



AULÃO 4 · CAPÍTULO 4 DA IMERSÃO

Linguagem natural

Na Fase 1 o texto virava número contando palavras. Palavra nunca vista sumia, e erro de digitação virava palavra desconhecida. Isso acabou.



COMO LER ESTE DOCUMENTO

Este PDF é o registro do Aulão 4. Cada página mostra uma tela do material interativo já aberta, num estado escolhido, e a faixa de baixo explica o que está acontecendo ali. Os laboratórios aparecem em vários estados, para vocês compararem o que muda quando um controle se mexe — é justamente o que não cabe numa imagem só. O material vivo, com todos os laboratórios para mexer, está no endereço da última página.

Palavra que o modelo não conhece

A mesma frase, nas duas fases. Troquem a frase e reparem no que some.

frase comum

com erro de digitação

com gíria

FASE 1 · CONTANDO PALAVRAS

o produto chegou certo e o atendimento foi bom

0 palavras jogadas fora por não estarem na lista. Com elas, foi embora o sentido.

FASE 2 · PEDAÇOS DE PALAVRA

o produto chegou certo e o atendimento foi bom

9 pedaços, e **nenhuma palavra jogada fora**. O modelo entende uma palavra que nunca viu porque entende as partes.

A mesma frase, lida das duas maneiras. Em cima, o jeito da Fase 1: contar palavras de uma lista fechada. Embaixo, o jeito desta fase: partir o texto em pedaços de palavra. Numa frase bem comportada os dois funcionam, e é por isso que esta página é o ponto de partida — a diferença só aparece quando a frase sai do comportado.

Palavra que o modelo não conhece

A mesma frase, nas duas fases. Troquem a frase e reparem no que some.

frase comum

com erro de digitação

com gíria

FASE 1 · CONTANDO PALAVRAS

o **preeto** chegou muito rápido, **recomendissimo**

2 palavras jogadas fora por não estarem na lista. Com elas, foi embora o sentido.

FASE 2 · PEDAÇOS DE PALAVRA

o **prod** · **##to** chegou muito rápido · **,** **recomend** ·
##issimo

9 pedaços, e **nenhuma palavra jogada fora**. O modelo entende uma palavra que nunca viu porque entende as partes.

Agora com um erro de digitação. Na Fase 1 a palavra errada simplesmente não está na lista e vai para o lixo — e com ela vai parte do sentido. Na Fase 2 ela é partida em pedaços que o modelo conhece, e nada se perde. Repare no contador dos dois lados: é a mesma frase, e um dos métodos jogou informação fora.

Palavra que o modelo não conhece

A mesma frase, nas duas fases. Troquem a frase e reparem no que some.

frase comum

com erro de digitação

com gíria

FASE 1 · CONTANDO PALAVRAS

chegou tudo ~~certinho,~~ top demais, ~~recomende~~ mto

5 palavras jogadas fora por não estarem na lista. Com elas, foi embora o sentido.

FASE 2 · PEDAÇOS DE PALAVRA

chegou tudo cert · ##inho · , top demais · ,
recomendo m · ##to

11 pedaços, e **nenhuma palavra jogada fora**. O modelo entende uma palavra que nunca viu porque entende as partes.

O caso extremo, e o que mais importa num acervo de texto escrito por gente de verdade: gíria, abreviação, palavra inventada. O modelo entende uma palavra que nunca viu porque entende as partes dela. É esta página que justifica tudo o que vem depois no caderno.

Os quatro blocos do desafio

A trilha é a mesma das outras duas. Cliquem num bloco para ver o que é novo aqui.

01 · EXPLORAR

02 · CONFIGURAR

03 · MODELAR

04 · AVALIAR

O QUE JÁ ERA ASSIM

Contar quantos textos tem cada classe e olhar as palavras mais frequentes.

O QUE É NOVO AGORA

A mesma contagem — e mais duas perguntas que só existem aqui: **como o texto é partido** em pedaços de palavra, e **quanto texto cabe** no limite do modelo.

Novo aqui: pedaços de palavra e o corte em 128. Os dois laboratórios a seguir são deste bloco.

O caderno da trilha tem quatro blocos, e só um deles tem código para escrever. No primeiro vocês olham o corpus antes de treinar qualquer coisa: como ele é, quanto vale um modelo que não aprende nada, e o que acontece quando se tenta resolver contando palavras iguais. A resposta dessa última pergunta decide o resto.

Os quatro blocos do desafio

A trilha é a mesma das outras duas. Cliquem num bloco para ver o que é novo aqui.

01 · EXPLORAR

02 · CONFIGURAR

03 · MODELAR

04 · AVALIAR

O QUE JÁ ERA ASSIM

Escolher o modelo e mexer nos botões de cada um.

O QUE É NOVO AGORA

Esta é a trilha mais simples de configurar, e por um motivo bonito: **o tradutor e o modelo são carregados pelo mesmo nome**. Trocar de modelo é trocar uma string.

Novo aqui: a regra que não pode ser quebrada — o tradutor e o modelo têm de ser o mesmo nome.

No segundo bloco não se escreve código: escolhe-se o modelo de linguagem e giram-se os botões. O que vocês decidirem aqui é o que as lacunas do Bloco 3 vão buscar.

Os quatro blocos do desafio

A trilha é a mesma das outras duas. Cliquem num bloco para ver o que é novo aqui.

01 · EXPLORAR

02 · CONFIGURAR

03 · MODELAR

04 · AVALIAR

O QUE JÁ ERA ASSIM

Chamar `.fit` e esperar.

O QUE É NOVO AGORA

As **mesmas três células**, pela terceira vez: A prepara, B treina, C avalia. Repararam que a forma nunca muda?

É aqui que ficam as lacunas do desafio. Todas elas.

O terceiro bloco é onde está todo o código a completar. Cada lacuna tem, logo acima, um comentário dizendo o que aquela linha precisa fazer e onde o assunto foi tratado. A pergunta que sobra para vocês é sempre o *como*.

Os quatro blocos do desafio

A trilha é a mesma das outras duas. Cliquem num bloco para ver o que é novo aqui.

01 · EXPLORAR

02 · CONFIGURAR

03 · MODELAR

04 · AVALIAR

O QUE JÁ ERA ASSIM

Olhar a métrica e a matriz de confusão.

O QUE É NOVO AGORA

Com **duas classes**, as quatro casas da matriz ganham nomes próprios — e os dois tipos de erro **não custam a mesma coisa**. Quem decide isso é gente, não métrica.

Novo aqui: F1 por grupo, os nomes do falso positivo e do falso negativo, e o recall como métrica de apoio.

O quarto bloco roda sozinho e é onde metade do relatório nasce: as curvas, a matriz de confusão com nome próprio, o desempenho por grupo de texto e a conversa sobre qual erro custa mais. A cliente da narrativa pediu uma coisa só, e é aqui que ela se responde.

01

BLOCO 1

Explorar

- Quantos textos tem cada classe, e como eles variam
- **Pedaços de palavra:** três tradutores, uma frase
- **Onde cortar:** quanto cabe em 128 pedaços



ABERTURA DO BLOCO 1

A partir daqui, os dois laboratórios que tratam do que acontece com o texto antes de o modelo ver qualquer coisa: como cada modelo parte a frase, e onde ele corta o que não couber.

Três tradutores, uma frase

Cada modelo parte o texto do seu próprio jeito. Não em palavras: em **pedaços de palavra**.

frase comum

com erro de digitação

com gíria

simulação do comportamento

DistilBERT multiling. 12 PEDAÇOS · VOCABULÁRIO DE 119 mil

o produto cheg ##ou certo e o atendim ##en ##to foi bom

Fala 104 idiomas e nenhum muito bem. Parte muito o português.

BERTimbau 9 PEDAÇOS · VOCABULÁRIO DE 30 mil

o produto chegou certo e o atendimento foi bom

Treinado só em português brasileiro. Conhece palavras nossas inteiras.

XLNet 10 PEDAÇOS · VOCABULÁRIO DE 250 mil

o produto cheg ##ou certo e o atendimento foi bom

Generalista em cem idiomas. Vocabulário gigante, e é dele boa parte do peso.

Os três modelos da trilha partindo a mesma frase. Repare que o número de pedaços já é diferente numa frase simples, e que o tamanho do vocabulário de cada um explica boa parte disso: 30 mil no especialista em português, 119 mil e 250 mil nos generalistas. Vocabulário maior não significa partir menos — significa cobrir mais idiomas.

Três tradutores, uma frase

Cada modelo parte o texto do seu próprio jeito. Não em palavras: em **pedaços de palavra**.

frase comum

com erro de digitação

com gíria

simulação do comportamento

DistilBERT multiling. 11 PEDAÇOS · VOCABULÁRIO DE 119 mil

o prod ##to cheg ##ou muito rápi ##do , recomend ##issimo

Fala 104 idiomas e nenhum muito bem. Parte muito o português.

BERTimbau 9 PEDAÇOS · VOCABULÁRIO DE 30 mil

o prod ##to chegou muito rápido , recomend ##issimo

Treinado só em português brasileiro. Conhece palavras nossas inteiras.

XLm-ROBERTa 10 PEDAÇOS · VOCABULÁRIO DE 250 mil

o prod ##to cheg ##ou muito rápido , recomend ##issimo

Generalista em cem idiomas. Vocabulário gigante, e é dele boa parte do peso.

A mesma comparação com a palavra escrita errada. Nenhum dos três joga nada fora: todos partem em pedaços conhecidos. O que muda é em quantos — e cada pedaço a mais é espaço gasto dentro do limite de 128 que o próximo laboratório vai discutir.

Três tradutores, uma frase

Cada modelo parte o texto do seu próprio jeito. Não em palavras: em **pedaços de palavra**.

frase comum

com erro de digitação

com gíria

simulação do comportamento

DistilBERT multiling. 13 PEDAÇOS · VOCABULÁRIO DE 119 mil

cheg ##ou tudo cert ##inho , top demais , recome ##ndo m ##to

Fala 104 idiomas e nenhum muito bem. Parte muito o português.

BERTimbau 11 PEDAÇOS · VOCABULÁRIO DE 30 mil

chegou tudo cert ##inho , top demais , recomendo m ##to

Treinado só em português brasileiro. Conhece palavras nossas inteiras.

XLm-ROBERTa 12 PEDAÇOS · VOCABULÁRIO DE 250 mil

cheg ##ou tudo cert ##inho , top demais , recomendo m ##to

Generalista em cem idiomas. Vocabulário gigante, e é dele boa parte do peso.

Aqui a diferença entre os três fica mais visível. O modelo treinado só em português brasileiro reconhece palavras nossas inteiras, enquanto os generalistas precisam quebrá-las. Isso é argumento a favor dele — e não é o único critério, porque custo de memória e tempo de treino também contam.

