CulturaQuant: Quantization and Brazilian Cultural Knowledge in Small Language Models

Cristhian Kapelinski[1], **Dionatan Schmidt**[1],
Aline Lunkes[1], **Diego Kreutz**[1]

[1]AI Horizon Labs, Federal University of Pampa (UNIPAMPA)

cristhianavila.aluno@unipampa.edu.br
{dionatanschmidt,alinelunkes,diegokreutz}@unipampa.edu.br

Abstract. *Quantization compresses small language models but erodes the long tail of rare knowledge. For a Brazilian deployment, that tail includes regional cultural factors, so a practitioner choosing a bit width needs to know whether any loss follows from the fact that it is Brazilian or merely from its being rare. We present* CULTURAQUANT*, a multiple-choice probe drawn from Wikidata that asks which region a Brazilian cultural figure or event originates from. Each item is paired with an equally rare cultural fact from another country, so the comparison isolates Brazilian origin rather than rarity. Across 11 open models at 8-bit and 4-bit, measuring the 6 that answer above chance, the Brazilian items erode no faster at 8-bit: the pooled differential is +0.2 percentage points (pp) and every per-model interval crosses zero, excluding gaps beyond about 1.8 pp. At 4-bit, a penalty appears only in the Qwen2.5 family. The bit-width budget safe for rare facts is also safe for rare Brazilian facts.*

1. Introduction

Portuguese is the language spoken by more than 200 million people in Brazil.[1] The open generative models that serve them run both in the cloud and, increasingly, on local hardware, such as consumer GPUs, laptops, and on-premise servers. In either setting, the dominant cost is inference: serving a model at scale is memory-bound and expensive [Kwon et al. 2023], and on local hardware the model must fit within the available resources. This constraint has motivated the adoption of small language models (SLMs), open models with roughly 0.5 to 9 billion parameters, for on-device and edge use [Nguyen et al. 2025]. Both challenges point to the same solution: compressing the model so it needs less memory and compute. A common such method is post-training weight quantization. Casting 16-bit weights to 8-bit integers after training halves the memory [Dettmers et al. 2022]; at 4 bits, which we also evaluate, the saving is larger.

Quantization of Portuguese-capable language models is an active research area: recent work studies how 8-bit floating-point quantization affects activation outliers in English versus Brazilian Portuguese [Silva et al. 2026], the translation performance of dozens of quantized SLMs at 8-bit and 4-bit [Tamiosso et al. 2026], and quantized models answering the Brazilian secondary-school exam on consumer hardware [Santos and Campelo 2023]. The Brazilian evaluation ecosystem for assessing these models has grown quickly, from broad Portuguese benchmark suites [Almeida et al. 2025b] to a multilingual benchmark of notable events that includes Brazil [Almeida et al. 2025c]. The models

1. IBGE, 2022 Population Census, 2023

deployed to the edge are both Portuguese-capable and quantized, making it a practical question to determine what knowledge they retain after compression.

Compression methods are evaluated almost entirely by a single metric: aggregate accuracy on multiple-choice benchmarks, scored using constrained log-likelihood [Wang et al. 2025a, Lee et al. 2025], which obscures the types of knowledge a compressed model retains and loses. The concern is justified, because the mechanism is established: compression leaves top-line accuracy nearly flat while degrading the underrepresented long tail of atypical examples [Hooker et al. 2019, Hooker et al. 2020], and pretraining frequency governs which facts a model holds, implying that rare facts are lost first during compression [Chang et al. 2024, He et al. 2025].

For a Brazilian deployment, part of this rare tail is regional cultural knowledge: which state an artist comes from, which region hosts a festival. Throughout, we operationalize that as the region of origin of Brazilian cultural figures and events, a deliberately narrow proxy that captures geography rather than cultural practice or meaning. Cultural knowledge sits in the long tail of what a model holds [Badhe et al. 2026], compression degrades that tail [Hooker et al. 2019, Hooker et al. 2020], and quantization affects languages unequally [Marchisio et al. 2024], so the worry that it erases cultural facts first is well-founded.

That worry hides two questions, and they lead to opposite advice. Cultural facts are rare, and rare facts erode under compression. If they erode because they are Brazilian, a deployment here requires more care than one in English. If they erode only because they are rare, the bit width that is safe for any rare fact is safe here too. Prior cross-lingual work identifies the tension without resolving it, even disagreeing on whether quantization disproportionately harms lower-resource languages [Marchisio et al. 2024, Borgersen and Goodwin 2025], and neither side compares rare Brazilian facts with equally rare facts from elsewhere.

