BR Sovereign LLM: An Auditable Reference Implementation and Experimental Protocol for Brazilian Portuguese Pretraining

Guilherme Monteiro
FIAP
ORCID: 0009-0008-5294-224X

Manuscript version 1.0.0, August 2026

Abstract

Language-model sovereignty is often reduced to accelerator location or provider nationality, neither of which establishes whether an institution can lawfully reconstruct data, restore software, recover training, or operate exported artifacts without a mandatory external service. We present BR Sovereign LLM, an implemented reference system and experimental protocol for auditable from-scratch Brazilian Portuguese pretraining. It separates scientific settings from execution topology, enforces document-level rights and transformation lineage, produces content-addressed release manifests, and reports sovereignty as a vector of evidence instead of a composite score. Validation of release 0.1.1 exercised the complete local artifact path: eight reviewed synthetic documents passed the fail-closed corpus boundary; a byte-level tokenizer had zero round-trip failures; and a 149,696-parameter model recovered an interrupted six-step CPU trajectory with an identical final tensor hash. The implementation passed 134 tests with 81.80% branch-aware test coverage. Three public repositories containing 188 files were then downloaded without authentication and matched the approved release byte for byte. These are engineering results, not evidence of model quality or national-HPC execution. Supercomputer performance, scientific corpus quality, data residency, energy, and target-scale outcomes remain `not_measured`.

1 Introduction

Brazilian institutions increasingly depend on language models whose training data, system software, checkpoints, and operating interfaces are controlled outside the institution. Hosting an inference endpoint domestically addresses only one part of that dependency. A training program can still be unable to prove rights over its corpus, resume an interrupted run, reconstruct its environment without a foreign package service, or export a model independently of a vendor-specific interface.

Brazilian Portuguese has established model and corpus work. TeenyTinyLlama released compact models trained from scratch [5]. Tucano released a multi-scale from-scratch family and the GigaVerbo corpus [6]. Tucano 2 extended the open Portuguese training stack and mixed from-scratch and continual-pretraining regimes [7]. Sabiá showed the value of Portuguese continual pretraining [19]. These results make broad claims such as "the first Portuguese LLM" indefensible.

We study a different question: which technical and legal capabilities can a Brazilian institution demonstrate while training a model on controlled infrastructure? Our contribution is a protocol that makes those capabilities falsifiable. It covers corpus rights, artifact custody, offline operation, rebuildable dependencies, scheduler portability, checkpoint recovery, model export, Brazilian Portuguese quality, and remaining vendor exposure.

This work makes three concrete contributions. First, it defines operational sovereignty as independently auditable capabilities with explicit missing-data semantics. Second, it implements those boundaries in a public reference system covering provenance, configuration fingerprints, checkpoint integrity, statistical analysis, release manifests, and fail-closed assurance gates. Third, it reports an end-to-end engineering validation, including deterministic recovery and anonymous reconstruction of the public release. The validation is deliberately small enough to reproduce locally and is not used as a proxy for target-scale quality or supercomputer performance.

The system is designed for public high-performance computing, but it does not assume an allocation, accelerator model, node count, or data-residency result. Those facts enter the evidence record only after measurement.

2 Related work

2.1 Brazilian Portuguese models and corpora

TeenyTinyLlama documents 160M and 460M parameter models for Brazilian Portuguese [5]. Tucano scales this approach across 160M, 630M, 1.1B, and 2.4B parameter models [6]. Its GigaVerbo dataset demonstrates that large Portuguese aggregates can be assembled, while also motivating a stricter distinction between repository metadata and the rights of underlying web documents.

Tucano 2 reports a suite between approximately 0.5B and 3.7B parameters [7]. The 0.6B base model is trained from scratch, while larger members use continual pretraining from Qwen3. This distinction matters when measuring control over base weights. Sabiá is a strong Portuguese continual-pretraining precedent, rather than a from-scratch training baseline [19].