Quanto cabe em 128 pedaços

O modelo lê um número fixo de pedaços por texto e corta o resto. Mover essa linha tem preço dos dois lados.

limite

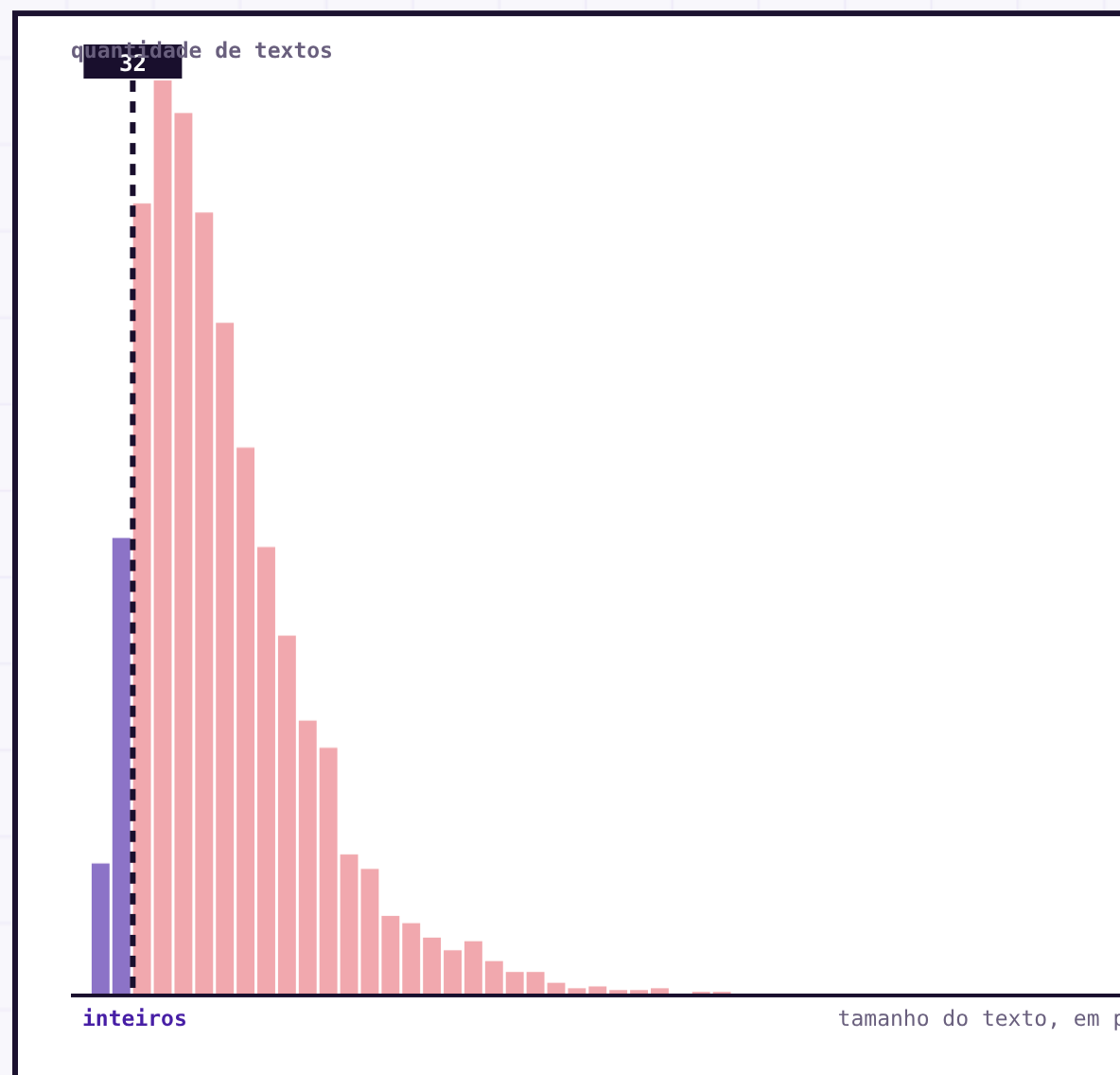
32

PERDEM TEXTO NO CORTE

92%

dos textos do conjunto.

O custo de cada passo cresce com o **quadrado** deste número.



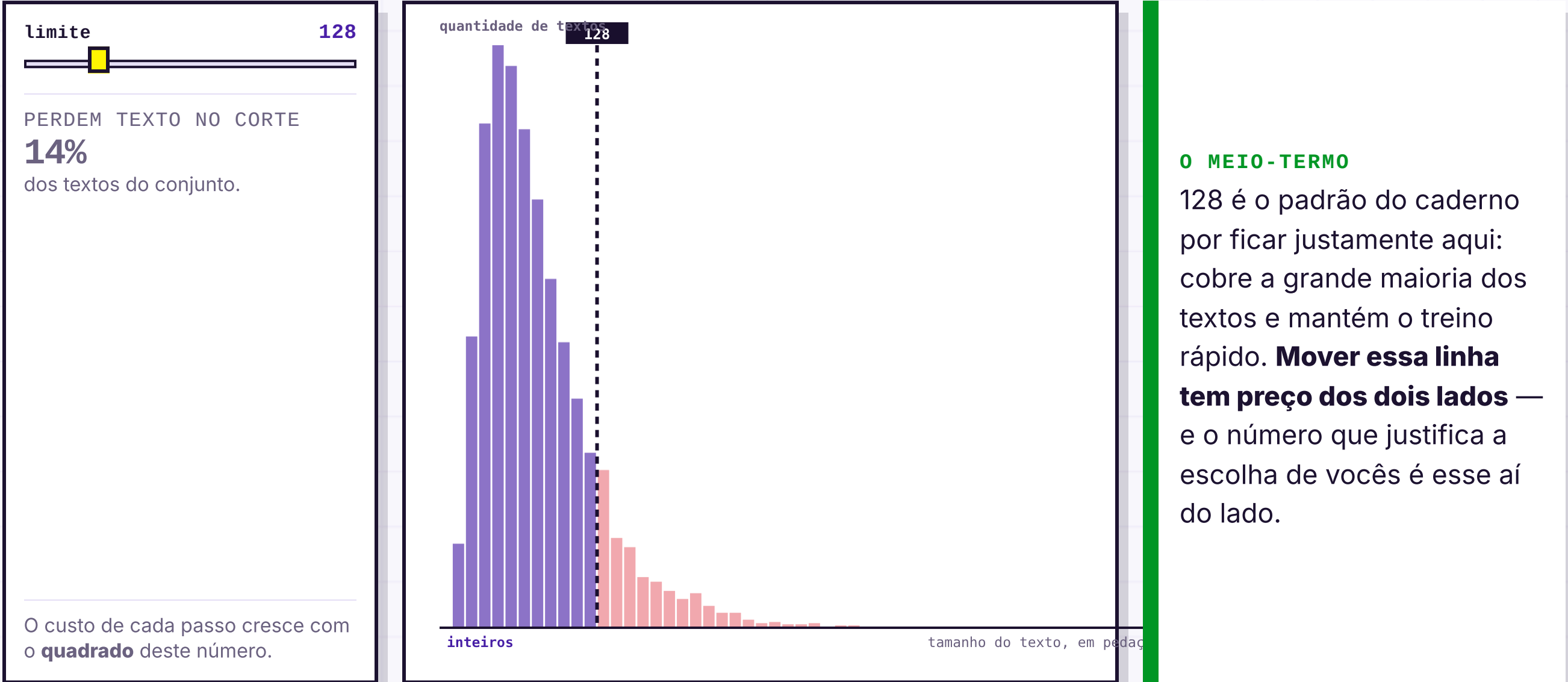
CURTO DEMAIS

Mais de um terço dos textos perde o fim. E o fim costuma ser onde está a conclusão — justamente o que decide a resposta.

O limite no mínimo. O treino fica rápido e barato, e boa parte dos textos é decapitada no meio. Em pares de frases isso é fatal: se a segunda frase for cortada, o modelo decide sem ter lido o que precisava comparar.

Quanto cabe em 128 pedaços

O modelo lê um número fixo de pedaços por texto e corta o resto. Mover essa linha tem preço dos dois lados.



O valor que o material adota. Cobre a grande maioria dos textos e mantém o treino rápido. O ponto da página não é decorar o 128 — é entender que ele foi **medido**, olhando quantos textos do corpus cabem, e não escolhido por tradição.

Quanto cabe em 128 pedaços

O modelo lê um número fixo de pedaços por texto e corta o resto. Mover essa linha tem preço dos dois lados.

limite512

PERDEM TEXTO NO CORTE

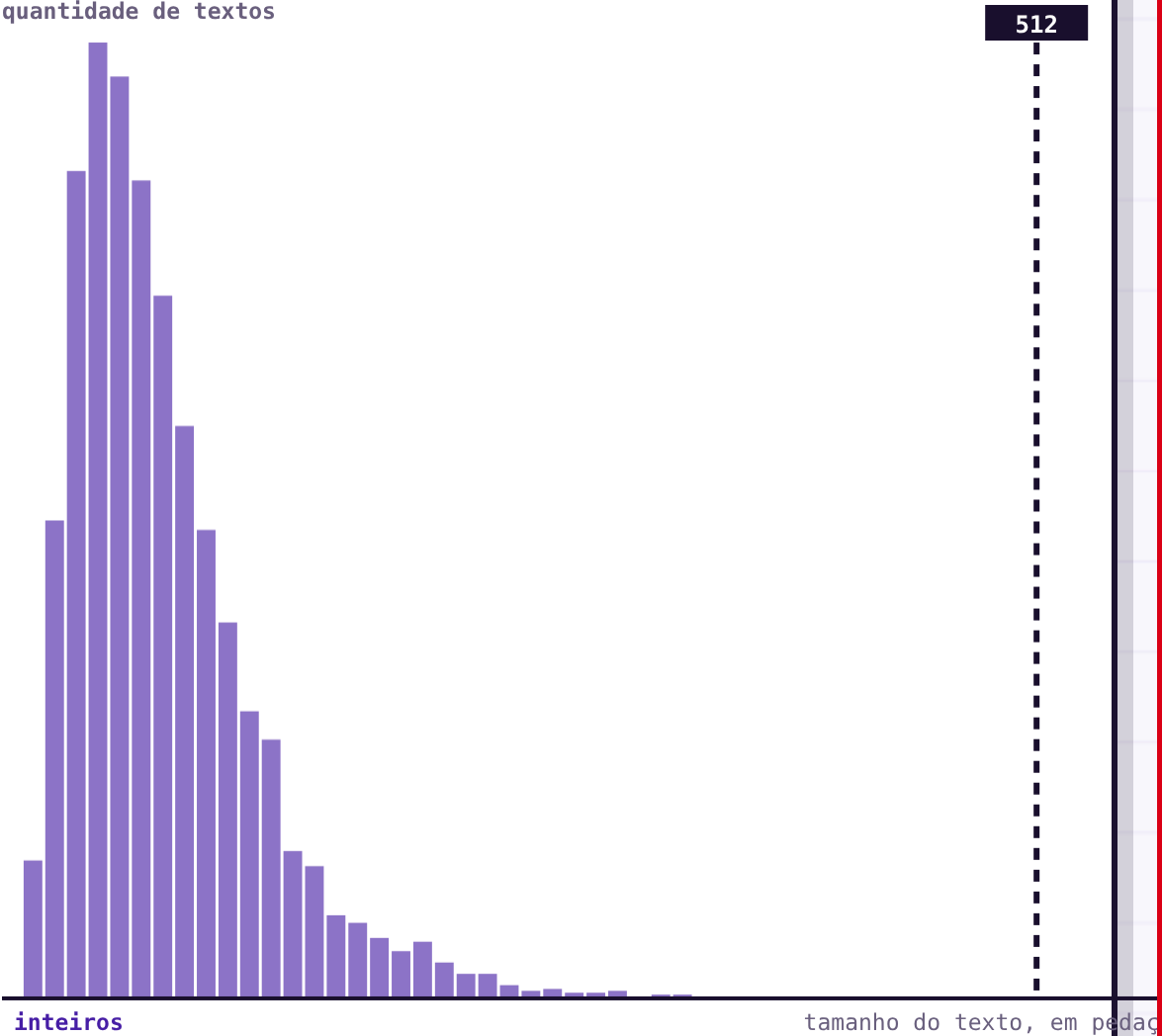
0%

dos textos do conjunto.

O custo de cada passo cresce com o **quadrado** deste número.

quantidade de textos

512



inteirostamanho do texto, em pedaços

LONGO DEMAIS

Quase nada é cortado, e cada passo de treino fica caríssimo: o custo cresce com o **quadrado** do limite. Vocês vão esperar muito para ganhar pouco.

O limite no máximo. Nenhum texto é cortado, e o preço aparece no tempo e na memória — o custo cresce rápido com o comprimento. Comparem as três páginas: mover essa linha tem preço dos dois lados, e a decisão certa depende do corpus de vocês, não de um número universal.

02

BLOCO 2

Configurar

- Três modelos: um **especialista** e dois generalistas
- Trocar de modelo é trocar **uma palavra**
- A regra que não pode ser quebrada
- Quatro botões, e só quatro



ABERTURA DO BLOCO 2

Daqui em diante, as três decisões do bloco de configuração: qual dos três modelos, a regra que não se quebra ao usá-los, e os quatro botões do treino.

Um especialista e dois generalistas

Os três já sabem português. A pergunta é quanto, e a que custo.

DistilBERT multilíngue

BERTimbau

XLNet

ENTRADA

Um modelo grande que foi encolhido para a metade das camadas, imitando o original.

Escolham quando vocês querem espaço para testar várias configurações em vez de esperar treino.

taxa	3e-5 a 5e-5
lote	32 ou 64
épocas	3 a 4
camadas	6

A PEGADINHA DESTA

Ele fala **104 idiomas e nenhum muito bem**. Perde para o especialista em português — e medir essa perda é parte do trabalho.

O QUE ESPERAR

Cerca de duas vezes mais rápido que os outros. Bom para as primeiras tentativas.