We draw that comparison with CULTURAQUANT, an openly released multiple-choice probe of rare Brazilian cultural facts in Portuguese, with each item paired with a non-Brazilian fact matched for rarity, so that any systematic difference between the groups reflects cultural content rather than the rarity of the facts. Across the 6 measurable models of the 11 evaluated, at 8-bit and 4-bit, cultural knowledge loses no more accuracy under quantization than the rarity-matched control: the pooled 8-bit gap is +0.2 percentage points (pp), every per-model interval crosses zero, and the design is powered to detect a gap larger than about 1.8 pp. For a Brazilian deployment, cultural facts warrant no more caution than any equally rare fact.

Contributions. We establish that, across the 6 measurable models at 8-bit, Brazilian cultural facts erode no faster than equally rare cultural facts from other countries (pooled +0.2 pp, any gap beyond about 1.8 pp excluded), with a cultural penalty at 4-bit confined to the Qwen2.5 family (Section 5). To make this comparison possible, we release a probe of rare Brazilian cultural facts in Portuguese, each matched item by item on rarity with a non-Brazilian cultural fact, drawn from Wikidata with full provenance and tagged by macro-region, under an open license (Section 3), and from it we provide a safe bit-width recommendation for practitioners (Section 6). We also relate the rise of cultural accuracy with popularity to a broader concern: cultures even less represented, such as Brazil's

Indigenous languages, may fall below chance already at full precision, a representation limit rather than a bit-width one (Section 6). The probe, with its scoring and analysis code, is the artifact we release.[2]

2. Related Work

Table 1 positions the directly comparable approaches on the axes that decide whether a prior result answers our question: *what each work measures*, the *language* it covers, whether it studies *quantization*, and the released *artifact* (D dataset, B benchmark, C code, N none). None of these works includes a control matched on rarity, which is what separates a cultural effect from a frequency effect, so none can on its own tell a practitioner whether quantization erases Brazilian cultural facts faster than equally rare non-Brazilian ones; that control is the gap this work closes. Building blocks and surveys that inform the design but do not compete on these axes stay in prose.

Table 1. Positioning against comparable approaches; (✔) marks an axis touched but not fully provided.

Work	What it measures	Lang.	Quant.	Artifact
Quantization and knowledge type				
[Wang et al. 2025b]	Factual recall	EN	✔	N
[Zhou et al. 2025]	Cognitive type	EN	✔	N
[Lee et al. 2025]	Task difficulty	EN	✔	C
[Rath and Maliakkal 2026]	Social bias	EN	✔	C
Multilingual and Portuguese quantization				
[Marchisio et al. 2024]	Translation	Multi.	✔	N
[Borgersen and Goodwin 2025]	Multilingual QA	Multi.	✔	N
[Silva et al. 2026]	LM fidelity	EN/PT-BR	(✔)	N
Cultural knowledge				
[Singh et al. 2025]	Cult. vs. agnostic	Multi.	✘	B
CULTURAQUANT (this work)	**BR vs. non-BR, rarity-matched**	**PT-BR**	✔	**D,C**

Quantization and knowledge type. Three results together set up our question. First, erosion is not uniform across facts: calibration-free integer casting preserves recall best, smaller models degrade more, and facts the full-precision model never mastered degrade further still [Wang et al. 2025b]. Second, the bit width matters in a step rather than a slope, since 8-bit is near-lossless while 4-bit markedly degrades knowledge-intensive tasks [Wang et al. 2025a, Mekala et al. 2025]. Third, erosion is uneven by knowledge type, whether the strata are cognitive [Zhou et al. 2025] or defined by a model's existing weaknesses [Lee et al. 2025]. All of this is English, and the strata are complexity levels within a capability rather than where a fact comes from. Closest in method is an item-level paired study of three of our model families, which finds that quantization shifts behavior that aggregate metrics hide [Rath and Maliakkal 2026]; we borrow that design and point it at factual cultural knowledge against a rarity-matched control rather than at bias.

Multilingual and Portuguese quantization. The multilingual evidence converges on two points and stops just short of ours. Harm is real but unevenly measured: automatic metrics understate what human raters perceive, and smaller models suffer most [Marchisio et al. 2024]. Yet when disproportionate multilingual harm is tested rather than assumed, it does not appear [Borgersen and Goodwin 2025]. Neither study reaches culture, since both

2. https://github.com/CristhianKapelinski/culturaquant-artifact

use geography-free prompts above the edge band, and the second evaluates one model on two languages. For Portuguese, the English-versus-Portuguese gap traces to naive FP8 casting, while integer quantization with outlier handling preserves both languages [Silva et al. 2026]; that is why we cast to integers. Brazilian work has since shown that quantized models can be evaluated on exam pools and on text classification [Santos and Campelo 2023, Jodas et al. 2024], that the calibration language matters [Chimoto et al. 2026], and that 4-bit harm is concentrated in low-capacity models [Tamiosso et al. 2026], whose non-parametric protocol we adopt. What none of them adds is an axis for the kind of knowledge at stake.

