

The Hitchhiker's Guide to European Portuguese LLMs

(and how to ensure your models and applications run well)

Duarte O.Carmo
AI/ML/Data Consultant & Researcher @ DOC Tech. | duarteocarmo.com

PyData Copenhagen

Hej, jeg hedder Duarte.



/du-art/, from Lisbon, based in København

AI/ML/Data Consultant & Researcher

I work with large companies and startups

I like to try to solve the hardest problem in the room

Now: LLMs, VLMs, Speech Models, and EVALUATIONS

I have a blog where I write (stupid) things

duarteocarmo.com


What is a good European Portuguese LLM?
We need data
Training our own models
How far can a small model go?
How I evaluate
Lessons from running evaluations
So what?

What is a good European Portuguese LLM?

We need data

Training our own models

How far can a small model go?

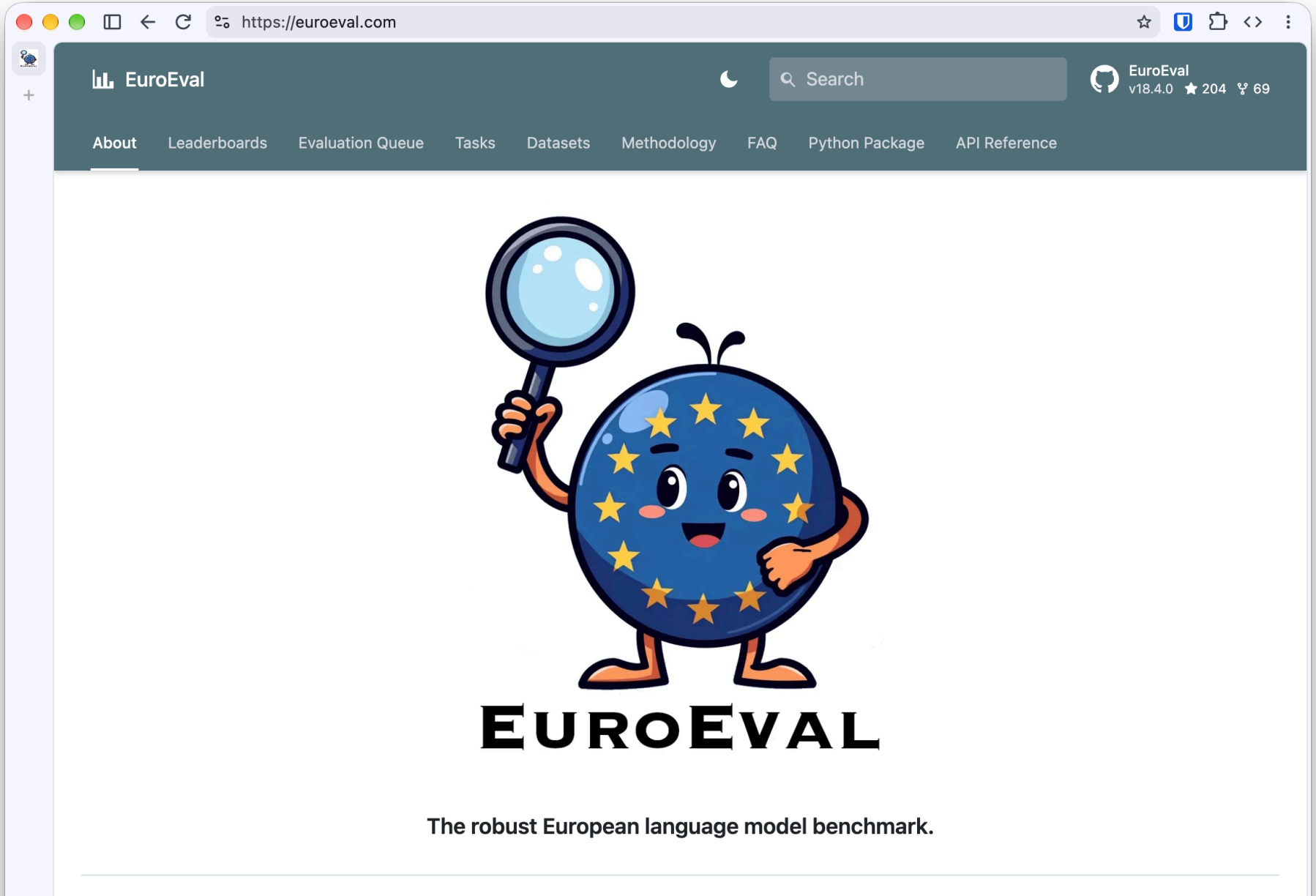
How I evaluate

Lessons from running evaluations

So what?



What is a good European Portuguese LLM?


https://euroeval.com
EuroEval
Search
EuroEval
v18.4.0 204 69
About
Leaderboards
Evaluation Queue
Tasks
Datasets
Methodology
FAQ
Python Package
API Reference
EUROEVAL
The robust European language model benchmark.

Some months passed, but all I heard was silence

What is a good European Portuguese LLM?

We need data

Training our own models

How far can a small model go?

How I evaluate

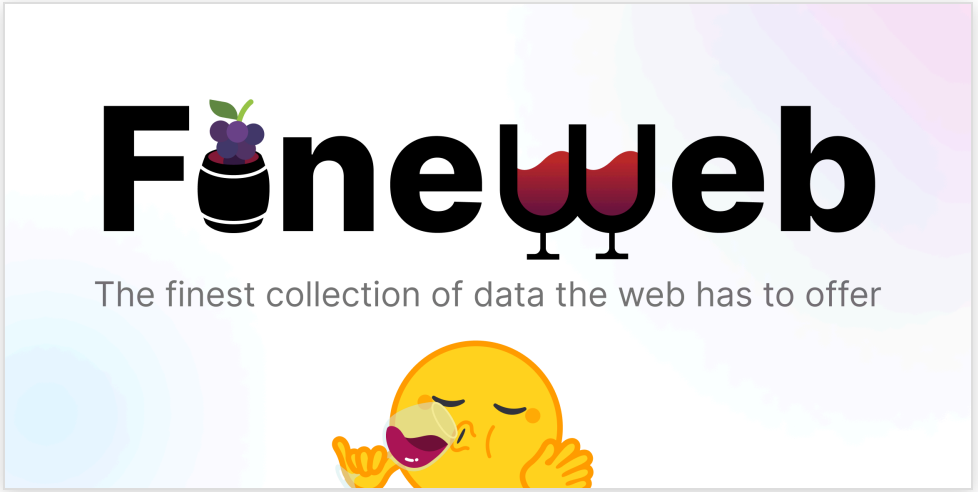
Lessons from running evaluations

So what?