OS TRÊS MODELOS · DISTILBERT MULTILÍNGUE

O mais leve dos três: fala 104 idiomas e nenhum muito bem. Treina rápido, cabe com folga na GPU gratuita e costuma dar o resultado mais modesto. É a escolha sensata para a equipe que quer uma primeira volta completa antes de arriscar.

Um especialista e dois generalistas

Os três já sabem português. A pergunta é quanto, e a que custo.

DistilBERT multilíngue

BERTimbau

XLNet-RoBERTa

ESPECIALISTA

Um modelo treinado só em português brasileiro, de textos da web do Brasil.

Escolham quando vocês querem o melhor resultado pelo menor risco. É a escolha natural desta trilha.

taxa	2e-5
lote	32
épocas	3
camadas	12

A PEGADINHA DESTA

Ele diferencia maiúscula de minúscula e usa os acentos. **Não passem o texto para minúsculas como na Fase 1:** isso destrói informação que ele usa.

O QUE ESPERAR

Tende a dar o melhor número da trilha. Conceitualmente, não traz nada de novo.

O especialista: treinado só em português brasileiro. Tende a dar o melhor número da trilha pelo menor risco. A pegadinha está escrita na própria ficha — ele usa maiúsculas e acentos, então passar o texto para minúsculas, como se fazia na Fase 1, destrói informação que ele usa.

Um especialista e dois generalistas

Os três já sabem português. A pergunta é quanto, e a que custo.

DistilBERT multilíngue

BERTimbau

XLM-RoBERTa

AVANÇADA

Um generalista treinado em 2,5 terabytes de texto em cem idiomas.

Escolham quando vocês querem testar a hipótese: será que muito mais treino compensa não ser especialista?

taxa	1e-5 a 2e-5
lote	16 ou 32
épocas	3
camadas	12

A PEGADINHA DESTA

Ele **não** é mais fundo que o BERTimbau: tem as mesmas 12 camadas. É mais pesado por causa do vocabulário de 250 mil pedaços, que sozinho responde por boa parte do tamanho.

O QUE ESPERAR

O mais pesado e o mais sensível à taxa. Pode ganhar, pode divergir no treino.

O generalista grande: cem idiomas e vocabulário de 250 mil, que é de onde vem boa parte do peso dele. Exige mais memória e é o mais sensível a configuração — em especial à taxa de aprendizado e ao aquecimento, como as curvas vão mostrar no Bloco 4.

0 tradutor e o modelo saem do mesmo nome

Escolham um tradutor e um modelo. Vejam o que chega do outro lado.

TRADUTOR

DistilBERT multiling.

BERTimbau

XLM-RoBERTa

MODELO

DistilBERT multiling.

BERTimbau

XLM-RoBERTa

1 · O TRADUTOR PARTE A FRASE

o produto chegou rapidinho

2 · CADA PEDAÇO VIRA UM NÚMERO

1006 3174 12865 7108

3 · O MODELO LÊ ESSES NÚMEROS NA TABELA DELE

o produto chegou rapidinho

✓ MESMO NOME DOS DOIS LADOS

Cada número chega no lugar certo, e o modelo lê a frase que vocês escreveram. **É a única regra inegociável desta trilha** — e é de graça: o tradutor certo vem junto, automaticamente, quando se usa o mesmo nome.

O tradutor parte a frase, cada pedaço vira um número, e o modelo lê esses números na tabela dele. Com o mesmo nome dos dois lados, cada número chega no lugar certo e o modelo lê a frase que vocês escreveram. É a única regra inegociável da trilha — e sai de graça, porque o tradutor certo vem junto quando se usa o mesmo nome.

O tradutor e o modelo saem do mesmo nome

Escolham um tradutor e um modelo. Vejam o que chega do outro lado.

TRADUTOR

DistilBERT multiling.

BERTimbau

XLM-RoBERTa

MODELO

DistilBERT multiling.

BERTimbau

XLM-RoBERTa

1 · O TRADUTOR PARTE A FRASE

o produto cheg ##ou rapid ##inho

2 · CADA PEDAÇO VIRA UM NÚMERO

1006 3174 12865 7108 15441 23056

3 · O MODELO LÊ ESSES NÚMEROS NA TABELA DELE

##ado ##mente três ##inho casa ##ado

× TRADUTOR E MODELO DIFERENTES

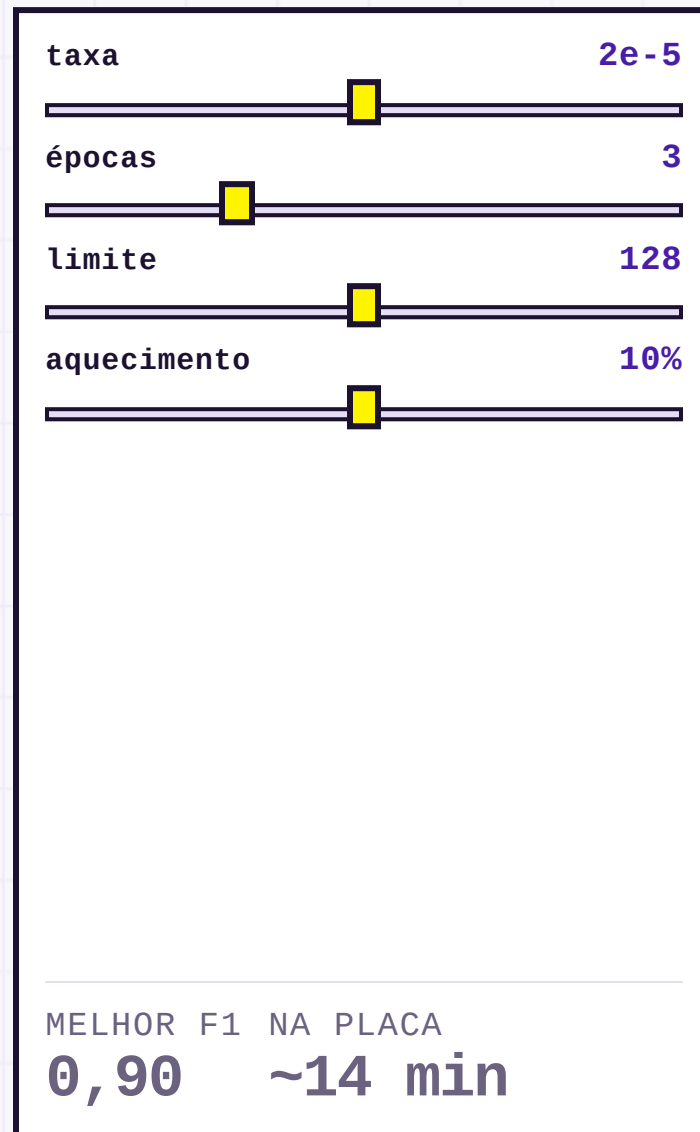
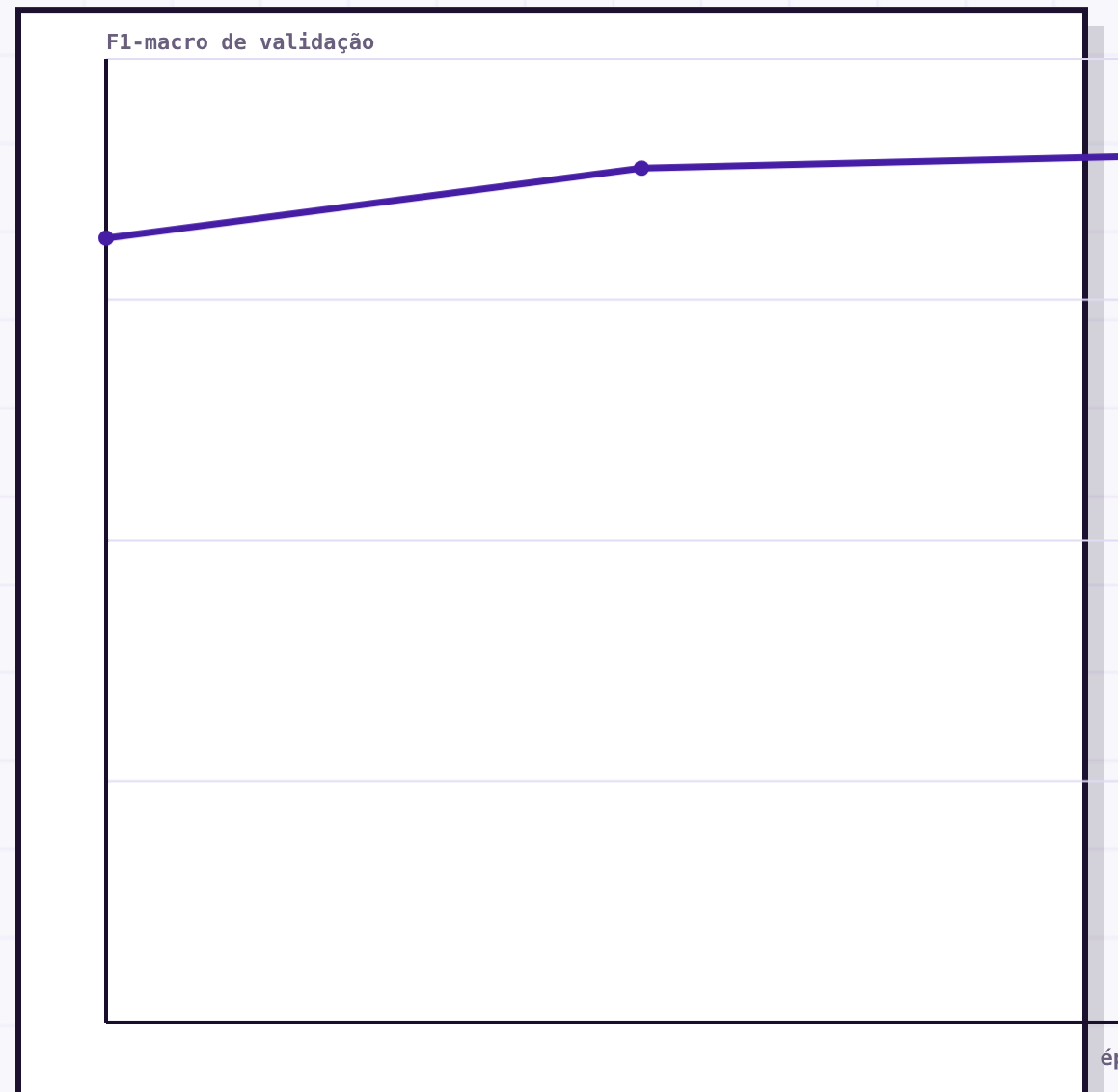
Os pedaços foram numerados pela tabela de um modelo e lidos na tabela de outro. O modelo recebeu a linha de baixo. **O código roda, não dá erro nenhum**, e o resultado despenca.

É o erro campeão desta trilha, e a atividade de caça ao erro do Bloco 3 é exatamente sobre ele.

Agora o tradutor de um modelo com os pesos de outro. Os pedaços não são os que o modelo aprendeu, então cada número chega numa linha errada da tabela e ele lê uma frase que ninguém escreveu. **Isto não dá erro**. O código roda inteiro, a barra de progresso anda, e o resultado despenca sem nenhuma mensagem para explicar por quê.

Quatro botões, e só quatro

Não existem mil coisas para ajustar. Existem estas — e a curva responde na hora.