Cultural knowledge and the long tail. That compression forgets the long tail is established for vision [Hooker et al. 2019, Hooker et al. 2020] and, for specialized knowledge under quantization in language, is now repeatedly reported [Wang et al. 2025b, Zhou et al. 2025]; what remains open is whether Brazilian cultural facts are lost faster than equally rare facts of the same kind from elsewhere. Two studies show that pretraining frequency, rather than the surface difficulty of a probe, governs which facts a model holds, and that a knowledge-graph rarity proxy is an imperfect surrogate for true corpus frequency [Chang et al. 2024, He et al. 2025], which both motivates matching the control on rarity and bounds how precisely that match can be made. Global MMLU establishes a cultural versus agnostic split at full precision across many languages [Singh et al. 2025], the same content axis that we carry under quantization with a rarity-matched control and Brazilian macro-region tags, while a cultural-literacy benchmark probes factual cultural knowledge across nine countries at full precision [Tanwar et al. 2025], a line our Brazilian, quantization-focused probe complements. A multilingual proverb evaluation warns that answer-order bias can substantially shift multiple-choice scores [Azime et al. 2025], which we control for by rotating the order of options.

Brazilian evaluation. Table 2 lists the Portuguese evaluation resources CULTURAQUANT builds on, the Role column marking a control source (C) or a design reference (R). None of them combines cultural knowledge, quantization, and a rarity-matched control. Our items and controls come from Wikidata, whose statements give a verifiable answer and a sitelink rarity measure for every fact. The Brazilian items are newly drawn rather than taken from exam pools, which reduces overlap with Portuguese pretraining corpora built from web crawls [Almeida et al. 2025a].

Table 2. Datasets referenced and the control source; Role R (referenced) or C (control).

Dataset	Size	Licence	Role
PoETa v2 [Almeida et al. 2025b]	44 tasks	varies	R
TiEBe [Almeida et al. 2025c]	23k QA	varies	R
BLUEX [Almeida et al. 2023]	1,260 MCQ	unstated	R
ALBA [Vieira et al. 2026]	800 items	unstated	R
Global MMLU [Singh et al. 2025]	14k MCQ	Apache 2.0	R
Wikidata	100M+ items	CC0	C

3. The CulturaQuant Probe

Figure 1 lays out the pipeline end to end, from matched-pair construction through cyclic scoring to differential erosion (the accuracy drop on cultural facts minus the drop on the control); this section details the probe and its control, and Section 4 the scoring and the statistics.

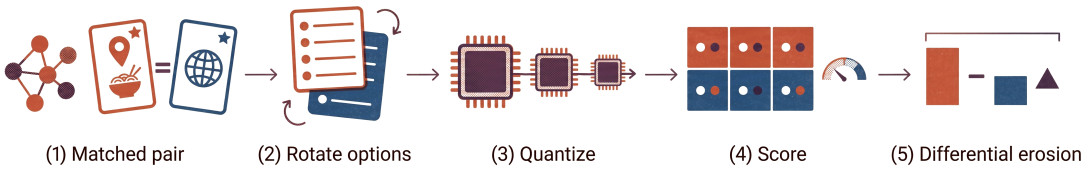


Figure 1. The CULTURAQUANT pipeline: matched-pair construction (each Brazilian cultural item paired one-to-one on rarity with a non-Brazilian control), cyclic constrained-log-likelihood scoring across the five option positions, and differential erosion (cultural drop minus control drop) computed per model and bit width.

We operationalize Brazilian cultural knowledge as the regional grounding of Brazilian cultural figures and events: which Brazilian state a notable artist, musician, or writer comes from, or where a regional festival is held. This proxy is narrow but verifiable: it captures the geography of cultural production, not practices or their meaning, a limit we return to in Section 6. CULTURAQUANT is a set of multiple-choice items in Brazilian Portuguese, each with five options and one correct answer, auto-graded by matching the chosen option against a fixed gold answer. Every item comes from a Wikidata statement, which gives a verified gold answer with no human authoring and a rarity signal we can match on. A cultural item asks in which Brazilian state an arts figure was born or a festival is held; the gold answer is that state and the four distractors are other Brazilian states. A wrong answer therefore reflects missing knowledge rather than an implausible option. One item asks, for the Calypso singer Joelma, in which state she was born, with gold answer *Pará*. Each item records its Wikidata identifier, supporting statement, region, and rarity, so the dataset remains fully traceable.

Having defined the cultural facts to test, we must ensure that the models actually hold them, because the design hinges on observing knowledge as it erodes. A fact the model never knew cannot be observed to erode, and a fact it always knows cannot either, so the probe must span the band where a full-precision model answers above chance but below ceiling. To locate that band, we quantify rarity by Wikidata sitelink count, the number of language editions that link to an entity. This public popularity proxy requires no access to a model's training data. Full-precision accuracy rises with this count, from 21.4% in the rarest band, at the five-way chance floor of 20%, to 52.1% in the most popular (Table 3). We therefore score the full gradient, 700 cultural items spread across seven rarity bands and the five macro-regions (N 105, NE 166, CO 113, SE 179, S 137), and rather than discard bands a priori, we report erosion per model only where that model clears chance; the per-model measurable range we make precise in Section 4. This gradient thus fixes the operating window the probe must cover; the control fixes the comparison group that isolates cultural content from rarity, and it is to that control that we now turn.

