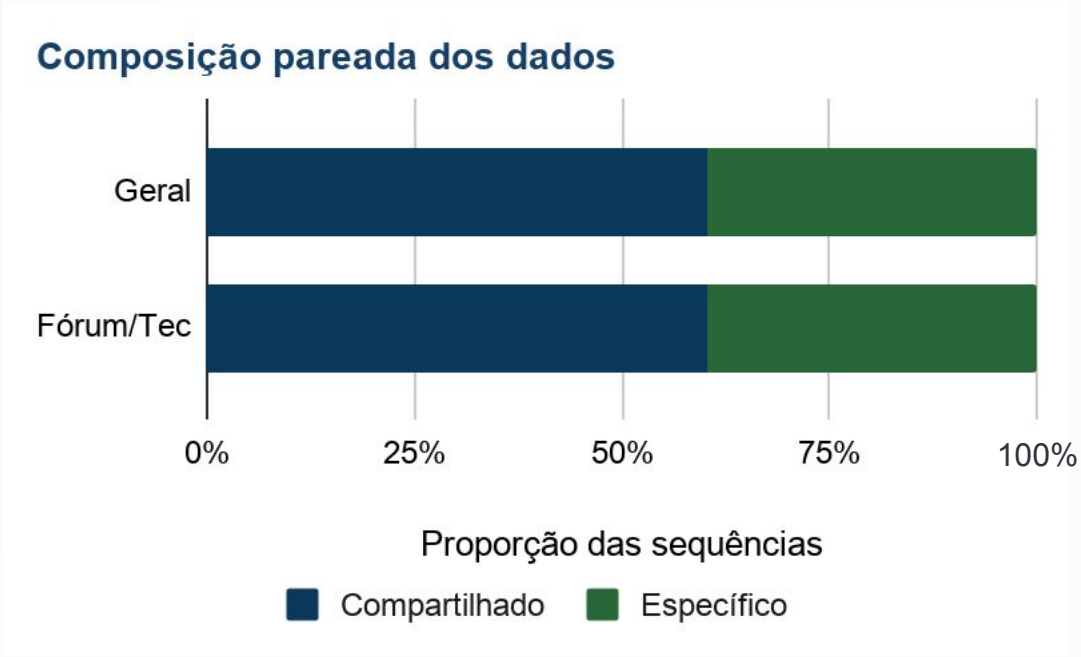


Figura 18. Modelo base e variantes refinadas.

Contraste	O que permite examinar
Geral - Base	Efeito do treinamento adicional com a composição geral.
Fórum/Tec - Geral	Efeito da substituição de uma parcela por textos de tecnologia.
Fórum/Tec - Base	Efeito total do refinamento com a composição Fórum/Tec.



60,20% compartilhados
Mesmas sequências, nas mesmas posições

39,80% substituídos
Composição alterada

Geral	BrWaC-CLEAN adicional
Fórum/Tec	Adrenaline + OuterSpace

Figura 19. Parcelas compartilhadas e específicas das composições.

Compartilhados: BrWaC-CLEAN, GigaVerbo, WackyWacky, MultiWOZ-PTBR

416.000	1.024	425.984.000	52.000
sequências	tokens por sequência	tokens por variante	passos por variante

Condições comuns às duas variantes

Mesmo checkpoint, arquitetura e tokenizador

Mesmos hiperparâmetros e semente

Atualização de todos os parâmetros

Separação entre treino e avaliação



Figura 20. Separação anterior ao packing.

Mesmo orçamento entre variantes. Controle de sobreposição por thread e grupo de títulos.

Avaliação	Dados e procedimento	Métrica
Intrínseca	6 corpora; 256 sequências de 1.024 tokens por corpus.	Loss e perplexidade
Classificação principal	Adrenaline, 6 categorias; título + primeiro post. Representações congeladas + regressão logística.	Macro-F1
Previsão em fóruns	1.800 threads; primeiro post condicionado ao título e aos tokens anteriores.	NLL condicional

Classificação: repetimos a seleção e a divisão dos dados cinco vezes, usando os mesmos exemplos para comparar os modelos.
Pré-treinamento contínuo: cada variante foi treinada uma única vez.

Menor perplexidade é melhor.

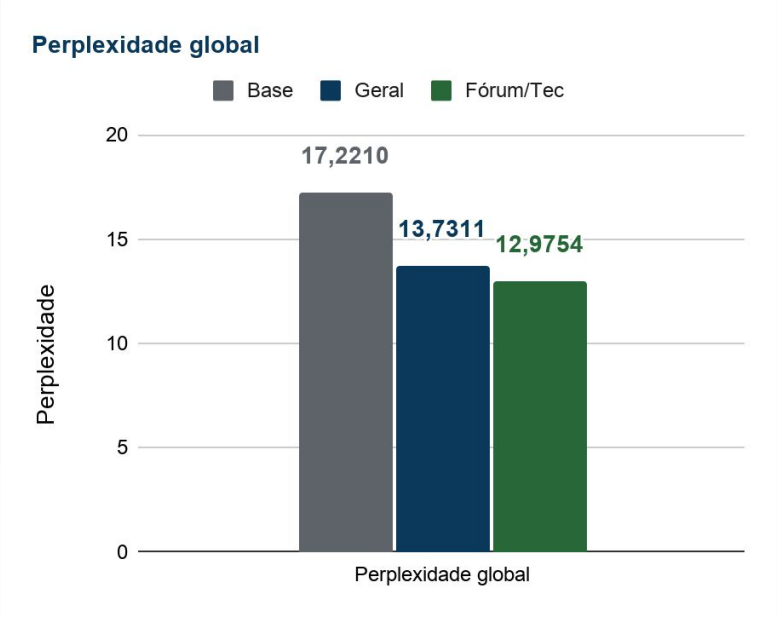


Figura 21. Perplexidade global dos três modelos.

Base 17,2210 - Geral 13,7311
Fórum/Tec 12,9754

Perplexidade por corpus

Corpus	Geral	Fórum/Tec
Adrenaline	**32,3824**	**27,8470**
BrWaC-CLEAN	15,0667	15,6238
GigaVerbo	12,9431	13,1389
MultiWOZ-PTBR	5,6755	5,6842
OuterSpace	**33,5156**	**26,1042**
WackyWacky	5,5797	5,6263

Fórum/Tec: menor PPL nos dois fóruns.
Geral: menor nos quatro demais corpora.

Comparação

Geral - Base

[+0,0010; +0,0101]

Fórum/Tec - Geral

[-0,0029; +0,0081]

Fórum/Tec - Base

[+0,0013; +0,0149]

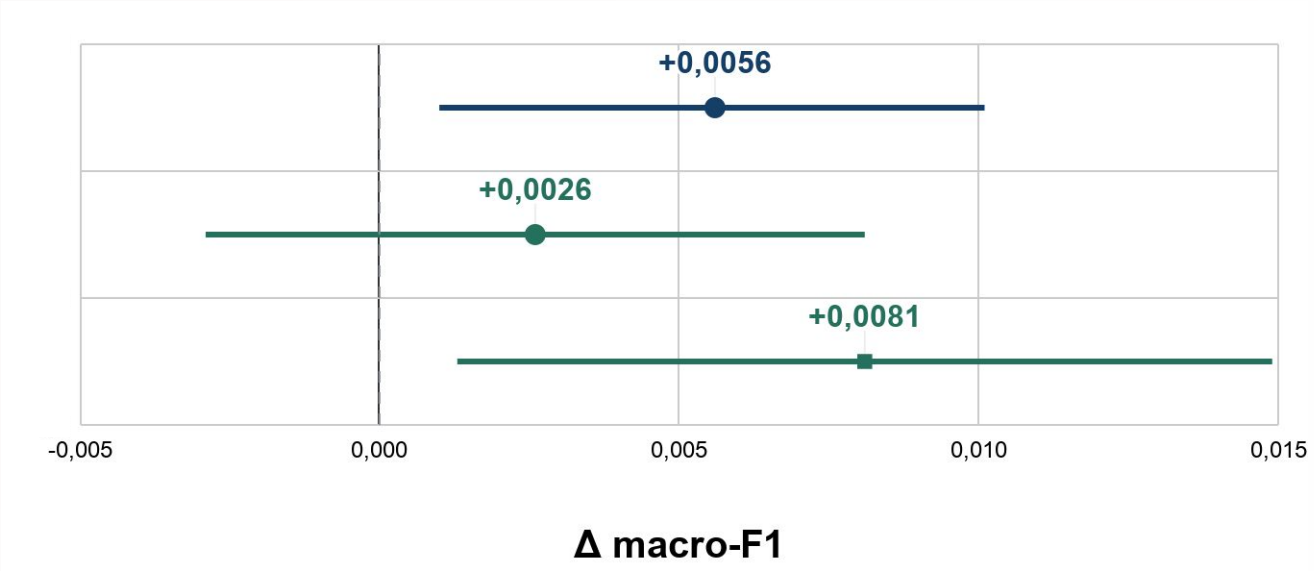


Figura 22. Diferenças de macro-F1 e intervalos descritivos de 95%.