FAIXA DE TRABALHO

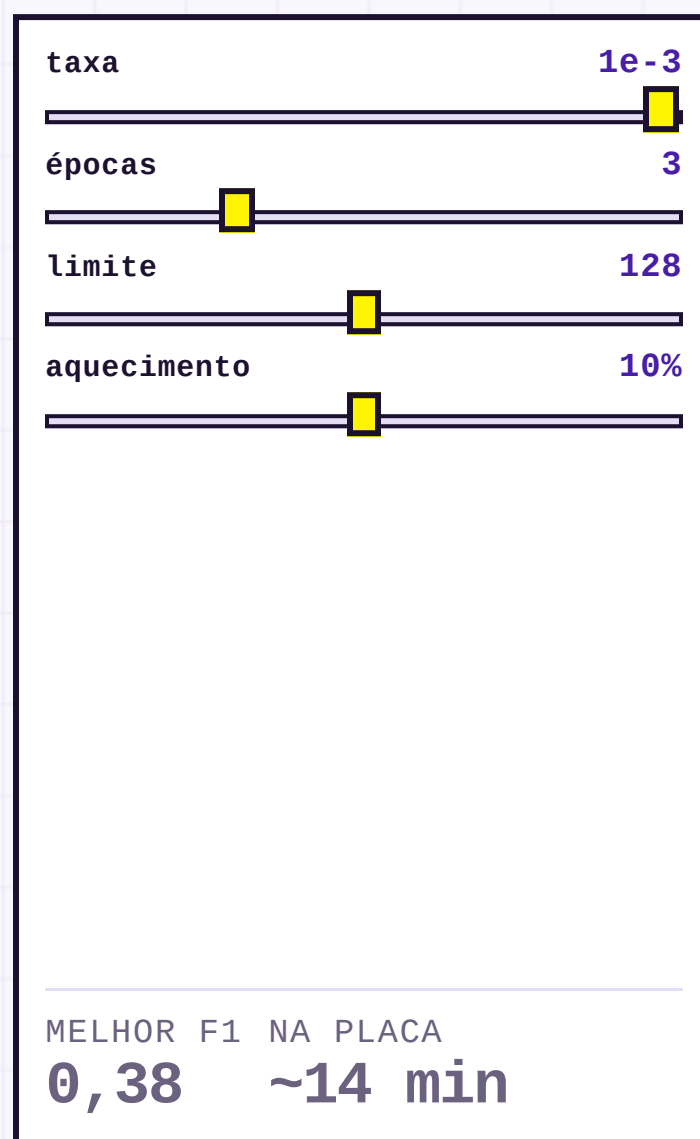
Três épocas, taxa entre $1e-5$ e $5e-5$, e o limite medido no Bloco 1. Mexam **um botão por vez** e anotem: é isso que vira a tabela de experimentos no relatório.

HIPERPARÂMETROS · A FAIXA DE TRABALHO

Os quatro botões no ponto recomendado: três épocas, taxa de $2e-5$, limite medido no Bloco 1 e aquecimento de 10%. Guardem esta página como referência — as três seguintes mostram o que cada botão estraga quando sai daqui.

Quatro botões, e só quatro

Não existem mil coisas para ajustar. Existem estas — e a curva responde na hora.



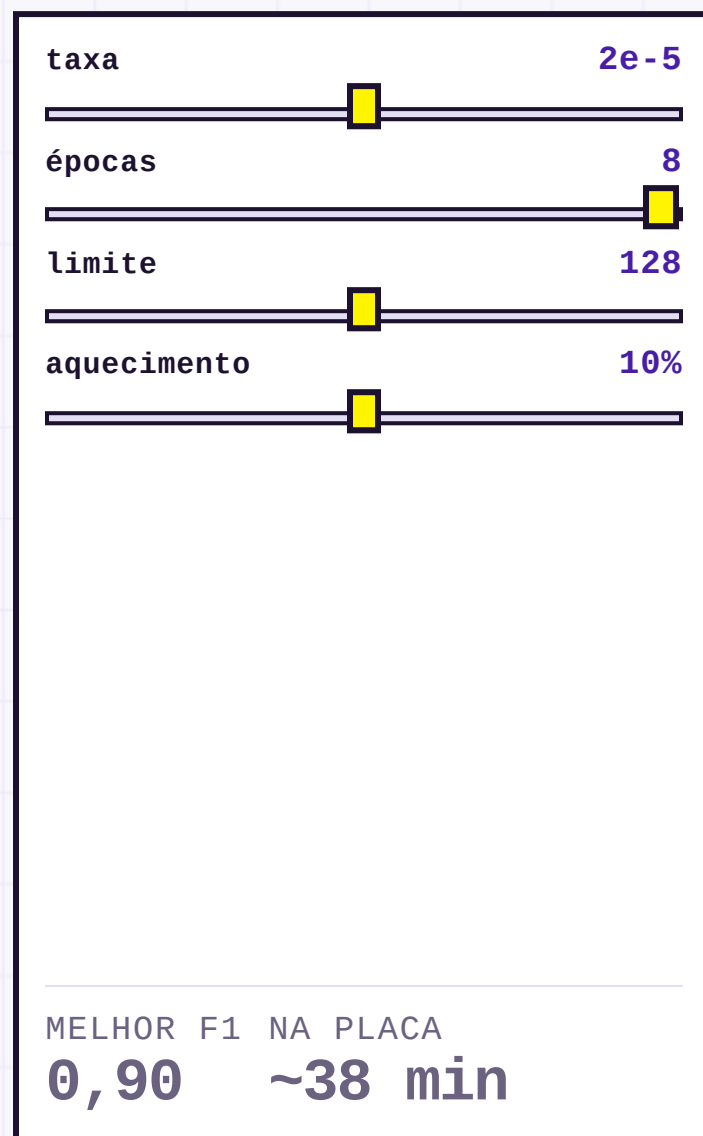
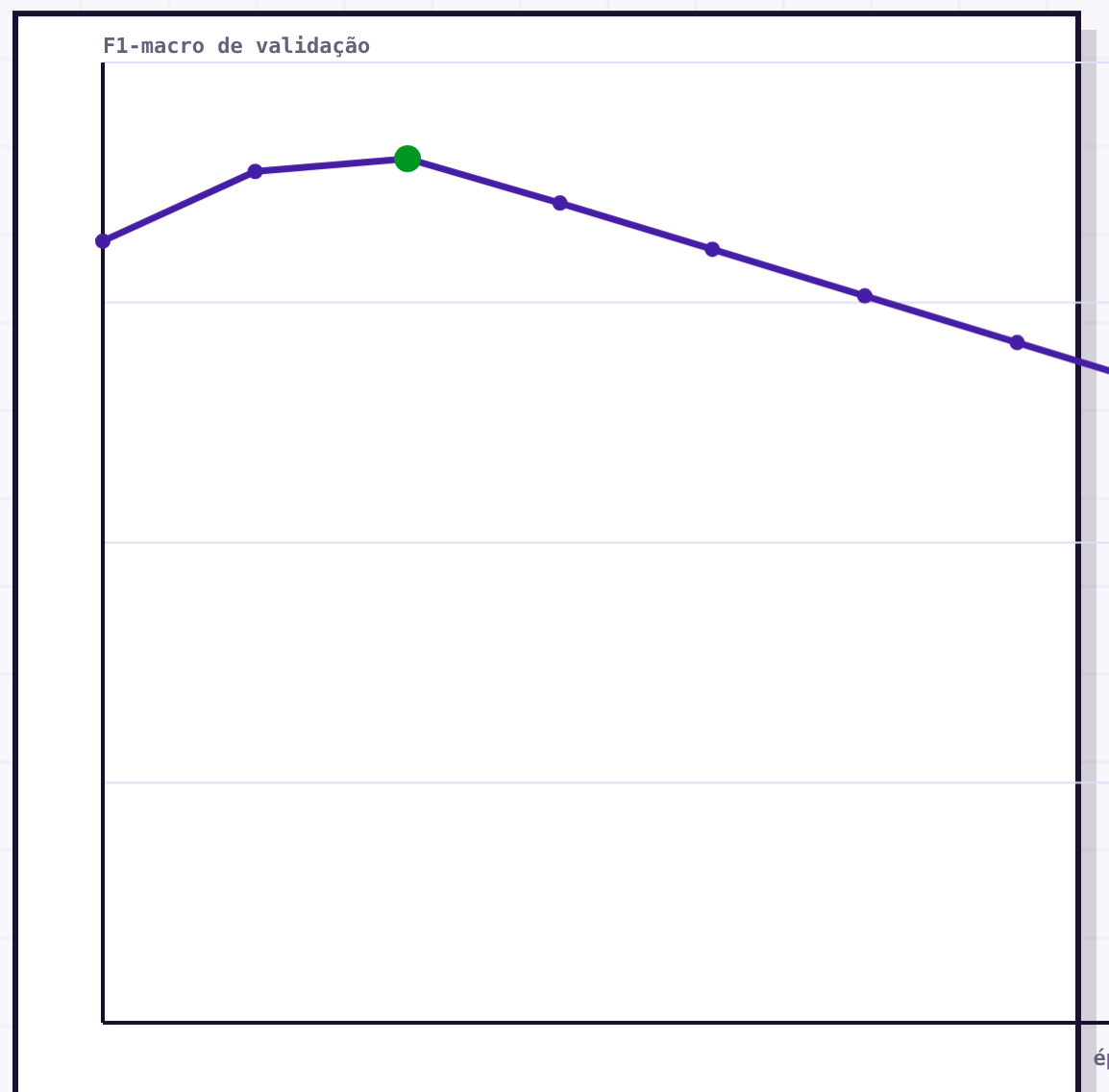
PASSO GRANDE DEMAIS

Mil vezes acima da faixa.
Apaga o pré-treino nas primeiras centenas de passos, e o que sobra não sabe português nem a tarefa.

Taxa de aprendizado no extremo de cima. Num modelo que já chega sabendo português, um passo desse tamanho não ajusta: atropela. O que ele aprendeu antes de chegar aqui é destruído nas primeiras épocas, e a perda serrilha sem descer.

Quatro botões, e só quatro

Não existem mil coisas para ajustar. Existem estas — e a curva responde na hora.