Carolina emphasizes provenance, typology, TEI metadata, and versioning for Brazilian Portuguese text [8]. Its source-specific licensing model supports the governance direction of this work. It also demonstrates why a corpus package should not be assigned a permissive blanket license when documents retain different terms.

2.2 Open and public-HPC training

Pythia provides controlled model scales, identical data order, and frequent checkpoints [2]. OLMo releases training data, code, and evaluation artifacts for scientific inspection [10]. EuroLLM demonstrates multilingual training on European public HPC [14]. Apertus provides a recent reference for public compute, data reconstruction, compliance, code, checkpoints, and multilingual coverage [24].

TorchTitan supplies a PyTorch-native distributed executor with composable parallelism and checkpointing [12]. We place it behind a project-owned backend boundary. This retains the option to replace it after target-cluster profiling without changing the scientific study specification.

3 Operational definition

We define operational sovereignty through independently reported capabilities. Table 1 gives the targets. We do not average them. An average could conceal a mandatory remote API or incomplete rights coverage behind strong performance elsewhere.

The study fingerprint includes corpus snapshot, tokenizer, model, optimizer, batch order, token budget, seed, and numerical policy. The execution fingerprint includes the container, hardware, world size, parallelism, drivers, kernels, and scheduler. A portability experiment can change only the execution fingerprint.

Capability	Measurement	Target
Provenance coverage	Training bytes linked to source, rights evidence, hashes, and transformations	100%
Redistribution coverage	Released bytes whose rights allow redistribution and derivatives	100%
Artifact custody	Required artifacts mirrored and hash verified locally	100%
Offline completion	Required post-snapshot steps completed without network	100%
Mandatory remote services	Remote APIs required by training, evaluation, or inference	0
Portability	Execution profiles satisfying specified equivalence criteria	fraction tested
Recovery	Successful fault-injected resumes	10/10
Export completeness	Tokenizer, configuration, weights, tests, and checksums in open formats	100%
Residency and energy	Direct infrastructure evidence	report if measured

Table 1: Capability-specific sovereignty evidence. Missing evidence is reported as `not_measured`.

4 Hypotheses

Table 2 summarizes six confirmatory hypotheses. The complete exclusion and analysis rules are specified in the prospective registration draft and must be frozen before confirmatory execution.

ID	Claim	Support criterion
H1	Local and Slurm single-node runs with one study fingerprint are equivalent.	95% interval within ± 0.02 nats/token
H2	Interrupted training restores all required state.	10/10 recoveries; same-runtime state hashes match
H3	Nano and micro scaling predicts the unrevealed next scale.	relative loss error $\leq 5\%$
H4	A licensed pt-BR-weighted allocation improves native pt-BR modeling under equal bytes and FLOPs.	at least 5% bits/byte improvement; guardrail $\leq$ 3% regression
H5	Legal-lineage controls have bounded overhead.	preprocessing $< 10\%$; steady-state training $< 2\%$
H6	The approved mirror supports a complete network-isolated workflow.	all six steps pass in three clean environments

Table 2: Prospectively specified hypotheses and primary support criteria.

5 Corpus construction

The pipeline begins with a source registry that records the rights basis, license URI, evidence digest, acquisition method, revision policy, attribution, and restrictions. Public access alone is insufficient. Unknown, noncommercial, no-derivatives, and ambiguous aggregate sources fail closed.

The initial eligible families are dated Portuguese Wikimedia text under its applicable CC BY-SA/GFDL terms, with project-, page-, and media-specific exceptions retained in the provenance manifest [30]; Portal da OBMEP material identified by the portal as CC BY 3.0 Brazil [22]; and Brazilian laws, regulations, judicial decisions, and other official acts covered by Article 8(IV) of Law 9,610/1998 [3]. ProEdu supports multiple Creative Commons options and directs users to inspect the permissions and restrictions recorded for each resource, so the pipeline verifies the declared license per item [26]. Current SciELO Brasil criteria make article full texts available under CC BY or, at the authors' request, CC0, and metadata under CC0; the pipeline likewise records and evaluates the license declared for each article [23]. Common Crawl, mC4, OSCAR, HPLT, CulturaX, BrWaC, Carolina as an aggregate, and GigaVerbo as an aggregate do not enter the main release without document-level reacquisition and evidence.