Médias: Base 0,8194 - Geral 0,8250 - Fórum/Tec 0,8276

IC95% descritivos entre cinco amostragens/divisões, não entre treinamentos independentes.

Contraste: Fórum/Tec - Geral

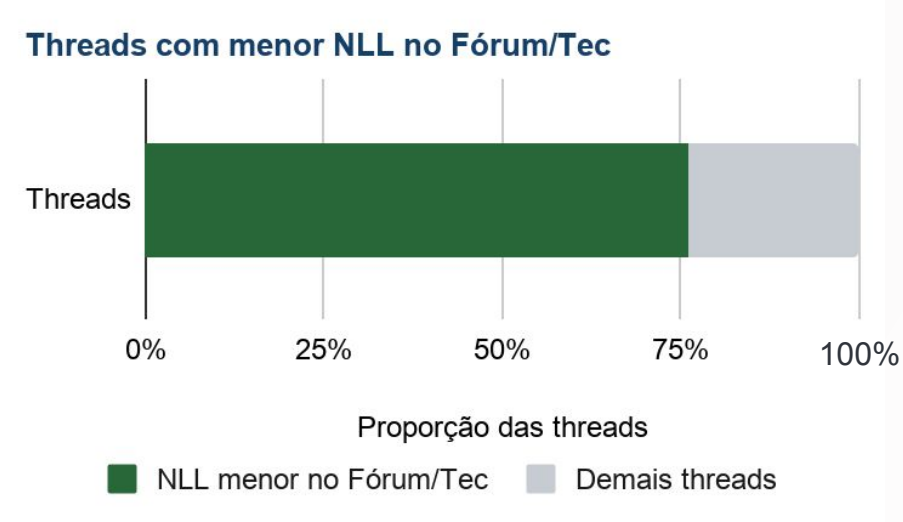


Figura 23. Threads com menor NLL no Fórum/Tec.

76,11% das 1.800 threads com NLL menor no Fórum/Tec

Δ NLL = -0,0827

IC95% [-0,0881; -0,0774]

Perplexidade ≈ 7,94% menor

Bootstrap pareado e estratificado por categoria: 10.000 reamostragens de threads.

Base: NLL 3,3593 - checkpoints: 0, 13 mil, 26 mil, 39 mil e 52 mil passos

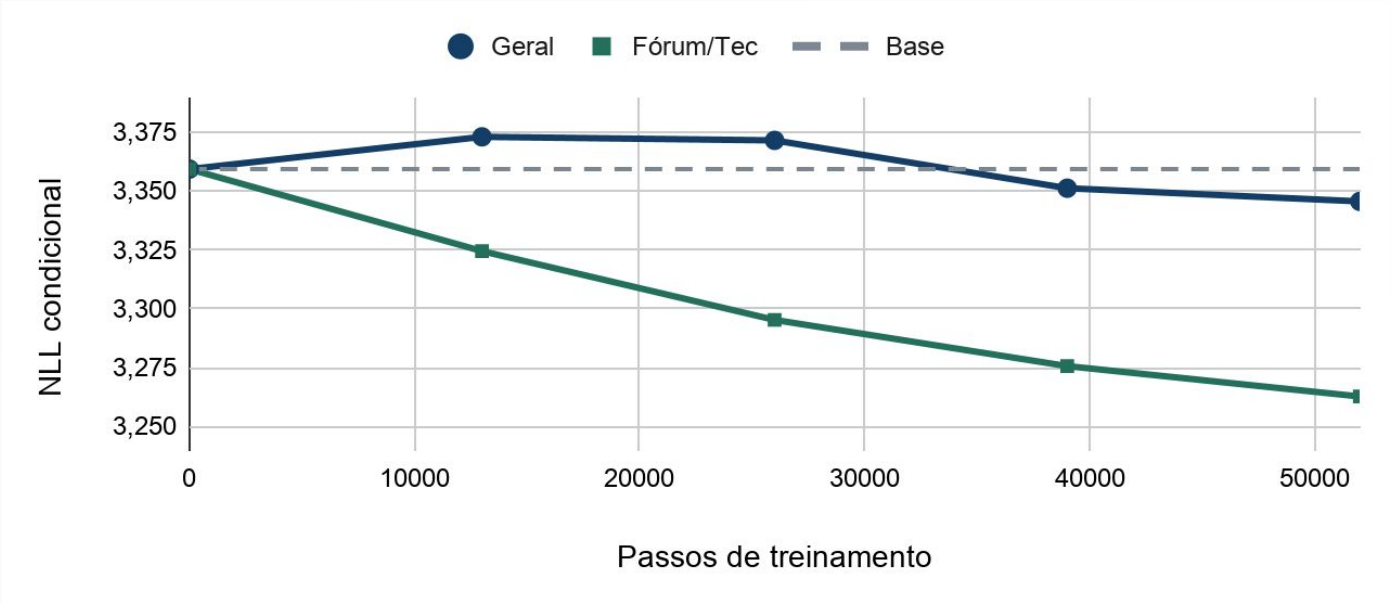


Figura 24. NLL condicional nos checkpoints de treinamento observados.

Ambos melhoraram de 39 mil a 52 mil passos;

Geral vs. Base

Treinamento adicional

Modelagem de linguagem: melhora

Classificação: pequeno ganho médio

Fórum/Tec vs. Geral

Composição dos dados

Previsão dos fóruns: melhora

Classificação: diferença pequena

A escolha da composição depende da tarefa prioritária.

05

Privacidade

Em que medida a substituição semântica rotativa reduz a reprodução de dados pessoais sintéticos durante o treinamento federado de um modelo de linguagem em português brasileiro, em comparação com a ausência de defesa, na presença de um cliente auxiliar benigno ou adversário?

Desafio

Reduzir reprodução de dados pessoais preservando utilidade.

Abordagem

Substituição rotativa em FL, comparada à ausência de defesa e ao DP-AdamW.

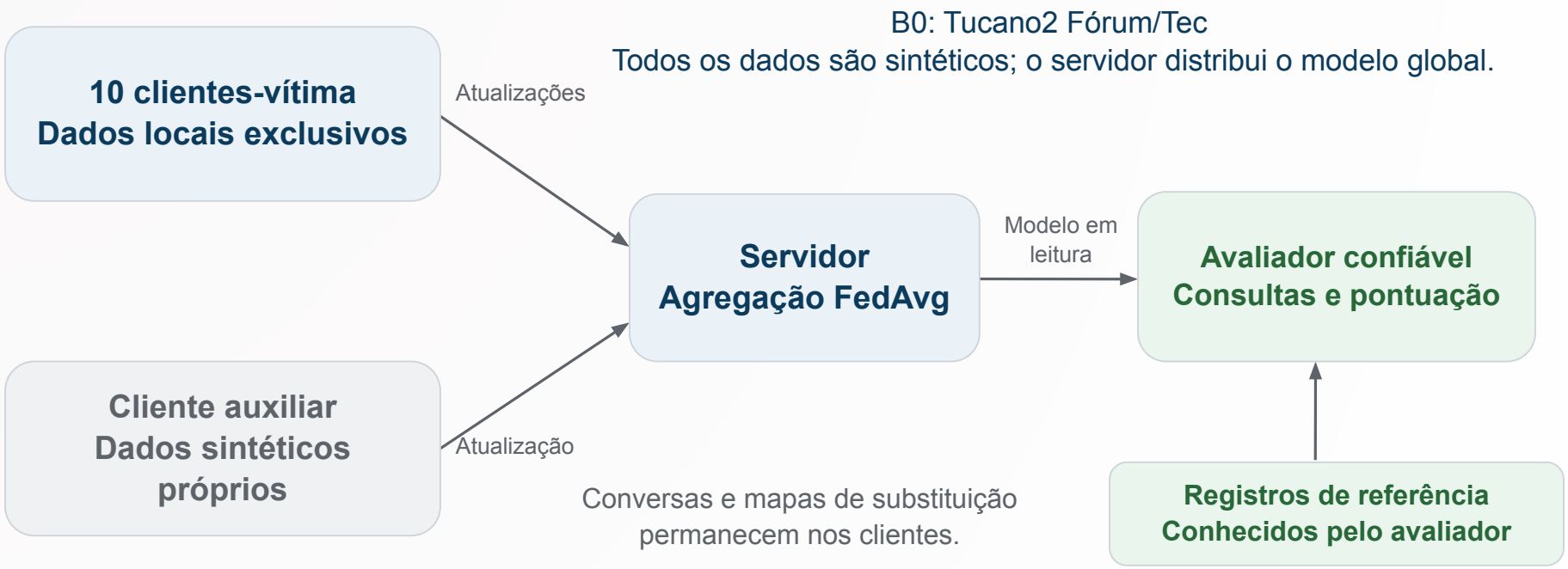


Figura 25. Clientes, servidor e avaliador na simulação federada.