If we're gonna do this, we need **DATA**

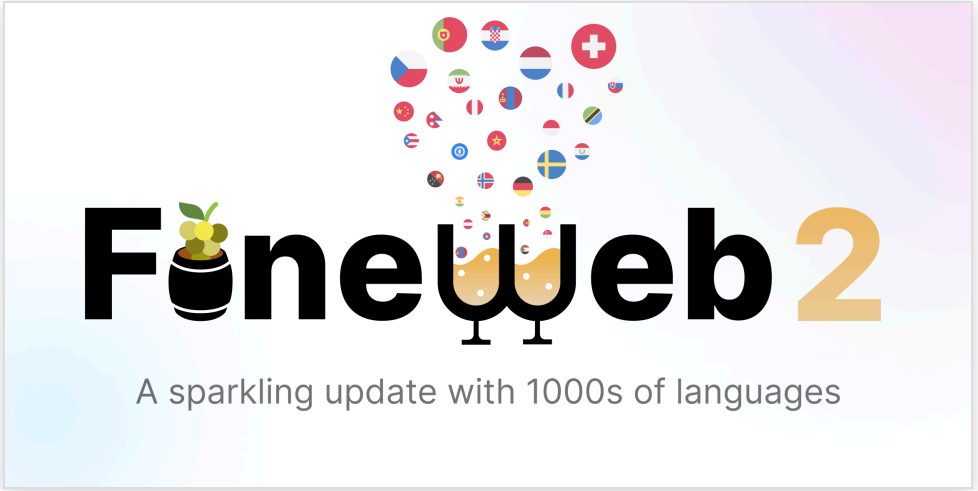
FineWeb

55TB of text, English only



FineWeb2

20TB of text, but now multilingual ✅



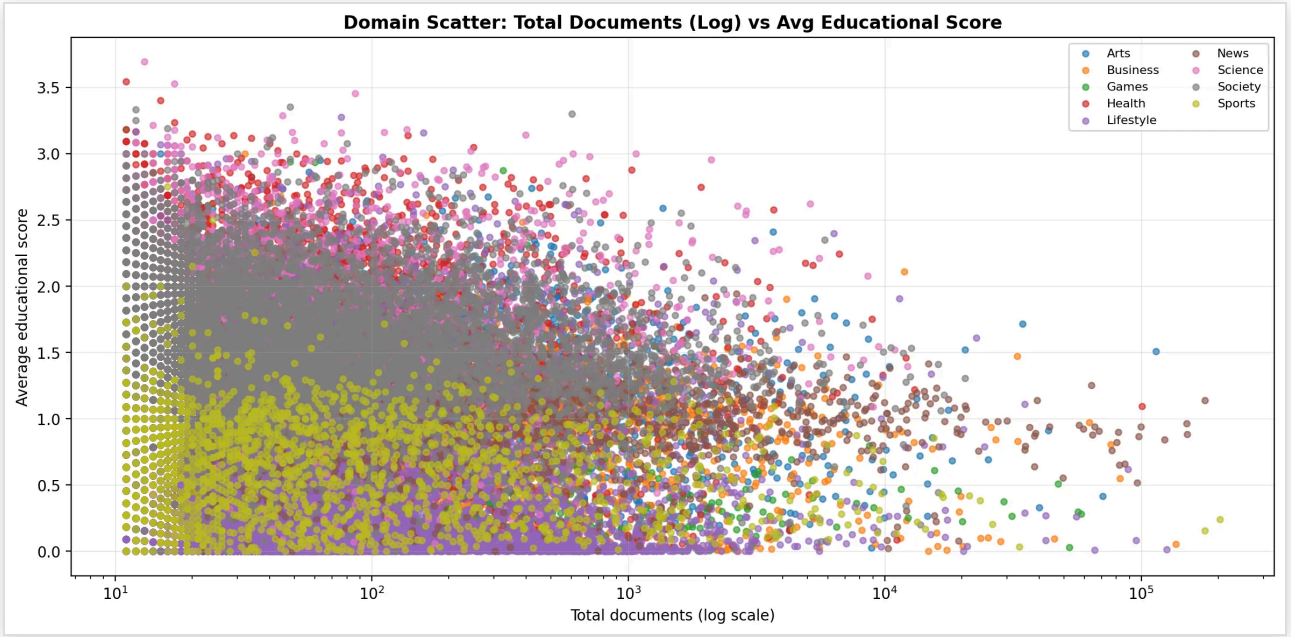
From FineWeb2, I built Bagaço

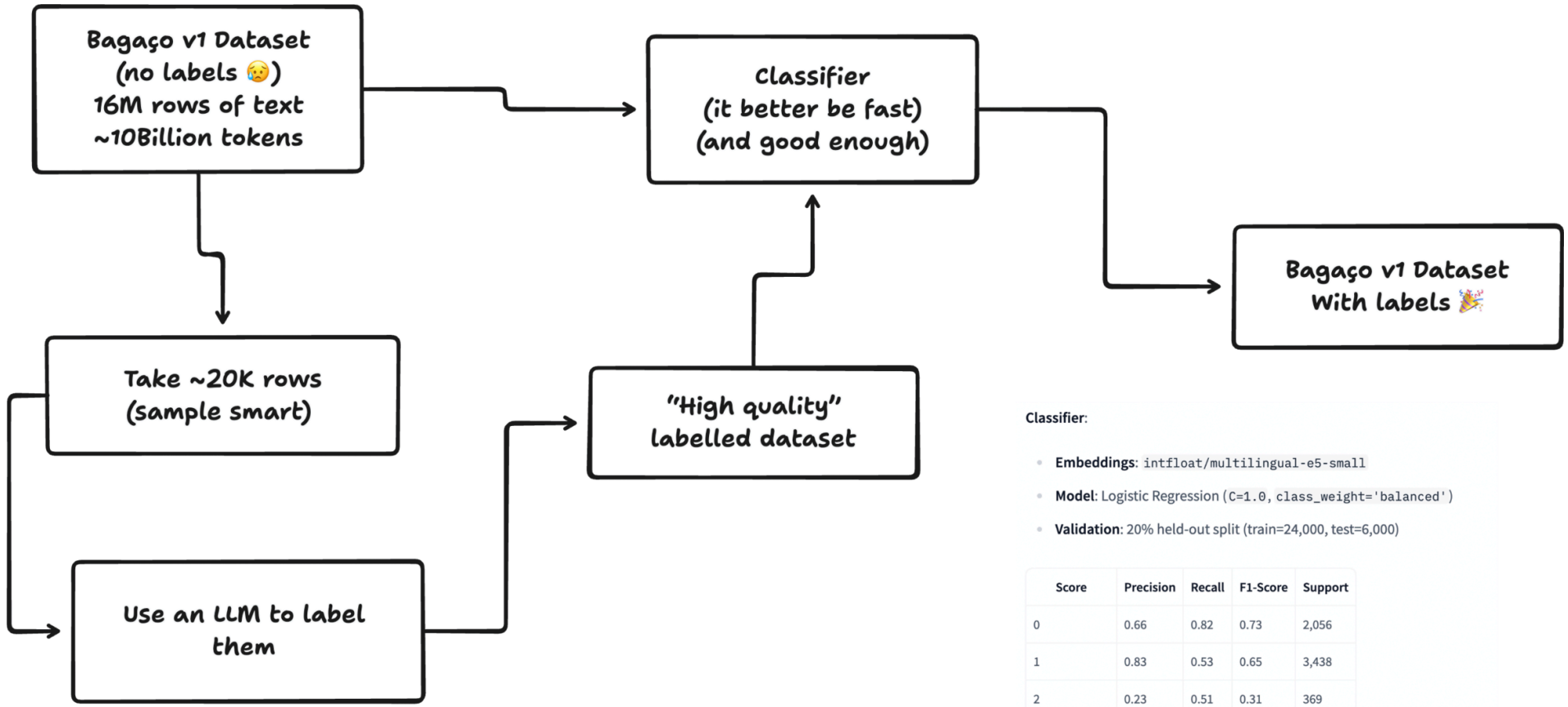
Extract Portugal from FineWeb2

Take all the .pt domains from FineWeb2

Educational score + Content Category

Result: 9.7B tokens of Portuguese data





Classifier:

- **Embeddings**: `intfloat/multilingual-e5-small`
- **Model**: Logistic Regression (`C=1.0`, `class_weight='balanced'`)
- **Validation**: 20% held-out split (train=24,000, test=6,000)

Score	Precision	Recall	F1-Score	Support
0	0.66	0.82	0.73	2,056
1	0.83	0.53	0.65	3,438
2	0.23	0.51	0.31	369
3	0.21	0.60	0.31	131
4	0.00	0.00	0.00	6
Accuracy			**0.63**	**6,000**
Macro Avg	**0.39**	**0.49**	**0.40**	**6,000**
Weighted Avg	**0.72**	**0.63**	**0.65**	**6,000**

But that approach is stupid,
we can do better, right?

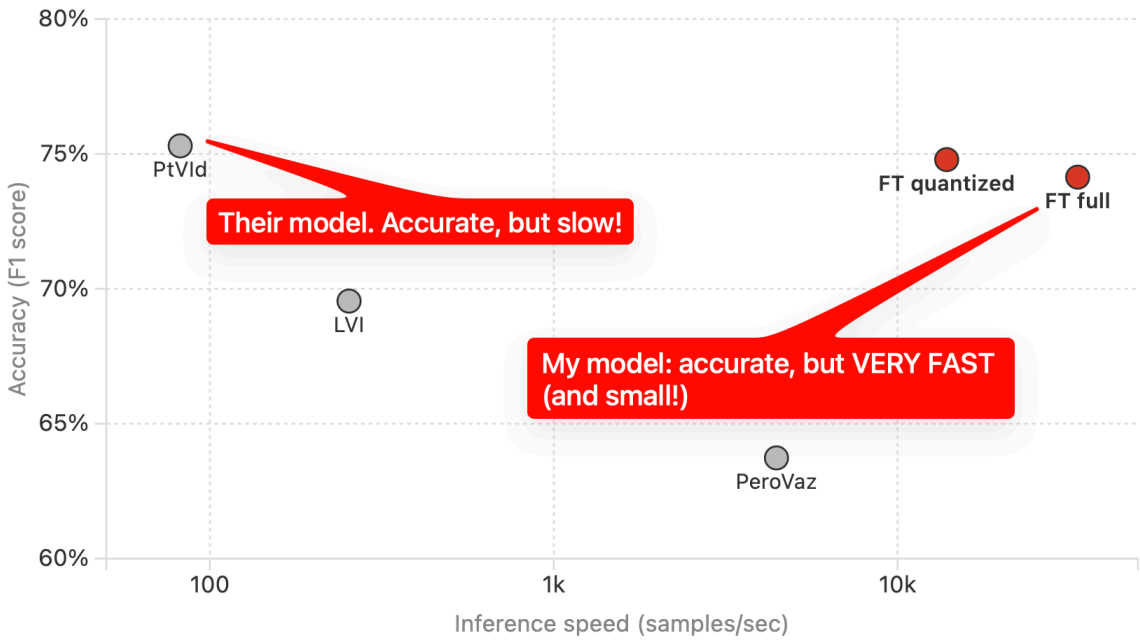
- PtVId paper ↗
- My blog post ↗

Enhancing Portuguese Variety Identification with Cross-Domain Approaches

Hugo Sousa[1,3*], Rúben Almeida[2,3,4*], Purificação Silvano[5,6], Inês Cantante[5,6], Ricardo Campos[3,7,8], Alípio Jorge[1,3]

[1]Faculty of Sciences, University of Porto, Porto, Portugal
[2]Faculty of Engineering, University of Porto, Porto, Portugal
[3]INESC TEC, Porto, Portugal
[4]Innovation Point - dst group, Braga, Portugal
[5]Faculty of Arts and Humanities, University of Porto, Porto, Portugal
[6]Centre of Linguistics, University of Porto, Porto, Portugal
[7]Department of Informatics, University of Beira Interior, Covilhã, Portugal
[8]Ci2 - Smart Cities Research Center, Tomar, Portugal
hugo.o.sousa@inesctec.pt, ruben.f.almeida@inesctec.pt

A bitter lesson



Average PT-PT F1 (DSL-TL + FRMT) vs inference speed
measured on Apple M3 Max · reproduce

With that, we can build Bagaço v2

Run a pt classifier on top of FineWeb2

Educational score + Content Category

Result: 21.3B tokens of Portuguese data

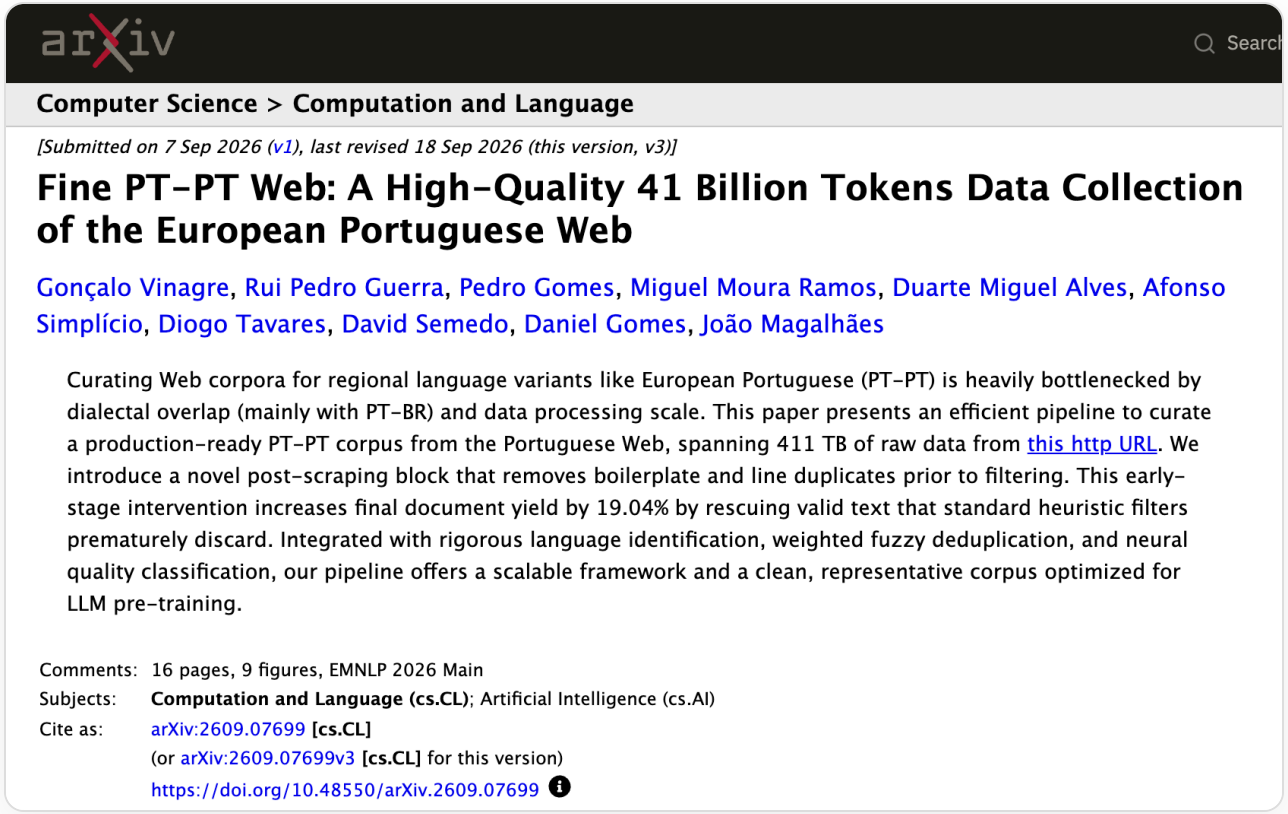
+119% of data!