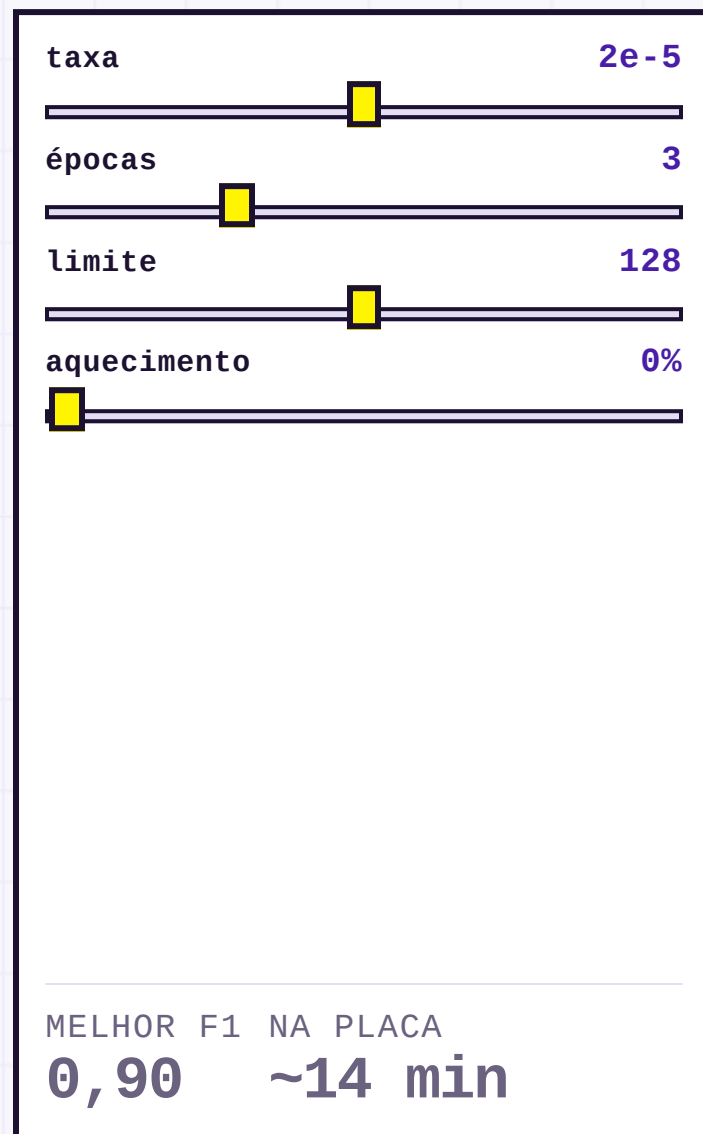
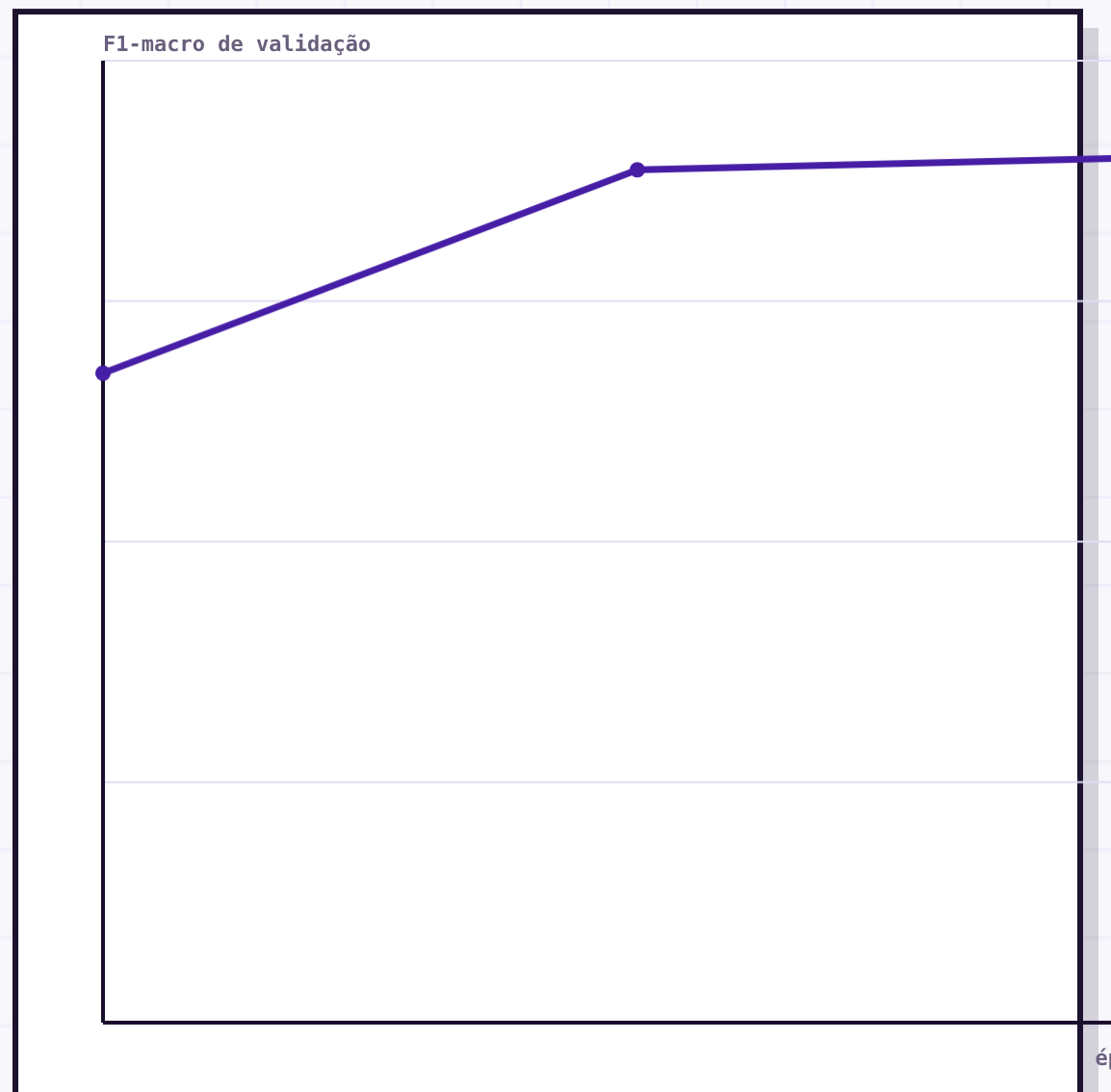
ÉPOCAS DEMAIS

Repare onde caiu o ponto verde. O modelo **já sabia português** antes de conhecer vocês: o ajuste fino não ensina a língua, ensina a tarefa. É pouca coisa — e por isso passar do ponto estraga rápido.

Oito épocas em vez de três. Esta é a armadilha mais comum do ajuste fino de modelo de linguagem: a perda de treino continua caindo e a validação já virou. Três épocas bastam, e a quarta costuma estragar — está escrito no fechamento do aulão por este motivo.

Quatro botões, e só quatro

Não existem mil coisas para ajustar. Existem estas — e a curva responde na hora.



SEM AQUECIMENTO

A primeira época sai pior. O aquecimento começa com passos minúsculos e vai crescendo — e o XLM-R é o que mais sofre sem ele.

HIPERPARÂMETROS · SEM AQUECIMENTO

Aquecimento em zero. A primeira época sai pior porque o modelo leva passos de tamanho cheio antes de se orientar. O aquecimento começa com passos minúsculos e vai crescendo — e é o XLM-R, o maior dos três, que mais sofre sem ele.

03

BLOCO 3

Modelar

- As **mesmas três células**, pela terceira vez
- Repararam que a forma nunca muda?
- **É aqui que ficam as lacunas do desafio.** Todas elas.



ABERTURA DO BLOCO 3

Este é o bloco que tem código para escrever. As quatro páginas seguintes mostram as duas atividades resolvidas, para vocês conferirem o raciocínio depois de terem tentado — não antes.

Construa o modelo que vocês escolheram

Escolham o modelo, cliquem num espaço amarelo e respondam no painel da direita.

DistilBERT multilíngue

BERTimbau

XLM-RoBERTa

7 de 7 certas · 0 em branco

LACUNA 7 DE 7 · JÁ RESPONDIDA

responder de novo

```
# == CÉLULA A · preparo ==  
NOME = "neuralmind/bert-base-portuguese-cased"  
tradutor = AutoTokenizer.from_pretrained(NOME)  
modelo = AutoModelForSequenceClassification.from_pretrained(NOME, num_labels=2)  
# este modelo usa maiúsculas e acentos. NÃO limpem o texto  
# como na Fase 1: isso destrói informação que ele usa  
textos_limpos = textos  
  
# == CÉLULA B · treino ==  
# faixa estreita: não mexam muito neste número  
otimizador = AdamW(modelo.parameters(), lr=2e-5)  
  
# == CÉLULA C · avaliação ==  
modelo.eval()  
print(f1_score(y_va, previsoes, average="macro"))
```

✓ CÉLULA COMPLETA

As 7 lacunas estão certas. Assim ela roda sem erro — e, mais importante, sem o erro silencioso que não dá mensagem nenhuma.

ATIVIDADE 3.1 · DISTILBERT (RESOLVIDO)

A célula de preparo e a de treino, com todas as lacunas preenchidas. Reparem na ordem das decisões: um nome só, usado dos dois lados; o texto entregue cru, sem a limpeza da Fase 1; o limite medido no Bloco 1; e a métrica da fase com a média que a fase escolheu. Cada uma dessas linhas é uma lacuna no caderno do desafio.

Construa o modelo que vocês escolheram

Escolham o modelo, cliquem num espaço amarelo e respondam no painel da direita.

DistilBERT multilíngue

BERTimbau

XLM-RoBERTa

7 de 7 certas · 0 em branco

LACUNA 7 DE 7 · JÁ RESPONDIDA

responder de novo

```
# == CÉLULA A · preparo ==
NOME = "neuralmind/bert-base-portuguese-cased"
tradutor = AutoTokenizer.from_pretrained(NOME)
modelo = AutoModelForSequenceClassification.from_pretrained(NOME, num_labels=2)
# este modelo usa maiúsculas e acentos. NÃO limpem o texto
# como na Fase 1: isso destrói informação que ele usa
textos_limpos = textos

# == CÉLULA B · treino ==
# faixa estreita: não mexam muito neste número
otimizador = AdamW(modelo.parameters(), lr=2e-5)

# == CÉLULA C · avaliação ==
modelo.eval()
print(f1_score(y_va, previsoes, average="macro"))
```

✓ CÉLULA COMPLETA

As 7 lacunas estão certas. Assim ela roda sem erro — e, mais importante, sem o erro silencioso que não dá mensagem nenhuma.

ATIVIDADE 3.1 · BERTIMBAU (RESOLVIDO)

A mesma atividade para o BERTimbau. Compare com a página anterior: o esqueleto é idêntico e o que muda é essencialmente o nome — porque é dele que saem, de uma vez, o tradutor e os pesos. É esse o padrão que a equipe precisa reconhecer para trocar de modelo sem reescrever a célula.

Construa o modelo que vocês escolheram

Escolham o modelo, cliquem num espaço amarelo e respondam no painel da direita.

DistilBERT multilíngue

BERTimbau

XLM-RoBERTa

6 de 6 certas · 0 em branco

LACUNA 6 DE 6 · JÁ RESPONDIDA

responder de novo

```
# == CÉLULA A · preparo ==
NOME = "xlm-roberta-base"
tradutor = AutoTokenizer.from_pretrained(NOME)
# o vocabulário deste tem 250 mil pedaços

# == CÉLULA B · treino ==
# é o mais instável dos três: passo pequeno e aquecimento longo
otimizador = AdamW(modelo.parameters(), lr=1e-5)
# o aquecimento começa com passos minúsculos e vai crescendo
agendador = get_linear_schedule_with_warmup(otimizador,
                                             num_warmup_steps=int(0.1 * total), num_training_steps=total)

# == CÉLULA C · avaliação ==
modelo.eval()
print(f1_score(y_va, previsoes, average="macro"))
```

✓ CÉLULA COMPLETA

As 6 lacunas estão certas. Assim ela roda sem erro — e, mais importante, sem o erro silencioso que não dá mensagem nenhuma.

E a terceira variação. Três modelos, o mesmo caderno, e a diferença concentrada numa linha. Quem entendeu o padrão das três páginas consegue trocar de modelo no meio do desafio sem medo.

Este código roda e o resultado despenca

Sem mensagem nenhuma. Cliquem na linha culpada.

```
NOME = "neuralmind/bert-base-portuguese-cased"

tradutor = AutoTokenizer.from_pretrained("distilbert-base-multilingual-
cased")

modelo = AutoModelForSequenceClassification.from_pretrained(NOME,
num_labels=2)

entradas = tradutor(textos, truncation=True, max_length=128)

treinador.train()
```

✓ É ESSA A LINHA

Aqui está. O tradutor é de **outro modelo**. Os pedaços de palavra que ele produz não são os que este modelo aprendeu, então cada número chega no lugar errado. O código roda e o resultado despenca, sem mensagem nenhuma.

E NO CADERNO DE VOCÊS

Antes da primeira época, o caderno compara o nome de origem do tradutor com o do modelo. Se forem diferentes, ele **para e diz o que aconteceu** — em vez de deixar vocês gastarem dias procurando no lugar errado.

Cinco linhas plausíveis, uma defeituosa. A errada é a do tradutor, que foi criado a partir de um nome diferente do nome dos pesos — exatamente a regra quebrada que o Bloco 2 mostrou em desenho, agora escrita em código. Nenhuma das cinco dá mensagem de erro, e é por isso que este tipo de defeito custa caro no dia do desafio.

04

BLOCO 4

Avaliar

- Treino curto muda a leitura das **duas curvas**
- Com duas classes, a matriz tem **nome próprio**
- Com quem o modelo erra mais
- F1-macro decide, recall aparece



ABERTURA DO BLOCO 4

O último bloco, e o que mais pesa no relatório. Quatro leituras: o que as curvas dizem, o nome próprio de cada casa da matriz, com que tipo de texto o modelo erra mais, e por que a métrica da fase é a que é.