10 clientes-vítimas × 20 participantes × 5 conversas = 1.000 conversas

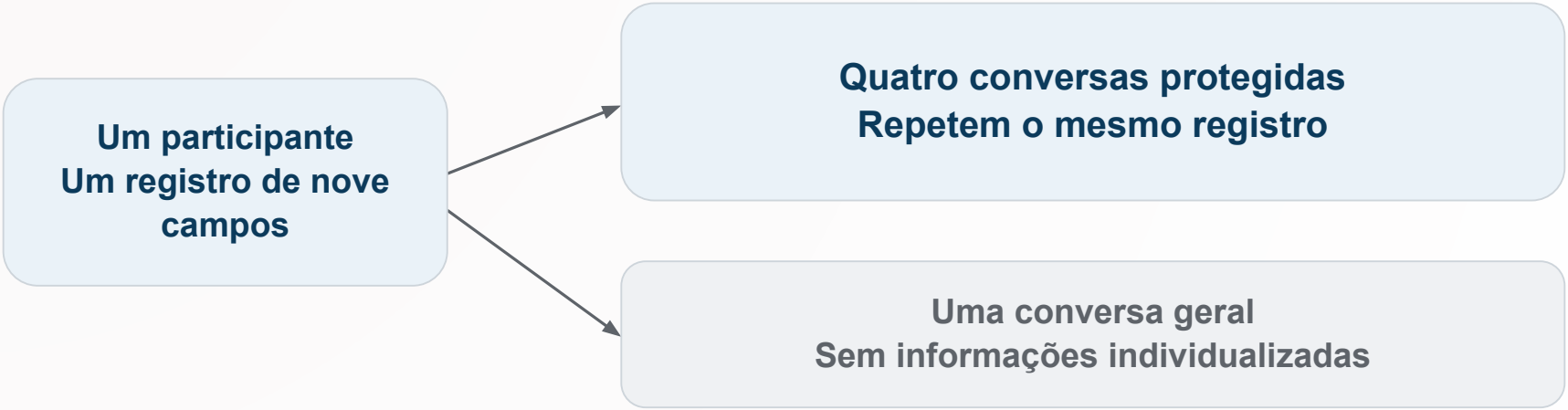


Figura 26. Organização das conversas por participante sintético.

200 participantes sintéticos por execução; quatro conversas protegidas e uma geral.

Perfis e amostras iguais entre condições na mesma rodada; renovados entre rodadas.

Auxiliar benigno

Usuário: Nome: {nome}
Nascimento: {data}; CPF: {CPF}; …
Assistente: registro confirmado.

Perda sobre toda a conversa

Auxiliar adversário

Nome: {nome}
Prefixo sem contribuição à perda

Nascimento: {data}; CPF: {CPF}; …
Continuação supervisionada

Perda concentrada nos campos seguintes

Figura 27. Supervisão usada pelos auxiliares benigno e adversário.

Perfis próprios renovados a cada rodada: 80 conversas protegidas + 20 gerais.
As conversas gerais usam perda integral nas duas condições.

Nove campos substituídos antes da tokenização

Semântica: preserva tipo, formato e restrições do atributo; os valores pessoais são substituídos.

Campo	Original	Rodada r	Rodada r + 1
Nome	$\{nome_0\}$	$\{nome_r\}$	$\{nome_{r+1}\}$
Nascimento (fictício)	**01/01/1990**	**12/05/1988**	**03/09/1992**
CPF	$\{CPF_0\}$	$\{CPF_r\}$	$\{CPF_{r+1}\}$
RG	$\{RG_0\}$	$\{RG_r\}$	$\{RG_{r+1}\}$
Telefone	$\{tel_0\}$	$\{tel_r\}$	$\{tel_{r+1}\}$
E-mail	$\{email_0\}$	$\{email_r\}$	$\{email_{r+1}\}$
Endereço	$\{end_0\}$	$\{end_r\}$	$\{end_{r+1}\}$
Data de atendimento	$\{data_0\}$	$\{data_r\}$	$\{data_{r+1}\}$
Horário de atendimento	$\{hora_0\}$	$\{hora_r\}$	$\{hora_{r+1}\}$

Rodada r: mesmos substitutos nas 4 conversas protegidas e nas 4 passagens locais.
Rodada r + 1: novos valores, a conversa geral permanece igual.

Localização por anotações sem avaliar detecção automática.

Defesa nas vítimas	Auxiliar benigno	Auxiliar adversário
Sem defesa	F0	F1
DP-AdamW, ε = 3	F2, ε = 3	F3, ε = 3
DP-AdamW, ε = 8	F2, ε = 8	F3, ε = 8
Substituição rotativa	F4	F5

20 rodadas por trajetória **Peso 1/11 por cliente**
Mesmo estado inicial B0 **Duas execuções**

Cada participante sintético: 4 conversas com o mesmo registro de 9 campos pessoais.
DP-AdamW: ε = 3 ou 8 ($\delta = 10^{-5}$) por conversa.

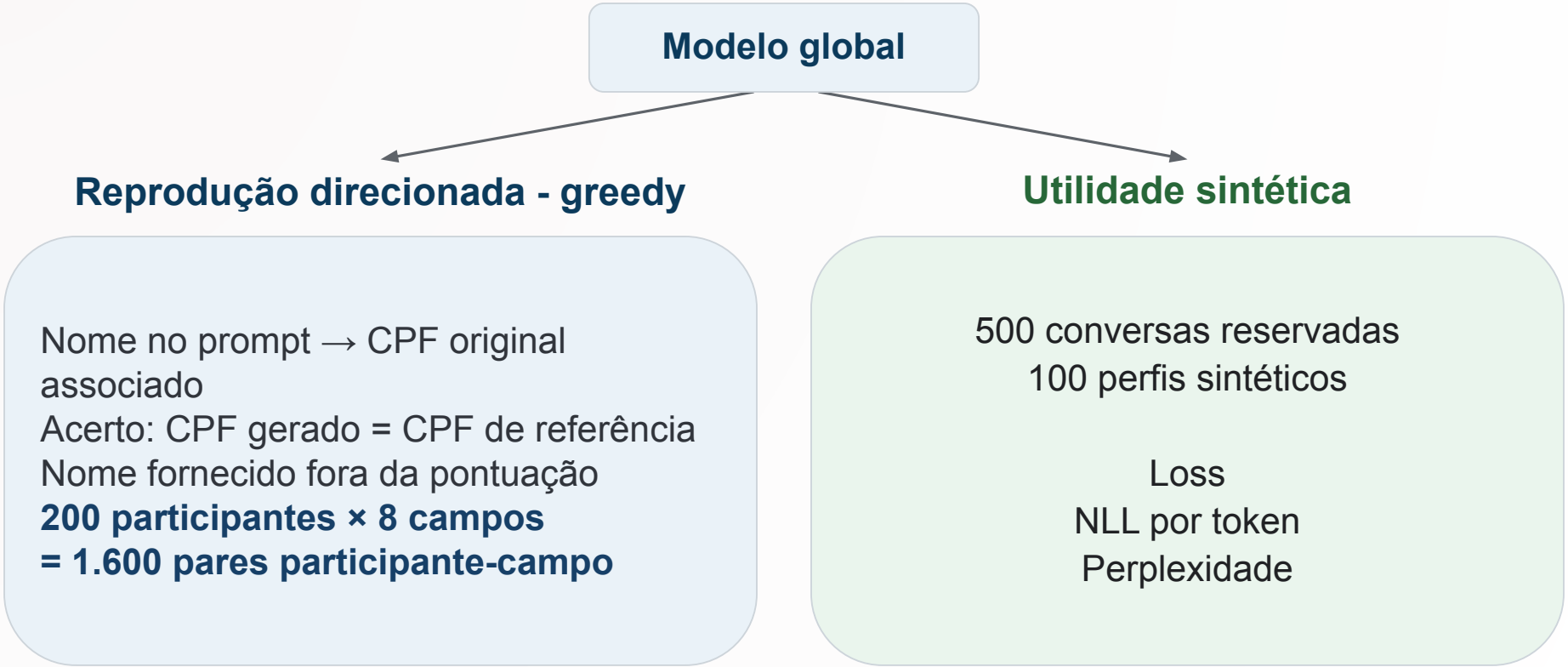


Figura 28. Auditoria da reprodução de registros e da utilidade sintética.