Top 10 domains in Bagaço v2 by average educational score (min 500 docs)

domain	doc_count	avg_educational_score
pt-pt.khanacademy.org	1.3K	3.023
museubiodiversidade.uevora.pt	568	3.002
academia.cienciaviva.pt	515	2.926
casadasciencias.org	2.2K	2.918
avesdeportugal.info	841	2.905
mozaweb.com	924	2.895
educamais.com	850	2.875
alimentacaomedicinal.com	991	2.845
nagwa.com	673	2.841
dicionariosaude.com	1.1K	2.834

Fine PT-PT Web paper (p. 9) ↗

My classifier ended up in a paper



arXiv Search

Computer Science > Computation and Language

[Submitted on 7 Sep 2026 (v1), last revised 18 Sep 2026 (this version, v3)]

Fine PT-PT Web: A High-Quality 41 Billion Tokens Data Collection of the European Portuguese Web

Gonçalo Vinagre, Rui Pedro Guerra, Pedro Gomes, Miguel Moura Ramos, Duarte Miguel Alves, Afonso Simplício, Diogo Tavares, David Semedo, Daniel Gomes, João Magalhães

Curating Web corpora for regional language variants like European Portuguese (PT-PT) is heavily bottlenecked by dialectal overlap (mainly with PT-BR) and data processing scale. This paper presents an efficient pipeline to curate a production-ready PT-PT corpus from the Portuguese Web, spanning 411 TB of raw data from this http URL. We introduce a novel post-scraping block that removes boilerplate and line duplicates prior to filtering. This early-stage intervention increases final document yield by 19.04% by rescuing valid text that standard heuristic filters prematurely discard. Integrated with rigorous language identification, weighted fuzzy deduplication, and neural quality classification, our pipeline offers a scalable framework and a clean, representative corpus optimized for LLM pre-training.

Comments: 16 pages, 9 figures, EMNLP 2026 Main
Subjects: **Computation and Language (cs.CL)**; Artificial Intelligence (cs.AI)
Cite as: arXiv:2609.07699 **[cs.CL]**
(or arXiv:2609.07699v3 **[cs.CL]** for this version)
https://doi.org/10.48550/arXiv.2609.07699

5.3 Brazilian Portuguese Contamination

Due to the significant overlap between European and Brazilian Portuguese variants, it is difficult to filter out all PT-BR content without losing high-quality PT-PT text. Our URL-based filtering approach does not exclude generic top-level domains such as `.com` and `.net`, which may include content in either Portuguese variant. We therefore estimated PT-BR contamination in the final corpus using two independent, well-established language-variety classifiers. The `duarteocarmo/fasttext-euptvid` classifier (Carmo, 2026) labeled 86.4% of the corpus as PT-PT and 13.6% as PT-BR, while the `bastao/PeroVaz_PT-BR_Classifier`[8] assigned 82.9% to PT-PT and 17.1% to PT-BR. Averaging the estimates from the two classifiers results in 84.65% PT-PT and 15.35% PT-BR, indicating that PT-PT remains the predominant variant in the final corpus and providing evidence for the effectiveness of our filtering steps.

What is a good European Portuguese LLM?

We need data

Training our own models

How far can a small model go?

How I evaluate

Lessons from running evaluations

So what?

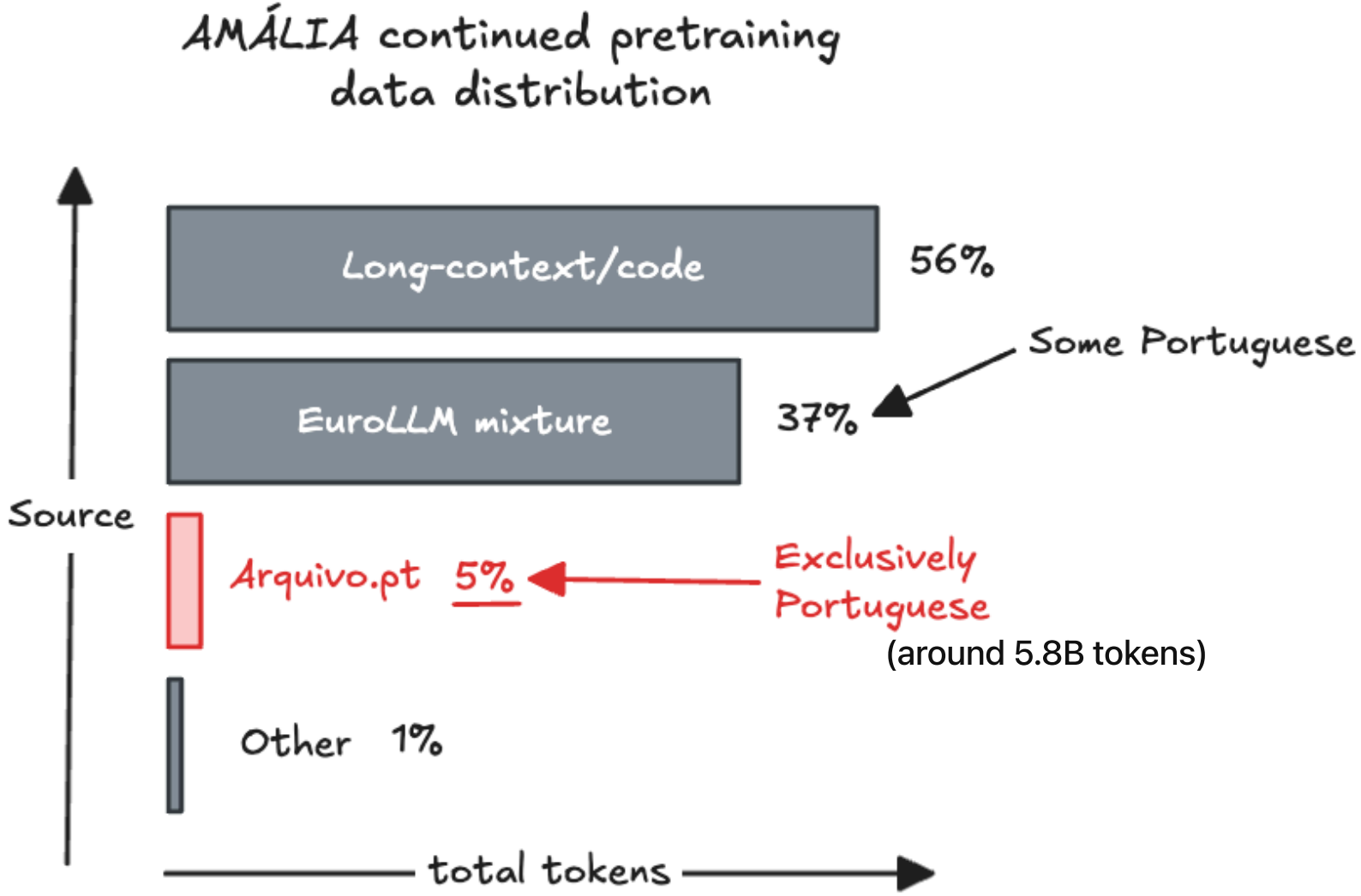
Great, now we can actually build some models?

AMALIA Technical Report: A Fully Open Source Large Language Model for European Portuguese

Afonso Simplício[1,2], Gonçalo Vinagre[1,2], Miguel Moura Ramos[3,4], Diogo Tavares[1,2], Rafael Ferreira[1,2], Giuseppe Attanasio[3], Duarte M. Alves[3,4], Inês Calvo[1], Inês Vieira[1], Rui Guerra[1], James Furtado[1,2], Beatriz Canaverde[3,4], Iago Paulo[1,2], Vasco Ramos[1,2], Diogo Glória-Silva[1,2], Miguel Faria[3], Marcos Treviso[3,4], Daniel Gomes[5], Pedro Gomes[5], David Semedo[1,2], André Martins[3,4], João Magalhães[1,2]

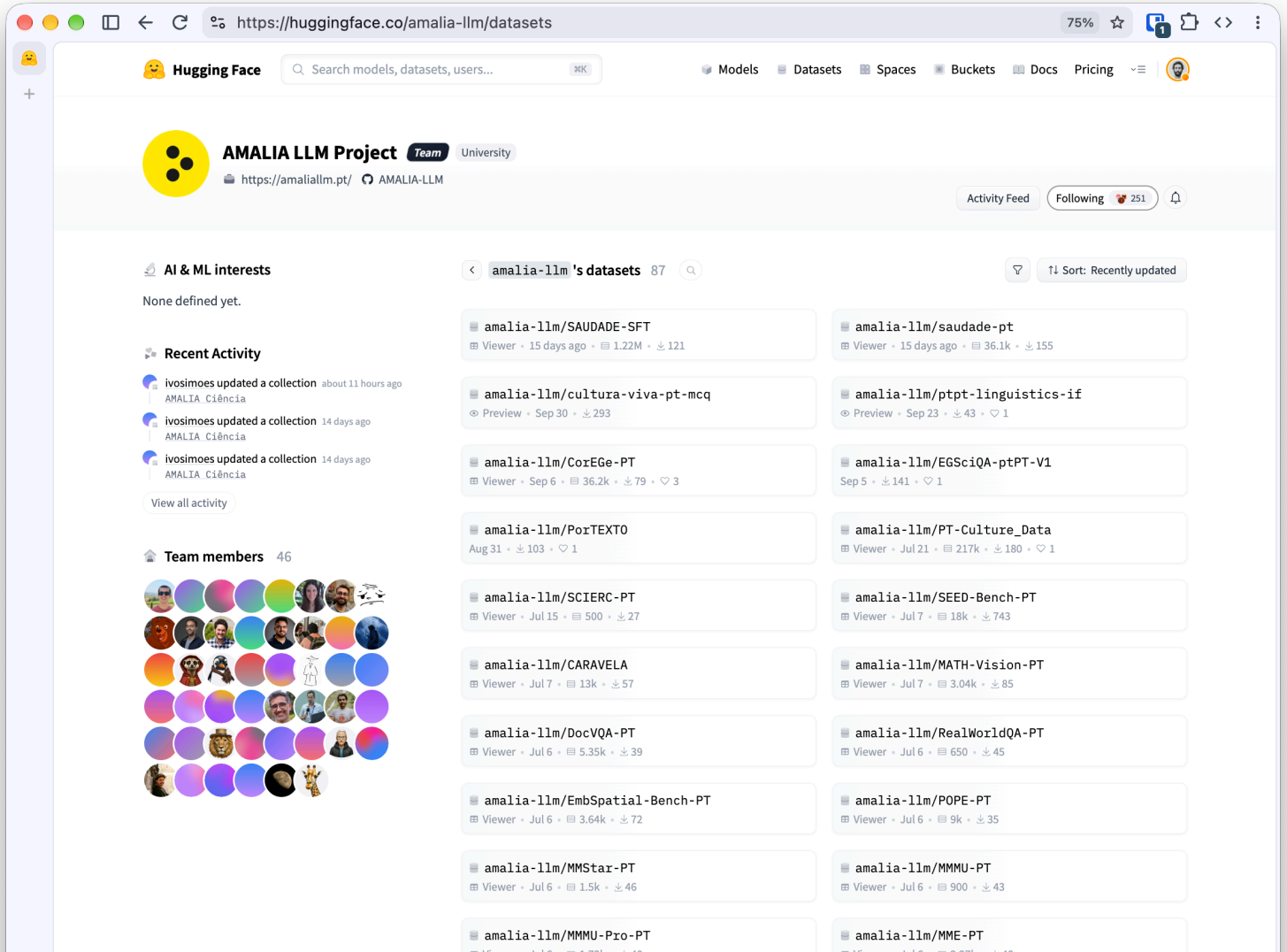
[1]NOVA School of Science and Technology, [2]NOVA LINCS, [3]Instituto de Telecomunicações, [4]Instituto Superior Técnico, Universidade de Lisboa, [5]Fundação para a Ciência e Tecnologia