The control consists of 700 non-Brazilian facts of the same shape: an arts figure born in a first-level administrative division of Italy, Spain, Germany, the United Kingdom, or Argentina, with that division as the gold answer and sibling divisions as distractors. We rotate the source country, so the control is not tied to any single foreign culture. Note what this contrast is and is not. Both arms hold cultural facts about people and events; only the country differs. The comparison therefore isolates Brazilian origin, not cultural content against neutral content. We match the two arms on rarity using two public popularity proxies: sitelink count (an integer) and log pageviews. Sitelinks define the discrete rarity

bands of the gradient above. Pageviews are the signal on which each Brazilian item is paired one-to-one, without replacement, with a non-Brazilian item.

That pairing is looser than it looks, and we state plainly what it delivers. Pageviews are near zero for most of these entities, so 615 of the 700 pairs (88%) match at the floor, zero against zero, rather than at a shared level of traffic. Sitelinks carry the real rarity signal, and it is here that the two arms differ: the median gap within a pair is 4 sitelinks, and the marginal band profiles are visibly different (Table 3). The comparison rests on the design rather than on the proxy. Differential erosion is a within-pair quantity, each pair's Brazilian drop minus its own control drop, so rarity is held fixed pair by pair; and because erosion does not grow with rarity (Section 5), an imperfect match cannot bias it. Section 5 tests exactly this by recomputing the result on the tightly matched pairs alone.

4. Evaluation Setup

We evaluate 11 open models from 0.5 to 7 billion parameters, the consumer-deployment band, across several pretraining lineages (the Qwen2.5 and Qwen3 families, Phi-3.5, Mistral, and the Brazilian-pretrained Tucano models [Corrêa et al. 2025]). Each runs at full precision (FP16), 8-bit integer, and 4-bit (NF4, double quantization), employing calibration-free integer casting, the method prior work reports as comparatively recall-preserving [Wang et al. 2025b]. Scoring is deterministic constrained log-likelihood, the length-normalized log-probability over the five options with no sampling, which removes any generation-format confound and makes each precision pair a paired test on identical inputs. All scoring is containerized on one reference machine: an AMD Ryzen 7 9700X (8 cores, 16 threads), 64 GB of RAM, and a single NVIDIA GeForce RTX 5080 with 16 GB of memory (VRAM).

Erosion is hard to observe where aggregate accuracy sits at chance: a drop measured there is mostly noise, and would dilute or fabricate an effect. We therefore report differential erosion for the 6 models whose full-precision accuracy on the Brazilian items clears the five-way chance floor (the Wilson lower bound, a conservative interval for a proportion, above 20%), and treat the other five (Qwen2.5-0.5B, Qwen3-0.6B, Mistral-7B, and both Tucano models) as unmeasurable at this rarity. Aggregate chance does not prove a model holds nothing, since it can know some facts and miss others systematically; Section 5 therefore repeats the analysis over all 11 models and with a continuous margin that registers movement at the floor. Notably, both Brazilian-pretrained Tucano models remain at chance on the cultural probe, so Portuguese pretraining alone does not bring these facts within reach.

Language models tend to favor particular letters of the alphabet regardless of content, a bias that can skew multiple-choice scores [Azime et al. 2025]. We neutralize it by cyclic rotation, scoring every item in all 5 rotations that place the gold answer at each position in turn so that no score can reward a position preference.

Accuracies are paired across precisions on identical items, and the quantity of interest is differential erosion: the drop in cultural accuracy from full precision to a quantized precision, minus the same drop for the rarity-matched control. A positive value means Brazilian items erode faster than their controls, whereas zero means the two erode at the same rate. Because an item's rotations are correlated, we obtain the 95% confidence interval from an item-clustered bootstrap that resamples items rather than rotations, and,

alongside each estimate, we report the gap that the design can exclude.

5. Results

The central result is a bounded null at 8-bit and, at 4-bit, a penalty confined to the Qwen2.5 family; we develop each in turn and then subject both to a series of robustness checks. Figure 2 organizes the evidence into three panels: full-precision accuracy rising with popularity, the per-model differential erosion with its 95% intervals, and absolute cultural accuracy across precisions.