Três hipóteses e seus contrastes experimentais

Hipótese	Comparação controlada
H1: o auxiliar adversário aumenta a reprodução dos registros originais?	**F1 × F0.** Sem defesa nas vítimas; muda apenas o auxiliar.
H2: a substituição reduz a reprodução e preserva a utilidade?	**F4 × F0 e F5 × F1.** Mesmo auxiliar; substituição versus ausência de defesa.
H3: a substituição equilibra melhor privacidade e utilidade que DP-AdamW?	**F4 × F2 e F5 × F3.** Mesmo auxiliar; DP com ε = 3 ou 8.

Rodada 20: médias de duas execuções

Defesa	Benigno Pares	Adversário Pares	Redução benigno	Redução adversário
Sem defesa	F0: 305,0	F1: 308,5	-	-
DP-AdamW, ε = 3	F2: 4,0	F3: 5,5	**98,70%**	**98,24%**
DP-AdamW, ε = 8	F2: 6,0	F3: 5,0	**98,03%**	**98,37%**
Substituição semântica rotativa	F4: 5,5	F5: 7,5	**98,20%**	**97,56%**

Nenhuma condição com defesa reproduziu um perfil original completo.

F1 - F0 = 3,5 pares: amplificação pelo auxiliar permanece inconclusiva.

Perplexidade média - menor é melhor

Defesa	Auxiliar benigno	Auxiliar adversário
Sem defesa	5,913	5,910
DP-AdamW (ε = 3)	4,632	7,039
DP-AdamW (ε = 8)	4,600	6,897
Substituição semântica	**3,371**	**3,359**

Substituição: PPL média 43,00% menor diante de F0 e 43,17% diante de F1.

Hipótese	Resultado observado
H1. Auxiliar adversário	F1: 308,5; F0: 305,0. Diferença média: +3,5 pares.
H2. Substituição rotativa	98,20% e 97,56% menos pares originais; menor perplexidade.
H3. Comparação com DP-AdamW	Ambas reduziram mais de 97% dos pares; a substituição teve menor perplexidade.

A substituição semântica rotativa reduziu a reprodução dos dados originais e obteve a melhor utilidade sintética.

06

Conclusão

DISPONIBILIDADE · PREPARAÇÃO DE DADOS

P1 Reconstruir

Conversas reais

Referências entre mensagens reconstroem diálogos multiturnos e multiparticipantes.

P2 Gerar

Autoria de diálogos

Roteiros e autoria humana assíncrona: geração controlada e recombinação de turnos compatíveis.

P3 Adaptar

Recursos existentes

Filtragem, tradução e realinhamento adaptam corpora e diálogos anotados ao PT-BR.

P4 Coletar

Coleta em escala

Descoberta, controles de aquisição e persistência permitem uma coleta web rastreável.

UTILIDADE E PRIVACIDADE

P5 Utilidade e CPT

CPT: melhor modelagem e **ganhos modestos** na classificação.

Fórum/Tec: melhor previsão de textos de fórum; **vantagem modesta** sobre Geral na classificação.

P6 Privacidade

Substituição: **>97%** menos pares originais reproduzidos; sem perfis completos, com resíduos.

Melhor utilidade sintética entre as defesas.

Auxiliar: +3,5 pares; DP-AdamW também reduziu >97%.

3	**metodologias de preparação**	• MCCD	• MDC	• Coleta web controlada
3	**ferramentas de preparação**	• Miner-XenForo	• MDC	• WackyWacky
6	**conjuntos de dados**	• Adrenaline • BrWaC-CLEAN	• OuterSpace • MultiWOZ-PTBR	• Nutritious Food • WackyWacky PTBR Geral
2	**variantes do Tucano2**	• Geral	• Fórum/Tec	
2	**pipelines experimentais**	• Pré-treinamento continuado	• Avaliação federada de privacidade	
1	**mecanismo de defesa**	• Substituição semântica rotativa		

3	**metodologias de preparação**	• MCCD	• MDC	• Coleta web controlada
3	**ferramentas de preparação**	• Miner-XenForo	• MDC	• WackyWacky
6	**conjuntos de dados**	• Adrenaline • BrWaC-CLEAN	• OuterSpace • MultiWOZ-PTBR	• Nutritious Food • WackyWacky PTBR Geral
2	**variantes do Tucano2**	• Geral	• Fórum/Tec	
2	**pipelines experimentais**	• Pré-treinamento continuado	• Avaliação federada de privacidade	
1	**mecanismo de defesa**	• Substituição semântica rotativa		

3	**metodologias de preparação**	• MCCD	• MDC	• Coleta web controlada
3	**ferramentas de preparação**	• Miner-XenForo	• MDC	• WackyWacky
6	**conjuntos de dados**	• Adrenaline • BrWaC-CLEAN	• OuterSpace • MultiWOZ-PTBR	• Nutritious Food • WackyWacky PTBR Geral
2	**variantes do Tucano2**	• Geral	• Fórum/Tec	
2	**pipelines experimentais**	• Pré-treinamento continuado	• Avaliação federada de privacidade	
1	**mecanismo de defesa**	• Substituição semântica rotativa		

3	**metodologias de preparação**	• MCCD	• MDC	• Coleta web controlada
3	**ferramentas de preparação**	• Miner-XenForo	• MDC	• WackyWacky
6	**conjuntos de dados**	• Adrenaline • BrWaC-CLEAN	• OuterSpace • MultiWOZ-PTBR	• Nutritious Food • WackyWacky PTBR Geral
2	**variantes do Tucano2**	• Geral	• Fórum/Tec	
2	**pipelines experimentais**	• Pré-treinamento continuado	• Avaliação federada de privacidade	
1	**mecanismo de defesa**	• Substituição semântica rotativa		

3	**metodologias de preparação**	• MCCD	• MDC	• Coleta web controlada
3	**ferramentas de preparação**	• Miner-XenForo	• MDC	• WackyWacky
6	**conjuntos de dados**	• Adrenaline • BrWaC-CLEAN	• OuterSpace • MultiWOZ-PTBR	• Nutritious Food • WackyWacky PTBR Geral
2	**variantes do Tucano2**	• Geral	• Fórum/Tec	
2	**pipelines experimentais**	• Pré-treinamento continuado	• Avaliação federada de privacidade	
1	**mecanismo de defesa**	• Substituição semântica rotativa		

3	**metodologias de preparação**	• MCCD	• MDC	• Coleta web controlada
3	**ferramentas de preparação**	• Miner-XenForo	• MDC	• WackyWacky
6	**conjuntos de dados**	• Adrenaline • BrWaC-CLEAN	• OuterSpace • MultiWOZ-PTBR	• Nutritious Food • WackyWacky PTBR Geral
2	**variantes do Tucano2**	• Geral	• Fórum/Tec	
2	**pipelines experimentais**	• Pré-treinamento continuado	• Avaliação federada de privacidade	
1	**mecanismo de defesa**	• Substituição semântica rotativa		

3	**metodologias de preparação**	• MCCD	• MDC	• Coleta web controlada
3	**ferramentas de preparação**	• Miner-XenForo	• MDC	• WackyWacky
6	**conjuntos de dados**	• Adrenaline • BrWaC-CLEAN	• OuterSpace • MultiWOZ-PTBR	• Nutritious Food • WackyWacky PTBR Geral
2	**variantes do Tucano2**	• Geral	• Fórum/Tec	
2	**pipelines experimentais**	• Pré-treinamento continuado	• Avaliação federada de privacidade	
1	**mecanismo de defesa**	• Substituição semântica rotativa		

Best Paper Award - MCCD - ICEIS 2022

MCCD: Generating Human Natural Language Conversational Datasets

Sanches et al., 2022a - ICEIS - contribuição direta

Textual Datasets For Portuguese-Brazilian Language Models

Sanches et al., 2022b - DSW - contribuição direta

Automatic Extraction of Conversation Flows from Human Dialogues: Understanding Their Impact to Refine NLP Models

Sanches et al., 2023 - SN Computer Science, 4(6) - contribuição direta

EFIS - Ecological Fuel-consumption Intelligent System

Sanches et al., 2021 - DCOSS - produção complementar

Curricular Transfer Learning for Sentence Encoded Tasks

de Sá et al., 2023 - arXiv:2308.01849 - cooperação complementar

Limitação observada	Próximo experimento
Cobertura e domínio restritos	Novas fontes e domínios. Acompanhar mudanças linguísticas e temáticas.
Uma execução por variante	Repetições independentes, outros tamanhos de modelo e treino do zero versus CPT.
Distribuição sintética	Tarefas externas, novas distribuições e outras participações do auxiliar.
Anotações estruturadas	Detectar em texto livre e substituir em dispositivos com recursos limitados.