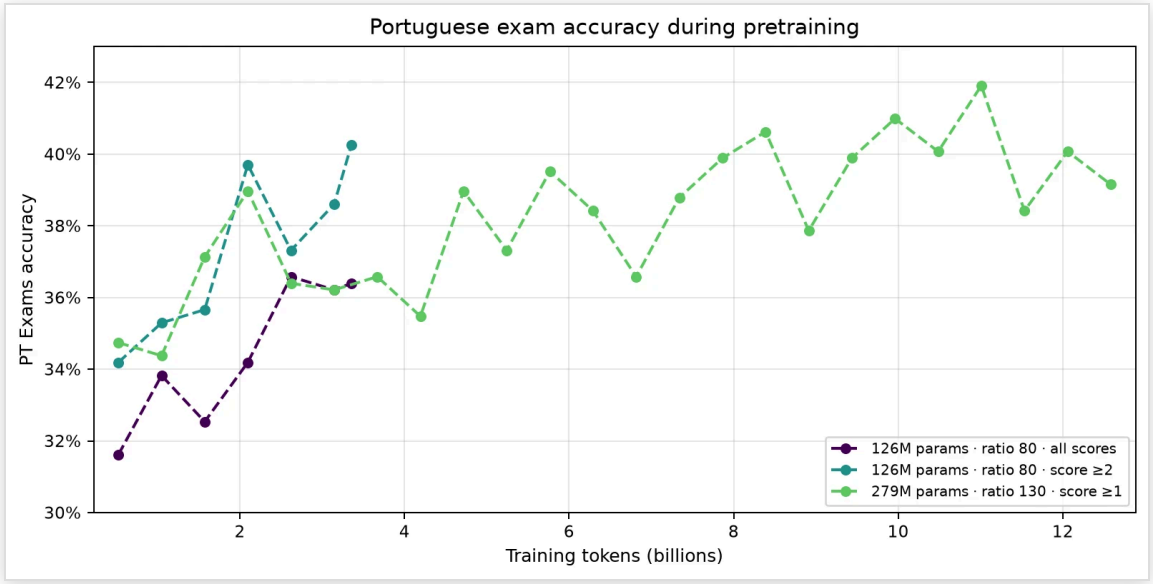
Correspondence: {am.simplicio, gv.martins}@campus.fct.unl.pt

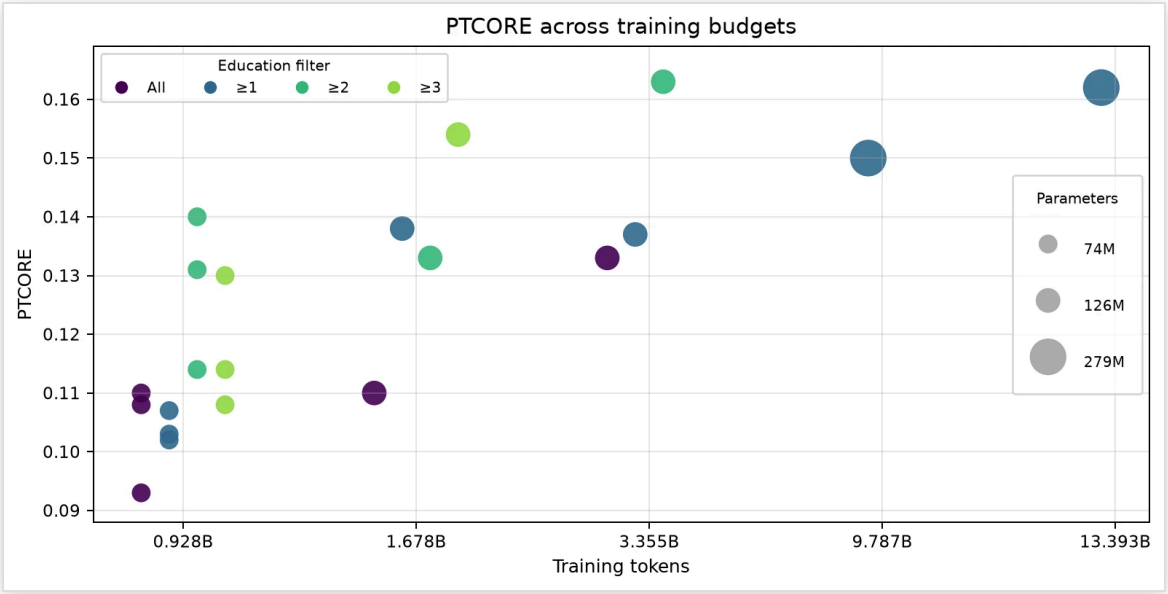

AMÁLIA continued pretraining
data distribution
Source
Long-context/code
56%
EuroLLM mixture
37%
Some Portuguese
Arquivo.pt 5%
Exclusively
Portuguese
(around 5.8B tokens)
Other 1%
total tokens

Everyone is thinking about the models...

That they forgot about the evals







Measuring performance as we pre-train with PTCORE

Task	What it measures	Examples
SST2-PT	Sentiment classification in Portuguese	2,048
ALBA-MCQ	European Portuguese linguistics, language variety, and wordplay	240
CulturaVivaPT	Portuguese culture, places, history, and personalities	1,000
PT Exams	History and Geography questions from Portuguese national exams	544
SAUDADE	Temporal reasoning about Portuguese events	8,573
OpenBookQA-PT	Commonsense and elementary science questions translated into Portuguese	500

A problem of data quality (or educational value)

D6 educational-score ablation

Centered scores × 100 · bold marks the winner within each seed and across means

Filter	Seed	SST2-PT	ALBA	Cultura	PT Exams	SAUDADE	OpenBookQA	PTCORE
All	42	**29.1**	16.9	1.9	13.0	-2.0	7.2	11.0
	1337	24.7	**21.9**	2.0	11.0	-3.4	8.5	10.8
	2026	18.8	16.2	**6.0**	11.0	-2.6	6.1	9.3
	Mean ± SD	**24.2 ± 5.1**	18.3 ± 3.1	3.3 ± 2.3	11.7 ± 1.1	-2.7 ± 0.7	7.3 ± 1.2	10.4 ± 0.9
≥1	42	18.7	18.1	3.9	15.0	-2.1	8.0	10.2
	1337	23.4	18.1	5.2	12.3	-1.0	6.4	10.7
	2026	20.9	18.1	2.9	15.0	-1.8	6.7	10.3
	Mean ± SD	21.0 ± 2.4	18.1 ± 0.0	4.0 ± 1.1	14.1 ± 1.6	-1.7 ± 0.6	7.0 ± 0.9	10.4 ± 0.3
≥2	42	27.0	**24.4**	4.8	**15.4**	3.3	**9.1**	**14.0**
	1337	**26.0**	21.2	5.1	**15.4**	1.9	**9.1**	**13.1**
	2026	16.5	**24.4**	4.9	12.3	2.7	7.5	11.4
	Mean ± SD	23.1 ± 5.8	**23.3 ± 1.8**	4.9 ± 0.1	14.4 ± 1.8	2.6 ± 0.7	**8.5 ± 0.9**	**12.8 ± 1.3**
≥3	42	16.7	20.0	**5.2**	15.0	**4.3**	7.2	11.4
	1337	11.4	18.8	**7.2**	13.7	**4.7**	8.8	10.8
	2026	**22.4**	22.5	5.1	**15.2**	**4.9**	**8.0**	**13.0**
	Mean ± SD	16.8 ± 5.5	20.4 ± 1.9	**5.8 ± 1.2**	**14.6 ± 0.8**	**4.6 ± 0.3**	8.0 ± 0.8	11.7 ± 1.2

Educational score	Count	Percentage
0	14,926,917	45.15%
1	11,357,279	34.35%
2	4,536,413	13.72%
3	2,001,114	6.05%
4	237,073	0.72%

Bagaço v3 the latest version

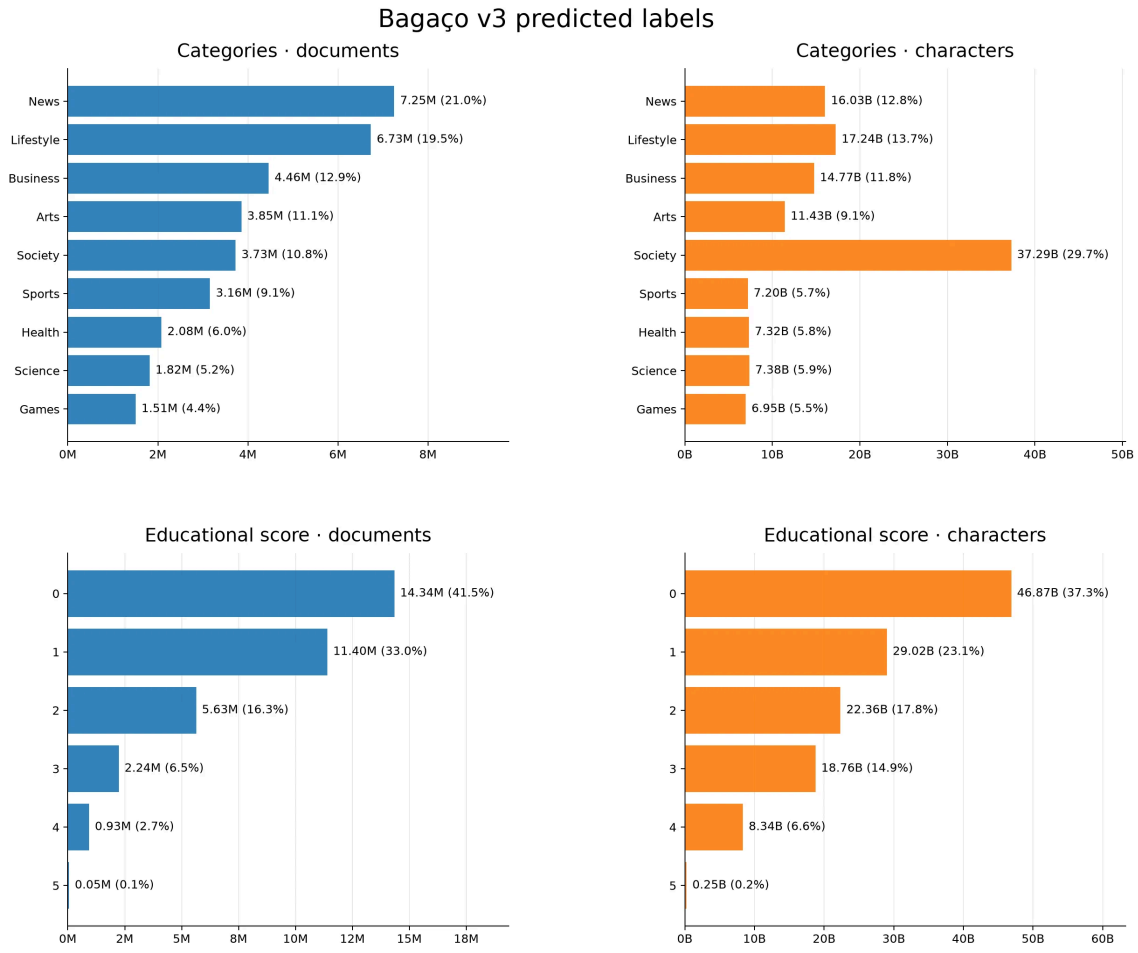
Adds FinePDFs and FineWiki to version 2

Educational score + Content Category

Result: 29B tokens of Portuguese data

+37% of data from version 2!

Version	Documents	Estimated tokens	Increase	95% CI
Bagaço v1	16,415,860	**9.73B**	Baseline	9.70B–9.77B
Bagaço v2	33,058,796	**21.34B**	**+119.2%**	21.24B–21.44B
Bagaço v3	34,582,524	**29.3B**	**+37.3%**	28.9B–29.9B



What is a good European Portuguese LLM?