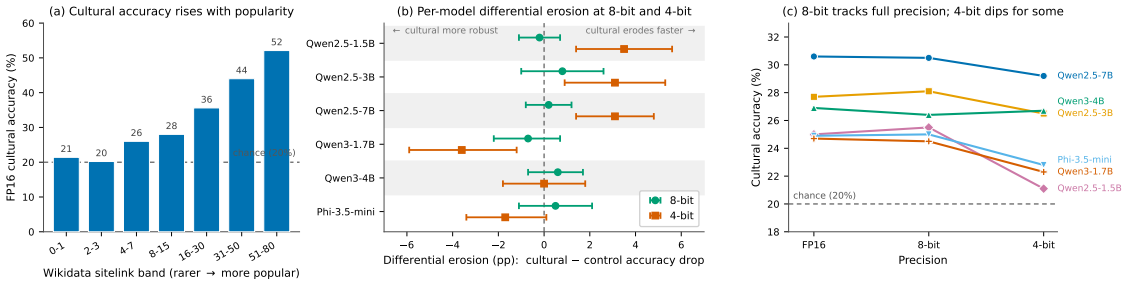


Figure 2. (a) full-precision accuracy by rarity band; (b) differential erosion per model and bit width, with 95% intervals; (c) cultural accuracy at full precision, 8-bit, and 4-bit.

Rarity sets that measurable band: full-precision cultural accuracy rises with popularity (Table 3, Figure 2(a)), so the rarest facts sit at chance and cannot erode while the most popular approach the ceiling. Five of the 11 models (Qwen2.5-0.5B, Qwen3-0.6B, Mistral-7B, and both Tucano models) never clear chance; that the 7B Mistral is among them shows size alone does not decide measurability.

Table 3. CULTURAQUANT composition by Wikidata rarity band: cultural and control item counts and pooled full-precision cultural accuracy.

Band (sitelinks)	0–1	2–3	4–7	8–15	16–30	31–50	51–80	All
Cultural items	153	150	146	123	78	36	14	700
Control items	372	55	33	13	57	107	63	700
FP16 accuracy (%)	21.4	20.2	26.0	28.0	35.6	44.0	52.1	–

Cultural knowledge does not erode faster at 8-bit. Table 4 reports, for the 6 measurable models, full-precision cultural and control accuracy and the differential erosion at each bit width with a 95% interval. At 8-bit, the pattern is uniform: every per-model interval crosses zero, the pooled differential is +0.2 pp [-0.4, 0.7], and the design excludes a pooled differential larger than about 1.8 pp in either direction (Figure 2(b)). Because absolute cultural accuracy barely moves under 8-bit, this is a bounded null rather than a coincidence of wide intervals: the design had the power to surface a cultural penalty of even modest size, and none appears. The measurable models hold cultural facts at 24.7% to 30.6% full-precision accuracy, a few to roughly ten points above the 20% chance floor, so the comparison is read over a modest but real band of dynamic range and the null reflects demonstrated equality within that band rather than absent signal.

The 4-bit effect is architecture-dependent. At 4-bit, the differential splits by architecture rather than vanishing into noise. All three Qwen2.5 models erode the Brazilian items faster than their controls, by 3.1 to 3.5 pp, with each interval excluding zero; thus, for this family, 4-bit produces a detectable cultural penalty rather than mere instability. We apply

Table 4. Differential erosion (pp) for the 6 measurable models; positive means cultural erodes faster, bold marks an interval excluding zero.

	FP16 acc (%)		Diff. erosion (pp)	
Model	Cult.	Ctrl	8-bit	4-bit
Qwen2.5-1.5B	25.0	20.8	-0.2 [-1.1,0.7]	**+3.5** [1.4,5.6]
Qwen2.5-3B	27.7	22.2	+0.8 [-1.0,2.6]	**+3.1** [0.9,5.3]
Qwen2.5-7B	30.6	23.8	+0.2 [-0.8,1.2]	**+3.1** [1.4,4.8]
Qwen3-1.7B	24.7	29.1	-0.7 [-2.2,0.7]	**-3.6** [-5.9,-1.2]
Qwen3-4B	26.9	20.6	+0.6 [-0.7,1.7]	0.0 [-1.8,1.8]
Phi-3.5-mini	24.9	19.9	+0.5 [-1.1,2.1]	-1.7 [-3.4,0.1]
Pooled			+0.2 [-0.4,0.7]	+0.7 [-0.1,1.6]

no formal multiple-comparison correction across the 6 models and two bit widths, but the Qwen2.5 finding does not rest on a single test, since three models in the family agree in sign and magnitude. That agreement is suggestive rather than decisive: we run no test at the level of the family itself, so the effect should be read as one that a direct family-level comparison still has to confirm. The remaining models do not share it: Qwen3-1.7B erodes them more slowly, by -3.6 pp, while Qwen3-4B and Phi-3.5 register no effect (Table 4). Pooled across the measurable models, the differential is +0.7 pp [-0.1, 1.6] and still crosses zero, so 4-bit is not uniformly harmful; the cultural penalty is established for the Qwen2.5 family and unestablished elsewhere. Figure 2(c) shows this in absolute terms: the 8-bit point sits on the full-precision point for every model, while at 4-bit the cultural accuracy of the Qwen2.5 models dips and the others hold.