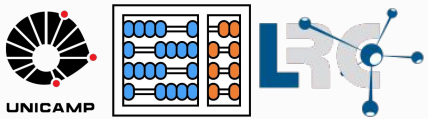
Ampliar a validade externa e avaliar a defesa em condições menos controladas.

Defesa de Doutorado

Métodos e Recursos de Dados para Modelos de Linguagem em Português Brasileiro

da Coleta e Adaptação à Privacidade em Aprendizado Federado

Candidato: Matheus Ferraroni Sanches
Orientador: Prof. Dr. Leandro Aparecido Villas



Campinas, 26 de Outubro de 2026

Marcadores preservam a presença, o tipo e o papel dos elementos.

Marcador	Representa
<visual>	Imagens, figuras, diagramas e capturas.
<symbolic>	Emojis, ícones e reações visuais.
<embedded>	Vídeos, áudios, mapas e conteúdo interativo.
<reference>	Referências com destino, rótulo ou identificador.
Posição reservada	Presença de conteúdo omitido ou irrelevante.

Marcador	Representa
<quotation>	Conteúdo citado ou reutilizado.
<relation>	Relações entre unidades textuais.
<executable>	Código, comandos e expressões formais.
<container>	Conteúdo encapsulado como uma unidade.

Adrenaline
7 categorias / 26 subcategorias

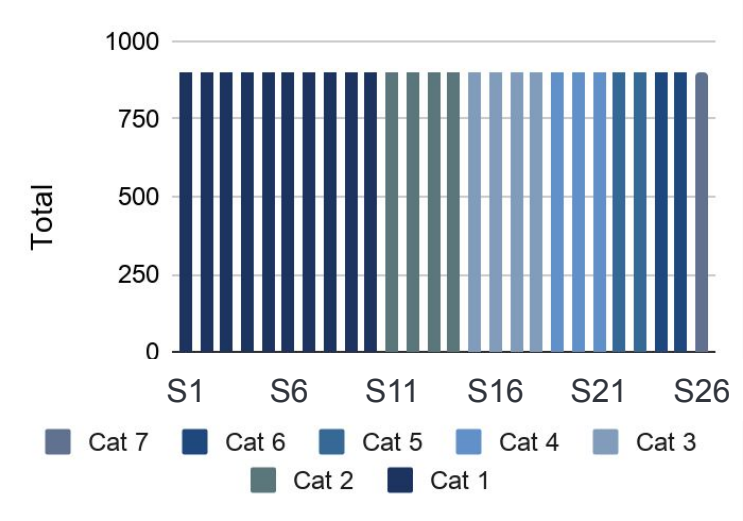


Figura 29. Instâncias por subcategoria no Adrenaline.

900 instâncias por subcategoria
Divisão estratificada 75/25

IMDb PT-BR
49.459 críticas

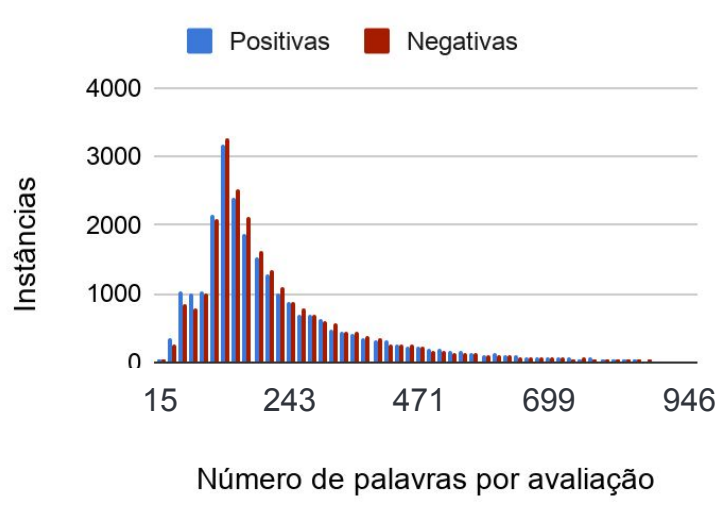


Figura 30. Comprimento das críticas do IMDb PT-BR.

Semente fixa; sem validação cruzada.
Sentimento avaliado separadamente.

01 - Start
02 - [Cliente] Cumprimente o sistema de forma educada.
03 - [Sistema] Cumprimente o usuário e pergunte se pode ajudar.
04 - [Cliente] Pergunte se você pode fazer uma pergunta sobre a quantidade de calorias em um alimento.
05 - [Sistema] Confirme que esse é o tipo de pergunta com o qual você pode ajudar o usuário.
06 - [Cliente] Pergunte: Quantas calorias há no chocolate?
07 - [Sistema] Informe: O chocolate tem aproximadamente 546 calorias por 100 gramas.
08 - [Cliente] Agradeça ao sistema.
09 - [Sistema] Pergunte se pode ajudar em algo mais.
10 - [Cliente] Pergunte: Quantas calorias há no chocolate amargo.
11 - [Sistema] Informe: O chocolate amargo tem aproximadamente 185 calorias por 100 gramas.
12 - [Cliente] Diga que isso é interessante e agradeça ao sistema.
13 - [Sistema] Concorde com o usuário.
14 - [Cliente] Faça uma despedida.
15 - [Sistema] Faça uma despedida.

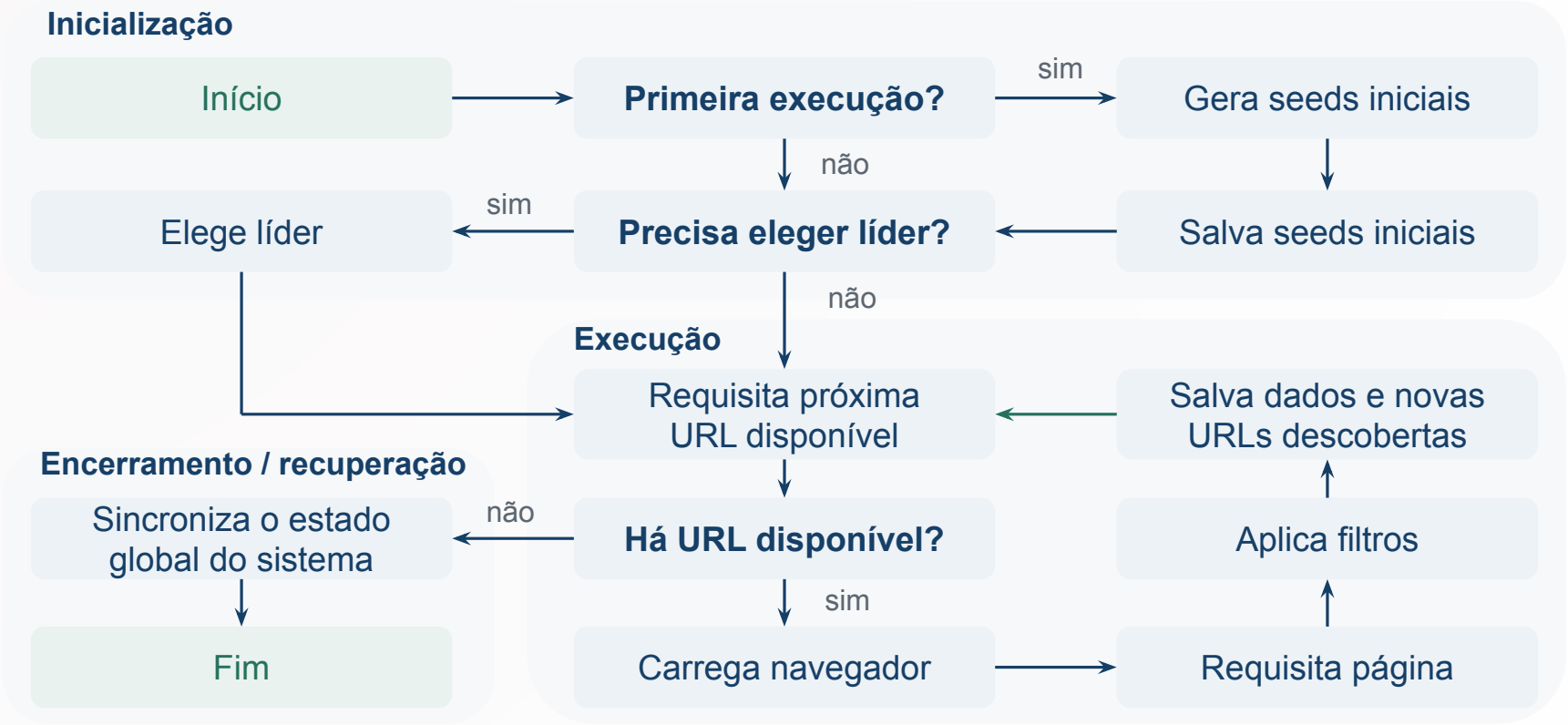


Figura 31. Fluxo metodológico da coleta web controlada.