In the protocol repository, digests of registry statements provide tamper-evident metadata but do not independently prove the asserted rights. Scientific admission additionally requires archived evidence bytes, media type, source revision, snapshot digest, and item attribution where applicable; legal review remains a separate release gate.

Each document records source identity, revision, URLs, authorship or issuing body, rights basis, exact license, evidence hash, raw and normalized hashes, language variant, domain, PII status, transformations, stable group, split, and takedown state. The flow applies immutable acquisition, lossless NFC normalization, language and policy classification, exact and near deduplication, benchmark decontamination, group-aware splitting, and license-separated Parquet output.

The training representation uses content-addressed unsigned 32-bit token shards with an index and manifest. Every shard must be reproducible from the document snapshot and tokenizer fingerprint.

6 Tokenizer and model ladder

A 32,768-entry byte-level BPE tokenizer is trained on a deterministic, source-balanced sample. Lossless NFC normalization preserves orthography, and byte fallback prevents out-of-vocabulary text. Selection uses round-trip tests, bytes per token, characters per token, tokens per word, vocabulary utilization, throughput, and downstream bits per byte across general, scientific, legal, educational, code, and mathematical domains.

All models use one decoder-only family with RMSNorm, RoPE, SwiGLU, and grouped-query attention. Table 3 is a planning ladder. The final stage is chosen only after the licensed token inventory, throughput, storage, and allocation are known.

Stage	Approx. parameters	Planning tokens	Role
Nano	20–30M	0.25–0.50B	pipeline and recovery
Micro	100–125M	2–3B	mixture and seed experiments
Small	350M	7B	scaling holdout
Medium	1.3B	26B	conditional multi-node target
Target	3B	60B	conditional national-HPC study
Stretch	7B	140B	future work

Table 3: Conditional scale ladder. Planning tokens are not completed runs.

7 Training and recovery

The canonical launcher is torchrun. Slurm supplies allocation and node mechanics but cannot override scientific parameters. The initial distributed candidate is TorchTitan with FSDP2. Tensor, pipeline, or context parallelism is introduced only when profiling demonstrates a need.

A complete checkpoint includes model, optimizer, scheduler, random generator state for every rank, data cursor, fingerprints, dependency lock, container digest, topology, and numerical flags. Distributed checkpoint data are written to a temporary generation and committed by writing the manifest last. Release checkpoints are separately consolidated into safetensors.

Structured JSONL telemetry includes experiment, rank, step, tokens, loss, gradient norm, learning rate, throughput, memory, data wait, checkpoint duration, and measured energy when available. Remote telemetry is optional.

8 Evaluation and analysis

The primary language-model outcome is bits per UTF-8 byte on a native, licensed, source-stratified Brazilian Portuguese holdout. Nats per token and perplexity remain secondary because tokenizer efficiency affects token boundaries.

The planned suite includes CALAME-PT [13, 16]; GSM8K-PT [4, 21]; IFEval-PT [20, 31]; Belebele [1]; MASSIVE [9]; evaluation-only FLORES+ v4.6 [15, 17]; FaQuAD [27]; Universal Dependencies Portuguese Bosque and PUD [25, 28, 29]; and MMMLU Portuguese [11, 18]. Native and translated results are reported separately. Examination-derived datasets such as ENEM and OAB wrappers remain quarantined until the underlying rights chain is reviewed.

Item-level paired bootstrap intervals use 10,000 prospectively specified resamples. Portability uses paired equivalence intervals. Missing telemetry is `not_measured`, not zero. A weak or null result is retained and cannot be excluded for failing to support a hypothesis.