O que as curvas dizem numa trilha de texto

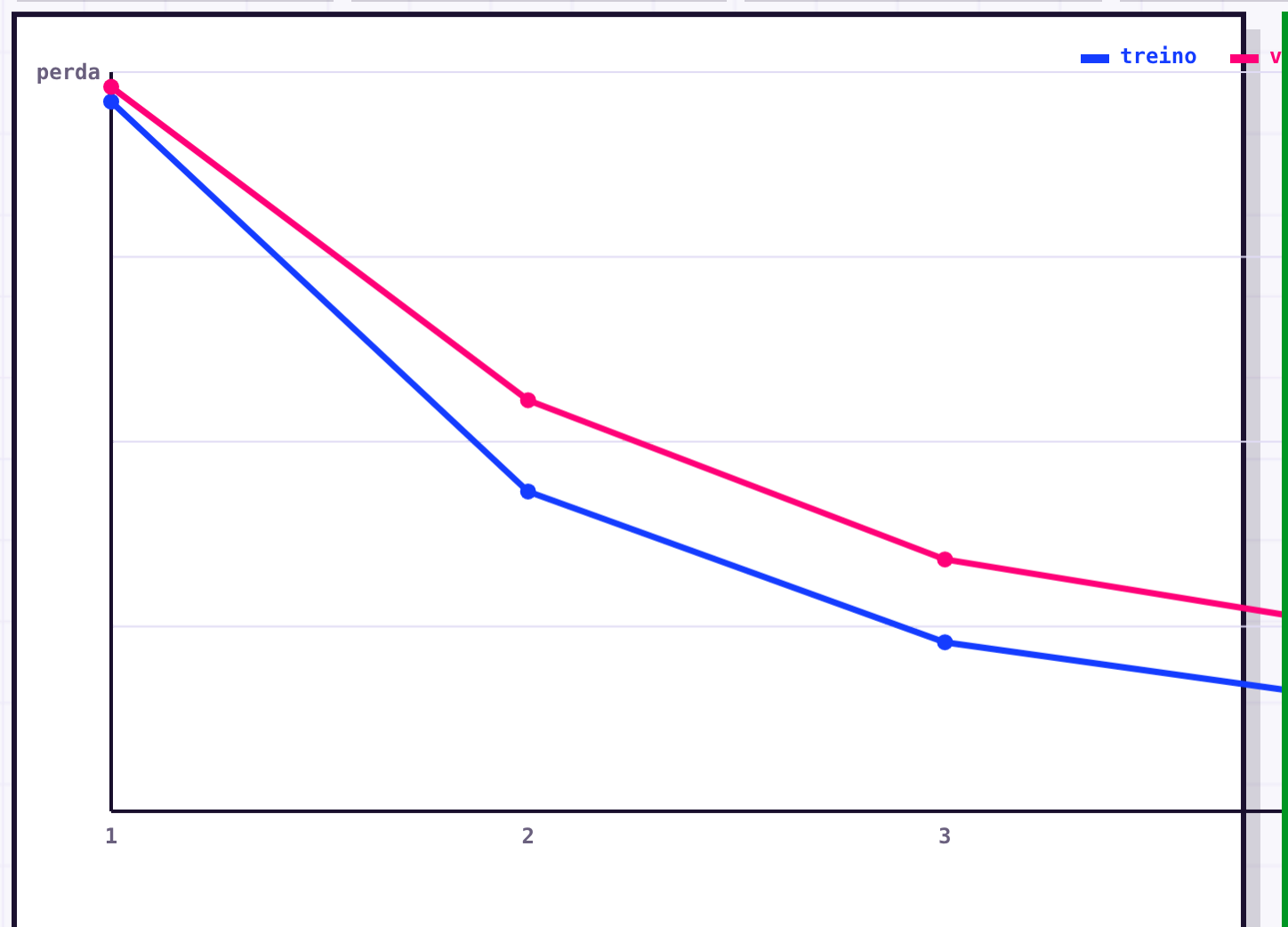
Aqui o treino é curto: três épocas, às vezes quatro. Isso muda a leitura.

Como deve ser

Uma época a mais

Sem aquecimento

Passo grande demais



O MELHOR PONTO VEM CEDO

Num modelo que **já sabia português**, a validação atinge o melhor ponto quase sempre na segunda ou terceira época. O ajuste fino não ensina a língua: ensina a tarefa. É pouca coisa a aprender.

As duas descem e a distância entre elas fica pequena. A validação assenta num patamar — e é a melhor época dela que vai para a entrega, não a última. Guardem esta página como referência: as três seguintes são o que dá errado.

O que as curvas dizem numa trilha de texto

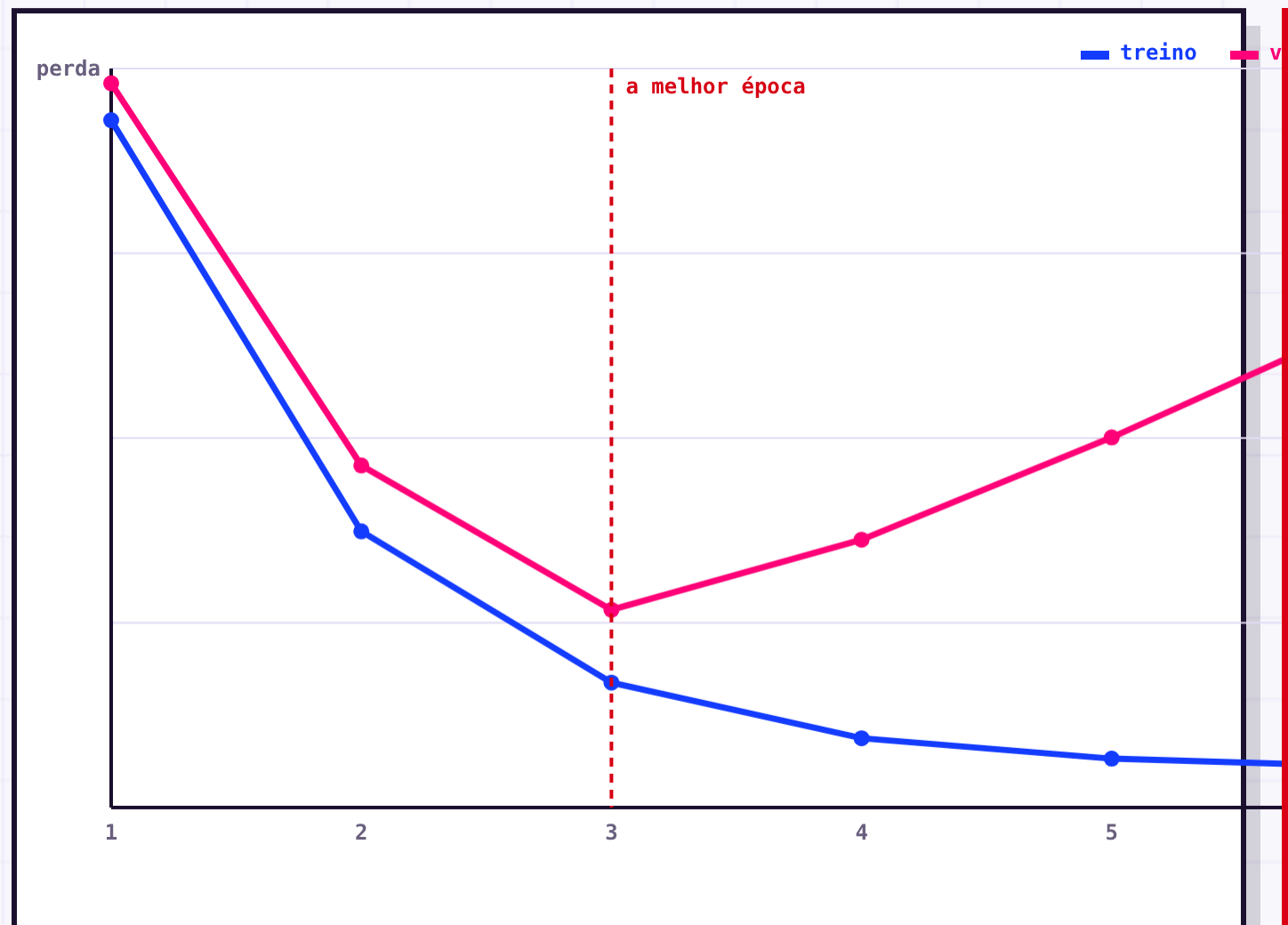
Aqui o treino é curto: três épocas, às vezes quatro. Isso muda a leitura.

Como deve ser

Uma época a mais

Sem aquecimento

Passo grande demais



PASSOU DO PONTO

Da quarta época em diante a perda de treino continua caindo e a de validação sobe. Com treino curto, **uma época a mais já estraga** — e é por isso que a entrega leva a melhor época, não a última.

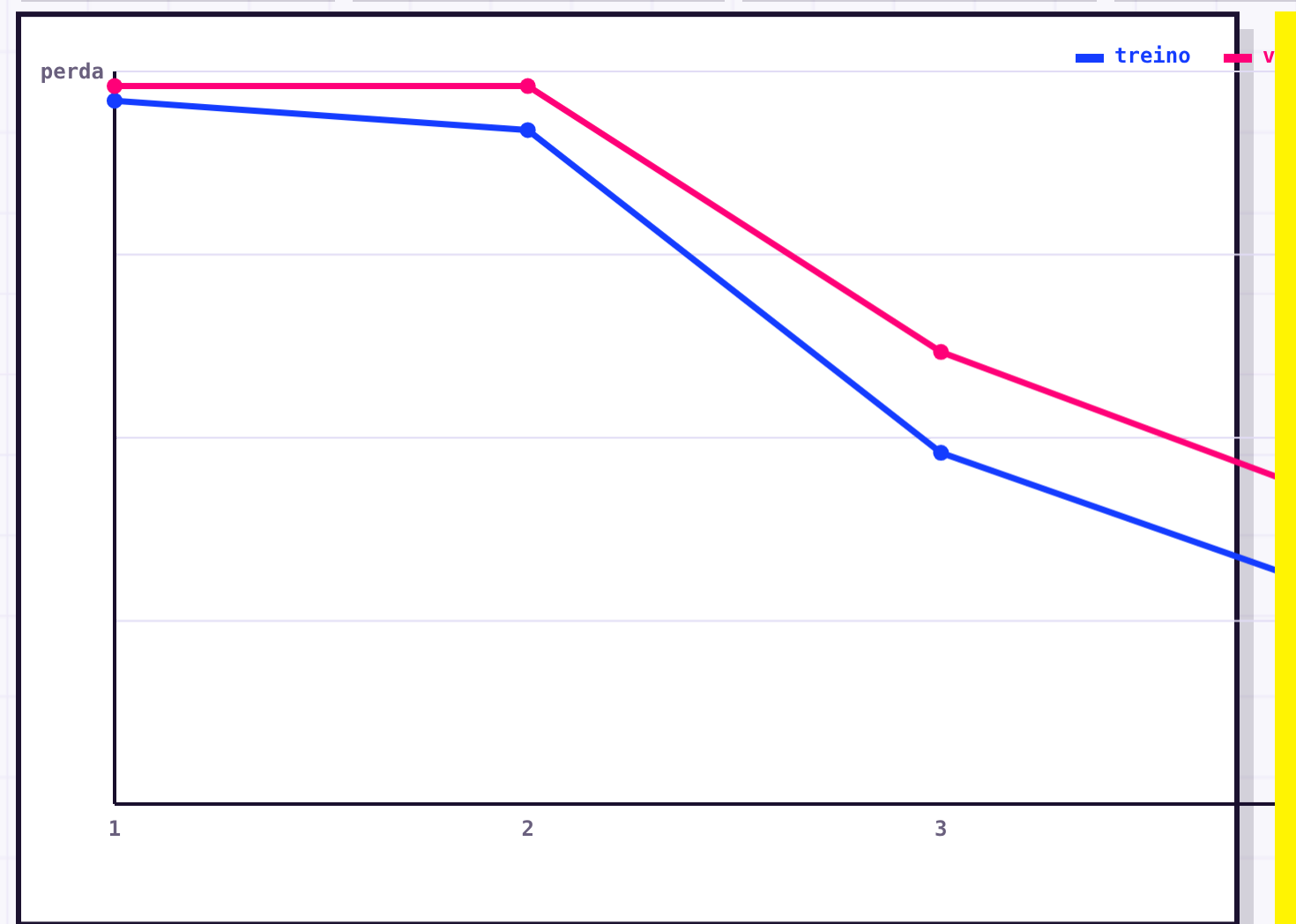
AS CURVAS · UMA ÉPOCA A MAIS

O treino continua descendo e a validação virou para cima. É o retrato da página dos hiperparâmetros com oito épocas, agora visto pelo gráfico. Em ajuste fino de modelo de linguagem a virada chega cedo, e quem não estiver olhando a curva vai entregar o modelo da última época achando que é o melhor.

O que as curvas dizem numa trilha de texto

Aqui o treino é curto: três épocas, às vezes quatro. Isso muda a leitura.

Como deve ser Uma época a mais **Sem aquecimento** Passo grande demais



A PRIMEIRA ÉPOCA SOFRE

A curva começa num degrau alto e demora a descer. O aquecimento começa com passos minúsculos e vai crescendo — e o XLM-R é o que mais se beneficia disso.

AS CURVAS · SEM AQUECIMENTO

O tranco da primeira época. A curva começa num lugar pior do que precisava e leva um tempo para se recuperar — às vezes não recupera dentro das três épocas. É o defeito mais barato de corrigir da trilha inteira: basta ligar o aquecimento.