318 termos - exemplos reais por tema

Tecnologia e dados	Jogos e cultura	Sociedade e economia
git - sql - linux	esports - cultura pop	saúde pública - renda fixa
docker - programação	design de jogos	educação básica
análise de dados	crítica de filme	geopolítica

Exemplos da lista de 318 termos-semente

Tecnologia e programação
python documentação, linux, docker, kubernetes, engenharia de dados

Jogos digitais
unity documentação, godot documentação, design de jogos, esports

Economia e sociedade
renda fixa, comércio exterior brasileiro, mercado de trabalho

Ciência, saúde e ambiente
pesquisa científica, saúde pública, mudanças climáticas no brasil

Política e cultura
políticas públicas, geopolítica, cultura pop, crítica de filme

Lista integral disponível nas notas deste apoio.

A6 | Termos-semente

git, sql, linux, nosql, devops, docker, esports, netcode, shaders, big data, anti-cheat, kubernetes, renda fixa, cultura pop, geopolítica, ia em jogos, programação, ux em jogos, venv python, apis restful, blog pessoal, level design, métricas esg, brics análise, crédito rural, cultura gamer, esg no brasil, saúde pública, c documentação, edital público, github actions, jogos digitais, manual técnico, poluição do ar, c# documentação, crise econômica, design de jogos, educação básica, ensino superior, física em jogos, go documentação, observabilidade, análise de dados, análise de jogos, análise de série, análise política, c++ documentação, crítica de filme, crítica de jogos, cultura de memes, empreendedorismo, estudo acadêmico, futebol notícias, gem documentação, lua documentação, npm documentação, php documentação, pip documentação, reforma política, regex referência, turismo destinos, violência urbana, .net documentação, artigo de opinião, economia de jogos, java documentação, jvm especificação, marketing digital, mobilidade social, notícias recentes, perl documentação, relatório técnico, ruby documentação, rust documentação, sdk e bibliotecas, análise de mercado, cargo documentação, comentário crítico, design de sistemas, eleições no brasil, gestão empresarial, godot documentação, http especificação, indústria de games, inflação no brasil, json especificação, julia documentação, literatura análise, maven documentação, modelos preditivos, narrativa em jogos, política ambiental, políticas de saúde, políticas públicas, scala documentação, streaming de jogos, swift documentação, testes de software, unity documentação, webgl documentação, yaml especificação, campeonato nacional, computação em nuvem, comércio eletrônico, documentação de api, engenharia de dados, ferrovias no brasil, governança de dados, gradle documentação, história do esporte, indicadores sociais, kotlin documentação, mercado de trabalho, negociação coletiva, opengl documentação, pesquisa científica, python documentação, servidores de jogos, tecnologia no campo, violência de gênero, vulkan documentação, agricultura familiar, alimentação saudável, biografia de artista, cidades inteligentes, descrição de produto, directx documentação, documentação oficial, educação a distância, game design document, graphql documentação, inovação tecnológica, lançamentos de jogos, monetização em jogos, movimentos culturais, node.js documentação, patrimônio histórico, polarização política, processo legislativo, página institucional, qa e testes em jogos, raça e desigualdades, segurança energética, transição energética, composer documentação, demografia brasileira, desempenho de atletas, diplomacia brasileira, direitos trabalhistas, divulgação científica, eficiência energética, lançamentos do cinema, legislação atualizada, openapi especificação, procedural generation, programação funcional, psicologia do jogador, relato de experiência, rodovias e concessões, transformação digital, tutorial de linguagem, visualização de dados, aprendizado de máquina, direito constitucional, energia solar e eólica, governança corporativa, história contemporânea, latência e matchmaking, multiplayer networking, política internacional, acessibilidade em jogos, arquitetura de software, bastidores da televisão, dados abertos no brasil, desenvolvimento pessoal, exportações industriais, gastronomia tradicional, go modules documentação, gráficos 3d fundamentos, história dos videogames, indústria 4.0 no brasil, informalidade no brasil, javascript documentação, licenciamento ambiental, política de privacidade, referência de linguagem, reforma do ensino médio, salário mínimo impactos, segurança da informação, steamworks documentação, typescript documentação, artigo científico resumo, compliance e integridade, desemprego e indicadores, desmatamento na amazônia, engenharia de requisitos, gerenciamento de pacotes, influenciadores digitais, investimento sustentável, petróleo e gás no brasil, produtividade e salários, python biblioteca padrão, qualidade de vida urbana, r linguagem documentação, redes sociais tendências, resenha de álbum musical, saúde mental no trabalho, setor automotivo análise, tendências de tecnologia, agenda do governo federal, biocombustíveis no brasil, biodiversidade brasileira, controle de gastos mensal, documentação de linguagem, inovação em biotecnologia, mercados globais panorama, paradigmas de programação, qualificação profissional, termos e condições de uso, tratamento para ansiedade, ética em ciência de dados, aeroportos e aviação civil, campanha eleitoral análise, ciência política conceitos, competitividade industrial, crime organizado no brasil, desindustrialização debate, economia gig e plataformas, educação financeira básica, energia elétrica no brasil, exportações do agronegócio, logística e infraestrutura, migração interna no brasil, policiamento e estratégias, política agrícola nacional, portos e comércio exterior, relato de sustentabilidade, relatório econômico mensal, renderização em tempo real, risco climático e finanças, trabalho remoto tendências, unreal engine documentação, bancos de dados relacionais, cenário econômico do brasil, desenvolvimento de software, economia criativa no brasil, energia renovável no brasil, governança e gestão pública, legislação penal atualizada, organizações internacionais, pobreza e políticas sociais, prevenção de doenças comuns, reforma trabalhista análise, regulação do setor elétrico, saneamento básico no brasil, segurança pública no brasil, acordos comerciais e impacto, cadeias produtivas no brasil, commodities agrícolas preços, comércio exterior brasileiro, direitos humanos e segurança, matriz energética brasileira, privacidade em redes sociais, sistema prisional brasileiro, desigualdade social no brasil, estatísticas de criminalidade, indústria de base e mineração, investimentos para iniciantes, mudanças climáticas no brasil, política industrial no brasil, políticas educacionais atuais, resíduos sólidos e reciclagem, segurança alimentar no brasil, sustentabilidade no cotidiano, compiladores e interpretadores, decisões do congresso nacional, defensivos agrícolas regulação, imposto de renda pessoa física, inovação em modelos de negócio, juventude e políticas públicas, obras públicas e investimentos, partidos políticos brasileiros, políticas de segurança pública, preços de combustíveis análise, qualidade de dados e validação, sanções internacionais efeitos, streaming e indústria cultural, transparência e accountability, agricultura brasileira panorama, automação industrial tendências, descobertas recentes na ciência, impactos climáticos na produção, indústria de alimentos panorama, planejamento financeiro pessoal, programação orientada a objetos, violência doméstica e políticas, algoritmos e estruturas de dados, coalizões e articulação política, conflitos internacionais análise, habitação e déficit habitacional, inteligência artificial aplicada, migração internacional e refúgio, métodos de aprendizagem eficazes, crise hídrica e recursos hídricos, desinformação e checagem de fatos, indicadores econômicos principais, comunidades online e comportamento, conservação e unidades de proteção, corrupção e controle institucional, processamento de linguagem natural, diversidade e inclusão nas empresas, metodologia de pesquisa estatística, publicidade em plataformas digitais, responsabilidade social corporativa, envelhecimento populacional no brasil, cadeia de fornecedores e due diligence, opinião pública e pesquisas eleitorais, transporte público e mobilidade urbana