Three checks probe the design choices most likely to have produced the 8-bit null, and each of their intervals still crosses zero. Restricting to the 286 pairs whose sitelink counts differ by at most two, the tightest match the data allow, gives -0.4 pp [-1.3, 0.4]. Scoring all 11 models rather than the 6 measurable ones, which removes any selection on full-precision performance, gives +0.0 pp [-0.3, 0.4]; we mark this exploratory, since accuracy near the floor carries little signal. Replacing binary accuracy with the margin between the correct option's log-probability and the best incorrect one's, which registers movement even at the floor, gives -0.006 nats.

Both findings also hold up under further robustness checks. The first alternative is that quantization erases knowledge simply because it is rare, in which case the rarest items would be eroded the most; they are not (Table 5). Pooled over the 6 measurable models, the 8-bit cultural drop stays within -1.0 to 1.3 pp across all seven rarity bands with no trend toward the rarest, and the 4-bit drop, though larger, is scattered from 0.0 to 5.9 pp and peaks in a mid-rarity band rather than the rarest. Even within Brazilian cultural knowledge, the small 4-bit loss is not concentrated on the rarest facts as a purely frequency-driven erosion would require.

Table 5. Absolute cultural-accuracy drop (pp) per Wikidata rarity band, pooled over the 6 measurable models, not the control-adjusted differential of Table 4.

Band (sitelinks)	0–1	2–3	4–7	8–15	16–30	31–50	51–80
Cultural items	153	150	146	123	78	36	14
8-bit drop (pp)	-0.5	-0.5	+0.5	+0.4	-0.2	+1.3	-1.0
4-bit drop (pp)	+1.5	0.0	+2.1	+3.1	+1.5	+5.9	+3.8

A second alternative is that erosion concentrates geographically, with some Brazilian macro-region forgotten first. Per-region cells of a few dozen items are too small for a stable estimate, so we group the five macro-regions into three: North-plus-Northeast, Centro-Oeste, and Southeast-plus-South. Their 8-bit drops run from -0.7 to +0.3 pp, and their 4-bit drops are +1.1, -0.1, and +3.2 pp, respectively; because the 8-bit differences are within noise, no region shows a stable, bit-width-consistent ordering. A final alternative is that the differential is an artifact of how the models distribute their answers across option letters. The cyclic rotation already makes the gold answer uniform over the five positions by construction, so a differential cannot arise from a position preference. A direct check confirms it: the spread of chosen positions (RStd) stays small at full precision, from 0.12 to 0.20 across the measurable models against zero for a uniform chooser, and it grows under 4-bit in only one model (Qwen3-1.7B, to 0.33), which is not the model that drives the 4-bit cultural differential. None of the three alternatives accounts for the results, leaving the bounded 8-bit null and the architecture-specific 4-bit penalty as the standing reading.

6. Discussion

The most immediate consequence is a concrete recommendation for practice. For the 6 measurable models, 8-bit integer quantization is the safe ship point: Brazilian cultural facts erode no faster than equally rare non-Brazilian ones, and the absolute loss on the Brazilian items stays within the noise of the control (Table 4). A Brazilian deployment therefore does not need a smaller bit width than an English one to protect cultural facts, aligning with the 8-bit-safe, 4-bit-breaks ladder reported for English knowledge tasks [Wang et al. 2025a, Mekala et al. 2025]. At 4-bit, the guidance must become conditional, because the picture splits by architecture: the Qwen2.5 family exhibits a detectable cultural penalty, with each interval excluding zero, whereas for the other models the effect is not established. A practitioner shipping a Qwen2.5 model at 4-bit should therefore verify accuracy on their own items before trusting it. For the other architectures we neither found the effect nor ruled it out, so the safe reading is that 4-bit deserves a check anywhere, and that our guidance is firm only at 8-bit. A calibration-based method or a model outside our set would extend the guidance [Chimoto et al. 2026].

The 4-bit method is not a neutral implementation detail. At these model sizes, whether a method tunes its rounding on a sample of text changes how much rare training content the compressed model reproduces verbatim [Kapelinski and Kreutz 2026]. That is a different question from ours, but it shows the method choice reaching what a compressed model retains of rare material, which is what we measure here.