O que as curvas dizem numa trilha de texto

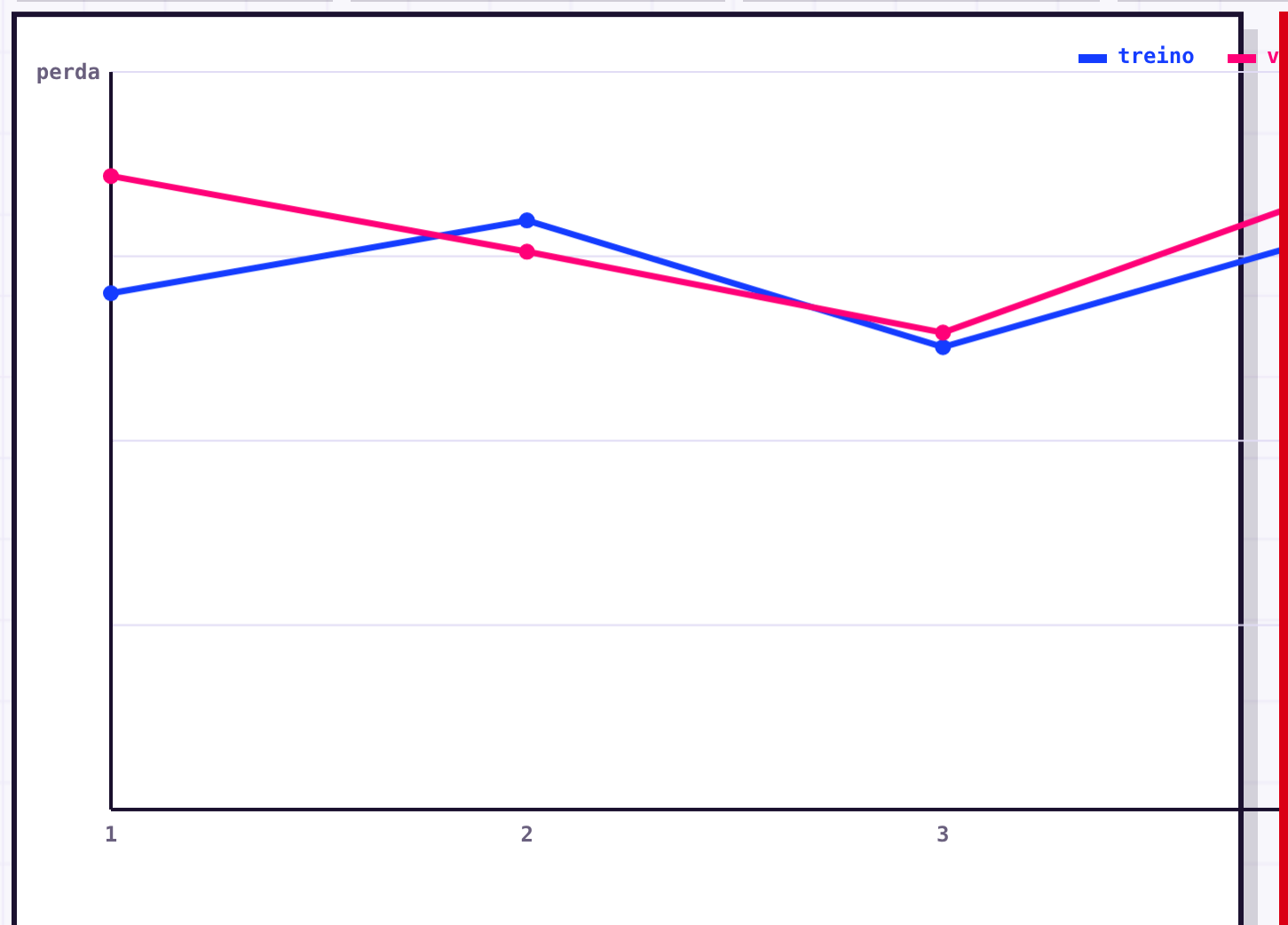
Aqui o treino é curto: três épocas, às vezes quatro. Isso muda a leitura.

Como deve ser

Uma época a mais

Sem aquecimento

Passo grande demais



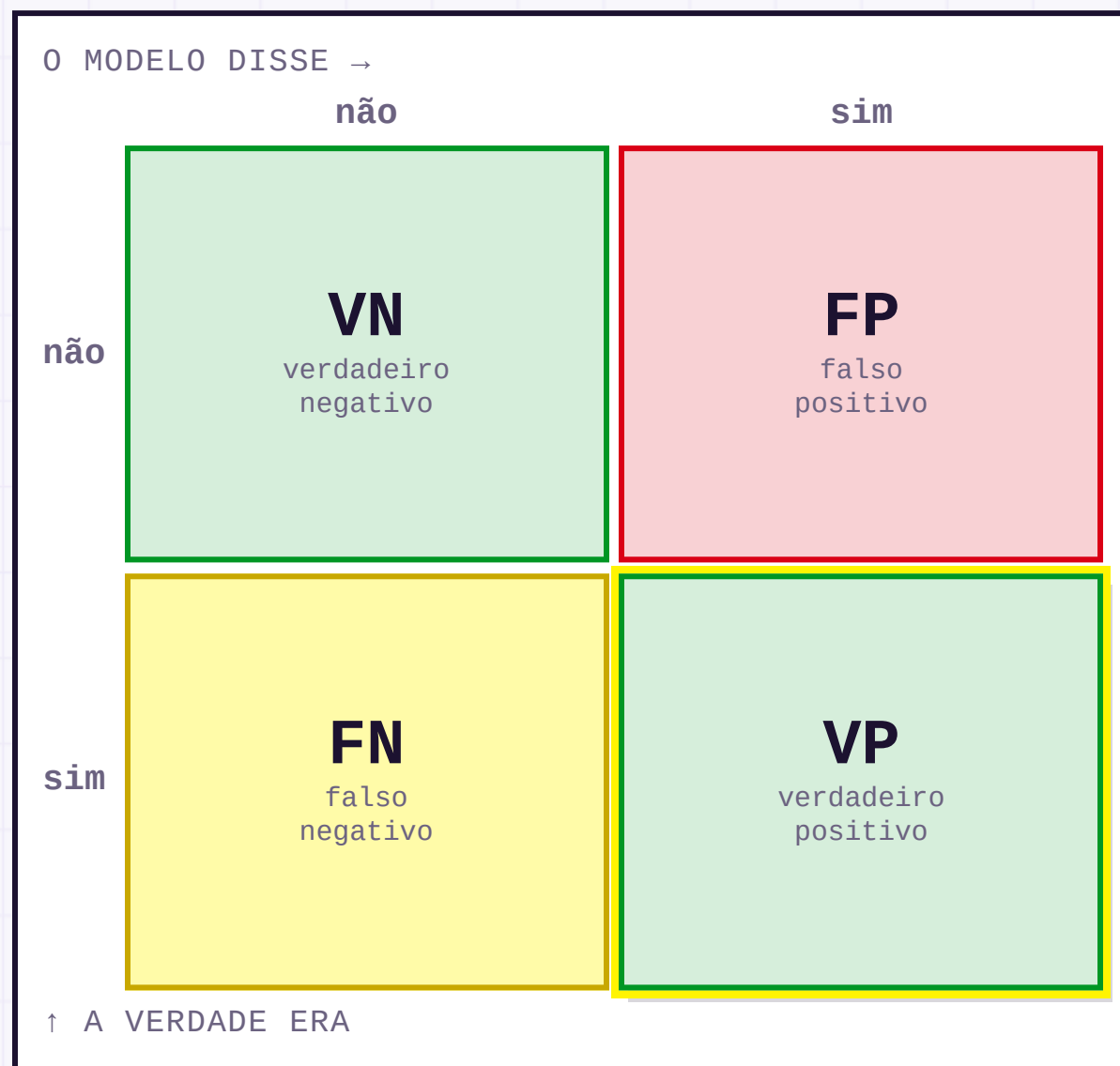
A PERDA NÃO DESCE

Com pesos pré-treinados, um passo mil vezes acima da faixa **apaga o que o modelo já sabia** nas primeiras centenas de passos. Diminuem a taxa antes de mexer em qualquer outra coisa.

As duas serrilham e não descem. Num modelo pré-treinado isso costuma significar que o passo apagou o que ele já sabia. Diminuem a taxa antes de mexer em qualquer outra coisa — e comparem com a página do passo grande nos hiperparâmetros: é a mesma história, contada pelo outro lado.

Com duas classes, a matriz tem nome próprio

Quatro casas, quatro nomes que vocês vão usar o tempo todo.



VERDADEIRO POSITIVO

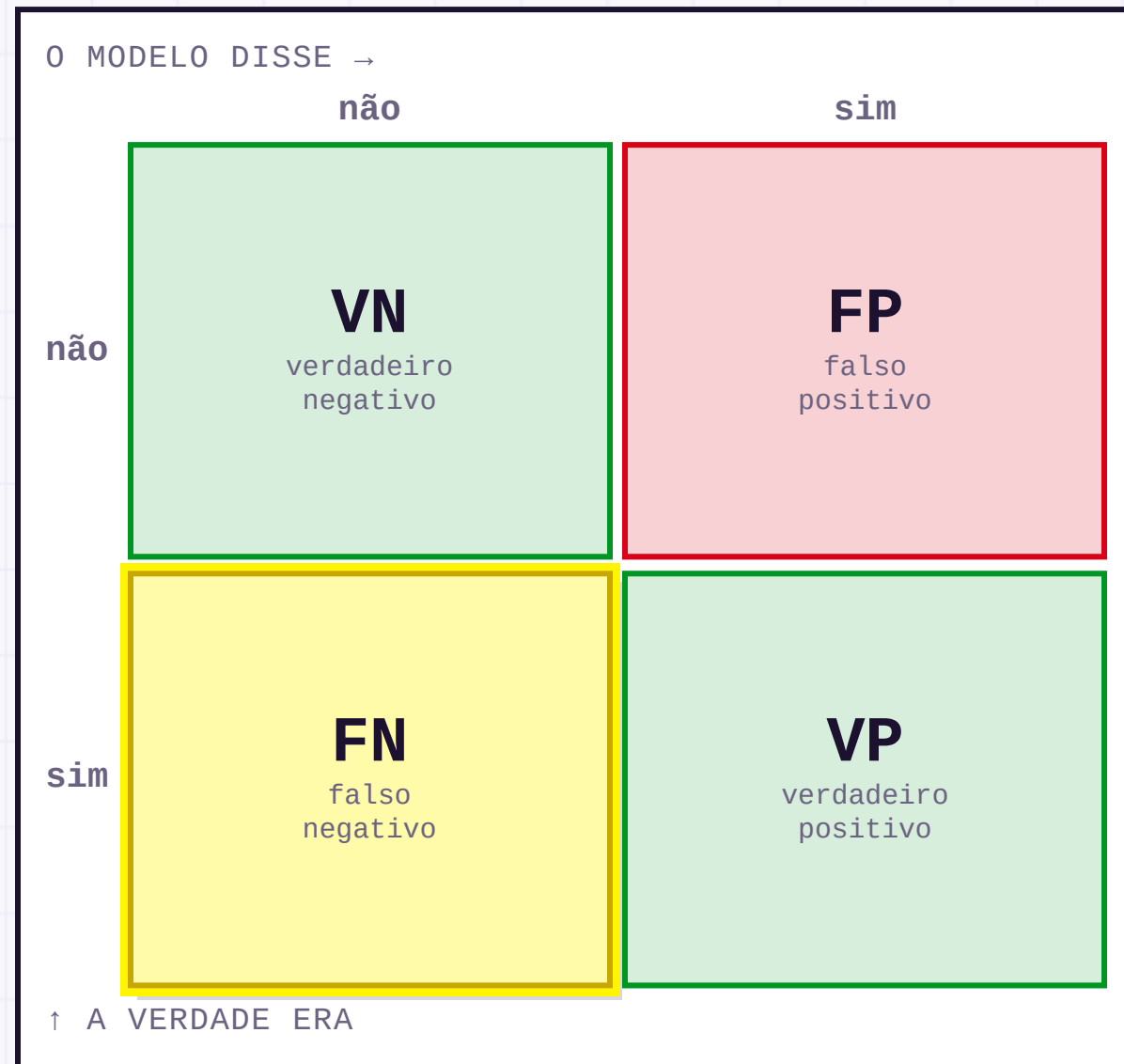
A resposta era **sim** e o modelo disse **sim**. Acertou.