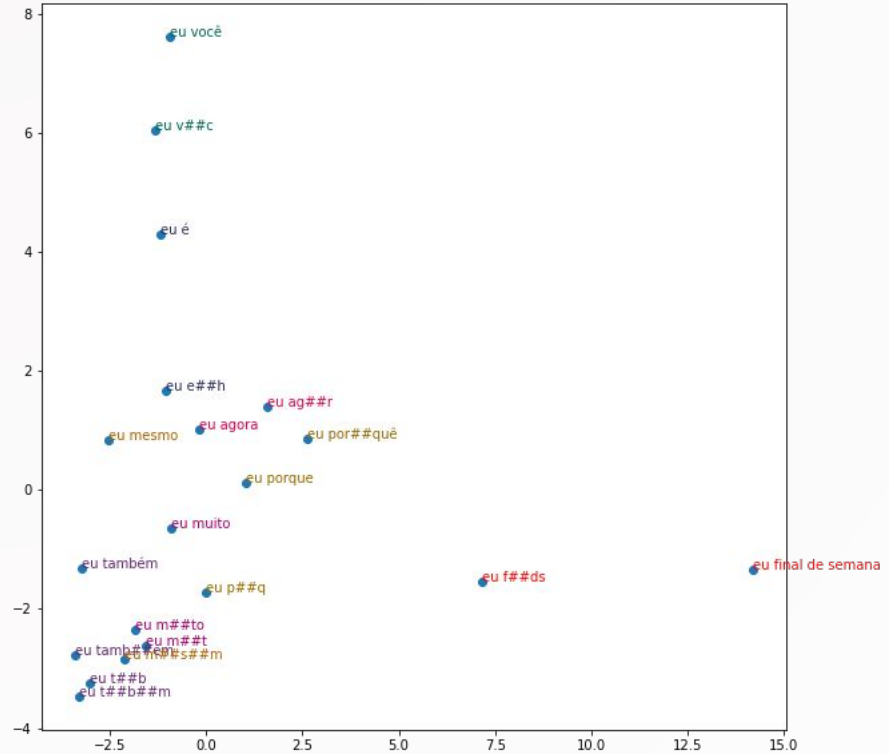


Figura 32. Distribuição dos termos

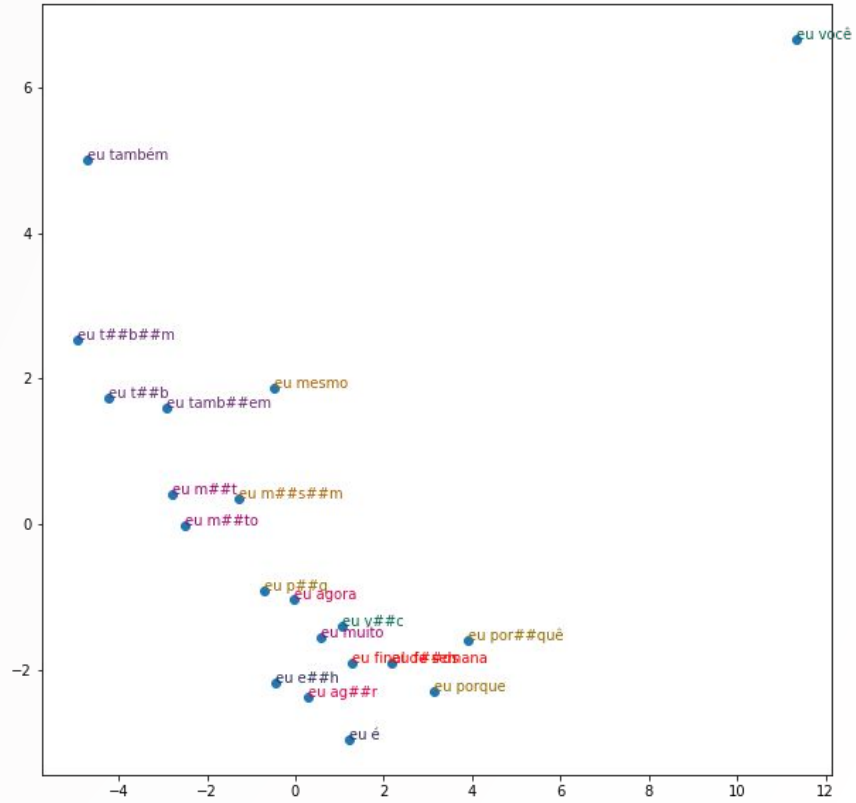


Figura 33. Distribuição dos termos

	eu final de semana	eu fds	eu agora	eu agr	eu eh	eu é	eu mto	eu mt	eu muito	eu também	eu tbm	eu tambem	eu tb	eu mesmo	eu msm	eu você	eu vc	eu porque	eu pq	eu porquê
eu final de semana	1.00	0.78	0.61	0.62	0.54	0.52	0.55	0.56	0.56	0.53	0.52	0.52	0.53	0.51	0.55	0.52	0.51	0.61	0.58	0.65
eu fds	0.78	1.00	0.80	0.83	0.76	0.71	0.79	0.80	0.76	0.75	0.76	0.76	0.77	0.68	0.77	0.70	0.74	0.79	0.81	0.81
eu agora	0.61	0.80	1.00	0.89	0.86	0.86	0.89	0.89	0.89	0.91	0.89	0.89	0.89	0.86	0.88	0.86	0.86	0.91	0.89	0.87
eu agr	0.62	0.83	0.89	1.00	0.87	0.85	0.86	0.87	0.84	0.85	0.82	0.84	0.83	0.80	0.83	0.80	0.82	0.88	0.88	0.86
eu eh	0.54	0.76	0.86	0.87	1.00	0.93	0.87	0.86	0.82	0.87	0.85	0.88	0.86	0.81	0.84	0.78	0.82	0.84	0.87	0.84
eu é	0.52	0.71	0.86	0.85	0.93	1.00	0.82	0.81	0.81	0.86	0.81	0.84	0.81	0.83	0.81	0.83	0.82	0.83	0.81	0.82
eu mto	0.55	0.79	0.89	0.86	0.87	0.82	1.00	0.97	0.94	0.90	0.90	0.92	0.91	0.84	0.91	0.79	0.82	0.88	0.91	0.84
eu mt	0.56	0.80	0.89	0.87	0.86	0.81	0.97	1.00	0.94	0.90	0.92	0.92	0.93	0.82	0.91	0.79	0.83	0.87	0.91	0.84
eu muito	0.56	0.76	0.89	0.84	0.82	0.81	0.94	0.94	1.00	0.88	0.84	0.86	0.86	0.80	0.84	0.81	0.80	0.88	0.86	0.83
eu também	0.53	0.75	0.91	0.85	0.87	0.86	0.90	0.90	0.88	1.00	0.94	0.97	0.94	0.89	0.91	0.82	0.82	0.86	0.87	0.83
eu tbm	0.52	0.76	0.89	0.82	0.85	0.81	0.90	0.92	0.84	0.94	1.00	0.97	0.99	0.84	0.93	0.78	0.82	0.84	0.88	0.80
eu tambem	0.52	0.76	0.89	0.84	0.88	0.84	0.92	0.92	0.86	0.97	0.97	1.00	0.97	0.87	0.92	0.78	0.82	0.84	0.88	0.82
eu tb	0.53	0.77	0.89	0.83	0.86	0.81	0.91	0.93	0.86	0.94	0.99	0.97	1.00	0.83	0.92	0.79	0.82	0.85	0.89	0.80
eu mesmo	0.51	0.68	0.86	0.80	0.81	0.83	0.84	0.82	0.80	0.89	0.84	0.87	0.83	1.00	0.90	0.80	0.79	0.83	0.82	0.81
eu msm	0.55	0.77	0.88	0.83	0.84	0.81	0.91	0.91	0.84	0.91	0.93	0.92	0.92	0.90	1.00	0.77	0.80	0.86	0.89	0.83
eu você	0.52	0.70	0.86	0.80	0.78	0.83	0.79	0.79	0.81	0.82	0.78	0.78	0.79	0.80	0.77	1.00	0.93	0.80	0.77	0.79
eu vc	0.51	0.74	0.86	0.82	0.82	0.82	0.82	0.83	0.80	0.82	0.82	0.82	0.82	0.79	0.80	0.93	1.00	0.79	0.81	0.78
eu porque	0.61	0.79	0.91	0.88	0.84	0.83	0.88	0.87	0.88	0.86	0.84	0.84	0.85	0.83	0.86	0.80	0.79	1.00	0.95	0.93
eu pq	0.58	0.81	0.89	0.88	0.87	0.81	0.91	0.91	0.86	0.87	0.88	0.88	0.89	0.82	0.89	0.77	0.81	0.95	1.00	0.90
eu porquê	0.65	0.81	0.87	0.86	0.84	0.82	0.84	0.84	0.83	0.83	0.80	0.82	0.80	0.81	0.83	0.79	0.78	0.93	0.90	1.00