9 Reference implementation validation

We evaluated release 0.1.1 as an engineering artifact on macOS arm64 with Python 3.13.12 and CPU execution for the recovery experiment. The dependency environment used PyTorch 2.13.0, Transformers 5.15.1, Tokenizers 0.22.2, Datasets 5.0.1, and PyArrow 25.0.1. Table 4 reports only observed quantities whose evidence is included in the public release or the content-addressed local assurance record.

The tokenizer smoke run encoded 249 words as 736 tokens, corresponding to 2.4008 UTF-8 bytes per token and 2.9558 tokens per whitespace-delimited word. Those values diagnose the executable path on eight short documents; they do not select the future 32,768-entry scientific tokenizer.

For the recovery test, a Llama-compatible decoder was initialized from seed 1701. The uninterrupted and interrupted-then-resumed trajectories began at loss 6.1313037872 and ended at loss 6.1140251160. Both final serialized tensor states had the same SHA-256 digest, recorded in the public smoke report. The test also restored Python and PyTorch random-generator state, observed finite nonzero gradients, and confirmed that weights changed. Because the interruption was simulated in one process, this result validates serialization semantics but does not support confirmatory hypothesis H2.

The release then traversed private staging, content-addressed verification, owner-approved public transition, and a new anonymous download. The public dataset loaded offline with one eight-row,

Boundary	Observed result	Interpretation limit
Repository quality	134 tests passed; 81.80% branch-aware test coverage; strict type checking over 49 source files	Software validation, not scientific model quality
Corpus admission	8/8 fixtures accepted; 0 rejected, quarantined, or marked duplicate	AI-assisted project fixtures only; no scientific corpus claim
Tokenizer path	736 tokens over 1,767 UTF-8 bytes; 0 round-trip failures; 0 unknown tokens	Vocabulary size 447; no domain or fertility-quality claim
Recovery path	149,696 parameters; 6 CPU steps; interruption after step 3; identical final tensor hash	Same-process simulation; no process, scheduler, or node failure claim
Public reconstruction	188/188 files downloaded anonymously with exact inventories and SHA-256 matches	Mirrors release artifacts; does not prove a network-isolated training workflow

Table 4: Measured validation results and their explicit claim boundaries.

21-column split; the exported model and 447-entry tokenizer loaded through the Transformers automatic interfaces. The public dataset viewer returned eight rows with the expected 21-feature schema. All 188 files across the research, corpus, and smoke-model repositories matched the approved candidate bytes. This provides a reproducible release-integrity result while leaving the stronger network-isolated workflow in H6 untested.

10 Validation and release

Fourteen ordered gates cover repository integrity, legal admissibility, acquisition, corpus quality, lineage, benchmark isolation, tokenizer quality, training semantics, HPC preflight, scientific freeze, model evaluation, offline reproduction, private release, and separately approved public release. A later gate cannot turn an earlier failure into a warning.

Hugging Face serves as a publication mirror rather than the authoritative archive. Dataset records retain source-specific licenses. Apache-2.0 applies to project-controlled code and documentation to the extent of the owner's rights. The scientific target-weight license decision is deferred until the corpus and legal review are complete.

11 Artifact availability

The versioned reference implementation, governance documents, experiment configurations, tests, and manuscript source are available in the research snapshot. The corresponding eight-document corpus snapshot and the 149,696-parameter smoke-model snapshot are independently downloadable. Each repository contains a strict release manifest that binds every file by relative path, byte size, and SHA-256 digest. The synthetic fixture and smoke weights are released under Apache-2.0 to the extent of the author's rights; future mixed corpora and scientific model weights require artifact-specific licensing decisions.

12 Threats to validity

The source allowlist may reduce domain and dialect coverage. Automated language, PII, and duplicate classifiers have false positives and false negatives. Single-node equivalence does not establish multi-node or cross-vendor portability. CUDA evidence does not establish ROCm behavior. Small-scale scaling predictions may fail under optimizer, network, or data-regime changes. Benchmark licenses can change, and wrapper licenses may not cover underlying content.