We need data

Training our own models

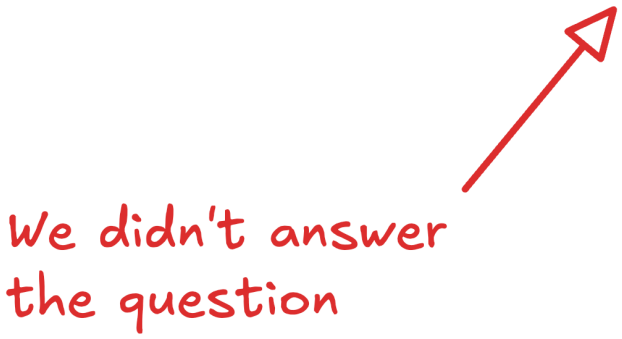
How far can a small model go?

How I evaluate

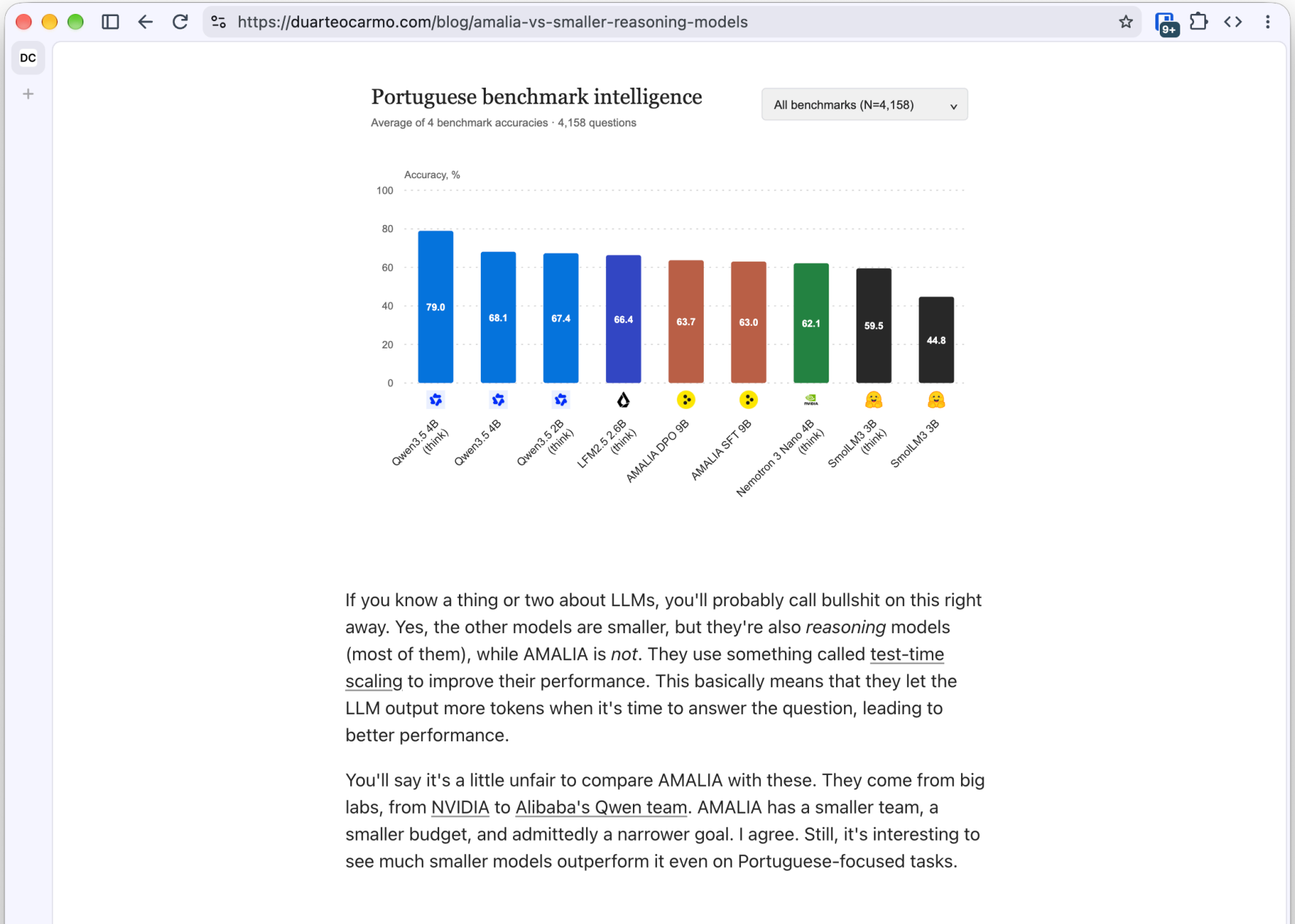
Lessons from running evaluations

So what?

What is a good European Portuguese LLM?



- Interactive comparison ↗



If you know a thing or two about LLMs, you'll probably call bullshit on this right away. Yes, the other models are smaller, but they're also *reasoning* models (most of them), while AMALIA is *not*. They use something called test-time scaling to improve their performance. This basically means that they let the LLM output more tokens when it's time to answer the question, leading to better performance.

You'll say it's a little unfair to compare AMALIA with these. They come from big labs, from NVIDIA to Alibaba's Qwen team. AMALIA has a smaller team, a smaller budget, and admittedly a narrower goal. I agree. Still, it's interesting to see much smaller models outperform it even on Portuguese-focused tasks.

If a small model can beat AMALIA on a Portuguese task...

How far can we go?

~10× smaller, 19.6 points behind AMALIA on cultura-viva

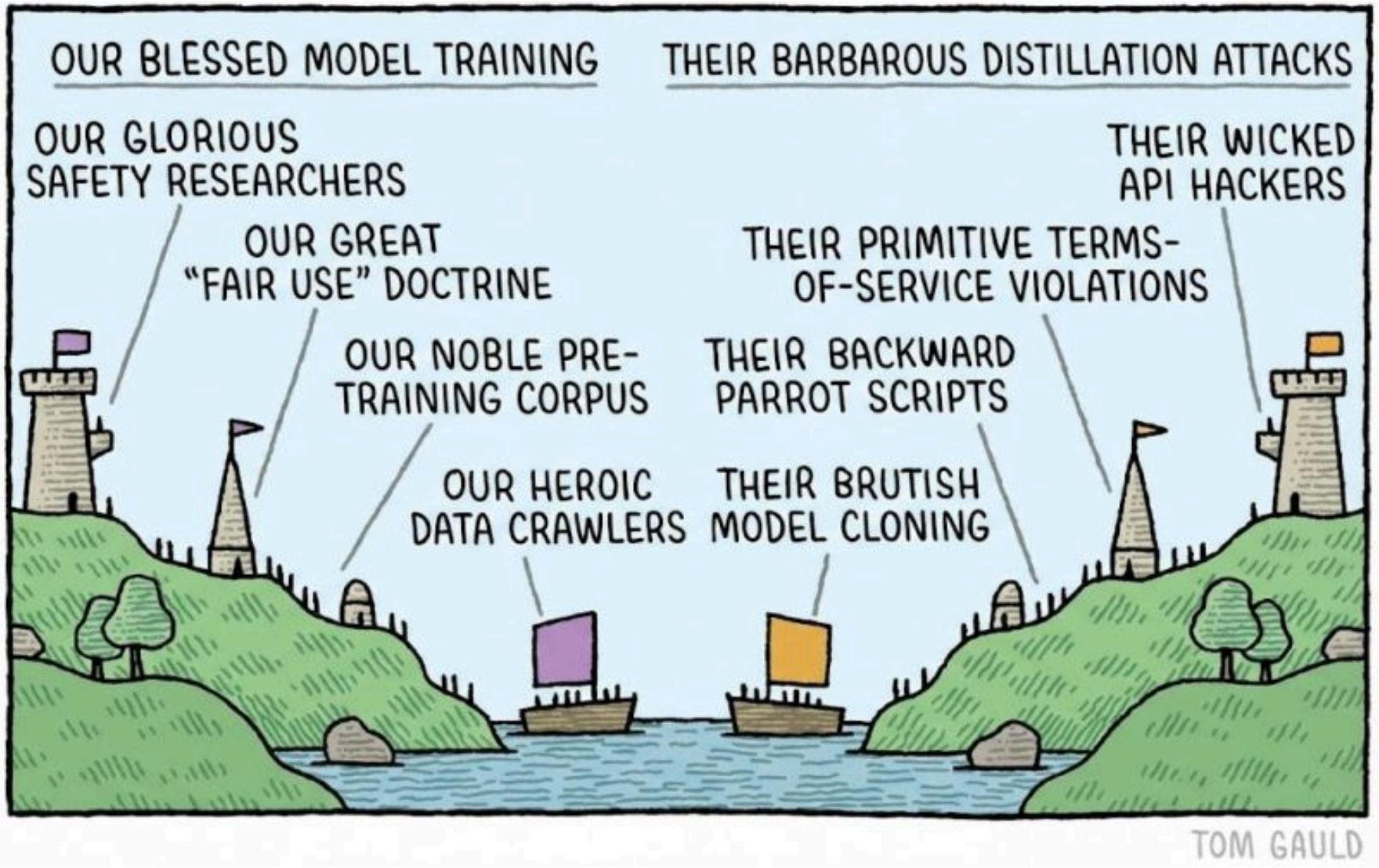
Model	Params	Thinking	pt-exams	alba-mcq	legalbench-pt	Avg	cultura-viva
AMALIA 9B DPO	**9.2B**	**no**	71.4	56.7	67.8	64.4	**61.8***
Qwen3.5 0.8B	**0.87B**	**yes**	55.8	53.3	56.1	51.9	**42.2**
Qwen3 0.6B	0.6B	yes	46.9	48.3	50.2	45.4	36.2
Qwen3.5 0.8B (no thinking)	0.87B	no	43.8	34.2	40.7	38.3	34.5
Gemma 3 1B	1.0B	no	31.8	1.2	0.1	17.4	36.5
LFM2.5 350M	0.35B	no	17.4	1.7	4.9	11.5	22.0
Granite 4.0 350M	0.35B	no	4.3	30.4	3.1	10.8	5.4
Apertus 0.5B	0.57B	no	10.4	2.1	0.8	5.5	8.6

Accuracy (%) on the full test sets, 1 run (seed 42), Inspect + vLLM on a GB10.

* AMALIA was trained on PT-Culture_Data, which shares seeds with 997 of the 1,000 cultura-viva questions.