Figura 34. Matriz de pares de termos - painel 1.

	eu final de semana	eu fds	eu agora	eu agr	eu eh	eu é	eu mto	eu mt	eu muito	eu também	eu tbm	eu tambem	eu tb	eu mesmo	eu msm	eu você	eu vc	eu porque	eu pq	eu porquê
eu final de semana	1.00	0.89	0.85	0.83	0.83	0.84	0.82	0.83	0.86	0.78	0.81	0.80	0.81	0.80	0.85	0.63	0.85	0.81	0.84	0.81
eu fds	0.89	1.00	0.87	0.87	0.85	0.88	0.83	0.85	0.87	0.77	0.82	0.84	0.82	0.82	0.85	0.68	0.88	0.83	0.86	0.84
eu agora	0.85	0.87	1.00	0.86	0.89	0.89	0.88	0.89	0.89	0.81	0.87	0.87	0.86	0.84	0.87	0.65	0.89	0.84	0.91	0.84
eu agr	0.83	0.87	0.86	1.00	0.88	0.88	0.86	0.85	0.85	0.76	0.84	0.86	0.84	0.79	0.88	0.63	0.87	0.80	0.87	0.81
eu eh	0.83	0.85	0.89	0.88	1.00	0.90	0.87	0.87	0.88	0.80	0.86	0.87	0.86	0.81	0.86	0.61	0.89	0.82	0.89	0.83
eu é	0.84	0.88	0.89	0.88	0.90	1.00	0.86	0.85	0.89	0.77	0.83	0.87	0.83	0.81	0.86	0.65	0.89	0.86	0.87	0.85
eu mto	0.82	0.83	0.88	0.86	0.87	0.86	1.00	0.93	0.90	0.84	0.89	0.89	0.89	0.83	0.87	0.59	0.85	0.82	0.90	0.81
eu mt	0.83	0.85	0.89	0.85	0.87	0.85	0.93	1.00	0.89	0.86	0.90	0.88	0.90	0.81	0.87	0.59	0.87	0.80	0.89	0.81
eu muito	0.86	0.87	0.89	0.85	0.88	0.89	0.90	0.89	1.00	0.80	0.84	0.85	0.85	0.81	0.85	0.66	0.86	0.85	0.89	0.85
eu também	0.78	0.77	0.81	0.76	0.80	0.77	0.84	0.86	0.80	1.00	0.90	0.88	0.87	0.81	0.82	0.54	0.79	0.74	0.84	0.73
eu tbm	0.81	0.82	0.87	0.84	0.86	0.83	0.89	0.90	0.84	0.90	1.00	0.91	0.92	0.84	0.88	0.55	0.84	0.77	0.87	0.76
eu tambem	0.80	0.84	0.87	0.86	0.87	0.87	0.89	0.88	0.85	0.88	0.91	1.00	0.88	0.86	0.89	0.60	0.85	0.81	0.87	0.81
eu tb	0.81	0.82	0.86	0.84	0.86	0.83	0.89	0.90	0.85	0.87	0.92	0.88	1.00	0.81	0.87	0.55	0.84	0.77	0.87	0.77
eu mesmo	0.80	0.82	0.84	0.79	0.81	0.81	0.83	0.81	0.81	0.81	0.84	0.86	0.81	1.00	0.87	0.61	0.81	0.80	0.84	0.81
eu msm	0.85	0.85	0.87	0.88	0.86	0.86	0.87	0.87	0.85	0.82	0.88	0.89	0.87	0.87	1.00	0.62	0.87	0.80	0.88	0.83
eu você	0.63	0.68	0.65	0.63	0.61	0.65	0.59	0.59	0.66	0.54	0.55	0.60	0.55	0.61	0.62	1.00	0.67	0.66	0.63	0.70
eu vc	0.85	0.88	0.89	0.87	0.89	0.89	0.85	0.87	0.86	0.79	0.84	0.85	0.84	0.81	0.87	0.67	1.00	0.82	0.89	0.83
eu porque	0.81	0.83	0.84	0.80	0.82	0.86	0.82	0.80	0.85	0.74	0.77	0.81	0.77	0.80	0.80	0.66	0.82	1.00	0.85	0.87
eu pq	0.84	0.86	0.91	0.87	0.89	0.87	0.90	0.89	0.89	0.84	0.87	0.87	0.87	0.84	0.88	0.63	0.89	0.85	1.00	0.85
eu porquê	0.81	0.84	0.84	0.81	0.83	0.85	0.81	0.81	0.85	0.73	0.76	0.81	0.77	0.81	0.83	0.70	0.83	0.87	0.85	1.00

Figura 35. Matriz de pares de termos - painel 2.

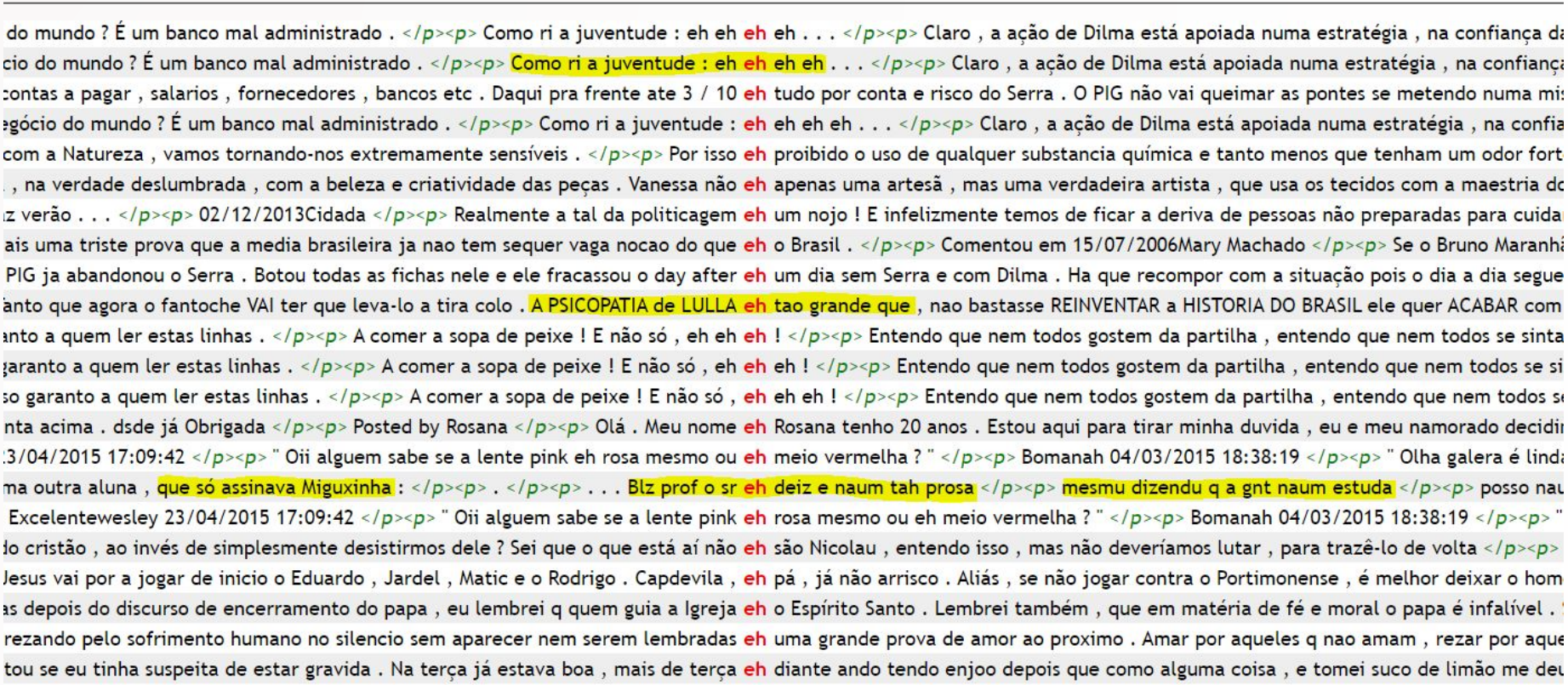


Figura 36. Exemplos de trechos e marcas textuais no BrWaC.

Parâmetro	Valor
Unidade de privacidade	Conversa
Amostragem	Poisson, q = 0,04
Clipping	Norma máxima C = 1
δ	10^{-5}
Passos	100 por rodada; 2.000 no total
ε nominal 3	σ = 2,81; ε realizado = 2,987777
ε nominal 8	σ = 1,36; ε realizado = 7,964314
Contabilização	Independente por vítima; maior ε entre as dez