National infrastructure still depends on foreign accelerators, firmware, and parts of the software stack. We therefore measure control and disclosure rather than claiming absolute independence.

13 Ethics and governance

The corpus minimizes personal data and initially excludes high-risk personnel records, children's information, and ambiguous web text. A takedown ledger connects removed documents to corpus and token-shard revisions. Model releases require contamination, memorization, safety, and limitation reports. The protocol cannot eliminate harmful model behavior and does not authorize deployment in high-impact decisions.

14 Scope of evidence and next experiments

This article reports a complete reference implementation and its bounded local and public-release validation. It does not report a scientific corpus snapshot, target-scale model, supercomputer throughput, energy, jurisdictional residency, or confirmatory quality result. Those outcomes remain `not_measured`, and none of H1–H6 is claimed as supported. The next stage is to freeze the confirmatory protocol, complete the remaining statistical and runtime prerequisites, and run the prospectively specified experiments only after a qualifying corpus and infrastructure allocation exist.

15 Conclusion

BR Sovereign LLM turns an otherwise broad sovereignty claim into explicit technical contracts, failure states, and content-addressed evidence. The reference release demonstrates that corpus admission, tokenizer export, checkpoint serialization, release gating, and anonymous reconstruction can be tested together without presenting a tiny smoke model as a scientific result. Its remaining value will depend on whether the same boundaries survive a licensed scientific corpus and real public-HPC execution. Publishing missing outcomes as `not_measured` makes that transition auditable: later evidence can extend the record but cannot silently rewrite what release 0.1.1 established.

Generative-AI disclosure

Generative AI language tools assisted literature triage, drafting, editing, code scaffolding, and test and audit orchestration. They are not authors. The named human author retains full responsibility for verifying and approving the manuscript, citations, data, code, affiliation, and submission.

References

[1] Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

[2] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, et al. Pythia: A suite for analyzing large language models across training and scaling, 2023. URL https://arxiv.org/abs/2304.01373.

[3] Brasil. Lei n. 9.610, de 19 de fevereiro de 1998: altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências, February 1998. URL https://www.planalto.gov.br/ccivil_03/leis/l9610.htm. Article 8, item IV; published in the Diário Oficial da União on 20 February 1998. Accessed 26 August 2026.

[4] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

[5] Nicholas Kluge Corrêa, Sophia Falk, Shiza Fatimah, Aniket Sen, and Nythamar de Oliveira. TeenyTinyLlama: Open-source tiny language models trained in Brazilian Portuguese. *Machine Learning with Applications*, 16:100558, 2024. doi: 10.1016/j.mlwa.2024.100558.

[6] Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. Tucano: Advancing neural text generation for Portuguese. *Patterns*, 2025. doi: 10.1016/j.patter.2025.101325.

[7] Nicholas Kluge Corrêa, Aniket Sen, Shiza Fatimah, Sophia Falk, Lennard Landgraf, Julia Kastner, and Lucie Flek. Tucano 2 Cool: Better open source LLMs for Portuguese, 2026. URL https://arxiv.org/abs/2603.03543.

[8] Maria Clara Ramos Morales Crespo, Maria Lina de Souza Jeannine Rocha, Mariana Lourenço Sturzeneker, et al. Carolina: A general corpus of contemporary Brazilian Portuguese with provenance, typology and versioning information, 2023. URL https://arxiv.org/abs/2303.16098.

[9] Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.235. URL https://aclanthology.org/2023.acl-long.235/.

[10] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, et al. OLMo: Accelerating the science of language models, 2024. URL https://arxiv.org/abs/2402.00838.

[11] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

[12] Wanchao Liang, Tianyu Liu, Less Wright, Will Constable, Andrew Gu, Chien-Chin Huang, Iris Zhang, Wei Feng, Howard Huang, Junjie Wang, Sanket Purandare, Gokul Nadathur, and Stratos Idreos. TorchTitan: One-stop PyTorch native solution for production ready LLM pre-training, 2024. URL https://arxiv.org/abs/2410.06511.