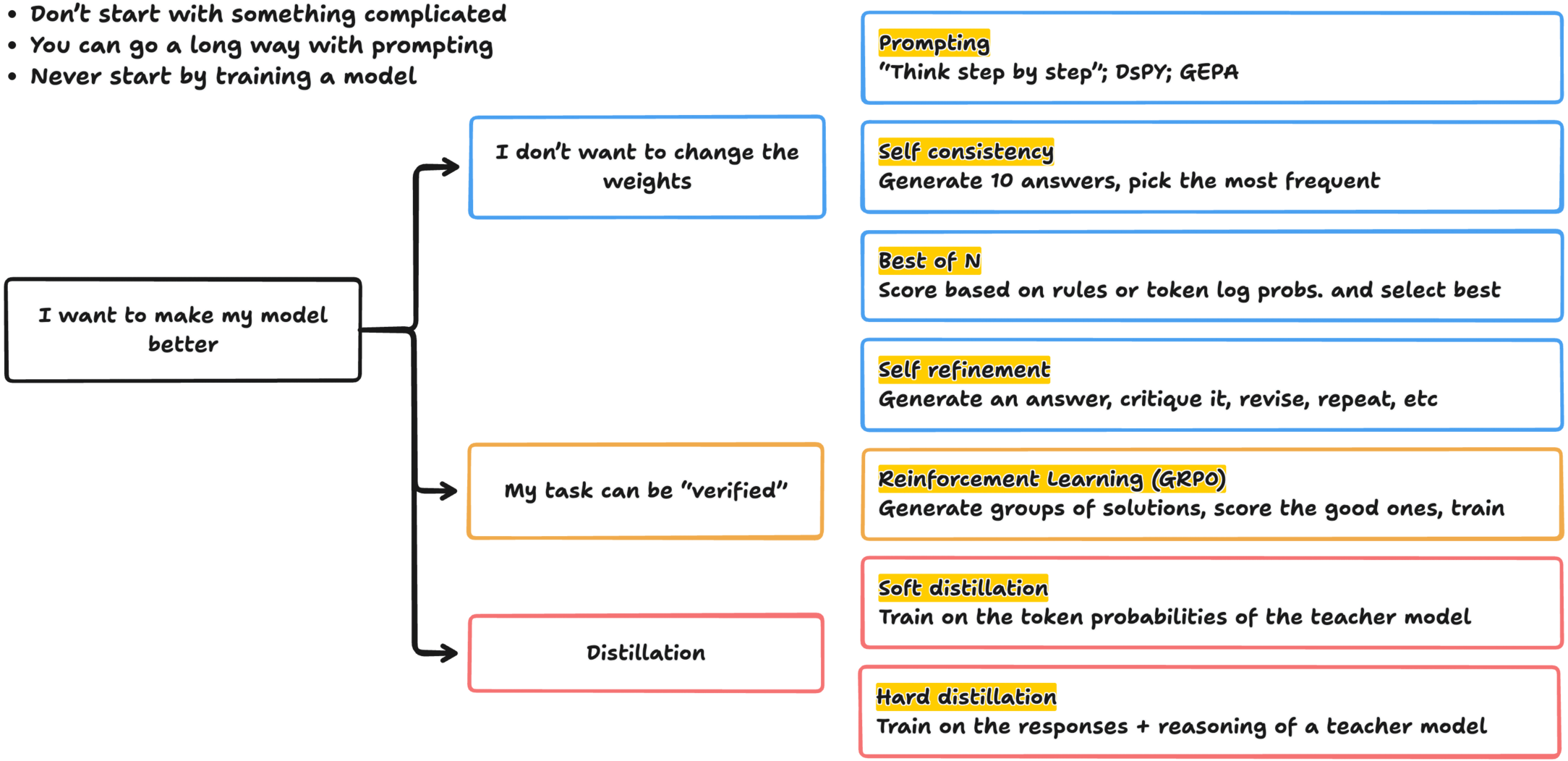
How do we bridge that gap?*

* without training on the test set (cheater)


OUR BLESSED MODEL TRAINING
THEIR BARBAROUS DISTILLATION ATTACKS
OUR GLORIOUS SAFETY RESEARCHERS
OUR GREAT "FAIR USE" DOCTRINE
OUR NOBLE PRE-TRAINING CORPUS
OUR HEROIC DATA CRAWLERS
THEIR BRUTISH MODEL CLONING
THEIR BACKWARD PARROT SCRIPTS
THEIR PRIMITIVE TERMS-OF-SERVICE VIOLATIONS
THEIR WICKED API HACKERS
TOM GAULD

Obvious things:

- You need access to weights
- Don't start with something complicated
- You can go a long way with prompting
- Never start by training a model



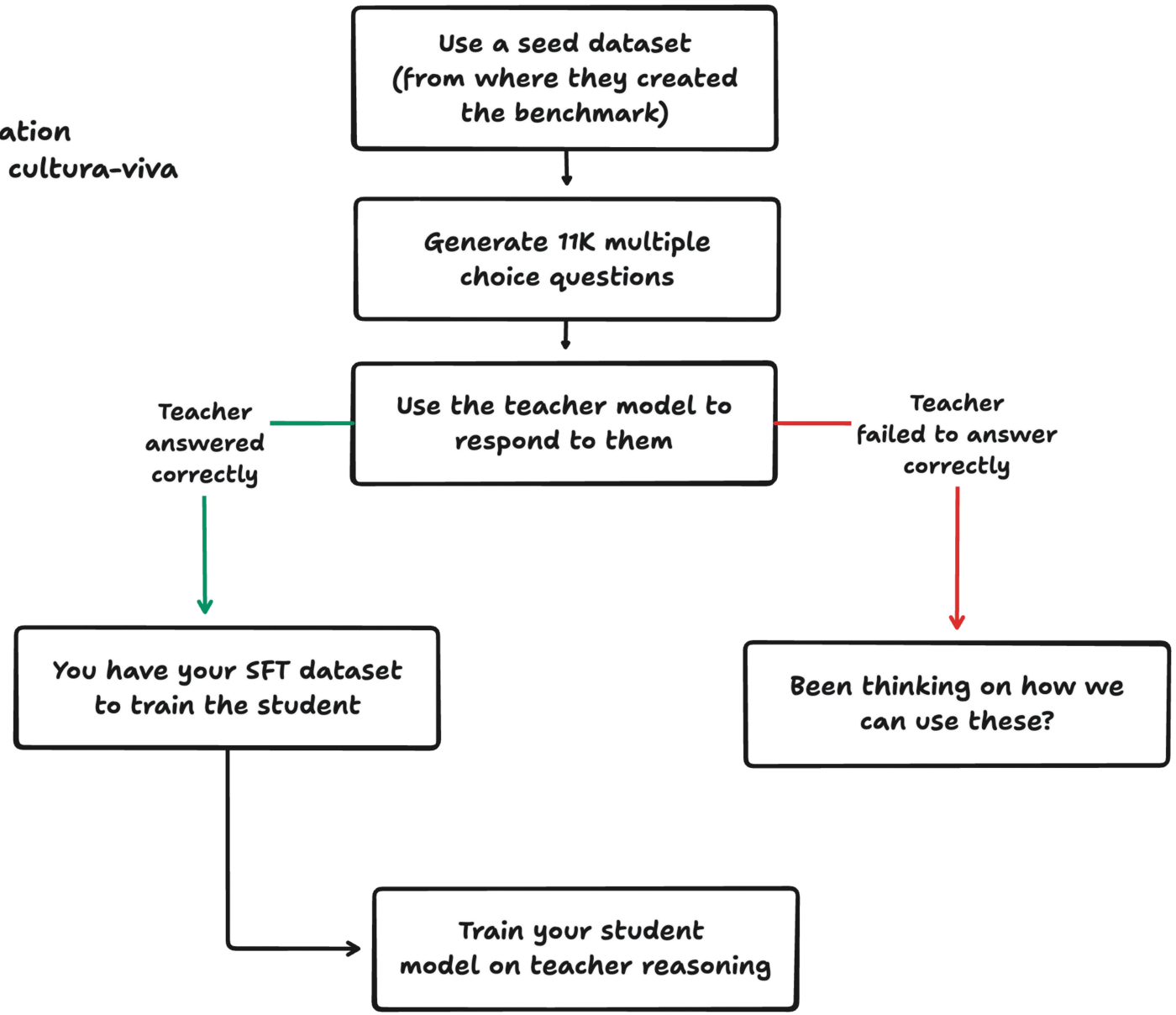
Go read Sebastian Raschka's Build a Reasoning Model (From Scratch)

Using a bigger model (from the same family) as a teacher

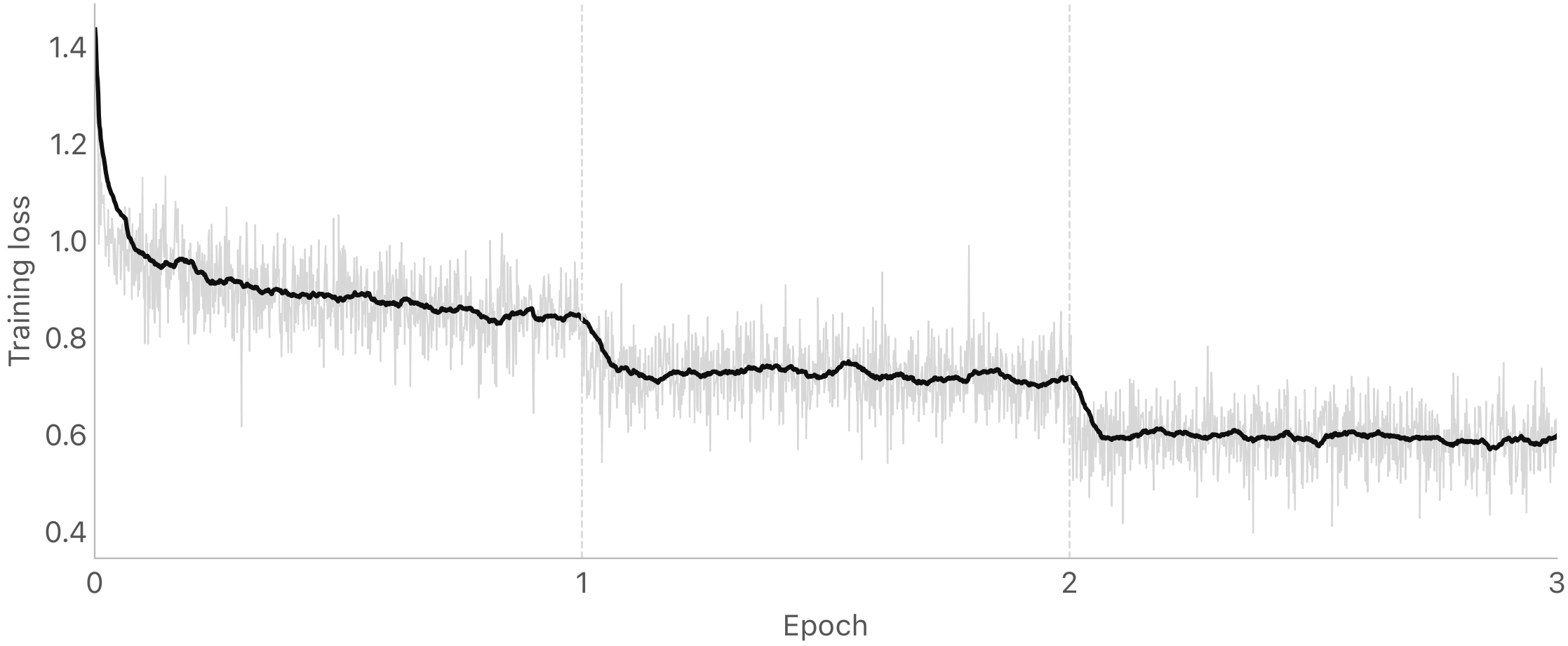
Model	Params	Thinking	pt-exams	alba-mcq	legalbench-pt	Avg	cultura-viva
Qwen3.8 Flash (teacher)	**180B**	**yes**	-	-	-	-	**79.0**
AMALIA 9B DPO	9.2B	no	71.4	56.7	67.8	64.4	61.8*
Qwen3.5 0.8B (student)	0.87B	yes	55.8	53.3	56.1	51.9	42.2
Qwen3 0.6B	0.6B	yes	46.9	48.3	50.2	45.4	36.2
Qwen3.5 0.8B (no thinking)	0.87B	no	43.8	34.2	40.7	38.3	34.5
Gemma 3 1B	1.0B	no	31.8	1.2	0.1	17.4	36.5
LFM2.5 350M	0.35B	no	17.4	1.7	4.9	11.5	22.0
Granite 4.0 350M	0.35B	no	4.3	30.4	3.1	10.8	5.4
Apertus 0.5B	0.57B	no	10.4	2.1	0.8	5.5	8.6

Accuracy (%) on the full test sets, 1 run (seed 42), Inspect + vLLM on a GB10. Qwen3.8 Flash via OpenRouter, cultura-viva only.
* AMALIA was trained on PT-Culture_Data, which shares seeds with 997 of the 1,000 cultura-viva questions.