A MATRIZ · O ACERTO

Com duas classes, as quatro casas ganham nome próprio. As duas da diagonal são acerto e não ensinam nada de novo. Esta página existe para fixar o vocabulário: é com ele que a equipe vai conversar com a cliente no relatório.

Com duas classes, a matriz tem nome próprio

Quatro casas, quatro nomes que vocês vão usar o tempo todo.



FALSO NEGATIVO

A resposta era **sim** e o modelo disse **não**. É o alarme falso: alguém vai conferir à toa. Custa tempo de gente, não um erro que vaza.

Qual dos dois custa mais caro? Não existe resposta técnica.

Depende do que o sistema faz com a previsão — e quem decide isso é gente, não métrica.

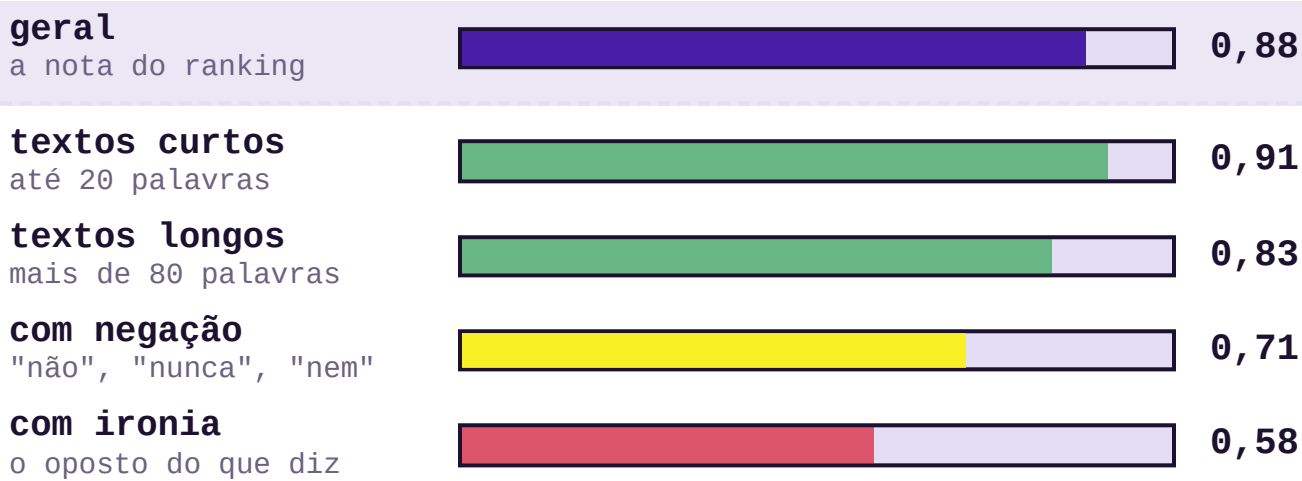
A MATRIZ · O ERRO QUE NINGUÉM VÊ

O falso negativo: existia e o modelo não encontrou. Na narrativa desta trilha é o erro caro — o que passa adiante sem ninguém perceber, porque não volta para a fila de revisão de ninguém. As outras duas casas não custam a mesma coisa, e dizer qual delas vocês escolheram tolerar é parte da entrega.

Com quem o modelo erra mais

Um número só esconde muita coisa. Exemplo ilustrativo — a tabela de vocês sai do próprio treino.

F1-MACRO POR GRUPO DE TEXTO



GERAL · 0,88

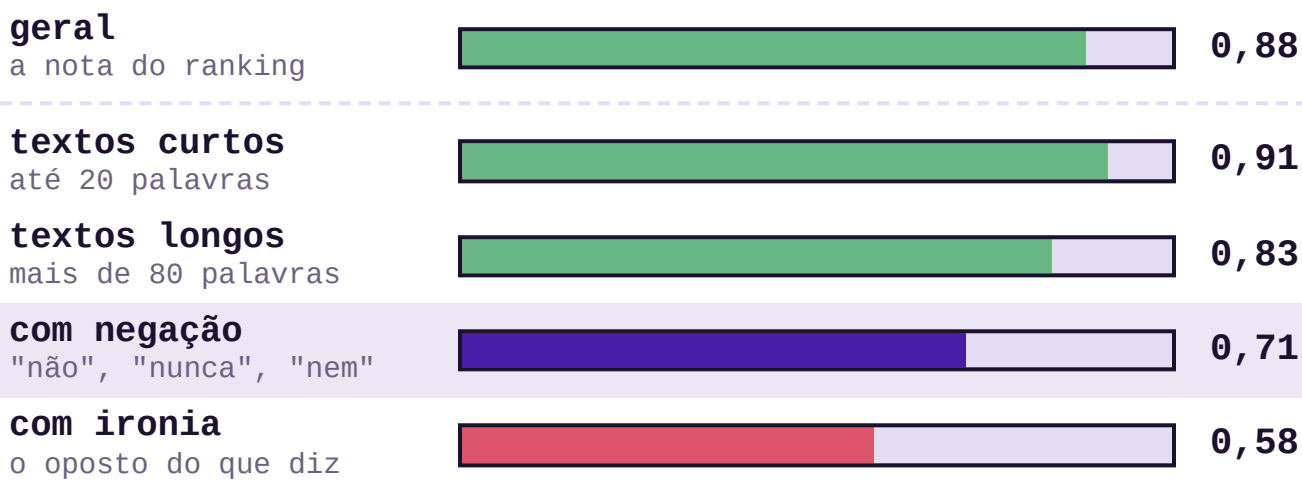
A nota que aparece no placar. Parece ótima — e esconde todo o resto desta lista.

A nota única do ranking. É com ela que as equipes são ordenadas — e é justamente por isso que esta página não pode ser a última coisa que vocês olham. Uma média esconde o que vem nas próximas duas páginas.

Com quem o modelo erra mais

Um número só esconde muita coisa. Exemplo ilustrativo — a tabela de vocês sai do próprio treino.

F1-MACRO POR GRUPO DE TEXTO



COM NEGAÇÃO · 0,71

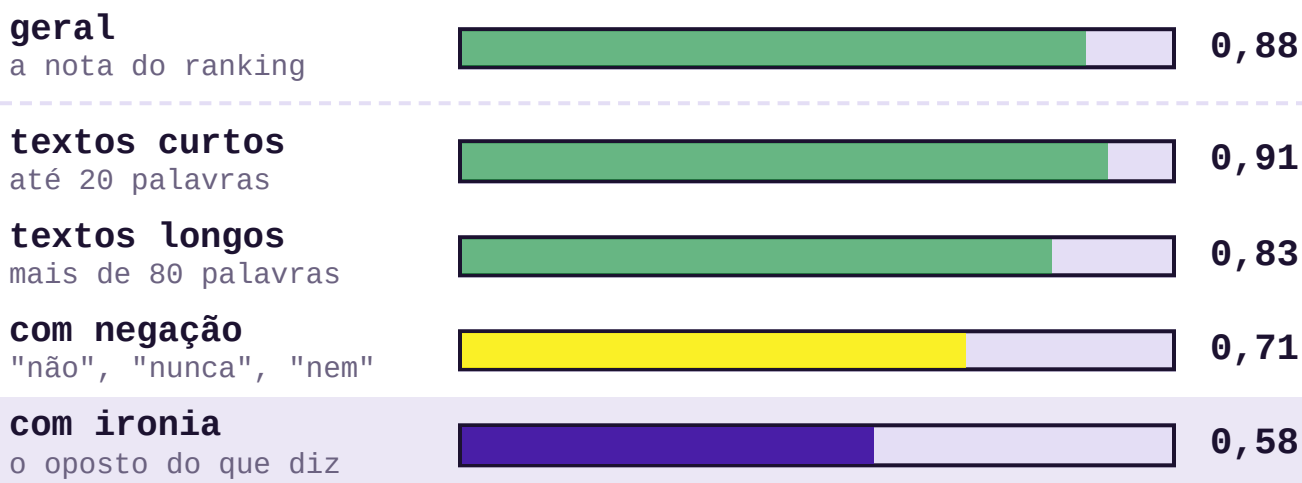
Aqui aparece o buraco. A negação inverte o sentido da frase inteira, e é exatamente onde contar palavras soltas falha.

O mesmo modelo, agora só nos textos que contêm "não", "nunca", "nem". A nota cai bastante. Faz sentido: negação muda o sentido inteiro mexendo em uma palavra só, e quase nada no vocabulário denuncia isso. É aqui que o relatório de vocês começa a ficar interessante.

Com quem o modelo erra mais

Um número só esconde muita coisa. Exemplo ilustrativo — a tabela de vocês sai do próprio treino.

F1-MACRO POR GRUPO DE TEXTO



COM IRONIA · 0,58

O pior grupo. As palavras apontam para um lado e a intenção para o outro. Nenhum dos modelos desta fase resolve bem.

O pior grupo de todos, e por larga margem. Ironia diz o oposto do que está escrito, então o sinal que o modelo aprendeu funciona ao contrário. Uma equipe que entrega só a média geral nunca descobre isto — e é exatamente o tipo de achado que distingue um bom relatório de um placar bonito.

F1-macro decide, recall aparece

Um modelo que responde **sempre a classe maior** e nunca acerta a menor. Que nota cada média dá a ele?

classe maior

70%

MÉTRICA PRINCIPAL

F1-macro

Ordena o ranking, nas três trilhas.

MÉTRICA DE APOIO

Recall

Dos casos que existiam, quantos o modelo encontrou. Só desempata — senão bastaria responder a classe positiva em tudo.

F1-macro



F1-micro



F1 ponderado



- **F1-macro** — média simples: a classe pequena pesa igual à grande
- **F1-micro** — junta tudo num bolo só
- **F1 ponderado** — pondera pelo tamanho de cada classe

POR QUE MACRO, NAS TRÊS TRILHAS

Quanto mais desbalanceado o problema, mais o micro e o ponderado premiam quem só acerta a classe grande. **Arrastem até 90%** e vejam a distância abrir.

Um modelo que responde sempre a classe maior e nunca acerta a menor. Com 70% de desequilíbrio as três médias já discordam: 0,41 de macro contra 0,70 de micro e 0,58 de ponderado. Guardem estes três números — a página seguinte mostra o que acontece com eles quando o desequilíbrio aperta.

F1-macro decide, recall aparece

Um modelo que responde **sempre a classe maior** e nunca acerta a menor. Que nota cada média dá a ele?

classe maior

95%

MÉTRICA PRINCIPAL

F1-macro

Ordena o ranking, nas três trilhas.

MÉTRICA DE APOIO

Recall

Dos casos que existiam, quantos o modelo encontrou. Só desempata — senão bastaria responder a classe positiva em tudo.

F1-macro



F1-micro



F1 ponderado



- **F1-macro** — média simples: a classe pequena pesa igual à grande
- **F1-micro** — junta tudo num bolo só
- **F1 ponderado** — pondera pelo tamanho de cada classe

O MODELO PREGUIÇOSO FICA BONITO

Com 95% numa classe só, um modelo que **responde sempre a classe maior** e nunca acerta a menor tira 0,95 de micro e de acurácia. O macro desmascara: 0,49.

O mesmo modelo inútil, agora num problema bem mais desbalanceado. O micro sai de 0,70 para 0,95 e o ponderado de 0,58 para 0,93 — os dois passam a premiar quem só acerta a classe grande. O macro sobe pouco, de 0,41 para 0,49, porque continua dando o mesmo peso à classe pequena. Comparem as duas páginas: o que interessa não é cada número, é a distância entre eles abrindo.

O QUE LEVAR DAQUI

Cinco frases e um endereço

Palavra nunca vista não some mais — vira pedaços.

O tradutor e o modelo saem do mesmo nome. Sempre.

Não limpem o texto como na Fase 1. Isso destrói pista.

Três épocas bastam. A quarta costuma estragar.

Os dois tipos de erro não custam a mesma coisa.

Agora abram o Capítulo 4 do material de
imersão e refaçam os laboratórios com calma.

olimpiadadeia.ceia.digital

olimpiada@ceia.ufg.br



FECHAMENTO

As cinco frases que resumem o aulão e o endereço do material vivo. Tudo o que está neste PDF pode ser mexido lá: as frases dos tradutores, o limite de corte, os quatro botões, as lacunas, as curvas e os grupos de texto. Quem quiser realmente entender, abre o material e refaz com calma — este documento serve para lembrar o que foi visto, não para substituir a mexida.