The same data carry a warning about what lies beyond the band the probe can measure. Even among nationally prominent Brazilian figures, full-precision accuracy rises with popularity, from 21.4% in the rarest band to 52.1% in the most popular (Figure 2(a)). Read forward, that gradient is a caution for cultures far less represented than these. Brazil's Indigenous peoples and languages, Tikuna with on the order of fifty thousand speakers and Nheengatu with around twenty thousand, occupy the extreme low-resource end with almost no support from digital tools [Pinhanez et al. 2024], a scarcity that pretraining corpora inherit [Joshi et al. 2020] and that models reflect on Indigenous American languages [Ebrahimi et al. 2022] and underrepresented cultures [Chiu et al. 2025]. The intuition that the cultural knowledge of smaller peoples should erode first under compression, because it is rarer, is plausible. Yet, our results reframe where the risk actually lies:

within the band we can measure, erosion does not grow with rarity, while full-precision accuracy does fall with it. Extrapolated to the rarest cultures, a model would sit at or below the chance floor already at full precision, so for those peoples the binding problem is not faster erosion under quantization but that the knowledge was never acquired, a representation gap that sits upstream of any bit width.

This reframing carries an ethical weight. Lossy compression of language models can, in principle, quietly and unequally erase the cultural representations of communities whose languages and knowledge are underrepresented in pretraining corpora. Our finding that erosion is governed by rarity rather than by the source of a fact is reassuring. Still, it does not eliminate the risk because cultural facts are rare. A fact's rarity is not an accident but a consequence of who dominates digital content production, so treating rarity-driven erosion as neutral is itself a choice. Our probe, moreover, covers only facts verifiable on Wikidata, so intangible cultural practices, oral expression, and undocumented knowledge fall outside it and may be the most vulnerable. We therefore recommend that quantization evaluations for regional deployments include cultural items as a routine check, in the same way that fairness metrics already are.

7. Limitations and Conclusion

Several limitations qualify these conclusions. The result is a bound rather than a proof of exact equality, since the intervals exclude a differential larger than about 1.8 pp but cannot exclude a smaller one. The match rests on a rarity proxy rather than on measured corpus frequency, which no public tool provides for these models [He et al. 2025], and we did not record which Wikimedia projects, languages, or time window the pageview counts came from. That provenance cannot now be reconstructed, and future work should record it at the time of collection. Residual frequency differences therefore cannot be ruled out, though the tight-pair check above bounds how much they could matter. The two arms are also not regionally balanced: the cultural items are unevenly spread across macro-regions, with Southeast and Northeast dominant and Centro-Oeste holding only 113 items, while the control draws mostly from two countries. The control is moreover matched on rarity but not on full-precision accuracy, which differs across models (for example, Qwen3-1.7B at 24.7% cultural against 29.1% control), so differing baselines could in principle bias the difference of drops through floor effects, the arm with more room above the floor having more room to fall. Because every measurable accuracy is only a few points above the 20% floor, the room for such distortion is small, not zero. Finally, the study covers one quantization family and one scoring mode on a single 16 GB GPU, so the upper deployment band and calibration-based methods remain untested.

The central question can now be answered. The introduction asked whether Brazilian cultural facts erode faster under quantization than equally rare facts from elsewhere, because if they did, a Brazilian deployment would need a more careful bit width than an English one. At 8-bit, the answer is no: across the measurable models we find no statistically significant cultural penalty, and the design bounds any cultural effect to a negligible margin of about 1.8 pp. At 4-bit, this conclusion holds for most architectures but not for the Qwen2.5 family, whose cultural penalty a deployment should verify directly. What erodes under quantization is governed mostly by how rare a fact is rather than by where it comes from, so a Brazilian deployment can otherwise reuse the same 8-bit budget that is safe for rare knowledge anywhere.

Two limits belong in the answer itself, because they bound what it means. The probe asks a model to rank five options for a fact Wikidata can verify, so it measures recall, not production. Cultural competence in a wider sense, idiomatic Portuguese, register, and reasoning about a practice rather than looking up where it happens could behave differently under compression. The control, in turn, contains cultural facts about other countries rather than culturally neutral material, so the comparison isolates Brazilian origin rather than cultural content as such. The result is therefore a bound on the differential recall of verifiable facts. That is what a practitioner choosing a bit width needs, and it does not guarantee that compression preserves culture. Three extensions follow from what the probe cannot yet reach. Calibration-based methods are the first, since they are what a deployer actually reaches for at 4 bits and our result does not transfer to them. The second is a probe that spans a wider above-chance band, which would allow pairs to be matched with full-precision accuracy as well as by rarity and would remove the floor effect that limits the margin analysis. The third is the harder one: an instrument for cultural competence rather than factual recall, built on free generation judged for idiom and register, which is the only way to test the claim this paper deliberately does not make. The released probe, with its rarity-matched control and macro-region tags, is the starting point for all three. For Brazilian institutions that place SLMs on local hardware, this turns bit-width selection into a frequency question they can audit using a popularity proxy, rather than a separate cultural risk to mitigate. At the same time, the harder exposure lies upstream, with people whose knowledge of full precision never acquired.[3]

References

Almeida, T. S. et al. (2023). BLUEX: A benchmark based on brazilian leading universities entrance exams. In *BRACIS*.