[13] Ricardo Lopes, Joao Magalhaes, and David Semedo. GlórIA: A generative and open large language model for Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese – Vol. 1*, pages 441–453, Santiago de Compostela, Galicia/Spain, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.propor-1.45/.

[14] Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, et al. EuroLLM-9B: Technical report, 2025. URL https://arxiv.org/abs/2506.04079.

[15] NLLB Team, Marta R. Costa-jussà, et al. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846, 2024. doi: 10.1038/s41586-024-07335-x. URL https://doi.org/10.1038/s41586-024-07335-x.

[16] NOVA Vision & Language. CALAME-PT: Context-aware language modeling evaluation for portuguese, 2023. URL https://huggingface.co/datasets/NOVA-vision-language/calame-pt. Dataset card; MIT License; accessed 26 August 2026.

[17] Open Language Data Initiative. FLORES+: Multilingual machine translation benchmark, n.d. URL https://huggingface.co/datasets/openlanguagedata/flores_plus. Living dataset; version 4.6, 230 language varieties, CC BY-SA 4.0; accessed 26 August 2026.

[18] OpenAI. MMMLU, 2024. URL https://huggingface.co/datasets/openai/MMMLU. Professionally translated MMLU test set; includes PT-BR; MIT License; accessed 26 August 2026.

[19] Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. Sabiá: Portuguese large language models, 2023. URL https://arxiv.org/abs/2304.07880.

[20] Polygl0t. IFEval-PT, 2025. URL https://huggingface.co/datasets/Polygl0t/IFEval-PT. Portuguese translation of IFEval; Apache License 2.0; accessed 26 August 2026.

[21] Polygl0t. GSM8K-PT, 2026. URL https://huggingface.co/datasets/Polygl0t/gsm8k-pt. Portuguese translation of GSM8K; MIT License; accessed 26 August 2026.

[22] Portal da OBMEP. Uso do material do portal, n.d. URL https://portaldaobmep.impa.br/index.php/site/usar. Undated. Links its Creative Commons 3.0 notice to CC BY 3.0 BR. Accessed 26 August 2026.

[23] Programa SciELO. Critérios, política e procedimentos para a admissão e a permanência de periódicos na coleção scielo brasil, September 2024. URL https://www.scielo.br/media/files/20240900-Criterios-SciELO-Brasil.pdf. September 2024 version; especially Sections 2.3 and 5.2.10.1. Accessed 26 August 2026.

[24] Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, et al. Apertus: Democratizing open and compliant LLMs for global language environments, 2025. URL https://arxiv.org/abs/2509.14233.

[25] Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206, Pisa, Italy, September 2017. Linköping University Electronic Press. URL https://aclanthology.org/W17-6523/.

[26] Repositório ProEdu do MEC. Termos de uso do repositório proedu, November 2020. URL https://proedu.rnp.br/page/terms. Effective November 2020; requires users to verify the license assigned to each resource. Accessed 26 August 2026.

[27] Hélio Fonseca Sayama, Anderson Viçoso de Araujo, and Eraldo Rezende Fernandes. FaQuAD: Reading comprehension dataset in the domain of brazilian higher education. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 443–448. IEEE, 2019. doi: 10.1109/BRACIS.2019.00084. URL https://doi.org/10.1109/BRACIS.2019.00084.

[28] Universal Dependencies. UD Portuguese Bosque, n.d.. URL https://github.com/UniversalDependencies/UD_Portuguese-Bosque. Treebank repository; CC BY-SA 4.0; accessed 26 August 2026.

[29] Universal Dependencies. UD Portuguese PUD, n.d.. URL https://github.com/UniversalDependencies/UD_Portuguese-PUD. Parallel test-only treebank; CC BY-SA 3.0; accessed 26 August 2026.

[30] Wikimedia Foundation. Wikimedia Foundation Terms of Use, June 2023. URL https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use/en. Effective 7 June 2023; Section 7, Licensing of Content. Accessed 26 August 2026.

[31] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.