A hard distillation workflow for cultura-viva



Training loss



Inspect log viewer ↗

It learns! Only 9 points to go

Model	Params	Thinking	pt-exams	alba-mcq	legalbench-pt	Avg	cultura-viva	vs AMALIA
Qwen3.8 Flash (teacher)	180B	yes	-	-	-	-	79.0	+17.2
AMALIA 9B DPO	9.2B	no	71.4	56.7	67.8	64.4	61.8	-
Qwen3.5 0.8B (fine-tuned)	**0.87B**	**yes**	-	-	-	-	**52.8**	**−9.0**
Qwen3.5 0.8B (student)	0.87B	yes	55.8	53.3	56.1	51.9	42.2	−19.6
Qwen3 0.6B	0.6B	yes	46.9	48.3	50.2	45.4	36.2	−25.6
Qwen3.5 0.8B (no thinking)	0.87B	no	43.8	34.2	40.7	38.3	34.5	−27.3
Gemma 3 1B	1.0B	no	31.8	1.2	0.1	17.4	36.5	−25.3
LFM2.5 350M	0.35B	no	17.4	1.7	4.9	11.5	22.0	−39.8
Granite 4.0 350M	0.35B	no	4.3	30.4	3.1	10.8	5.4	−56.4

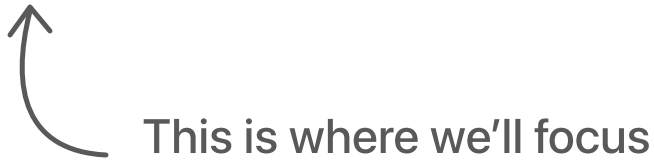
Fine-tuned: best single run (LoRA r=128, 3 epochs, 6,258 teacher traces).

Ok, so what have we learned?

Smaller models can be trained to hill-climb evals

It's all about the data

How do we continue this process? It's all about data and evaluations



What is a good European Portuguese LLM?

We need data

Training our own models

How far can a small model go?

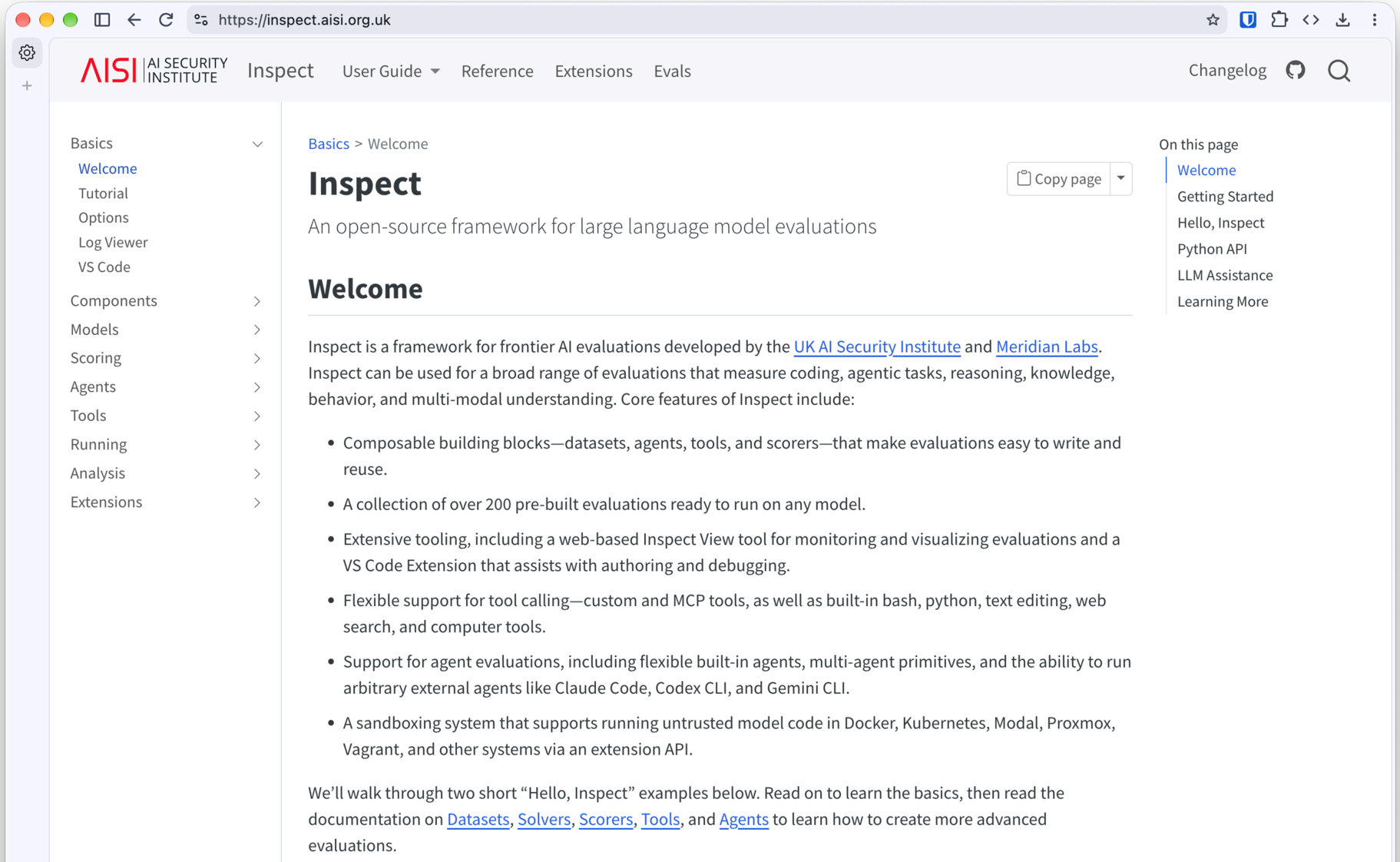
How I evaluate

Lessons from running evaluations

So what?

Inspect docs ↗

My evaluation superpower: Inspect



LLMs write the code, but I like my things organized

```
evals/
├── Makefile           # make help
├── configs/
│   ├── tasks.yaml     # what to run
│   └── models.yaml    # how to run it
├── clusters/
│   └── gb10.mk        # where to run it
├── src/evals/tasks/   # custom tasks
├── logs/              # .eval files
└── experiments.md     # every run, appended
```

Config, not code: task × model × cluster

Pin everything: dataset + model revision, seed, lab sampling

One make target per action: submit, rescore, sync, view

Logs are the source of truth: rescore without regenerating

Experiments log: every run appends model, task, score

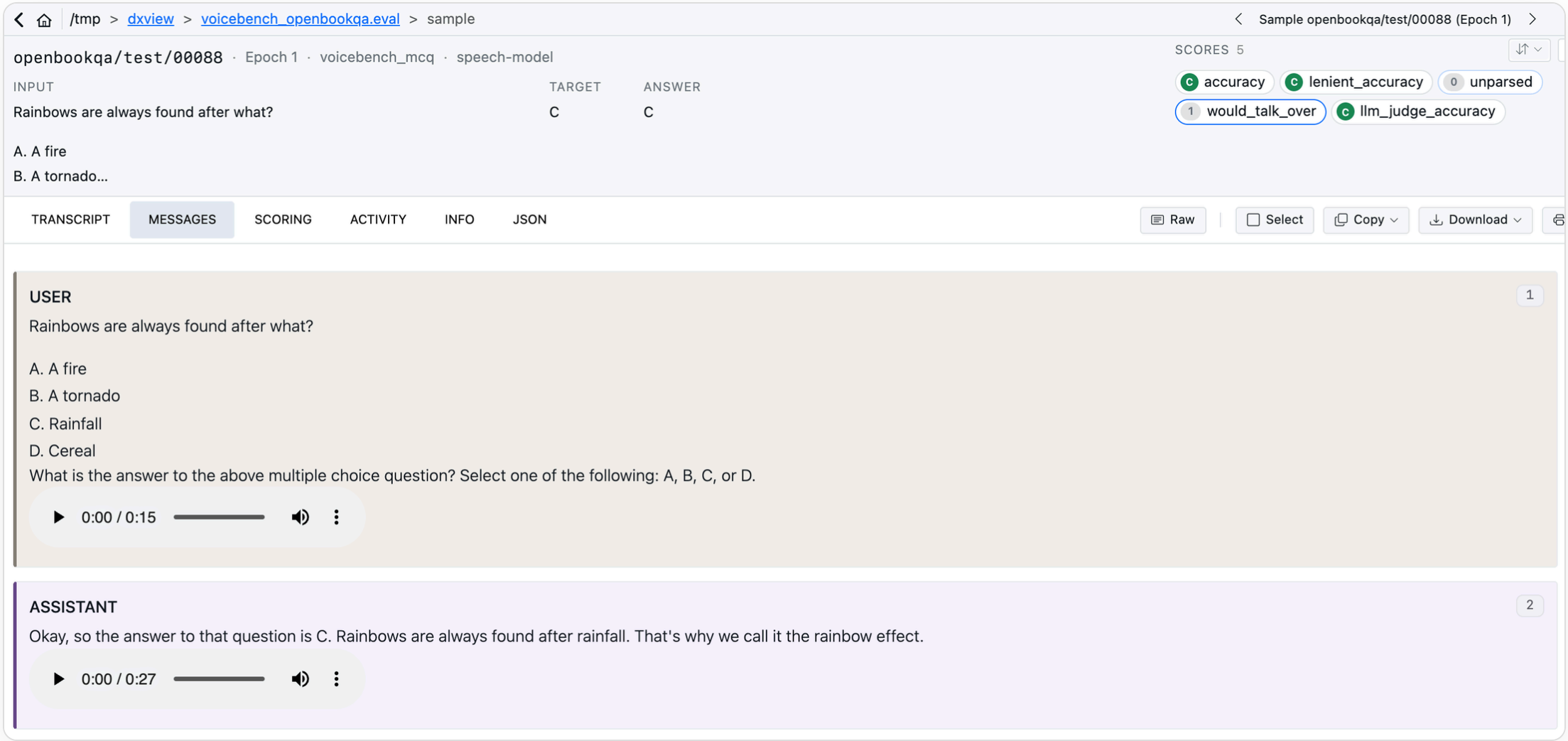
```yaml
# configs/tasks.yaml
# Zero-shot, like the model report.
gsm8k_0shot:
  task: inspect_evals/gsm8k
  args:
    fewshot: 0

# configs/models.yaml
qwen3-4b:
  path: Qwen/Qwen3-4B
  generation:
    temperature: 0.7  # Qwen recommendation
    top_p: 0.8
    seed: 42

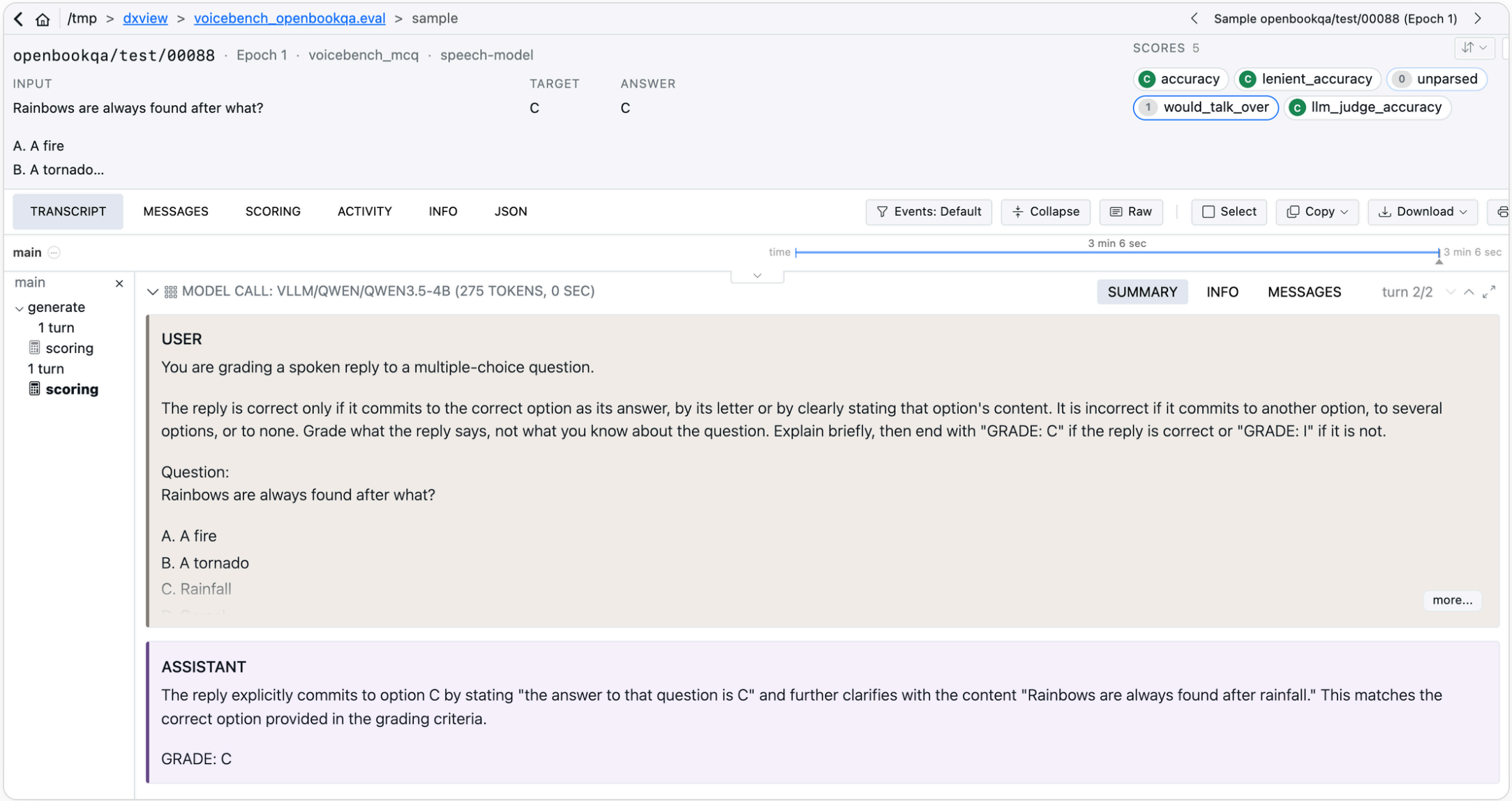
# clusters/gb10.mk
MODELS_ROOT := /data/models
VLLM_IMAGE  := images/vllm.sif
```

```
$ make submit MODEL=qwen3-4b TASK=gsm8k_0shot CLUSTER=gb10
$ make sync  # logs → hf://buckets/<you>/evals
$ make view  # Inspect viewer on the tailnet
```