Almeida, T. S. et al. (2025a). Building high-quality datasets for Portuguese LLMs: From common crawl snapshots to industrial-grade corpora. *arXiv:2509.08824*.

Almeida, T. S. et al. (2025b). PoETa v2: Toward more robust evaluation of LLMs in portuguese. *arXiv:2511.17808*.

Almeida, T. S. et al. (2025c). TiEBe: Tracking language model recall of notable worldwide events through time. *arXiv:2501.07482*.

Azime, I. A. et al. (2025). ProverbEval: Exploring LLM evaluation challenges for low-resource language understanding. In *Findings of NAACL*.

Badhe, S. et al. (2026). Long-tail knowledge in large language models: Taxonomy, mechanisms, interventions and implications. *arXiv:2602.16201*.

Borgersen, K. A. and Goodwin, M. (2025). English k-quantization of LLMs does not disproportionately diminish multilingual performance. *arXiv:2503.03592*.

Chang, H. et al. (2024). How do large language models acquire factual knowledge during pretraining? In *NeurIPS*.

Chimoto, E. A. et al. (2026). Calibrating beyond english: Language diversity for better quantized multilingual LLMs. In *EACL*.

Chiu, Y. Y. et al. (2025). CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-ai red-teaming. In *ACL*.

3. **Acknowledgments.** This work was partially supported by Capes, CNPq (Grant No. 409743/2025-9), and FAPERGS (Grant Nos. 24/2551-0001368-7 and 24/2551-0000726-1). We thank the anonymous reviewers of ENIAC 2026 for their suggestions and comments. The authors also acknowledge the use of Generative AI as a supporting tool for reviewing and improving the code and manuscript, and for generating the schematic pipeline figure (Figure 1); they verified every number and take responsibility for all content.

Corrêa, N. K. et al. (2025). Tucano: Advancing neural text generation for portuguese. *Patterns*.

Dettmers, T. et al. (2022). LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *NeurIPS*.

Ebrahimi, A. et al. (2022). AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *ACL*.

He, Y. et al. (2025). Supposedly equivalent facts that aren't? entity frequency in pre-training induces asymmetry in LLMs. In *COLM*.

Hooker, S. et al. (2019). What do compressed deep neural networks forget? *arXiv:1911.05248*.

Hooker, S. et al. (2020). Characterising bias in compressed models. *arXiv:2010.03058*.

Jodas, D. S. et al. (2024). Impact of quantization on large language models for Portuguese classification tasks. In *CIARP*.

Joshi, P. et al. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *ACL*.

Kapelinski, C. and Kreutz, D. (2026). Not all 4-bit quantizers are equal: Deployment-time mitigation of PII leakage in fine-tuned small language models. In *SBSeg*.

Kwon, W. et al. (2023). Efficient memory management for large language model serving with PagedAttention. In *SOSP*.

Lee, J. et al. (2025). Exploring the trade-offs: Quantization methods, task difficulty, and model size in large language models from edge to giant. In *IJCAI*.

Marchisio, K. et al. (2024). How does quantization affect multilingual LLMs? In *Findings of EMNLP*.

Mekala, A. et al. (2025). Does quantization affect models' performance on long-context tasks? In *EMNLP*.

Nguyen, C. V. et al. (2025). A survey on small language models. In *RANLP*.

Pinhanez, C. et al. (2024). Harnessing the power of artificial intelligence to vitalize endangered indigenous languages: Technologies and experiences. *arXiv:2407.12620*.

Rath, P. K. and Maliakkal, R. (2026). Quantization undoes alignment: Bias emergence in compressed LLMs across models and precision levels. *arXiv:2605.15208*.

Santos, M. L. O. and Campelo, C. E. C. (2023). Benchmarking quantized LLaMa-based models on the brazilian secondary school exam. In *CBIC*.

Silva, G. et al. (2026). Lost in quantization: Activation outliers explain language-specific FP8 sensitivity in Llama-3. In *PROPOR*.

Singh, S. et al. (2025). Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *ACL*.

Tamiosso, G. L. et al. (2026). Evaluating small language models for english-to-portuguese translation: Impact of model scale and quantization. In *PROPOR*.

Tanwar, E. et al. (2025). Do you know about my nation? investigating multilingual language models' cultural literacy through factual knowledge. In *EMNLP*.

Vieira, I. et al. (2026). ALBA: A european portuguese benchmark for evaluating language and linguistic dimensions in generative LLMs. In *PROPOR*.

Wang, J. et al. (2025a). SLMQuant: Benchmarking small language model quantization for practical deployment. In *ACM Multimedia Workshops*.

Wang, Q. et al. (2025b). Through a compressed lens: Investigating the impact of quantization on factual knowledge recall. *arXiv:2505.13963*.

Zhou, C. et al. (2025). Task-stratified knowledge scaling laws for post-training quantized large language models. *arXiv:2508.18609*.