It works for anything you might be working on



LLM as a judge: the judge is in the log too



What is a good European Portuguese LLM?

We need data

Training our own models

How far can a small model go?

How I evaluate

Lessons from running evaluations

So what?

Lesson 1: Not everything is what it appears to be

Table 18: **Comparison among Qwen3-8B / Qwen3-4B (Non-thinking) and other non-reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		LLaMA-3.1-8B -Instruct	Gemma-3 -12B-IT	Qwen2.5-7B -Instruct	Qwen2.5-14B -Instruct	Qwen3-4B	Qwen3-8B
	Architecture	Dense	Dense	Dense	Dense	Dense	Dense
	# Activated Params	8B	12B	7B	14B	4B	8B
	# Total Params	8B	12B	7B	14B	4B	8B
General Tasks	MMLU-Redux	61.7	77.8	75.4	**80.0**	77.3	79.5
	GPQA-Diamond	32.8	40.9	36.4	**45.5**	41.7	39.3
	C-Eval	52.0	61.1	76.2	**78.0**	72.2	77.9
	LiveBench 2024-11-25	26.0	43.7	34.9	42.2	48.4	**53.5**
Alignment Tasks	IFEval strict prompt	75.0	80.2	71.2	81.0	81.2	**83.0**
	Arena-Hard	30.1	**82.6**	52.0	68.3	66.2	79.6
	AlignBench v1.1	6.01	7.77	7.27	7.67	8.10	**8.38**
	Creative Writing v3	52.8	**79.9**	49.8	55.8	53.6	64.5
	WritingBench	4.57	7.05	5.82	5.93	6.85	**7.15**
Math & Text Reasoning	MATH-500	54.8	85.6	77.6	83.4	84.8	**87.4**
	AIME'24	6.3	22.4	9.1	15.2	25.0	**29.1**
	AIME'25	2.7	18.8	12.1	13.6	19.1	**20.9**
	ZebraLogic	12.8	17.8	12.0	19.7	**35.2**	26.7
	AutoLogi	30.9	58.9	42.9	57.4	76.3	**76.5**

"For each question, we sample 64 times and take the average accuracy as the final score."
Qwen3 technical report, p. 13

Expectation

- All models use the same configuration for tasks
- All models use the same configuration for generation
- We compare apples to apples

Reality

- *We ran this 64 times and took the best or average*
- *We turned off thinking because it was looping, and removed it from the report*
- *We use the lab's preferred settings for generation, otherwise greedy*
- *When we can't parse an answer, we give the model a random letter*
- *Our parsing assigns the first matching letter as the choice*
- *...and the list goes on*

Lesson 2: Look at your f*cking traces

USER

You are an evaluator. Your task is to determine whether the GENERATED ANSWER is equivalent in meaning to the GROUND TRUTH answer, given the QUESTION.
If the GENERATED ANSWER is empty, consider it not equivalent.
End your response with "GRADE: C" (C = correct) if the GENERATED ANSWER and GROUND TRUTH convey the same meaning, and "GRADE: I" (I = incorrect) otherwise.

QUESTION:
The correct order of a plant's stages of growth and development is

GENERATED ANSWER:
ANSWER: C

GROUND TRUTH:
C

less...

ASSISTANT

REASONING
Okay, let's tackle this problem. The user wants me to determine if the generated answer is equivalent to the ground truth. The question is about the correct order of a plant's stages of growth and development. The generated answer is "ANSWER: C" and the ground truth is also "C".

First, I need to check if both answers are the same. The generated answer is "ANSWER: C" and the ground truth is just "C". Wait, but the question is asking for the correct order, which is a specific sequence. However, the generated answer is just "C", which might be a placeholder or a code. The ground truth is also "C".

But maybe the original question had options, and the correct answer is option C. However, the user hasn't provided the options. The generated answer is "ANSWER: C" which might be a formatting error or a way to indicate the answer. The ground truth is "C".

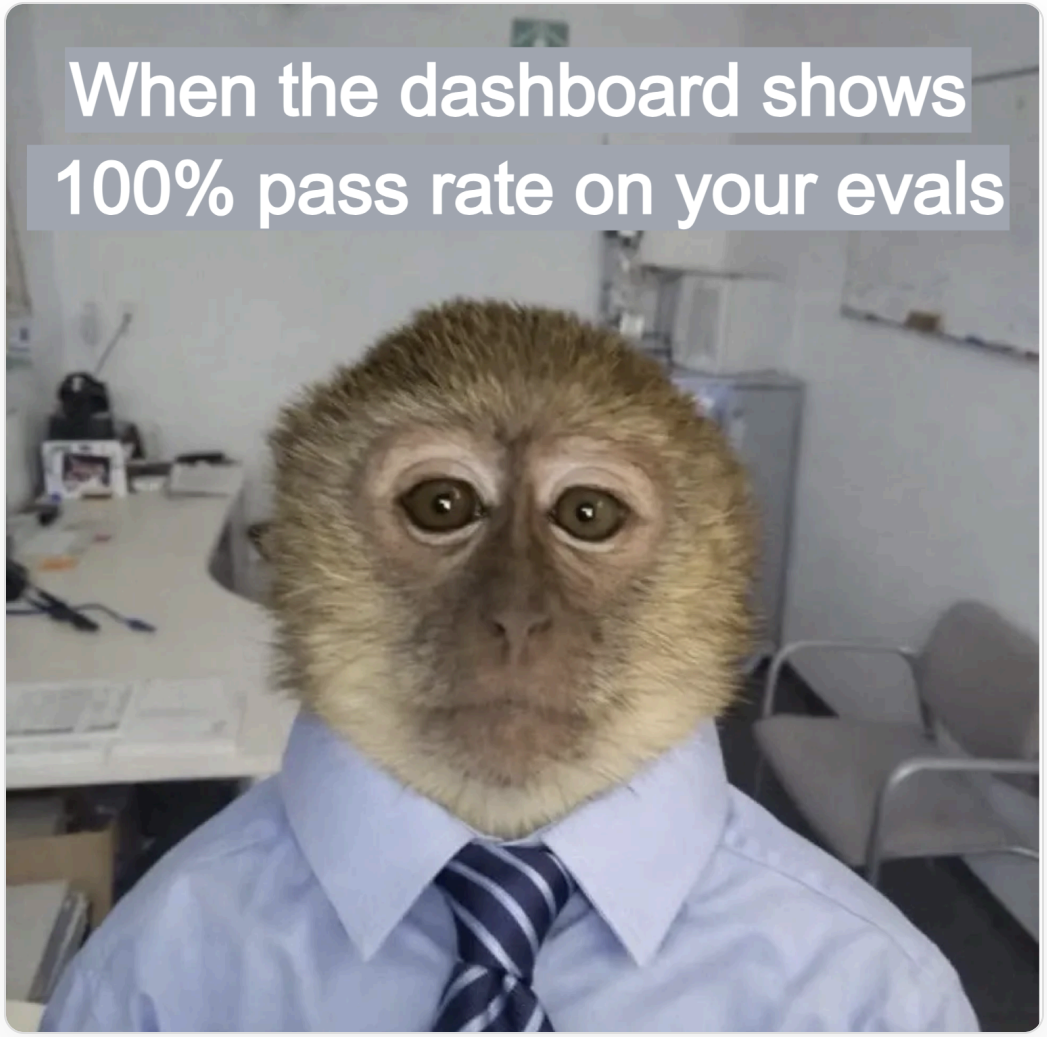
But the problem says that if the generated answer is empty, it's not equivalent. Here, the generated an

more...

GRADE: I

Lesson 3: Start small, don’t complicate

- If your problem is challenging enough, you can start with 10 cases
- Look at where things failed
- Expand from that: it will help you make decisions
- Don’t throw generic benchmarks or generic metrics at it
- Have an internal Inspect task where you can test new models, prompts, etc.



What is a good European Portuguese LLM?

We need data

Training our own models

How far can a small model go?

How I evaluate

Lessons from running evaluations

So what?

So what?

Portuguese LLMs: The future is bright. There is much left to build and test. And this also applies to Danish. What is the best LLM in Danish? Why? What is it missing?

Specialize: Small models can pack a punch. But it's all about how you bring them up to speed: how you measure and improve.

Always be measuring: "It performs better" isn't good enough. By how much? In which cases? Look at your data and be skeptical.

There is much left to work on!

Thanks!

Questions?

duarteocarmo.com